Position: A Call to Action for a Human-Centered AutoML Paradigm

Marius Lindauer [* 1 2]   **Florian Karl** [* 3 4 5]   **Anne Klier** [3]   **Julia Moosbauer** [4 5]   **Alexander Tornede** [1]
Andreas Mueller [6]   **Frank Hutter** [7]   **Matthias Feurer** [4 5]   **Bernd Bischl** [4 3 5]

Abstract

Automated machine learning (AutoML) was formed around the fundamental objectives of automatically and efficiently configuring machine learning (ML) workflows, aiding the research of new ML algorithms, and contributing to the democratization of ML by making it accessible to a broader audience. Over the past decade, commendable achievements in AutoML have primarily focused on optimizing predictive performance. This focused progress, while substantial, raises questions about how well AutoML has met its broader, original goals. In this position paper, we argue that a key to unlocking AutoML's full potential lies in addressing the currently underexplored aspect of user interaction with AutoML systems, including their diverse roles, expectations, and expertise. We envision a more human-centered approach in future AutoML research, promoting the collaborative design of ML systems that tightly integrates the complementary strengths of human expertise and AutoML methodologies.

1. Introduction

Over the last decade, Automated machine learning (AutoML, see Hutter et al., 2019; Bergstra & Bengio, 2012; Snoek et al., 2012; Thornton et al., 2013; Escalante, 2021) has proven its potential to improve machine learning (ML) systems by automating parts of the data science workflow, in particular the selection and configuration of pipelines of ML algorithms of various sorts, by providing new and efficient hyperparameter optimization (HPO) procedures (Feurer & Hutter, 2019; Bischl et al., 2023), neural architecture search (NAS) methods (Elsken et al., 2019; White et al., 2023) and the construction of powerful ensembles (Erickson et al., 2020). AutoML success stories are numerous – to name a few of them: the substantial contribution of AutoML to AlphaGo (Chen et al., 2018); many AutoML systems with a total of over 100.000 downloads each month, e.g., AutoGluon (Erickson et al., 2020), Auto-Sklearn (Feurer et al., 2022), Auto-Weka (Thornton et al., 2013), Auto-Prognosis (Alaa & van der Schaar, 2018), SMAC (Hutter et al., 2011; Lindauer et al., 2022); the routine usage of hardware-aware NAS for automatic design of neural architectures with hardware constraints in industry (Benmeziane et al., 2021); learned optimizers like LION (Chen et al., 2023); the learned Swish activation function (Ramachandran et al., 2018); learned data augmentation strategies (Cubuk et al., 2019); and prior-fitted networks (PFNs) for learning classification algorithms (Hollmann et al., 2023a). Because of that, AutoML research has grown rapidly over the last years, probably most evident in NAS (Elsken et al., 2019; White et al., 2023). At the same time, most big IT companies have developed large software packages enabling AutoML, including Google (Golovin et al., 2017; Song et al., 2022), Amazon (Erickson et al., 2020), Meta (Balandat et al., 2020), IBM (Wang et al., 2020), Oracle (Yakovlev et al., 2020) and Microsoft (Wang et al., 2021a).

Despite these successes, after more than a decade of research on AutoML, it is time to reflect on whether the AutoML community has achieved its original goals, whether those goals really addressed all the needs of all the targeted user groups in the first place and what is currently missing. AutoML, in its current form, arguably aims at (i) accelerating the development of well-performing ML pipelines in applications by effectively addressing the problems of model selection and configuration (incl. neural architectures and hyperparameters); (ii) supporting research on new ML algorithms by automatically configuring the entire system and thus building the best possible system – but also providing strong and appropriate baseline comparisons via essentially the same mechanism; (iii) contributing to the democratization of ML for domain experts with little to no ML expertise.

Although many important challenges remain (e.g., regarding

[*]Equal contribution   [1]Institute of Artificial Intelligence (LUH|AI), Leibniz University Hannover, Germany [2]L3S Research Center, Hannover, Germany [3]Fraunhofer Institute for Integrated Circuits IIS, Fraunhofer IIS, Nuremberg, Germany [4]Ludwig-Maximilians-Universität München, Munich, Germany [5]Munich Center for Machine Learning, Munich, Germany [6]Microsoft, Redmond, USA [7]Albert-Ludwigs-Universität Freiburg, Freiburg, Germany. Correspondence to: Marius Lindauer <m.lindauer@ai.uni-hannover.de>, Florian karl <florian.karl@iis.fraunhofer.de>.

the scaling of AutoML to large foundation models, or the expressiveness of NAS methods), the potential and partial success of the first two goals was shown in many studies (Chen et al., 2018; Ramachandran et al., 2018; Guyon et al., 2019; Erickson et al., 2020, and many more, see the references above), especially if we define "performance" narrowly in terms of "efficiently optimizing predictive performance".

Nevertheless, one aspect that has not been sufficiently considered in large parts of AutoML research is that there are several user groups that could benefit from AutoML, each of which has very different needs and expectations: First, there are domain experts who would mainly like to communicate their general goals and domain knowledge to the AutoML system. They are typically very interested in understanding the final model – or rather, a population-level understanding regarding their task that allows valid inferences about general relationships. Then, there are ML practitioners and data scientists who deal with deeper and more technical and mathematical issues in applied model building. They usually like being in control of the ML process but want to automate away repetitive and mundane work (which human experts are less good at anyways), e.g., technical aspects of model selection (but not all of them) and especially HPO. Last but not least, there are ML researchers who focus on developing new ML approaches (and underlying theories). They nearly always need to be in full control and usually care much more about the effectiveness of the ML components; so explanations of HPO are more relevant to them than interpretations of the final models. AutoML researchers or experts interact with an AutoML system mainly during its development and setup but could be considered an additional user group. They are mostly interested in information and visualizations to analyze the performance and behavior of AutoML systems.

However, several studies also showed (Bouthillier & Varoquaux, 2020; Hasebrook et al., 2023; Simon et al., 2023; Lee et al., 2020) that AutoML has not fully permeated all these user groups and, thus, has not been able to reach its full potential. We attribute this to the following open challenges:

1. Full AutoML systems were constructed too rigidly. Regarding their use by domain experts, automating the full process of "data science" is arguably a complex problem, and current AutoML systems provide an oversimplified and inflexible solution for this task, when it comes to e.g., the expression of (auxiliary) goals, model preferences and domain knowledge. Likely, although more challenging, it would be more desirable for domain scientists to express such aspects via a natural interface into the system and also have results explained back to them in the same manner. Furthermore, the design of AutoML systems as a software application (rather than a library) also complicates their use

as a subcomponent in more complex systems and code bases, which is, in particular, relevant for data scientists and ML experts. In such scenarios, HPO packages are far more convenient than monolithic AutoML software applications.

2. Current AutoML systems address a narrow task in the data science process by mainly optimizing predictive performance. The shift from optimizing only the predictive model to optimizing full ML pipelines mitigated this problem to some degree, but data science encompasses a lot more than simply optimizing predictive performance. In many applications, aspects like interpretability, causality, fairness and robustness matter greatly, but these are hard to express in a single-objective metric a-priori. Additionally, AutoML systems often cannot handle data organized in multiple tables or non-i.i.d. observations (time and curve data), which occur extremely often in practice.

3. AutoML is often not designed as an iterative process with human interaction but as a press-the-button-once system that returns a single design. However, ML practitioners and data scientists are often unaware of hidden constraints and preferences a-priori that nonetheless matter for the task at hand. Often, this can only be figured out in an interactive process, where intermediate results are discussed with domain experts, implying that AutoML tools should support such an interactive workflow. While HPO tools already provide benefits to ML researchers, not all of their needs are fully addressed, especially the search for scientific insights. For example, many ML researchers require an understanding of hyperparameter sensitivity and the impact of new components.

4. Last but not least, AutoML tools would benefit from further efficiency improvements, especially for the challenging tasks at the cutting edge of ML research. When ML researchers are required to train the largest possible model on currently available hardware, they cannot afford many runs. At the same time, any kind of interactivity between users and AutoML would require a reasonably low response time from the AutoML tool.

Our Position: We believe that most of these challenges are connected: they are caused by ignoring the interaction between users and AutoML systems, and the different roles, expectations, workflows, goals and valuable expertise of users. Although AutoML can support practitioners in many different ways, the strengths of AutoML approaches and users' expertise are complementary. **Therefore, we argue for a more human-centered paradigm in AutoML in this position paper, enabling efficient and collaborative design of ML systems by leveraging the best of both worlds, human experts, and systematic AutoML.**

2. Related Work

To put our advocacy of a more human-centered AutoML paradigm into context, we give an overview of the ML community's shift towards a human-centered paradigm and the first few steps taken by the AutoML community.

2.1. Human-Centered Machine Learning

In recent years, human-centered ML has gained significant momentum, driven by an increasing awareness of social and ethical implications of ML technologies. Leading the charge on human-centered ML are interpretable ML (including transparent decision making), interactive human-in-the-loop approaches and fair ML. As the understanding of terms like interpretability, fairness and transparency is still evolving, these research fields develop and change fast.

Interpretable ML encompasses principles and methods that aim at offering explanations of why an ML model makes certain decisions (Lipton, 2018; Molnar, 2022). These include various approaches, like the development of interpretable models (Rudin, 2019), model-specific methods for deep neural networks (Zhang et al., 2021), model-agnostic techniques to visualize feature effects such as partial dependence plots (Friedman, 2001) or accumulated local effects plots (Apley & Zhu, 2020), methodologies for assessing feature importance (Casalicchio et al., 2019; Hooker et al., 2021; Ewald et al., 2024), and example-based explanations tailored to individual instances, such as Shapley values (Lundberg & Lee, 2017; Sundararajan & Najmi, 2020) or strategies like examining adversarial examples (Goodfellow et al., 2015) and counterfactual explanations (Wachter et al., 2018; Dandl et al., 2020), among others. Given possible systematic errors or unwanted shortcuts taken by ML models, understanding them is crucial in high-stakes scenarios and legal requirements are increasingly mandating audits that rely on interpreting and verifying model behavior (European Union, 2021).

Alongside transparency, fairness in ML has emerged as an important topic (Barocas et al., 2023) to mitigate the risks of unlawful and socially detrimental discrimination. In particular, it attempts to detect, avoid or at the very least mitigate biases of an ML model. Finally, cooperative or interactive ML attempts to integrate human experts into the ML process. This is important from two perspectives: First, to increase performance by injection of expert knowledge and second, to increase the trust in these models by granting experts greater oversight of the systems and a deeper comprehension of the learning process (Wu et al., 2022).

2.2. Human-Centered AutoML

As a human-centered paradigm increasingly permeates the field of ML, it becomes imperative to extend AutoML in this direction.[1] While AutoML research mainly focuses on improving the computational efficiency of AutoML systems (c.f. Guyon et al., 2022; Faust et al., 2023), there are already a few select papers proposing methods related to human-centered AutoML, especially in the realm of interpretable (Biedenkapp et al., 2018; Ono et al., 2020; Moosbauer et al., 2021, see also Section 3.1) and interactive AutoML (Anastacio & Hoos, 2020; Souza et al., 2021; A V et al., 2022; Giovanelli et al., 2024; Hvarfner et al., 2024). We highlight specific works alongside future research directions in Section 4. These works are complemented by publications from Human Computer Interaction (HCI) researchers, e.g. Gil et al. (2019), who collect interface requirements that need to be fulfilled to ensure that human-centered AutoML systems can recreate traditional ML workflows. The position paper by Pfisterer et al. (2019) focuses on the consequences of the popularity of AutoML and possible interfaces for human-centered AutoML systems. Similar to this work, De Bie et al. (2022) argue that automated data science needs to be designed with humans in mind and should only support users, not replace them. However, they put more emphasis on the earlier (but arguably important and currently in AutoML neglected) stages of the data science workflow, i.e., data exploration and problem formalization concerns. Finally, there are several user studies on AutoML, which give valuable insights into user requirements (Wang et al., 2019a; Drozdal et al., 2020; Crisan & Fiore-Gartland, 2021; Xin et al., 2021; Wang et al., 2021b; Hasebrook et al., 2023; Sun et al., 2023). Some previous user studies focus on human-centered AutoML (Lee et al., 2020; Khuat et al., 2023; Xanthopoulos et al., 2020) and examine the current AutoML state-of-the-art and landscape from the user side, e.g., suggesting that AutoML systems be judged increasingly by how much users can interact with them. Our work puts emphasis on investigating currently neglected research directions within the AutoML community that are related to the absence of a human-centered paradigm.

3. The Case for Human-Centered AutoML

In our view, there are five main goals for AutoML tools: (i) Predictive Performance, (ii) Optimization Speed, (iii) Transparency and Interpretability, (iv) Customizability and Flexibility, and (v) Usability and Interaction.

The first two of these requirements (i.e., predictive performance and optimization speed) have been the primary focus of a large portion of AutoML research, and while further re-

---

[1] We note that human-centered can be understood in two ways: "humans as users of ML" and "humans being impacted by ML". Both are equally important, but from an AutoML perspective, we focus on the former view, while the latter is mostly out of scope for this work.

search is required (especially to scale AutoML for the age of foundation models), progress is already well underway for them based on multi-fidelity optimization (Li et al., 2018; Falkner et al., 2018; Wistuba et al., 2022; Kadra et al., 2023), exploiting user priors (Hvarfner et al., 2022; Mallik et al., 2023; Hvarfner et al., 2024), and transfer learning (Wistuba et al., 2018; Feurer et al., 2018; Wistuba & Grabocka, 2021). In contrast to a fully automated approach, we note that a human-centered approach can incur time costs for humans. Both leveraging human expertise and automation via efficient optimization can aid in obtaining increased efficiency of machine learning workflows; we argue that combining them will not result in increased costs but, through their complementary nature, achieve this common goal. A more detailed discussion on this tradeoff can be found in Appendix Section C. With that in mind, this paper focuses on the last three goals, that are in our opinion currently understudied in the AutoML community.

Overall, the machine learning workflow CRISP-ML(Q), as described by Studer et al. (2021), consists of six major phases: 1. business and data understanding, 2. data preparation, 3. model engineering, 4. model evaluation, 5. model deployment, and 6. model monitoring and maintenance. Figure 2 visualizes these phases and puts (interactive) AutoML research in the context of the CRISP-ML(Q) workflow. Due to the inherent complexity, even data science experts need to iterate this workflow, potentially returning to a much earlier stage if the problem has been understood better (Xin et al., 2018). Rapid prototyping, involving swift development and model testing, helps establish an initial baseline. Subsequently, users engage in an often lengthy, iterative trial-and-error process where they experiment with different configurations of the ML workflow to achieve satisfactory outcomes. Therefore, it is not surprising that AutoML is often used for establishing first baselines and model refinement after further insights about the problems at hand are obtained. This iterative and collaborative nature has been overlooked when designing past AutoML systems, and incorporating it explicitly into future systems promises to increase user productivity.

We posit several hypotheses about key insights that have not yet been sufficiently addressed in AutoML.

3.1. Hypothesis 1: Transparency and Interpretability Are Key for ML and AutoML in Many Applications and on Many Levels

Transparency and interpretability are closely linked and are an important source of trust for users in AutoML systems (Wang et al., 2019a; Drozdal et al., 2020). Some studies report interpretability as a concrete requirement requested by users (Xin et al., 2021; Hasebrook et al., 2023; Sun et al., 2023; Wang et al., 2019a). E.g., a lack of inter-
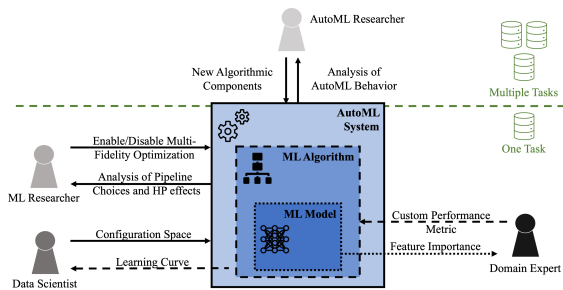

Figure 1. Selected (one-time) interactions of different user groups with AutoML, ML algorithms and models.

pretability of AutoML tools led to study participants choosing manual development for projects with higher stakes Xin et al. (2021). We expand the three levels of interpretability of Moosbauer (2023) by a fourth level, see Figure 1.

1 - ML model interpretability deals with the interpretability of the final model and addresses common questions within the field of interpretable ML. Often, understanding the underlying relations in the data-generating process is the primary target of modeling, instead of simply creating a black box predictor (Shmueli, 2010; Bzdok et al., 2017), as, e.g., stated by data scientists participating in the study by Wang et al. (2019a). Even if maximizing predictive performance is the primary objective, being able to explain and audit a model is usually of high value.

2 - ML algorithm interpretability focuses on the behavior of the learning algorithm, aiming to explain aspects like convergence and the interpretation of learning curves.

3 - AutoML system interpretability deals with understanding why and how an AutoML system has chosen a certain element or pipeline and how hyperparameters influence the final result. This question is usually relevant to a per task / per data set context.

4 - AutoML comparative performance interpretability deals with the performance of the AutoML system itself and could help users understand e.g., what algorithmic choices on the AutoML level may improve an AutoML system for certain applications and why (Dang et al., 2018; Lindauer et al., 2019; Moosbauer et al., 2022b). These questions naturally arise in scientific contexts, where performance comparisons over multiple data sets are of interest.

The relevance of the above requirements varies among user groups. For example, domain experts may prioritize transparency and interpretability of the model returned by the AutoML system, whereas ML practitioners are also interested in understanding why a certain model was returned.

In contrast, ML researchers are often more interested in ablation studies, the robustness of model performance con-

cerning hyperparameters and the effect of hyperparameters. Since AutoML typically collects a lot of data about the performance of different configurations, this data can build the foundation to create such insights (Hutter et al., 2014; Fawcett & Hoos, 2016; Biedenkapp et al., 2017; 2018; Moosbauer et al., 2021; Sass et al., 2022; Watanabe et al., 2023; Theodorakopoulos et al., 2024). At the same time, traditional interpretable ML methods might not be directly applied to performance data collected by AutoML systems because the AutoML process is typically biased towards well-performing configurations and thus, the performance data is biased too (Moosbauer et al., 2021; Segel et al., 2023), at least when one is interested in a global analysis across the whole configuration space.

3.2. Hypothesis 2: Customizability and Flexibility Are Essential to Leverage the Potential of AutoML for Different User Groups

Different AutoML tools operate on different levels of abstraction, each providing varying degrees of customization. Fully automated AutoML systems usually provide the least flexibility but the largest amount of automation. NAS approaches generally allow users to provide a pipeline specification but are still limited in their expressiveness and the building blocks they provide. HPO approaches provide the most flexibility but require users to set up pipelines, configuration spaces and evaluation themselves. However, even for fully automated systems, users might have to configure them correctly for their particular use case to obtain the best performance possible (Lindauer et al., 2019; Moosbauer et al., 2022a; Feurer et al., 2022; Neutatz et al., 2023).

Recent studies showed that the ML practitioner user group criticized current AutoML tools in terms of these customization options (Xin et al., 2021; Sun et al., 2023; Zöller et al., 2023). In the study by Xin et al. (2021) where users rated attributes of current AutoML tools, customizability was among the qualities that received the lowest ratings. Studies however identified different needs when it comes to different customization options. Both Hasebrook et al. (2023) and Wang et al. (2019a) found that current tools for HPO only optimize for predictive performance, whereas practitioners have several additional objectives, such as increasing model comprehension, which could be accounted for using multi-objective optimization (Karl et al., 2023; Horn & Bischl, 2016; Binder et al., 2020). Furthermore, Zöller et al. (2023) found that users have a concrete need to adapt the configuration space to include what they learned from previous AutoML runs, which also very directly hints at the fact that users have a strong need for more interactive workflows, as requirement specifications in data science projects are usually not perfectly precise in a first try. A partial reason for this request is likely also the desire to inject domain knowledge and/or to speed up the optimization, for

example, through warmstarting (Anastacio & Hoos, 2020; Souza et al., 2021; Hvarfner et al., 2022; Mallik et al., 2023; Hvarfner et al., 2024).

Overall, the users of AutoML are diverse, and different applications demand different functionalities of AutoML and, specifically, different types of user interactions. For example, a biologist's primary objective might be accurate predictions of certain processes for which they would like to contribute domain expert knowledge (e.g., a specific kernel or distance function for gene data), while a bank employee, in addition, has to satisfy regulations on interpretability. In the case of a data scientist conducting unsupervised learning, an interactive AutoML approach based on preferences might be required if the ideal performance metric cannot be defined explicitly. For a more in-depth description of the aforementioned applications we refer the interested reader to Appendix Section B.

Due to this diversity in users and applications, another possible approach to AutoML, aside from a customizable platform solution, is a modular one. Frameworks like *GAMA* (Gijsbers & Vanschoren, 2021) or the recently proposed *AutoML Toolkit*[2] aim to provide a toolbox that allows users to design the AutoML solution geared towards their individual applications and beyond simply choosing one black-box optimization algorithm over another. This design philosophy is related to human-centered AutoML in two ways. First, a similar design philosophy could also be extended to interactive, explainable and overall human-centered AutoML by including modules that correspond to certain interactions (position in the data science lifecycle, interpretability level etc.); to the best of our knowledge, a framework adopting this approach specifically for human-centered elements of AutoML has not been proposed. Second, simply by adopting a modular approach, the AutoML process itself becomes more human-centered. Conscious decisions have to be made about design choices of the AutoML solution, which cannot or should not be automated, such as fairness (Weerts et al., 2024). At the end of the day, this will lead to a new abstraction layer of how to build ML systems, hiding many tedious and error-prone design decisions such as hyperparameters and allowing users to focus on the essential decisions by combining different high-level modules, leading to responsible and trustworthy use of (Auto)ML.

3.3. Hypothesis 3: AutoML Tools Have to Integrate with the Data Science Workflow Allowing for an Iterative Interaction with the User

The user experience, i.e. usability, and the options for interacting with current AutoML tools are important and, as for some of the other hypotheses, corresponding require-

---

[2] https://github.com/automl/amltk

ments vary among user groups (Xin et al., 2021; Wang et al., 2021b; Crisan & Fiore-Gartland, 2021). Hasebrook et al. (2023) observe that an important reason for ML practitioners choosing grid search or random search over Bayesian optimization for HPO is the ability to easily integrate these methods into their workflow. We speculate that this is due to additional work that is required to integrate more advanced HPO, especially iterative and synchronous techniques, such as Bayesian optimization, into the technical workflow.

Importantly, the degree of automation or the number of potential interaction points, respectively, for a user drastically depends on the user group (Lee et al., 2020; Crisan & Fiore-Gartland, 2021; Wang et al., 2021b). For example, Crisan & Fiore-Gartland (2021) show that users with a higher technical expertise tend to prefer less automation than users with a less technical background. Similarly, Wang et al. (2021b) highlight that the desired level of automation over the different stages of the machine learning workflow varies for different data science-related roles.

Interaction with an AutoML system can typically happen on the levels introduced in Section 3.1 and can take various forms: a user may desire to give input to, receive output from, or mutually interact with the AutoML system throughout the optimization process, and the details of the requirements for interaction depend again on the user group. For example, domain experts may want to ingest their expert knowledge to inform the ML algorithm, e.g., we might need to learn the expert's internal loss, which they cannot precisely specify, or to restrict a Pareto set to guide the optimization process towards practically desired directions. Another example would be users including their preferences to guide the optimization process either implicitly through prior beliefs on promising pipelines (Mallik et al., 2023; Hvarfner et al., 2024) or explicitly by relative preferences between proposed pipelines (Kulbach et al., 2020; Giovanelli et al., 2024). The latter was, e.g., recently integrated in *Optuna* (Akiba et al., 2019) since Version 3.4 through preferential Bayesian optimization.

We refer the interested reader to Appendix Section A for a discussion of how machine learning operations (MLOps) relates to human-centered AutoML.

3.4. Hypothesis 4: Since Human Experts Are Essential to ML Processes, AutoML Will Only Reach Its Full Potential by Collaborating with Them

Experts - mostly in the form of domain experts and data scientists - shape the lifecycle of an ML model in diverse ways. Khuat et al. (2023) even argue: "systems cannot be considered optimal if they do not welcome and make use of optional human input". The time of these experts is, however, a precious resource in the development of ML solutions; human experts have to be integrated into the Au-

toML process, but the effort on their part should be minimal. We base the discussion on the groups identified in Section 1.

Domain experts can provide invaluable context through their knowledge of the application domain. The user study by Xin et al. (2021) concludes that participants mainly use their domain knowledge for data pre-processing steps, such as feature engineering, and for validating the resulting model. Khuat et al. (2023) detail many possible uses of domain knowledge in every step of an ML workflow, ranging from defining success criteria in the beginning to selecting a configuration space for model search and monitoring the deployed model for possible biases. It can easily be argued that the more domain knowledge a practitioner has, the greater their need to customize and influence the AutoML system (Wang et al., 2019b; Sun et al., 2023).

A lot of data science and ML projects are successful precisely because of the collaboration of domain experts and data scientists (Mao et al., 2019). So, even if data scientists can work more effectively with AutoML tools, domain experts are still integral to the success of ML projects. A common workaround to handle the perceived inflexibility of AutoML systems is often to inject domain knowledge in the optimized objective of the system in a rather technical manner (Sun et al., 2023). Based on this, the authors suggest that instead AutoML systems should be developed either for specific domains and applications and/or support an interactive approach so that users can supply their domain expertise into the process.

ML experts / data scientists are largely receptive to the advantages of using AutoML in a supportive role as outlined in Section 3.5. In practice, ML projects are rarely represented by a linear workflow where a dataset is presented and a model chosen, but through an iterative process. Often, new data is acquired and labeled because of the information gained through baseline performances, which then may require an adaption in model selection and tuning. A data scientist may also decide to include unlabeled data, which requires special modeling techniques. The complexity and variety of these processes make navigating a given ML project's optimal workflow challenging. Many data scientists agree that these types of strategic decisions or the investigative mindset of an expert cannot be fully automated in the near future (Wang et al., 2019a; De Bie et al., 2022). By the same reasoning, if human experts do not fully comprehend the process leading up to finding and training an optimal model, a lot of information will be lost that might have sparked further improvements for future iterations.

ML researchers are the main factor for moving ML forward. They have invaluable knowledge about their fields of expertise that would be foolish to ignore by any AutoML system. While AutoML systems could, in principle, learn to self-evolve and outcompete human ML researchers (Clune, 2019;

Huang et al., 2023), we believe it to be unlikely that AutoML systems alone will yield major novel research breakthroughs (on the level of discovering transformers) without human experts being closely in the loop anytime soon. Nevertheless, we fully expect that, based on the increasing computational efficiency of AutoML and its ability to seamlessly reason about thousands of short and cheap experiments (with down-scaled models, less data and fewer epochs, combined with extrapolation models), it will become ever more standard for ML researchers to use AutoML in their research, leading to a speedup in ML progress by AutoML. Early examples of this already exist in the literature, with the identification of new state-of-the-art deep neural network architecture variants (So et al., 2021), activation functions (Ramachandran et al., 2018), variants of weight updates (Real et al., 2020) or neural optimizers (Andrychowicz et al., 2016; Chen et al., 2023), and we will likely see more such works in the future as AutoML methods become increasingly powerful and convenient to use.

Across user groups In general, interaction and communication between the user and the AutoML system leads to increased trust (Crisan & Fiore-Gartland, 2021; Drozdal et al., 2020), which makes a human-centered paradigm necessary for the widespread adoption of AutoML. This is especially challenging for those without a background in ML (Wang et al., 2019a), but obviously, this depends on the form of communication used (with the current form of rather mathematical and model-based communication being harder to understand than potential natural language output). Another key role that human experts fill in ML projects and that goes hand in hand with the issues of trust and transparency, is that of a regulatory body. Sanity checks (e.g., "maybe there is data leakage because this model performs too well"), ethical concerns (e.g., "this model may be biased against a certain population subgroup") and safety standards (e.g., trained physicians giving the final approval for a course of therapy suggested by ML) are all important to ensure the quality, fairness and safety of ML (Khuat et al., 2023). Moreover, for certain aspects such as fairness, it is highly debatable, whether they can be cast into a metric and then automatically optimized (Weerts et al., 2024), which further underscores the need for a human-in-the-loop for model validation.

Taking all these factors into account, we believe that AutoML, in its current form, is bound to reach a lower ceiling than it could reach if a human-centered paradigm is adopted - in terms of performance, number of applications as well as ethical and safety standards. Thus, allowing interactions, such as providing expert priors on well-performing designs, updates of data or the design space will be crucial to bringing humans back into the AutoML loop.

3.5. Hypothesis 5: Human-Centered AutoML Empowers Users Instead of Making Them Dependent on a System They Do Not Understand

Automation, particularly AutoML, can bear risks when used without care, e.g., discrimination of groups. This can be especially problematic for users of AutoML as evidenced by Zöller et al. (2023) who observe the trend of users with little ML knowledge overestimating their understanding of a model proposed by an AutoML system. This can be seen as an instance of automation bias, i.e., the tendency to place too much trust in automated recommendations (Skitka et al., 1999). The combination of this overly high level of trust and lower barriers to using ML may lead to ML being used for more and more applications where it may not be desirable. This concern of automating bad decisions is shared by participants in the user study by Crisan & Fiore-Gartland (2021). Naturally, one would hope that manually developed ML applications (without the use of AutoML) would include better oversight of experts and thus allow for several stopping points if the ML application turns out to be problematic. We further discuss the effects of bias of humans and automation in Appendix Section D.

With a shift of the focus from pure automation to a human-centered approach, the next generation of AutoML tools can avert many of these dangers and potentially even lead to more positive changes. In particular, we believe that more research in the direction of human-centered AutoML has the potential to empower people instead of making them data-science-illiterate. With the original promise of AutoML of lowering the entry barrier of ML for users, transparency and interpretability allow users to understand which parts have been automated and why certain outputs are obtained. As such, human-centered AutoML could focus on (i) the automation of the tedious and error-prone repetitive task of choosing a well-performing model and optimizing the corresponding hyperparameters and (ii) the interpretability of this process and its outcome. It allows users to focus on those parts of the data science workflow where less automation is possible or desirable.

In the same spirit, human-centered AutoML also allows different data science teaching paradigms: future data scientists might be able to focus much more on the data at the start of their education instead of manually trying tens of different learners without gaining valuable insights by doing so. With the emergence of data-centric artificial intelligence / machine learning (DCAI/DCML), which (re-)emphasizes the importance of data quality, this has even more merit. Additionally, the interpretability of results combined with a good explanation interface could allow users to learn something about the modeling task from the output of an AutoML tool. As such, while they focus on other parts of the workflow, they can still benefit from applying the tool by acquiring

knowledge. This vision is also prevalent in the recent literature, with Wang et al. (2019a) arguing that AutoML should also fulfill an educational role in the future and Xin et al. (2021), who find that several users took models given by AutoML tools as an opportunity to learn more about ML techniques.

4. Ongoing Work and Future Directions

In the following, we discuss existing and modern work in human-centered AutoML and highlight several specific opportunities for new research in this area.

4.1. Existing Work on Trust and Interpretability in the Context of AutoML

In Section 3.1, we argued that interpretability and transparency are some of the core features that facilitate trust in AutoML systems. In fact, during the last few years, a growing amount of publications targeting interpretability in AutoML have been published, usually in the context of extending HPO. On the first interpretability level, multi-objective optimization (Karl et al., 2023) can be used for finding tradeoffs between accurate and less complex/interpretable models (Igel, 2005; Molnar et al., 2020; Binder et al., 2020; Schneider et al., 2023) or fairness-aware models (Perrone et al., 2021; Weerts et al., 2024). On the third interpretability level, multiple approaches exist to increase hyperparameter interpretability. Approaches include measuring hyperparameter importance (Hutter et al., 2014; Jin, 2022; Watanabe et al., 2023), or expressing hyperparameter effects (Moosbauer et al., 2021; Segel et al., 2023). Furthermore, Li & Adams (2020) constrain the points explored by Bayesian optimization to ensure explainability. Lastly, on the fourth level, there is also work on analyzing configuration spaces for HPO and hyperparameter importance across datasets (van Rijn & Hutter, 2018; Probst et al., 2019), but also work on understanding the optimization algorithms themselves (Dang et al., 2018; Lindauer et al., 2019; Moosbauer et al., 2022b).

Similarly, multiple tools directly explain AutoML systems or their outputs (compared to the more methodological work above). Model LineUpper (Narkar et al., 2021) allows users to compare candidate models on multiple information levels. PipelineProfiler (Ono et al., 2020) visualizes ML pipelines produced by AutoML systems. DeepCAVE (Sass et al., 2022) and XAutoML (Zöller et al., 2023) can visualize optimizer runs. ATMSeer (Wang et al., 2019b) explains through visualization what models have been evaluated and how they performed; it also offers a visualization tool to support users in adapting the search space.

Nevertheless, we believe that there is still quite a bit of road ahead. In particular, we encourage the AutoML community to continue working towards understanding what trust, transparency, interpretability and related terms signify in the context of AutoML specifically, particularly when not applied to the model-level but rather the higher levels as discussed in Section 3.1.

4.2. Bridging the Gap Between Algorithmic and HCI AutoML Research

Above, we focused on technical publications on methods for explainability and interaction in the context of AutoML. At the same time, the field of HCI has produced insightful data on user needs and recommendations on how to fulfill them, both for human-centered AutoML systems specifically (Gil et al., 2019; Khuat et al., 2023) and for human-AI interaction in general (Amershi et al., 2019; Yang et al., 2018), often with a focus on trust and interpretability (Liao et al., 2020; Hoffman et al., 2018; Vössing et al., 2022). We encourage more collaboration between these two communities. Few publications have tackled the issues from both sides, but some examples exist. In particular, Zöller et al. (2023) examine users' needs for visualization w.r.t. transparency in AutoML and then propose a framework that satisfies these requirements. Although some survey participants felt overwhelmed by the amount of information presented in their framework XAutoML, it is certainly an important step in the right direction.

4.3. Better Interfaces for AutoML

Thus far, a major obstacle to harnessing the power of ML for people with a non-technical background has been the lack of a sufficiently intuitive way for users to formulate their tasks, domain knowledge, preferences or constraints. In current systems, the interface language is largely mathematical/statistical regarding problem, goal and model specification, and program code in terms of implementation.

This constitutes an obstacle for domain experts, who, e.g., might have a good understanding of how ML models should be evaluated due to their domain knowledge, but may not be able to specify, e.g., the loss or performance metric as a precise formula. Lee et al. (2020) and Bakshy (2023) outline the additional problem that domain knowledge can be so complex that it is difficult to map it to simple constraints. An interesting approach to tackle this issue might be preference learning or preferential optimization (Kulbach et al., 2020; Diaz & López-Ibáñez, 2021; Ungredda & Branke, 2023; Giovanelli et al., 2024), where users are only asked to provide relative feedback by indicating their preference for one outcome over another. This way, they can indirectly communicate important aspects of evaluation but do not have to formally specify objectives and constraints.

Arguably, large language models (LLMs) offer another opportunity to provide an easy-to-use interface to AutoML
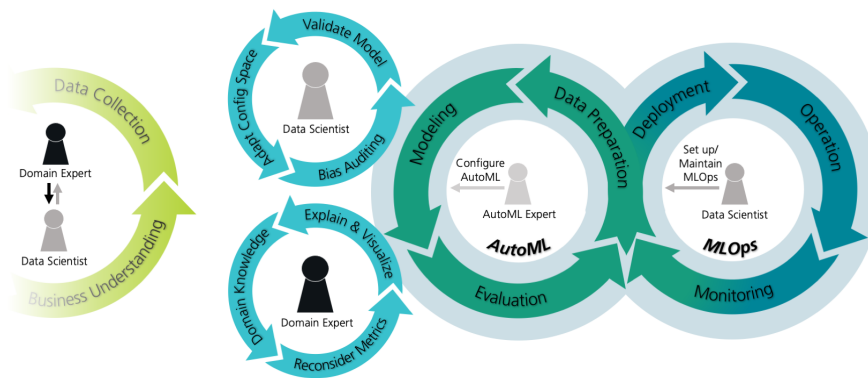
Figure 2. CRISP-human-centered AutoML Cycle, inspired by (Visengeriyeva et al.), with a focus on iterative interactions.

methods (Tornede et al., 2024). The success of LLMs with user groups beyond data science and ML suggests that such a text-based interface increases usability for a wide variety of potential users which could be a key piece of the puzzle for widespread adoption of AutoML systems. LLMs may provide exactly what is needed to allow AutoML to provide the power of ML to many potential users without sacrificing the benefits of a human-centered paradigm: a suitable interface. This is reflected by Karmaker et al. (2021), who see a natural language interface as a prerequisite for domain experts to interact comfortably with AutoML tools. First attempts have been made to use LLMs as an interface to AutoML systems for feature engineering (Hollmann et al., 2023b) or even full ML pipelines (Zhang et al., 2023). We encourage the community to explore this topic further and to facilitate an interactive approach to AutoML through LLMs and other generative multi-modal models. In this context, it could be very promising to couple very fast AutoML systems, such as TabPFN (Hollmann et al., 2023a), with LLMs to allow for a highly interactive user experience.

4.4. From Human-Centered AutoML to Human-Centered Automated Data Science

As with AutoML research in general, most interactive and explainable AutoML approaches are focused on the modeling part of a machine learning project; but AutoML may well be extended beyond modeling. In fact, De Bie et al. (2022) have argued along similar lines as we do in the direction of a human-centered automated data science approach. They argue that some parts of the data science workflow rely on human input (e.g., domain expertise in data acquisition and labeling) and require human oversight (e.g., task definition through business understanding). In fact, data scientists often argue that the early stages of the data science workflow are the most important and time consuming ones; in contrast, the modeling part often contributes considerably less or is less time-consuming (Press, 2016). Thus, there is great potential in taking a human-centered approach

to AutoML across the data science lifecycle. An initial work in this direction is CAAFE (Hollmann et al., 2023b), a human-understandable feature engineering approach for semi-automated data science. Furthermore, an AutoML solution could help users navigate the often non-linear and iterative process through the data science lifecycle. As shown in Figure 2, AutoML can support users in certain aspects regarding model development but requires a human-centered component to intervene if necessary. A human-centered AutoML solution could furthermore help users decide if a data-centric paradigm is most promising moving forward (e.g., labeling of additional data) or a model-centric paradigm (e.g., allocating additional computing resources to modeling) and thus support users in decision making.

5. Conclusion

AutoML had a great success story over the last few years, with a plethora of impressive community achievements. Nevertheless, there is still potential for improvement and further research, in particular in the underexplored area of human interaction with AutoML systems. In this position paper, we proposed and elaborated on a more human-centered approach to AutoML. In particular, we have formed a series of hypotheses on the need for more transparency, interpretability, customizability, flexibility, usability and the integration of human-centered aspects into AutoML research in general. Based on these hypotheses, we analyzed the status quo and encouraged specific future research topics to move the field into the direction we anticipate with this work. In particular, we encouraged more work in interpretability of AutoML, better user interfaces for AutoML, LLMs as an interface to AutoML and extending the human-centered automation paradigm to other parts of the data science workflow.

Impact Statement

Since machine learning is arguably becoming extremely important in research, applications, and industry, and potentially even our personal lives, it is also crucial to push for a democratization of machine learning. AutoML has always aimed to contribute to this democratization, but we believe that the research on AutoML and its tools has underappreciated the requirements and expectations of its different user groups. We believe that this will generate new important research stimulus and eventually lead to a new generation of AutoML tools that are more useful to their users. Since human-centered AutoML tools will also enable to take ethical considerations into account (e.g., interpretability and transparency), we further believe that this paradigm will contribute to the responsible use of machine learning.

Acknowledgements

References

A V, A. K., Rana, S., Shilton, A., and Venkatesh, S. Human-ai collaborative bayesian optimisation. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Proceedings of the 36th International Conference on Advances in Neural Information Processing Systems (NeurIPS'22)*, pp. 16233–16245. Curran Associates, 2022.

Adriaensen, S., Biedenkapp, A., Shala, G., Awad, N., Eimer, T., Lindauer, M., and Hutter, F. Automated dynamic algorithm configuration. *Journal of Artificial Intelligence Research (JAIR)*, 75:1633–1699, 2022.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

Alaa, A. and van der Schaar, M. AutoPrognosis: Automated clinical prognostic modeling via Bayesian optimization with structured kernel learning. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, volume 80, pp. 139–148. Proceedings of Machine Learning Research, 2018.

Amershi, S., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., and Bennett, P. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2019.

Anastacio, M. and Hoos, H. Model-based algorithm configuration with default-guided probabilistic sampling. In Bäck, T., Preuss, M., Deutz, A., Wang, H., Doerr, C., Emmerich, M., and Trautmann, H. (eds.), *Proceedings of the Sixteenth International Conference on Parallel Problem Solving from Nature (PPSN'20)*, Lecture Notes in Computer Science, pp. 95–110. Springer, 2020.

Andrychowicz, M., Denil, M., Colmenarejo, S., Hoffman, M., Pfau, D., Schaul, T., and de Freitas, N. Learning to learn by gradient descent by gradient descent. In Lee, D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Proceedings of the 30th International Conference on Advances in Neural Information Processing Systems (NeurIPS'16)*, pp. 3981–3989. Curran Associates, 2016.

Apley, D. and Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1059–1086, 2020.

Bakshy, E. Beyond loss efficient optimization of living machine learning, 2023. URL https://www.youtube.com/watch?v=MFj8BH0nUBM. Keynote talk at the 2nd Conference on Automated Machine Learning.

Balandat, M., Karrer, B., Jiang, D., Daulton, S., Letham, B., Wilson, A., and Bakshy, E. Botorch: A framework for efficient monte-carlo Bayesian optimization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., and Lin, H. (eds.), *Proceedings of the 34th International Conference on Advances in Neural Information Processing Systems (NeurIPS'20)*. Curran Associates, 2020.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

Benmeziane, H., Maghraoui, K., Ouarnoughi, H., Niar, S., Wistuba, M., and Wang, N. Hardware-aware neural architecture search: Survey and taxonomy. In Zhou, Z. (ed.), *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI'21)*, pp. 4322–4329, 2021.

Bergstra, J. and Bengio, Y. Random search for hyperparameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

Biedenkapp, A., Lindauer, M., Eggensperger, K., Fawcett, C., Hoos, H., and Hutter, F. Efficient parameter importance analysis via ablation with surrogates. In Singh, S. and Markovitch, S. (eds.), *Proceedings of the Thirty-First Conference on Artificial Intelligence (AAAI'17)*, pp. 773–779. AAAI Press, 2017.

Biedenkapp, A., Marben, J., Lindauer, M., and Hutter, F. CAVE: Configuration assessment, visualization and evaluation. In Battiti, R., Brunato, M., Kotsireas, I., and Pardalos, P. (eds.), *Proceedings of the International Conference on Learning and Intelligent Optimization (LION)*, Lecture Notes in Computer Science. Springer, 2018.

Binder, M., Moosbauer, J., Thomas, J., and Bischl, B. Multi-objective hyperparameter tuning and feature selection using filter ensembles. In Ceberio, J. (ed.), *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'20)*, pp. 471—479. ACM Press, 2020.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A., Deng, D., and Lindauer, M. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pp. e1484, 2023.

Bouthillier, X. and Varoquaux, G. Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020. Research report [hal-02447823], Inria Saclay Ile de France, 2020.

Bzdok, D., Krzywinski, M., and Altman, N. Machine learning: a primer. *Nature methods*, 14(12):1119, 2017.

Casalicchio, G., Molnar, C., and Bischl, B. Visualizing the feature importance for black box models. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pp. 655–670. Springer, 2019.

Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Liu, Y., Pham, H., Dong, X., Luong, T., Hsieh, C.-J., Lu, Y., and Le, Q. V. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems*, volume 35, 2023.

Chen, Y., Huang, A., Wang, Z., Antonoglou, I., Schrittwieser, J., Silver, D., and de Freitas, N. Bayesian optimization in alphago. *arXiv:1812.06855 [cs.LG]*, 2018.

Clune, J. Ai-gas: Ai-generating algorithms, an alternate paradigm for producing general artificial intelligence. *arXiv:1905.10985v2 [cs.AI]*, 2019.

Crisan, A. and Fiore-Gartland, B. Fits and starts: Enterprise use of AutoML and the role of humans in the loop. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021.

Cubuk, E., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR'19)*, pp. 113–123. Computer Vision Foundation and IEEE Computer Society, IEEE, 2019.

Dandl, S., Molnar, C., Binder, M., and Bischl, B. Multi-objective counterfactual explanations. In Bäck, T., Preuss, M., Deutz, A. H., Wang, H., Doerr, C., Emmerich, M. T. M., and Trautmann, H. (eds.), *Parallel Problem Solving from Nature - PPSN XVI - 16th International Conference, PPSN*, volume 12269 of *Lecture Notes in Computer Science*, pp. 448–469. Springer, 2020.

Dang, N., Cáceres, L., De Causmaecker, P., and T. Stützle, T. Configuring irace using surrogate configuration benchmarks. In Bosman, P. (ed.), *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'17)*, pp. 243–250. ACM Press, 2018.

De Bie, T., De Raedt, L., Hernández-Orallo, J., Hoos, H., Smyth, P., and Williams, C. Automating data science. *Commun. ACM*, 65(3):76–87, 2022.

Diaz, J. E. and López-Ibáñez, M. Incorporating decision-maker's preferences into the automatic configuration of bi-objective optimisation algorithms. *Eur. J. Oper. Res.*, 289(3):1209–1222, 2021.

Drozdal, J., Weisz, J., Wang, D., Dass, G., Yao, B., Zhao, C., Muller, M. J., Ju, L., and Su, H. Trust in AutoML: exploring information needs for establishing trust in automated machine learning systems. In Paternò, F., Oliver, N., Conati, C., Spano, L. D., and Tintarev, N. (eds.), *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI'20)*, pp. 297–307. ACM, 2020.

Elsken, T., Metzen, J., and Hutter, F. Neural Architecture Search: A survey. *Journal of Machine Learning Research*, 20(55):1–21, 2019.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv:2003.06505 [stat.ML]*, 2020.

Escalante, H. Automated machine learning–a brief review at the end of the early years. In Pillay, N. and Qu, R. (eds.), *Automated Design of Machine Learning and Search Algorithms*, pp. 11–28. Springer, 2021.

European Union. Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. European Union, 2021. URL https://eur-lex.europa.eu/eli/reg/2021/xxxx/oj.

Ewald, F. K., Bothmann, L., Wright, M. N., Bischl, B., Casalicchio, G., and König, G. A guide to feature importance methods for scientific inference. *arXiv:2404.12862 [stat.ML]*, 2024.

Falkner, S., Klein, A., and Hutter, F. BOHB: Robust and efficient Hyperparameter Optimization at scale. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, volume 80, pp. 1437–1446. Proceedings of Machine Learning Research, 2018.

Faust, A., Garnett, R., White, C., Hutter, F., and Gardner, J. R. (eds.). *Proceedings of the Second International Conference on Automated Machine Learning*, volume 228 of *Proceedings of Machine Learning Research*, 2023. PMLR.

Fawcett, C. and Hoos, H. Analysing differences between algorithm configurations through ablation. *Journal of Heuristics*, 22(4):431–458, 2016.

Feurer, M. and Hutter, F. Hyperparameter Optimization. In Hutter et al. (2019), chapter 1, pp. 3 – 38. Available for free at http://automl.org/book.

Feurer, M., Letham, B., and Bakshy, E. Scalable meta-learning for Bayesian optimization. *arXiv:1802.02219v1 [stat.ML]*, 2018.

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., and Hutter, F. Auto-Sklearn 2.0: Hands-free automl via meta-learning. *Journal of Machine Learning Research*, 23(261):1–61, 2022.

Friedman, J. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, pp. 1189–1232, 2001.

Gijsbers, P. and Vanschoren, J. Gama: A general automated machine learning assistant. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*, pp. 560–564. Springer, 2021.

Gil, Y., Honaker, J., Gupta, S., Ma, Y., D'Orazio, V., Garijo, D., Gadewar, S., Yang, Q., and Jahanshad, N. Towards human-guided machine learning. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 614–624, 2019.

Giovanelli, J., Tornede, A., Tornede, T., and Lindauer, M. Interactive hyperparameter optimization in multi-objective problems via preference learning. In *Proceedings of the Thirty-Eighth Conference on Artificial Intelligence (AAAI'24)*, 2024.

Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., and Sculley, D. Google Vizier: A service for black-box optimization. In Matwin, S., Yu, S., and Farooq, F. (eds.), *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*, pp. 1487–1495. ACM Press, 2017.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In Bengio, Y. and LeCun, Y. (eds.), *Proceedings of the International Conference on Learning Representations (ICLR'15)*, 2015. Published online: iclr.cc.

Guyon, I., Sun-Hosoya, L., Boullé, M., Escalante, H., Escalera, S., Liu, Z., Jajetic, D., Ray, B., Saeed, M., Sebag, M., Statnikov, A., Tu, W., and Viegas, E. Analysis of the AutoML Challenge Series 2015-2018. In Hutter et al. (2019), chapter 10, pp. 177–219. Available for free at http://automl.org/book.

Guyon, I., Lindauer, M., Schaar, M., Hutter, F., and Garnett, R. (eds.). *Proceedings of the First International Conference on Automated Machine Learning*, volume 188 of *Proceedings of Machine Learning Research*, 2022. PMLR.

Hasebrook, N., Morsbach, F., Kannengießer, N., Zöller, M., Franke, J., Lindauer, M., Hutter, F., and Sunyaev, A. Practitioner motives to select hyperparameter optimization methods. *arXiv:2203.01717 [cs.LG]*, 2023.

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. Metrics for explainable AI: challenges and prospects. *arXiv:1812.04608 [cs.AI]*, 2018.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR'23)*, 2023a. Published online: `iclr.cc`.

Hollmann, N., Müller, S., and Hutter, F. LLMs for semi-automated data science: Introducing CAAFE for context-aware automated feature engineering. In *Advances in Neural Information Processing Systems*, volume 35, 2023b.

Hooker, G., Mentch, L., and Zhou, S. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31:1–16, 2021.

Horn, D. and Bischl, B. Multi-objective parameter configuration of machine learning algorithms using model-based optimization. In Likas, A. (ed.), *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8. IEEE, 2016.

Huang, Q., Vora, J., Liang, P., and Leskovec, J. Benchmarking large language models as ai research agents. *arXiv:2310.03302 [cs.LG]*, 2023.

Hutter, F., Hoos, H., and Leyton-Brown, K. Sequential model-based optimization for general algorithm configuration. In Coello, C. (ed.), *Proceedings of the Fifth International Conference on Learning and Intelligent Optimization (LION'11)*, volume 6683 of *Lecture Notes in Computer Science*, pp. 507–523. Springer, 2011.

Hutter, F., Hoos, H., and Leyton-Brown, K. An efficient approach for assessing hyperparameter importance. In Xing, E. and Jebara, T. (eds.), *Proceedings of the 31th International Conference on Machine Learning, (ICML'14)*, pp. 754–762. Omnipress, 2014.

Hutter, F., Kotthoff, L., and Vanschoren, J. (eds.). *Automated Machine Learning: Methods, Systems, Challenges*. Springer, 2019. Available for free at `http://automl.org/book`.

Hvarfner, C., Stoll, D., Souza, A., Nardi, L., Lindauer, M., and Hutter, F. πBO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR'22)*, 2022. Published online: `iclr.cc`.

Hvarfner, C., Hutter, F., and Nardi, L. A general framework for user-guided bayesian optimization. In *The Twelfth International Conference on Learning Representations*, 2024.

Igel, C. Multi-objective Model Selection for Support Vector Machines. In Coello, C., Aguirre, A., and Zitzler, E. (eds.), *Evolutionary Multi-Criterion Optimization*, pp. 534–546. Springer, 2005.

Jin, H. Hyperparameter importance for machine learning algorithms. *arXiv:2201.05132 [stat.ML]*, 2022.

Kadra, A., Janowski, M., Wistuba, M., and Grabocka, J. Scaling laws for hyperparameter optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=ghzEUGfRMD`.

Karl, F., Pielok, T., Moosbauer, J., Pfisterer, F., Coors, S., Binder, M., Schneider, L., Thomas, J., Richter, J., Lang, M., Garrido-Merchán, E., Branke, J., and Bischl, B. Multi-objective hyperparameter optimization – an overview. *Transactions of Evolutionary Learning and Optimization*, 3(4):1—50, 2023.

Karmaker, S., Hassan, M., Smith, M., Xu, L., Zhai, C., and Veeramachaneni, K. AutoML to date and beyond: Challenges and opportunities. *ACM Computing Surveys (CSUR)*, 54(8):1–36, 2021.

Khuat, T. T., Kedziora, D. J., and Gabrys, B. The roles and modes of human interactions with automated machine learning systems: A critical review and perspectives. *Foundations and Trends in Human-Computer Interaction*, 17(3-4):195–387, 2023.

Kulbach, C., Philipp, P., and Thoma, S. Personalized automated machine learning. In Lang, J., Giacomo, G. D., Dilkina, B., and Milano, M. (eds.), *Proceedings of the Twenty-fourth European Conference on Artificial Intelligence (ECAI'20)*, June 2020.

Lee, D., Macke, Stephen an Xin, D., Lee, A., Huang, S., and Parameswaran, A. A human-in-the-loop perspective on AutoML: Milestones and the road ahead. *IEEE Data Engineering Bulletin*, 2020.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: A novel bandit-based approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

Li, M. and Adams, R. Explainability constraints for bayesian optimization. In Eggensperger, K., Feurer, M., Weill, C., M.Lindauer, Hutter, F., and Vanschoren, J. (eds.), *ICML workshop on Automated Machine Learning (AutoML workshop 2020)*, 2020.

Liao, Q., Gruen, D., and Miller, S. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–15, 2020.

Lindauer, M., Feurer, M., Eggensperger, K., Biedenkapp, A., and Hutter, F. Towards assessing the impact of bayesian optimization's own hyperparameters. In De Causmaecker, P., Lombardi, M., and Zhang, Y. (eds.), *IJCAI 2019 DSO Workshop*, 2019.

Lindauer, M., Eggensperger, K., Feurer, M., Biedenkapp, A., Deng, D., Benjamins, C., Ruhkopf, T., Sass, R., and Hutter, F. SMAC3: A versatile bayesian optimization package for Hyperparameter Optimization. *Journal of Machine Learning Research*, 23(54):1–9, 2022.

Lipton, Z. C. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43, 2018.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Mallik, N., Bergman, E., Hvarfner, C., Stoll, D., Janowski, M., Lindauer, M., Nardi, L., and Hutter, F. Priorband: Practical hyperparameter optimization in the age of deep learning. In *Advances in Neural Information Processing Systems*, volume 35, 2023.

Mao, Y., Wang, D., Muller, M., Varshney, K. R., Baldini, I., Dugan, C., and Mojsilović, A. How data scientists work together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proceedings of the ACM on Human-Computer Interaction*, 3: 1–23, 2019.

Molnar, C. *Interpretable Machine Learning*. Self-Published, 2 edition, 2022. Available at https://christophm.github.io/interpretable-ml-book/.

Molnar, C., Casalicchio, G., and Bischl, B. Quantifying model complexity via functional decomposition for better post-hoc interpretability. In Cellier, P. and Driessens, K. (eds.), *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD'19)*, volume 1167 of *Communications in Computer and Information Science*, pp. 193–204. Springer, 2020.

Moosbauer, J. *Towards explainable automated machine learning*. PhD thesis, Ludwig-Maximilians-Universität München, 2023.

Moosbauer, J., Herbinger, J., Casalicchio, G., Lindauer, M., and Bischl, B. Explaining hyperparameter optimization via partial dependence plots. In Ranzato, M., Beygelzimer, A., Nguyen, K., Liang, P., Vaughan, J., and Dauphin, Y. (eds.), *Proceedings of the 35th International Conference on Advances in Neural Information Processing Systems (NeurIPS'21)*. Curran Associates, 2021.

Moosbauer, J., Binder, M., Schneider, L., Pfisterer, F., Becker, M., Lang, M., Kotthoff, L., and Bischl, B. Automated benchmark-driven design and explanation of hyperparameter optimizers. *IEEE Transactions on Evolutionary Computation*, 26(6):1336–1350, 2022a.

Moosbauer, J., Binder, M., Schneider, L., Pfisterer, F., Becker, M., Lang, M., Kotthoff, L., and Bischl, B. Automated benchmark-driven design and explanation of hyperparameter optimizers. *IEEE Transactions on Evolutionary Computation*, 26(6):1336–1350, 2022b.

Narkar, S., Zhang, Y., Liao, Q., Wang, D., and Weisz, J. Model lineupper: Supporting interactive model comparison at multiple levels for AutoML. In *26th International Conference on Intelligent User Interfaces*, pp. 170–174, 2021.

Neutatz, F., Lindauer, M., and Abedjan, Z. AutoML in heavily constrained applications. *VLDBJ*, 2023.

Ono, J., Castelo, S., Lopez, R., Bertini, E., Freire, J., and Silva, C. Pipelineprofiler: A visual analytics tool for the exploration of AutoML pipelines. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):390–400, 2020.

Perrone, V., Donini, M., Zafar, M., Schmucker, R., Kenthapadi, K., and Archambeau, C. Fair Bayesian optimization. In Fourcade, M., Kuipers, B., Lazar, S., and Mulligan, D. (eds.), *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES'21)*, pp. 854—-863, 2021.

Pfisterer, F., Thomas, J., and Bischl, B. Towards human centered AutoML. *arXiv:1911.02391 [cs.HC]*, 2019.

Press, G. Cleaning big data: Most time-consuming, least enjoyable data science task, survey says, 2016. URL https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/.

Probst, P., Boulesteix, A., and Bischl, B. Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, 20(53): 1–32, 2019.

Ramachandran, P., Zoph, B., and Le, Q. Searching for activation functions. In *Proceedings of the International Conference on Learning Representations (ICLR'18)*, 2018. Published online: iclr.cc.

Real, E., Liangn, C., So, D., and Le, Q. AutoML-zero: evolving machine learning algorithms from scratch. In Daume III, H. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, volume 98. Proceedings of Machine Learning Research, 2020.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019.

Sass, R., Bergman, E., Biedenkapp, A., Hutter, F., and Lindauer, M. Deepcave: An interactive analysis tool for automated machine learning. In Mutny, M., Bogunovic, I., Neiswanger, W., Ermon, S., Yue, Y., and Krause, A. (eds.), *ICML Adaptive Experimental Design and Active Learning in the Real World (ReALML Workshop 2022)*, 2022.

Schneider, L., Bischl, B., and Thomas, J. Multi-objective optimization of performance and interpretability of tabular supervised machine learning models. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '23, pp. 538–547, 2023.

Segel, S., Graf, H., Tornede, A., Bischl, B., and Lindauer, M. Symbolic explanations for hyperparameter optimization. In Faust, A., White, C., Hutter, F., Garnett, R., and Gardner, J. (eds.), *Proceedings of the Second International Conference on Automated Machine Learning*. Proceedings of Machine Learning Research, 2023.

Shmueli, G. To Explain or to Predict? *Statistical Science*, 25(3):289 – 310, 2010.

Simon, S., Kolyada, N., Akiki, C., Potthast, M., Stein, B., and Siegmund, N. Exploring hyperparameter usage and tuning in machine learning research. In *2nd IEEE/ACM International Conference on AI Engineering - Software Engineering for AI, CAIN*, pp. 68–79. IEEE, 2023.

Skitka, L., Mosier, K., and Burdick, M. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006, 1999.

Snoek, J., Larochelle, H., and Adams, R. Practical Bayesian optimization of machine learning algorithms. In Bartlett, P., Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems (NeurIPS'12)*, pp. 2960–2968. Curran Associates, 2012.

So, D., Mańke, W., Liu, H., Dai, Z., Shazeer, N., and Le, Q. Primer: Searching for efficient transformers for language modeling. In Ranzato, M., Beygelzimer, A., Nguyen, K., Liang, P., Vaughan, J., and Dauphin, Y. (eds.), *Proceedings of the 35th International Conference on Advances in Neural Information Processing Systems (NeurIPS'21)*. Curran Associates, 2021.

Song, X., Perel, S., Lee, C., Kochanski, G., and Golovin, D. Open source vizier: Distributed infrastructure and API for reliable and flexible blackbox optimization. In *International Conference on Automated Machine Learning, AutoML*, volume 188, pp. 8/1–17. PMLR, 2022.

Souza, A., Nardi, L., Oliveira, L., Olukotun, K., Lindauer, M., and Hutter, F. Bayesian optimization with a prior for the optimum. In Oliver, N., Pérez-Cruz, F., Kramer, S., Read, J., and Lozano, J. A. (eds.), *Machine Learning and Knowledge Discovery in Databases. Research Track*, volume 12975 of *Lecture Notes in Artificial Intelligence*, pp. 265–296. Springer-Verlag, 2021.

Studer, S., Bui, T., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., and Müller, K.-R. Towards crisp-ml(q): A machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction*, 3(2):392–413, 2021.

Sun, Y., Song, Q., Gui, X., Ma, F., and Wang, T. AutoML in the wild: Obstacles, workarounds, and expectations. In Schmidt, A., Väänänen, K., Goyal, T., Kristensson, P., Peters, A., Mueller, S., Williamson, J., and Wilson, M. (eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 1–15. ACM Press, 2023.

Sundararajan, M. and Najmi, A. The many shapley values for model explanation. In Daume III, H. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, volume 98, pp. 9269–9278. Proceedings of Machine Learning Research, 2020.

Theodorakopoulos, D., Stahl, F., and Lindauer, M. Hyperparameter importance analysis for multi-objective automl. *arXiv:2405.07640 [cs.LG]*, 2024.

Thornton, C., Hutter, F., Hoos, H., and Leyton-Brown, K. Auto-WEKA: combined selection and Hyperparameter Optimization of classification algorithms. In Dhillon, I., Koren, Y., Ghani, R., Senator, T., Bradley, P., Parekh, R., He, J., Grossman, R., and Uthurusamy, R. (eds.), *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13)*, pp. 847–855. ACM Press, 2013.

Tornede, A., Deng, D., Eimer, T., Giovanelli, J., Mohan, A., Ruhkopf, T., Segel, S., Theodorakopoulos, D., Tornede, T., Wachsmuth, H., and Lindauer, M. AutoML in the age of large language models: Current challenges, future opportunities and risks. *Transactions of Machine Learning Research (TMLR)*, 2024.

Ungredda, J. and Branke, J. When to elicit preferences in multi-objective bayesian optimization. *Proceedings of the Companion Conference on Genetic and Evolutionary Computation (GECCO '23)*, 2023.

van Rijn, J. and Hutter, F. Hyperparameter importance across datasets. In Guo, Y. and Farooq, F. (eds.), *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'18)*, pp. 2367–2376. ACM Press, 2018.

Visengeriyeva, L., Kammer, A., Bär, I., Kniesz, A., and Plöd, M. https://ml-ops.org/content/crisp-ml.

Vössing, M., Kühl, N., Lind, M., and Satzger, G. Designing transparency for effective human-ai collaboration. *Information Systems Frontiers*, 2022.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 31(2), 2018.

Wang, C., Wu, Q., Weimer, M., and Zhu, E. Flaml: A fast and lightweight automl library. In Smola, A., Dimakis, A., and Stoica, I. (eds.), *Proceedings of Machine Learning and Systems 3*, volume 3, pp. 434–447, 2021a.

Wang, D., Weisz, J., Muller, M., Ram, P., Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H., and Gray, A. Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. In Lampinen, A., Gergle, D., and Shamma, D. (eds.), *Proceedings of the ACM on Human-Computer Interaction*. ACM Press, 2019a.

Wang, D., Ram, P., Weidele, D., Liu, S., Muller, M., Weisz, J., Valente, A., Chaudhary, A., Torres, D., Samulowitz, H., and Amini, L. AutoAI: Automating the end-to-end ai lifecycle with humans-in-the-loop. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, pp. 77–78, 2020.

Wang, D., Liao, Q. V., Zhang, Y., Khurana, U., Samulowitz, H., Park, S., Muller, M., and Amini, L. How much automation does a data scientist want? *arXiv:2101.03970 [cs.LG]*, 2021b.

Wang, Q., Ming, Y., Jin, Z., Shen, Q., Liu, D., Smith, M., Veeramachaneni, K., and Qu, H. Atmseer: Increasing transparency and controllability in automated machine learning. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–12, 2019b.

Watanabe, S., Bansal, A., and Hutter, F. PED-ANOVA: efficiently quantifying hyperparameter importance in arbitrary subspaces. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI*, pp. 4389–4396. ijcai.org, 2023.

Weerts, H., Pfisterer, F., Feurer, M., Eggensperger, K., Bergman, E., Awad, N., Vanschoren, J., Pechenizkiy, M., Bischl, B., and Hutter, F. Can fairness be automated? guidelines and opportunities for fairness-aware AutoML. *Journal of Artificial Intelligence Research*, 79:639–677, 2024.

White, C., Safari, M., Sukthanker, R., Ru, B., Elsken, T., Zela, A., Dey, D., and Hutter, F. Neural architecture search: Insights from 1000 papers. *arXiv:2301.08727 [cs.LG]*, 2023.

Wistuba, M. and Grabocka, J. Few-shot bayesian optimization with deep kernel surrogates. In *Proceedings of the International Conference on Learning Representations (ICLR'21)*, 2021. Published online: iclr.cc.

Wistuba, M., Schilling, N., and Schmidt-Thieme, L. Scalable Gaussian process-based transfer surrogates for Hyperparameter Optimization. *Machine Learning*, 107(1): 43–78, 2018.

Wistuba, M., Kadra, A., and Grabocka, J. Supervising the multi-fidelity race of hyperparameter configurations. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Proceedings of the 36th International Conference on Advances in Neural Information Processing Systems (NeurIPS'22)*. Curran Associates, 2022.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381, 2022.

Xanthopoulos, I., Tsamardinos, I., Christophides, V., Simon, E., and Salinger, A. Putting the human back in the AutoML loop. In *Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference, Copenhagen, Denmark, March 30, 2020*, volume 2578 of *CEUR Workshop Proceedings*, 2020.

Xin, D., Ma, L., Song, S., and Parameswaran, A. G. How developers iterate on machine learning workflows - A survey of the applied machine learning literature. *arXiv:1803.10311 [cs.LG]*, 2018.

Xin, D., Wu, E. Y., Lee, D. J.-L., Salehi, N., and Parameswaran, A. Whither AutoML? understanding the role of automation in machine learning workflows. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2021.

Yakovlev, A., Moghadam, H., Moharrer, A., Cai, J., Chavoshi, N., Varadarajan, V., Agrawal, S., Idicula, S., Karnagel, T., Jinturkar, S., and Agarwal, N. Oracle AutoML: a fast and predictive AutoML pipeline. *Proceedings of the VLDB Endowment*, 13(12):3166–3180, 2020.

Yang, Q., Suh, J., Chen, N.-C., and Ramos, G. Grounding interactive machine learning tool design in how non-experts actually build models. In *Proceedings of the 2018 designing interactive systems conference*, pp. 573–584, 2018.

Zhang, S., Gong, C., Wu, L., Liu, X., and Zhou, M. AutoML-GPT: Automatic machine learning with gpt. *arXiv:2305.02499 [cs.CL]*, 2023.

Zhang, Y., Tiño, P., Leonardis, A., and Tang, K. A survey on neural network interpretability. *IEEE Trans. Emerg. Top. Comput. Intell.*, 5(5):726–742, 2021.

Zhou, N., Jiang, Y., Bergquist, T., Lee, A., Kacsoh, B., Crocker, A., Lewis, K., Georghiou, G., Nguyen, H., Hamid, M., et al. The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, 20:1–23, 2019.

Zöller, M.-A., Titov, W., Schlegel, T., and Huber, M. F. XAutoML: A visual analytics tool for understanding and validating automated machine learning. *ACM Transactions on Interactive Intelligent Systems*, 2023.

A. Human-centered AutoML and Machine Learning Operations (MLOps)

Both MLOps and AutoML aim to make machine learning more accessible and more convenient for potential users. Where AutoML aims at automating machine learning processes - especially the configuration of machine learning pipelines - MLOps is centered around the technical aspects as well as the deployment and monitoring of machine learning models in production. In Figure 2, the tasks often associated with MLOps are represented in the right circle (though MLOps extends into experimentation and validation as well through e.g., reproducibility of experiments). By integrating AutoML within an MLOps framework, we can streamline the entire lifecycle of a machine learning project – from data preparation to model deployment and monitoring – eventually leading to *AutoMLOps*. In view of Hypothesis 3, this strong automation nevertheless comes with a multitude of challenges in which humans are still needed, including defining business objectives (e.g., choosing success metrics) and problem framing, data management, interpreting results, making strategic decisions based on monitoring, ethical oversight, compliance, and continuous improvement by feedback loops. The multitude of tools coming together to form such a complex (semi-) automated system requires strong modularity of components, interfaces between components and users, compatibility between tools, and data management on experiments and versions; a good MLOps setup therefore becomes increasingly important.

B. Selected AutoML Users with Different Needs

Overall, the users of AutoML are diverse and different applications demand different functionality of AutoML and specifically different types of user interactions. To put emphasis on this variety, in the following we showcase four distinct examples for potential users and their applications:

A manufacturing expert in predictive maintenance:
Machine learning can reduce maintenance costs and unexpected downtime of heavy machinery through predictive maintenance. While this can easily be formulated as a classification, regression or survival analysis problem, domain expertise is important to create an effective machine learning model. Domain experts provide important information for feature selection and engineering, about constraints and even when it comes to evaluation of a machine learning model through their technical expertise and knowledge of the circumstances surrounding the machinery. While something like root cause analysis can be interesting in its own right, the predictive maintenance task usually only concerns itself with

predictive performance and does not care about the interpretability of predictions.

A biologist working with gene or gene-expression data: In such tasks, ML models are often adapted through domain knowledge. This domain knowledge usually comes in the form of string (or graph) kernels, gene (dis)similarity functions, or information from e.g., the gene ontology database (grouping of genes, and a hierarchical / DAG structure on groups). It is well known that using such domain knowledge can improve model performance (Zhou et al., 2019). In terms of goals, users are, in many cases, interested in sparse solutions (i.e., few genes or few groups), as these results drive subsequent (expensive) experiments. Interpretability is often a goal, but not always.

A data scientist in a bank predicting credit risk: Here, in contrast to the previous example, less customized ML models are often used, e.g., often linear models or other interpretable models; sometimes nonlinear forests or boosting-type models. Next to high predictive performance (maximizing profits), explainability of the machine learning model is often either a (legal) hard constraint or at least highly desirable. This needs to be incorporated into the AutoML process because otherwise very complex models, which are only marginally better than a simple one, might be selected. An additional consideration is fairness, which is notoriously hard to automate and quantify in single metrics (Weerts et al., 2024).

A data scientist conducting unsupervised learning: In contrast to supervised learning, it is notoriously hard to a-priori define ideal performance metrics that can simply be optimized against. Experienced data scientists usually diagnose and evaluate such models through visualization. In this case, an interactive approach along the lines of preference learning would be useful in AutoML systems for such tasks, see e.g., (Giovanelli et al., 2024).

C. Scalability and Cost of Human Efforts

In the age of deep learning and large models, scalability is indeed a very important topic to consider in the context of AutoML. A lot of current AutoML research is centered around scalability to large models (incl. LLMs). On the one hand, AutoML and especially NAS algorithms are often designed to identify configurations with high-predictive power as well as energy or memory-efficient pipelines through, e.g., multi-objective optimization. On the other hand, research focuses on specific optimization methods like multi-fidelity optimization or scaling laws to make the search for a desirable pipeline as efficient as possible. We identify this as a current challenge for AutoML and call for further efficiency improvements to accommodate current architectures. (Tornede et al., 2024) provides an in-depth discussion on this for LLMs.

Two main factors influence the final quality during model selection and configuration: Prior knowledge and available budget for optimization. The tighter the available budget, the more important the reliance on prior knowledge becomes – one-shot configuration would be an extreme case. Unfortunately, many AutoML systems (with few modern exceptions) do not really allow the incorporation of very flexible and custom priors, which reflect human domain knowledge. We can already see that in the domain of large (language) models. When, for example, configuring an LLM, there are only so many attempts that can realistically be made at finding a good architecture and/or hyperparameters, so – due to a lack of better alternatives – we often rely on human experts here. Even worse, the training of a large model over several days or weeks requires human supervision, e.g., to prevent divergence of training – in fact, no one would risk training a model on thousands of GPUs for a month without actively monitoring it. Automating this process as much as possible is still desirable.

As a concrete example of a synergetic workflow, let us assume that only a single training of a large model is feasible and there is a constant risk of divergence. We envision a future in which a new kind of human-centered AutoML provides suggestions on how to adapt training when being at risk, but human supervision is necessary in view of the lack of extrapolation capabilities of AutoML for unseen future training steps. There are very first steps in this direction with approaches such as dynamic algorithm configuration (Adriaensen et al., 2022).

Comparing the two extremes, manual vs. fully automated systems, it should be obvious that both lack the advantages of the other. AutoML will support users in tedious and error-prone tasks, such as manually trying out a sequence of different architectures or hyperparameters, whereas users will support AutoML in providing prior knowledge on high-level concepts and adapting overall objectives. Combining the complementary strengths of both can minimize the overall cost of designing and training a complex model.

D. Bias through Human Intervention

The question of whether bias will be introduced through human intervention is an interesting but subtle topic. First of all, just because a decision process is based on data, ML and automation, this does not imply it is free of bias, in the sense of "bad" or "improper" human bias, as humans usually select the data source, humans specify goals, humans specify side constraints, and humans interpret results. We

know that this can lead to drastic problems, if not properly accounted for, e.g., from the current fair-ML literature, but even earlier, there are many instances in applied data science where this has been observed. Automation can partially mitigate such biases, but it never fully eliminates them. And unfortunately, there is also a counter-effect of automation. If we accept that these biases can and will occur in many instances, and there is no silver bullet to avoid them a-priori, the next best thing is careful analysis, checking, and auditing. This is done by human experts, but the more automated, the less transparent a system is, the harder this becomes. In a fully autonomous system, a human expert is more detached from the machine learning process and may be able to fulfill oversight responsibilities to better prevent bias. However, with the increasing automation of black boxes and lack of transparency, it becomes harder for human experts to spot biases and related problems. Additionally, automation bias, the tendency to place too much trust in automated recommendations (Skitka et al., 1999), may be reinforced through AutoML (Weerts et al., 2024). Finally, the previous point more or less refers to better, potential output of AutoML systems – to enable better human auditing. When it comes to humans selecting configurations during an interactive process, this can both introduce human biases or mitigate them. For example, a human could artificially drive the system to an unfair solution – or detect that the currently created solution is unfair and intervene. It is our belief that a transparent process is necessary to minimize this as much as possible. Nevertheless, this will not be a perfect system simply because humans are not perfect. However, removing human oversight and reason from the equation is, in our opinion, far more dangerous.