AI in the Public Sector

Use Cases, Challenges, & Solutions

www.dataiku.com

Part I:
Challenges & Opportunities

Introduction

The use of data science, machine learning, and ultimately AI within the public sector is on the rise, and the opportunities for even greater adoption appear boundless. The benefits of AI extend to every imaginable realm, from crime prediction to simple bureaucratic process efficiency.

State, local and education government agencies (SLED) have an interest in more efficiently and effectively delivering their mission. In contrast to businesses, public sector agencies typically benefit from the knowledge that they will be around permanently and they don't have to worry about delivering quarterly results to shareholders. This allows them to put a greater emphasis on the long-term benefits of an investment, and that includes investments in AI.

Challenges and Solutions

Though space is rife with opportunity, there are challenges as well. Some of them are the same ones faced by businesses in the private sector, including (but certainly not limited to):

- Barrier to getting started, including the pace at which technology in the AI space is being adopted and dropped, making it difficult to understand in which tools it is best to invest.
- Questions surrounding regulation and privacy, which can make it difficult to understand how teams can work with data within these parameters.
- The level of trust in machine learning systems, which as a **recent AI study by EY** revealed, is not a small hurdle - 33 percent of US CEOs cite employee trust as one of the greatest barriers to AI adoption.

In addition to these more widespread questions, the public sector faces additional challenges specific to their domain, including:

A MORE WIDESPREAD SKEPTICISM OR FEAR OF AI

Fortunately, it seems the tide is slowly turning - **a recent study from NetApp and Arrow** found that more than half of SLED IT managers feel "optimistic" or "excited" about the prospect of adopting or increasing AI in their organizations, and only 12% feel "skeptical." Of course, there is still work to do in trickling these sentiments horizontally to non-IT managers; education and upskilling initiatives around data science and machine learning can help.

BUREAUCRATIC ROADBLOCKS

Many organizations want to innovate, but it can take several years to go through procurement cycles. And given the rapidly changing nature of the AI landscape (especially in terms of the rotating cast of popular technologies used in machine learning projects - **see this post on Technoslvia to get a taste**), this consistently leaves them at the tail end of the adoption curve. Data science, machine learning, and AI platforms - like Dataiku - can help in this regard, as they integrate with various tools and technologies, allowing organizations to provide a consistent and safe experience of building data projects (no matter what the underlying technology).

HANDLING ORGANIZATIONAL CHANGE

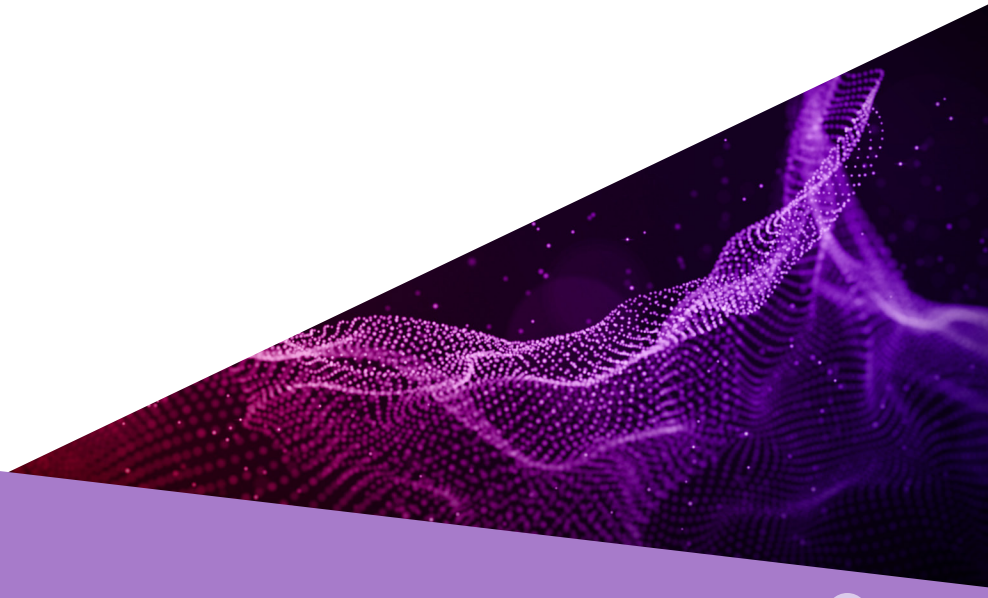
Though this is a challenge to different extents across industries, it is perhaps more pronounced in the public sector, and particularly the federal government. Moving into the age of AI isn't just about hiring a few additional people with data skills; it's about coordinating and undergoing fundamental change throughout the organization to incorporate data into all processes for a variety of use cases. Technology (specifically data science, machine learning, and AI platforms) can help address this challenge, but it's not a magic bullet.

LACK OF DATA TALENT

Hiring data scientists is hard. After being named the **sexiest job of the 21st century**, the demand for the role has skyrocketed - **a LinkedIn report** from the end of 2018 suggested that there were 150,000 fewer data scientists than needed to fill open jobs in that profession, and that number continues to climb. Unfortunately, the challenge can be even bigger for the public sector, where budgets are tight and cutting-edge IT technology and tools - **one of the key ways to retain data scientists** - are lagging. The good news is that it's not just a matter of hiring the best data scientists that money can buy; it's also a matter of leveraging existing staff from across the organization, whether or not they have formal training in data science.

Each of these challenges is complex, and it's outside the scope of this white paper to address each in full. However, for follow-up recommended reading, we recommend **Defining a Successful AI Project: A Framework for Choosing the RIght Use Cases**, **Why Enterprises Need Data Science, Machine Learning, and AI Platforms**, **Insights on Top Challenges from 100+ Data Professionals**, **Executing Data Privacy-Compliant Data Projects**, **Ovum: Profit from AI and Machine Learning**, and **The 6 Key Challenges to Building a Successful Data Team**.

The remainder of this white paper will focus on top use cases in the public sector, including detailing specific stories and examples to bring into focus the potential impact.



Part II:
Top Use Cases

LAW ENFORCEMENT

Crime is an issue that is present everywhere; it's in every city and in every country in the world. Preventing, intervening in, and rectifying harm done by crime costs $10 billion annually in the United States alone.

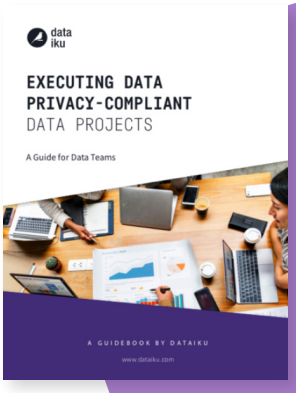
In the U.S., facial recognition technology has been around for many years, but it has not yet become a familiar law enforcement tool. Two of the notable early uses came in Tampa, Florida in 2001 at the Super Bowl, where cameras were deployed in the hopes of identifying terrorists or others who might cause trouble among the crowd of 100,000.

The cameras instead flagged 19 people with outstanding warrants for low-level crimes and **resulted in some false positives**. Later that year, the city of Tampa **put in place cameras** with facial recognition software on the streets of its entertainment district. They only lasted a year and **the police chief at the time later said** that the cameras had not been useful.

Suffice it to say that the technology has progressed dramatically since 2001. Not only has the software become much more adept at analyzing the contours of a person's face, but thanks to social media, the internet is now crawling with images that the software can use to match to the person spotted by the camera on the street.

In 2014, the FBI launched the "Next Generation Identification" system database that aims, among other things, to bolster the ability of law enforcement agencies to identify suspects based on biometric data, such as iris scans and facial recognition. The database draws on photos from a number of sources, including mugshots and employment records. The most obvious use would be to verify the identity of a person who has been arrested and to see whether the suspect is wanted for other offenses.

The FBI has stated the technology will only be used for investigations rather than to produce evidence that will be used in court to prosecute someone. However, privacy advocates are on high alert, concerned that the widespread collection of biometric and biographical data could be used in the future to build a system of surveillance that more closely resembles what exists today in China.



GO FURTHER:

Get the Guidebook to Working with Data Under Privacy Constraints

In the coming years, more and more law enforcement agencies will begin to embrace virtual assistants (or chatbots) in much the same manner that private industry has. For instance, if a citizen writes “I found an injured dog,” a chatbot could advise the person to call 911 only if the animal is being subjected to cruelty but otherwise to contact the animal services department. The result is that citizens are less likely to take up the police department’s time with questions that should be addressed elsewhere.

Advanced analytics may play a big role in helping police departments decide where to allocate precious resources, particularly their officers. Algorithms based on years worth of crime data can pinpoint crime hot spots with far greater precision than a human. Its multidimensional analysis can pick up on trends that would likely elude a team of human analysts, such as certain times of day or certain days of the week when crimes are more likely to occur in a given location. With this analysis, police departments can deploy officers to the areas most likely to generate crime. That increases the chance that offenders will be caught but more importantly, it helps to deter crime in the long-term.

Predictive analytics, however, are not without controversy. Critics worry that algorithms will simply reinforce the overpolicing of areas that have traditionally been more harshly treated by the police, such as those occupied by racial minorities. That doesn’t mean the technology shouldn’t be used; it does, however, give greater importance to using data science, machine learning, and AI platforms that introduce more transparency (white-box AI) as well as help identify potential biases in models.

FEATURED USE CASE

RENT, RAIN, & REGULATIONS: WHAT PREDICTS CRIME IN PORTLAND?

Over the years, there have been many attempts to predict crime, of course. But in general, people are still not very good at knowing how to intervene and where it will happen next. There are more than 200,000 scholarly articles on the underlying causes of crime, and past studies connect crime to poverty, education levels, and many other parameters that we can't actually actively change.

In September 2016, the National Institute of Justice (NIJ) launched the **Real-Time Crime Forecasting Challenge** to predict crime hotspots in the city of Portland, Oregon and kick start innovation around the topic of crime prediction. The team at Dataiku made a submission to the challenge and won in a few categories, predicting Auto Theft with a higher accuracy than others. This feature will give a brief overview of the project and how it was executed.

This is a simplified explanation of this project and use case; for a technical deep-dive on the process, **read the post on Data from the Trenches**.

HOW DATAIKU DID IT

For this project, the team at Dataiku started by observing that one of the key characteristics of crime data is that it has both a geographic dimension and a temporal dimension (crimes happen in different places at different times). It can also be affected by many different types of features — weather, city infrastructure, population demographics, public events, government policy etc.

So they started with the data released by the NIJ and enriched this data with a variety of public data sets including police reports, U.S. census data, data from the Foursquare application, data from newspapers, and the weather.

THE CHALLENGE: WHY PREDICT CRIME

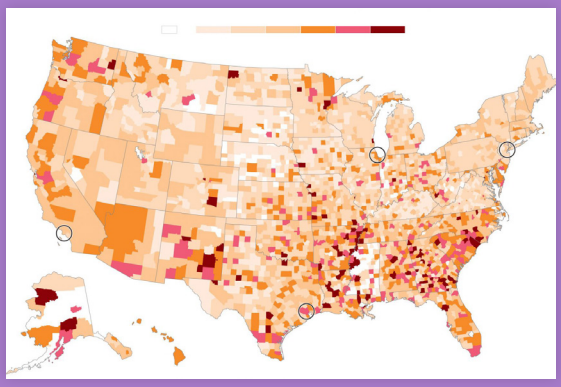
Recently, preventative policing has become a topic of interest -- can we predict, and therefore prevent, crime before it occurs? To be clear, the goal of projects like the NIJ challenge is not to strike crime before it happens à-la Minority Report, or to increase community stigma. The goal is to provide actionable techniques for policy makers and police officers to anticipate crime.

Indeed, past research on preventive policing suggests that randomly increasing police presence isn't as effective in reducing crime rate as targeting crime hotspots, or areas with a high likelihood of future crime occurring. Moreover, evidence suggests that these hotspot interventions are not merely displacing the crime to nearby areas, but actually decreasing the overall crime and incarceration rates in the larger area.

Also, understanding what features drive crime in an area is an interesting tool to support police efforts with certain policies to change those factors.

FINDING HOTSPOTS: CRIME AS A TECHNICAL CHALLENGE

So, how does one actually find these crime hotspots?



SOURCE: **The Washington Post**

Looking at a map of the United States, one can see that there is clearly a spatial structure to criminal events, but the temporal dimension is harder to visualize. Obviously, not every point in time and space is equally likely to host a crime. But it can be very hard to leverage both of these structures to accurately predict future crime, and it can be especially hard to understand the underlying factors that are driving these crimes.

There have already been many different approaches employed to attempt to predict crime, including interviews with convicted criminals and analyses of large-scale social media data. Work in criminology and sociology has laid much of the foundation for these types of analyses, from figuring out how to define a hotspot to measuring the effects of police intervention in these areas to coming up with targeted hypotheses of features that drive crime.

However, traditional crime prediction approaches like these face challenges like:

- Relatively small data sets
- Looking at temporal or spatial data, but not both.
- A small number of features used to explain crime

These challenges are exactly what AI-based approaches are designed to overcome. And, in fact, there are some really cool projects that have come out of data science and machine learning approaches to crime.

However, these projects also have their own limitations. Because they are relatively new approaches, they lack targeted analysis techniques (even though the data that goes into thinking about crime is complex and requires more specific techniques). Also, these projects often treat the spatial or the temporal aspect, but not both simultaneously.

So with all of that in mind, the goal of the Dataiku team's project was to:

- Use only open source data.
- Engineer spatial and temporal features, and take advantage of the dependencies in these types of data.
- Use targeted statistical techniques to handle these specific types of data.
- Model and predict crime.
- Understand those predictions and generate features for further testing and exploration.

WHAT TO PREDICT?

The data released by the NIJ was the basis for the project, as that is the data that needs to be predicted. It's a dataset of all the 911 calls that have led to police interventions for street crimes, burglary, and motor vehicle theft between March 2012 and March 2017. This is what the team used as a target to predict a crime.

This data comes in the form of a geopoint, a time and date, and the features of the call. When trying to figure out where crime might occur, it's important to consider both geographic features (what makes one neighborhood different from another?) and temporal features (what makes one point in time different from another?).

WHAT DATA CAN BE USED TO PREDICT CRIME?

OpenStreetMap: The first data the team brought in was points of interest from OpenStreetMap. Along with roads and geographic features, OSM marks points of interest (POIs). Features that occupy a particular point that are tagged with user-generated information, including category tags such as Amenity, Hotel, Shop, Restaurant, Mailbox, etc. The team extracted these geopoints and tags.

Foursquare: More than just points of interest, the Dataiku team wanted to look into how people actually use different neighborhoods in Portland. To get a sense of how people spend their time in each neighborhood, they pulled data from the Foursquare API and extracted checkins in restaurants and other businesses.

Police Precincts: Next, they grabbed the map of police precincts in Portland, which came in the shape of a tiling of the city showing what bit of land belongs to what precinct.

U.S. Census Data: The U.S. Census collects over 20,000 features on each neighborhood in the United States (between 600-3,000 people per precinct) including the number of people, gender, average income, number of cars per household, etc. To choose from among these 20,000+ features, the team only extracted the features that correlated with our target (911 calls) over the years that we had (2014, 2015, 2016). This data is both geographical (its tiles) and historical since one can observe the variations over the year.

Weather Data: To explore the temporal aspect of crime, the team extracted weather from the NOAA weather API. This gave them pretty much everything we wanted to know about the weather (temperature, precipitation, wind, etc).

Holidays and Events: The goal here was to clean up the noise in the data linked to the day of the week or events (do people commit less crimes around Christmas? More crimes?).

READY FOR MORE?

If you want more details on exactly what the team did with this data to win the contest, check out the more technical article. Dataiku also led a project mapping and predicting real rates of crime in greater London - see the details here. Of course, for each of these projects, the teams used Dataiku DSS for everything from data ingestion to cleaning to modeling.

PREDICTING EVICTIONS IN KANSAS CITY

The Kansas City Eviction Project is a collaborative effort involving researchers, community organizers, neighborhood leaders, lawyers, and policymakers. They aim to advance a housing justice agenda in Kansas City and to build sustainable political power with people affected by housing injustice.

Back in 2017, VP of Field Engineering Jed Dougherty worked with the organization to blend a Jackson County data set that included 173,720 eviction records spanning 17 years with U.S. Census, education and neighborhood data — all based on the geo-coded location of each eviction — to create a map on which to overlay all the information.

And with the assistance of experts from Code for Kansas City and the resources of Eric Roche and the City of Kansas City's Open Data KC site, the team was able to build a model that predicts whether a residence will have an eviction in the next year.

Watch the video to get the full story of the project from Jed.

OPERATIONAL EFFICIENCY

Virtual assistants are one very promising tool that could relieve many government employees from the drudgery of answering the same mundane questions over and over again. Case and point: Eric Ellis, the former chief technology officer for the state of North Carolina, told Government Technology magazine in 2017 that 80 percent of the calls to the Help Desk he set up were from people who wanted to change their passwords. That is hardly a worthy use of government manpower when that question could be easily addressed by a chatbot.

Just as patients are increasingly interacting with robots to refill their prescriptions, those interacting with government agencies will increasingly get prompt service from automated assistants who can help them obtain their benefits, enroll in training programs, register to vote, check out a book from the library or perform any other number of important government functions.

The greater the capacity for data analysis, the better government agencies will be able to crack down on waste and fraud. Currently, the best case scenario for public sector entities is that they are able to hire dedicated auditors to review their budgets and other important information, such as payroll data, to identify wrongdoing or potential savings. However, due to resource constraints, comprehensive audits are rare and even thorough investigations leave some stones unturned.

One notable example comes in the European Union, which recently built a system, DIGIWHIST, that automatically incorporates all public procurement announcements across the 28-member federation. The system includes over 20 million procurement contracts. The system automatically scores each contract based on a number of different metrics that correlate with corruption, such as the number of bidders, the length of the solicitation and the amount of information published about the bid. The system will help the EU identify contracts that may have been obtained via improper means and bring transparency to the procurement process.

TRANSPORTATION, PLANNING, AND MANAGEMENT

Perhaps the most substantial impact of AI in the public sector will come in the realm of transportation planning and management. Transportation systems power the world economy, and they are almost always built and maintained by public sector entities, from municipal agencies to national governments. While each year brings exciting innovations in transportation, the two main issues are far from being

resolved: soul-crushing congestion in major metropolitan areas and the more than 1 million drivers, transit users, bicyclists, and pedestrians a year who die on the world's roads.

While the private sector will forge ahead with improvements via autonomous vehicles and the like, the public sector can leverage AI to make transportation safer and more efficient.

ADDITIONAL RESOURCES

WATCH the talk from Luca Maria Aiello, Senior Research Scientist @ Nokia Bell Labs

The Spirit of the City: Capturing Network-Generated Data for Better Cities

Consider one of the most vexing decisions cities face when building infrastructure: how long should the light stay red at an intersection? Traditionally, the decision is based on information collected by a contractor hired to stand next to the intersection and count the number of cars going by. Now, city leaders are deploying intelligent traffic signals that are not only more accurate counters, but are taking into account a complex matrix of factors to determine signalization. The result are intersections that move vehicles more quickly and more safely. That's significant not just for busy intersections, but for slower ones, where drivers regularly sit and wait for the light to turn green even when there's not another vehicle nearby. In such instances, the signal can sense that only one car is present and allow it to pass.

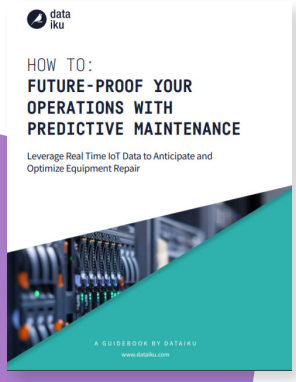
Autonomous vehicles (which are already being used by militaries worldwide) offer a significant opportunity to make public transit cheaper and more reliable. A number of manufacturers are already experimenting with autonomous buses. That technology will provide a smoother, safer ride and be able to seamlessly platoon multiple vehicles simultaneously in response to increased demand. Because of the platooning capability, cities will be able to implement bus systems with the same carrying capacity as rail, which is traditionally the best option for high-capacity transit. This new system, however, will be significantly cheaper simply because it doesn't require the substantial investment in tracks, tunneling and the other major construction associated with rail.

In the realm of transportation there is also the topic of AI-based predictive maintenance, which can apply both to public transit systems as well as fleets (cars, vans, etc.) owned by SLED agencies. Predictive maintenance algorithms offer the prospect of both reduced maintenance costs as well as extended equipment lifetime.

Just like car owners, the public sector currently typically maintains equipment on set schedules based on age and mileage. That approach makes sense –- until there is something better, such as AI applications that can monitor the condition of a bus, train, truck, or any other expensive piece of machinery in real time and is able to predict when maintenance issues will arise so that technicians can get ahead of the issue.

GO FURTHER:

Get the Step-by-Step Guide to Introducing AI into Equipment Maintenance



DEFENSE

It is hardly surprising that the largest and most advanced military in the world is investing heavily in artificial intelligence. The 2020 budget for the U.S. Department of Defense allocated nearly $1 billion for AI, including $208 million for DoD's Joint Artificial Intelligence Center, which oversees any AI project exceeding $15 million in all branches of the military.

While only tangentially related to the use of AI in the public sector, the two have a lot in common. The use of AI in the military ranges from mundane cost-savings measures to the highest-stakes issues imaginable to mankind, such as averting devastating nuclear attacks.

The world's top militaries currently deploy an enormous amount of human labor to identify and react to threats. And yet, even with the best expertise at their disposal, the fact is that they may not be equipped to react fast enough in the event of an attack. Ultimately, they still have a lot of work to do when it comes to technology; for example, many organizations still rely heavily on spreadsheets, which are notoriously error-prone (many studies in the last decade suggest that up to 80 percent of spreadsheets contain errors).

The more intelligent systems become, the less dependent they will be on the time-consuming process of humans analyzing the data and communicating it to each other. At the same time, that doesn't mean taking people entirely out of the loop; responsible AI is all about enhancing - not entirely replacing - uniquely human abilities.

"A defense agency could have just eight to 10 minutes to decide whether a launch represents a threat, share findings with allies, and decide what to do. The use of countermeasures has to happen quickly, given that missiles could impact 16 minutes after launch."

SOURCE: KPMG

Part III: Conclusion

HUMAN-CENTERED AI

The more people are involved in AI processes, the better the outcome. The goal in introducing data science, machine learning, and - eventually - AI into the public sector is truly to enhance human abilities, not to replace them. In order for AI efforts to be a success, this has to be a message that is clear from the top down and put into practice from the bottom up by those working in public sector organizations every day.

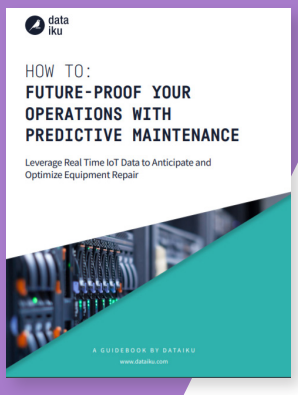
Ultimately, crafting an AI strategy that is inclusive (i.e., involving everyone in an organization, not siloed to data teams or data scientists) and human-centered (i.e., values the input of humans over blindly pushing decisions to machines) is the key to unlocking scalable AI.

NEXT STEPS

There are a few best practices the public sector can adopt to improve their approach to AI and ensure its longevity:

1. **Take steps toward machine learning**. Deterministic decision models based on static rules (e.g., If X, then Y) are comforting in their simplicity, but lag far behind more sophisticated models, which can weigh a host of variables and concerns in order to make recommendations. Probabilistic modeling and partial dependency plots provide weighted recommendations and demonstrate the extent to which specific variables contributed to that recommendation, providing transparency into elusive mathematical models.

GO FURTHER



2. **Embrace Open Source**. While privacy is paramount when it comes to data and systems in the public sector, open source technologies offer exceptional AI technology for free. The collaborative development community and widespread buy-in leads to software that performs better and is more reliable than exclusively private alternatives. The most innovative tools when it comes to building and applying machine learning are open source, but when it comes to operationalizing projects, organizations do need another proprietary layer to best integrate with their use cases.

3. **Invest in Tools that Empower**. The right data tools demystify AI and enable everyone to understand how AI systems arrive at their recommendations and why. Additionally, a central environment with auditing capacity offers the collaborative experience that data analytics and decision-making flourish in, while restricting user access to sensitive data. This enables engagement without compliance risk.

GO FURTHER

data iku

Your Path to **Enterprise AI**

Dataiku is the platform democratizing access to data and enabling enterprises to build their own path to AI. To make this vision of Enterprise AI a reality, Dataiku is the only platform on the market that provides one simple UI for data wrangling, mining, visualization, machine learning, and deployment based on a collaborative and team-based user interface accessible to anyone on a data team, from data scientist to beginner analyst.

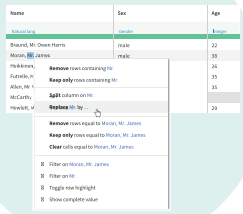
300+
CUSTOMERS

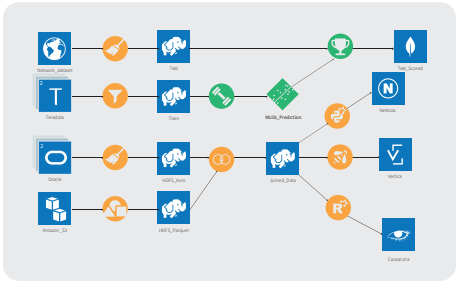
30,000+
ACTIVE USERS

*data scientists, analysts, engineers, & more

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

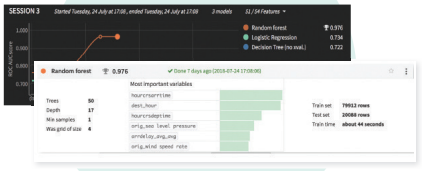
1. Clean & Wrangle



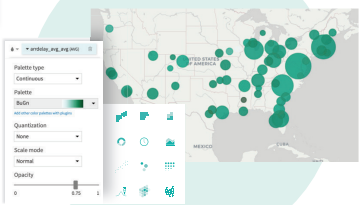


5. Monitor & Adjust

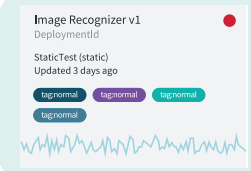
2. Build + Apply Machine Learning



3. Mining & Visualization



4. Deploy to production



WHITE PAPER

www.dataiku.com