The Future of AutoML and The Impact of Bayesian Methods on AutoML

Author

Hussain Abbas MSc, CEO & Chief Data Scientist at Stats AI

PhD Student in Statistics & Data Science at University of Nevada, Reno

Abstract

AutoML is one of the hottest trends in Artificial Intelligence research and is poised to completely upend the standard model currently used in the Data Science industry. Ironically, few understand what it is, how it works, and its relation to Bayesian Statistics.

In this paper, we will discuss the future of AutoML and the impact of Bayesian Methods on AutoML. Specifically, we will discuss:

- What AutoML is
- A Critique on the Current Academic & Software Approach to AutoML
- A Resolution to the Ethical Quandary at the Core of AutoML
- How AutoML is poised to completely upend the standard model currently used in the Data Science industry
- How Bayesian Methods play a critical role in AutoML
- Future areas of research & application

Keywords: automated machine learning, AutoML, future of AutoML, ethics, AI ethics, machine learning, ML, statistics, AI, bayesian optimization, hyperparameter optimization, data science, meta-learning, Artificial Intelligence, auto-sklearn

Introduction

In this essay, we present a novel approach to understanding the current and future state of AutoML.

To the best of our knowledge, nobody has yet discussed the points that we have in the manner that we do. Thus, this represents the first foray into this subject matter.

We show the advantages of AutoML, as well as discuss pitfalls and critical areas that have been overlooked by both academic researchers and software designers to date.

Ethics has become a major part of the conversation with regards to data science. However, the issue of ethics has been entirely non-existent with respect to AutoML research.

As ethical AutoML researchers, we can no longer afford to disregard the ethics of the utilization of AutoML.

Our goal in this paper, is not only to raise awareness, but to start that conversation.

To that effect, we introduce a new standard for AutoML that rectifies unresolved deficiencies in its structure. We show how AutoML as is currently defined leads to an inevitable ethical dilemma. To cure this deficiency, we introduce a standard which enables data science practitioners to ethically utilize AutoML.

We show from the data scientists' perspective how AutoML has already changed the data science landscape and the impact of Bayesian Methods on AutoML systems.

We discuss where AutoML is, where it is going, and where it needs to go.

We rigorously define true general end-to-end AutoML and demonstrate how its existence will fundamentally change the way data scientists work for the better.

Contrary to the belief that AutoML will replace data scientists with AI, we believe that such a system will increase their productivity, make them more efficient, and free them up to be more creative in their work.

We believe that true generalized end-to-end AutoML will completely revolutionize the field of data science, advance the state of the discipline, and move us closer to true artificial general intelligence.

Here at Stats AI, we look forward to partnering with both the public and private sectors and the academic research community to accomplish this task.

What is Auto ML?

In 2016, a Crowdflower survey indicated that the typical Data Scientist spends roughly 80% of their time gathering and cleansing data, and only 20% of the time on actual science [1]. The median Data Scientist salary as of 2019 is roughly $120,000 a year [2].

Thus, a Data Scientist ironically spends very little time doing actual science.

The data gathering and data cleansing tasks which occupy most of the Data Scientist's time do not actually require a college degree. Yet, ironically, a PhD or a Master's degree is widely considered to be a prerequisite for entry into the field of data science, with 90% of data scientists holding at least a Master's degree [3].

Furthermore, these data gathering and data cleansing tasks are the ones that data scientists dislike the most [4]. This underutilization of skills results in churn amongst data scientists as they seek employment in positions which actually utilize the skills they went to school for, increasing labor costs for firms.

Data Science is an iterative process, so the faster we can iterate, the faster we can innovate.

Thus, our inability to quickly iterate through non-creative tasks such as data gathering and data cleansing is the biggest barrier to innovation. Were we to overcome this barrier, it would be akin to a dam breaking and a releasing of the floodwaters.

But how could we overcome this barrier?

Enter AutoML.

Since 2016, there has been a plethora of software, some proprietary and some open source, which have sought to reduce the time spent by Data Scientists on the Data Science pipeline ranging from data gathering and cleansing to ML [19], [24], [26], [27].

As is often the case with proprietary software solutions, they offer tangible benefits, but at the cost of a hefty price tag that is beyond the realistic budget of most firms [25].

The fact that an updated 2018 survey indicated that most Data Scientists do not use proprietary platforms, does not bode well [5]. This low utilization is partially explained by the fact that these platforms often do not have a coherent value proposition, in that firms who would benefit most from utilization of the software are often unable to afford it while firms that are able to afford it usually have ready alternatives and tend not to use it, even after purchasing it.

Thus, there is a need for an end-to-end open source AutoML platform which can transform the Data Scientist into a person who does real science most of the time.

Thus, we define true general end-to-end AutoML to be a set of technological methods that:

1) Reduces the time spent on uninteresting tasks such as data cleansing and data gathering.
2) Can perform at expert level, independent of the problem and data, any supervised learning, unsupervised learning, reinforcement learning task.
3) Accelerates the scientific research process itself, thereby enabling data scientists to rapidly innovate, decrease time-to-results, and build more robust models.

It is important to note that there exist functions within packages scattered across the data science landscape which already perform AutoML, but just aren't marketed as AutoML. For example, the excellent forecast library, developed by Hyndman in R, has a function, auto.arima, which automatically fits an ARIMA model to time series data [6]. We consider this to be AutoML as it fits well within the scope of our above definition.

The good news is that many AutoML frameworks already exist [7], [20], [23], although not at the level which they need to be, and not in an end-to-end fashion as defined above.

As is usually the case with a new technology still in its relative infancy, these frameworks are at present decentralized and disconnected, each having a different focus, often utilizing different software environments, backed by firms and entities having different, and often opposing goals. This naturally makes it difficult for the frameworks to "talk to" each other and realize the goal of true general AutoML.

This is why here at Stats AI, our goal is to research and develop such a system.

We intend to build a system that realizes the goal of having a one-stop shop that performs true end-to-end general AutoML.

Our experience in researching and developing successful end-to-end fully autonomous artificially intelligent inferential systems for pipe leak prediction, oil and gas fraud detection, and electricity fraud detection has shown us that not only that it can be done, but that it should be done.

Experience has shown us that the ecosystem approach works best for end-to-end fully autonomous artificially intelligent inferential systems, a subset of AutoML.

Thus, rather than AutoML being a massive one-size fits all system, we believe that it is most likely to be successful if designed as an ecosystem of individual subsystems.

Contrary to the belief that AutoML will replace data scientists with AI, we believe that such a system will increase their productivity, make them more efficient, and free them up to be more creative in their work.

We believe that such a system will completely revolutionize the field of data science, advance the state of the discipline, and move us closer to true artificial general intelligence.

A Critique on the Current Academic & Software Approach to AutoML

The problem with the current academic and software approaches to AutoML is that they all assume that the data which is the input into the AutoML process already exists in a nice Kaggle-like format provided by a benefactor who serves the data up on a silver platter to the AutoML which serves as a substitute for the data scientist from this point onwards, sans putting the model into production [15], [16], [18], [19], [20], [23].

Unfortunately, this is rarely, if ever the case in practice.

On the contrary, the norm is the situation in which *somebody* knows where *some* of the data is, but *nobody* knows where *all of the data* is.

Furthermore, that *somebody* usually has no idea what the data *contains* since their job is not to *analyze* it but to *store* it.

That is, the data scientist rather than having access to a preexisting data-dictionary, usually is the one who has to construct the data dictionary as a means to conduct ML.

Recall from [1] that roughly 80% of the time is spent on data gathering and data cleansing and only 20% of the time on actual science.

The 80% includes, but is not limited to the following:

1. Identify what data exists
2. Identify where the data resides.
3. Identify how to connect the various data sources
4. Understand what the data means
5. Understand the quality of the data
6. Understand the legality of using the data
7. Cleanse the data
8. Construct data sets that are ready to be used for ML purposes

Current academic and software approaches to AutoML assume that most, if not all, of the above steps have already been completed and the data is ready to go, kit and kaboodle. Even data cleansing assumes the data already exists.

It is widely understood by ML practitioners that the 20% of the time that is spent on actual science is usually the easiest part of the Machine Learning process, whereas the above 80% is the most time consuming, the most difficult, and the most boring.

And yet the 20% is the focus of current AutoML Research rather than the 80%.

Why is that the current academic and software approaches to AutoML focus on automating the easy stuff which doesn't take that much time instead of focusing on automating the hard stuff which does take up most of the time?

A Resolution to the Ethical Quandary at the Core of AutoML

Liu in [8] argues that most of the existing research in AutoML falls under what he calls "narrow AutoML", which he defines as AutoML that reduces but does not eliminate the need for experts. In contrast, Liu defines "generalized AutoML" as AutoML that does not require the need for experts.

Liu argues that this definition of AutoML stems from the first AutoML workshop in 2014 [9], which states that, "*machine learning has achieved considerable successes in recent years and an ever-growing number of disciplines rely on it. However, this success crucially relies on* ***human machine learning experts, who select appropriate features, workflows, machine learning paradigms, algorithms, and their hyperparameters.*** *As the complexity of these tasks is often beyond non-experts, the rapid growth of machine learning applications has created a demand for off-the-shelf machine learning methods that can be used easily and without expert knowledge. We call the resulting research area that targets progressive automation of machine learning AutoML*."

The problem with the above definition of AutoML and Liu's interpretation of it is that it defines human machine learning experts only in terms of the tasks that currently need to be performed by humans.

We believe that with AutoML, the most ethical scenario is one in which human experts are still utilized. We believe that human experts should always be kept in the loop (especially if lives depend upon it).

Already, removing humans from the loop of automated ML processes has resulted in some spectacular failures and backlash against what we would refer to as AutoML. Examples include, but are not limited to: Microsoft's chatbot Tay which had to be unplugged due to its racist remarks [10]; Uber's self-driving car program which was placed on hold after a self-driving car killed a pedestrian; and IBM's Watson AI which recommended procedures that Doctors stated would have caused harm or death [11].

Clearly then, the goal of removing human experts entirely from the equation is unethical. And yet, as of 2019, the removal of human experts entirely from the equation is still the stated goal of AutoML research [15] [16].

Had the Doctors blindly followed Watson, they would have been prosecuted for criminal negligence. Removing them from the equation would have resulted in deaths.

The point of commonality of all well-designed automated systems is that a human controller can intervene when things go awry, examples include but are not limited to manufacturing, high-frequency trading, driverless cars, etc.

Thus, in contrast, we define a human machine learning expert to be a person who possesses knowledge of how to perform the aforementioned tasks, rather than someone who actually does them or needs to do them.

Clearly then, as ethical AutoML researchers, our goal shouldn't be to exclude ML experts from the equation, but to enable them to intervene when things go awry.

How AutoML is poised to completely upend the standard model currently used in the Data Science industry

In Data Science, it is said that data begets data. As more data is collected, more products and services can be offered, which in turn, generate more data. Models themselves generate data.

Today's data scientists are drowning in data, and the bad news is that the deluge of data and data sources is only going to get worse [4].

True general end-to-end AutoML will enable data scientists for the first time in history to not only overcome this issue, but fundamentally change the way they interact with data.

For example, in order to build a model with data, a data scientist must know what the data is, where it resides, and what it means. Usually that involves talking to people and getting a data dictionary (assuming one even exists – it usually doesn't). Usually that involves meetings. Meetings take time. You get the idea.

However, all this data is stored somewhere. If it could be accessed, the AutoML system could not only construct the relationships between the data sources, but construct models autonomously using them.

Just having accomplished the above, the AutoML system would have done in seconds, what would have taken a data scientist a few days of work to do.

Usually, since the majority of a data scientist's energies are devoted to data cleansing and data gathering [1], the data scientist will often settle for rudimentary models that are often "good enough". This is less the case at larger firms with large teams of data scientists who have the luxury of specialization but is usually the case at small startups and firms which have only a single data scientist, i.e., a jack-of-all trades.

True general end-to-end AutoML changes all of this.

By outsourcing the data gathering and data cleansing tasks to the AutoML system, the data scientist is now freed to focus on building better models, better meaning more advanced, more robust, more accurate, etc.

In addition, the data scientist can now construct many models in the time that it took to build a single model. Holding everything else constant, with true general end-to-end AutoML, a single data scientist at a small startup can now meet or even exceed the productivity of a large team of data scientists at a large firm operating without it.

Thus, true general end-to-end AutoML decreases opportunity costs for data scientists by increasing their productivity. Rather than spending time writing code to create models, the data scientist now simply leapfrogs off what the AutoML has built. At the same time, the data scientist serves as check on the AutoML, i.e., a person who can understand and edit what the AutoML has built, in a sense, an editor to the AutoML.

With this newfound time, the data scientist is now free to:

1) Go deeper down the rabbit hole by researching more advanced mathematical techniques not yet integrated into the AutoML system and then integrating them into AutoML system.
2) Focus on the conclusions generated by the AutoML system and communicating those conclusions to business stakeholders, thereby enabling them to make better decisions.
3) Iterate faster through the pipeline of planned projects and deliver solid results.

Thus, we see that true general end-to-end AutoML fundamentally changes the way data scientists work for the better. Data science is done today roughly the same way it was done 100 years ago.

True general end-to-end AutoML changes the entire game by enabling data scientists to go deeper and broader than hitherto otherwise without increasing labor costs.

Thus, with the onset of true general end-to-end AutoML, firms not utilizing it will not be allocating their resources and capital wisely. Investors will take notice of this and shift their capital to firms which efficiently allocate capital, i.e., firms that use AutoML.

Clearly, true general end-to-end AutoML will not only provide a major boost to the startup economy, but to the national economy as well. Thus, it is in the primary interest of policy makers and industrial stakeholders to support its development.

How Bayesian Methods play a critical role in AutoML

In Machine Learning, given a model, the goal is to find the set of hyperparameters which minimizes a loss function/objective function which measures the difference between the actual observed values and the values predicted by the model. Optimization procedures are simply the means by which we perform this task.

Current AutoML approaches can be described via a structure consisting of an outer and inner loop [8]. The outer loop selects the candidate model and the inner loop searches through the set of hyperparameter choices for the model. As this process can be quite computationally burdensome, a method that can reduce this burden is preferred.

Bayesian optimization is such a method.

Bayesian optimization (BO) is an optimization procedure that can best be described as a fire and forget method analogous to how a homing missile zeroes in on its target.

Thus, it is not a coincidence, but a deliberate design choice, that two of the most well-known open-source AutoML platforms Auto-WEKA [17], [18] and Auto-Sklearn [19] both utilize Bayesian Optimization under the hood.

Furthermore, BO is particularly useful in situations in which evaluation takes a long time (minutes or hours to evaluate) [29], [30], [31].

BO falls under the set of optimization methods known as sample-based optimization methods. Sample based optimization methods improve over *grid search* and *random search* by directly utilizing sample information to guide the optimization procedure [15].

BO is a derivative-free black-box optimization method. It is *derivative-free* since it does not require the use of derivatives. This differs from *gradient descent*, a popular optimization method which utilizes gradient information to find a minimum value of a loss function. It is *Black-box* since it does not presume to know apriori the functional form of the objective function. Rather, it forms a probabilistic model of what the loss function looks like.

In traditional statistics, we assume that some population follows a probability distribution indexed by a fixed set of hyperparameters. Since the entire population is characterized by a distribution and set of hyperparameters, knowing these things enables us to make statements about the entire population [28].

Bayesian statistics provides us with a more flexible framework which enables us to express uncertainty about what the hyperparameters are. Thus, rather than assuming the hyperparameters are fixed and something to be discovered, we assume that the hyperparameters are themselves random variables which follow some probability distribution [28].

Bayesian methods follow a very general framework consisting of three components:

i) Prior: The expert's opinion before any data is collected.
ii) Data: The data the expert collects.
iii) Posterior: The expert's opinion now that data has been obtained.

Bayes rule is simply the means by which the prior is updated using the data to generate the posterior.

BO, which utilizes the above framework, enables us to move from an ***initial configuration*** consisting of Model X and Hyperparameters set Y to a ***final configuration*** consisting of Model X and Hyperparameter set Z that minimizes the objective function. But why stop there? Why limit us to only Model X? Why not let the model itself be a hyperparameter? The goal then becomes to find the configuration consisting of a Model and corresponding hyperparameter set which leads to the smallest cross-validation loss.

The above is called the CASH (Combined Algorithm Selection and Hyperparameter optimization) problem and is exactly the problem solved by Auto-WEKA, which treats the CASH problem as a blackbox optimization problem to be solved via BO. Naturally, there are many implementations of BO such as Gaussian Processes based BO and tree-based BO. Auto-WEKA uses the tree-based BO rather than Gaussian processes, since tree-based BO methods perform well in high dimensional, highly conditional situations, such as that faced in AutoML. For a complete implementation, see [18].

A natural question is how may we select an ***initial configuration***?

Enter meta-learning

Meta-learning assumes that given two identical datasets, a configuration should perform identically.

But what does it mean for two datasets to be identical?

Enter meta-data

Meta-data is data about data.

Meta-learning characterizes datasets themselves by capturing and storing the dataset's meta-data in a database. A record in the database corresponds to a specific dataset. The columns corresponding to this record describe the characteristics of the data. Examples of characteristics are number of columns, number of continuous variables, number of discrete variables, etc. These characteristics represent the data's meta-data [21].

Applying configurations to the Meta-learning database results in a new database which maps configurations to dataset performance.

Thus, if a configuration consisting of Model X and hyperparameter set Y performs well on dataset A, then applying the same configuration to a new dataset B with identical meta-data ought to result in identical performance.

Future areas of research & application

Figure 1 below shows that there has been a marked uptick in searches on Google for AutoML, with roughly linear growth since 2017, indicating that interest in AutoML continues to build worldwide [12].

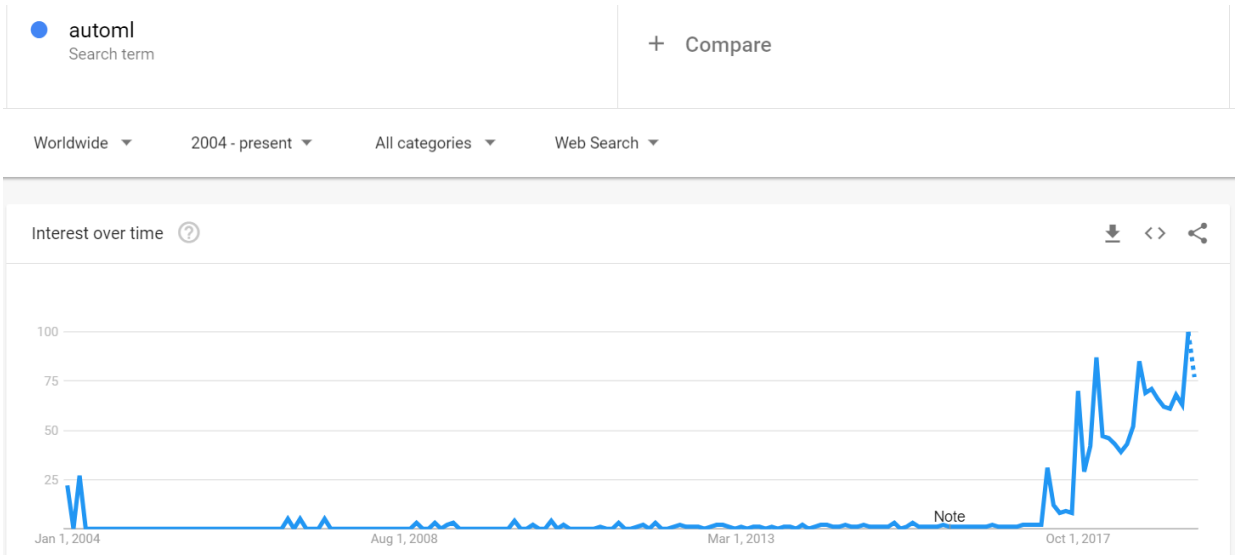

Figure 1

Figure 2 below shows that most of the interest in AutoML has come from Asia, with the US in 11th place.

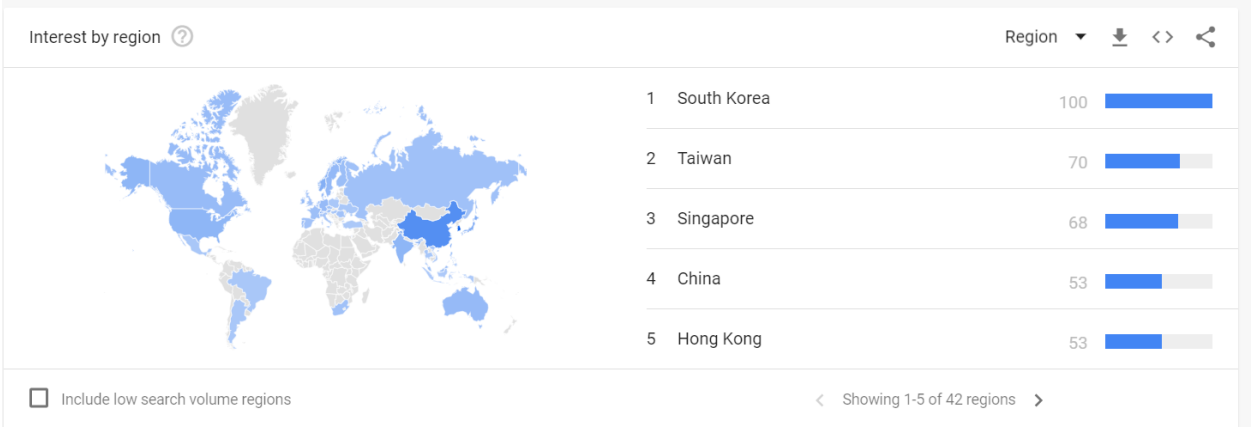

Figure 2

Figure 3 below shows that interest in AutoML from China (50) is almost triple that of the US (20).

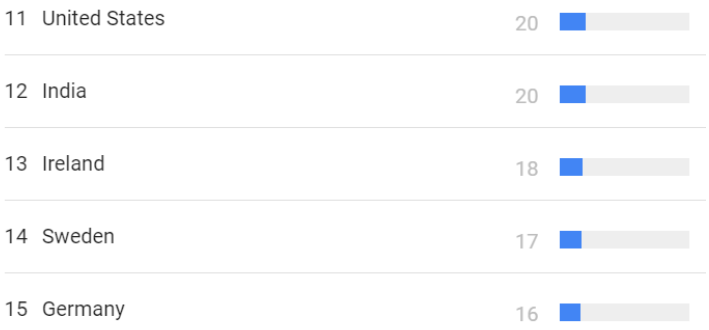

Figure 3

In February of 2019, the President of the US signed an executive order directing the government to prioritize R&D in AI [13], an action widely understood to be in response to the US falling behind China in terms of investment in AI research [14].

Our research shows that if the US is serious about committing to AI research, it should first and foremost direct its resources into the development of true general end-to-end AutoML.

Here at Stats AI, we look forward to partnering with both the public and private sectors and the academic research community to accomplish this task.

References

[1] G. Press, “Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says”, https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#1b2193006f63, 2016.

[2] Glassdoor, https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm, 2019.

[3] D. Ramel, “Supply of More 'Junior' Data Scientists Levels Off Salaries, Says Recruiter”, https://adtmag.com/articles/2017/06/08/data-scientist-salaries.aspx, 2017.

[4] Crowdflower, “2017 Data Scientist Report”, https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport.pdf, 2017.

[5] Figure Eight , “Data Scientist Report 2018”, https://www.datasciencetech.institute/wp-content/uploads/2018/08/Data-Scientist-Report.pdf, 2018.

[6] R. J. Hyndman, “Software I've written”, https://robjhyndman.com/software/, 2019.

[7] A. Balaji, A. Allen, “Benchmarking Automatic Machine Learning Frameworks”, https://arxiv.org/pdf/1808.06492.pdf, 2018.

[8] B. Liu, “A Very Brief and Critical Discussion on AutoML”, https://arxiv.org/pdf/1811.03822.pdf, 2018.

[9] F. Hutter, R. Caruana, R. Bardenet, M. Bilenko, I. Guyon, B. Kégl, H. Larochelle, “AutoML workshop @ ICML'14”, https://sites.google.com/site/automlwsicml14/, 2014.

[10] J. Vincent, “Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day”, https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist, 2016.

[11] T. Peng, M. Sarazen, “2018 in Review: 10 AI Failures”, https://medium.com/syncedreview/2018-in-review-10-ai-failures-c18faadf5983, 2018.

[12] Google Trends, https://trends.google.com/trends/explore?date=all&q=automl

[13] White House, “President Donald J. Trump Is Accelerating America’s Leadership in Artificial Intelligence”, https://www.whitehouse.gov/briefings-statements/president-donald-j-trump-is-accelerating-americas-leadership-in-artificial-intelligence/

[14] C. Metz, “Trump Signs Executive Order Promoting Artificial Intelligence”, https://www.nytimes.com/2019/02/11/business/ai-artificial-intelligence-trump.html

[15] Q. Yao et al. “Taking the Human out of Learning Applications: A Survey on Automated Machine Learning”, https://arxiv.org/pdf/1810.13306.pdf, 2019

[16] M. Zoller, M. Huber, “Survey on Automated Machine Learning”, https://arxiv.org/pdf/1904.12054.pdf, 2019

[17] C. Thornton, F. Hutter, H. Hoos, K. Leyton-Brown, “Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms”, https://www.cs.ubc.ca/labs/beta/Projects/autoweka/papers/autoweka.pdf, 2013

[18] L. Kotthoff, C. Thornton, H. Hoos, F. Hutter, K. Leyton-Brown, “Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA”, https://www.cs.ubc.ca/labs/beta/Projects/autoweka/papers/16-599.pdf, 2016

[19] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, F. Hutter, “Efficient and Robust Automated Machine Learning”, https://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf, 2016

[20] F. Santiago, "Auto is the new black—Google AutoML, Microsoft Automated ML, AutoKeras and auto-sklearn",https://medium.com/@santiagof/auto-is-the-new-black-google-automl-microsoft-automated-ml-autokeras-and-auto-sklearn-80d1d3c3005c ,2018

[21] A. Sharma, "Contest 2nd Place: Automating Data Science",https://www.kdnuggets.com/2016/08/automating-data-science.html, 2016

[22] M. Feurer, A. Klein, F. Hutter, "Contest 2nd Place: Automating Data Science","Contest Winner: Winning the AutoML Challenge with Auto-sklearn”, https://www.kdnuggets.com/2016/08/winning-automl-challenge-auto-sklearn.html, 2016

[23] R. Olson, R. Urbanowicz, P. Andrews, N. Lavender, L. Kidd, J. Moore, "Automating biomedical data science through tree-based pipeline optimization”, https://arxiv.org/pdf/1601.07925.pdf, 2016

[24] A. Nelson, “Cleaning Dirty Data with Pandas & Python”, http://www.developintelligence.com/blog/2017/08/data-cleaning-pandas-python/, 2017

[25] Feature Tools, https://www.featurelabs.com/featuretools/

[26] M. Dowle, https://cran.r-project.org/web/packages/data.table/data.table.pdf, 2019

[27] P. Pandey, “AutoML: The Next Wave of Machine Learning”, https://heartbeat.fritz.ai/automl-the-next-wave-of-machine-learning-5494baac615f, 2018

[28] G. Casella, R. Berger, Statistical Inference, 2nd edition, 2001

[29] P. Frazier, "A Tutorial on Bayesian Optimization", https://arxiv.org/abs/1807.02811, 2018

[30] I. Dewancker, M. McCourt, S. Clark, "Bayesian Optimization Primer", https://app.sigopt.com/static/pdf/SigOpt_Bayesian_Optimization_Primer.pdf,

[31] W. Koehrsen, "An Introductory Example of Bayesian Optimization in Python with Hyperopt", https://towardsdatascience.com/an-introductory-example-of-bayesian-optimization-in-python-with-hyperopt-aae40fff4ff0, 2018