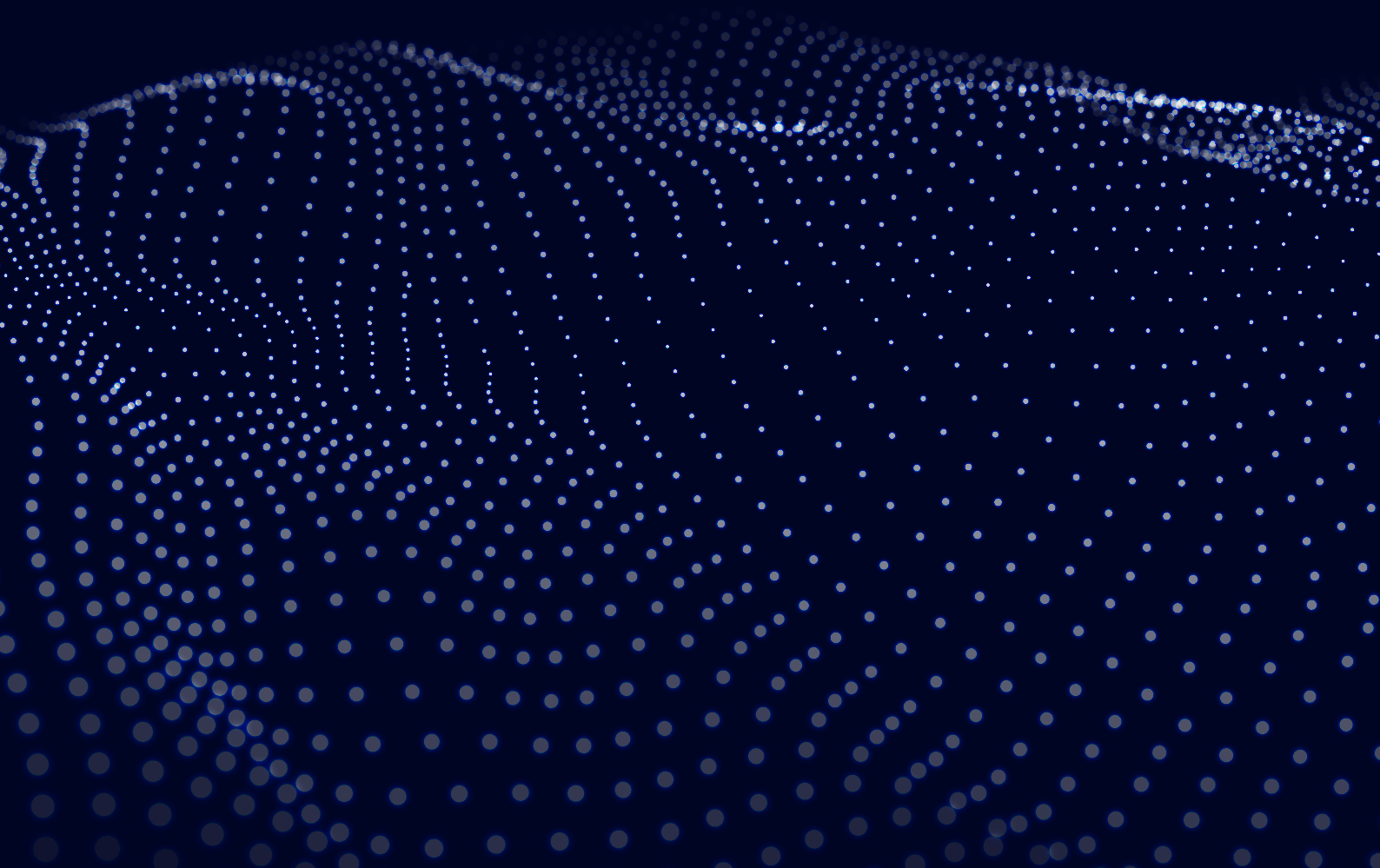
VIANOPS™

AN OVERVIEW OF

MLOps



Contents

Executive Summary

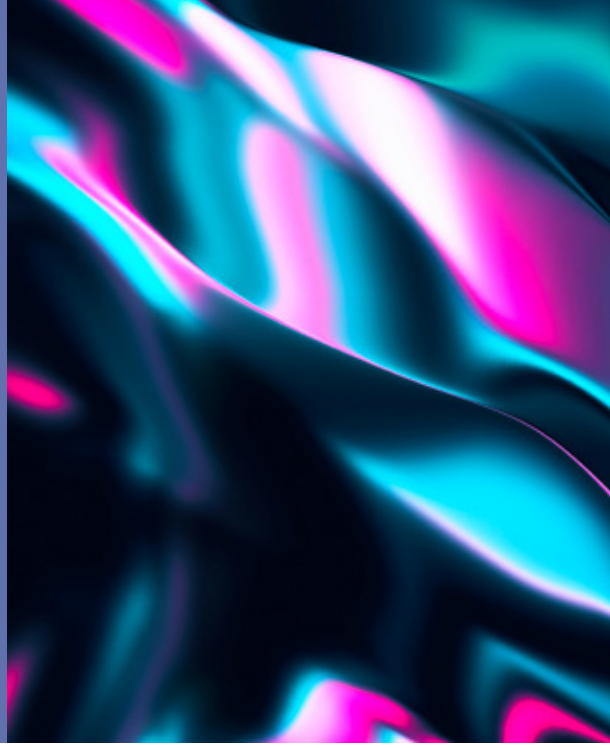
To better understand and gain insights from their vast amounts of data, companies have accelerated their adoption of Artificial Intelligence (AI) and Machine Learning (ML) with the goal of making better predictions, creating innovative services for customers, and delivering business value faster. But most are struggling to move ML models from design to production. Gartner states that only 47% of all AI and machine learning models make it into production, and they take an average of nine months to get there. 451 Research focuses specifically on ML projects, noting that 34% fail to reach production. IDC's perspective shows an even worse scenario, stating that only 31% of organizations have functioning models in production, and only one-third of those have reached a mature state of adoption and are delivering enterprise-wide value.

The fundamental reason for many of these challenges is that models are different from traditional software, yet most organizations do not have tools and processes in place that account for these differences. And, stakeholders from business, data science, IT operations, risk management and other teams are involved in taking a model from design to production, often leading to communication gaps among people that typically don't speak the same business, technical or process-oriented language, or use the same tools. This paper will show how using a unified MLOps platform can drive collaboration across these disparate teams, mitigate financial, reputational and regulatory risk, and deploy models at scale to more quickly deliver business value.

It will introduce the VIANOPS platform, an extensible, enterprise-grade machine learning operations platform that facilitates the orchestration of people, processes, and technologies across a model's design to production lifecycle – from optimization, validation and deployment, to monitoring and retraining – until the model is retired. Our platform is uniquely positioned to help organizations deploy more models into production faster, optimize for performance and throughput gains on commodity hardware as well as edge devices, and monitor for model and data drift, uncertainty, bias and more.

What is MLOps?

MLOps is set of best practices for communicating, collaborating, and efficiently orchestrating the many tasks and teams involved in deploying, monitoring, and assessing the viability and performance of ML models in production. Just as DevOps bridges gaps in the discrete steps across development and operations teams, MLOps combines machine learning with operations by bringing automation and orchestration to people, process, and technology to manage models at scale.



Why do we need MLOps?

Companies have made significant investments to capture and leverage data, largely from a historical point of view. Now with recent advancements in artificial intelligence and machine learning tools, companies are increasingly turning to AI/ML models to accelerate insights from data, more quickly pivot to new business opportunities, rapidly identify problems, and build the solutions to solve them. ML models are an efficient way to extract business value from data. But to realize this value, these models must be put into production quickly using repeatable processes, and at scale.

Therein lies the challenge. As discussed earlier, less than 50% of models make it into production, and the ones that do take an average of nine months to get there. The current path from model creation to testing to production is often a complex, disjointed one that requires multiple ad-hoc workflows that span disparate, siloed teams of data engineers, data scientists, IT operations, auditors, business domain experts and ML engineering teams. And there are over 300 tools used by these various stakeholders, some that span more than one set of tasks, others that are more narrowly focused. This makes it difficult for organizations to standardize on workflows.

To drive efficiency and accelerate the ability to realize business value, a repeatable, streamlined processes is required to create, manage, optimize, deploy, and monitor models while accounting for myriad factors such as versioning, usage, governance, regulatory scrutiny, and potential security risks.

This is what MLOps provides. It helps organizations orchestrate disparate systems into a custom-built toolchain that connects different pieces of complex applications, and help companies become truly data driven.

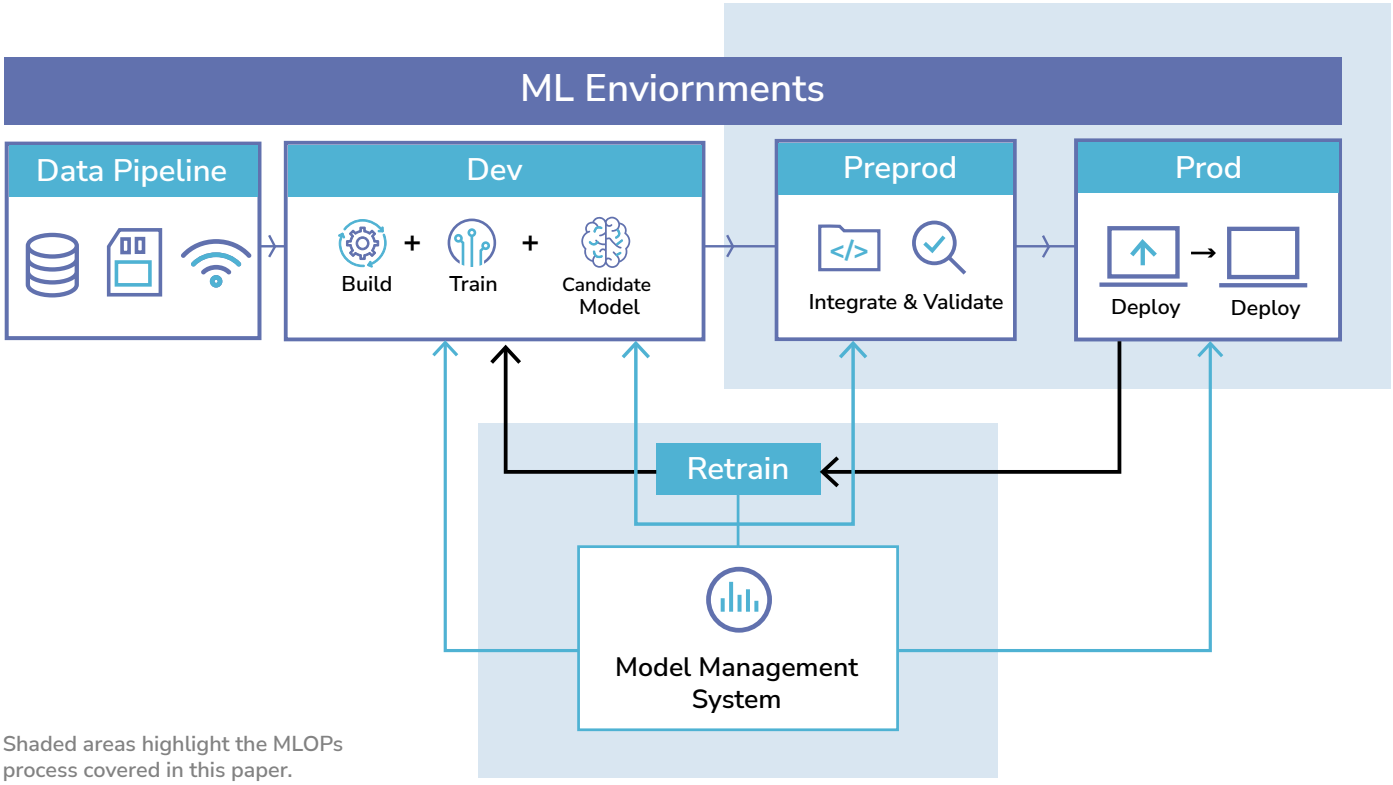
Biggest pitfalls in leveraging AI/ML:

AI/ML Annual Research Report 2022, Rackspace

ML Models are Different from Traditional Software

The fundamental reason for challenges across the MLOps process is that ML models are different from traditional software. And yet, most organizations do not have processes in place that account for these differences. Unlike traditional software, ML models are highly dependent on data that is constantly changing. And that data impacts models. While applications remain fairly static once they are deployed, a model's performance and behavior changes after it is deployed to production. Monitoring these changes is critical to understanding the reliability, trustworthiness and risk level of a model as well as knowing when a model needs to be retrained with new data, modified, or retired.

Traditional software code is tested in a rigorous manner and deployed to a known environment. In contrast, ML models are probabilistic, dependent on data, and may be deployed to many different environments, often unlike the environments where they were built and tested. The process of testing models uses both training data and testing data which may or may not reflect real-world data. Tests are done in iterative experiment runs with the goal of identifying acceptable thresholds for accuracy and performance.



Shaded areas highlight the MLOPs process covered in this paper.

And unlike code, a model's results, which are predictions, change over time as the real-world input it receives also changes. And predictions may become less accurate if the assumptions upon which they are based change. This happened during COVID-19 when consumer buying patterns and many other behaviors shifted dramatically, rendering the signals and features upon which models were built to be unreliable. And, models can be deployed to many different environments, often to end points that vary significantly from the environments in which they were trained.

MLOps is Different Across Organizations

Every company will implement the process of operationalizing models a bit differently, using a variety of tools and roles across the different steps. The following two examples provide insight into why a unified and modular platform is needed to manage the entire model's lifecycle.

In some companies, data scientists own models throughout the entire lifecycle, by choice or by necessity. However, this highly skilled and highly paid talent often ends up spending valuable time trying to get help from IT to access, configure, and manage infrastructure to run experiments. Instead, this process should be automated, freeing up the data scientist to bring more business value by focusing on creating models, not operationalizing them.

In other organizations, models may be created by a separate group, or even an outside organization. Teams that bring these models to production typically lack insight to understand a model's context and lineage – what it does, how it was trained, what its data looks like and what compliance, security, and regulatory requirements it may have. And, once a model is in production, teams must monitor models for drift, analyze its performance, and retrain when necessary.

> "Successful enterprise ML at scale demands the careful orchestration of a complex tapestry made up of people, processes, and platforms, an effort that does not end when an ML solution goes live but instead continues for the life of the solution."
>
> — Omdia

A unified MLOps platform can ensure that regardless of which person in an organization is performing a specific task or set of tasks, they will be able to do so seamlessly with access to the data they need and with an intuitive UI to complete each step. And in turn, this will improve productivity, efficiency, and mitigate risk. For the data scientist, streamlining the deployment process eliminates downtime; for the person importing a model, it provides them a way to explore a model's provenance, lineage, and governance requirements, mitigating risk and ensuring that compliance requirements are met.

An Overview of the MLOps Process within the Full Model Lifecycle

We view MLOps as a process that streamlines and operationalizes taking machine learning models into production and maintaining and monitoring them. But sometimes people use the term MLOps to refer to the entire lifecycle of a model. It's important to understand what happens earlier in a model's lifecycle before the MLOps process begins because these steps – including data gathering and wrangling, model creation, and experimentation – are often performed by people from different teams who may or may not regularly interact with each other, or with MLOps tools and teams. This lack of communication and collaboration can create gaps in the MLOps process because at each stage during a model's lifecycle, decisions are made and actions are taken that may impact a model's performance, governance, security, reproducibility, and traceability.

The full lifecycle of a machine learning model is outlined below, with an overview of the important tasks that occur before the MLOps process. Depending on an organization's maturity level, the tasks may be done by different teams, and in other cases the same people may perform more than one of these sets of tasks.

01 - Data Acquisition and Preparation

There are a number of steps in this process, including data wrangling, where data engineers gather data from a variety of sources and transform it into a format that can be used by data scientists.

Data is merged, cleansed by removing duplicates, unnecessary data points and extreme outliers, enriched by adding additional values from other data sets, and validated. Data pipelines are created, and data is stored efficiently and securely so users can access and analyze data while ensuring compliance with regulatory and organizational governance.

02 - Model Build, Training, & Selection

Next, data scientists explore the data to make sense of it, extract key features, test variables with correlation and identify trends. Feature engineering – or refining the features or values that are critical to a model's ability to perform predictions – is part of an iterative process that includes training a model with the goal of finding the best performing model to be deployed into production. Training is done with experiments, and experiments include multiple runs. These runs involve changes to the model or its artifacts, such as transformers or pipelines, and teams often use a model registry or catalog to efficiently manage and track the multiple versions created.

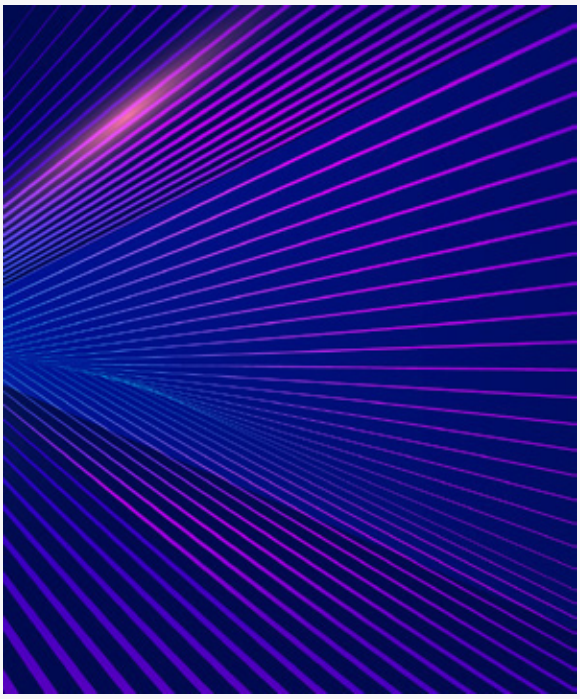
An Overview of the MLOps Process (con't)

03 - Packaging & Deployment

This is the first step in the MLOps process. Once models have been trained, the best performing model is typically handed off to the operations team who can optimize the model for its target environment and deploy it along with the required artifacts. There are several types of strategies that can be used to safely deploy a model, and to rapidly discover any problems that might have been missed during the build and test phases.

Challenges:

- Comparing model metrics side-by-side is difficult when many changes are made in a highly iterative process
- Packaging a model in a way that ensures reproducibility is difficult when data, assets and metadata are stored across different frameworks
- Integrating with one or more business workflows or applications can be complex
- Model execution is often a slow, manual, ad-hoc process that may take days or even months for a single model



04 - Monitoring & Observability

Once models have been deployed into production and are serving inferences based on real-worl data, they need to be monitored for accuracy, performance, drift and bias. Unlike traditional software systems, a model's behavior is influenced by code, training data and the real-world data it is being fed in production – three key areas where things can go wrong. Organizations need the ability to monitor model performance and behavior, identify problems and gain insight into the root cause of problems so they can be addressed, and the model put back into production where it is delivering business value.

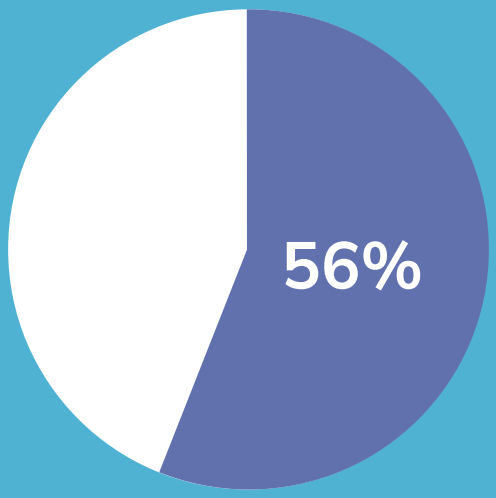
Challenges:

- Model performance in production can differ greatly from testing environments in latency, throughput, performance, prediction accuracy, bias, & fairness
- Many organizations lack MLOps maturity and do not continuously monitor models
- To keep models in production and delivering business value, teams must know as soon as possible when model behavior changes so they can pinpoint and address the root cause, minimizing its downtime
- Teams may lack access to traceable and repeatable components, such as pipelines, to rapidly retrain a model

05 - Risk & Governance

Model risk can vary greatly – some low-risk models are recommendation engines to determine what to show on a web page, others may make critical decisions about the maintenance requirements for equipment or the likelihood that a transaction is fraudulent.

Organizations need to assess and implement processes to manage risk, both to ensure public trust and to comply with regulatory, privacy or other requirements. Companies must define processes for the governance of data as well as the entire model lifecycle from defining the business goal, data discovery, choice of algorithm, data processing, through model building, testing, deployment and monitoring.



Organizations that find implementing governance a top challenge - Algorithmia

Governance of data includes things like the time it is collected, any terms of use, data accuracy and types of data such as Personally Identifiable Information (PII) or other types of sensitive data. For models, it covers everything from access control for all production models and the systems they run on including on-prem, private cloud, public cloud or the edge, workflow approval flows, versioning, logging and auditability.

Models are highly dependent upon data, presenting many opportunities for risk, including risk that an incorrect prediction is given for a set of data or that the model's accuracy or its fairness decreases over time. And there is a risk of legal consequences when critical software does not operate properly, and its output does not align with established standards or requirements.

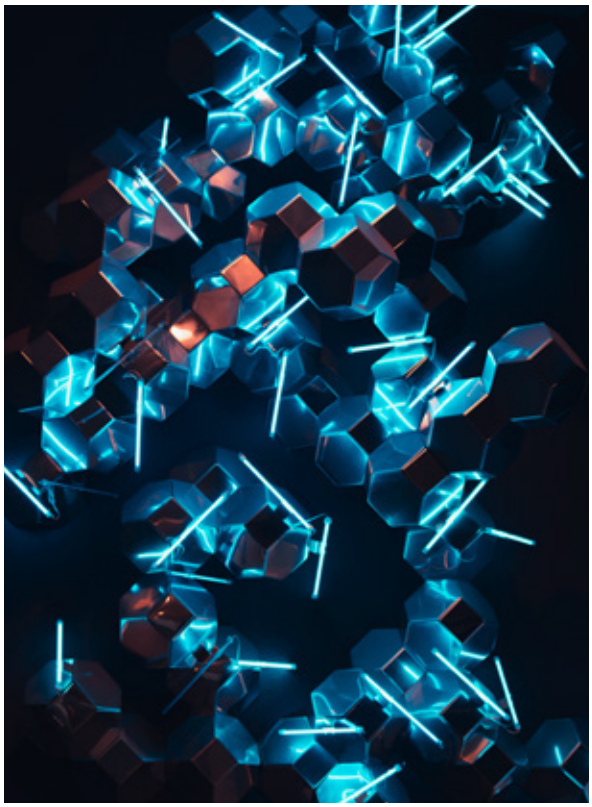
To evaluate a model, organizations need to understand its intended purpose, and have knowledge of the many decisions made during data gathering, processing, testing, and training. And, it's only possible to truly assess a model's performance when it is in production using real-world data.

Challenges:

- Lack of well-established best practices for model governance
- Model signoff required to ensure adherence to governance and security policies – 67% of companiesiv must comply with multiple regulations (ISO, HIPAA, PCI, GDPR, etc.)
- Governance has a broad reach that spans each step in a model's lifecycle, from data to various engineering frameworks and algorithms, to how models are deployed

The Need for a Unified MLOps Platform

As we have seen, MLOps is often a disjointed set of steps more than a cohesive process. It involves different stakeholders across various parts of an organization, such as data scientists, data engineers, ML or MLOps engineers, IT and business owners. There are many tasks that require collaboration across these teams to share expertise and information in order to operationalize models. And, postdeployment, models must be continuously monitored by people that know what to look for and understand how a model should perform and will be alerted to changes in behavior or drift.



But collaboration across these disparate teams is difficult as teams use different tools, focus on different tasks, have different goals and KPIs, and are often not even aware of what is involved across different steps in the process. In many organizations, models are created by siloed teams or external partners and then handed off to another team to bring them into production. And today's tools do not adequately enable an IT operations team to receive a model created by a data scientist and validate or recreate a model with confidence. Similarly, when a model is finally validated and put into production, today's tools lack the ability to continuously monitor for risk over time, and quickly alert users with critical information about model drift or performance decay so that they can address the root cause and retrain or replace the model.

An overall lack of trust, transparency, and governance of models over their entire lifecycles, exacerbated by rising infrastructure costs, has significantly slowed the time-to-value of ML models – preventing enterprises from realizing the full potential of AI.

Organizations need a unified platform that brings together all the disparate teams and tools, none of which were designed to work together, but must do so seamlessly to deploy, monitor and govern ML models at scale.

The VIANOPS Platform

At Vianai, our purpose and vision are to bring human-centered AI to businesses worldwide. For us, this means abstracting away the complexity for the various stakeholders across the ML model lifecycle, dramatically improving the cost-performance of AI systems and delivering trustworthy, explainable and observable machine learning models at scale.

With the VIANOPS Platform companies can:

- Amplify the knowledge, experiences, and decision-making power of people
- Build trust, transparency, and accountability of AI to drive adoption at scale
- Deliver AI in a responsible way, reducing cost, environmental impact, and risk

With our platform, teams collaborate across all steps in the MLOps process using their preferred tools and technologies. They get more trusted models into production – faster and cheaper – with our Performance Optimization. And they can implement governance, deliver explainability to ensure compliance, and continuously monitor deployed models for risk, drift, and bias. And with dashboards that provide actionable insights to problems, teams can rapidly take steps to retrain and redeploy models, minimizing downtime, and maximizing business value.

Key capabilities within the Platform include:

- **Collaboration -** Facilitates communication and sharing of information across teams with an intuitive, unified UX that guides users to complete tasks
- **Monitoring and Observability -** Provides Model Risk Management through continuous monitoring, custom alerts, and dashboards with drill-down insight into which models are problematic and why performance degradation or drift is occurring – and the ability to take action such as retraining
- **Model Performance Optimization -** With AI infrastructure costs skyrocketing, Vian Performance Optimization enables companies to cut through the cost barriers and leverage existing hardware to improve performance and throughput
- **Deployment Operation -** Bring models into production efficiently, leveraging deployment strategies such as Canary, A/B, and Shadow/Challenger to help teams evaluate model performance, capabilities and discover any issues with the model

An Open and Extensible Platform that Drives Collaboration

The VIANOPS Platform delivers unified access to the many disparate ML frameworks, coding environments, and other tools that span an ML model's lifecycle. This means all users will be productive right from the start using their tools and language of choice. And it's easy to integrate other tools through our APIs so your team remains productive with their existing skills and preferred tools. The user experience is tailored for each stakeholder accessing the platform through an intuitive interface that provides access to the data & information needed for the task at hand.

Model Onboarding | Model Validation | Model Testing | Model Optimization

Model Deployment | Model Inference | Model Monitoring | Model Rebuild/ Retrain

Continous Monitoring

- Auditability, Observability, Explainability
- Pre-deployment & Post-deployment Risk Analysis
- Continuous Monitoring

Enterprise Readiness & Scalability

- Collaboration across all stakeholders & project phases
- Simplifies model-based business solutions as-a-service
- Persona-based UX abstracts complexity across multiple tools

Modular Architecture

We built our platform so users can access the capabilities they need, when they need them, without having to work in linear order. For example, operations teams who receive a model from data scientists can leverage our proprietary optimization technology to improve speed and throughput before deploying a model into production. Or they may use our automated monitoring and retraining for models built outside of our platform.

Data Scientist

Use different Data Science platforms and tools to design models

Self-service to deploy models to production with out dependence on ML or MLOps engineers

ML Engineer

Take models to production - import, register, optimize, package, deploy and serve.

Optimize model execution to minimize running cost

MLOps Engineer

Manage models in production

Continuously monitor model health to mitigate failure and risk

Model Validator

Risk assessment and compliance certification for model release to production.

Periodically review model risk in production and approve risk mitigation plan

The VIANOPS platform was built to enable any user to perform a set of tasks, starting wherever they need to start, while seamlessly spanning one or more modules

Monitoring, Observability, and Model Management

VIANOPS provides a rich set of model risk monitoring capabilities for data quality and integrity, drift, uncertainty, bias, explainability, and more. Leveraging integrated open source as well as Vian proprietary technologies, each model can follow a customized risk monitoring plan after it is deployed.

Once in production, monitoring models for bias, drift, and performance is critical but ownership of this process is often unclear in organizations. Our focus on human-centered collaboration helps all stakeholders understand monitoring metrics and communicate efficiently using data to make decisions about maintaining models in production. We inform users when they should not trust the output of their models, and we make it easy to configure and set customer-defined alerts that can automate workflows once these thresholds are met, such as automated model retraining. Dashboards provide users meaningful insight to the critical data points you're monitoring, filtering out the noise, so you can quickly identify which models are affected, drill down to understand root cause, and take appropriate action.

To mitigate operational and reputational risk, we aim to provide the most comprehensive set of tools for monitoring and addressing risk. We currently deliver solutions for tabular and time-series data with support for NLP and images planned for future releases.

Deployment

Our platform makes it easy for any user to deploy models into production. Our intuitive UI provides details about each model to ensure users are selecting the correct model to deploy and automate the process of optimizing the model for its target environment and deploying it along with all the artifacts it needs to run inferencing. We also support several types of deployment strategies that can be used to safely deploy a model, and to rapidly discover any problems that might have been missed during the build and test phase.

For example, A/B testing can be used to generate quick results that identify and eliminate lowperforming models by pushing out two models that differ somewhat in their features, but this may not be a reliable method as model complexity increases. Using shadow deployment, teams can test a model with new features on real-world data by deploying it alongside the current model in production. Both receive the same requests and make predictions, but only the existing production model delivers the predictions. This can minimize risk by monitoring the shadow model in a production environment before its actual deployment. This type of strategy is often used in combination with a champion– challenger framework where there are multiple shadow models that are tested and compared with the current model in production.

Governance & Risk

Our platform makes it easy to integrate with other systems to provide capabilities needed to deliver repeatable and reportable processes for ML Models, and control access to models and the related data.
In addition to monitoring models in production, we provide the following support for model governance:

Model Lineage

Traceability of where the model comes from, which feature set was used to train the model, and which dataset was used to engineer the feature set. Provides critical information for reproducibility and tracks versioning.

Access Control

Through LDAP integration and role-based access control (RBAC) administrators set user roles and restrictions to limit the scope of access to align with policy.

Model Provenance

Describes the origin and ownership of the model, and processing steps applied to it, accessible for legal & regulatory compliance as audit reports.

A Platform Designed to Evolve

With VIANOPS, we have removed the biggest barriers to bringing ML to the enterprise at scale by:

- Operationalizing ML models, enabling collaboration across disparate teams, and delivering an intuitive UX for each stakeholder
- Providing a comprehensive set of risk monitoring capabilities for evaluating data quality and integrity, drift, uncertainty, bias, explainability and more
- Eliminating the need for cost prohibitive infrastructure to support model deployment whether on-prem, the cloud, or at the edge

Our extensible platform with our intuitive user experience anticipates user actions and drives collaboration across teams. As new ML solutions emerge, these capabilities will be even more critical as users adopt new technologies but lack the skills and expertise to leverage them. VIANOPS will continue to guide users through unfamiliar tasks for efficiency and productivity. The intentional design will continue to focus on augmenting and amplifying human intelligence and deliver new opportunities to enable collaboration. Going forward, in close collaboration with customers, the VIANOPS Platform will deliver a wide range of advanced ML tools and techniques, to help enterprises pursue their most strategic opportunities by leveraging human-centered AI.

— Enterprise AI, June 8, 2022

— Enterprise AI, June 8, 2022

— IDC Survey Illustrates Growing Importance of Purpose-built AI Infrastructure in the Modern Enterprise, February, 2022.

— 2021 Enterprise Trends in Machine Learning, Algorithmia

VIANOPS