AutoAIViz: Opening the Blackbox of Automated Artificial Intelligence with Conditional Parallel Coordinates

Daniel Karl I. Weidele*, Justin D. Weisz, Erick Oduor, Michael Muller, Josh Andres, Alexander Gray and Dakuo Wang
IBM Research - daniel.karl@ibm.com

ABSTRACT
Artificial Intelligence (AI) can now automate the algorithm selection, feature engineering, and hyperparameter tuning steps in a machine learning workflow. Commonly known as AutoML or AutoAI, these technologies aim to relieve data scientists from the tedious manual work. However, today’s AutoAI systems often present only limited to no information about the process of how they select and generate model results. Thus, users often do not understand the process, neither they trust the outputs. In this short paper, we build an experimental system AutoAIViz that aims to visualize AutoAI’s model generation process to increase users’ level of understanding and trust in AutoAI systems. Through a user study with 10 professional data scientists, we find that the proposed system helps participants to complete the data science tasks, and increases their perceptions of understanding and trust in the AutoAI system.

CCS CONCEPTS
• Human-centered computing → Empirical studies in visualization; • Computing methodologies → Continuous space search: Machine learning approaches.

KEYWORDS
AutoAI, visualization, parallel coordinates, human-AI collaboration

ACM Reference Format:
Daniel Karl I. Weidele, Justin D. Weisz, Enos Oduor, Michael Muller, Josh Andres, Alexander Gray and Dakuo Wang. 2020. AutoAIViz: Opening the Blackbox of Automated Artificial Intelligence with Conditional Parallel Coordinates. In IUI '20: ACM Intelligent User Interfaces, March 17–20, 2020, Cagliari, Italy. ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION
Data scientists apply machine learning techniques to gain insights from data in support of making decisions [25]. The challenge is that this process is labor intensive, requiring input from multiple specialists with different skill sets [1, 29, 32, 45]. As a result, AI and Human-Computer Interaction (HCI) researchers have investigated how to design systems with features that support data scientists in creating machine learning models [23, 27, 32, 39, 40, 42]. This work is often referred to as human-centered machine learning or human-in-the-loop machine learning [8, 27, 42].

Recently, techniques have been developed to automate various steps of the data science process, such as preparing data, selecting models, engineering features, and optimizing hyperparameters [21, 24, 26, 28, 46]. These techniques – known as AutoAI or AutoML – are being incorporated into commercial products (e.g. [10, 11, 15, 30, 37]) as well as open source packages [4, 6, 34]. Although the ability for AutoAI to speed up the model-building process is very promising, there are challenges due to the opaque nature of how AutoAI actually creates models [27, 42, 43]. Users do not necessarily understand how and why AutoAI systems make the choices they make when selecting algorithms or tuning hyperparameter values. This “blackbox” nature of AutoAI operation hinders the collaboration between data scientists and AutoAI systems [43].

Several solutions have been proposed to make AutoAI more transparent in its operation in order to increase trust. These proposals include documenting the process by which machine learning models were created [12, 31] and using visualizations to show the search space of candidate algorithms or hyperparameter values [9, 35, 43]. However, these solutions either document the state of an AI model after it has been created, or focus only on one step of the model generation workflow (e.g., search algorithms or selecting hyperparameters). There remains a gap in providing users with an overview of how an AutoAI system works in operation, from the moment data are read to the moment candidate models are produced.

We propose a new kind of visualization – Conditional Parallel Coordinates (CPC) [44] – to open up the “black box” of AutoAI operation. CPC is a variation of the parallel coordinates visualization [17, 18], that provides interactive views of categorical and continuous value dimensions, allowing the user to expose increasing amounts of detail added to a conventional Parallel Coordinates display. We implemented CPC into an existing AutoAI platform [42] and conducted a usability study with 10 professional data scientists to understand its usefulness in showing fine-grained details about the operation of a live AutoAI system. Our results suggest that data scientists are able to successfully glean information from the visualization in order to gain a better understanding of how AutoAI makes decisions, while those models are being built. Our work shows how increased process transparency leads to improved collaboration between data scientists and AutoAI tools.

2 RELATED WORK

2.1 Human-in-the-loop Data Science and Automated Machine Learning
Many studies have focused on understanding the work practices of data scientists. Muller et al. and Hou & Wang both reported that data
science workflows operate as an iterative conversation between data scientists and other stakeholders, focused around data and model artifacts [14, 32]. As such, data science work is a confluence of people with many different roles, skills, and needs. Amershi et al. highlighted the importance of adopting a human-centered approach to building data science tools [2, 3]. Kross & Guo users expressed a strong desire for an integrated user interface with both code and narrative. These needs are perhaps best captured in the narrative uses [23] of the Jupyter Notebook environment [19, 20], and researchers have conducted numerous studies of how data scientists incorporate notebooks into their workflows [36, 39], how they conduct version control for notebooks [22], and how they enable simultaneous multi-user editing in notebooks [40].

What remains less well-understood is how data scientists will incorporate automated machine learning technologies into their workflows. While initial work suggests that data scientists see AutoAI tools as collaborative [42], much about the underlying process by which AutoAI operates remains obscure to data scientists [9, 27, 35, 43]. Our efforts in this work focus on increasing trust in AutoAI by increasing the transparency of its operation.

### 2.2 Trust via Transparency in AI Systems

Recent criticisms of “blackbox” machine learning models (e.g. [28, 38]) begin with the notion that if one does not understand how a model produces a recommendation, then that recommendation may not be trustworthy. The issue of trust may be crucial for AutoAI systems. Not only must a model’s recommendations be explainable, but also the process by which the model was created.

Visualizations provide a promising way for increasing transparency of operation, and several systems have already been created that focus on specific points within the data science workflow. Vizier [9] and VisualHyperTuner [35] both focus on hyperparameter tuning, showing the relationships between a model’s hyperparameters and performance. ATMSeer [43] provides a multi-granularity visualization of model selection and hyperparameter tuning, enabling users to monitor the process and adjust the search space in real time. While these visualizations improve transparency of specific components of the AutoAI process, no visualization has yet tackled a representation of the entire process, from data ingestion to model evaluation. Our approach for depicting the entire process applies the Conditional Parallel Coordinates (CPC) visualization, which presents data scientists with (1) a real-time overview of the machine learning pipelines as they are being created, and (2) the ability to view detailed information at each step in the AutoAI process.

### 3 CONDITIONAL PARALLEL COORDINATES FOR AUTOAI

We provide a first user-evaluation of a recently proposed Conditional Parallel Coordinates (CPC) component [44]. Compared to classic Parallel Coordinates (PC) visualizations [7, 16], CPC introduces new layers of details recursively.

Pipeline configuration data generated by an AutoAI search process is considered conditional data. We say hyperparameters are conditioned on the choice of a pipeline step. This conditionality enables CPC, which allows to inspect conditional layers. AutoAI first chooses a sequence of transformation steps before training an estimator. Then, the hyperparameters of these choices are fine-tuned. Further, the user can bound the AutoAI search space by setting constraints, such as group disparity against model bias, or prediction runtime for inference speed [28]. Typically, during the process pipelines are being optimized towards a performance measure, in our case ROC AUC.

Figure 1 shows our suggested mapping of this data to CPC. On the top level we generate an axis for each of the following attributes: Pipeline ID, Transformer 1, Transformer 2, Estimator, Group Disparity, Prediction Runtime, and ROC AUC on training and holdout split.
respectively. In the next level of detail, for categorical choices, we construct a conditional axis per hyperparameter. The figure shows the visualization after the user expanded the Sparse Random Projection Transformer in the Transformer 2 step. Here, the visualization further allows to differentiate pipelines across the 5 hyperparameters dense_output, density, eps, n_components and random_state.

Thus, in our CPC visualization users have the ability to drill down into a particular step of the machine learning pipeline by clicking it. This reveals further details about this step, which in turn are again visualized in the form of PC. We assume this recursive process together with seamless preservation of intuitive interaction patterns from PC will enable users to familiarize themselves with the limited additional complexity rather quickly.

Next to the CPC component, towards the bottom of the interface we feature a pipeline leaderboard, showing pipeline configuration data sorted according to ROC AUC descendingly. For each pipeline row, from left to right, we display the rank, ROC AUC, Group Disparity, Prediction Time, ID and information on the steps.

Color indicates the choice of the estimator, as well for pipelines in CPC as for rows in the leaderboard. In principle the color mapping could be left up to the user to support focusing on different axes, however, for this study we intended to limit further complexities.

Lastly, AutoAIVis is capable of reflecting the AutoAI search status live, i.e. as the search progresses in the background latest pipelines are added to the visualizations.

4 USER-EVALUATION STUDY

To evaluate the CPC visualization in the context of AutoAI, we recruited 10 professional data scientists across three continents. We developed a set of 17 informational questions for participants to answer using AutoAIVis (Table 1), based on a binary classification task. Each question asked about one factual piece of information contained in the interface. Although all questions could be answered with the CPC visualization alone, we included the Leaderboard visualization (shown at the bottom of Figure 1) in order to represent existing AutoAI user interfaces, and gauge which representation was preferable for participants. Our questions included finding the total number of pipelines generated by AutoAI (Q1), finding the training data ROC AUC score for a given pipeline (Q9), and finding the most frequently searched transformer (Q13).

Participation in our evaluation took approximately 30 minutes. Participants were allowed a few minutes to freely explore the interface before they were presented with the set of questions. Due to our desire to gauge CPC’s self-explanatory nature, participants were given minimal explanation about the visualization or how to interact with it. Participants worked for about 10-15 minutes in answering the questions. As in other studies of visualizations (e.g. [5, 13, 33, 41]), participants were encouraged to think aloud as they explored the interface and ask any questions they had.

We tracked how participants answered each question, either by consulting the CPC visualization (C), the Leaderboard (L), or both (B). All but one of our questions were factual, and thus had either correct and incorrect answers; one question asked participants to select a single pipeline and then answer 4 questions specific to it.

In addition to the questions, we had participants self-assess their familiarity with, understanding of, and trust in AutoAI systems on 5-point Likert scales. Familiarity was assessed prior to the study; understanding and trust were measured at the beginning and the end. Finally, we asked participants to reflect on elements they liked and disliked during the study, as well as their desire to use AutoAI in the future as part of their work practice.

We primarily examine the extent to which participants were able to correctly find each piece of information and which visualization they used to find it, as well as their understanding of and trust in AutoAI changed in the study.

5 RESULTS

5.1 Ability to Interpret CPC

Across all 17 questions, we find that CPC has been used three times more often than the leaderboard ($\chi^2 = 42.050$, p<.001). However, to learn about the results on the components more thoroughly, let us compare which of the two components has been favorably consulted depending on the types of questions.

Questions on total counts, but also counts grouped by transformers or estimators have mostly been answered using CPC. Further, all candidates were able to intuitively learn how to reveal hyperparameter information, with only 2 people trying to look for the answer to such questions in the leaderboard. Interestingly, even question 17, concerned with the difference between the best and the worst pipeline, has been answered favorably using CPC only, making participants feel comfortable enough to avoid a scroll to the far end of the leaderboard. One participant also managed to answer all questions correctly solely based on CPC.

We further find the leaderboard serves a differentiating purpose. For questions where a single pipeline needed to be identified based on the result metrics, participants preferred to consult the leaderboard. This is particularly the case for Q6, where we put our participants into a conflicting scenario by asking for the best pipeline in their opinion. 40% would not overwrite the top pipeline of the leaderboard. 50% opted for a slightly less accurate pipeline roughly cutting group disparity in half and predicting twice as fast. Only one subject opted for a pipeline with group disparity of 0%, setting back ROC AUC to 0.500. For this question not a single participant based their decision on CPC only. Surprisingly, over the course of the next 3 questions where only single data of the just chosen pipeline row was asked for, up to 6 participants hovered back to instead read the results from the CPC visualization. Overall, we further leave with the impression that subjects did not have difficulties switching between the components.

With respect to the quality of provided answers, across all questions, we were not able to determine a statistically significant dependency between the question outcome (correct / incorrect) and the used component ($\chi^2 = 0.42$, p < .81), which means CPC did not introduce systematic error compared to the leaderboard.

Regarding the questions most frequently answered incorrectly, some participants found the mix of pipeline steps with result metrics confusing in the beginning (Q2). We believe this can be improved visually by separating the axes more clearly, or adding tool tips. Q11 is asking for the top estimator, however, participants slipped and reported the top pipeline. This seems more of an issue with the study than the components themselves. Up to 40% struggled giving...
correct answers on hyperparameter questions (Q15, Q16), ignoring numerical hyperparameter axis, or counting values instead of axis.

5.2 Effect of CPC on Trust and Understanding

Participants reported some initial familiarity with AutoAI prior to the study (M (SD) = 2.5 (1.51) of 5). Ratings of trust fell in the middle of the scale prior to using CPC, with a non-significant increase after using CPC (pre: M (SD) = 2.7 (1.06) of 5, post: M (SD) = 3.3 (1.06), t(9) = 1.77, p < .11). By contrast, understandability increased significantly (pre: M (SD) = 2.4 (0.97), post: M (SD) = 3.6 (0.52), t(9) = 3.67, p < .005). We speculate that, with a larger sample size, the difference in trust might also have achieved significance.

5.3 Qualitative Feedback

After the study participants were further asked to give their subjective opinion on what they liked or disliked in AutoAIViz.

On the positive side, 8 out of the 10 participants found the interactive CPC visualization useful, mostly for its ability to simultaneously display pipeline steps, hyperparameters and results. The remaining 2 candidates more commended the overall system, specifically as a way to generate competitive baselines with little effort.

On the negative side, candidates sometimes found the CPC display to be crowded, thus making it harder to perceive individual or subsets of pipelines, or hyperparameters. Participants were also not always familiar with the terminologies used in AutoAIViz. Lastly, 2 subjects wished they were given more power to fine-tune the AutoAI manually, i.e. through code.

We also wanted to know if participants could imagine AutoAIViz as part of their job in the future. Half of our participants could clearly see themselves work with AutoAIViz after some minor improvements, with one candidate even already relying on a similar system. With only a single subject arguing against AutoAI correlated to lacking trust, the remaining four expressed they would consult AutoAI moderately, as a baseline to kick-off new projects, or in cases when they are less informed.

6 DISCUSSION

With CPC being consulted most frequently to answer our questions, we believe AutoAI visualization systems should display this component in a prominent position in the user interface design. The fact that some questions, however, were favorably addressed using the alternative leaderboard suggests that other components should also be displayed. We think next steps need to taken towards brushing and linking, to better enable the interplay of other components with CPC.

Based on participants’ feedback, we believe we are on the right track to promote a better understanding of AutoAI systems. Some subjects, however, would like to see improvements in AutoAIViz before adding it to their job routine. We can mostly address these by placing further visual elements to increase the readability of metrics, where constraints could be indicated as bars in the background of the corresponding axes. Further, interaction to zoom to intervals of numerical axes will allow users to hide outliers distorting the view, such as shown in the axis Prediction time in Figure 1. Another useful feature could be a close all expanded steps button, so less trained users can quickly reset in case they get lost.

Lastly, we found it hard to significantly increase trust among the participants. This could be due to the fact that we allowed participants to only work with the system for 15 minutes, but building trust and requires more experience. Further, subjects were exposed to a dataset they are not necessarily very familiar with. We speculate that allowing participants to work with their own dataset could have led to more trust into AutoAIViz.

7 CONCLUSION

AutoAIViz with CPC has been accepted well and significantly contributed to users’ understanding of AutoAI. The system did not significantly increase trust, but we identified measures to improve towards this on the component level. Overall, CPC has proven to be of central value in opening the AutoAI blackbox, especially when augmented with other visualization techniques depending on the task, such as the leaderboard. Our study leads to future work towards enabling brushing and linking in AutoAIViz, exploring further techniques suggested for classic PC to improve CPC, and providing insights into other steps of the machine learning pipeline.
