MLOPS WORLD 2022

# MACHINE LEARING INTO PRODUCTION

## A GLOBAL EXPLORATIVE GATHERING

Workshops, case-studies
best practices & lessons
learned

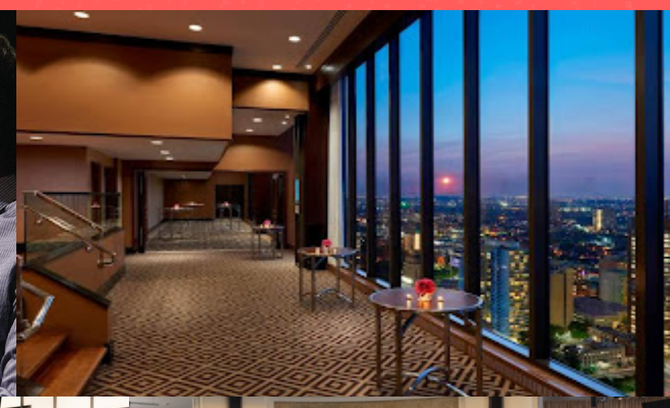
DATA MINING

www.vistacompany.ir

**June 7-10th**

www.mlopsworld.com

# MLOPS WORLD 2022




## 80+ Learn Technical and Strategic Leaders From

Flyte · shopify · Meta · Hugging Face · salesforce

DOORDASH · aporia · ebay · Pinterest · zoom

## Join here

## June 7-10th

virtual & in-person

3rd annual

**MLOps World Machine Learning in Production**

# Day 1: Virtual Workshops June 7th

**Bonus WORKSHOPS Days** June 7-8th Held Virtually for ticket holders Reg Here

***Colors Don't Apply**

10: 00 AM

## Machine Learning Monitoring in Production: Lessons learned from 30+ Use Cases

*Lina Weichbrodt, Lead Machine Learning Engineer,* ***DKB bank***

10: 00 AM

## Implementing MLOps Practices on AWS using Amazon SageMaker

*Shelbee Eigenbrode, Principal AI/ML Specialist Solutions Architect / Bobby Lindsey, AI/ML Specialist Solutions / Kirit Thadaka, ML Solutions Architect,* ***Amazon Web Services***

10: 00 AM

## Workshop Session with Aporia

11: 00 AM

## Automated Machine learning & Tuning with FLAML

*Qingyun Wu, Assistant Professor,* ***Penn State University*** *and Chi Wang, Principal Researcher,* ***Microsoft Research***

12: 00 AM

## Workshop Session with Iterative

12: 00 PM

## Deploy High Scale ML Models Without the Hustle

*Pavel Klushin, Head of Solutions Architecture,* ***Qwak***

12: 00 PM

## Taking MLOps 0-60: How to Version Control, Unify Data and Manage Code Lifecycles

*Jimmy Whitaker, Chief Scientist of AI,* ***Pachyderm***

1: 00 PM

## Production ML for Mission-Critical Applications

Robert Crowe, TensorFlow Developer Engineer, **Google**

2: 00 PM

## Workshop With Robust Intelligence

2: 00 PM

## Implementing a Parallel MLOps Test Pipeline for Open Source Development

*Miguel González-Fierro, Data Science Manager, Microsoft*

2: 00 PM

## Workshop With IBM

3: 00 PM

## Workshop TBA

4: 00 PM

## Deep Dive: How to Treat Your Data Platform Like a Product: 5 Key Best Practices

*Barr Moses, CEO & Co-Founder of* ***Monte Carlo***

4:00 PM

## Deep Dive: WarpDrive: Orders-of-Magnitude Faster Multi-Sgent Deep RL on a GPU

*Stephan Zheng, Lead Research Scientist; Tian Lan, Senior Applied Scientist and Sunil Srinivasa, Research Engineer,* ***Salesforce***

4:00 PM

## Workshop with Petuum

5:00 PM

## Deep Dive: Scaling ML Embedding Models to Serve a Billion Queries

*Senthilkumar Gopal, Senior Engineering Manager (Search ML), Deepika Srinivasan,* ***Senior MTS, (Search ML) Ebay Inc***

3rd annual

**MLOps World Machine Learning in Production**

# Day 2: Virtual Workshops June 8th

**Bonus WORKSHOPS Days** June 7-8th Held Virtually for ticket holders Reg Here

***Colors Don't Apply**

10: 00 AM

## Accelerating Transformers with Hugging Face Optimum and Infinity Workshop

*Lewis Tunstall, Machine Learning Engineer, Philipp Schmid, Technical Lead,* ***Hugging Face***

10: 00 AM

## Personalized Recommendations and Search with Retrieval and Ranking at Scale on Hopsworks

*Jim Dowling, CEO,* ***Hopsworks***

10: 00 AM

## Session with Seldon

11: 00 AM

## Deep Dive: Parallelizing your ETL with Dask on Kubeflow

Jacob Tomlinson *Host and* ***Company Name***

12: 00 PM

## Workshop With ClearML

12: 00 PM

## Workshop: What's in the Box? Automatic ML Model Containerization

Clayton Davis, Head of Data Science, **Modzy**

12: 00 PM

## A Zero-Downtime Set-up for Models: How and Why

***Anouk Dutrée,*** *Product Owner,* ***UbiOps***

1: 00 PM

## Deep Dive: Critical Use of MLOps in Finance: Using Cloud-managed ML Services that Scale

*Vinnie Saini, Director, Enterprise Architecture (Data, Analytics & Cloud Strategy,* ***Scotiabank***

2: 00 PM

## Workshop With Ocado

2: 00 PM

## Building Real-Time ML Features with Feast, Spark, Redis, and Kafka

*Danny Chiao, Software Engineer, Achal Shah, Engineering Lead* ***Tecton / Feast***

2: 00 PM

## Workshop With Genesis

3: 00 PM

## Deep Dive: Lessons Learned from DAG-based Workflow Orchestration

**Kevin Gregory Kho**

4: 00 PM

## Defending Against Decision Degradation with Full-Spectrum Model Monitoring : Case Study and AMA

*Mihir Mathur, Machine Learning Product Lead,* ***Lyft***

4: 00 PM

## Workshop: A Guide to Building a Continuous MLOps Stack

*Itay Ben Haim, ML Engineer,* ***Superwise***

4:00 PM

## Workshop TBA

5: 00 PM

## Deep Dive: Scale and Accelerate the Distributed Model Training in Kubernetes Cluster

*Jack Jin, Lead ML Infra Engineer,* ***Zoom***

3rd annual

# MLOps World Machine Learning in Production

www.MLOpsWorld.com

**SESSION TRACKS**

**For Ticket Holders REG HERE**

| | |
|---|---|
| Business & Strategy | Technical |
| Case Studies | Workshops |

**Free Public Tracks REG HERE**

Exhibitior Demo Presentations & Speakers Corner

## June 9th: In Person, Toronto Canada

| | | | | | | |
|---|---|---|---|---|---|---|
| 9:30 am | **Opening Welcome:** David Scharbach, Executive Director, ***MLOps World*** | | | | | |
| 9:45 am | 1st Keynote | | | | | |
| | 10.15 am Break | | | | | |
| 10:40 am | **Building Real-Time ML Features with a Feature Platform**<br>Mike Del Balso, CEO & Co-Founder<br>Willem Pienaar, **Feast** Committer and Tech Lead<br>**Tecton** | **Understanding Foundation Models: a New Paradigm for Building and Productizing AI Systems**<br>Hagay Lupesko<br>Director of Engineering<br>**Meta AI** | **A GitOps Approach to Machine Learning**<br>Amy Bachir, Senior MLOps Engineer, Stephan Brown, MLOps Engineer<br>**Interos** | **Workshop:**<br>**Leaner, Greener and Faster Pytorch Inference with Quantization**<br>Suraj Subramanian Developer Advocate,<br>**PyTorch** | **Workshop:**<br>Hands-on : A Beginner-Friendly Crash Course to Kubernetes<br>Mohamed Sabri, Chief MLOps Officer, Eric Hammel, ML Engineer Asim Sultan<br>**Rocket Science** | |
| 11:25am | **Don't Fear Compliance Requirements & Audits: Implementing SecMLOps at Every Stage of the Pipeline**<br>Ganesh Nagarathnam Head of Machine Learning Engineering & Analytics<br>**S&P Global** | **One Cluster to Rule Them All - ML on the Cloud Using Ray on Kubernetes and AWS**<br>Victor Yap<br>MLOps Engineer<br>**Rev.com** | **How MLOps Tools Will Need to Adapt to Responsible and Ethical AI: Stay Ahead of the Curve**<br>Patricia Thaine Co-Founder & CEO<br>**Private AI** | | | |
| | 12:30 pm Lunch Break \| Exhibition \| Brain-Dates via Whova App | | | | | |
| 1:00 pm | **Panel: Embedding Diversity and Fairness Into Your Model Governance**<br>Andrea Olgiati Chief Engineer, **AWS SageMaker**<br>Arthur Berrill,Head of Location IntelligenceHead of Location Intelligence, **RBC** | **Solving MLOps From First Principles**<br>Dean Pleban Co-Founder & CEO<br>**DagsHub** | **Low-latency Neural Network Inference for ML Ranking Applications**<br>Ryan Irwin Engineering Manager Rajvinder Singh<br>**Yelp Inc.** | **Workshop:**<br>**Building Production ML Monitoring from Scratch: Live Coding Session**<br>Alon Gubkin CTO<br>**Aporia** | **Workshop:**<br>Hands-on : A Beginner-Friendly Crashcourse to Kubernetes<br>Mohamed Sabri, Chief MLOps Officer, Eric Hammel, ML Engineer Asim Sultan<br>**Rocket Science** | |
| 1:45 pm | **The Key pillars of ML Observability and How to Apply them to *Your* ML Systems**<br>Aparna Dhinakaran, Chief Product Officer, **Arize AI**<br>Gandalf Hernandez, Senior Machine Learning Engineering Manager, **Spotify** | **Eliminating AI Risk, One Model Failure at a Time**<br>Yaron Singer CEO and Co-Founder<br>**Robust Intelligence** | ***Advanced Technical***<br>**Company** | | | |
| | 2:30 pm Break \| Exhibition \| Brain-Dates via Whova App | | | | | |
| 3:00 pm | **Top 5 Lessons Learned in Helping Organizations Adopt MLOps Practices**<br>Shelbee Eigenbrode Principal AI/ML Specialist Solutions Architect<br>**Amazon Web Services** | **Scotiabank's Path Towards Accelerated Analytics Through GCP**<br>Shimona Narang, Vipul Upadhye, Data Scientists<br>**Scotiabank** | **Robustness and Security for AI and the Dangerous Dismissal of Edge Cases**<br>James Stewart CEO<br>**TrojAI Inc.** | **Workshop:**<br>**MLOps Delivery Platform: Transform Your Use Cases from Concept to Production**<br>Ryan Gillard, Cloud Machine Learning Elvin Zhu, AI Engineer<br>**Google** | **Workshop:**<br>**Introduction to Model Deployment with Ray Serve**<br>Jules Damji, Lead Developer Advocate<br>Archit Kulkarni<br>**Anyscale Inc.** | |
| 3:45 pm | ***Advanced Technical***<br>**Company** | **Tips on a Successful MLOps Adoption Strategy: A DoorDash Case Study**<br>Hien Luu, Sr. Engineering Manager at DoorDash<br>**DoorDash** | **A Framework for a Successful Continuous Training Strategy**<br>Or Itzary, Chief Architect<br>**Superwise** | | | |

### Exhibition Floor

**10:15 AM**
Speakers Corner Presentation & Demo

**11:00 AM**
Speakers Corner Presentation & Demo

**11:45 AM**
Speakers Corner Presentation & Demo

**12:30 PM**
Speakers Corner Presentation & Demo

**1:15 PM**
Speakers Corner Presentation & Demo

**2:30 PM**
Speakers Corner Presentation & Demo

**3:15 PM**
Speakers Corner Presentation & Demo

**4:00 PM**
Speakers Corner Presentation & Demo

## Social Events

| | |
|---|---|
| | Day 1 Program Ends Networking |
| 4:30 pm | Dinner Break & Networking |
| 5:00 - 6:00 pm | Start-up Expo , Makers Fest, Career Fair<br>25+ Cocktail tables/Presentations |
| 6:00 - 8:00 pm | |
| 8:00 - 10:00 pm | Day 1 Wrap- up & Evening Party<br>Location to be announced |

3rd annual

# MLOps World Machine Learning in Production

www.MLOpsWorld.com

**SESSION TRACKS**

**Tracks For Ticket Holders; REG HERE**

| | |
|---|---|
| Business & Strategy | Technical |
| Case Studies | Workshops |

**Free Session for Public; REG HERE**

Exhibitior Demo Presentations & Speakers Corner

## June 10th: In Person, Toronto Canada

**9:30 am** — **Day 2 Recap:** David Scharbach, Executive Director, ***MLOps World***

**9:45 am** — 1st Keynote

10.15 am Break

**10:45 am**

- ***A Guide for Start-ups; How to Scale a PoC to Production System and Not Go Up in Smoke***
  Maia Brenner, AI Specialist
  **Tryolabs**
- **MLOps at Rovio for Personalization (Self-service Reinforcement Learning in Production)**
  Ignacio Amaya de la Peña, Lead ML Engineer
  **Rovio**
- **CyclOps - A Framework for Data Extraction, Model Evaluation and Drift Detection for Clinical Use-cases**
  Amrit Krishnan,Senior Applied ML Specialist
  **Vector Institute,**
  Vallijah Subasri, Researcher at **The Hospital for Sick Children**
- **Workshop:**
  **Concretes Guidelines to Improve ML Model Quality, Based on Future ISO Certifications**
  Olivier Blais
  VP Decision Science
  **Moov AI**
- **Workshop:**
  **Feature Engineering Made Simple Part 1**
  Anindya Datta,
  Founder & CEO
  **Mobilewalla**

**11:25am**

- ***Monitoring AI Fairness with Facebook & Borealis AI?***
  **Company**
- **Shopify's ML Platform Journey Using Open Source Tools. Case study building Merlin**
  Isaac Vidas, Machine Learning Platform Tech Lead
  **Shopify**
- **MLOps for Deep Learning**
  Diego Klabjan, Professor
  Yegna Jambunath, MLOps Researcher
  **Northwestern University**

12.15 pm Lunch Break | Exhibition | Brain-Dates

**1:00 pm**

- ***Managing a Data Science Team During the Great Resignation***
  Jessie Lamontagne
  Data Science Manager,
  **Kinaxis**
- **Hold Spot:**
  **MLOps at Proctor & Gamble**
  Michele Floris,Data Scientist
  **Procter & Gamble**
- ***Advanced Technical***
  **Company**
- ***Women in AI Session***
- **Workshop: How to Continuously Improve ML Models, the Data-centric Way**
  Bernease Herman, Data Scientist
  **WhyLabs**

**1:45 pm**

- **The Critical Things you HAVE to Build to Transform Your Company to be ML Driven**
  Yuval Fernbach, Co-Founder & CTO
  **QWAK**
- **Loblaws Case Study: Overcoming Challenges Using Vertex AI**
  Mefta Sadat, Senior ML Software Engineer
  **Loblaw Digital**
- µlearn: a Microframework for Building Machine Learning Applications
  Niels Bantilan
  ML Engineer at **Union**, Core Maintainer , **Flyte**

2.30 pm Break | Exhibition | Brain-Dates

**3:00 pm**

- **Preventing Stale Models in Production**
  Milecia McGregor
  Developer Advocate
  **Iterative.ai**
- **DataOps For the Modern Computer Vision Stack**
  James Le
  Data Advocate
  **Superb AI**
- **SLA-Aware Machine Learning Inference Serving on Serverless Computing Platforms**
  Nima Mahmoudi,Data Scientist,
  TELUS Communications Inc.
- ***In person Workshop***
- ***In person Workshop***

**3:45 pm**

- **Strategy & Policy Deep Dive**
  **Company**
- **Machine Learning Infrastructure at Meta Scale**
  Shivam Bharuka, Senior AI Infra Engineer
  **Facebook (Meta)**
- ***Advanced Technical***
  **Company**

## Exhibition Floor

**10:15 AM** Speakers Corner Presentation & Demo

**11:00 AM** Speakers Corner Presentation & Demo

**11:45 AM** Speakers Corner Presentation & Demo

**12:30 PM** Speakers Corner Presentation & Demo

**1:15 PM** Speakers Corner Presentation & Demo

**2:30 PM** Speakers Corner Presentation & Demo

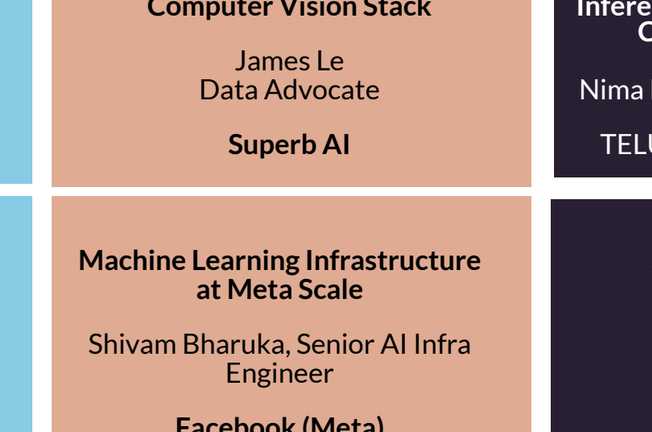
**3:15 PM** Speakers Corner Presentation & Demo

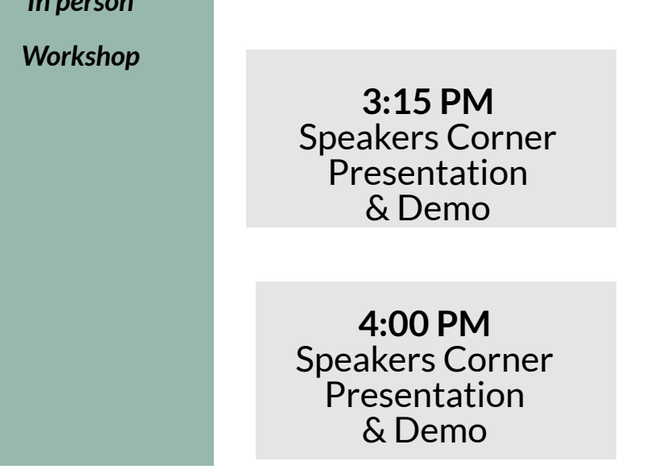
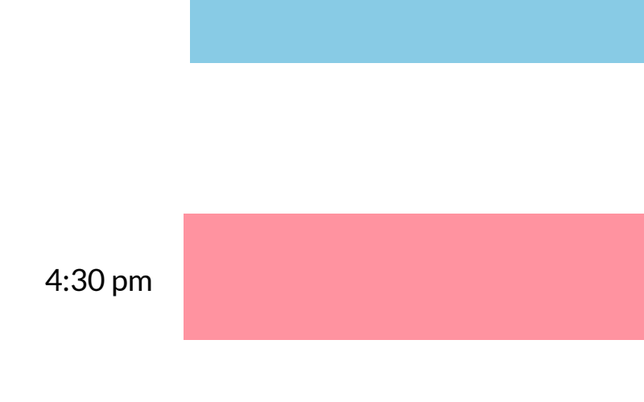
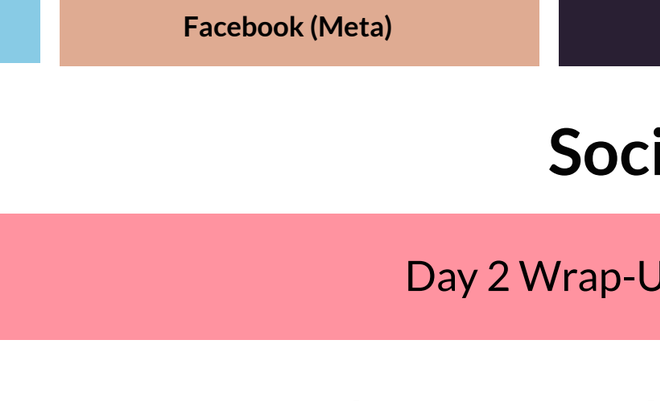
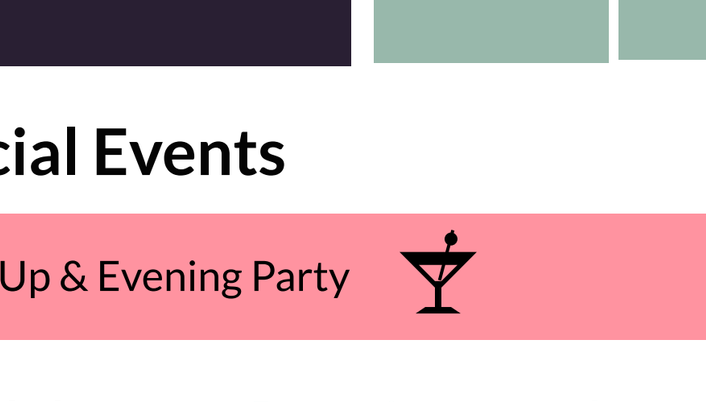
**4:00 PM** Speakers Corner Presentation & Demo

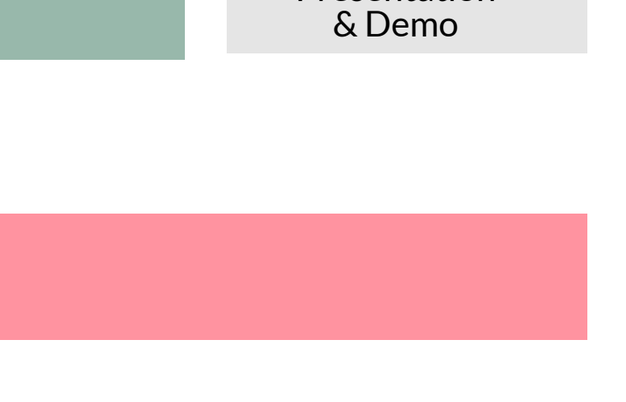
## Social Events

**4:30 pm** — Day 2 Wrap-Up & Evening Party

www.mlopsworld.com Register here



# JUNE 7th WORKSHOPS

**10: 00 AM - 11:00 AM EST**

**Machine Learning Monitoring in Production: Lessons Learned from 30+ Use Cases,** *Lina Weichbrodt, Lead Machine Learning Engineer,* ***DKB Bank***

**Abstract:**
Traditional software monitoring best practices are not enough to detect problems with machine learning stacks. How can you detect issues and be alerted in real-time?

This talk will give you a practical guide on how to do machine learning monitoring: which metrics should you implement and in which order? Can you use your team's existing monitoring and dashboard tools, or do you need an MLOps Platform?

**Technical Level: 5/7**

**What you will learn:**
- Monitor the four golden signals + add machine learning monitoring
- For ml monitoring prioritize monitoring the response of a service
- You often don’t need a new tool, use the tools you already have and add a few metrics

**10: 00 AM - 12:30 PM EST**

**Implementing MLOps Practices on AWS using Amazon SageMaker,** *Shelbee Eigenbrode, Principal AI/ML Specialist Solutions Architect / Bobby Lindsey, AI/ML Specialist Solutions Architect / Kirit Thadaka, ML Solutions Architect,* ***Amazon Web Services (AWS)***

**Abstract:**
In this workshop, attendees will get hands-on with SageMaker Pipelines to implement ML pipelines that incorporate CI/CD practices.

**Technical Level: 5/7**

**What you will learn:**
Familiarity with end-to-end features of Amazon SageMaker used in implementing ML pipelines

**What is unique about this talk:**
The opportunity to get hands-on

**11: 00 AM - 2:00 PM EST**

**Automated Machine Learning & Tuning with FLAML,** *Qingyun Wu, Assistant Professor,* ***Penn State University,*** *and Chi Wang, Principal Researcher,* ***Microsoft Research***

**Abstract:**
In this tutorial, we will provide an in-depth and hands-on training on Automated Machine Learning & Tuning with a fast python library FLAML. FLAML finds accurate machine learning models automatically, efficiently and economically. It frees users from selecting learners and hyperparameters for each learner.

**Technical Level: 4/7**

**What you will learn:**
1. How to use FLAML to find accurate ML models with low computational resources for common ML tasks.
2. How to leverage the flexible and rich customization choices provided in FLAML to customize your AutoML or tuning tasks.

**What is unique about this talk:**
In addition to a set of hands-on examples, the speakers will also share some rule-of-thumbs, pitfalls, open problems, and challenges learned from AutoML practice.

**1: 00 PM - 1:40 PM EST**

**Production ML for Mission-Critical Applications,** *Robert Crowe, TensorFlow Developer Engineer,* ***Google***

**Abstract:**
Deploying advanced Machine Learning technology to serve customers and/or business needs requires a rigorous approach and production-ready systems. This is especially true for maintaining and improving model performance over the lifetime of a production application. Unfortunately, the issues involved and approaches available are often poorly understood. An ML application in production must address all of the issues of modern software development methodology, as well as issues unique to ML and data science. Often ML applications are developed using tools and systems which suffer from inherent limitations in testability, scalability across clusters, training/serving skew, and the modularity and reusability of components. In addition, ML application measurement often emphasizes top level metrics, leading to issues in model fairness as well as predictive performance across user segments.

Rigorous analysis of model performance at a deep level, including edge and corner cases is a key requirement of mission-critical applications. Measuring and understanding model sensitivity is also part of any rigorous model development process.

We discuss the use of ML pipeline architectures for implementing production ML applications, and in particular we review Google’s experience with TensorFlow Extended (TFX), as well as available tooling for rigorous analysis of model performance and sensitivity. Google uses TFX for large scale ML applications, and offers an open-source version to the community. TFX scales to very large training sets and very high request volumes, and enables strong software methodology including testability, hot versioning, and deep performance analysis.

**Technical Level: 5/7**

**What you will learn:**
- How Production ML is fundamentally different from Research or Academic ML
- Methods and architectures for creating an MLOps infrastructure that adapts to change
- Review of several approaches to implementing MLOps in production settings

**What are some of the infrastructure you plan to discuss?**
TFX, Kubernetes, Kubeflow, Apache Beam, TensorFlow

**What is unique about this talk:**
Probably just having it all pulled together and put into context. You can find nearly anything online.

**2: 00 PM - 3:00 PM EST**

**Implementing a Parallel MLOps Test Pipeline for Open Source Development,** *Miguel Gonzalez-Fierro, Principal Data Science Manager,* ***Microsoft***

**Abstract:**
GitHub has become a hugely popular service for building software, open source or private. As part of the continuous development and integration process, frequent, reliable and efficient testing of repository code is necessary. GitHub provides functionality and resources for automating testing workflows (GitHub Workflows), which allow for both managed and self-hosted test machines.

However, managed hosts are of computational size that is limited for many machine learning workloads. Moreover, they don't include GPU hosts currently. As for self-hosted machines, there is the inconvenience and cost of keeping machines online 24 x 7. Another issue is that it is cumbersome to distribute test jobs to multiple machines.

Our goal is to leverage Azure Machine Learning along with GitHub Workflows in order to address these issues. With AzureML, we can access powerful compute with both CPU and GPU. Bringing the compute online is automatic and on demand for all the testing jobs. Moreover, we can easily distribute testing jobs to multiple hosts, in order to limit the end-to-end execution time of the workflow.

We show a configuration for achieving the above programmatically, which we have developed as part of the Microsoft Recommenders repository (https://github.com/microsoft/recommenders/), which is a popular open-source repository that we maintain and develop. In our setting, we have three workflows that trigger nightly runs as well as a workflow triggered by pull requests.

Nightly workflows, in particular, include smoke and integration tests and are long (more than 6 hours) if run sequentially. Using our parallelized approach on AzureML, we have managed to bring the end-to-end time down to less than 1 hour. We also discuss how to divide the tests into groups in order to maximize machine utilization.

We also talk about how we retrieve the logs associated with runs from AzureML and register them as artifacts on GitHub. This allows one to view the progress of testing jobs from the GitHub Actions dashboard, which makes monitoring and debugging of errors easier.

**Technical Level: 5/7**

**What you will learn:**

People who attend this session will learn about:
- Best practices on testing GitHub repositories of Python code, which are based on our experience with the Microsoft/Recommenders repository
- Guidelines on testing in economical ways
- How to use GitHub workflows for setting up their testing pipelines
- How to benefit from Azure Machine Learning capabilities in order to automate testing jobs that run in parallel.

**What are some of the infrastructure you plan to discuss?** GitHub and AzureML

**What is unique about this talk:**
We have one of the most sophisticated test pipelines of GitHub repositories related to machine learning

**4: 00 AM - 4:30 PM EST**

**How to Treat Your Data Platform Like a Product: 5 Key Best Practices,** *Barr Moses, CEO & Co-Founder,* ***Monte Carlo***

**Abstract:**
Your team just migrated to a data mesh (or so they think). Your CTO is all in on this "modern data stack," or as he calls it: "The Enterprise Data Discovery." To satisfy your company's insatiable appetite for data, you may even be building a complex, multi-layered data ecosystem: in other words, a data platform. Still, it's one thing to build a data platform, but how do you ensure it actually drives value for your business?

In this fireside chat, Barr Moses, CEO & co-founder of Monte Carlo, will walk through why best in class data teams are treating their data platforms like product software and how to get started with reliability and scale in mind.

**Technical Level: 3/7**

**What you will learn:**
5 best practices (across technology, processes, and culture) for treating your data platform like a scalable, measurable product with machine learning and automation.

**What are some of the infrastructure you plan to discuss?** Cloud data warehouses, data lakes, reverse ETL, data observability, dbt

**What is unique about this talk:**
I've never discussed these best practices before at a public talk or in a blog article, and they're pulled from my own experience at Monte Carlo working with 100s of data teams attempt to build their own data platforms.

**4: 00 PM - 5:00 PM EST**

**WarpDrive: Orders-of-Magnitude Faster Multi-Agent Deep RL on a GPU,** *Stephan Zheng, Lead Research Scientist; Tian Lan, Senior Applied Scientist; Sunil Srinivasa, Research Engineer,* ***Salesforce***

**Abstract:**
Reinforcement learning is a powerful tool that has enabled big technical successes in AI, including superhuman gameplay, optimizing data center cooling, nuclear fusion control, economic policy analysis, etc. For wider real-world deployment, users need to be able to run RL workflows efficiently and quickly. WarpDrive is an open-source framework that runs multi-agent deep RL end-to-end on a GPU. This enables orders of magnitude faster RL.

In this talk, we will review how WarpDrive works and several new features introduced since its first release in Sep 2021. These include automatic GPU utilization tuning, distributed training on multiple GPUs, and sharing multiple GPU blocks across a simulation. These features result in throughput scaling linearly with the number of devices, to a scale of millions of agents. WarpDrive also provides several utility functions that improve quality-of-life and enable users to quickly implement and train RL workflows.

**Technical Level: 6/7**

**What you will learn:**
How WarpDrive enables you to run reinforcement learning orders of magnitude faster.

**What is unique about this talk:**
Accessible explanations of the latest features, demos, and future roadmap.

**5: 00 PM - 5:40 PM EST**

**Scaling ML Embedding Models to Serve a Billion Queries,** *Senthilkumar Gopal, Senior Engineering Manager (Search ML),* ***eBay Inc.***

**Abstract:**
This talk is aimed at providing a deeper insight into the scale, challenges and solutions formulated for powering embeddings based visual search in eBay. This talk walks the audience through the model architecture, application archite for serving the users, the workflow pipelines produced for building the embeddings to be used by Cassini, eBay's search engine and the unique challenges faced during this journey. This talk provides key insights specific to embedding handling and how to scale systems to provide real time clustering based solutions for users.

**Technical Level: 5/7**

**What you will learn:**
The audience will learn how to productionize embedding based data pipelines, key challenges and potential solutions, introduction to different quantization algorithms and their advantages/disadvantages. The audience will also get a deeper view on how data pipelines and workflows are modeled for optimal scale.

**What are some of the infrastructure you plan to discuss?** Hadoop, Airflow, Pyspark, Kafka, Flink, Pytorch

**What is unique about this talk:**
Most of the online content, dwells on pieces of the infrastructure required without providing an end to end coherent picture. Most critically, the content does not relate to the model architecture and how the pipelines and model architecture/parameters are influenced by each other. This talk also goes into the aspects of a large scale search engine and how the application architecture influences the operational aspects to enable the scale required.

# JUNE 8th WORKSHOPS

**10: 00 AM - 2:00 PM EST**

**Personalized Recommendations and Search with Retrieval and Ranking at scale on Hopsworks,** *Jim Dowling, CEO,* ***Hopsworks***

**Abstract:**
Personalized recommendations and personalized search systems at scale are increasingly being built on retrieval and ranking architectures based on the two-tower embedding model. This architecture requires a lot of infrastructure. A single user query will cause a large fanout of traffic to the backend, with hundreds of database lookups in a feature store, similarity search in an embedding store, and model outputs from both a query embedding model and a ranking model. You will also need to index your items in the embedding store using an item embedding model, and instrument your existing systems to store observations of user queries and the items they select.

**Technical Level: 6/7**

**What you will learn:**
How to build a state-of-the-art two tower model for personalized recommendations that scales with Hopsworks.

**What is unique about this talk:**
The only integrated open-source platform for scalable retrieval and ranking systems.

**10: 00 AM - 11:30 AM EST**

**Accelerating Transformers with Hugging Face Optimum and Infinity,** *Philipp Schmid, Machine Learning Engineer and Lewis Tunstall, Machine Learning Engineer,* ***Hugging Face***

**Abstract:**
Since their introduction in 2017, Transformers have become the de facto standard for tackling a wide range of NLP tasks in both academia and industry. However, in many situations accuracy is not enough — your state-of-the-art model is not very useful if it's too slow or large to meet the business requirements of your application.

**Technical Level: 5/7**

**What you will learn:**
How Hugging Face Optimum and Infinity provide developers with the tools to easily optimise Transformers with techniques such as quantization and pruning.

**12: 00 PM - 1:30 PM EST**

**What's in the box? Automatic ML Model Containerization,** *Clayton Davis*, Head of Data Science, ***Modzy***

**Abstract:**
This talk will include a deep dive on building machine learning (ML) models into container images to run in production for inference. Based on our experience setting up ML container builds for many customers, we'll share a set of best practices for ensuring secure, multi-tenant image builds that avoid lock-in, and we'll also cover some tooling (chassis.ml) and a standard (Open Model Interface (OMI)) to execute this process. Data scientists and developers will walk away with an understanding of the merits of a standard container specification that allows for interoperability, portability, and security for models to seamlessly be integrated into production applications.

**Technical Level: 5/7**

**What you will learn:**
Prerequiste: Basic familiarity with ML models and/or common ML frameworks (pytorch, scikit learn, etc.)

**12: 00 PM - 1:00 PM EST**

**A Zero-Downtime Set-up for Model: How and Why,** *Anouk Dutrée, Product Owner,* ***UbiOps***

**Abstract:**
When a model is in production you ideally want zero-downtime. Whenever the model is needed it should be ready to respond. This issue is two-sided, on one hand you need to make sure that there is no down-time when updating your model, on the other hand you need to ensure that a request can be processed even if your model itself fails. In this talk we will take you through the set-up we use to ensure zero-downtime when updating models, and how this set-up can be expanded to ensure you can handle failing models as well.

**Technical Level: 4/7**

**What you will learn:**
- How to create an easy to work with zero-downtime set-up for data science models using smart routing
- How to expand this set-up to a champion challenger set-up to ensure there is always a model available that can take over if a model fails unexpectedly
- What a champion challenger set-up is

**What are some of the infrastructure you plan to discuss?** Deployment and Serving infrastructure

**What is unique about this talk:**
I personally find most of the articles on this topic to be to specific to one part of the chain. In this talk I want to go over the entire process as a whole, and cover the two sides of downtime. (i.e. downtime caused by maintenance and downtime because the model fails)

**1: 00 PM - 1:40 PM EST**

**Critical Use of MLOps in Finance: Using Cloud-managed ML Services that Scale,** *Vinnie Saini, Director and Senior Principal , Enterprise Data Architecture & Cloud Strategy,* ***Scotiabank***

**Abstract:**
With ML Engineering being a superset of Software Engineering, treating Data as a first class citizen is key to ML Engineering.
The talk will be focused on how leveraging MLOps is a key to improve the quality and consistency of machine learning solutions, managing the lifecycle of your models with the goal of:
- Faster experimentation and development of models
- Faster deployment of models into production
- Quality assurance and end-to-end lineage tracking

With trained machine learning models deployed as web services in the cloud or locally, we'll see how deployments use CPU, GPU, or field-programmable gate arrays (FPGA) for inferencing- using different compute targets:
- Container Instance
- Kubernetes Service
- development environment

**Technical Level: 4/7**

**What you will learn:**
This talk is intended for technology leaders and enterprise architects who want to understand the details about what MLOps in practice:
- Capture the governance data for the end-to-end ML lifecycle.
- Monitor ML applications for operational and ML-related issues. Compare model inputs between training and inference, explore model-specific metrics, and provide monitoring and alerts on your ML infrastructure.
- Automate the end-to-end ML lifecycle with Pipeline to continuously roll out new ML models alongside your other applications and services

**What is unique about this talk:**
The session will provide a deeper dive into the themes of scale and automate to illustrate the requirements for building and operationalizing ML systems. This will include a special emphasis on using cloud managed ML services that scale with large amounts of data and large numbers of data processing and ML jobs, with reduced operational overhead.

**3:00 PM EST**

**Lessons Learned from DAG-based Workflow Orchestration,** *Kevin Kho, Senior Open Source Community Engineer,* ***Scotiabank***

**Abstract:**
Workflow orchestration has traditionally been closely coupled to the concept of Directed Acyclic Graphs (DAGs). Building data pipelines involved registering a static graph containing all the tasks and their respective dependencies. During workflow execution, this graph would be traversed and executed. The orchestration engine would then be responsible for determining which tasks to trigger based on the success and failure of upstream tasks.

**Technical Level: 4/7**

**What you will learn:**
They will learn about workflow orchestration, and why pinning it to the Directed Acylic Graph concept proved to be limiting. They will learn how to spin up their own free open-source orchestrator.

**What are some of the infrastructure you plan to discuss?** Docker/Kubernetes

**What is unique about this talk:**
A lot of the content here will come from supporting the Prefect community over the last 3 years and the difficulties we recognized with traditional orchestration systems. There are not a lot of people with the experience of supporting thousands of use cases and extracting insight from that.

**4:00 PM - 4:40 PM EST**

**Defending Against Decision Degradation with Full-Spectrum Model Monitoring : Case Study and AMA** , *Mihir Mathur, Product Manager, Machine Learning,* ***Lyft***

**Abstract:**
ML models at Lyft make millions of high stakes decisions per day including decisions for real-time pricing, physical safety classification, fraud detection, and much more. Preventing models from degrading and making ineffective decisions is therefore critical. Over the past two years, we've invested in building a full-spectrum model monitoring solution to catch and prevent model degradation. In this talk, we'll discuss our suite of approaches for model monitoring including real-time feature validation, performance drift detection, anomaly detection, and model score monitoring as well as the cultural change needed to get ML practitioners to effectively monitor their models. We'll also discuss the impact our monitoring system delivered by catching problems.

**Technical Level: 4/7**

**What you will learn:**
Many different aspects of Model Monitoring:
Why it's needed
Challenges in building a model monitoring system
How to prioritize among a plethora of things that can be built
Overview of Lyft's model monitoring architecture
How to cause cultural change at a company for better AI/ML practices

**What are some of the infrastructure you plan to discuss?** Technical architecture for realtime feature validation, performance drift detection, anomaly detection, and model score monitoring

**What is unique about this talk:**
Often, great MLOps tools themselves don't lead to business impact. What's needed is prioritization and cultural change to ensure that tools are effectively used.

**4:00 PM - 5:30 PM EST**

**A Guide to Building a Continuous MLOps Stack,** *Itay Ben Haim, Machine Learning Engineer,* ***Superwise***

**Abstract:**
In this workshop, we'll take a dive into MLOps CI/CD pipeline automation with GCP, Superwise, and retraining/auto-resolution notebooks. In part 1, we'll focus on how to put together a continuous ML pipeline to train, deploy, and monitor models. Part 2 will focus on automations and production-first insights to detect and resolve issues continuously.

**Technical Level: 3/7**

**What you will learn:**
How to build a continuous MLOps stack Platform and tool alternatives for each step Considerations for scaling up Production-first insights and automations

**What is unique about this talk:**
This is a practical live coding session together with the participants that shows how to implement MLOps level two

# JUNE 9th TALKS

**10: 45 AM - 11:25 PM EST**

**Building Real-Time ML Features with a Feature Platform,** *Mike del Balso, Co-Founder & CEO and Willem Pienaar, Feast Committer and Tech Lead,* ***Tecton***

**Abstract:**
Deploying ML in production is hard, and data is often the hardest part. Production ML pipelines are different from traditional analytics pipelines. They need to process both historical data for training, and fresh data for online serving, often using streaming or real-time data sources. They must ensure training/serving parity, provide point-in-time correctness, and serve data with production service levels. These challenges are difficult to tackle with traditional ETL tools, and can often add weeks or months to project timelines.

In this session, Mike Del Balso and Willem Pienaar will present the challenges faced when building the core ML infrastructure at Uber and Gojek faced, and how their teams built feature stores to scale their ML efforts to thousands of models in production. Uber and Gojek used these internal ML platforms to power every aspect of their business: ride ETAs, demand forecasting, pricing, and restaurant recommendations.

Feature stores have now emerged as the tool of choice to solve the challenges of production ML. At their core, they provide a simple solution to store, serve and share features. However, feature stores are not enough. Teams still need to create bespoke data pipelines to process raw data into features in real-time. To solve the data problem for ML, organizations need a complete feature platform, which extends a feature store to include automated ML data pipelines that can transform data from batch and real-time sources. Mike and Willem will share their views on the evolution of feature stores to feature platforms that can manage the complete lifecycle of real-time ML features.

**Technical Level: 4/7**

**What you will learn:**
In this session, attendees will learn about the data challenges faced by ML teams at Uber and Gojek, and how they were solved with feature stores. Attendees will also get a hands-on example of how a feature platform can be used to build and operationalize enterprise-grade feature pipelines for a fraud detection use case.

We'll show how to:
- Define features as code
- Transform data and materialize feature values
- Store values in offline and online store
- Serve data for training
- Serve data online for real-time inference
- Monitor pipeline health, data drift, and online service levels

**What are some of the infrastructure you plan to discuss?** SQL and Python

**What is unique about this talk:**
Mike and Willem are early pioneers of the feature store category, having respectively created Uber Michelangelo and the Feast open source feature store. They are uniquely positioned to discuss the challenges that operational ML teams face, and the path to overcoming them. And, for the past few years, they've been working with hundreds of organizations on solving similar problems in their respective capacities as co-founder of Tecton and creator of Feast. They bring a unique depth of perspective to the discussion. The notion of extending feature stores with a complete feature platform is new and hasn't been talked about online before. Attendees will learn about the evolution of feature stores and how a feature platform can solve the challenges associated with the complete lifecycle of ML features.

**10: 45 AM - 11:25 PM EST**

**Understanding Foundation Models: a New Paradigm for Building and Productizing AI Systems,** *Hagay Lupesko, Director of Engineering,* ***Meta AI***

**Abstract:**
The term Foundation Models was coined in a 2021 technical report published by dozens of Stanford researchers, describing Foundation Models as no less than a new paradigm for building AI systems. In this session we will unpack this bold concept and identify practical ways for companies to start leveraging foundation models today.

**Technical Level: 4/7**

**What you will learn:**
- What are Foundation Models
- Real world examples of Foundation Models
- How companies can leverage Foundation Models today

**What is unique about this talk:**
The combination of a new paradigm with real world examples of how this is leveraged today

**10: 45 AM - 11:25 PM EST**

**A GitOps Approach to Machine Learning,** *Amy Bachir, Senior MLOps Engineer and Stephan Brown, MLOps Engineer,* ***Interos***

**Abstract:**
The focus of this talk is the application of GitOps principles to machine learning in production. At Interos we use GitOps for most of our MLOps work, storing our ML configurations as code. GitOps has many benefits including traceability, stability, reliability, consistency, enhanced productivity, and provides a single source of truth. We apply GitOps to our deployment configurations, onboarding process, monitoring configurations, and use it all stages of the model lifecycle. The portable and declarative nature of GitOps has led to increased traceability, and as a small team has increased our development capacity.

**Technical Level: 5/7**

**What you will learn:**

**What are some of the infrastructure you plan to discuss?** SQL and Python

**What is unique about this talk:**
Mike and Willem are early pioneers of the feature store category, having respectively created Uber Michelangelo and the Feast open source feature store. They are uniquely positioned to discuss the challenges that operational ML teams face, and the path to overcoming them. And, for the past few years, they've been working with hundreds of organizations on solving similar problems in their respective capacities as co-founder of Tecton and creator of Feast. They bring a unique depth of perspective to the discussion. The notion of extending feature stores with a complete feature platform is new and hasn't been talked about online before. Attendees will learn about the evolution of feature stores and how a feature platform can solve the challenges associated with the complete lifecycle of ML features.

**10: 45 AM - 12:15 PM EST**

**Hands-on : A Beginner-Friendly Crashcourse to Kubernetes,** *Eric Hammel, ML Engineer, Mohamed Sabri, Senior Consultant in MLOps, and Asim Sultan,* ***Rocket Science***

**Abstract:**
Have you ever wondered what kubernetes and Cloud Native applications are? Here is the perfect opportunity to get exposed to these complex yet powerful tools. You will discover concepts and tools such Container Orchestration, Cloud Native, Kubernetes and application deployment.

**Technical Level: 4/7**

**What you will learn:**
The participants will get a crash course about Kubernetes and Cloud Native concepts. They learn how to deploy an application on a managed kubernetes cluster thanks to the presented abstractions.

# JUNE 9th TALKS

**11: 30 AM - 12:15 PM EST**

**Don't Fear Compliance Requirements & Audits: Implementing SecMLOps at Every Stage of the Pipeline,** *Ganesh Nagarathnam,* ***S&P Global***

**Abstract:**
Over the last few years, MLOps - as a discipline has made significant inroads in operationalizing and democratizing ML for a variety of use cases spanning across industries. With so many tools being available for the conventional ML pipeline, organizations have stayed away from a 'swiss army knife 'kind of a tool mind set and made better choices in choosing the right tool stack for their problems in the pipeline. With this kind of proliferation, managing, building and monitoring security in the ML pipeline poses unique challenges. The speaker dissects the ML pipeline and applies core drivers for incorporating security at every stage and proposes an extension framework - namely SecMLOps

**Technical Level: 4/7**

**What you will learn:**
They will learn how to integrate security early in to the ML development process enabling them to come up with their own core set of drivers for SecMLOps as needed. Product Managers, Program Managers, Application security Managers in the organization will feel empowered to be a part of ML Development cycle. Compliance Requirements and audits will not be feared and it would rather be simplified when such a framework is in place !

**11: 30 AM - 12:15 PM EST**

**One Cluster to Rule Them All - ML on the Cloud Using Ray on Kubernetes and AWS,** *Victor Yap, MLOps Engineer,* ***Rev.com***

**Abstract:**
Distributed compute clusters (aka HPC) are fundamental to machine learning in order to scale data processing, model training, model serving and more. However, each of these areas require diverse compute resources, both in quantity (10s-1000s) and type (CPU/GPU/Memory). On top of all of that, data scientists can face significant friction when trying to run their experiments across their local environment and the cluster's environment. This talk will cover how to build a single cluster, on AWS with Ray and Kubernetes, that can dynamically scale any resource type to any quantity, bridge the gap between local and cluster environments, and describe how Rev uses it to handle any compute problem.

**Technical Level: 6/7**

**What you will learn:**
How to build a dynamic, scalable distributed cluster on AWS with Ray and Kubernetes. The audience will learn why distributed compute clusters are required in ML, how to implement one that dynamically creates instances, and how Rev uses one to run all distributed computing needs. The audience will learn about Ray, Kubernetes and AWS, and what they have to offer in the space of MLOps.

**11: 30 AM - 12:15 PM EST**

**How MLOps Tools Will Need to Adapt to Responsible and Ethical AI: Stay Ahead of the Curve,** *Patricia Thaine, Co-Founder & CEO,* ***Private AI***

**Abstract:**
We are at the dawn of a new age for responsible AI: there's a flourishing field studying its benefits and harms, and the EU is actively legislating AI ethics. But while MLOps platforms have grown in capability and complexity, their consideration of responsible/ethical AI have lagged significantly behind. In this talk, we'll dive into the ethical guardrails every MLOps solution should implement to be prepared for a fast-approaching future and into how they can help with GDPR compliance (data residency, data security, data privacy) as well as cater to future regulatory requirements.

**Technical Level: 4/7**

**What you will learn:**
How legislators are thinking about regulating AI and how the requirements fit into MLOps, including privacy and explainability.

**1: 00 PM - 1:40 PM EST**

**Solving MLOps From First Principles: A Framework to Reduce Complexity**, *Dean Pleban, Co-Founder & CEO,* ***DagsHub***

**Abstract:**
One of the hardest challenges data teams face today is selecting which tools to use in their workflow. Marketing messages are vague, and you continuously hear of new buzzwords you "just have to have in your stack". There is a constant stream of new tools, open-source and proprietary that make buyer's remorse especially bad. I call it "MLOps Fatigue". This talk will not discuss a specific MLOps tool, but instead present guidelines and mental models for how to think about the problems you and your team are facing, and how to select the best tools for the task. We will review a few example problems, analyze them, and suggest Open Source solutions for them. We will provide a mental framework that will help tackle future problems you might face and extract the concrete value each tool provides.

**Technical Level: 4/7**

**What you will learn:**
You'll learn what signals to watch for to notice you might have MLOps fatigue. How to define the challenge you're facing and which questions to ask in order to build a "decision tree" for selecting the best suited tools for the task. A few examples for using this framework in practice on challenges involving data management and automating training/pipeline tasks

**1: 00 PM - 1:40 PM EST**

**Low-latency Neural Network Inference for ML Ranking Applications: Yelp Case Study,** *Ryan Irwin, Engineering Manager and Rajvinder Singh,* ***Yelp, Inc.***

**Abstract:**
At Yelp, we train and deploy models for a variety of business applications requiring low-latency model inference. At first we focused on streamlining support for XGboost and LR models built in Spark to support business recommendations, search, ads, restaurants, and trust & safety use-cases. However, we didn't have a way of supporting low-latency neural network models with Tensorflow. Such models usually relied on batched model inference in support of models used for photo classification [1] and popular dishes [2].

**Technical Level: 6/7**

**What you will learn:**
The audience will learn how different technologies in MLOPs can be used to solve low-latency ranking problems.

**1: 00 PM - 1:40 PM EST**

**Panel: Embedding Diversity and Fairness Into Your Model Governance,** *Andrea Olgiati, Chief Engineer, AWS SageMaker,* ***Amazon Web Services***

**Abstract:**

**Technical Level: 4/7**

**What you will learn:**

**1: 00 PM - 2:30 PM EST**

**Building Production ML Monitoring from Scratch: Live Coding Session**, *Alon Gubkin, CTO,* ***Aporia***

**Abstract:**
In this session, together we will build a cloud native ML Monitoring stack using open-source tools. We'll start from explaining the basic principles of Machine Learning monitoring in production, and then create a live web dashboard to measure model drift (training compared to prediction), feature statistics, and performance metrics in production. This monitoring stack will also enable us to integrate Python-based custom metrics which will be displayed on the dashboard. The code will be available on GitHub after the workshop.

**Technical Level: 5/7**

**What you will learn:**
They'll learn how to build a cloud native ML Monitoring stack using open-source tools that will integrate into their MLOps platform.

**1: 00 PM - 2:30 PM EST**

**Hands-on : A Beginner-Friendly Crashcourse to Kubernetes,** *Eric Hammel, ML Engineer, Mohamed Sabri, Senior Consultant in MLOps, and Asim Sultan,* ***Rocket Science***

**Abstract:**
Have you ever wondered what kubernetes and Cloud Native applications are? Here is the perfect opportunity to get exposed to these complex yet powerful tools. You will discover concepts and tools such Container Orchestration, Cloud Native, Kubernetes and application deployment.

**Technical Level: 4/7**

**What you will learn:**
The participants will get a crash course about Kubernetes and Cloud Native concepts. They learn how to deploy an application on a managed kubernetes cluster thanks to the presented abstractions.

**1: 45 PM - 2:30 PM EST**

**The Key Pillars of ML Observability and How to Apply them to Your ML Systems,** *Aparna Dhinakaran, Chief Product Officer,* ***Arize AI,*** *and Gandalf Hernandez, Senior Machine Learning Engineering Manager,* ***Spotify***

**Abstract:**
"If you build it, they will come" is a totally bogus way to approach building an ML platform. All the time, teams learn the hard way that the details -- justifying the platform, identifying the key components that matter, how it fits into the broader whole, business impact, etc. -- are what determines success, not unnecessarily technical specifications or wasting time building a product that will only be irrelevant once it's done. It's about fundamentals, and getting those right is hard. Compared to DevOps or data engineering, MLOps is still relatively young as a discipline and best practices are often learned on the fly…so sometimes it pays to buy over build. In this session Gandalf Hernandez, Senior Machine Learning Manager at Spotify, and Aparna Dhinakaran – Chief Product Officer at Arize AI share best practices, war stories, and debate questions such as:

- How do you justify building an ML platform internally?
- What are the key components that matter to your team?
- And why is ML infrastructure necessarily distinct from software infrastructure?

**Technical Level: 4/7**

**What you will learn:**
The audience will gain insight into what goes on from behind the scenes when building an ML platform at scale. From the specific tools required for machine learning to the rationale behind build versus buy for MLOps tools, audience members can use this talk to help frame their evaluations of various tools or internal efforts to stand up ML infrastructure for their organizations. Audience members will learn common challenges and problems from real-world examples, and how engineers approached those challenges head-on.

**1: 45 PM - 2:30 PM EST**

**Eliminating AI Risk, One Model Failure at a Time**, *Yaron Singer, CEO & Co-Founder,* ***Robust Intelligence***

**Abstract:**
As organizations adopt AI they inherent AI risk. AI risk often manifests itself in AI models that produce erroneous predictions that go undetected and result in serious consequences for the organization and individuals affected by the decisions. In this talk we will discuss root causes for AI models going haywire, and present a rigorous framework for eliminating risk from AI. We will show how this methodology can be used as building blocks for building an AI firewall that can prevent and model AI model failures.

**Technical Level: 5/7**

**What you will learn:**
How to eliminate AI failure from their model pipelines.

**3: 00 PM - 3:40 PM EST**

**Top 5 Lessons Learned in Helping Organizations Adopt MLOps Practices,** *Shelbee Eigenbrode, Principal AI/ML Specialist Solutions Architect,* ***Amazon Web Services***

**Abstract:**
In this session, I'll cover the top 5 lessons learned in helping organizations implement MLOps practices at scale. Here you'll learn about some of the common challenges encountered as well as recommendations in how to mitigate those challenges.

**Technical Level: 2/7**

**What you will learn:**
In this session, the audience will learn pitfalls to avoid based on large scale adoption of MLOps as well as technical implementations.

**3: 00 PM - 3:40 PM EST**

**Scotiabank's Path Towards Accelerated Analytics Through GCP,** *Shimona Narang and Vipul Upadhye, Data Scientists,* ***Scotiabank***

**Abstract:**
Investing in data and analytics has been critical for financial institutions for years, but it has risen to the forefront during the pandemic as a critical tool for assisting customers during difficult times. Recently, Scotiabank has partnered with Google Cloud Platform to strengthen the bank's cloud-first strategy and accelerate its global data and analytics efforts. As part of this partnership, AIML team at International Banking, has been leveraging GCP for analytics experiments, model training, and for operationalizing them. With GCP in place, we achieved enhanced performance on bank operations in Peru Analytics by reducing bottlenecks between data science and engineering teams. Our architecture on GCP also facilitates governing ML artifacts to support auditability, traceability, and compliance.

In this talk, we will share our journey of onboarding our machine learning use case to GCP at Scotiabank.

**Technical Level: 5/7**

**What you will learn:**
Key takeaways:
- Scotiabank's architecture, MLOps lifecycle and development on Google Cloud
- Leveraging GCP's Vertex AI for model training and serving pattern
- Secure handling of production data on Google Cloud Platform
- Advanced customer analytics success stories in International Banking

**3: 00 PM - 3:40 PM EST**

**Robustness and Security for AI and the Dangerous Dismissal of Edge Cases,** *James Stewart, CEO,* ***TrojAI Inc.***

**Abstract:**
As we look to deploy AI to mission critical systems, it is no longer acceptable to dismiss AI limitations as edge cases. Doing so is akin to ignoring the robustness and security of models based on the fallacy that there are a very small finite number of edge cases that need to be addressed. The truth is, there are a near limitless number of edge cases and, paired with the probable improbable, AI systems across the world will encounter edge cases every single day. An edge case is something that will rarely occur in practice and can be both naturally occurring or malicious like an adversarial attack. It is tempting to dismiss edge cases as unlikely to happen again once addressed. Sometimes the situation is so obscure that even humans may have been confused except that most won't and those that are will typically deal with the confusion more gracefully than AI. Conversely, when an AI is confused by an edge case, all AI in the system will be confused. The problem of edge cases is amplified because we cannot predict model performance using traditional accuracy metrics like recall, precision and F1-Score, which do not translate well from the lab to the real world.

In this talk, we present examples of both naturally occurring and malicious edge cases and discuss possible strategies for avoiding the situation where a new model is more accurate but more brittle to failure. Robustness metrics provide insight into problem classes and model failure bias which can reduce risk by shaping models towards more benign failure cases. Regulations and significant penalties are emerging around Responsible AI requiring industry to articulate what could go wrong with models and what steps have been taken to mitigate the risks and ultimately protect the pace of innovation.

**Technical Level: 2/7**

**What you will learn:**
The limitations and risks of AI and the coming regulations for Responsible AI.

**3: 00 PM - 4:30 PM EST**

**MLOps Delivery Platform: Transform Your Use Cases from Concept to Production,** *Ryan Gillard, Cloud Machine Learning and Elvin Zhu, AI Engineer,* ***Google***

**Abstract:**
We present an end-to-end MLOps delivery platform on GCP for 6 different ML frameworks and custom accelerators to transform your ML use cases from concept to production. The platform offers 3 steps of MLOps starting with the "Dev" stage where we unit test the ML components required in your use case. This is followed by the "Test" stage where we build ML pipelines with Kubeflow Pipelines and Cloud Composer using the ML components built in the Dev stage. Finally the "Prod" stage deploys the ML components into production with model monitoring and CI/CD. We demonstrate several use cases where we use this platform to deliver ML applications. The platform is supported by code in GitHub with examples.

**Technical Level: 5/7**

**What you will learn:**
New platform for MLOps on Google Cloud

**3: 00 PM**
**-**
**4:30 PM EST**

**Workshop/Tutorial: Introduction to Model Deployment with Ray Serve,** *Jules Damji, Lead Developer Advocate and Archit Kulkarni, Software Engineer,* ***Anyscale Inc.***

**Abstract:**
This is a two-part introductory and hands-on guided tutorial of Ray and Ray Serve.

Part one covers a hands-on coding tour through the Ray core APIs, which provide powerful yet easy-to-use design patterns (tasks and actors) for implementing distributed systems in Python.

Building on the foundation of Ray Core APIs, part two of this tutorial focuses on Ray Serve concepts, what and why Ray Serve, scalable architecture, and model deployment patterns. Then, using code examples in Jupyter notebooks, we will take a coding tour of creating, exposing, and deploying models to Ray Serve using core deployment APIs.

And lastly, we will touch upon Ray Serve's integration with model registries such as MLflow, walk through an end-to-end example, and discuss and show Ray Serve's integration with FastAPI.

**Technical Level: 5/7**

**What you will learn:**
- Use Ray Core APIs to convert Python function/classes into a distributed setting
- Learn to use Ray Serve APIs to create, expose, and deploy models with Ray Server APIs
- Access and call deployment endpoints in Ray Serve via Python or HTTP
- Configure compute resources and replicas to scale models in production
- Learn about Ray Serve integrations with MLflow and FastAPI

**3: 45 PM**
**-**
**4:30 PM EST**

**Tips on a Successful MLOps Adoption Strategy: DoorDash Case Study,** *Hien Luu, Head of Machine Learning Platform,* ***DoorDash***

**Abstract:**

MLOps is one of the hottest topics being discussed in the ML practitioner community. Streamlining the ML development and productionalizing ML are important ingredients to realize the power of ML, however it requires a vast and complex infrastructure. The ROI of ML projects will start only when they are in production. The journey to implementing MLOps will be unique to each company. At DoorDash, we've been applying MLOps for a couple of years to support a diverse set of ML use cases and to perform large scale predictions at low latency. This session will share our approach to MLOps, as well as some of the learnings and challenges.

**Technical Level: 5/7**

**What you will learn:**
A strategy for adoption MLOps

# JUNE 10th TALKS

**10: 45 AM - 11:25 AM EST**

**MLOps at Rovio for Personalization (Self-Service Reinforcement Learning in Production),** *Ignacio Amaya dela Pena, Lead Machine Learning Engineer,* ***Rovio***

**Abstract:**
Rovio's game teams leverage Beacon, our internal cloud services platform which among other things enables them to leverage data to grow their games. Machine Learning is part of Beacon's offering. With a few clicks games can start using Reinforcement Learning models with "Personalized rules" which aim to replace complex sets of rules and heuristics that currently are still common across all industries.

**Technical Level: 6/7**

**What you will learn:**
From a business point of view, you will learn about the games personalization use case and how Rovio ML product offering helps growing the games. From a technical point of view, you will learn about the MLOps required to run Reinforcement Learning use cases in production (both contextual bandits and deep reinforcement learning) and what are the main challenges we faced

**10: 45 AM - 11:25 AM EST**

**A Guide for Start-ups; How to Scale a PoC to Production System and Not Go Up in Smoke**, *Maia Brenner, AI Specialist,* ***Tyrolabs***

**Abstract:**
Nowadays, AI is more than a promising idea; it has become imperative to get a competitive advantage. Companies started to look at their data insights to stay on track, giving their first steps into their AI journey with ad-hoc pilot and PoC projects.

But, without the proper roadmap and building blocks, many of these efforts will fall short and project will never get into production. According to McKinsey, only 8% of companies have integrated AI in core practices that support widespread adoption.

How to avoid falling into that category and succeed as an AI organization?

In this talk, we will introduce best practices to no get trapped during the PoC phase.

Attendees will learn practical approaches for:

- Avoiding common pitfalls when starting a PoC
- Understanding the feasibility, impact, and ROI of different AI initiatives
- Building a simple framework for PoC development to break away from the pack and be able to put the system in production

**Technical Level: 6/7**

**What you will learn:**
- MLOps best practices need to be adopted and integrated from the very beginning.
- Avoiding common pitfalls when starting a PoC
- Understanding the feasibility, impact, and ROI of different AI initiatives
- Building a simple framework for PoC development to break away from the pack and be able to put the system in production

**10: 45 AM - 11:25 AM EST**

**CyclOps - A framework for Data Extraction, Model Evaluation, and Drift Detection for Clinical Use-Cases,** *Amrit Krishnan, Senior Applied ML Specialist, and Vallijah Subasri, Graduate Researcher & Applied Machine Learning Intern,* ***Vector Institute***

**Abstract:**
The ever-growing applications of Machine Learning (ML) in healthcare emphasizes the increasing need for a unified framework that harmonizes the various components involved in the development and deployment of robust clinical ML models. Namely, data extraction and model robustness are primary challenges in the healthcare domain. Data extraction is particularly convoluted due to a lack of standardization in Electronic Health Record (EHR) systems used across hospitals. Building robust clinical ML systems has also proven difficult, attributed to dataset shifts that change feature distributions and lead to spurious predictions. Rigorous evaluation of ML models across time, hospital sites and diverse patient cohorts is critical for identifying model degradation and informing model retraining.

**Technical Level: 6/7**

**What you will learn:**
We wish to share our rationale and technical design of the framework, with the broader MLOps community. Our framework will be open-source, and takes a different approach compared to enterprise solutions that are trying to solve similar problems in the healthcare domain. Furthermore, we are building this on top of one of Canada's largest retrospective databases collected for clinical use-cases, with several hospitals in the Greater Toronto Area as partners, which makes it uniquely positioned.

**10: 45 AM - 12:15 AM EST**

**Concretes Guidelines to Improve ML Model Quality, Based on Future ISO Certifications,** *Olivier Blais, VP Decision Science,* ***Moov AI***

**Abstract:**
Only 15% of AI projects will yield results in 2022. That's bad. The good news: there is a better way. We can deliver high-level quality AI systems that meet business objectives and drive adoption. Olivier Blais is Head of Decision Science and Editor of the international AI ISO project on quality evaluation guidelines for AI Systems. He lives and breathes to redefine the quality of AI systems and apply it to real-world business challenges. He'll share a new quality evaluation approach supported by the upcoming ISO standards that redefines how you deliver your ML models and AI systems.

**Technical Level: 5/7**

**What you will learn:**
- Current validation and testing methodologies are often not sufficient
- There are new quality evaluation processes and tools
- Proper quality evaluation enhance delivery success likelihood as well as adoption

**11: 30 AM - 12:15 PM EST**

**Shopify's ML Platform Journey Using Open Source Tools. Case study building Merlin & AMA,** *Isaac Vidas, Machine Learning Platform Tech Lead,* ***Shopify***

**Abstract:**
Merlin, Shopify's new machine learning platform is based on an open source stack and tooling end-to-end. In this talk I will share a deeper look at the process, architecture, and how Merlin is helping us scale our ML work. This talk will be based on the following blog post with additional details on our architecture, technologies and tools - https://shopify.engineering/merlin-shopify-machine-learning-platform

**Technical Level: 3/7**

**What you will learn:**
How to build a machine learning platform with open source tools (Ray, Kubernetes, ML libraries, etc.)

**11: 30 AM - 12:15 PM EST**

**MLOps for Deep Learning,** *Diego Klabjan, Professor and Yegna Jambunath, MLOps Researcher,* ***Northwestern University, Center for Deep Learning***

**Abstract:**
In model serving, two important decisions are when to retrain the model and how to efficiently retrain it. Having one fixed model during the entire often life-long inference process is usually detrimental to model performance, as data distribution evolves over time, resulting in a lack of reliability of the model trained on historical data. It is important to detect drift and retrain the model in time. We present an ensemble drift detection technique utilizing three different signals to capture data and concept drifts. In a practical scenario, ground truth labels of samples are received after a lag in time, which we consider appropriate. Our framework automatically decides what data to use to retrain based on the signals. It also triggers a warning indicating a likelihood of drift.

**Technical Level: 6/7**

**What you will learn:**
- The practical challenges in Model serving for Deep Learning
- Possible algorithmic and modeling solutions
- How to use our open-source project which incorporates these aspects.

**1: 00 PM - 1:40 PM EST**

**Managing a Data Science Team During the Great Resignation,** *Jessie Lamontagne, Data Science Manager,* ***Kinaxis***

**Abstract:**
The COVID-19 pandemic fundamentally changed the way we work and no industry has seen as much change as the tech industry, with many tech giants committing to continue to support remote work for their employees, and hybrid or remote work becoming the new normal. In this talk I cover the challenges and opportunities we face as leader when managing an increasingly remote workforce which now has the opportunity to tap into global labour demand for tech talent. How we retain, motivate, and grow our teams requires rethinking the relationship we have with the firm and with each other, and what it means to build trust.

**Technical Level: 1/7**

**What you will learn:**
Keeping good talent requires treating them as individuals - each with unique goals, dreams and aspirations.

**3: 00 PM - 3:40 PM EST**

**How to Conquer Data Drift & Prevent Stale Models in Production using DVC,** *Milecia McGregor, Developer Advocate* ***Iterative AI***

**Abstract:**
Deploying a machine learning model production is not the end of the project. You have to constantly monitor the model for model drift and the underlying data drift that causes it. That means you have to re-train your model on new datasets often.

In this talk, we'll cover how you can use DVC to track all of the changes to your dataset across each model that gets trained and deployed to production. You'll see how to reproduce experiments and how you can share experiments and their results with others on your team. By the end of the talk, you should feel comfortable switching between datasets as you keep your model up to date.

**Technical Level: 4/7**

**What you will learn:**
How to version their data as they get insights from production and use that to deploy updated models

**3: 00 PM - 3:40 PM EST**

**DataOps For the Modern Computer Vision Stack,** *James Le, Data Advocate,* ***Superb AI***

**Abstract:**
Implementing state-of-the-art architectures, tuning model hyperparameters, and optimizing loss functions are the fun parts of computer vision. Sexy as it may seem, behind each model that gets deployed into production are data labelers and data engineers responsible for building a high-quality training dataset that serves as the model's input. This talk will provide an overview of DataOps for computer vision, outline the challenges that any computer vision teams have to deal with, and propose specific functions of an ideal DataOps platform to address these challenges.

**Technical Level: 5/7**

**What you will learn:**
The audience will learn the origin of DataOps from the data analytics world, why it is important to bring the DataOps discipline to the computer vision world, building blocks of the DataOps lifecycle, and the future of the modern computer vision stack.

# Thank you to our Sponsors/Exhibitors

## Platinum Sponsor

aporia

## Gold Sponsors

Pachyderm

iterative

Qwak

Superwise

iguazio

modzy

CLEAR|ML

SELDON

IBM

## Silver Sponsors

tecton

ROBUST INTELLIGENCE

tenstorrent

arize

Petuum

agnostiq

Genesis Cloud

ocado TECHNOLOGY

## Bronze Sponsors

SharpestMinds

UbiOps

PRIVATE AI

TROJ.AI

## Community Partner

VECTOR INSTITUTE