

# 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 do they trust the outputs. In this short paper, we provide a first user evaluation by 10 data scientists of an experimental system, AutoAIViz, that aims to visualize AutoAI's model generation process. We find that the proposed system helps users to complete the data science tasks, and increases their understanding, toward the goal of increasing trust in the AutoAI system.

## CCS CONCEPTS

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

## KEYWORDS

AutoAI, AutoML, visualization, parallel coordinates, human-AI collaboration, democratizing AI

### ACM Reference Format:

Daniel Karl I. Weidele, Justin D. Weisz, Eno Oduor, Michael Muller, Josh Andres, Alexander Gray and Dakuo Wang. 2020. AutoAIViz: Opening the Blackbox of Automated Artificial Intelligence with Conditional Parallel Coordinates. In *25th International Conference on Intelligent User Interfaces (IUI '20), March 17–20, 2020, Cagliari, Italy*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3377325.3377538>

## 1 INTRODUCTION

Data scientists apply machine learning techniques to gain insights from data in support of making decisions [26]. The challenge is that this process is labor intensive, requiring input from multiple specialists with different skill sets [1, 31, 34, 47, 48]. 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 [24, 28, 34, 41, 42, 44]. This

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*IUI '20, March 17–20, 2020, Cagliari, Italy*

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-7118-6/20/03...\$15.00  
<https://doi.org/10.1145/3377325.3377538>

work is often referred to as **human-centered machine learning** or **human-in-the-loop machine learning** [9, 28, 44].

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 [22, 25, 27, 30, 50]. These techniques – known as **AutoAI** or **AutoML** – are being incorporated into commercial products (e.g. [11, 12, 16, 32, 39]) as well as open source packages [4, 7, 36]. 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 [28, 44, 45]. 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 [45].

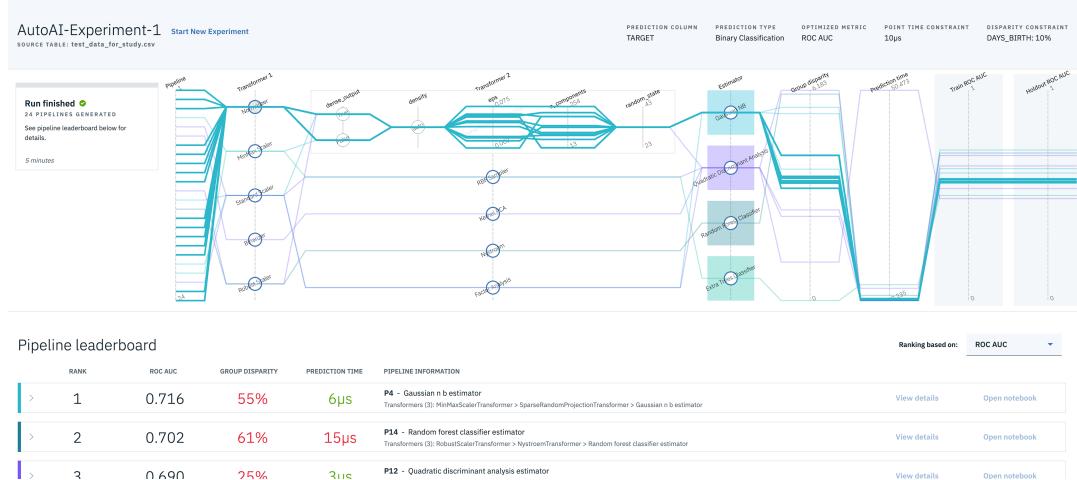
Several solutions have been proposed to make AutoAI more transparent in its operation in order to increase trust. These proposals include documenting the process by which machine learning models were created [13, 33] and using visualizations to show the search space of candidate algorithms or hyperparameter values [10, 37, 45]. 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.

In this paper, we provide a first user evaluation of a new kind of visualization – **Conditional Parallel Coordinates (CPC)** [46] – to open up the “black box” of AutoAI operation. CPC is a variation of the parallel coordinates visualization [18, 19], that allows the user to expose increasing amounts of detail added to a conventional Parallel Coordinates display interactively. We integrated CPC into an existing AutoAI platform [44] and conducted a usability study with 10 professional data scientists to understand its usefulness. 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, as a step toward democratizing AI for non-technical users.

## 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



**Figure 1: Screenshot showing the CPC visualization (top) and the leaderboard (bottom, partially shown). In CPC, each colored line running from left to right represents a machine learning pipeline, and corresponds to a row in the leaderboard. In this example, all pipelines consist of three steps: transformation step 1, transformation step 2, and estimator. Pipelines are evaluated on four metrics: group disparity, prediction time, and ROC AUC scores on the training data and the holdout data.**

data scientists and other stakeholders, focused around data and model artifacts [15, 34]. As such, data science work is a confluence of people with many different roles, skills, and needs [48]. 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 [24] of the Jupyter Notebook environment [20, 21], and researchers have conducted numerous studies of how data scientists incorporate notebooks into their workflows [38, 41], how they conduct version control for notebooks [23], and how they enable simultaneous multi-user editing in notebooks [42].

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 [44], much about the underlying process by which AutoAI operates remains obscure to data scientists [10, 28, 37, 45]. 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. [29, 30, 40, 49]) 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 [6]. 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 [10] and VisualHyperTuner [37] both focus on hyperparameter tuning, showing the relationships between a model’s hyperparameters and performance. ATMSeer [45] 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 single components in the AutoAI process, no visualization has yet tackled a representation of the entire process, from data ingestion to model evaluation. To visualize the entire process, we use Conditional Parallel Coordinates (CPC), which offers data scientists, (a) a real-time overview of the machine learning pipelines as they are being created, and (b) 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 [46]. Compared to classic Parallel Coordinates (PC) visualizations [8, 17], 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 [30]. 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, AutoAIViz 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 AutoAIViz (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, 14, 35, 43]), 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 ( $\chi^2_1 = 42.050$ ,  $p < .001$ ). 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 UI. The fact that some questions were favorably addressed using the alternative leaderboard suggests that other components should also be displayed. We think next steps need to be taken towards brushing and linking, to better enable the interplay of other components with CPC. We will also explore "scaling-up" issues involving larger and more diverse collections of pipelines.

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

**Table 1: Quantitative results of the study. In case of multiple answers we highlight the correct answer **bold**. Note [Q6] is subjective, so all answers are correct. For [Q6-10] we provide aggregated results of correct (✓) and incorrect answers (✗).**

	Result	#	#C	#L	#B
<b>Q1</b> Find the total number of pipelines generated by AutoAI	24	10	6	0	4
3	6	3	1	2	
7	4	4	0	0	0
<b>Q3</b> Find the tot. num. of transf. evaluated in transf. step 1	5	10	10	0	0
<b>Q4</b> Find the tot. num. of transf. evaluated in transf. step 2	5	10	10	0	0
<b>Q5</b> Find the tot. num. of estim. evaluated in the estim. step	4	10	10	0	0
<b>P12</b>	5	0	4	1	
<b>P4</b>	4	0	1	3	
<b>P21</b>	1	0	1	0	
<b>Q7</b> Find the group disparity score of this pipeline (Q6)	✓	10	2	7	1
<b>Q8</b> Find the prediction time of this pipeline (Q6)	✓	10	3	6	1
<b>Q9</b> Find the Train. data ROC AUC score of this pipeline (Q6)	✓	10	6	3	1
<b>Q10</b> Find the Holdout data ROC AUC score of this pipeline (Q6)	✓	9	5	0	4
✗	1	0	1	0	
<b>Q11</b> Find the best performing estimator regarding Training ROC AUC performance score	<b>GNB</b>	6	3	1	2
P4	3	2	1	0	
QDA	1	0	1	0	
<b>Q12</b> Is there a pipeline that satisfies both constraints?	<b>No</b>	9	1	4	4
Yes	1	1	0	0	
<b>Q13</b> Find the most frequently searched transformer	<b>SRP</b>	8	8	0	0
Norm.	2	2	0	0	
<b>Q14</b> Find the most frequently searched estimator	GNB	10	9	0	1
5	6	6	0	0	
4	2	1	0	1	
2	2	2	0	0	
<b>Q16</b> Find the hyperparameter that most influenced the Quadr. Discr. Analysis Estimator	<b>tol</b>	7	6	0	1
N/A	2	1	0	1	
s_cov	1	1	0	0	
<b>Q17</b> Find the difference in Holdout ROC AUC between the best and the worst perf. pipelines	<b>0.231</b>	8	6	2	0
0.138	1	0	0	1	
N/A	1	0	0	1	
<b>TOTAL</b>	170	108	33	29	
✓	149	94	30	25	
✗	21	14	3	4	

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.

## REFERENCES

- [1] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: a case study. In *ICSE '19*. IEEE Press, 291–300.
- [2] Saleema Amershi, Bongshin Lee, Ashish Kapoor, Ratul Mahajan, and Blaine Christian. 2011. Human-guided machine learning for fast and accurate network alarm triage. In *Twenty-Second International Joint Conference on Artificial Intelligence*.
- [3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collison, Jina Shu, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *CHI '19*. ACM, 3.
- [4] auto-sklearn. [n. d.]. *auto-sklearn*. Retrieved 06-Oct-2019 from <https://automl.github.io/auto-sklearn/master/>
- [5] Luca Chittaro and Lucio Ieronutti. 2004. A visual tool for tracing users' behavior in Virtual Environments. In *Proceedings of the working conference on Advanced visual interfaces*. ACM, 40–47.
- [6] Jamie Drozdal, Justin Weisz, Dakuo Wang, Dass Gaurave, Bingsheng Yao, Changrui Zhao, Michael Muller, Lin Ju, and Hui Su. 2020. Exploring Information Needs for Establishing Trust in Automated Data Science Systems. In *IUI '20*. ACM, in press.
- [7] EpistasisLab. [n. d.]. *tpot*. Retrieved 3-April-2019 from <https://github.com/EpitasisLab/tpot>
- [8] H Gannett and FW Hewes. 1883. General summary showing the rank of states by ratios 1880. *Scribner's statistical atlas of the United States showing by graphic methods their present condition and their political, social and industrial development*, Hewes FW,(Ed.), United States. Census Office (1883), 151.
- [9] Yolanda Gil, James Honaker, Shikhar Gupta, Yibo Ma, Vito D'Orazio, Daniel Garijo, Shruti Gadewar, Qifan Yang, and Neda Jahanshad. 2019. Towards human-guided machine learning. In *IUI '19*. ACM, 614–624.
- [10] Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D Sculley. 2017. Google Vizier: A Service for Black-Box Optimization. In *KDD '17*. ACM, 1487–1495.
- [11] Google. [n. d.]. Cloud AutoML. Retrieved 3-April-2019 from <https://cloud.google.com/automl/>
- [12] H2O. [n. d.]. *H2O*. Retrieved 3-April-2019 from <https://h2o.ai>
- [13] Michael Hind, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, and Kush R Varshney. 2018. Increasing Trust in AI Services through Supplier's Declarations of Conformity. *arXiv preprint arXiv:1808.07261* (2018).
- [14] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, 164–168.
- [15] Youyang Hou and Dakuo Wang. 2017. Hacking with NPOs: collaborative analytics and broker roles in civic data hackathons. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 53.
- [16] IBM. [n. d.]. *AutoAI*. Retrieved 06-Oct-2019 from <https://www.ibm.com/cloud/watson-studio/autoai>
- [17] Alfred Inselberg. 1985. The plane with parallel coordinates. *The visual computer* 1, 2 (1985), 69–91.
- [18] Alfred Inselberg. 2009. *Parallel coordinates*. Springer.
- [19] Alfred Inselberg and Bernard Dimondale. 1990. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the 1st conference on Visualization '90*. IEEE Computer Society Press, 361–378.
- [20] Project Jupyter. [n. d.]. *Jupyter Notebook*. Retrieved 3-April-2019 from <https://jupyter.org>
- [21] Project Jupyter. [n. d.]. *JupyterLab*. <https://www.github.com/jupyterlab/jupyterlab>
- [22] James Max Kanter and Kalyan Veeramachaneni. 2015. Deep feature synthesis: Towards automating data science endeavors. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 1–10.
- [23] Mary Beth Kery, Bonnie E John, Patrick O'Flaherty, Amber Horvath, and Brad A Myers. 2019. Towards Effective Foraging by Data Scientists to Find Past Analysis Choices. In *CHI '19*. ACM, 92.
- [24] Mary Beth Kery, Marissa Radensky, Mahima Arya, Bonnie E John, and Brad A Myers. 2018. The story in the notebook: Exploratory data science using a literate programming tool. In *CHI '18*. ACM, 174.
- [25] Udayan Khurana, Deepak Turaga, Horst Samulowitz, and Srinivasan Parthasarathy. 2016. Cognito: Automated feature engineering for supervised learning. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 1304–1307.
- [26] Sean Kross and Philip J Guo. 2019. Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 263.
- [27] Hoang Thanh Lam, Johann-Michael Thiebaut, Mathieu Sinn, Bei Chen, Tiep Mai, and Oznur Alkan. 2017. One button machine for automating feature engineering in relational databases. *arXiv preprint arXiv:1706.00327* (2017).
- [28] Doris Jung-Lin Lee, Stephen Macke, Doris Xin, Angela Lee, Silu Huang, and Aditya Parameswaran. 2019. A Human-in-the-loop Perspective on AutoML: Milestones and the Road Ahead. *Data Engineering* (2019), 58.
- [29] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM.
- [30] Sijia Liu, Parikshit Ram, Deepak Vijaykeerthy, Djallel Bouneffouf, Gregory Bramble, Horst Samulowitz, Dakuo Wang, Andrew Conn, and Alexander Gray. 2019. An ADMM Based Framework for AutoML Pipeline Configuration. *arXiv:cs.LG/1905.00424*
- [31] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R. Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilovic. 2019. How Data Scientists Work Together With Domain Experts in Scientific Collaborations. *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (Dec 2019).
- [32] Microsoft. [n. d.]. Azure Machine Learning Studio. Retrieved 06-Oct-2019 from <https://azure.microsoft.com/en-us/services/machine-learning-studio/>
- [33] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 220–229.
- [34] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *CHI '19*. ACM, 14 pages.
- [35] Judith S Olson, Dakuo Wang, Gary M Olson, and Jingwen Zhang. 2017. How people write together now: Beginning the investigation with advanced undergraduates in a project course. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 1 (2017), 4.
- [36] Randal S Olson and Jason H Moore. 2016. TPOT: A tree-based pipeline optimization tool for automating machine learning. In *Workshop on Automatic Machine Learning*, 66–74.
- [37] Heungsik Park, Jinwoong Kim, Minkyu Kim, Ji-Hoon Kim, Jaegul Choo, Jung-Woo Ha, and Nako Sung. 2019. VisualHyperTuner: Visual analytics for user-driven hyperparameter tuning of deep neural networks. In *Demo at SysML Conference*.
- [38] Samir Passi and Steven J Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 136.
- [39] Data Robot. [n. d.]. *Data Robot: Automated Machine Learning for Predictive Modeling*. Retrieved 06-Oct-2019 from <https://datarobot.com>
- [40] Cynthia Rudin. 2018. Please stop explaining black box models for high stakes decisions. *arXiv preprint arXiv:1811.10154* (2018).
- [41] Adam Rule, Aurélien Tabard, and James D Holland. 2018. Exploration and explanation in computational notebooks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 32.
- [42] APRIL YI WANG, ANANT MITTAL, CHRISTOPHER BROOKS, and STEVE ONEY. 2019. How Data Scientists Use Computational Notebooks for Real-Time Collaboration. (2019).
- [43] Dakuo Wang, Judith S. Olson, Jingwen Zhang, Trung Nguyen, and Gary M. Olson. 2015. DocuViz: Visualizing Collaborative Writing. In *Proceedings of CHI '15*. ACM, New York, NY, USA, 1865–1874.
- [44] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *CSCW'19* (2019).
- [45] Qianwen Wang, Yao Ming, Zhihua Jin, Qiaomo Shen, Dongyu Liu, Micah J Smith, Kalyan Veeramachaneni, and Huamin Qu. 2019. Atimseer: Increasing transparency and controllability in automated machine learning. In *CHI '19*. ACM, 681.
- [46] Daniel Karl I. Weidle. 2019. Conditional Parallel Coordinates. *IEEE transactions on visualization and computer graphics* 26, 1 (2019).
- [47] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding interactive machine learning tool design in how non-experts actually build models. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM.
- [48] Amy Zhang, Michael Muller, and Dakuo Wang. 2020. How do Data Science Workers Collaborate?: Roles, Workflows, and Tools. In *CSCW'20*. ACM, in press.
- [49] Yunfeng Zhang, Q. Vera Liao, and Rachel KE Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM.
- [50] Marc-André Zöller and Marco F Huber. 2019. Survey on Automated Machine Learning. *arXiv preprint arXiv:1904.12054* (2019).