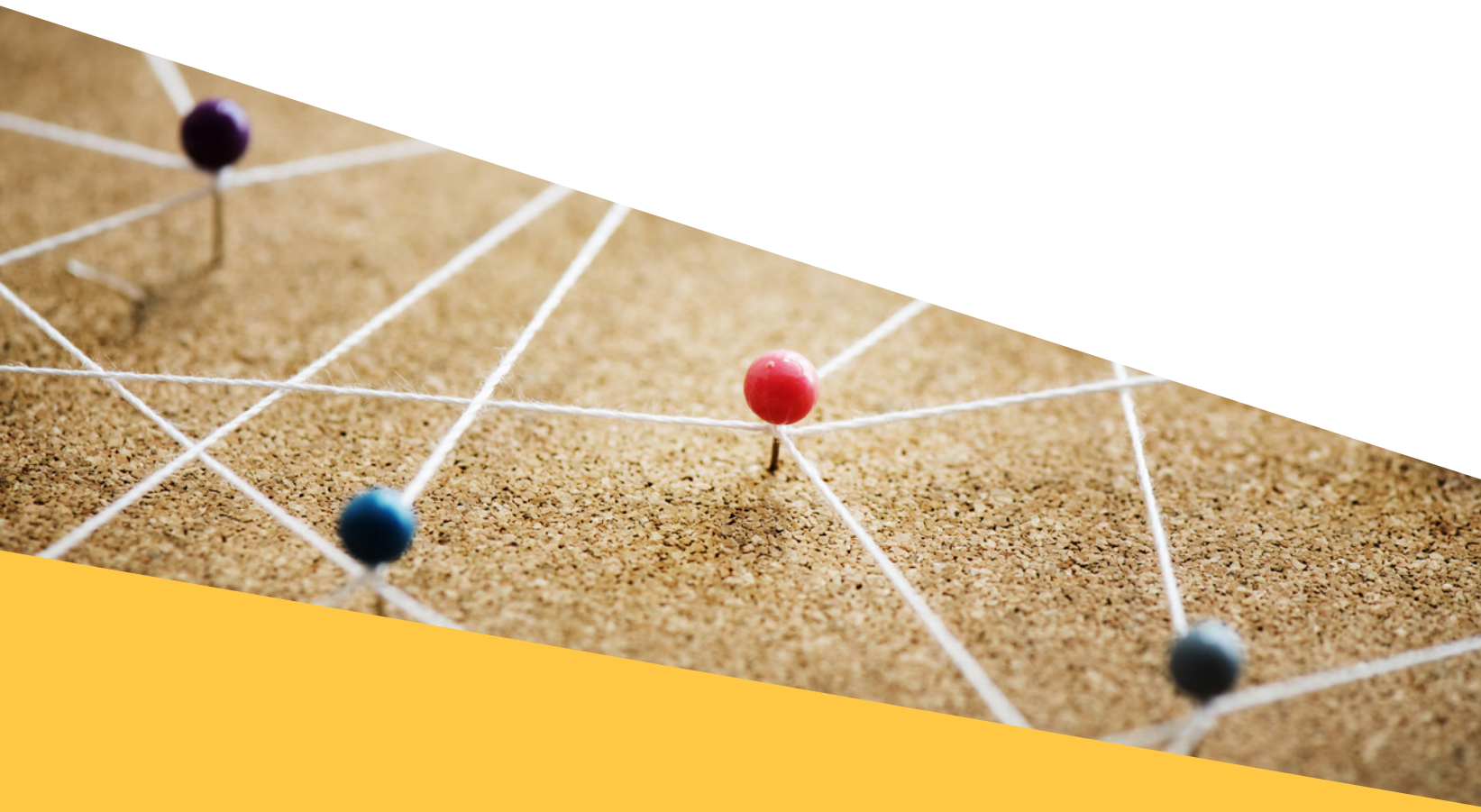
How to Improve Data Quality

With an Efficient Data Labeling Process



WHITE PAPER

www.dataiku.com

Introduction

The good news: Data is a foundational requirement to any successful machine learning project, and there is no shortage of diverse data sources at the disposal of today's organizations. They leverage data for varying project scopes, both day-to-day and long term — from high-level, more visualization-based efforts to lengthy, in-the-trenches machine learning models.

The not-so-good news: Organizations across all industries face a myriad of challenges when it comes to data quality, including but not limited to unlabeled data, poorly labeled data, inconsistent or disorganized data, an inundation of data sources, a lack of tools to adequately address data quality issues, and process bottlenecks. As data gets even cheaper to collect and store, it will become even more of a burden for data professionals to plow through mass amounts of raw data.

Many of these data concerns stem from a lack of data governance within the organization, illustrating the need for a sound data governance strategy that allows existing data storage systems to be connected in a centralized, controlled environment and provides enterprise-level security.

Not only does the way people work with data need to be consistent and secure, but the data itself needs to be high quality and properly labeled to alleviate time (and talent) wasted on unnecessary or repetitive steps. Ensuring a thorough data cleaning process allows for more efficient data wrangling, workflow visualization, and model deployment processes. The refreshing news is, though, that adopting AI tools and processes can help usher in defining improvements to data quality.

Why Data Labeling Matters

In order to create successful machine learning models, organizations need tools and people to enrich various datasets so models can be trained, validated, and ultimately, operationalized and scaled.

Bad data — missing data, errors, unlabeled data, and so on — can skew results, making it harmful to overall AI efforts. By preparing data, both features and labels, in an efficient way, models can perform better, increasing the business value of the output.

A significant part of data science involves identifying areas of the data-to-insights pipeline where efficiencies can be identified. Active learning — a process that automates data labeling through machine learning algorithms — helps address one such inefficiency: bottlenecks associated with data cleaning and labeling. When these processes are usur ping a massive volume of time and resources from teams that could instead be building their next machine learning models or pushing models into production, active learning can be used to minimize the number of labeling tasks necessary.

From a business perspective, data labeling is never usually a budget-friendly undertaking, so any steps organizations can take to reduce costs associated with data labeling can prove beneficial. With an abundance of unlabeled data and budget constraints to get the data labeled effectively, implementing an active learning strategy to identify which rows of data should be labeled can help maximize model performance subject to such established budget boundaries.

The market for AI and machine learning data preparation solutions is slated to reach $1.2 billion by the end of 2023.[1]

In this white paper, we will discuss the significance of data quality in any end-to-end AI project, with a specific focus on the need for data labeling through active learning. Key topics will include:

- The benefits of active learning, namely the ability to lower the number of labor-related tasks and cost of data labeling necessary for a model to reach the required accuracy
- Challenges associated with active learning and how to surmount them
- Use cases that illustrate the massive business opportunity active learning presents and why having labeled data is such a valuable asset

[1] https://www.cognilytica.com/2019/03/06/report-data-engineering-preparation-and-labeling-for-ai-2019/

What Is Active Learning?

In the early 2000s, data labeling was a frequently discussed topic with an influx of literature produced around it. However, it gradually fizzled out over time, as organizations weren't presented with actionable ways to jumpstart their end-to-end data efforts, including cleaning and labeling. Major corporations like Google and Facebook had the budget and infrastructure to have their data labeled for them, whether it was done in-house or was outsourced. Researchers also prioritized experimenting with data labeling for their more pressing use cases. However, as more and more organizations are incorporating AI strategies into their company's blueprint today, data labeling is having a resurgence.

Before running any cutting-edge machine learning algorithms or deploying any models, the data involved needs to be high quality and labeled. However, data collection, including cleaning and wrangling, is a tedious, time consuming, and iterative process that typically involves data labeling and model training.

According to a 2019 survey from Cognilytica[2], data preparation and engineering tasks represent over 80% of the time consumed in most AI and machine learning projects. Many techniques exist to help make the data labeling process more efficient, which is where active learning comes into the picture.

Active learning is a framework allowing users to reduce the cost of data labeling necessary for a model to reach the required accuracy. It can be used:

- When not all data can be annotated because it is too costly or complicated.
- To speed up the labeling procedure by leveraging previously labeled data.
- To optimize the order in which unlabeled data is processed.

The active learning algorithm can be summarized in the following steps:

1. Label enough data to train an initial model.
2. Train the initial model on labeled data.
3. Pick the batch of data samples that the model is most "uncertain" about.
4. Label each sample and add it to a training set.
5. Retrain the model on a new training set.
6. Monitor labeling process and performance accuracy.
7. Repeat from step two onward until the required accuracy is reached.[3]

Active learning is the science of applying algorithms to automate data labeling.

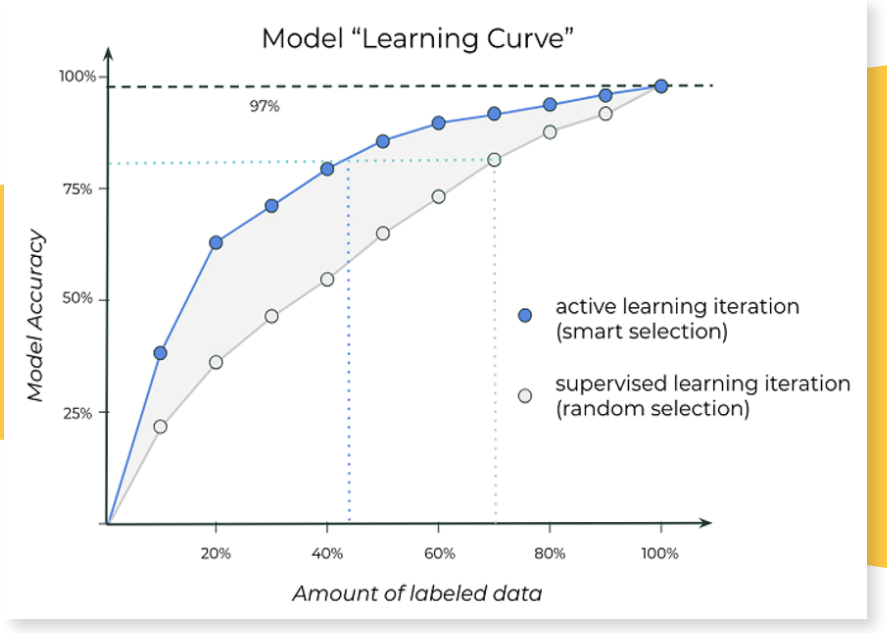
[2] https://www.cognilytica.com/2019/03/06/report-data-engineering-preparation-and-labeling-for-ai-2019/
[3] https://towardsdatascience.com/learn-faster-with-smarter-data-labeling-15d0272614c4

When it comes to filtering your email, for example, machine learning can effectively filter spam with an 80-90% accuracy. That level of accuracy is optimized when the human user corrects the machine's output by relabeling messages that are in fact *not* spam, and vice versa. The relabeled messages are fed back into the classifier's training data to be fine-tuned with future emails.

By using a classifier that estimates its own uncertainty (whether an email is spam or not spam) and asks the user for labeling feedback only at the most uncertain times, the labels will be more effective at training the classifier than randomly selected ones. Over time, the classifier learns and more effectively identifies what should and should not be classified as spam.

On a fundamental level, active learning algorithms can potentially achieve a higher level of accuracy while using fewer training labels if they have the ability to choose the data they want to learn from. The algorithms can pose queries during the training process, such as with unlabeled data instances, which will then be labeled by a human annotator — signifying a true example of the importance of human-centric AI.

Learn More With Less



The more data that is collected and used for model training, the higher the model's accuracy — regardless of whether active learning is used or not. However, in instances where active learning is used, accuracy can still be achieved while using significantly less data. In the image to the left[4], an 80% accuracy can be achieved with only 45% of the total volume of available data versus 70% in the case of regular supervised learning.

[4] https://www.kdnuggets.com/2018/10/introduction-active-learning.html

Benefits of Active Learning

Active learning can be incredibly useful, particularly in instances where there is a significant amount of unlabeled data that would be extremely expensive or impossibly time consuming to label by hand. Resultantly, it can reduce bottlenecks that arise during the data labeling process by selecting the most informative data sample to label next, leading to enhanced performance and cost effectiveness for labeling projects.

According to O'Reilly[5], active learning works best in cases where there's plenty of cheap, unlabeled data, such as tweets, news articles, and images. With the example of fraud detection that we will highlight later in the paper, expert investigation is necessary for labeling, demonstrating why the process is so costly. Due to the fact that data labeling is such an expensive undertaking, knowing what and how much to label is critical to the process. Organizations should label only the data that will have the greatest impact on a given model's training data.

Further, active learning can drive value to various personas at an organization, not just those performing the hands-on machine learning functions, such as data executives who tend to be more concerned with achieving higher-level KPIs of the business. While the technical aspect is crucial and a data science team will set up the experiment to see if the labeling performance is improving over time, data labeling represents a significant business opportunity for cost reduction.

[5] https://visit.figure-eight.com/rs/416-ZBE-142/images/Real-World_Active_Learning_CF.pdf?src=CIOMagazine&medium=CPC&campaign=Real-World-Active-Learning&content=Whitepaper&term=Asset

At a certain point, continuing to label a pool of unlabeled data samples will no longer drive additional value to a given project. By using active learning to maintain a pulse on when to stop labeling or to perform labeling in a budget-friendly way, data science teams can gain buy-in from data team leads and executives to demonstrate the tangible business impact of AI efforts and open the door to additional use cases.

Interestingly, a 2019 Dataiku survey revealed that **29% of IT professionals** across various industries plan to implement active learning within the next year. This trend will likely continue to grow as more businesses recognize the vast scalability associated with data science and machine learning platforms as well as the importance of democratized AI as a business-defining principle.

Go Further With Dataiku

When you need to manually label rows for a machine learning classification problem, active learning can help optimize the order in which you process the unlabeled data. Dataiku's ML-assisted plugin enables active learning techniques in Dataiku by aiding the labeling process whether data is tabular, images, or even sound.

Check Out the Active Learning Plugin ↗

How Active Learning Works

More often than not, deciding whether or not to query a specific label comes down to deciding whether the benefits from obtaining the label outweighs the cost associated with collecting that information. Essentially, active learning involves incrementally labeling data during model training in order to enable the algorithm to identify the label that would be most beneficial for it to learn faster. We will cover two key types of active learning querying strategies:

Uncertainty Sampling

Uncertainty sampling, which makes sense only on classification tasks, involves training a model on a fairly small sample of labeled data. Then, the model is applied on the unlabeled remainder of the dataset. The algorithm then chooses which instances to label over the next active learning loop based on how uncertain the classifier is on a given sample.

Data scientists frequently use uncertainty sampling to sample items for human review. For example, let's say you are responsible for a machine learning model to help an autonomous vehicle understand and interpret traffic. You may have millions of unlabeled images taken from cameras on the front of cars, but you only have the time or budget to label 1,000 of them. If you sample randomly, you may get images that are mostly from highway driving, where the autonomous vehicle is already confident and does not need further training.

You can use uncertainty sampling to pinpoint the 1,000 most "uncertain" images where your model is the least confident. It is likely that they will come from non-highway driving, because the initial model was not trained enough on such images. When you update the model with the newly labeled examples, its performance should improve.

It is important to note that there are several variations of uncertainty-like measures, such as uncertainty sampling, margin sampling (which selects the instance with the smallest difference between the first and second most probable labels), and entropy sampling (the instance with the largest entropy value is queried).

What exactly is a classification task?

First of all, classification is the process of identifying to which set of categories (or subpopulations) a new observation belongs. Essentially, it refers to predicting categorical values.

A classification task, then, is the process of predicting the class for a given unlabeled item — and the class must be selected among a set of predefined classes.

For example, a classification task could be dividing a group of customers into various buckets based on their predicted customer lifetime value (CLV), where predefined CLVs existed such as "less than $1,000", "$1,000 to $4,999," "$5,000 to $9,999," or "more than $10,000."

Diversity Sampling

While uncertainty sampling targets data that is obviously confusing for a model in its current state because of low confidence predictions, diversity sampling targets data that allows for a better exploration of the space. For example, let's say someone is attempting to build a voice-activated device with close to 100% coverage of the English language (when the average English speaker only knows about 40,000 words from English's vast 200,000-word vocabulary).

The person has unlabeled recordings he can label, but some people use very rare words. If he randomly sampled the recordings, he would miss out on the rare words. Therefore, he needs to explicitly try to get training data that covers as broad a spectrum of words as possible. He may also want to see what words are frequently used when people talk to their voice-activated devices and sample some of those.

Further, he wants to pay attention to demographic diversity — the recordings he has come from mostly one gender and people in a select number of locations, meaning that the resulting models are likely to be more accurate for the specific accents in those locations and that one gender. He should sample as fairly as possible from various demographics in order for the model to be equally accurate across all demographics.

The Need for Diversity

Uncertainty and representativeness are two criteria of a triad that active learning methods strive to optimize: diversity, representativeness, and uncertainty. In our Data From the Trenches article on diverse mini-batch active learning, discover why diversity is necessary to compensate for the lack of exploration of uncertainty-based methods.

Go Further on Active Learning ↗

Active Learning Examples and Use Cases

Active learning can be beneficial for a diverse range of machine learning projects, from the mundane (like the email filtering example from earlier) to the lifesaving — using the framework to assist with biomedical imaging.

Biomedical Imaging

Historically, in order to accurately annotate a medical image, the labeling had to be done exclusively by trained biomedical experts, which requires extensive bandwidth from both a time and cost perspective. Further, in this instance, the images are subject to human error.

Here, an active learning framework can be applied in order to train a deep neural network with fewer annotated samples. If radiologists only had a scan or an image, without knowing if the asset in hand was good or bad, it would do them no good. Tumor image segmentation is critical to know the size and shape of annotated objects in order to allow for proper analysis and give an accurate diagnosis or cancer stage to the patient.

9 out of 10 researchers who have attempted some work involving active learning claim that their expectations were met either fully or partially.[6]

Fraud Detection

Similarly, active learning can be used to power fraud detection in the banking industry. Fraud detection uses anomaly detection to uncover behavior intended to mislead or misrepresent an actor. Common examples include check and credit card fraud, but it can also occur in other financial sectors like insurance.

Here, human intervention can be used to label new examples and improve classification accuracy, such as with credit card fraud detection, a subset of fraud detection that is particularly specific due to transactional data.

In this instance, stakeholders at these financial organizations can establish a framework that decides what data the model will be trained on and attempt to find the optimal model with the smallest amount of data, with the goal of minimizing costs associated with sending examples to an analyst who labels them.

[6] https://www.kdnuggets.com/2018/10/introduction-active-learning.html

Human investigator analysts contact a small number of cardholders associated with the most high-risk transactions and obtain the class (whether it's fraud or legitimate) of the transactions in question. Once each transaction is labeled by the analysts, they are added back into the labeled data pool and a machine learning model can be trained using this data.

While this feedback received by the analysts is certainly helpful, it is not exclusive. Spontaneous feedback from cardholders identifies fraud that went undetected by the machine learning system. This also allows random transactions to be labeled, helping to keep a balanced learning dataset and ensure that the updated model is able to recognize new types of fraud.

Industrial Equipment Recognition in Images

Organizations can use active learning to accurately label various kinds of industrial equipment and their providers in images. With cars, for example, there is already a multitude of labeled databases with cars everywhere so it's easy to do. With industrial equipment, it's harder to recognize because nothing is pre-existing — organizations have to do it themselves.

Go Further: Named Entity Recognition (NER)

NER, a fundamental element to building natural language processing systems, seeks to locate and classify named entities in text into predefined categories such as the names of people, organizations, quantities, dates, locations, monetary values, percentages, and so on.

For example, if there's a mention of "Paris" in your data, NER would classify that as "location." One goal of NER is to make information easier to locate, by locating named entities and categorizing them under predefined labels. Then, the data can be aggregated within those labels for rapid information retrieval. Job listing data, for example, can contain categories like "organization" or "location" and candidates could then search by those specific categories.

While pre-trained models exist for commonly used entities and languages, they may not exist for specific business domain entities or a specific language. Because NER is a problem that can be solved with supervised machine learning techniques, active learning — which can be used on any kind of supervised machine learning problem — can be leveraged with NER.

Alternatives to Active Learning

While active learning can be extremely useful, it is a topic that is still regularly being researched and tested to garner insights on when exactly it works, on which data, and which techniques perform the best. Given this information, there are several other techniques that can be used as an alternative to active learning, such as transfer learning and semi-supervised learning.

The principle of **transfer learning** makes sense for specific image-related tasks such as computer vision and sometimes natural language processing, but is not applicable for structured data such as in email spam or fraud detection. Active learning and transfer learning both involve efficient and effective use of available data, are both learning accelerators, and are both related to the training of deep learning networks.

While active learning predominantly focuses on expediting the data labeling step to build effective supervised learning models, transfer learning uses existing labeled data from one task to help learning-related tasks for which limited labeled data is available. Transfer learning allows users to retrain a model to specialize it on a particular set of images. It is important to note that the quality of transfer learning for image classification is dependent upon the proximity of the tasks, as well as how close the two tasks (or sets of images) are. A labeled dataset is still needed.

Another technique is **semi-supervised learning**, which uses both labeled and unlabeled samples to perform prediction. It can be used to improve model performance, particularly for smaller samples of labeled data. The technique falls between unsupervised learning (which has no labeled training data) and supervised learning (which only has labeled training data).

Go Further With Dataiku

With Dataiku's deep learning for images plugin, you can receive step-by-step instructions on how to perform transfer learning to retrain a model on a particular set of images.

See the Transfer Learning Tutorial ↗

Conclusion

One of the primary difficulties of active learning revolves around its vulnerability to biases, such as sampling bias. As model training occurs and data points are queried based on increasing confidence, the training set is prone to drift from its underlying data.

While active learning is still being tested and refined today as a result of this, it can be used as a baseline to help determine and prioritize the data that should be labeled and enforce internal guidelines for when resources should and should not be used for labeling.

Sometimes not all of the data is needed — when you train a model, the difference between having 95% and 100% of the data labeled is rather minimal. Additionally, though, because a labeled data is very valuable, it can be a key asset for an organization and, if leveraged properly, can drive a return on investment when used in conjunction with real-world, impact-generating models.

At Dataiku, we pride ourselves on the deep collaboration and granular explainability that our end-to-end platform brings to Enterprise AI. By breaking down silos between teams and encouraging transparency and visibility, data and AI projects can be effectively democratized within an organization. In our latest product release, Dataiku 7, we provide a machine learning-assisted labeling plugin for active learning. With a human-in-the-loop approach, the plugin provides a suite of Dataiku web apps to ease the labeling process whether the data is tabular, images, or even sound.

As a growing number of organizations continue to adopt AI technology to more effectively tackle new business opportunities and uncover pockets to expedite the time to business value, there is an increasing importance in ensuring from the get-go that the data pipeline is prioritized. Higher-quality data generates stronger model performance, so any processes associated with data quality — particularly labeling — should be explainable and scalable, especially if many labelers are involved. The definitions have to be clearly and uniformly applied, as machine learning will not be able to learn concepts that aren't explicit. As model development accelerates, organizations will be able to spend less time on data preparation and further emphasize on long-term business growth.

data iku

Your Path to **Enterprise AI**

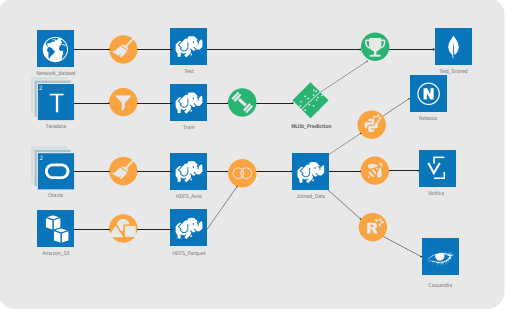
Dataiku is one of the world's leading AI and machine learning platforms, supporting agility in organizations' data efforts via collaborative, elastic, and responsible AI, all at enterprise scale. Hundreds of companies use Dataiku to underpin their essential business operations and ensure they stay relevant in a changing world.

300+
CUSTOMERS

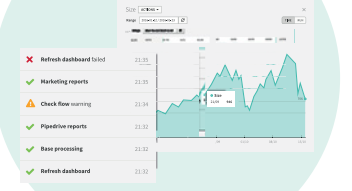
30,000+
ACTIVE USERS

*data scientists, analysts, engineers, & more

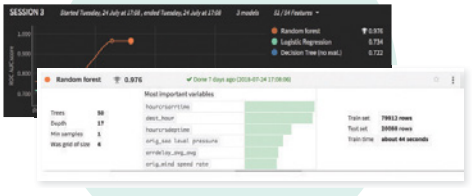
1. Clean & Wrangle



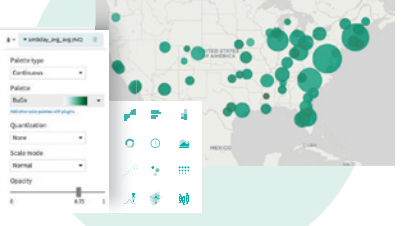
5. Monitor & Adjust



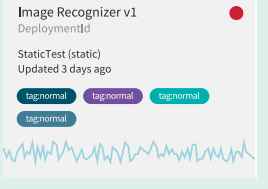
2. Build + Apply Machine Learning



3. Mining & Visualization



4. Deploy to production



WHITE PAPER

www.dataiku.com