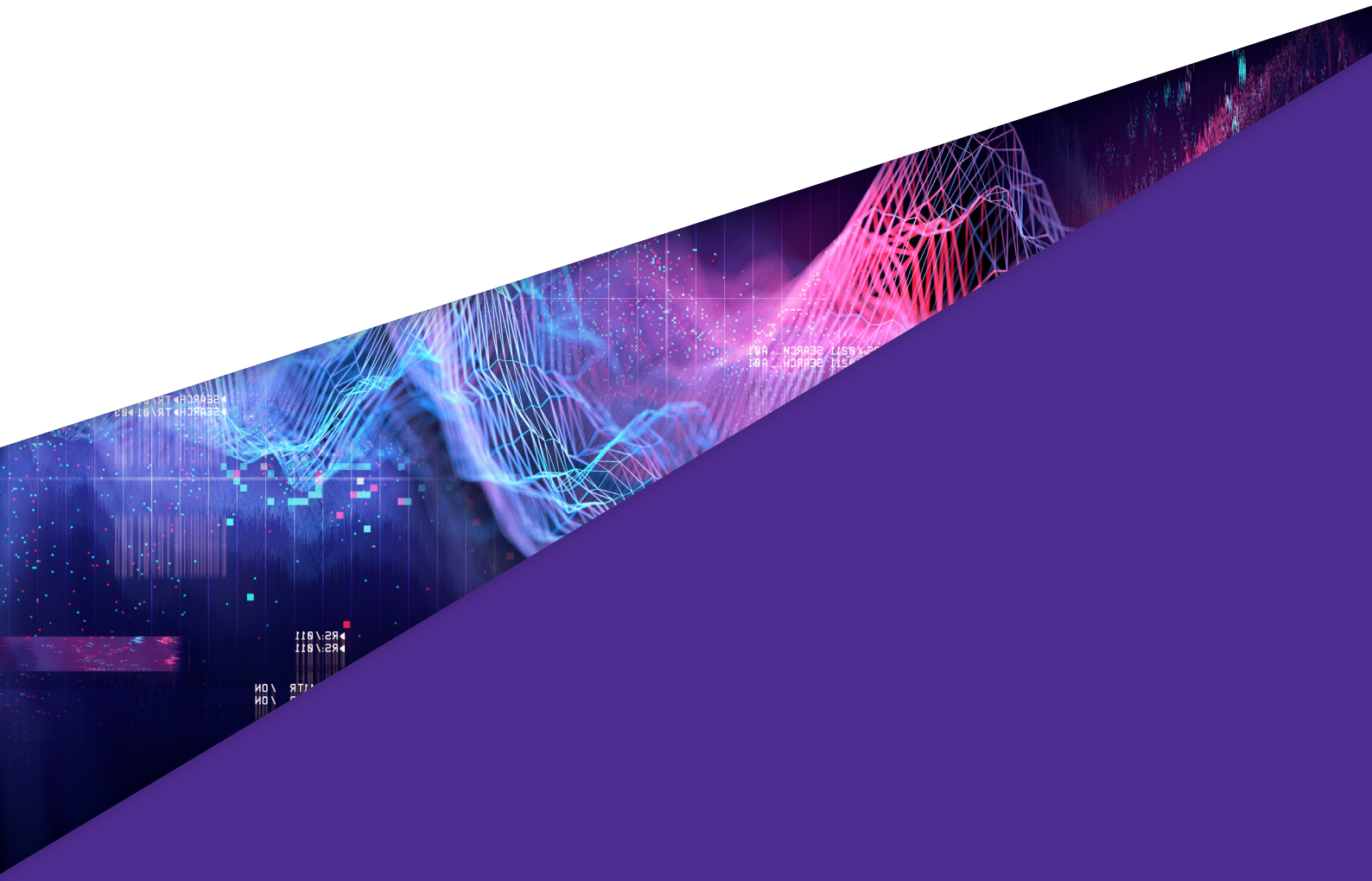
data iku

THE IMPORTANCE OF AUTOML FOR AUGMENTED ANALYTICS

And the Rise of the Citizen Data Scientist



A WHITE PAPER BY DATAIKU

www.dataiku.com

INTRODUCTION

Since the 1960s, machines have been replacing humans[1] by way of automation. Yet it's only in the last few years – with the rise of artificial intelligence (AI), specifically in the enterprise – that automation became a word more synonymous with fear and loss than with development and progress.

It's impossible to deny that when it comes to automation in the enterprise, whether it's inventory management, cyber security vulnerability detection, or any one of the growing examples of business process automation, that technology can save people from the often boring parts of their jobs[2].

In fact, various studies from across industries over the years have shown that not only are humans notoriously bad at repetitive work (they fatigue and make mistakes[3]), but they also (importantly) don't find it fulfilling[4] .

"Today there is a deep mistrust of user-facing automation and automatic AI systems. As a consequence, capabilities that can reduce the human-intensive nature of operations go unused, and analysts and operators often engage in manual unassisted tradecraft because it is what they know and understand."

-HUMAN-AI DECISION SYSTEMS ALEX (SANDY) PENTLAND
MIT MEDIA LABORATORY, CAMBRIDGE, MA[5]

This white paper will take a deep-dive into an important and burgeoning area of enterprise automation across industries, which is automated machine learning (also known as MetaML or, more commonly, AutoML). Specifically, it will explore how AutoML has developed, the larger part it plays in augmented analytics, and what role it will have in the rise of Enterprise AI and the elusive citizen data scientist.

THE ROAD TO AUTOML

In order to understand AutoML's larger impact (like its influence on staff makeup – from data scientist to citizen data scientist – as well as the types of work different profiles are doing), it's helpful to take a step back and understand more completely what AutoML actually is and how it came to be.

At a very high level, AutoML is about using machine learning techniques to automatically do machine learning. Or in other words, it means automating the process of applying machine learning. Early on, AutoML was almost exclusively used for the automatic selection of the best-performing algorithms for a given task and for tuning the hyperparameters of said algorithms (in a nutshell, hyperparameters are like knobs that need to be tuned when tuning a machine learning model - find a more in-depth definition here[6]).

At a very high level, AutoML is about using machine learning techniques to automatically do machine learning.

The oldest open-source AutoML library, AutoWEKA, was released in 2013 and was quickly followed by many others, including auto-sklearn and H2O AutoML.

Today, automated analytics can add efficiency to large swaths of the data pipeline.

The practice of applying automation to the data science process has been around for more than five years now. However, these technologies still narrowly focus primarily on algorithm selection and hyperparameter tuning, which is helpful in automating some of the work of data scientists, but not very useful for, say, the day-to-day work of a data analyst.

Yet AutoML can have a broader scope with later versions of auto-sklearn and tpot (and has). Its development has spurred the application of automation to the whole data-to-insights pipeline, from cleaning the data to tuning algorithms through feature selection and feature creation, even operationalization. At this larger scale, it's no longer AutoML, but augmented analytics. Today, automated analytics can add efficiency to large swaths of the data pipeline, with the potential to impact the entire process and influence the structure of data teams long term.

AUGMENTED ANALYTICS & ENTERPRISE AI

Indeed, the vision for the future of augmented analytics is one of complete (or nearly complete) automation, where one could feed a dataset and a target to an automated pipeline and get back cleaned data with engineered features, together with the best performing model on top. This automation of machine learning projects – whether those projects touch company operations, processes, or product development – is the essence of the idea of Enterprise AI and would allow for greatly accelerated AI modeling.

If Enterprise AI means the ability to embed AI methodology into the very core of the organization, then AutoML is essential to that vision and to companies trying to get to that end. In other words, Enterprise AI and augmented analytics (and by extension, AutoML) are intrinsically linked as a goal and a way to get there.

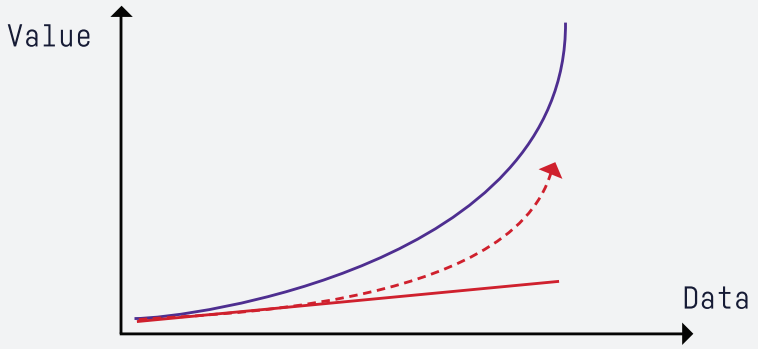
Does all of this necessarily mean that Enterprise AI and augmented analytics render the job of a data scientist obsolete? Well – yes and no. Yes in that the work they're doing today probably will not be the work they will be doing five years from now, but no in that data scientists will still exist and be essential for other specialized and high-impact tasks (more on this topic later – see section The Shifting Role of the Data Scientist).

	Data Preparation & Feature Engineering	Algorithms Comparison & Parameter Optimization	Validation & Assembling
The Manual Way	Write code to enrich, parse, and normalize each individual variable, then manually combine and select them.	Test different algorithms and different parameters.	Choose a way to cross-validate and possibly run through multiple train/evaluate strategies on subsets. Test different ways to combine and tune trained models to get meaningful and optimal results.
The AI Way [vis-à-vis AutoML]	Automatically transform any type of variable and automatically select and combine them for a optimal representation for learning.	Automatically compare algorithms & parameters, preselecting only those that make sense given the data, and select the best performing one(s).	Automatically choose adequate cross-validation techniques, evaluate variable contribution, and protect against data leaks. Automatically combine models to get optimal models, and re-tune them to get an optimal, stable output.

But perhaps more interesting is to look at is what this will mean (and indeed, is already starting to mean) for employees who are data- and business-savvy – whether they be on the business side or some kind of data analyst – but not formally trained data scientists. Here is where the story of the citizen data scientist plays a role in how these profiles will shift the landscape of responsibilities when it comes to Enterprise AI.

THE RISE OF THE CITIZEN DATA SCIENTIST

As the amount of data owned by organizations began (and continues to) to take off and grow exponentially, those businesses often find themselves stuck with data insights that increase only linearly, unable to capitalize on the massive amounts of data at their disposal.



New data = EXPONENTIAL
New insights = LINEAR

Turning linear insights into exponential insights (otherwise known as accelerated modeling) is critical for the movement into Enterprise AI, and it's a matter of scale – using more available data for more data projects, faster (and, of course, automating whatever and wherever possible).

Generally this happens best by implementing a combination of machine learning model operationalization and self-service analytics programs[7]. But in any case, it cannot happen without expanding the breadth of people that have access to and work with data on a day-to-day basis.

Yet for most companies, hiring exponentially more data scientists (who are not only expensive, but notoriously difficult to find, hire, and – importantly – keep on staff) is out of the question. And it is out of this combination of cultural shift toward a data-driven culture and economic reality of data scientist hireability that the citizen data scientist is born.

Gartner predicts that, by 2020, due in large part to the automation of data science tasks, citizen data scientists will surpass data scientists in the amount of advanced analysis produced.

- GARTNER, HYPE CYCLE FOR THE DIGITAL WORKPLACE, 18 JULY 2018

In fact, top companies who are well on the path to Enterprise AI – like Pfizer[8] and Daimler[9] – are shifting toward this model in order to provide the scalability necessary to support rapidly accelerated data efforts and growing number of machine learning projects in production.

THE SYNERGY OF CITIZEN DATA SCIENTISTS & AUGMENTED ANALYTICS

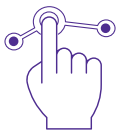
AutoML existed before the rise of the citizen data scientist, and similarly, citizen data scientists can (and do) exist without leveraging augmented analytics. But it's their synergy that makes them powerful – that is, their interaction produces a combined effect greater than the sum of their separate effects.

Having citizen data scientists do more advanced work vis-à-vis AutoML or, more broadly, augmented analytics, frees up data scientists to work on more specialized tasks, which is ultimately a win all around:

- A win for citizen data scientists (whether they be analysis or business users), who can contribute with more valuable (and less mundane) work.
- A win for data scientists, who can automate or leverage citizen data scientists for simpler tasks and stay focused (read: interested and not leaving the organization) on more challenging projects and tasks.
- A win for the business, who thanks to this arrangement, can scale data efforts and release more data projects with a large staff of many citizen data scientists supported by some data scientists – in other words, accelerated AI modeling.

Teams can reach this synergy, but not without some changes. Regardless of actual title (citizen data scientist, data analyst, business analyst, marketing analytics manager, etc.), incorporating the work of non-data scientists into data projects in meaningful ways requires a fundamental shift in mindset around data tooling. By nature, these profiles generally don't have the skills for advanced feature engineering, parameter optimization, algorithm comparison, etc. What they do bring to the table is intimate knowledge of the problems at hand and business questions that need to be answered.

But as AutoML and, more importantly, augmented analytics technologies arise, not everyone involved in the data pipeline needs to have these skills – parts of the data pipeline can be automated. But in order to add such efficiencies, AutoML - and augmented analytics more broadly – must necessarily become applicable outside of the role of data scientist by adding:



Usability: The system should be easily usable by non-developers with minimal technical skill. Look for a system that supports augmented analytics by providing contextual help and explanation for different parts of the data process and a visual, code-free user interface.

Stability: Users without intimate knowledge of data storage technologies need to be able to execute augmented analytics using a system that can be reliably leveraged from one step of the data pipeline to another. For example, local changes in the data should not result in complex change of algorithms (the system should make these changes seamless).

Transparency: Harkening back to the introduction of this paper, it's important to consider that it's difficult for anyone to trust something - especially technology - that they don't understand. So with AutoML, providing a system that gives an accurate description of algorithms used (and why they were chosen) provides the right level of knowledge necessary for citizen data scientists not only to trust outcomes, but to determine if they are right for the project at hand (and make adjustments if not). And transparency is important not just when it comes to details about algorithms - insight into the entire process (from raw data up until project deployment) when it comes to augmented analytics is an essential component of trusting results and being able to property supplement automation with business knowledge where and when it matters.

Adaptability: Even though the idea behind augmented analytics being leveraged by citizen data scientists is that they can contribute to data projects on their own, it doesn't mean that the projects they build won't be used or ever touched by others (namely data scientists). The automated system needs to be able to be used as a starting point for custom development and dedicated learning by data scientists - for example, outputs should be able to be translated into Python code for the full learning, including feature transformation and cross-validation.

Of course, it cannot be overstated that adding these features in an augmented analytics or AutoML platform doesn't mean that just anyone should be able to create models and push them into production without any kind of oversight, review, or input from someone extremely specialized in the field (like a data scientist).

Even with augmented analytics, it is still very possible to make big mistakes – like feeding it the wrong data – and there is nothing that will tell you that this is not the data you'll be receiving live when you deploy. So it's worth noting that advancements in automation don't replace data scientists; rather, they change the role (more on this later).

AUGMENTED ANALYTICS:

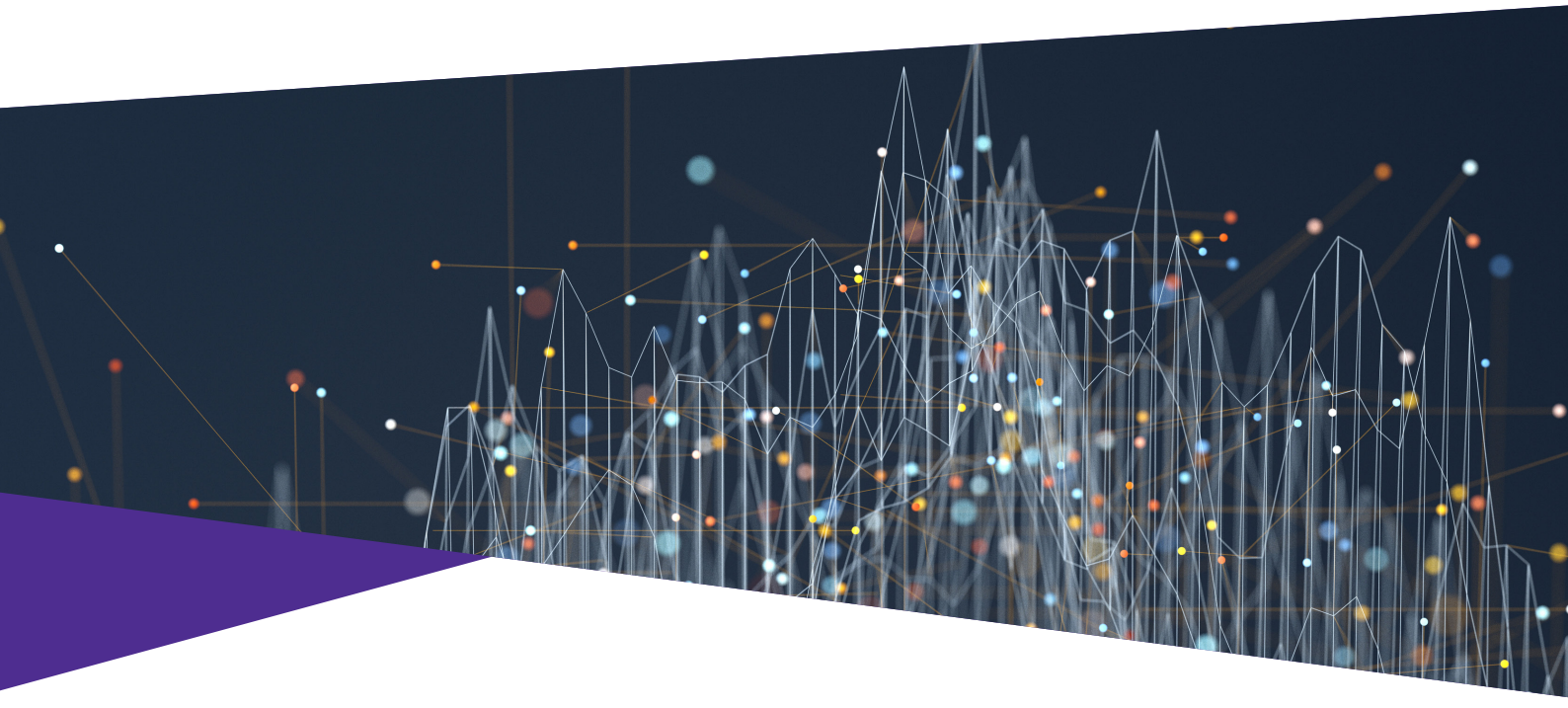
DIFFERENT DEGREES OF AUTOMATION

The vision for the future of augmented analytics is one of complete (or nearly complete) automation, but it's important to point out that it is still just that: a vision. The reality of most augmented analytics or AutoML tools and systems today is that they are not completely automatic – yet.

And this is probably a good thing, for now – there is still a learning curve as citizen data scientists and even data scientists adjust to this shift and ensure that quality of data projects remains high even as processes become increasingly automated. However, teams can start to evaluate how automated their systems are from an end-user perspective using this scale:

Manual	The system does not help you to do it - code your way in!
Tooled	The system provides tools or components that can be combined to perform the task.
Assisted	The system helps or guides you along the way in a simple fashion, but some important choices are still up to you.
Auto	The system does everything, end-to-end.

These levels of automation may also be useful for organizations evaluating systems that offer augmented analytics or AutoML capabilities – do the contenders offer the level of automation expected or required based on the needs of the team and the business?



THE SHIFTING ROLE OF THE DATA SCIENTIST

One of the main criticisms of the rise of the citizen data scientist (see the infamous article The Mirage of a Citizen Data Scientist by Gregory Piatetsky[10]) is that it implies that any untrained person can simply step in and do the job of a data scientist, thus devaluing their (highly paid) skills.

Yet it's often cited – and widely confirmed – that data scientists today spend about 80 percent of their time on relatively mundane tasks like finding the right data, cleaning it, and wrangling it. Data scientists by nature are curious people, always looking to use the next cool technology or work on the next challenging project. And this nature doesn't exactly jive with the current reality.

So if the AI dream would be to feed a dataset and a target to an automated pipeline and get back cleaned data with engineered features, together with the best performing model on top, that might eliminate most of the work data scientists are doing today – but wouldn't that be nice? Because a lot of the work they're doing today frankly isn't the fun, creative, or interesting part of the job and can be automated. Ask any data scientist, and you'd be hard pressed to find one that disagrees.

Additionally, it's worth re-emphasizing that providing citizen data scientists with the right tools to empower them in smart ways with this part of the data pipeline doesn't mean that they will be able to push any data project into a production environment without working with an expert – that is, a formally trained data scientist.Over time, that means the role of the data scientist will morph from more of a one-man show, spending 80 percent of his or her time on data preparation, to one who refines and fine-tunes projects worked on by citizen data scientists (with the help of AutoML) throughout the data pipeline.

And even with accelerated AI modeling meaning that the number of projects completed will skyrocket, this movement toward complete data team collaboration aided by augmented analytics means data scientists will have time to also dive into the most challenging projects at the business requiring a true specialist, experimenting with new technologies and techniques.

It's also worth noting that a critical part of this equation is to empower citizen data scientists in smart ways. That doesn't just mean allowing them to crank out models without proper training or understanding of the process such that those models are totally disconnected from the business questions they're trying to answer. As pointed out previously, the element of transparency is an important one to rectify this concern both when making this transition and when choosing a tool for augmented analytics or AutoML.

THE PSYCHOLOGY OF SHIFTING ROLES AND ORGANIZATIONAL CHANGE

With the advancement of AutoML and augmented analytics as well as the rise of the citizen data scientist, thus the outmoding of much of the work of today's data scientist and shifting in roles, the next logical question is: what's next? Will the citizen data scientist then be outmoded as well?

It's important that the implementation of augmented analytics not degrade the quality of data projects or models that they produce.

And how can companies deal with the dichotomy between implementing technology that will speed up data processes but also potentially alienating employees working on data projects who have negative feelings toward the augmentation or automation of their work?

The key in this puzzle is time. It's only natural for humans to fear technologies that they feel might fundamentally change – or worse, eliminate – their work. Clearly, abruptly flipping a switch to full end-to-end data project automation is not the answer (both in terms of organizational change management and in terms of bottom-line best practice for augmented analytics implementation).

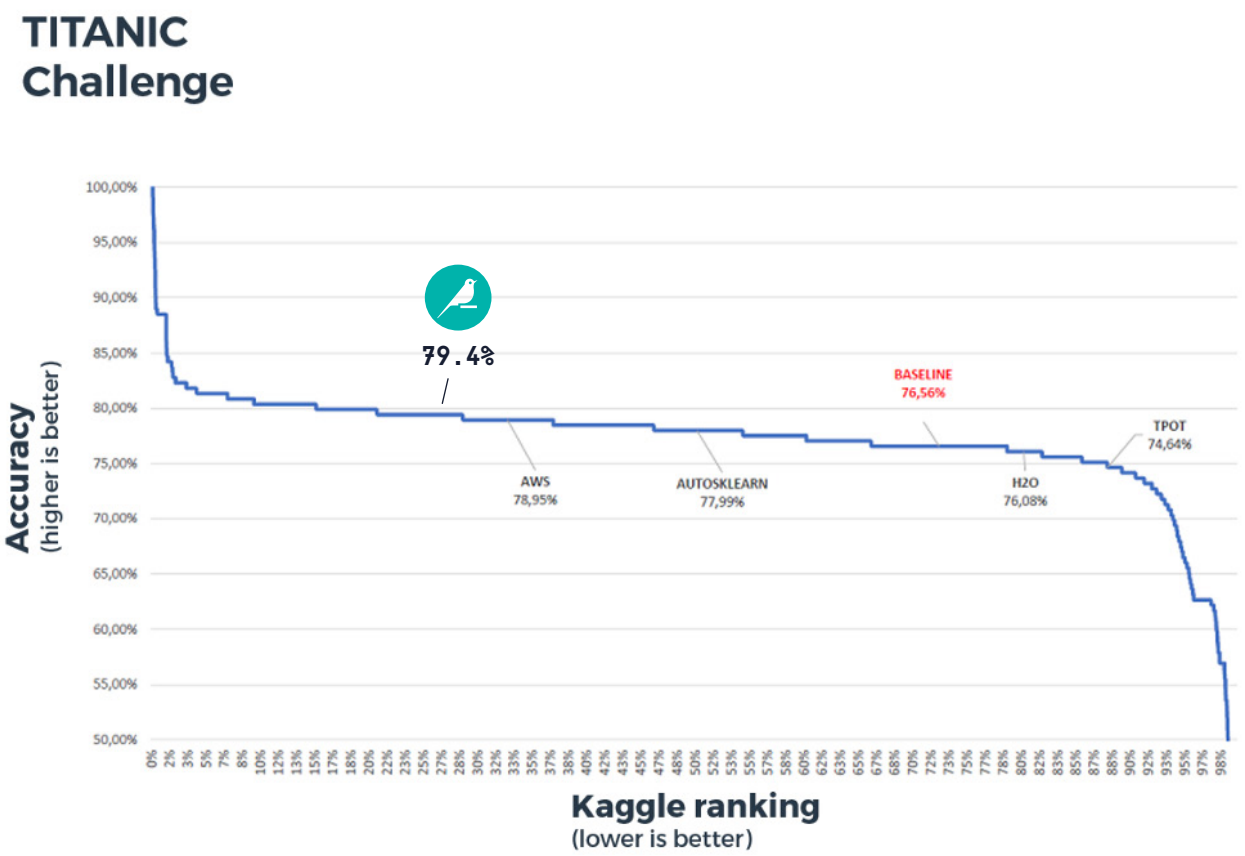
Instead, the recommended approach is a gradual shift toward augmented analytics, starting perhaps first at the core with AutoML and slowly expanding augmentation capabilities from there. Again, this helps both citizen data scientists and data scientists adjust workflows accordingly, but it also allows for – at an organizational level – time between changes to evaluate effects on the business itself. In other words, it's important that the implementation of augmented analytics not degrade the quality of data projects or models that they produce.

CASE STUDIES:

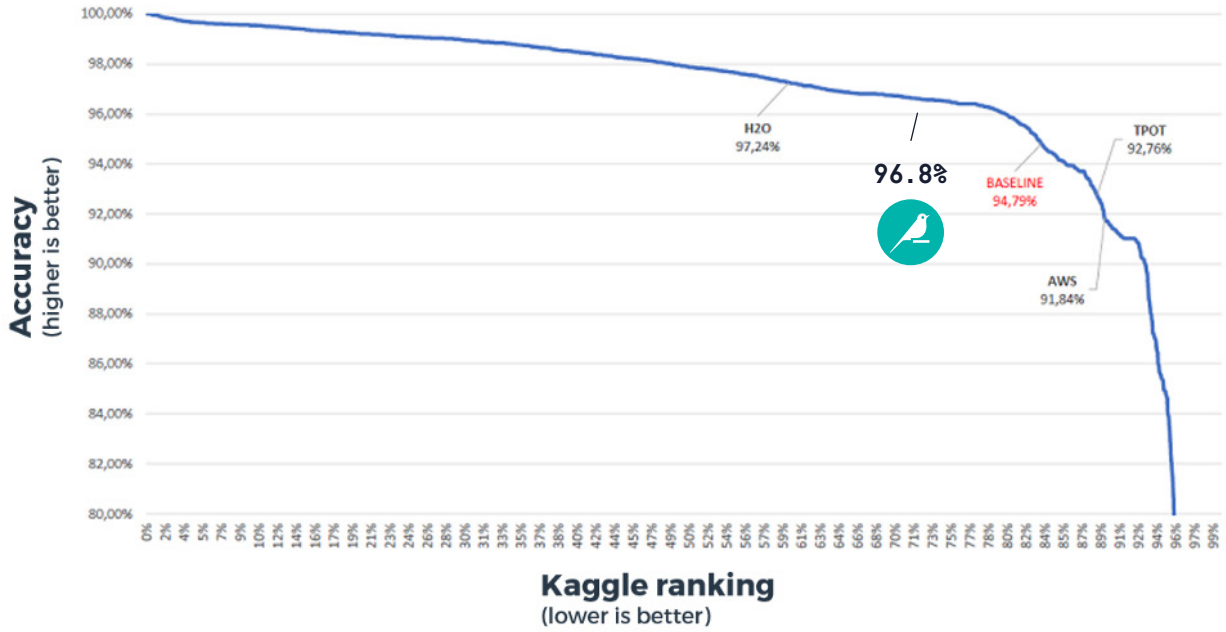
AUTOML VS. HUMANS, PUT TO THE TEST

In order to fully understand the power of AutoML (and augmented analytics, by extension), it's valuable to see how it compares to humans completing the tasks from scratch. A group of data scientists at Dataiku (the centralized data platform that moves businesses along their data journey from analytics at scale to Enterprise AI) took the CDiscount challenge and put Dataiku's AutoML capabilities to the test.

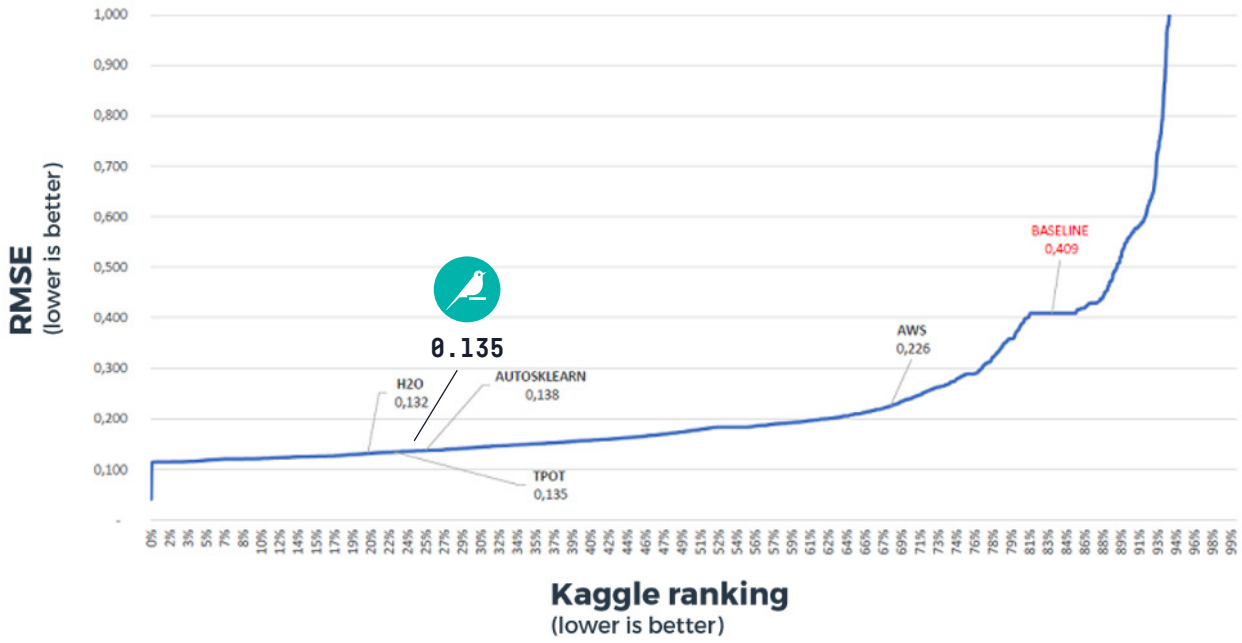
The group used only Dataiku visual AutoML on three different Kaggle challenges[11] , and here are the results (in the following diagrams, "baseline" is a human:



MNIST Challenge



HOUSE PRICES Challenge



FEATURE:

AUGMENTED ANALYTICS WITH DATAIKU

Since the early days of the product (all the way back to 2014), Dataiku has proposed a visual machine learning suite that guides the user through all of the machine learning steps (train-test split, feature handling, metrics to optimize, and different templates of pre-set algorithms).

In a nutshell, the interface offers a one-button option, simply called «Train» – this will automatically infer the feature handling, pre-select a collection of algorithms, and returns the best performing one. But of course, it's still up to the user to tune those parameters and select the best possible settings based on their experience.

In addition to this basic functionality, Dataiku offers:

DATA PREPARATION AND FEATURE ENGINEERING

Feature	Level of Automation	Details
Normalization of Numerical Data	Auto	Automated statistical normalization, per variable distribution
Normalization of Textual Data	Auto	TFIDF/ Text Normalization, Vectorization as add-on. Coming soon: Vectorization (Word2Vec) as built-in
Normalization of Time Data	Assisted	Parse date, extract time based data in a point of click manner. Coming soon: Automatic extraction and selection based on calendar information, business hours, etc.
Enrichment with Third-Party Data	Assisted	Join with third party data based on Key
Normalization of Geographical Data	Assisted	Match geographical position, match per geo point, country, etc.
Normalization of Image Data	Tooled	Add-on for image data vectorization. Coming soon: Automatic selection of pre-trained embedding model
Feature Selection	Assisted	Feature reduction via correlation with target, tree-based, principal components analysis (PCA), and LASSO regression
Feature Transformation Selection	Manual	Coming soon: Ability to test / select several transformations by original feature
Feature Space Transformation	Assisted	PCA Transformation
Feature Generation	Assisted	Generate feature by numerical combination. Coming soon: Automatic feature generation and selection by interaction and impact mapping

DATA PREPARATION AND FEATURE ENGINEERING

Feature	Level of Automation	Details
Support for tree-based methods	Auto	Support for Random Forest, Random Boosted Trees, etc., From Scikit / MLLib / H2O
Support for neural network algorithms	Auto	Support neural networks (from H2O and TensorFlow) Coming soon: GPU Monitoring Support
Support for out-of-memory/ distributed algorithms	Auto	Support for MLLib / H2o Algorithms + training in scalable Docker / Kubernetes clusters
Model selection	Auto	Automatically selects the best model per the metric selected by the user or per a business metric
Smart bootstrap	Assisted	Find an optimized starting point for optimization algorithms, leveraging knowledge gathered on other datasets
Support for grid search	Auto	Search exhaustively for optimal parameters
Support for random search	Auto	Search for optimal parameters using smart, non-exhaustive search
Support for resources/ time-based learning	Assisted	Automatically stop the search, optimizing per the amount of resources / time available

VALIDATION AND ASSEMBLING

Feature	Level of Automation	Details
K-folders and cross validationt	Assisted	Support for cross-validation schemes and K-fold
Time-based cross validation	Tooled	Enabled cross-validation per a variable or existing datasets, in a point-or-click fashion
Variable importance and contribution	Auto	Analyze the importance and contribution of each variable
Data leak detection	Assisted	Automatically detect data leaks and remove such variable
Unbalanced datasets	Auto	Automatically detect unbalanced datasets and leverage an adapted cross-validation strategy
Merge models	Assisted	Ability to merge existing models in a point-and-click fashion
Model calibration	Manual	Tune the model in order to match the behavior of the model to the expected behavior. Coming soon: Automated model calibration based on real distribution

CONCLUSION

The value the citizen data scientist still brings even in the theoretical world of completely augmented (that is, automation from raw data to production) analytics is the deep business knowledge that allows them to determine which business questions are important to answer, and then whether that answer adequately addresses the problem at hand. Certainly, there will always be a human involved.

However, it is not a stretch to imagine a next step of AutoML which some parts of the process pre-data collection. For example, a system that suggests possible data projects based on what kind of data is available to an enterprise (e.g., it looks like you have CRM data and transaction data, would you like to predict customer churn?). A sort of recommendation engine for data projects, if you will.

Another area of AutoML that is still very much in the realm of "what's next" and not "what's now" is in the area of deep learning. For many reasons, AutoML for deep learning is much more challenging, which means that we're probably a bit farther away from automating it. However, there is lots of work being done - here is a good resource and overview for those curious to learn more[12] .

All of this to say that progress in automation will continually move forward, which means those not taking advantage will fall farther behind in model creation and ability to scale, hindering the path to Enterprise AI. AutoML is one of the essential keys to the future of artificial intelligence, and now is the time to get started on this shift.

Endnotes

1 https://robohub.org/the-evolution-of-assembly-lines-a-brief-history/
2 Insert claire's article on AI + boredom from blog, when published
3 https://www.ncbi.nlm.nih.gov/pubmed/7245925 or https://www.techtransfer.com/blog/err-human/
4 https://www.emeraldinsight.com/doi/abs/10.1108/03074801211273939?fullSc=1&journalCode=nlw
5 https://thehumanstrategy.mit.edu/blog/human-ai-decision-systems
6 https://towardsdatascience.com/what-are-hyperparameters-and-how-to-tune-the-hyperparameters-in-a-deep-neural-network-d0604917584a
7 https://pages.dataiku.com/operationalization-ssaa
8 https://pages.dataiku.com/scaling-data-science-for-a-global-analytics-team
9 https://pages.dataiku.com/building-immersive-data-function-large-organizations
10 https://www.kdnuggets.com/2016/03/mirage-citizen-data-scientist.html
11 https://techblog.cdiscount.com/a-brief-overview-of-automatic-machine-learning-solutions-automl/
12 https://towardsdatascience.com/everything-you-need-to-know-about-automl-and-neural-architecture-search-8db1863682bf

data iku

Your Path to Enterprise AI

Dataiku is the centralized data platform that moves businesses along their data journey from analytics at scale to enterprise AI. Data-powered businesses use Dataiku to power self-service analytics while also ensuring the operationalization of machine learning models in production.

20,000+

ACTIVE-USERS

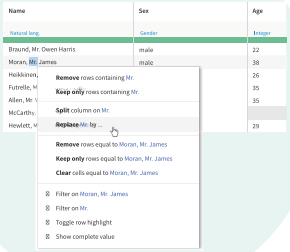
*data scientists, analysts, engineers, & more

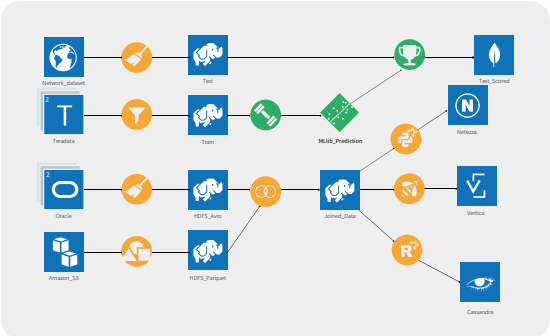
200+

CUSTOMERS

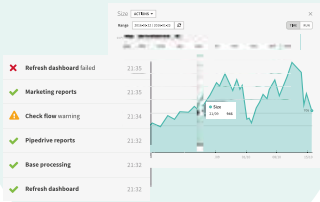
BNP PARIBAS · Santander · paloalto NETWORKS · PREMERA BLUE CROSS · FOX NETWORKS GROUP · Unilever · NXP · SEPHORA

1. Clean & Wrangle



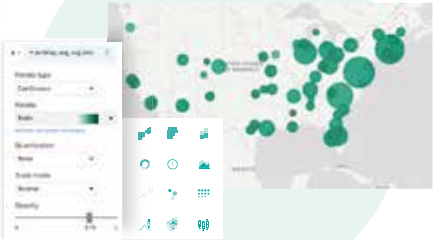


5. Monitor & Adjust

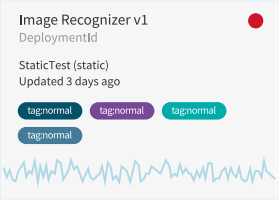


2. Build + Apply Machine Learning

3. Mining & Visualization



4. Deploy to production



WHITE PAPER