AWS Announces Six New Amazon SageMaker Capabilities

Amazon SageMaker Canvas expands access to machine learning by providing business analysts the ability to generate more accurate machine learning predictions using a point-and-click interface—no coding required

Amazon SageMaker Ground Truth Plus offers a fully managed data labeling service that uses a highly skilled workforce and built-in workflows to deliver high-quality annotated data for training machine learning models faster at lower cost

Amazon SageMaker Studio now makes data engineering, analytics, and machine learning workflows accessible within a universal notebook

Amazon SageMaker Training Compiler helps customers train deep learning models up to 50% faster by automatically compiling code to make it more efficient

Amazon SageMaker Inference Recommender automatically suggests the optimal AWS compute instances for running machine learning inference with the best price performance

Amazon SageMaker Serverless Inference offers serverless compute for machine learning inference at scale

LAS VEGAS—December 1, 2021—Today, at AWS re:Invent, Amazon Web Services, Inc. (AWS), an Amazon.com, Inc. company (NASDAQ: AMZN), announced six new capabilities for its industry-leading machine learning service, Amazon SageMaker, that make machine learning even more accessible and cost effective. Today's announcements bring together powerful new capabilities, including a no-code environment for creating accurate machine learning predictions, more accurate data labeling using highly skilled annotators, a universal Amazon SageMaker Studio notebook experience for greater collaboration across domains, a compiler for machine learning training that makes code more efficient, automatic compute instance selection machine learning inference, and serverless compute for machine learning inference. To get started with Amazon SageMaker, visit aws.amazon.com/sagemaker.

Driven by the availability of virtually infinite compute capacity, a massive proliferation of data in the cloud, and the rapid advancement of the tools available to developers, machine learning has become mainstream across many industries. For years, AWS has focused on making machine learning more accessible to a broader audience of customers. Today, Amazon SageMaker is one of the fastest growing services in AWS history with tens of thousands of customers, including AstraZeneca, Aurora, Capitol One, Cerner, Discovery, Hyundai, Intuit, Thomson Reuters, Tyson, Vanguard, and many more customers who use the service to train machine learning models of all sizes, some of which on the extreme now consist of billions of parameters capable of making hundreds of billions of predictions every month. As customers further scale their machine learning model training and inference on Amazon SageMaker, AWS has continued to invest in expanding the service's capability, delivering more than 60 new Amazon SageMaker features and functionalities in the past year alone. Today's announcements build on these advancements to make it even easier to prepare and gather data for machine learning, train models faster, optimize the type and amount of compute needed for inference, and expand machine learning to an even broader audience.

- **Amazon SageMaker Canvas no-code machine learning predictions:** Amazon SageMaker Canvas expands access to machine learning by providing business analysts (line-of-business employees supporting finance, marketing, operations, and human resources teams) with a visual interface that allows them to create more accurate machine learning predictions on their own—without requiring any machine learning experience or having to write a single line of code. As more companies seek to reinvent their businesses and customer experiences with machine learning, more people in their organizations need to be able to use advanced machine learning technology across different lines of business. However, machine learning has typically required specialized skills that can require years of formal education or intensive training with a challenging and evolving curriculum. Amazon SageMaker Canvas solves this challenge by providing a visual, point-and-click user interface that makes it easy for business analysts to generate predictions. Customers point Amazon SageMaker Canvas to their data stores (e.g. Amazon Redshift, Amazon S3, Snowflake, on-premises data stores, local files, etc.), and the Amazon SageMaker Canvas provides visual tools to help users intuitively prepare and analyze data. Amazon SageMaker Canvas then uses automated machine learning to build and train machine learning models without any coding. Business analysts can review and evaluate models in the Amazon SageMaker Canvas console for accuracy and efficacy for their use case. Amazon SageMaker Canvas also lets users export their models to Amazon SageMaker Studio, so they can share them with data scientists to validate and further refine their models.
- **Amazon SageMaker Ground Truth Plus expert data labeling:** Amazon SageMaker Ground Truth Plus is a fully managed data labeling service that uses an expert workforce with built-in annotation workflows to deliver high-quality data for training machine learning models faster and at lower cost with no coding required. Customers need increasingly larger datasets that are correctly labeled to train ever more accurate models and scale their machine learning deployments. However, producing large datasets can take anywhere from weeks to years and often requires companies to hire a workforce and create workflows to manage the process of labeling data. In 2018, AWS launched Amazon SageMaker Ground Truth to make it easier for customers to produce labeled data using human annotators through Amazon Mechanical Turk, third-party vendors, or their own private workforce. Amazon SageMaker Ground Truth Plus expands on this capability with a specialized workforce with specific domain and industry expertise, as well as qualifications to meet customers' data security, privacy, and compliance requirements for highly accurate data labeling. Amazon SageMaker Ground Truth Plus has a multistep labeling workflow that includes pre-labeling powered by machine learning models, machine validation of human labeling to detect errors and low-quality labels, and assistive labeling features (e.g. 3D cuboid snapping, removal of distortion in 2D images, predict-next in video labeling, and auto-segment tools) to reduce the time required to label datasets and help reduce the cost of procuring high-quality annotated data. To get started, customers simply point Amazon SageMaker Ground Truth Plus to their data source in Amazon Simple Storage Service (Amazon S3) and provide their specific labeling requirements (e.g. instructions for how medical experts should label anomalies in radiology images of lungs). Amazon SageMaker Ground Truth Plus then creates a data labeling workflow and provides dashboards that allow customers to follow data annotation progress, inspect samples of completed labels for quality, and provide feedback to generate high-quality data so customers can build, train, and deploy highly accurate machine learning models more quickly.
- **Amazon SageMaker Studio universal notebooks:** A universal notebook for Amazon SageMaker Studio (the first complete IDE for machine learning) provides a single, integrated environment to perform data engineering, analytics, and machine learning. Today, teams across different data domains want to collaborate using a range of data engineering, analytics, and machine learning

workflows. The practitioners of these domains often cross areas of knowledge like data engineering, data analytics, and data science and want to be able to work across the various workflows without needing to switch data exploration tools. However, when customers are ready to integrate data across analytics and machine learning environments, they often have to juggle multiple tools and notebooks, which can be cumbersome, time consuming, and prone to error. Amazon SageMaker Studio now allows users to interactively access, transform, and analyze a wide range of data for multiple purposes all from within a universal notebook. With built-in integration with Spark, Hive, and Presto running on Amazon EMR clusters and data lakes running on Amazon S3, customers can now use Amazon SageMaker Studio to access and manipulate data in a universal notebook without having to switch services. In addition to developing machine learning models using their preferred framework (e.g. TensorFlow, PyTorch, or MXNet) to build, train, and deploy machine learning models in Amazon SageMaker Studio, customers can browse and query data sources, explore metadata and schemas, and start processing jobs for analytics or machine learning workflows—without leaving the universal Amazon SageMaker Studio notebook.

- **Amazon SageMaker Training Compiler for machine learning models:** Amazon SageMaker Training Compiler is a new machine learning model compiler that automatically optimizes code to use compute resources more effectively and reduce the time it takes to train models by up to 50%. Today's state-of-the-art deep learning models are so large and complex that they require specialized compute instances to accelerate training and can consume thousands of hours of graphical processing unit (GPU) compute time to train a single model. To further accelerate training times, data scientists typically try to augment training data or tune hyperparameters (variables that govern the machine learning training process) to find the best performing and least resource-intensive version of a model. This work is technically complicated, and data scientists often do not have time to optimize the frameworks needed to train models to run on GPUs. Amazon SageMaker Training Compiler is a new machine learning model compiler that is integrated with the versions of TensorFlow and PyTorch in Amazon SageMaker that have been optimized to run more efficiently in the cloud, so data scientists can use their preferred frameworks to train machine learning models through more efficient use of GPUs. With a single click, Amazon SageMaker Training Compiler automatically optimizes the trained model and compiles it to execute training up to 50% faster.
- **Amazon SageMaker Inference Recommender automatic instance selection:** Amazon SageMaker Inference Recommender helps customers automatically select the best compute instance and configuration (e.g. instance count, container parameters, and model optimizations) to power a particular machine learning model. For large machine learning models commonly used for natural language processing or computer vision, selecting a compute instance with the best price performance is a complicated, iterative process that can take weeks of experimentation. Amazon SageMaker Inference Recommender removes the guesswork and complexity of determining where to run a model and can reduce the time to deploy from weeks to hours by automatically recommending the ideal compute instance configuration. Data scientists can use Amazon SageMaker Inference Recommender to deploy the model to one of the recommended compute instances, or they can use the service to run a performance benchmark simulation across a range of selected compute instances. Customers can review benchmark results in Amazon SageMaker Studio and evaluate the tradeoffs between different configuration settings including latency, throughput, cost, compute, and memory.
- **Amazon SageMaker Serverless Inference for machine learning models:** Amazon SageMaker Serverless Inference offers pay-as-you-go pricing inference for machine learning models deployed in production. Customers are always looking to optimize costs when using machine

learning, and this becomes increasingly important for applications that have intermittent traffic patterns with long idle times. For example, applications like personalized recommendations based on consumer purchase patterns, chatbots fielding incoming customer calls, and forecasting demand based on real-time transactions can have peaks of activity based on external factors like weather conditions, promotional offerings, or holidays. Providing just the right amount of compute for machine learning inference is a difficult balancing act. In some cases, customers over-provision capacity to accommodate peak activity, which allows for consistent performance but wastes money when there is no traffic. In other cases, customers under-provision compute to constrain costs at the expense of providing enough compute capacity to perform inference when conditions change. Some customers try manually adjusting computing resources on the fly to accommodate changing conditions, but this is tedious and manual work. Amazon SageMaker Serverless Inference for machine learning automatically provisions, scales, and turns off compute capacity based on the number of inference requests. When customers deploy their machine learning model into production, they simply select the serverless deployment option in Amazon SageMaker, and Amazon SageMaker Serverless Inference manages compute resources to provide the precise amount of compute needed. With Amazon SageMaker Serverless Inference, customers only pay for the compute capacity they use for each request and the amount of data processed, without having to manage the underlying infrastructure.

"Customers across all industries and sizes are excited about how Amazon SageMaker has helped them scale their use of machine learning such that it has become a core part of their operations and allows them to invent new products, services, and experiences for the world," said Bratin Saha, Vice President of Amazon Machine Learning at AWS. "We're excited to expand our industry-leading machine learning service to an even broader group of customers, so they too can drive innovation in their business and help solve challenging problems. With these new Amazon SageMaker tools, we're introducing a whole new group of users to the service while also providing additional capabilities for existing customers to make it easier to transform data into valuable insights, accelerate time to deployment, improve performance, and save money throughout the machine learning journey."

The BMW Group, headquartered in Munich, Germany, is a global manufacturer of premium automobiles and motorcycles, covering the brands BMW, BMW Motorrad, MINI, and Rolls-Royce. It also provides premium financial and mobility services. "The use of artificial intelligence as a key technology is an integral element in the process of digital transformation at the BMW Group. The company already employs AI throughout the value chain, enabling it to generate added value for customers, products, employees, and processes. In the past few years, we have industrialized many top BMW Group use cases, measured by business value impact," said Marc Neumann, Product Owner, AI Platform at The BMW Group. "We believe Amazon SageMaker Canvas can add a boost to our AI/ML scaling across the BMW Group. With SageMaker Canvas, our business users can easily explore and build ML models to make accurate predictions without writing any code. SageMaker also allows our central data science team to collaborate and evaluate the models created by business users before publishing them to production."

Siemens Energy is energizing society. They are transforming in key focus areas of environmental, social, and governance (ESG) and their innovation is making the future of tomorrow different today, for both their partners—and their people. "The core of our data science strategy at Siemens Energy is to bring the power of machine learning to all business users by enabling them to experiment with different data sources and machine learning frameworks without requiring a data science expert. This enables us to

increase the speed of innovation and digitalization of our energy solutions such as Dispatch Optimizer and Diagnostic services," said Davood Naderi, Data Science Team Lead at Industrial Applications for Siemens Energy. "We found Amazon SageMaker Canvas a great addition to the Siemens Energy machine learning toolkit, because it allows business users to perform experiments while also sharing and collaborating with data science teams. The collaboration is important because it helps us productionalize more ML models and ensure all models adhere to our quality standards and policies."

Airbnb is one of the world's largest marketplaces for unique, authentic places to stay and things to do, offering over 7 million accommodations and 40,000 handcrafted activities, all powered by local hosts. "At Airbnb, we are increasingly integrating ML across all aspects of our business. As a result, our teams consistently need to generate and maintain high-quality data in order to train and test ML models," said Wei Luo, Data Scientist at Airbnb China. "We were looking for a way to generate high quality text classification data results on one hundred thousand paragraphs of customer service logs in Mandarin so we can better serve our customers and reduce dependencies on our customer service team. With Amazon SageMaker Ground Truth Plus, the AWS team built a customized data labeling workflow, which included a customized ML model that was able to achieve 99% classification accuracy."

The National Football League is America's most popular sports league, comprised of 32 franchises that compete each year to win the Super Bowl, the world's biggest annual sporting event. "At the NFL, we continue to look for new ways to use machine learning to help our fans, broadcasters, coaches, and teams benefit from deeper insights," said Jennifer Langton, SVP, Player Health and Innovation at NFL. "Football is a fast moving sport where plays can happen in a split second. While coaches and referees carefully watch the game, it can be difficult to watch all players on a field for safety. Computer vision allows us to accurately detect player safety incidents, but developing these algorithms requires expertly labeled data. Now with Amazon SageMaker Ground Truth Plus, we have custom workflows and user interfaces for sophisticated labeling tasks, which helps us improve player safety."

Founded and headquartered in Orange County, California, VIZIO's mission is to deliver immersive entertainment and compelling lifestyle enhancements that make its products the center of the connected home. VIZIO is driving the future of televisions through its integrated platform of cutting-edge Smart TVs and powerful SmartCast operating system. VIZIO's platform gives content providers more ways to distribute their content and advertisers more tools to target and dynamically serve ads to a growing audience that is increasingly transitioning away from linear TV. "At VIZIO, we consistently look for ways to leverage ML to create personalized experiences for our customers. We were looking for a way to continuously review ad videos and generate commercial metadata for efficient ads classification," said Zeev Neumeier, Chief Innovation Officer at VIZIO. "With the use of Amazon SageMaker Ground Truth Plus's streaming capability, we can now use a custom template which provides video classification, metadata collection, and an automated system that enables data collection in real time as ads air. With Amazon SageMaker Ground Truth Plus, we are able to review the results in less than one business day."

Litterati is a data science company empowering people to 'crowdsource-clean' the planet. Litterati's platform empowers people to create better solutions for the litter and waste problems our world faces by developing behavioral insight, mapping problem areas, and mitigating future risk. From schools to scientists, environmental organizations, brands, and city governments, people are coming together using Litterati for the greater good to create a litter-free world. "For us, machine learning brings light to unseen challenges. In the US alone, each year billions of dollars are spent cleaning up litter," said Sean Doherty, CTO at Litterati. "With computer vision models, we transform images of litter all around the

world into data, so cities can better allocate their litter management resources. However, building object detection models requires access to object, material, and brand information, as well as localized knowledge due to datasets being spread across the globe. Amazon SageMaker Ground Truth Plus allows us to create a hierarchical annotation interface that captures these precise features within that localized context. In addition, the SageMaker Ground Truth Plus expert workforce created localized image annotations, which provides a standardized solution increasing our data labeling efficiency by up to 20%, accelerating our ability to ingest annotated results into our database by 200%, and reducing post-processing time by 90%."

Provectus helps its customers build end-to-end data and machine learning engineering experiences from raw datasets, enterprise data lakes, and machine learning models. "We have been waiting for a feature to create and manage Amazon EMR clusters directly from Amazon SageMaker Studio so that our customers could run Spark, Hive, and Presto workflows directly from Amazon SageMaker Studio notebooks," said Stepan Pushkarev, CEO at Provectus. "We are excited that Amazon SageMaker has now natively built this capability to simplify management of Spark and machine learning jobs. This will help our customers' data engineers and data scientists collaborate more effectively to perform interactive data analysis and develop machine learning pipelines with EMR-based data transformations."

The Vanguard Group, Inc., is an American registered investment advisor based in Malvern, Pennsylvania, with about $7 trillion in global assets under management. Vanguard is redefining the industry by doing what's right for investors and creating change for millions of clients worldwide. "We're excited that our Vanguard data scientists and data engineers can now collaborate in a single notebook for analytics and machine learning," said Doug Stewart, Senior Director of Data and Analytics at Vanguard. "Now that Amazon SageMaker Studio has built-in integrations with Spark, Hive, and Presto all running on Amazon EMR, our development teams can be more productive. This single development environment will allow our teams to focus on building, training, and deploying machine learning models."

Quantum Health is on a mission to make healthcare navigation smarter, simpler, and more cost-effective for everyone. They use Amazon SageMaker for use cases like text classification, text summarization, predictive models, classification problems, and Q&A to help the Quantum team and the members they serve. "Iterating with NLP models can be a challenge because of their size. Long training times bog down workflows and high costs can discourage our team from trying larger models that might offer better performance," said Jorge Lopez Grisman, Senior Data Scientist at Quantum Health. "Amazon SageMaker Training Compiler is exciting because it has the potential to alleviate these frictions. Achieving a speedup with Amazon SageMaker Training Compiler is a real win for our team that will make us more agile and innovative moving forward."

Guidewire is the platform property and casualty insurers trust to engage, innovate, and grow efficiently. The company combines digital, core, analytics, and AI to deliver its platform as a cloud service, and it enables its customers to do advanced analytics and machine learning for their industry-specific workloads. More than 450 insurers, from new ventures to the largest and most complex in the world run on Guidewire. "One of Guidewire's services is to help customers develop cutting-edge NLP models for applications like risk assessment and claims operations. Amazon SageMaker Training Compiler is compelling because it offers time and cost savings to our customers while developing these NLP models," said Matt Pearson, Principal Product Manager, Analytics and Data Services at Guidewire Software. "We expect it to help us reduce training time by more than 20% through more efficient use of GPU resources. We are excited to implement Amazon SageMaker Training Compiler in our NLP workloads, helping us to accelerate the transformation of data to insight for our customers."

Musixmatch is a leading music data company providing data, tools, and services that enrich the way we experience music such as searching for songs and sharing song lyrics. Musixmatch is the largest service of this kind in the world with over 80 million users and over 8 million distinct lyrics. “Musixmatch uses Amazon SageMaker to build natural language processing and audio processing models, and is experimenting using Hugging Face with Amazon SageMaker. We choose Amazon SageMaker because it allows data scientists to iteratively build, train, and tune models quickly without having to worry about managing the underlying infrastructure, which means data scientists can work more quickly and independently,” said Loreto Parisi, AI Engineering Director at Musixmatch. “As the company has grown, so too have our requirements to train and tune larger and more complex NLP models. We are always looking for ways to accelerate training time while also lowering training costs which is why we are excited about Amazon SageMaker Training Compiler. SageMaker Training Compiler provides more efficient ways to use GPUs during the training process and, with the seamless integration between SageMaker Training Compiler, PyTorch, and high-level libraries like Hugging Face, we have seen a significant improvement in training time of our transformer-based models going from weeks to days as well as lower training costs.”

Loka, a machine learning consulting firm, helps its clients harness and build ML into their products across a wide range of use cases to deliver better customer experiences. “We spend a lot of time and effort optimizing models, tuning servers, and testing instance types to deliver performant, scalable, and cost effective ML environments for its client,” said Bobby Mukherjee, CEO at Loka. “Now using Amazon SageMaker Inference Recommender, our engineers are able to get an ML model deployed to production within minutes from any location.”

Holmusk, a digital health company, launched its FoodDX app to help people improve their diet and health. “Our food image recognition algorithms need low latency to ensure our users get the right diet recommendations at the right time. To achieve low latency, we were over-provisioning GPUs, which was expensive,” said Sai Subramanian, CTO at Holmusk. “Using Amazon SageMaker Inference Recommender, we can now easily conduct load tests across different instances and determine an instance configuration within hours to reduce our compute costs significantly while maintaining latency requirements. This is a huge win for our team and lets our ML scientists focus on creating algorithms to help people live healthier lives rather than managing infrastructure.”

Qualtrics is an experience management company that helps extract information from customer surveys using natural language processing (NLP) models. "Amazon SageMaker Inference Recommender improves the efficiency of our MLOps teams with the tools required to test and deploy machine learning models at scale,” said Samir Joshi, ML Engineer at Qualtrics. “With Amazon SageMaker Inference Recommender, our team can define latency and throughput requirements and quickly deploy these models faster, while also meeting our budget and production criteria."

iFood, a leading player in online food delivery in Latin America fulfilling over 60 million orders each month, uses machine learning to make restaurant recommendations to its customers ordering online. “We have been using Amazon SageMaker for our machine learning models to build high-quality applications throughout our business," said Ivan Lima, Director of Machine Learning and Data Engineering at iFood. "With Amazon SageMaker Serverless Inference, we expect to be able to deploy even faster and scale models without having to worry about selecting instances or keeping the endpoint active when there is no traffic. With this, we also expect to see a cost reduction to run these services.”

About Amazon Web Services

For over 15 years, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud offering. AWS has been continually expanding its services to support virtually any cloud workload, and it now has more than 200 fully featured services for compute, storage, databases, networking, analytics, machine learning and artificial intelligence (AI), Internet of Things (IoT), mobile, security, hybrid, virtual and augmented reality (VR and AR), media, and application development, deployment, and management from 81 Availability Zones within 25 geographic regions, with announced plans for 27 more Availability Zones and nine more AWS Regions in Australia, Canada, India, Indonesia, Israel, New Zealand, Spain, Switzerland, and the United Arab Emirates. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—trust AWS to power their infrastructure, become more agile, and lower costs. To learn more about AWS, visit aws.amazon.com.

About Amazon

Amazon is guided by four principles: customer obsession rather than competitor focus, passion for invention, commitment to operational excellence, and long-term thinking. Amazon strives to be Earth's Most Customer-Centric Company, Earth's Best Employer, and Earth's Safest Place to Work. Customer reviews, 1-Click shopping, personalized recommendations, Prime, Fulfillment by Amazon, AWS, Kindle Direct Publishing, Kindle, Career Choice, Fire tablets, Fire TV, Amazon Echo, Alexa, Just Walk Out technology, Amazon Studios, and The Climate Pledge are some of the things pioneered by Amazon. For more information, visit amazon.com/about and follow @AmazonNews.