The Forrester Wave™: Notebook-Based Predictive Analytics And Machine Learning Solutions, Q3 2018

The Nine Providers That Matter Most And How They Stack Up

by Kjell Carlsson, Ph.D. and Mike Gualtieri

September 5, 2018 | Updated: September 7, 2018

Why Read This Report

In our 24-criteria evaluation of notebook-based predictive analytics and machine learning (PAML) solutions providers, we identified the nine most significant ones — Anaconda, Civis Analytics, Cloudera, Databricks, Domino Data Lab, Google, H2O.ai, OpenText, and Oracle — and researched, analyzed, and scored them. This report shows how each provider measures up and helps application development and delivery (AD&D) professionals make the right choice.

Key Takeaways

Oracle And Domino Data Lab Lead The Pack

Forrester's research uncovered a market in which Oracle and Domino Data Lab are Leaders; H2O.ai, Databricks, Civis Analytics, OpenText, and Cloudera are Strong Performers; and Anaconda and Google are Contenders. All of these vendors add enterprise features to support teams of data scientists rather than just a lone data scientist.

Notebook-Based Solutions Offer Open Source Without The Hassles

To leverage the latest innovation in open source machine learning frameworks and tap the vast ranks of data scientists who want to develop models in Python and R, enterprises need notebook-based PAML solutions to manage these tools, drive data scientist productivity, and accelerate model deployment.

Use The Right Notebook-Based PAML Solution For The Job

Each evaluated vendor shines for different data science use cases, whether it be supporting a wide range of PAML tools, automated machine learning, integrated machine learning on Spark and Hadoop, customer analytics, analysis of enterprise information application data, or expertise with the underlying machine learning algorithms.

The Forrester Wave™: Notebook-Based Predictive Analytics And Machine Learning Solutions, Q3 2018

The Nine Providers That Matter Most And How They Stack Up

by Kjell Carlsson, Ph.D. and Mike Gualtieri
with Srividya Sridharan, Aldila Yunus, and Robert Perdoni

September 5, 2018 | Updated: September 7, 2018

Table Of Contents

Related Research Documents

The Forrester Wave™: Multimodal Predictive Analytics And Machine Learning Solutions, Q3 2018

Now Tech: Predictive Analytics And Machine Learning Solutions In China, Q3 2018

Now Tech: Predictive Analytics And Machine Learning Solutions, Q2 2018

FORRESTER®

Forrester Research, Inc., 60 Acorn Park Drive, Cambridge, MA 02140 USA
+1 617-613-6000 | Fax: +1 617-613-5000 | forrester.com

For Machine Learning At Enterprise Scale, You Need PAML Solutions

Machine learning is an elemental core competency that every enterprise must have. The reasons are many. Machine learning gives enterprises the power to predict. It is a fundamental building block to AI. It can learn from data and find hidden insights. Most importantly, it can make even the hugest of enterprises gain the agility of disruptive upstarts by injecting scalable intelligence into hyperpersonalized customer experiences, business processes, operational applications, and employee decisions. None of this happens, though, without businesspeople, data scientists, data engineers, software developers, and AI engineers working together. That's where enterprise predictive analytics and machine learning solutions for data science teams and their friends are necessary. Forrester defines enterprise PAML as:

> *Software that provides enterprise data scientist teams and stakeholders with 1) tools to analyze data; 2) workbench tools to build predictive models using statistical and machine learning algorithms; 3) a platform to train, deploy, and manage analytical results and models; and 4) collaboration tools for extended enterprise teams including businesspeople, data engineers, application developers, DevOps, and AI engineers.*[1]

"Notebook-Based" Is One Of Three PAML Segments Identified By Forrester

The focus of this Forrester Wave™ is on evaluating notebook-based PAML solutions. However, this is just one of three market segments identified in the "Now Tech: Predictive Analytics And Machine Learning Solutions, Q2 2018" Forrester report, and most enterprises will find it beneficial to use solutions from all three market segments. We define these segments as follows:

- **Notebook-based PAML solutions favor a code-first approach.** These solutions provide workbench tools centered on coding in R, Python, and other programming languages using open source Jupyter or a proprietary interface that makes coding more efficient. The vendors in this segment add significant, differentiated features, such as environment provisioning, project management, deployment, model management, visualization tools, and more. We evaluate this PAML market segment in this report.
- **Multimodal PAML solutions provide the widest breadth of workbench tools.** These solutions offer multiple user-interface paradigms and the broadest set of workbench tools, such as graphical user interfaces (GUIs), configuration wizards, automation, and coding environments. Many of these solutions also provide tools for non-data scientists to build data pipelines, create machine learning models, and collaborate with data science teams. We've evaluated multimodal PAML solutions in the "The Forrester Wave™: Multimodal Predictive Analytics And Machine Learning Solutions, Q3 2018" Forrester report.
- **Automation-focused PAML solutions help non-data scientists build models.** This segment focuses on tools to automate the steps in the model-building life cycle. These solutions enable data scientists and non-data scientists to build models by configuration, rather than coding and

specifying each step in a data science pipeline. Some multimodal and notebook-based vendors offer automation as well, but because they offer other approaches to building models, they are not exclusively included in this automation-focused segment. The vendors in this category focus specifically on an automation approach to machine learning. Automation-focused PAML solutions are not evaluated here. Forrester plans a separate Forrester Wave evaluation focused on this subsegment in 2019.

Notebook-Based PAML Solutions Are Purpose-Built For Open Source

Open source machine learning is evolving at a dizzying pace. The number of open source machine learning frameworks — TensorFlow, Keras, Caffe, Theano, and PyTorch, to name just a few — has exploded, and they are evolving daily. New generations of data scientists grow up programming in open source languages like Python and R, and every data scientist is looking to build their skills with the latest and greatest deep learning algorithms. For enterprises, the opportunity is immense, and so are the challenges when every new library or updated machine learning framework can break applications or lead to hours of rework. Driving outcomes with data science requires collaboration between teams, getting models into production quickly, and managing them at scale. How do you do this with teams that are looking to use an ever-increasing, constantly changing set of open source tools? Enter a new class of PAML solutions. Notebook-based PAML solutions:

- **Enable teams of data scientists to use notebooks without the hassles.** Notebooks, such as Jupyter and Zeppelin, are coding environments that incorporate results, visualizations, and documentation in line with code. They've emerged as the standard workbench for data scientists to develop, record, and share their work when using open source programming languages. However, they're cumbersome to standardize, instantiate, and provision resources for. To make them effective for enterprise data science teams, notebook-based PAML vendors combine notebook support with containerization frameworks, such as Kubernetes, making it easier to manage the development, training, and deployment of PAML models.
- **Provide differentiated features that make data science teams more productive.** The new notebook-based PAML solutions are pushing the envelope on features for collaboration, experiment tracking and model reproducibility, as well as automated machine learning. Some even make it easy to manage multiple PAML tools, including proprietary multimodal solutions like SAS, from within the same platform.

Notebook-Based PAML Solutions Are Your (Data Scientists') New Best Friends

Notebook-based PAML solutions are like a collection of brilliant teenagers. They are developing, maturing, and learning from each other at breakneck speed. Also, like teenagers, there always seems to be more and more of them, with vendors ranging from cloud giants to data science consulting firms vying to launch their own platforms. Thanks in large part to their open source foundations, they've replicated in months much of the functionality that it took traditional PAML vendors years to develop.

And they've leapfrogged the older generation on many features by embracing best practices from software development. If your data science teams want to code using open source languages, and you want them to be productive, consider a notebook-based PAML solution. While they are:

- **Still playing catch-up . . .** Large gaps remain relative to the wide range of features offered by established multimodal PAML solutions. Most notably, these solutions typically have few features for making data discovery and data preparation easy, have only barebones features for managing models in deployment, and lack a catalog of battle-tested modules tailored for common business problems.
- **. . . they are driving immediate value.** The value of notebook-based PAML solutions — boosting your data scientists' productivity by enabling them to use the tools that they want to use, but in a more efficient and easily managed fashion — far outweighs the benefits of waiting. And, because of these solutions' open source foundations, you're less likely to be locked in to any particular vendor.

Notebook-Based PAML Solutions Evaluation Overview

To assess the state of the notebook-based PAML market and see how the vendors stack up against each other, Forrester evaluated the strengths and weaknesses of top vendors. After examining past research, user need assessments, and vendor and expert interviews, we developed a comprehensive set of 24 evaluation criteria, which we grouped into three high-level buckets:

- **Current offering.** Each vendor's position on the vertical axis of the Forrester Wave graphic indicates the strength of its current offering. Key criteria for these solutions include their capabilities for managing notebooks, workbench features for driving data science productivity, and the scalability of the architecture.
- **Strategy.** Placement on the horizontal axis indicates the strength of the vendors' strategies. Given the rapidly evolving nature of this segment, we considered the vendor's solution road map to be most important, followed by the ability of its partnerships to propel its growth.
- **Market presence.** Represented by the size of the markers on the graphic, our market presence scores reflect each vendor's number of enterprise customers, the degree to which enterprise buyers are aware of the vendor, and the vendor's product revenues.

Evaluated Vendors And Inclusion Criteria

Forrester included nine vendors in the assessment: Anaconda, Civis Analytics, Cloudera, Databricks, Domino Data Lab, Google, H2O.ai, OpenText, and Oracle. Each of these vendors (see Figure 1):

- **Is identified as a notebook-based PAML solution.** Forrester has determined that all evaluated vendors offer a notebook-based PAML solution, as defined in the report "Now Tech: Predictive Analytics And Machine Learning Solutions, Q2 2018." Vendors solutions that Forrester has identified as multimodal PAML or automation-focused PAML solutions are not included in this evaluation.

- **Offers a comprehensive, differentiated, notebook-based PAML solution.** Included vendors must offer a solution that can operate on large data sets; provide capabilities for data acquisition and preparation; and provide statistical and machine learning algorithms, a differentiated user interface to build models, and model management features. If a vendor offers a PAML based in whole or in part on open source, the vendor must have value-added differentiation. For example, we would not include a business intelligence (BI) or database vendor that offers the ability to run R scripts and/or Python code. A vendor that offers a cloud service for an undifferentiated Jupyter notebook would also not be included.
- **Markets a standalone notebook-based PAML solution to enterprise data scientists.** Forrester only included solutions that are marketed toward enterprise data science, AI engineers, and or application development teams that use machine learning algorithms to analyze data and create predictive models. Forrester did not include PAML solutions that we deem are technologically embedded in any other particular application, such as a BI, data preparation, or ETL application, or a middleware stack.[2]
- **Meets our install base and revenue requirements.** Included vendors have at least four paying, named enterprise customers using the PAML solution. Vendors also provided Forrester with three customer references who agreed to fill out a confidential survey and/or participate in a telephone interview. Included vendors also had a trailing 12-month revenue of at least $4 million.
- **Has sparked client inquiries and/or has market momentum.** Forrester clients often discuss the vendors and products through inquiries; alternatively, the vendor may, in Forrester's judgment, warrant inclusion or exclusion in this evaluation because of technology trends, market presence, or lack of client interest.

FIGURE 1 Evaluated Vendors And Product Information

Vendor	Product evaluated	Version evaluated
Anaconda	Anaconda Enterprise	5
Civis Analytics	Civis Platform	N/A
Cloudera	Cloudera Data Science Workbench (CDSW)	1.4
Databricks	Unified Analytics Platform	N/A
Domino Data Lab	Domino Data Science Platform	2.6
Google	Cloud Datalab	N/A
H2O.ai	H2O-3	3.0.18
OpenText	OpenText Magellan	16.4
Oracle	Oracle Data Science Platform	6.2

Vendor Profiles

We intend this evaluation of the notebook-based predictive analytics and machine learning solutions market to be a starting point only and encourage clients to view detailed product evaluations and adapt criteria weightings to fit their individual needs through the Forrester Wave Excel-based vendor comparison tool (see Figure 2 and see Figure 3). Click the link at the beginning of this report on Forrester.com to download the tool.

FIGURE 2 Forrester Wave™: Notebook-Based Predictive Analytics And Machine Learning Solutions, Q3 2018

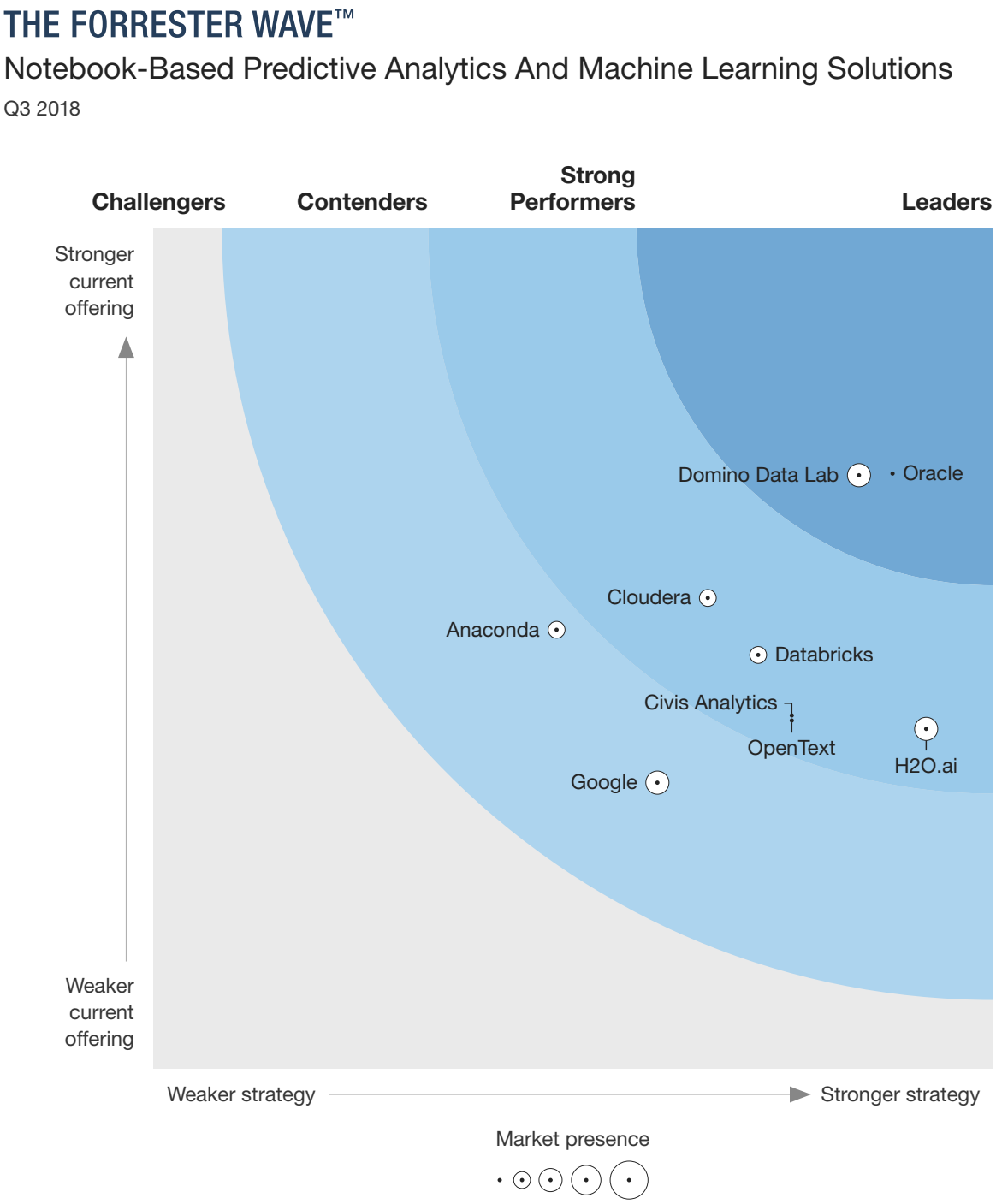

FIGURE 3 Forrester Wave™: Notebook-Based PAML Solutions Scorecard, Q3 2018

	Forrester's weighting	Anaconda	Civis Analytics	Cloudera	Databricks	Domino Data Lab	Google	H2O.ai	OpenText	Oracle
Current offering	50%	2.61	2.10	2.80	2.46	3.53	1.70	2.02	2.07	3.54
Platform	30%	2.00	1.00	3.00	2.00	5.00	0.50	1.00	2.00	5.00
Workbench	30%	2.20	3.00	3.00	2.20	3.10	1.15	1.90	2.90	2.80
Model operations (ModelOps)	10%	1.50	2.00	2.00	1.00	2.00	1.00	1.50	1.00	2.00
Open source algorithms	10%	5.00	3.00	1.00	5.00	3.00	5.00	5.00	1.00	3.00
Architecture	10%	4.00	1.00	4.00	3.00	4.00	3.00	3.00	3.00	4.00
Business solutions	10%	3.00	3.00	3.00	3.00	2.00	3.00	2.00	1.00	3.00
Strategy	50%	2.40	3.80	3.30	3.60	4.20	3.00	4.60	3.80	4.40
Ability to execute	10%	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
Solution road map	50%	1.00	5.00	3.00	3.00	5.00	3.00	5.00	5.00	5.00
Implementation support	10%	1.00	1.00	5.00	3.00	3.00	1.00	3.00	1.00	3.00
Pricing and acquisition	10%	5.00	3.00	0.00	5.00	1.00	5.00	5.00	3.00	3.00
Partners	20%	5.00	3.00	5.00	5.00	5.00	3.00	5.00	3.00	5.00
Market presence	0%	1.66	0.67	1.66	1.68	2.32	2.34	2.32	0.66	0.66
Customer adoption	34%	1.00	1.00	1.00	3.00	1.00	3.00	1.00	0.00	0.00
Evaluated product revenue	33%	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Market awareness	33%	3.00	0.00	3.00	1.00	5.00	3.00	5.00	1.00	1.00

All scores are based on a scale of 0 (weak) to 5 (strong).

Leaders

› **Oracle seeks to dominate machine learning in the cloud.** With its recent acquisition of DataScience.com, Oracle has added a critical missing component to its portfolio of machine learning solutions in the cloud. With the Oracle Data Science Platform, teams can leverage the three most common open source notebooks (Jupyter, Zeppelin, R Studio) and seamlessly provision and instantiate their own customized machine learning environments. At the same time, it provides the standardization and controls that enterprises need. The platform also makes it easy to put models into production by offering visual tools to create APIs with automatic load balancing. By

combining one of the most promising startups in the notebook-based PAML space with its own related offerings for infrastructure, data management, third-party data sources, BI, and a suite of business applications, Oracle is instantly a force to be reckoned with in the PAML market.

› **Domino Data Lab lets you conduct experiments with your tool of choice.** Domino Data Lab is all about making your data science teams more productive by working with them rather than forcing them to use just one notebook interface. Data scientists love their different tools, so Domino lets users code in Jupyter, Zeppelin, and R Studio but also in proprietary tools like SAS and DataRobot. It enables this through a unified platform that enables self-service provisioning of resources and allows teams to both standardize their environments and modify them when experimentation is called for. The platform also fosters collaboration by making it easy to find and share assets across projects, and it offers integrated features for commenting, providing feedback, and notifying users of recent activity. Since data scientists work through iterative experiments, Domino provides a visual workbench to track, monitor, and share the details of experiments that also ensures reproducibility. All of this is enabled through slick visual tools that users will want to use.

Strong Performers

› **H2O.ai's future is automated machine learning.** H2O.ai is best known for developing open source, distributed machine learning algorithms at a time (2011) when big data demanded them but no one else had them. H2O.ai's algorithms still shine, and many other machine learning vendors integrate them into their solutions. However, in this evaluation, we evaluated H2O.ai's Sparkling Water and Flow UI products. These are capable but basic tools that data science teams can use to take advantage of Apache Spark and, of course, H2O.ai algorithms. Enterprises looking for a notebook-based PAML solution will find better solutions from the other vendors in this evaluation that integrate H2O.ai's great algorithms. However, H2O.ai's bright future is in its Driverless AI product, which automates large swaths of the data science life cycle to automatically build hundreds or even thousands of models. Forrester plans to evaluate automation-focused PAML solutions separately in 2019.

› **Cloudera marries big data and machine learning.** Cloudera is a well-known big data platform vendor that offers a Hadoop/Spark-based Enterprise Data Hub to manage enterprise data of all kinds and from all sources. With Cloudera Data Science Workbench, Cloudera provides data science teams with a solution that is deeply integrated with Cloudera Enterprise, including data governance, data security, queries, streaming, and batch processing. What enterprise data science team doesn't need that? For enterprises that already have invested in Cloudera Enterprise, Cloudera Data Science Workbench is a compelling offer. However, some data science teams seeking notebook-based solutions may wish that the platform was able to instantiate open source notebooks such as Jupyter. Cloudera plans to decouple Data Science Workbench from Enterprise Data Hub to make the platform appeal to enterprises that have invested in another big data platform.

- **Databricks — nobody autoscales machine learning in Spark better.** Founded by many of the core contributors to the Apache Spark framework, Databricks provides a cloud-based machine learning platform that streamlines much of the work of managing data and machine learning pipelines on Spark. The platform offers cloud-hosted Spark clusters on demand, and its killer feature is the ability to autoscale to accommodate any size workload. Databricks Delta provides a layer of data preparation to help data engineers and data scientists work with their data faster, while MLflow provides new features for managing and ensuring the reproducibility of machine learning experiments. However, putting models into deployment requires a separate solution. Also, while Databricks' proprietary notebook offers unique features — such as the Google-Docs-like ability to collaborate on the same code simultaneously — some data scientists may balk at not being able to use their favorite development environments.
- **Civis Analytics helps you really get to know people.** Civis is the platform to beat for data scientists who want a notebook-based platform to work with person-level data. And it should be; it's the platform Civis Analytics built for its own data science consulting business. It has a host of unique features to handle notoriously messy data sets on individuals — such as a proprietary tool for matching people across data sets, address correction, deduplication, and imputation. It also provides an all-in-one solution to create and share the results of your analysis, as it white-labels Tableau and supports interactive web apps developed in R Shiny. It's secure, too, being one of the few HIPAA-compliant vendors.[3] But the pièce-de-résistance is its integration with Civis' proprietary data sets on US customers, enabling another level of insight into customers. Civis is not yet the platform for everyone, though, as it is currently cloud only, it doesn't support many machine learning frameworks out of the box, and Spark is still in the pipeline.
- **OpenText's data science tackles the gnarliest, most underleveraged valuable data.** OpenText surprisingly, yet delightfully, burst onto the machine learning scene in the past year with Magellan — a PAML solution that leverages the open source Jupyter notebook and an impressive number of its existing enterprise information products. Jupyter is the workbench interface many data science teams expect, and OpenText adds value by also giving them access to unstructured data with sophisticated text analytics. The real benefit of using OpenText Magellan is the ability to operationalize models in business-friendly interfaces in its vast portfolio of enterprise information applications and services, including customer experience management, business network, digital process automation, and content management. OpenText Magellan is compelling for existing OpenText customers, but the vendor must continue to invest in the platform, including more sophisticated model management and automation features, to draw wider appeal.

Contenders

- **Anaconda goes enterprise.** The creators of the most famous Python distribution for machine learning now offer an end-to-end platform for data scientists. In addition to enterprise support for the Anaconda distribution, the Anaconda Enterprise platform integrates support for Jupyter notebooks and manages containers for machine learning. The platform also provides unique

features for managing the wide range of frameworks, libraries, and packages used by different teams and ensuring the ability to share and reproduce models. As you would expect from a pillar of the open source community, the platform supports the largest set of open source machine learning frameworks out of the box, and customers praise Anaconda's in-depth knowledge of the underlying open source code. The platform also has a wide range of options for getting model outputs into the hands of business users through interactive web applications (like R Shiny and Python Flask).

› **Google's machine learning strategy today is tech, not tools.** Google has developed a host of cloud services for machine learning — such as BigQuery, Cloud ML Engine, Compute Engine, Cloud TPU, Cloud Vision API, and many others — and world-class machine learning technologies, such as TensorFlow.[4] Particularly exciting is Cloud AutoML, which automatically creates bespoke deep learning models.[5] However, Google's notebook-based PAML solution evaluated in this report, Cloud Datalab, has little to differentiate it from a standard Jupyter notebook, providing a basic interface to Google's cloud technology. Beyond that, it does little to improve data scientist productivity, such as through project capabilities, team collaboration features, and other modeling tools that are important criteria in this evaluation. Consequently, enterprise data science teams that wish to access all of Google's machine learning technology goodness are best advised to access it through other vendor tools until and if Google invests more heavily in a notebook-based PAML solution.

Supplemental Material

Online Resource

The online version of Figure 2 is an Excel-based vendor comparison tool that provides detailed product evaluations and customizable rankings. Click the link at the beginning of this report on Forrester.com to download the tool.

Data Sources Used In This Forrester Wave

Forrester used a combination of three data sources to assess the strengths and weaknesses of each solution. We evaluated the vendors participating in this Forrester Wave, in part, using materials that they provided to us by June 11, 2018.

- **Vendor surveys.** Forrester surveyed vendors on their capabilities as they relate to the evaluation criteria. Once we analyzed the completed vendor surveys, we conducted vendor calls where necessary to gather details of vendor qualifications.

- **Product demos.** We asked vendors to conduct demonstrations of their products' functionality. We used findings from these product demos to validate details of each vendor's product capabilities.
- **Customer reference surveys.** To validate product and vendor qualifications, Forrester fielded a reference survey with three of each vendor's current customers.

The Forrester Wave Methodology

We conduct primary research to develop a list of vendors that meet our criteria for evaluation in this market. From that initial pool of vendors, we narrow our final list. We choose these vendors based on 1) product fit; 2) customer success; and 3) Forrester client demand. We eliminate vendors that have limited customer references and products that don't fit the scope of our evaluation. Vendors marked as incomplete participants met our defined inclusion criteria but declined to participate or contributed only partially to the evaluation.

After examining past research, user need assessments, and vendor and expert interviews, we develop the initial evaluation criteria. To evaluate the vendors and their products against our set of criteria, we gather details of product qualifications through a combination of lab evaluations, questionnaires, demos, and/or discussions with client references. We send evaluations to the vendors for their review, and we adjust the evaluations to provide the most accurate view of vendor offerings and strategies.

We set default weightings to reflect our analysis of the needs of large user companies — and/or other scenarios as outlined in the Forrester Wave evaluation — and then score the vendors based on a clearly defined scale. We intend these default weightings to serve only as a starting point and encourage readers to adapt the weightings to fit their individual needs through the Excel-based tool. The final scores generate the graphical depiction of the market based on current offering, strategy, and market presence. Forrester intends to update vendor evaluations regularly as product capabilities and vendor strategies evolve. Vendors marked as incomplete participants met our defined inclusion criteria but declined to participate in or contributed only partially to the evaluation. For more information on the methodology that every Forrester Wave follows, please visit The Forrester Wave™ Methodology Guide on our website.

Integrity Policy

We conduct all our research, including Forrester Wave evaluations, in accordance with the Integrity Policy posted on our website.

Endnotes

[1] DevOps: development plus operations.

[2] ETL: extract, transform, load.

[3] HIPAA: Health Insurance Portability And Accountability Act.

[4] Google's cloud services for machine learning include BigQuery, Cloud ML Engine, Compute Engine, Cloud TPU, Cloud Vision API, and many others.

[5] Forrester plans to evaluate automation-focused PAML solutions separately in 2019. For an overview of the automation-focused PAML segment, see the Forrester report "Now Tech: Predictive Analytics And Machine Learning Solutions, Q2 2018."

FORRESTER®
CHALLENGE THINKING. LEAD CHANGE.

We work with business and technology leaders to develop customer-obsessed strategies that drive growth.

PRODUCTS AND SERVICES

- Core research and tools
- Data and analytics
- Peer collaboration
- Analyst engagement
- Consulting
- Events

Forrester's research and insights are tailored to your role and critical business initiatives.

ROLES WE SERVE

Marketing & Strategy Professionals

CMO
B2B Marketing
B2C Marketing
Customer Experience
Customer Insights
eBusiness & Channel Strategy

Technology Management Professionals

CIO
- Application Development & Delivery

Enterprise Architecture
Infrastructure & Operations
Security & Risk
Sourcing & Vendor Management

Technology Industry Professionals

Analyst Relations

CLIENT SUPPORT

For information on hard-copy or electronic reprints, please contact Client Support at +1 866-367-7378, +1 617-613-5730, or clientsupport@forrester.com. We offer quantity discounts and special pricing for academic and nonprofit institutions.