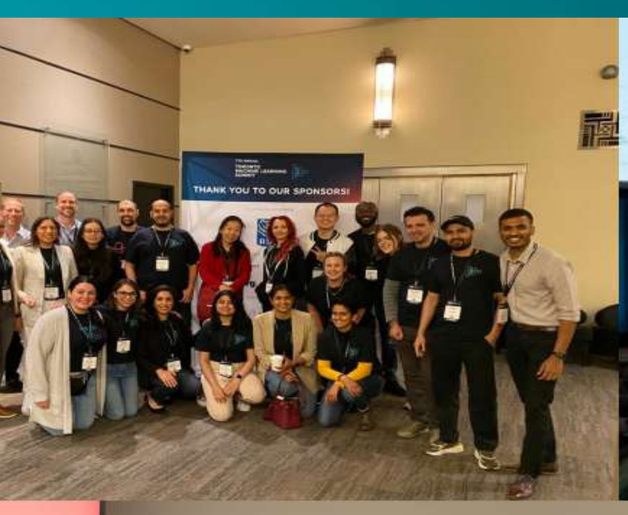
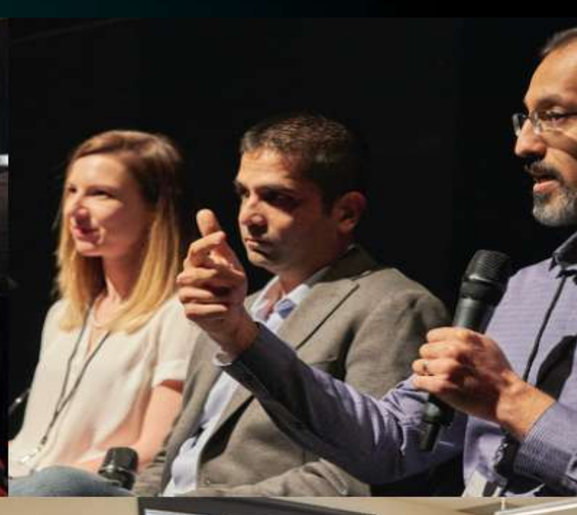
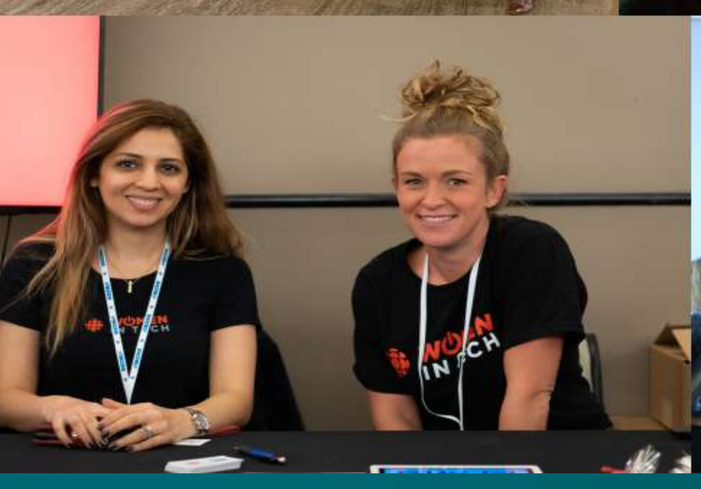
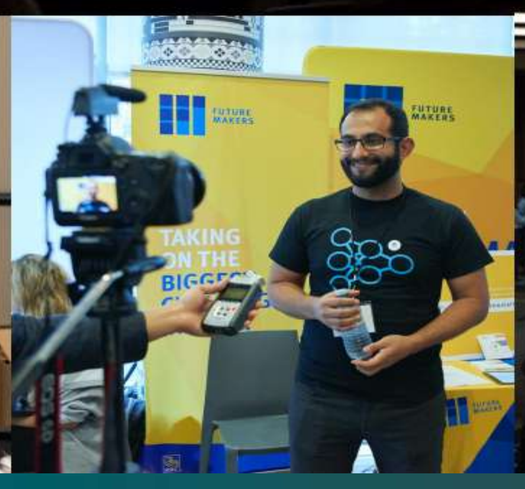
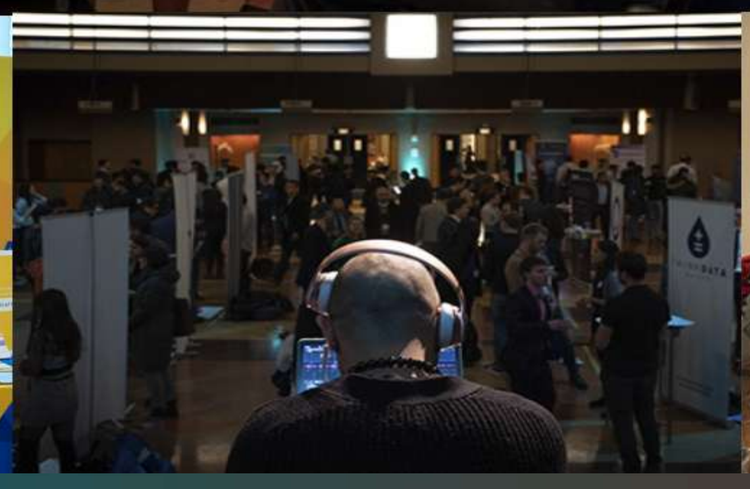
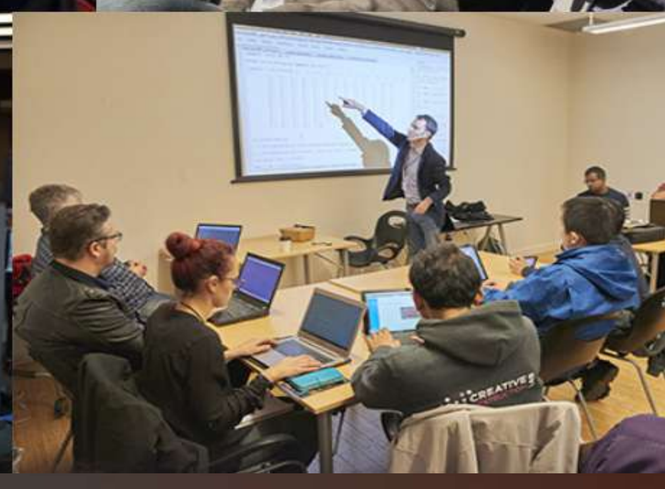
GENERATIVE AI SUMMIT

4th Annual
MLOPS WORLD CONFERENCE & EXPO





Join here

October 25th - 26th

Austin, Texas

Technical and Strategic Use Cases and Workshops from

Spotify · instacart · kaggle · lyft · cohere · Pinterest · EY · OpenAI · NASA · Capital One

www.mlopsworld.com

4th Annual MLOPS WORLD

GENERATIVE AI SUMMIT

www.mlopsworld.com
www.generative-ai-summit.com

SESSION TRACKS

Business & Strategy	Research/Technical
Case Studies	Networking & Additional
Exhibitior Demo Presentations	

Tickets give access to all sessions. Build your own program, your way.

October 24th: Virtual Talks

9:00am	***Sponsored: Building a Conversation-focused LLM on Communication Data*** Kartik Talamadupula, Director of AI Research, **Symbl.ai**
9:30 am	***Sponsored: Finetuning a large language model on a custom dataset*** Aniket Maurya, Developer Advocate **Lightning AI**
10:00am	***Sponsored: Feature Stores in Practice: Train and Deploy an End-to-End Fraud Detection Model with Featureform, Redis, and AWS.*** Simba Khadder, Founder & CEO **Featureform**
10:30am	***Sponsored: Applying GitOps principles at every step of an E2E MLOps project - an interactive workshop*** **Tibor Mach,** Machine Learning Solutions Engineer **DVC**
11:00am	***Sponsored: Deploying generative AI models: best practices and an interactive example*** Anouk Dutrée, Product Owner **UBIOPS**
11:30am	***Sponsored: LLMs in practice: a guide to recent techniques and trends*** Ville Tuulos, CEO Eddie Mattia, Data Scientist **Outerbounds**
12:00pm	***Sponsored: Introduction to Building ML Microservices: A hands-on approach with examples from the music industry*** Ramon Perez, Developer Advocate **Seldon**
12:30pm	***Sponsored: Learn Your Codebase: Fine-tuning CodeLlama with Flyte… to Learn Flyte*** Niels Bantilan, Chief ML Engineer **UNION.AI**
1:00pm	**Valohai**
1:30pm	***Quality Assurance (QA) in Machine Learning*** Serg Masis, Lead Data Scientist, **Syngenta** & Bestselling Author
2:00pm	***Avoid ML OOps with ML Ops:*** ***A modular approach to scaling Forethought's E2E ML Platform*** Salina Wu, Senior Machine Learning Infrastructure Engineer, **Forethought** ***The New AI Engineer: Building with Legos*** Aarushi Kansal, Principal Engineer, **Bumble**
2:30pm	***Lessons Learned Productionising LLMs for Stripe Support*** Sophie Daly, Staff Data Scientist, **Stripe** ***Lessons Learned: The Journey to Real-Time Machine Learning at Instacart*** Guanghua Shu, Staff Machine Learning Engineer, **Instacart**
3:00pm	***Spend Less Time Troubleshooting ML Production Issues*** Alon Gubkin, CTO & Co-Founder **Aporia**
3:30pm	***Applying Responsible AI with the Open-Source LangTest Library*** David Talby, CTO **John Snow Labs**
4:00pm	***Making ChatGPT Funny with Prompt Optimization*** Mike Taylor, Owner **Saxifrage** ***Synthetic Data: Generative AI for Enhanced Data Quality in the Era of Foundational Models*** Fabiana Clemente, Chief Data Officer, Gonçalo Ribeiro , CEO, **YData**
4:30pm	***How Booking.com is approaching LLMOps: Approaches & Considerations, Pro's & Con's*** Sanchit Juneja, Director-Product (Big Data & ML), **Booking.com** ***LLM economics : The Cost of leveraging Large Language Models*** Nikunj Bajaj, CEO **Truefoundry**
5:00pm	***Learning from Extremes: What Fraud-Fighting at Scale Can Teach Us About MLOps Across Domains*** Greg Kuhlmann, CEO, **Sumatra**
5:30pm	***How to Design and Build Resilient Machine Learning Systems*** Dan Shiebler, Head of Machine Learning, **Abnormal Security**
6:00pm	***Ingredients of a Comprehensive MLOps Program*** Mac Macoy, Chick-fil-A, Senior Lead ML Engineer, **Chick-fil-A** Lauren Face, Senior Technical Program Lead, **Chick-fil-A**

4th Annual MLOPS WORLD

GENERATIVE AI SUMMIT

www.mlopsworld.com
www.generative-ai-summit.com

October 25th: In Person, Austin Texas

SESSION TRACKS

- Business & Strategy
- Research/Technical
- Case Studies
- Workshop
- Exhibitior Demo Presentations

Tickets give access to all sessions. Build your own program, your way.

Registration and Exhibits Open

Opening Remarks: David Scharbach, Executive Director, ***MLOps World***

Keynote Special Guest

10:25 pm Break | Exhibition |

Removing the Roadblocks to Build Great GenAI Products

Liran Hason, Co-Founder & CEO, **Aporia**

Hyper-scaling Real-time Personalization with Privacy via on-Device Computing

NimbleEdge

Business Panel: GenAI Use-cases Across Industry Verticles. Early Trends and ROI

Surbhi Rathore, CEO & Co-founder, **Symbl.ai**
Shingai Manjengwa, Head of AI Education **ChainML,**
Kamelia Aryafar, Senior Engineering Director **Google Cloud AI**
Manas Bhuyan, **Deloitte Consulting LLP**

LLMOps: An Emerging Stack to Productionalize LLM Applications

Hien Luu, Head of ML Platform

DoorDash

Low-latency Model Inference in Finance

Vincent David, Senior Director, Machine Learning| Michael Meredith, Lead Software Engineer, **Capital One**

Supercharging Recommender Systems: Unleashing the Power of Distributed Model Training

Susrutha Gongalla, Principal Machine Learning Engineer, **Stitch Fix**

Evolution of ML Training and Serving Infrastructure at Pinterest

Aayush Mudgal, Senior Machine Learning Engineer, **Pinterest**

MLOps on Highly Sensitive Data - Strict Confinement, Confidential Computing and Tokenization Protecting Privacy

Andreea Munteanu, AI/ML Product Manager | Maciej Mazur, Principal ML Engineer, **Canonical**

MLOps for Graph-Based Recommender Systems: Orchestrating Intelligent Connections

Bandish Shah, Engineering Manager, Advanced Technical/Research, MosaicML/Databricks

WORKSHOP:

Retrieval Augmented Generation (RAG) with LangChain: "ChatGPT for Your Data" with Open-Source Tools Part 1

Dr. Greg Loughnane, Chris Alexiuk Founder & CEO, Head of LLMs at AI Makerspace

12:25 pm Lunch Break | Exhibition

Lightning Talks

Cracking the Code: Why SWEs Should Embrace Prompt Engineering (and what they risk by sticking to their comfort zone)

PatrickMarlow, Conversational AI Engineer, Cloud AI Incubator,

From Analytics to AI: How GenAI can Unlock Data Insights and Transform Decision-making

Shingai Manjengwa, Head of AI Education, ChainML

From Model T to Machine Learning: A Glimpse into Ford's MLOps and Hybrid Infrastructure Strategy

Muller Mu,Solution Architect / Senior Scientist
Naiel Samaan, Senior Product Owner, AI Platform
Valmir Bucal, AI/ML Platform Product Owner

Fine-tune LLMs or Integrate 3rd Party APIs? A Financial Case-study

Hannes Hapke, Principal Machine Learning Engineer, **Digits**

Panel: How to Finetune your LLM's and Evaluate Performance

Shaun Hillin, Global Head of Solutions Architecture **Cohere**
Hien Luu, Head of ML Platform, **DoorDash**
Savin Goyal, CEO **Outerbounds**

From Prototype to Product: Rapid iteration and ML model deployment at Dropbox

Richie Frost Software Engineer, ML Foundations, **Dropbox**

Operationalizing Data-centric AI: Practical Algorithms + Software to Quickly Improve ML Datasets

Jonas Mueller, Chief Scientist & Co-Founder, **Cleanlab**

How Many Labelled Examples do you Need for a BERT-sized Model to Beat GPT-4 on Predictive Tasks?

Matthew Honnibal,Founder & CTO, Writer, **ExplosionAI**

Your AI applications need Guardrails: Here's how to build them

Shreya Rajpal, Founder, **Guardrails AI**

Using Scouter Models to monitor Model Drift - A novel approach

Kumaran Ponnambalam, Principal Engineer - AI, **Cisco Systems Inc., Emerging Tech & Incubation**

WORKSHOP:

Evaluation Techniques for Large Language Models

Rajiv Shah
Machine Learning Engineer, **Hugging Face**

Sponsored **WORKSHOP**

Build Your Own ChatGPT with Open Source Tooling

Andreea Munteanu, MLOps product manager

Sponsored **WORKSHOP**

Production ML Serving & Monitoring in Kubernetes

Seldon

Round Table Networking + Happy Hour

OPENING DAY PARTY

4th Annual MLOPS WORLD

GENERATIVE AI SUMMIT

www.mlopsworld.com
www.generative-ai-summit.com

SESSION TRACKS

Business & Strategy	Research/Technical
Case Studies	Networking & Additional
Exhibitior Demo Presentations	

Tickets give access to all sessions. Build your own program, your way.

October 26th: In Person, Austin Texas

Registration and Exhibits Open

Opening Remarks: David Scharbach, Executive Director, **MLOps World**

Keynote: Is it Too Much to Ask for a Stable Baseline?
D. Sculley, CEO Kaggle

Getting Higher ROI on MLOps Initiatives: 5 Lessons Learned While Building out the MLOps Platform for 100+ Data Scientists

Stefan Krawczyk
CEO & Co-founder, **DAGWorks Inc.**

DAGWORKS

Beyond the Kaggle Paradigm: Future of End-to-End ML Platforms

Norm Zhou, Stealthmode Startup (Formerly at Meta)

Gen AI and Lightning: Accelerating AI Innovation while Ensuring Scalability and Security

William Falcon, Founder & CEO, **Lighting AI**

Lightning AI

WORKSHOP:

Retrieval Augmented Generation (RAG) with LangChain: "ChatGPT for Your Data" with Open-Source Tools Part 2

Dr. Greg Loughnane, Chris Alexiuk
Founder & CEO, Head of LLMs at AI Makerspace

WORKSHOP:

Chat with MLOpsWorld: Engineering an LLM Application

Charles Frye, Deep Learning Educator,

THE FULL STACK

Lessons Learned from Implementing GenAI at Large Enterprises

Ilyas Iyoob PhD, Faculty, Data Science & Artificial Intelligence, **The University of Texas at Austin**

TEXAS
The University of Texas at Austin

Creating the World's Premier Biological Foundation Model

Jess Leung, Staff Machine Learning Engineer, Recursion

RECURSION

Data Versioning in Generative AI: A Pathway to Cost-Effective ML

Dmitry Petrov, CEO, **DVC**

Open Source - Open Opportunities: What "Building Computers" for AI requires

David Bennett Chief Customer Officer, EVP, **Tenstorrent**

tenstorrent

Supercharging Search with LLMs: The Instacart Journey

Prakash Puttal, Staff Software Engineer, **Instacart**

instacart

Panel: Ethics, Compliance & Addressing Uncertainties with GenAI

12:25 pm Lunch Break | Exhibition

Lightning Talks

Amumu brain; How League of Legends uses machine learning an applied data science

Ian Schweer, Staff Software Engineer, **Riot Games**

RIOT GAMES

WORKSHOP:

Stable Diffusion for Your Images: Custom Dream

Sandeep Singh, Head of Applied AI, Beans.AI

BEANS.AI

***Sponsored* WORKSHOP:**

MLOps for Production-ready LLM — Putting LLMs into Production Through Iterative Training, Fine-Tuning, and Serving

Jay Chun, Co-founder & CTO,

VESSL AI

Evolved Structures: Using AI and Robots to Build Spaceflight Structures at NASA

Ryan McClelland, Research Engineer

NASA Goddard Space Flight Center

Introduction to LangChain and Retrieval Augmented Generation (RAG)

Sophia Yang, Senior Data Scientist **Anaconda**

ANACONDA.

Future of AI in Production/ GenAI DEMO Sessions

Closing Remarks

Thank you to Sponsors

Platinum: aporia

Gold: tenstorrent, featureform, Lightning AI (Creators of PyTorch Lightning), SELDON, DVC, VESSL AI, Canonical, neptune.ai, Outerbounds

Silver: Genesis Cloud, UNION, symbl.ai, Valohai

Bronze: UbiOps, elastic, Deepgram, xethub, ivy, HOPSWORKS

OCT 24TH

9:00 AM - 10:00 AM EST

Building a Conversation-focused LLM on Communication Data
Kartik Talamadupula, Director of AI Research, Symbl.ai

Workshop

Abstract: Communication data is a powerful resource for building large language models (LLMs). In this talk, we will discuss the challenges and opportunities of using communication data to build a conversation-first LLM.
We will introduce Nebula, a conversation-focused LLM from Symbl.ai that is trained and fine-tuned on a large dataset of communication data. Nebula is able to generate and analyze text, translate languages, write different kinds of creative content, and answer questions in an informative way, all with a focus on conversation.
We will discuss the challenges of using communication data to train LLMs, including the need for high-quality data, the need to address data privacy and security concerns, and the need to develop ML Ops pipelines that can handle the scale and complexity of communication data.

We will also discuss the opportunities that communication data presents for building LLMs. For example, communication data can be used to train LLMs that are more aware of the context of a conversation, and that are better able to understand and respond to human emotions.

Speakers Bio: Aniket is a Developer advocate at Lightning AI. He is an open source enthusiast and contributor to some popular repos like Lit-GPT and Gradsflow.

How technical is this talk?: 3/7

What You'll Learn: Specializing generative models for conversations, and the challenges and opportunities there-in

Pre Requisite Knowledge: Basic knowledge of AI and ML terminologies

9:30 AM - 10:30 AM EST

Finetuning a large language model on a custom dataset
Aniket Maurya, Developer Advocate, **Lightning AI**

Workshop

Abstract: This is a hands-on workshop for finetuning large language models using custom dataset. By the end of this workshop, you will learn about parameter efficient finetuning, optimised inference and tricks to finetune models at scale.

Speakers Bio: Aniket is a Developer advocate at Lightning AI. He is an open source enthusiast and contributor to some popular repos like Lit-GPT and Gradsflow.

How technical is this talk?: 5/7

What You'll Learn: Parameter efficient finetuning and LLM optimisations for very large models.

Pre Requisite Knowledge: Python, PyTorch basics

10:00 AM - 11:00 AM EST

Feature Stores in Practice: Train and Deploy an End-to-End Fraud Detection Model with Featureform, Redis, and AWS.
Simba Khadder, Founder & CEO, **Featureform**

Workshop

Abstract: In this workshop, we'll go through the process of building a fraud detection model from scratch using Featureform's open-source feature store alongside a handful of other tools like Redis and Sagemaker. We'll both train and deploy the model through this workshop. We'll deep dive into where feature stores fit into the MLOps stack, the value they provide, and how to use them in practice.

Speakers Bio: Simba Khadder is the founder & CEO of Featureform. He started his ML career in recommender systems where he architected a multi-modal personalization engine that powered 100s of millions of user's experiences. He later open-sourced and built a company around their feature store. Featureform is the virtual feature store. It enables data scientists to define, manage, and serve model features using a Python API. Simba is also a published astrophysicist, an avid surfer, and ran a marathon in basketball shoes.

How technical is this talk?: 4/7

What You'll Learn: Participants will learn how to:
- Use Featureform to build, manage, and serve their model features from fraud detection data.
- Use Redis, Spark, and Sagemaker to train and deploy a random forest model.
- Use Terraform and other best practice DevOps, DataOps, and MLOps through the process.

Pre Requisite Knowledge: Basics of cloud, networking, databases, and machine learning.

10:30 AM - 12:00 PM EST

Applying GitOps principles at every step of an E2E MLOps project - an interactive workshop
Tibor Mach, Machine Learning Solutions Engineer, **Iterative**

Workshop

Abstract: With the emergence of IaC (infrastructure as code) tools, we have seen GitOps become an increasingly popular DevOps pattern that facilitates automation, reproducibility, and security. While hugely beneficial, applying the same principles in MLOps is not straightforward due to the specific aspects of the field such as the need to work with large amounts of data and the experimental nature of ML development. In this talk, we will see how we can bridge these gaps by using tools such as DVC. Step by step, we will create an end-to-end MLOps pipeline which is centered around the git repository as its single source of truth.

Speakers Bio: Tibor Mach is a Machine Learning Solutions Engineer at Iterative.ai He has been working in ML and MLOps in the past 5 years. Tibor has a Ph.D in mathematics from the University of Göttingen and had published papers in the field of probability theory prior to refocusing to ML.

How technical is this talk?: 4/7

What You'll Learn? In this largely interactive workshop you can learn how you can use your git repositories to keep track of your ML experiments, version data and models, maintain a model registry and handle model deployment

Pre Requisite Knowledge: Basics of working with git and conceptual understanding of GitHub Actions or GitLab CI.

11:00 PM
-
12:00 PM EST

Deploying generative AI models: best practices and an interactive example
Anouk Dutrée, Product Owner, **UbiOps**

Workshop

Abstract: Generative AI models are all the hype nowadays, but how do you actually deploy them in a scalable way? In this talk we will discuss best practices when moving models to production, as well as show an interactive example of how to deploy one using UbiOps. UbiOps is a serverless and cloud agnostic platform for AI & ML models, built to help data science teams run and scale models in production. We will pay special attention to typical hurdles encountered in deploying (generative) AI models at scale. Python knowledge is all you need for following along!

Speakers Bio: Anouk is the Product Owner at UbiOps. She studied Nanobiology and Computer Science at the Delft University of Technology, and did a Master's in Game Development at Falmouth University, which spiked her interest in Machine Learning. Next to her role at UbiOps, she frequently writes for Towards Data Science about various MLOps topics and she co-hosts the biggest Dutch data podcast, de Dataloog. Her efforts in tech have been awarded twice with the T500 award, in both 2020 and 2021

How technical is this talk?: 4/7

What You'll Learn? Deployment at scale doesn't have to be difficult. Participants will learn how to deploy a generative AI model to the cloud themselves, and how to select the right hardware for your use case (CPU,GPU,IPU etc.).

Pre Requisite Knowledge: Python knowledge and a basic understanding of AI/ML models

10:30 AM
-
12:00 PM EST

LLMs in practice: a guide to recent techniques and trends
Ville Tuulos, CEO, Outerbounds
Eddie Mattia, Data Scientist, Outerbounds

Workshop

Abstract: In this workshop, attendees will learn about methods for working with LLMs. Our stories will be guided by examples you can run on your laptop or in a (free) hosted cloud environment provided to attendees. Developers will expand their awareness of how researchers and product designers are working with LLMs, with emphasis on connecting high-level concepts such as fine-tuning and vector databases to the fundamental math and APIs data scientists should understand. Business-minded executives can either get hands - on or follow the higher-level stories to deepen their sense of what is possible with LLMs, the technicalities behind risks they introduce, and how they fit into the arc of ML. The primary value of this workshop will be as a guide to help teams set reasonable goals in the complex and fast-moving world of LLMs, and understand what you need to successfully support your team's next LLM projects.

Speakers Bio: Ville has been developing infrastructure for machine learning for over two decades. He has worked as an ML researcher in academia and as a leader at a number of companies, including Netflix where he led the ML infrastructure team that created Metaflow, a popular open-source framework for data science infrastructure. He is a co-founder and CEO of Outerbounds, a company developing modern human-centric ML. He is also the author of the book, Effective Data Science Infrastructure, published by Manning.
Eddie Mattia is a data scientist at Outerbounds who began using Python to teach applied math in grad school. Since then, Eddie has worked at startups and at Intel building machine learning software

How technical is this talk?: 5/7

What You'll Learn? There are cheap (e.g., APIs) and expensive (e.g., fine-tuning, training) ways to build on top of LLMs. The methods you choose have consequences in apps you can build and how your dev team works. We will learn how to think about these choices as we develop basic apps you can use as templates for future genAI projects. Learners have the option to follow along in a provided dev environment where we will unpack these choices and make the tradeoffs and decision space concrete.

Pre Requisite Knowledge:
*Basic knowledge of Python
* Ability to use the command line
* Ability to use common ML algorithms in a notebook environment

12:00 PM
-
3:00 PM EST

Introduction to Building ML Microservices: A hands-on approach with examples from the music industry
Ramon Perez, Developer Advocate, **Seldon**

Workshop

Abstract: Serving a machine learning model is not particularly easy, especially if we add two or three models in parallel to the mix, in which case, a single model deployment recipe might start to crumble. To tackle the challenges around serving individual or multiple models in production, we have handy tools like MLServer and Seldon Core. The former is a python library that allows us create machine learning microservices with one or multiple models in the same service, and the latter allows us to build simple-to-complex inference graphs that can help us handle A/B testing, shadow and canary deployment, feature transformations, and model monitoring. If you want to learn how to use open-source tools to build microservices based on your different use cases and model recipes, come and join this hands-on workshop and get started with several of the key steps in the machine learning workflow as we walk through fun examples from the broader music industry.

Speakers Bio: Ramon is a data scientist and educator currently working in the Developer Relations team at Seldon in London. Prior to joining Seldon, he worked as a freelance data professional and as a Senior Product Developer at Decoded, where he created custom data science tools, workshops, and training programs for clients in various industries. Before Decoded, Ramon wore different research hats in the areas of entrepreneurship, strategy, consumer behavior, and development economics in industry and academia. Outside of work, he enjoys giving talks and technical workshops and has participated in several conferences and meetup events.

How technical is this talk?: 3/7

What You'll Learn? The core of the workshop will teach participants how to create machine learning microservices and inference graphs, and how to monitor the predictions made by these services. The main use case we'll follow throughout the workshop comes from the music industry, so this will be a fun and content-rich 3 hours to go through.

Throughout the workshop, we will be building a creative ML platform in several incremental steps. In the first 50 minutes of the workshop, we will set up the user interface and the back-end our application, and then we'll spin up the first model we will interact with. In the second 50-minute section, we will start adding different functionalities to our platform by running new machine learning models inside our inference server. Lastly, we'll create different replicas of each model, develop an inference graph to come up with unique tunes, and conduct AB testing on our service to assess and evaluate the output of different models when compared with real songs.

Within our 3 hours together, we'll have two 10- to 15-minute breaks and there will be plenty of exercises for participants to complete.

12:30 PM - 2:00 PM EST

Learn Your Codebase: Fine-tuning CodeLlama with Flyte… to Learn Flyte
Niels Bantilan, Chief ML Engineer, Union.ai

Workshop

Abstract: Today, foundation LLMs can only be trained by a handful of organizations possessing the compute resources required to pre-train models with more than a hundred billion parameters over internet-scale data. These foundation models are then fine-tuned by the wider ML community for specific applications. Even though fine-tuning can be more compute and memory efficient than full parameter tuning, a significant challenge to fine-tuning is provisioning the appropriate infrastructure.

In this session, Niels will demonstrate how to use Flyte, a Linux Foundation open-source orchestration platform to fine-tune a LLM on the Flyte codebase itself 🤯. Flyte allows for the declarative specification of the infrastructure needed for a broad range of ML workloads, including fine-tuning LLMs with limited resources by leveraging multi-node, multi-gpu distributed training.

Speakers Bio: Niels is the Chief Machine Learning Engineer at Union.ai, and core maintainer of Flyte, an open source workflow orchestration tool, author of UnionML, an MLOps framework for machine learning microservices, and creator of Pandera, a statistical typing and data testing tool for scientific data containers. His mission is to help data science and machine learning practitioners be more productive.

He has a Masters in Public Health with a specialization in sociomedical science and public health informatics, and prior to that a background in developmental biology and immunology. His research interests include reinforcement learning, AutoML, creative machine learning, and fairness, accountability, and transparency in automated systems.

How technical is this talk?: 5/7

What You'll Learn? Attendees will gain hands-on experience using Flyte to leverage state-of-the-art deep learning tools such as `torchrun` distributed training, LoRA, 4/8-bit quantization, and FSDP, while benefiting from Flyte's reproducibility, versioning, and cost management capabilities. At the end of this talk, you'll be able to take the code and adapt it to learn your own code base to help to answer user-support questions, create boilerplate starter code, or whatever downstream task you're interested in!

Pre Requisite Knowledge:
Intermediate Python
Intermediate Machine Learning
Familiarity with Command-line Tools

1:30 PM - 2:00 PM EST

QA in ML
Serg Masis, Lead Data Scientist & Bestselling Author, **Syngenta**

Workshop

Abstract: Trust is mission-critical for any technology, so if AI/ML solutions are to supplant and complement software, AI must reach the reliability standards currently expected from software. The difference is Quality Assurance (QA) has existed in software for three decades, and the burgeoning field of ML has barely begun to perform quality controls:
1) We will take a journey through the history of QA, discuss why it is crucial, and what lessons from other disciplines and industries we can apply to machine learning.
2) Then, will discuss what important role Explainable AI methods, not to mention best practices in MLOps, data engineering, and data science, can play.
3) Lastly, we will discuss the challenge ahead. Given the many steps in an Machine learning (ML) and the many qualities to assess in an ML model, choreographing and standardizing tasks in a QA effort is a challenging undertaking. New roles for ML QA will likely appear within DevOps, SecOps, and MLOps teams to ensure increased reliability and robustness. Still, also, the roles of data scientist and Machine Learning engineer will evolve to enhance quality.
Thus session is ultimately about what business stakeholders and practitioners can do to make AI/ML more trustworthy to the end-users of this technology.

Speakers Bio: Bestselling author of ML/AI books. Lead Data Scientist at multinational agribusiness company creating models for sustainable practices in agriculture.

How technical is this talk?: 2/7

What You'll Learn?
- What the history of quality assurance (QA) teaches us about how QA can be implemented in ML.
- What tools and roles already exist in ML that can enforce QA.
- What's missing to make QA work much better.

2:00 PM - 2:30 PM EST

Avoid ML OOps with ML Ops: A modular approach to scaling Forethought's E2E ML Platform
Salina Wu, Senior Machine Learning Infrastructure Engineer, **Forethought**

Workshop

Abstract: As Machine Learning becomes more ubiquitous in business and product applications, the need for a cost-efficient, scalable, and automated infrastructure to support the end-to-end ML lifecyle becomes mission critical. However, a scalable and reusable ML Ops platform is often an afterthought in productionizing ML models, due to urgency of business needs and lack of resources or experience. A very common scenario is for ML Ops to be ad-hoc and de-centralized, with no good way to reproduce or automate ML processes. It can be challenging, especially for smaller teams, to identify and foresee specific ML Ops needs and understand how to address them.

Forethought is an enterprise company building AI-powered customer experience (CX) solutions. Our products require training customer-specific language models and deploying them on low-latency, high-uptime endpoints. With ML at the heart of our business, our infrastructure supporting it is pivotal to our growth and success. At Forethought, we took a close look at our initial ML infrastructure, aiming to identify key areas of improvement and anticipate future requirements. Through a step-by-step approach, we gradually replaced our existing infrastructure with improved, modular components to arrive at a much more mature system. This case study will dive into which areas we identified as critical to replace as well as the steps we took to enhance them. In particular, we will look at the following:

- Streamlining ML training and migrating to the Sagemaker training platform
- Achieving efficient model serving with Sagemaker Serverless and Multi-Model Endpoints
- Orchestrating our ML processes with automated pipelines on Dagster
- Centralizing ML feature engineering across our datalake using Spark
- Building intuitive model management tooling with Retool

Through this talk, we'll show a real-world scenario of bringing a rudimentary v0 ML architecture to an enhanced v1 architecture. We will also share our plans and progress building towards our v2 vision, including automated re-training and LLM support. Key takeaways will include:

- Understanding the different components of a solid ML infrastructure
- Identifying and proactively addressing bottlenecks and opportunities for growth in your ML lifecycle
- Learning how to improve and migrate your ML infrastructure in stages
- Understanding the goals and best practices of a stable end-to-end ML infrastructure

Speakers Bio: Salina Wu is a Sr. Machine Learning Infrastructure engineer at Forethought.ai. She works closely with the Machine Learning team to build and maintain their end-to-end training, serving, and data infrastructures. She is particularly motivated by introducing new ways to improve efficiency and reduce cost across the ML space. When not at work, Salina enjoys surfing, pottery, and being in nature.

How technical is this talk?: 5/7

2:00 PM
-
2:30 PM EST

What does it mean to be an "AI Engineer" in the world of Generative AI. Merging Business Strategy and the Technical Ecosystem.
Aarushi Kansal, Principal Engineer, Bumble

Workshop

Abstract: In this talk the audience will learn about the rapidly changing ecosystem and what it now means to be an ""AI Engineer"" in the world of Generative AI. It will cover the technical eco system while also touching on business strategy.
The key function of a healthy team will be emphasized, that is, when both engineers and business decision-makers understand the AI ecosystem and how it affects their respective remits.
Currently, there's a lot of hype around AI but not enough understanding and education about what is possible, where certain tools fit in and how both engineers and business decision-makers can make use of this new boom.
This talk will tie in use cases and business strategies with both technical & business considerations as they impact this rapidly changing eco-system

Speakers Bio: Aarushi is a principal engineer at Bumble, she is heavily focused and experienced in cloud engineering, generative AI, Golang and Python

How technical is this talk?: 4/7

What You'll Learn? Key learnings: applying traditional engineering to AI engineering, key skills needed from both a hiring and being hired point of view. Key tools / concepts to know as an AI engineer

3:00 PM
-
4:00 PM EST

Lessons Learned: The Journey to Real-Time Machine Learning at Instacart
Guanghua Shu, Staff Machine Learning Engineer, **Instacart**

Workshop

Abstract: Instacart incorporates machine learning extensively to improve the quality of experience for all actors in our “four-sided marketplace” — customers who place orders on Instacart apps to get deliveries in as fast as 30 minutes, shoppers who can go online at anytime to fulfill customer orders, retailers that sell their products and can make updates to their catalog in real time, and the brand partners that participate in auctions on the Instacart Advertising platform to promote their products.
A typical shopping journey at Instacart is powered by hundreds of machine learning models. Many decisions/actions happen in real time, which means leveraging machine learning in real-time can provide significant value to the business. One of the major changes we have gone through is transitioning many of our batch-oriented ML systems into real-time. In this talk, we describe ML platform at Instacart with a focus on the journey of real-time ML. We will discuss both fundamental infrastructures and important use cases, review main challenges and decisions, and draw important lessons that could help others learn from our experience.

Speakers Bio: Guanghua Shu is a Staff Machine Learning Engineer at Instacart, where he focuses on building end-to-end machine learning solutions to gain actionable insights from data in the e-commerce domain. Guanghua has spent over six years in applied machine learning, and worked on recommender systems for product recommendation, leveraging ML to improve cloud security, and Data/AI platforms.

Guanghua holds a PhD in ECE from University of Illinois at Urbana-Champaign. He has published over 30 research papers and received over 10 patents. Through academia and industry experience, Guanghua has explored different abstract levels of the technology stack, ranging from ASIC design, computer architecture, distributed software systems, and applied machine learning. He believes in the power of technology and its outsized impact on business, society and beyond.

How technical is this talk?: 4/7

What You'll Learn?
Key problems to consider for real-time ML.
Important foundations to support real-time ML in a e-commerce platform.
Avoid pitfalls and take aways good lessons in building real-time ML from our experience.

3:00 PM
-
4:00 PM EST

Spend Less Time Troubleshooting ML Production Issues
Alon Gubkin, CTO & Co-Founder, **Aporia**

Workshop

Abstract: Business stakeholders are unhappy with the model decisions again? Manual triage takes up a lot of bandwidth from your team every single time? In this workshop, you’ll learn how ML leaders identify and troubleshoot ML issues in production faster than ever. By being more proactive about common types of ML-specific production issues such as model drift, you’ll be able to spend significantly less time troubleshooting and gain peace of mind to focus on cooler, mission-critical projects.

Speakers Bio: Alon is the CTO and Co-Founder at Aporia, an ML Observability platform designed to empower organizations to trust their AI. Alon spent the last decade leading multiple software engineering teams, working closely with various organizations on their Data & Machine Learning platforms.

How technical is this talk?: 4/7

What You'll Learn? Spend less time and resources on troubleshooting ML production issues.
Model drift awareness and monitoring
Improving model decision processes

What is the prerequisite knowledge or skills from those participating?
Basic understanding of machine learning concepts including model development and deployment.
Experience with ML production environments and awareness of common issues such as model drift.
Familiarity with data science research processes and techniques.
Ability to apply analytical and problem-solving skills to identify and address production issues proactively.

3:30 PM
-
4:00 PM EST

Applying Responsible AI with the Open-Source LangTest Library
David Talby, CTO, **John Snow Labs**

Workshop

Abstract: While there's a lot of work done on defining the risks, goals, and policies for Responsible AI, less is known about what you can apply today to build safe, fair, and reliable models. This session introduces open-source tools and examples of using them in real-world projects - to address three common challenges.

The first is robustness - testing and improving a model's ability to handle accidental or intentional minor changes in input that can uncover model fragility and failure points. The third is bias - testing that a model performs equally across gender, age, race, ethnicity, or other population groups. The third is data leakage, in combination with leakage caused by using personally identifiable information in training data. The open-source LangTest library is used to demonstrate how to generate tests, run tests, augment data, and integrate these evaluations into MLOps workflows.

This session is intended for data science practitioners and leaders who need to know what they can do today to build AI & LLM applications that work safely and reliably in the real world.

Speakers Bio: David Talby is the Chief Technology Officer at John Snow Labs, helping companies apply artificial intelligence to solve real-world problems in healthcare and life science. David is the creator of Spark NLP – the world's most widely used natural language processing library in the enterprise. He has extensive experience building and running web-scale software platforms and teams – in startups, for Microsoft's Bing in the US and Europe, and to scale Amazon's financial systems in Seattle and the UK. David holds a Ph.D. in Computer Science and Master's degrees in both Computer Science and Business Administration. He was named USA CTO of the Year by the Global 100 Awards and GameChangers Awards in 2022.

How technical is this talk?: 5/7

What You'll Learn: This session is intended for data science practitioners and leaders who need to know what they can & should do today to build AI systems that work safety & correctly in the real world. Background Knowledge: Basic familiarity with machine learning is assumed.

Why is this takeaway(s) important? While much effort has been done to define the risks goals, and policies for Responsible AI, little is known about what you can do now to build safe, fair, and reliable models. This talk provides open-source technologies and real-world examples of how you can address those challenges.

What is unique about this, which can't be found online? It is important to test beyond accuracy in your NLP system. This is because the business requirements for the system include robustness, reliability, fairness, toxicity, efficiency, lack of bias, lack of data leakage, and safety. Therefore, your test suites should reflect these requirements. A comprehensive review of definitions and metrics for these terms in different contexts is provided in the Holistic Evaluation of Language Models [Liang et. al 2022], which is well worth reading. However, you will need to write your own tests to determine what inclusiveness means for your specific application.

4:00 PM
-
4:30 PM EST

Synthetic Data: Generative AI for Enhanced Data Quality in the Era of Foundational Models
Fabiana Clemente, Chief Data Officer, **YData**

Workshop

Abstract: A thought-provoking session that explores the transformative potential of synthetic data in the age of foundational language models. As Language Models (LMs) like GPT-3 and its successors continue to reshape the landscape of AI, the quality of training data becomes paramount. This session will deep dives into the synergy between synthetic data and foundational models exploring the impact of data quality in foundational models, the role of synthetic data, data augmentation, bias and data privacy.

Speakers Bio: Fabiana Clemente is cofounder and CDO of YData, combining data understanding, causality, and privacy as her main fields of work and research, with the mission of making data actionable for organizations. As an enthusiastic data practitioner she hosts the podcast When Machine Learning Meets Privacy and is a guest speaker on the Datacast and Privacy Please podcasts. She also speaks at conferences such as ODSC and PyData.

How technical is this talk?: 5/7

What You'll Learn: This session aims to provide a good understanding into core concepts such as foundational models, generative AI, and the process of synthetic data generation. The objective is to provide a comprehensive perspective of how synthetic data can enhance data quality and fuel the capabilities of Machine Learning models, such as LLMs.

4:00 PM
-
4:30 PM EST

Making ChatGPT funny with Prompt Optimization
Mike Taylor, Saxifrage, Owner

Business Strategy

Abstract: A recent study found ChatGPT repeated the same 25 jokes 90% of the time. As is often the case with popular narratives about the limits of AI, ChatGPT is capable of so much more... you just have to know how to ask! Using principles of prompt engineering, I try to get ChatGPT to make you laugh, while arming you with techniques for overcoming similar supposed limitations, when working with AI.

Speakers Bio: I'm a data-driven, technical marketer who built a 50 person marketing agency (Ladder), and 300k people have taken my online courses (LinkedIn, Udemy, Vexpower). I now work freelance on generative AI projects, and I'm writing a book on Prompt Engineering for (O'Reilly Media).

How technical is this talk?: 2/7

What You'll Learn: AI is capable of a lot more than people realize, and they'd get better results if they learned prompt engineering.

4:30 PM - 5:00 PM EST

LLM economics : The Cost of leveraging Large Language Models
Nikunj Bajaj, TrueFoundry, CEO & Cofounder

Business Strategy

Abstract: Most of us are using LLMs and some of us are getting to the point where LLMs are going to production. Honeymoon phase is going to get over soon and practical realities like cost & maintainability are going to become mainstream.

However, the cost of running LLMs is not well understood or often not put in perspective. In this talk we will dive deep into what type of costs are involved in building LLM based apps. How do these compare when you run RAG vs Fine tuning, what happens when you use Open Source vs Commercial LLMs?
Spoiler- If you wanted to summarize the entire Wikipedia to half its size using GPT-4 8k context window, it would cost a whopping $360K!

Speakers Bio: Nikunj is the co-founder and CEO of TrueFoundry, a platform empowering ML developers to deploy and optimize Language Models. Prior to this role, he served as a Tech Lead for Conversational AI at Meta, where he spearheaded the development of proactive virtual assistants. His team also put Meta's first deep learning model on-device. Nikunj also led the Machine Learning team at Reflektion, where he built an AI platform to enhance search and recommendations for over 600 million users across numerous eCommerce websites. Fun Fact about Nikunj? He learnt scuba diving and swimming in parallel. His instructor laughed at him saying- "You don't know how to swim? And you thought it would be just fine to jump in the middle of Pacific with 70 lb gear on your back

How technical is this talk?: 3/7

What You'll Learn: The highlighted takeaways from our session provide valuable insights specific to LLM pricing. By understanding the economics of different models, businesses and LLM users can make informed decisions tailored to their needs. Comparing costs across models, such as GPT-4, Anthropic Claude V1, InstructGPT-DaVinci, Curie, and self-hosted 7B models, reveals the significant variations in pricing. This knowledge enables organizations to evaluate the trade-off between cost and performance effectively, ensuring optimal resource allocation. Additionally, we'll provide the breakdown of pricing levers, including input and output costs based on the number of tokens, which allows businesses or individuals to analyze their usage patterns and manage expenses efficiently. Understanding the cost implications of fine-tuning models empowers organizations to make strategic decisions based on their specific requirements and budget.

The introduction of TrueFoundry's solutions, such as the compression API for reducing OpenAI costs and simplified deployment of open-source LLMs, offers tangible strategies for optimizing expenses. These unique approaches provide actionable steps that businesses can implement to reduce LLM-related costs while maintaining desired performance levels.

5:00 PM - 5:30 PM EST

How Booking.com is approaching LLMOps: Approaches & Considerations, Pro's & Con's
Sanchit Juneja, Director-Product, Booking.com

Advanced/Technical Research

Abstract: Demarcate how organizations can work with Large Language Models; demarcate pros and cons of each approach, and give future trendlines

Speakers Bio: 17+ years of professional experience across The US, South-east Asia, Africa, and Europe across organizations such as National Instruments, Rocket Internet, AppsFlyer, Gojek, and Booking.com. 12+ years of technical leadership experience, after 5 years as an Applications Engineer. currently working as Director-Product (Big Data & ML/AI) at Booking.com

How technical is this talk?: 4/7

What You'll Learn: How a large organization scales its MLOps ecosystem, and pros and cons of the approaches

5:00 PM - 5:30 PM EST

Lessons Learned Productionising LLMs for Stripe Support
Sophie Daly, Staff Data Scientist, Stripe

Case Study

Abstract: Large Language Models are an especially exciting opportunity for Support Operations: they excel at answering questions, completing sentences, and summarising text while requiring ~100x less training data than the previous generation of models.
In this talk, Sophie discusses lessons learned productionising Stripe's first application of Large Language Modelling - providing answers to user questions for Stripe Support.

Speakers Bio: Sophie is a Staff Data Scientist working on improving user experience and efficiency for Stripe's Support Operations team Stripe. Her favourite thing about being a data scientist is getting to work on a huge variety of business problems, and using analysis and machine learning to solve them. Her hobbies include trying to keep up with the impossibly fast-growing world of LLMs (inside work) and binge watching The Office (outside work).

How technical is this talk?: 4/7

What You'll Learn: Three important lessons to keep top of mind when solving a business problem using LLMs.

5:00 PM
-
5:30 PM EST

Learning from Extremes: What Fraud-Fighting at Scale Can Teach Us About MLOps Across Domains
Greg Kuhlmann, CEO, **Sumatra**

Advanced/Technical Research

Abstract: The engineers behind large-scale anti-fraud platforms, faced with extreme demands for low-latency inference, feature freshness, and agile redeployment, have been the quiet pioneers at the cutting edge of MLOps. One might assume the architectures and practices developed for these intense problems would be overkill in less operationally-demanding domains. However, we will challenge this assumption, and discuss how the real-time-first approach taken by these systems actually simplifies architectures by eliminating many complex pipelines. Further, we'll show how the observability and replay technlogies developed to respond quickly to unpredictable attacks can be applied broadly to make ML teams more agile across the board.

Speakers Bio: Greg Kuhlmann is Co-founder and CEO of Sumatra, a realtime customer data platform that helps growth teams optimize conversions through on-site personalization and recommendations. He formerly led data science teams for the App Store and Apple Pay. He holds a PhD in machine learning from UT Austin.

How technical is this talk?: 5/7

What You'll Learn: - The canonical architecture for modern, large-scale, real-time fraud prevention systems
- A comparison of the "real-time-first" vs. "make-batch-faster" approaches
- How log-time denormalization, unified online/offline feature transformation engines, and backfill on demand, are the keys to rapidly deploying model improvements in non-stationary domains

5:30 PM
-
6:00 PM EST

How to Design and Build Resilient Machine Learning Systems
Dan Shiebler, Head of Machine Learning, Abnormal Security

Advanced/Technical Research

Abstract:
The real world is messy. Systems fail, pipelines break, services go down, engineers push bugs, and users behave erratically. Software is hard exactly because these problems always happen. Effective systems must gracefully handle these events and smoothly degrade without catastrophic failure.

Unfortunately, ML systems are more likely to break than bend. Just like a boxer who only punches a bag will fail in the ring, an ML model that only learns with clean data may fail in production. Most ML models are trained with clean data, and when failures occur feature distributions can shift in ways that the model has never seen during training. This can cause strange and unexpected behavior.

In this talk we will explore how to build resilience into ML systems. We will discuss several types of production-specific risks and how these risks tend to manifest. These risks are common across many domains, but we will primarily use examples from our experience at Abnormal Security to demonstrate how we can detect, mitigate, and overcome these risks.

Speakers Bio: Hi, I'm Dan Shiebler. I like math, history podcasts, fantasy novels, riding my bicycle, and traveling. I live in NYC.

Today I work as the Head of Machine Learning at Abnormal Security. I lead our team of 40+ detection engineers to build AI systems that fight cybercrime. We use a combination of foundational data engineering and advanced ML to detect and remediate cyberattacks. Our technology protects many of the world's largest companies.

Previously, I managed the Web Ads Machine Learning team at Twitter. Before that I worked as a Staff ML Engineer at Twitter Cortex and a Senior Data Scientist at TrueMotion.

I've also spent some time in Academia. My PhD at the University of Oxford focused on applications of Category Theory to Machine Learning (advised by Jeremy Gibbons and Cezar Ionescu). Before that I worked as a Computer Vision Researcher at the Serre Lab.

How technical is this talk?: 5/7

What You'll Learn: How to design machine learning systems that are resilient to the kinds of problems that occur in production systems

6:00 PM
-
6:30 PM EST

Ingredients of a Comprehensive MLOps Program
Mac Macoy, Chick-fil-A, Senior Lead ML Engineer, Chick-fil-A
Lauren Face, Senior Technical Program Lead, Chick-fil-A

Advanced/Technical Research

Abstract: Our industry focuses a lot on leveraging the latest tech to deliver ML models in our businesses, but technology is only one ingredient in a successful MLOps community within an organization. We'll talk about the principles and pillars that make up Chick-fil-A's MLOps Program.

Speakers Bio: Lead Machine Learning Engineer responsible for enabling the ML lifecycle in production

How technical is this talk?: 2/7

What You'll Learn: Foundational principles and practical applications to build an MLOps Program at any organization

OCT 25TH

10:55 AM - 12:25 PM EST

Retrieval Augmented Generation (RAG) with LangChain: "ChatGPT for Your Data" with Open-Source Tools
Dr. Greg Loughnane, Founder & CEO, **AI Makerspace**
Chris Alexluk, Head of LLMs, **AI Makerspace**

Workshop

Abstract: Retrieval Augmented Generation (RAG) - or "ChatGPT for your private data" is the most popular LLM application being built today. RAG systems are question-answering tools that return coherent, fact-checked answers. These answers are used to augment the initial question/prompt before it is fed into an LLM. During this workshop, we will walk through each component of a simple RAG system. You'll learn about how vector stores, embedding models, and LLMs are held together with LLM Ops infrastructure, and you'll get all of the code to do it yourself in no time! We will also evaluation of RAG systems and the emerging best-practices for optimizing the quality of RAG outputs. We will use LangChain tooling to embed own own documents using a [leading embedding model](https://huggingface.co/blog/mteb), which we'll store in a Pinecone vector database, and following context retrieval we'll pass our augmented prompts to Llama 2!

All demo code will be provided via GitHub and/or Colab!

Speakers Bio: - Dr. Greg Loughnane (https://www.linkedin.com/in/gregloughnane/) is the Founder & CEO of AI Makerspace, here he serves as lead instructor for their [LLM Ops: LLMs in Production] (https://maven.com/aimakerspace/llmops) course. Since 2021 he has built and led industry-leading Machine Learning & AI bootcamp programs. Previously, he has worked as an AI product manager, a university professor teaching AI, an AI consultant and startup advisor, and ML researcher. He loves trail running and is based in Dayton, Ohio.

Chris Alexiuk (https://www.linkedin.com/in/csalexiuk/), is the Head of LLMs at AI Makerspace, where he serves as a programming instructor, curriculum developer, and thought leader for their flagship [LLM Ops: LLMs in Production] (https://maven.com/aimakerspace/llmops) course. During the day, he's a Founding Machine Learning Engineer at Ox. He is also a solo YouTube creator, Dungeons & Dragons enthusiast, and is based in Toronto, Canada.

How technical is this talk?: 4/7

What You'll Learn: To understand how to use LangChain to build complex LLM applications.
To build "ChatGPT for their own data"
To understand how LangSmith can be used for productionizing LLM apps.
What "LLM Ops" actually means!

10:55 AM - 11:25 AM EST

Removing the roadblocks to build great GenAI products
Liran Hason, CEO & Co-Founder, **Aporia**

Business Strategy

Abstract: Generative AI represents an exciting new frontier, but deploying successful generative AI products comes with challenges. In this talk, we'll examine common roadblocks organizations face and provide a framework to overcome them. By having the right approach we can accelerate time-to-market and drive impact.

Speakers Bio: Liran Hason is the Co-Founder and CEO of Aporia, the AI Performance Platform. Aporia is trusted by Fortune 500 companies and data science teams in every industry to enable responsible AI and monitor, improve, and scale AI products. Prior to founding Aporia, Liran was an ML Architect at Adallom (acquired by Microsoft), and later an investor at Vertex Ventures. Liran founded Aporia after seeing first-hand the risks of AI without guardrails. In 2022, Forbes named Aporia as the "Next Billion-Dollar Company.

How technical is this talk?: 3/7

What You'll Learn: The audience will learn how to navigate organizational challenges in starting a AI/ML ecosystem at a large company with deeply rooted traditions. This will include pitfalls to avoid, building for your customer, and navigating team topology that works best for your ecosystem.

10:55 AM - 11:25 AM EST

LLMOps: An emerging stack to productionalize LLM Applications
Hien Luu, Head of ML Platform, **DoorDash**

Business Strategy

Abstract: MLOps, as an engineering discipline, has been widely adopted by enterprises to productionalize ML applications. Generative AI and LLMs can bring transformative values to businesses, but is there an emerging stack that can effectively support and operationalize LLM applications? Can MLOps take on those additional responsibilities?

This presentation first argues that the current form of MLOps is insufficient in supporting LLM applications by highlighting the unique set of challenges they bring. These challenges include:
The ambiguity inherent in interacting with LLMs in natural languages
The diverse needs of different types of LLM applications
The new risks associated with incorporating LLM into customer-facing applications.

The presentation will then conclude by providing an overview of an emerging stack called LLMOps.

Speakers Bio: Passionate about the intersection of big data & machine learning. Extensive working experience in designing and building big data applications and scalable web-based applications. Have 3 plus years of technical leadership experience in managing multiple data infrastructure related projects. Passion for architecting scalable and highly available big data applications and systems.

Specialties: Big data, web application framework, cloud computing, RESTful web services and cryptography.

Part-time Passion:
* Instructor at UCSC Extension: Apache Spark, Apache Hadoop, Spring Framework, Design Patterns and Java Comprehensive.

How technical is this talk?: 3/7

What You'll Learn: The unique challenges that LLM applications bring and an emerging LLMOps stack can help w/ those challenges to support and operationalize LLM applications.

10:55 AM - 11:25 AM EST

Evolution of ML Training and Serving Infrastructure at Pinterest
Aayush Mudgal, Senior Machine Learning Engineer, **Pinterest**

Case Study

Abstract: Join us for an insightful talk as we delve into the fascinating evolution of training and serving infrastructure at Pinterest Ads over the past 5+ years. Witness the remarkable progression from logistic regression-based models to the cutting-edge implementation of large transformer-based models, efficiently served using GPU technology. Throughout this transformative journey, we encountered numerous challenges and invaluable lessons that have shaped the very core of this critical business undertaking. Prepare to be inspired by our experiences as we share the triumphs and tribulations that ultimately led to a revolution in Pinterest Ads' capabilities.

Speakers Bio: Aayush Mudgal is a Senior Machine Learning Engineer at Pinterest, currently leading the efforts around Privacy Aware Conversion Modeling. He has a successful track record of starting and executing 0 to 1 projects, including conversion optimization, video ads ranking, landing page optimization, and evolving the ads ranking from GBDT to DNN stack. His expertise is in large-scale recommendation systems, personalization, and ads marketplaces. Before entering the industry, Aayush conducted research on intelligent tutoring systems, developing data-driven feedback to aid students in learning computer programming. He holds a Master's in Computer Science from Columbia University and a Bachelor of Technology in Computer Science from Indian Institute of Technology Kanpur.

How technical is this talk?: 5/7

What You'll Learn:
1. How to best structure training and serving infrastruture.
2. How to balance infrastructure costs and performance
3. Learn from real industrial system serving users at scale and the design choices that were made.

11:30 AM - 12:00 PM EST

Low-latency Model Inference in Finance
Vincent Michael, Senior Director - Machine Learning, ***Capital One***

ML in Production: Implementation, Tooling & Engineering, Data/ML Ops

Abstract: Model Inference at Capital One as across the Fintech sector is a key aspect of the Model Development Lifecycle. In order to reap the benefits of machine learning models trained by Data Scientists, we are required to deploy said models in a production environment. This enables critical business applications, such as credit decisions or fraud detection. Increasingly, we are faced with demanding non-functional requirements for high resilience and low latency service response leading us to invest into service oriented architectures for model inference. Seldon has emerged over the last years a key solution to address these challenges.
In this presentation we will take a close look at the recently released Seldon V2 and our findings for its application for financial services at enterprise scale, comparing it to its V1 specification. We will summarize our findings as they relate to the benefits of novel improvements as well as challenges we see in establishing much needed controls to establish this new release in a highly regulated environment.

Speakers Bio: Vincent David, Senior Director, Machine Learning - Capital One
Vincent is Senior Direct, Machine Learning at Capital One. He is an experienced Machine Learning & Engineering leader with a history of working in Fintech and Entertainment. Strong data and quantitative background, with deep knowledge of complex systems and experiences working in multiple industries. Skilled in Machine Learning, Cloud Engineering. Passionate about using technology to solve high-value business problems.

How technical is this talk?: 6/7

What You'll Learn: Practical learnings, insights, and considerations for more effectively deploying models in a production environment - especially complex production environments.

Why is this takeaway(s) important?: Increasingly, ML experts in large regulated companies are faced with demanding non-functional requirements for high resilience and low latency service response leading us to invest into service oriented architectures for model inference. Seldon has emerged over the last years a key solution to address these challenge -- and this talk is intended to provide real-world and practical advice and best practices for how the audience can apply this to their own field, regardless of where they work or study.

What is unique about this talk: This provides an inside look at how an ML practitioner is addressing this problem in a large-scale, multi-tenant environment within a regulated industry like finance.

11:30 AM - 12:00 PM EST

MLOps on Highly Sensitive Data - Strict Confinement, Confidential Computing and Tokenization Protecting Privacy
Andreea Maciej, AI/ML Product Manager, ***Canonical***

Which talk track does this best fit into?: ML in Production: Implementation, Tooling & Engineering, Data/ML Ops

Abstract:MLOps is used in various organizations that operate on very sensitive datasets. For instance, pharmaceutical and life science companies handle human DNA samples, healthcare institutions training models on patient data, or highly regulated environments like telecom and financial companies. End users, machine learning engineers or data scientists can be concerned about cloud-native workloads would expose them more to software vulnerabilities, data leaks, or any lack of data protection measures. In reality, it's just the opposite. This presentation will cover how you can improve compliance and security with features like Kubernetes strict confinement, blockchain-based tokenization and privacy-enhancing technologies like confidential computing. The talk will feature a case study by a life sciences company that created customized treatments using these technology building blocks. After the talk, you will understand how you can apply them yourself on cloud envorinments MLOps using Kubeflow.

Speakers Bio: Andreea Munteanu, AI/ML Product Manager - Canonical
Andreea Munteanu is a Product Manager at Canonical, leading the MLOps area. With a background in Data Science in various industries, she used AI techniques to enable enterprises to benefit from their initiatives and make data-driven decisions. Nowadays, Andreea is looking to help enterprises to get started with their AI projects and then deploy them to production, using open-source, stable solutions.

Co-speaker, Maciej Mazur

How technical is this talk? 5/7

What You'll Learn: In this talk, we will discuss how you can set up a secure foundation for machine learning with open-source building blocks. We will cover how confidential computing on the public cloud helps you address run time insecurity. You will then learn how Kubernetes' strict confinement helps you get complete isolation, up to a minimum access level to the host resources. Finally, we will cover how tokenization can enable you to avoid data leaks, and allows at the same time achieving high system productivity. We will demonstrate how this works in practice with a life sciences use case powered by Charmed Kubeflow, an open-source community-driven end-to-end MLOps platform.

Why is this takeaway(s) important? MLOps runs often on highly sensitive data and having ways to protect it ensures project succes.

What is unique about this talk: Secure MLOps is still a hot topic - it is being approached, but not often for highly sensitive data, even if industries have this clear need.

12:05 PM - 12:35 PM EST

Supercharging Recommender Systems: Unleashing the Power of Distributed Model Training
Susrutha Gongalla, Principal Machine Learning Engineer, ***Stitch Fix***

Case Study

Abstract: Stitch Fix utilizes a sophisticated multi-tiered recommender system stack, encompassing feature generation, scoring, ranking, and business policy decision-making. This presentation delves into the training architecture of the scoring model, a deep learning model that predicts the likelihood of a user purchasing an item. I will walk through our journey, detailing the transition from training on a single GPU to leveraging multiple GPUs through pytorch's Distributed Data Parallel (DDP) strategy. Additionally, I will share empirical results highlighting the efficiency of GPU utilization as we scale up with DDP across multiple GPUs.

Speakers Bio: Susrutha Gongalla is an experienced Machine Learning Engineer with over 8 years of experience in developing end-to-end machine learning models. She is currently a Principal Machine Learning Engineer in the recommender systems team at Stitch Fix, where she leads the optimization of the end-to-end model lifecycle stack. Her primary focus is on developing and improving the scoring model that generates purchase probabilities for clothing items. Prior to joining Stitch Fix, Susrutha worked as a tech lead at Intuit, where she led the development of recommender systems to improve customer engagement and reduce churn. Besides recommender systems, she also worked on projects using Natural Language Processing techniques to derive insights from unstructured data. Her passion lies in using machine learning to drive business impact and data-driven decision making.
Susrutha holds a Master's degree from Carnegie Mellon University and a Bachelor's degree from Indian Institute of Technology Indore. She has several patents in applications of machine learning and is a recognized leader in the field.

How technical is this talk? 6

What You'll Learn: I would like to show the implementation details of moving from using one GPU to multiple GPUs for model training. I hope this will give the attendees enough knowledge to implement it themselves, allowing them to train bigger (and better) machine learning models.

12:05 PM - 12:35 PM EST

MLOps for Graph-Based Recommender Systems: Orchestrating Intelligent Connections
Bandish Shah, Engineering Manager, Advanced Technical/Research, MosaicML/Databricks

Abstract: Training large AI language models is a challenging task that requires a deep understanding of natural language processing, machine learning, and distributed computing. In this talk, we will go over lessons learned from training models with billions of parameters across hundreds of GPUs. We will discuss the challenges of handling massive amounts of data, designing effective model architectures, optimizing training procedures, and managing computational resources. This talk is suitable for ML researchers, practitioners, and anyone curious about the "sausage making" behind training large language models.

Speaker Bio: Bandish Shah is an Engineering Manager at MosaicML/Databricks, where he focuses on making generative AI training and inference efficient, fast, and accessible by bridging the gap between deep learning, large scale distributed systems and performance computing. Bandish has over a decade of experience building systems for machine learning and enterprise applications. Prior to MosaicML, Bandish held engineering and development roles at SambaNova Systems where he helped develop and ship the first RDU systems from the ground up and Oracle where he worked as an ASIC engineer for SPARC-based enterprise servers.

How technical is this talk? 7/7

What You'll Learn:
- The challenges in training models with billions of parameters across hundreds of GPUs
- Tips and tricks to make the training work and the model to achieve high accuracy

1:35 PM - 3:15 PM EST

Evaluation Techniques for Large Language Models
Rajiv Shah, Machine Learning Engineer, ***Hugging Face***

Which talk track does this best fit into?: Advanced/Technical Research

Abstract: Large language models (LLMs) represent an exciting trend in AI, with many new commercial and open-source models released recently. However, selecting the right LLM for your needs has become increasingly complex. This tutorial provides data scientists and machine learning engineers with practical tools and best practices for evaluating and choosing LLMs.

The tutorial will cover the existing research on the capabilities of LLMs versus small traditional ML models. If an LLM is the best solution, the tutorial covers several techniques, including evaluation suites like the EleutherAI Harness, head-to-head competition approaches, and using LLMs for evaluating other LLMs. The tutorial will also touch on subtle factors that affect evaluation, including role of prompts, tokenization, and requirements for factual accuracy. Finally, a discussion of model bias and ethics will be integrated into the working examples.

Attendees will gain an in-depth understanding of LLM evaluation tradeoffs and methods. Jupyter Notebooks will provide reusable code for each technique discussed.

Speakers Bio: Rajiv Shah is a machine learning engineer at Hugging Face who focuses on enabling enterprise teams to succeed with AI. Rajiv is a leading expert in the practical application of AI. Previously, he led data science enablement efforts across hundreds of data scientists at DataRobot. He was also a part of data science teams at Snorkel AI, Caterpillar, and State Farm.

Rajiv is a widely recognized speaker on AI, published over 20 research papers, been cited over 1000 times, and received over 20 patents. His recent work in AI covers topics such as sports analytics, deep learning, and interpretability.

Rajiv holds a PhD in Communications and a Juris Doctor from the University of Illinois at Urbana Champaign. While earning his degrees, he received a fellowship in Digital Government from the John F. Kennedy School of Government at Harvard University. He has recently started making short videos, @rajistics, with several million views.

How technical is this talk? 5/7

What You'll Learn: Ways to quickly start evaluating models

1:35 PM
-
2:05 PM EST

Operationalizing data-centric AI: Practical algorithms + software to quickly improve ML datasets
Jonas Mueller, Chief Scientist, ***Cleanlab***

Which talk track does this best fit into?: ML in Production: Implementation, Tooling & Engineering, Data/ML Ops

Abstract: In applied ML projects, experienced data scientists know that improving data brings higher ROI than tinkering with models. However the process of finding and fixing problems in a dataset is highly manual (ad hoc ideas explored in Jupyter notebooks). Cleanlab develops open-source software to help make this process more: efficient (via novel algorithms that automatically detect certain issues in data) and systematic (with better coverage to detect different types of issues).
This talk will describe how high-level ideas from data-centric AI can be operationalized across a wide variety of datasets (image, text, tabular, etc). I will introduce novel algorithmic strategies to automatically identify various issues in data that we have researched and published papers on with extensive benchmarks. These include detection of label errors, bad data annotators, out-of-distribution examples, and other dataset problems that once identified can be easily addressed to significantly improve trained models. Thousands of data scientists have started using this sort of data-centric AI software, and results from a few case studies will be presented. I will conclude with a discussion of where the data-centric AI movement is headed next, and key obstacles that deserve more attention.

Speakers Bio: Jonas Mueller, Chief Scientist at Cleanlab
Jonas Mueller is Chief Scientist and Co-Founder at Cleanlab, a software company providing data-centric AI tools to efficiently improve ML datasets. Previously, he was a senior scientist at Amazon Web Services developing AutoML and Deep Learning algorithms which now power ML applications at hundreds of the world's largest companies. In 2018, he completed his PhD in Machine Learning at MIT, also doing research in NLP, Statistics, and Computational Biology. Jonas has published over 30 papers in top ML and Data Science venues (NeurIPS, ICML, ICLR, AAAI, JASA, Annals of Statistics, etc). This research has been featured in Wired, VentureBeat, Technology Review, World Economic Forum, and other media. He loves contributing to open-source, and helped create the fastest-growing open-source software for AutoML (https://github.com/awslabs/autogluon) and Data-Centric AI (https://github.com/cleanlab/cleanlab). An avid educator, he also taught the first-ever course on data-centric AI at MIT: https://dcai.csail.mit.edu/

How technical is this talk? 5/7

What You'll Learn: How to best practice data-centric AI in real-world ML projects. This covers automated methods to check the dataset for various issues common in ML data as well as how to efficiently address the issues to improve the dataset and subsequent ML model. I will cover novel algorithms invented by our research team and case studies which showcase the benefits of data-centric AI in real-world ML applications.
The intended audience is folks with experience in supervised learning who want to develop the most effective ML for messy, real-world applications. Some of the content will be technical, but not require a deep understanding of how particular ML algorithms/model work (having completed one previous ML course/project should suffice). The topics should be of interest to anybody working in: computer vision, natural language processing, audio/speech or tabular data, and other standard supervised learning applications, as well as DataOps folks.

Why is this takeaway(s) important? While there is a lot of buzz about data-centric AI, there has not been nearly as much scientific innovation in practical algorithms for data quality improvement. This talk will share new algorithms that are simple and highly effective to improve ML datasets, developed through over the past year by our researchers. This not a research talk though, I will showcase how these methods are implemented in real ML projects and the impact they have had.

What is unique about this talk: Cleanlab is the most popular open-source tool for data-centric AI, but a talk like this demonstrating its various functionalities for improving different types of data cannot be found anywhere online. This talk combines how the methods work, why they are useful, and how to apply them together in practice, in a way that has never been presented before. Many of the methods presented are brand new too having only been published months prior.

1:35 PM - 1:40 PM EST

The Evolution of ML Monitoring in Production: From ML 1.0 to LLMs
Gon Rappaport, MLOps Solutions Architect, **Aporia**

Abstract: In this talk, we'll explore the landscape of ML Monitoring in production, highlighting best practices for tracking real-world AI products like Recommender Systems and Fraud Detection models ("ML 1.0") and LLM-based applications such as Chatbots. We'll further explore the intricacies of LLMs, focusing on monitoring applications that use Retrieval Augmented Generation (enriching LLMs with external insights), LLMs serving as controllers capable of activating external APIs or other ML models, as well as LLM Guardrails.

Speaker Bio: Gon Rappaport is an MLOps Solutions Architect at Aporia, the AI Performance Platform, providing Observability and Guardrails to drive responsible, high-performing AI products. A software engineer at heart, Gon works closely with ML teams in every industry, seamlessly blending his knack for meme-based executive reports with his commitment to overcoming the visible and hidden challenges of machine learning models in production.

1:40 PM - 1:45 PM EST

LLMs from hallucinations to relevant responses
Eddie Mattia, Data Scientist, **Outerbounds**

Abstract: Present a taxonomy of methods for controlling LLMs. Listeners will learn the broad strokes of how apps based on retrieval-augmented generation (RAG) and instruction tuning work and where they fit into the big picture of generative AI. We focus on how these techniques can be used to make generated responses more relevant.

Speaker Bio: Eddie Mattia is a data scientist working on Metaflow and foundation models at Outerbounds. He began using Python to teach applied math in grad school. Since then, Eddie has worked at startups and at Intel building machine learning software.

1:50 PM - 1:55 PM EST

Supporting Community Competitions to Develop LLMs
Rick Izzo, Tech Lead, **Lighting AI**

Abstract: Lightning AI Supported the NEURIPS LLM Challenge by providing the base LIT-GPT model upon which contestants were asked to improve training of & finetune. This talk describes what it's like to support a supporting a community competition focused on llm efficiency; From setting up the leaderboards, to creating a model challenge that allows participants to demonstrate that you can meaningfully finetune a model on a single GPU (using quantization) in one day, and get to the best evaluation possible.

Speaker Bio: Rick Izzo was a Ph.D. candidate at the University at Buffalo Endovascular Device Development laboratory before co-founding Tensor[werk] Inc, a startup focused on building core infrastructure to support ML model training & deployment. After developing RedisAI & Hangar (tensor storage & version control system), Tensor[werk] was acquired by Lighting AI, where he and the team have continued to build and improve the core infrastructure needed to develop and train ML/DL Models.

1:55 PM - 2:00 PM EST

Generative AI: the open source way
Andreea Munteanu, MLOps Product Manager, **Canonical**

Abstract:
Generative AI is probably the topic of the year. Leaders across the world feel the pressure of missed opportunities related to the latest technology. Professionals are also left worried about the impact that latest technology will have on their roles. Data scientists and machine learning engineers are challenge and need to quickly upskill. With such a stretched picture in mind, will generative AI deliver up to the great promises that it has right now?
This lighting talk will depict the opportunities that open source gives to generative AI to accelerate innovation. Between anxiety and enthusiasms, the latest technologies bring a new angle to the market. Let's learn together about genAI with open source: from models to tooling to applications

Speaker Bio: Andreea is a Product Manager at Canonical, leading the MLOps area. With a background in Data Science in various industries, such as retail or telecommunications, Andreea used AI techniques to enable enterprises to benefit from their initiatives and make data-driven decisions. Andreea is looking to help enterprises get started with their AI projects and then deploy them to production, using open-source, secure, stable solutions.

2:00 PM - 2:05 PM EST

LLMs, Big Data, and Audio: Breaching an Untapped Gold Mine
Jose Nicholas Francisco, ML Developer Advocate, **Deepgram**

Abstract: Large language models like those in the GPT and Llama series are primarily trained on massive amounts of *text* data. However, the vast majority of language and communication doesn't take place over text, but rather through voice. Cues in vocal tone carry information that the plaintext cannot convey—think about the last time you've witnessed or experienced a miscommunication over text/email/Slack. Thus, in this talk, I argue that training language models on audio data is the next step to improving them. Then, I'll propose a way of integrating audio data with text data in a larger dataset that can then be used for training various LLMs.

Speaker Bio: Jose is a Developer Advocate at Deepgram, aiming to demystify the inner workings of AI. He has a background in software engineering, with projects focused on fraud detection and prevention. Jose earned a bachelor's and master's degree in computer science—with a specialization on AI and NLP—from Stanford University. He currently lives in San Francisco with his friends.

2:10 PM - 2:40 PM EST

How many Labelled Examples do you need for a BERT-sized Model to Beat GPT4 on Predictive Tasks?
*Matthew Honnibal, Founder and CTO, **Explosion AI***

Which talk track does this best fit into?: Advanced/Technical Research

Abstract: Large Language Models (LLMs) offer a new machine learning interaction paradigm: in-context learning. This approach is clearly much better than approaches that rely on explicit labelled data for a wide variety of generative tasks (e.g. summarisation, question answering, paraphrasing). In-context learning can also be applied to predictive tasks such as text categorization and entity recognition, with few or no labelled exemplars.
But how does in-context learning actually compare to supervised approaches on those tasks? The key advantage is you need less data, but how many labelled examples do you need on different problems before a BERT-sized model can beat GPT4 in accuracy?

The answer might surprise you: models with fewer than 1b parameters are actually very good at classic predictive NLP, while in-context learning struggles on many problem shapes --- especially tasks with many labels or that require structured prediction. Methods of improving in-context learning accuracy involve increasing trade-offs of speed for accuracy, suggesting that distillation and LLM-guided annotation will be the most practical approaches. Implementation of this approach is discussed with reference to the spaCy open-source library and the Prodigy annotation tool.

Speakers Bio:
Major work includes:
* ExplosionAI GmbH: Founder and CTO
* spaCy: Open-source NLP library in use by thousands of companies, with over 100m downloads. Particularly known for efficiency and API design.
* Prodigy: Developer-focussed annotation tool, with active learning and scriptability features. Licenses purchases by almost 1000 companies.
* Thinc: Open-source ML library built for spaCy, designed around function composition.
Entered the field in 2005 as a linguist, transitioning towards computer science over PhD and post-doctoral research. Left academia in 2014. Originally from Sydney, now in Berlin.

How technical is this talk? 5/7

What You'll Learn: On predictive tasks, LLMs currently perform much worse than you'd expect if you benchmark them directly against other approaches. If you're working on a task where it's not worth putting 20-40 hours of effort into the model, just use an LLM. If it's a problem worth doing well, prototype with the LLM, but also create training and evaluation data, and compare approaches.

2:10 PM - 2:40 PM EST

Fine-tune LLMs or Integrate 3rd party APIs? A financial Case-study
Hannes Hapke, Principal Machine Learning Engineer**, *Digits***

Which talk track does this best fit into?: Case Study

Abstract: "Almost a year ago, with the introduction of ChatGPT and, subsequently GPT-4, the sphere of machine learning transformed completely. These advancements and LLMs unlocked the capability to address previously unsolvable problems.

In this talk, Hannes explains how Digits' machine learning team has adapted to the new world of LLMs, how their MLOps processes have changed, and the team's learning around fine-tuning and deploying LLMs as part of a small, highly focused ML team. He will also discuss the key questions you must ask to determine if you should fine-tune open-source LLMs or integrate with a 3rd party API, the challenges and ethical concerns of using advanced language models via APIs, and how these risks in your projects can be mitigated through engineering."

Speakers Bio: "As one of Digits' principal machine learning engineers, Hannes Hapke is developing innovative machine learning systems to give accountants and business owners real-time insights into their businesses. Before joining Digits, Hannes solved machine learning infrastructure problems in various industries, including healthcare, retail, recruiting, and renewable energies.

Hannes is an active contributor to TensorFlow's TFX Addons project and has co-authored multiple machine learning publications, including the book ""Building Machine Learning Pipelines"" by O'Reilly Media. He has also presented state-of-the-art ML work at conferences like Google Developer Connect. He is excited about the recent developments around Large Language Models and ML Engineering."

How technical is this talk? 4/7

What You'll Learn: "* When to use model APIs and when to avoid them
* When to fine tune LLMs
* How to deploy LLMs effectively "

2:45 PM - 3:15 PM EST

Your AI applications need Guardrails: Here's how to build them
Shreya Rajpal, Founder, **Guardrails AI**

Which talk track does this best fit into?: Advanced/Technical Research

Abstract:Large Language Models (LLMs) such as ChatGPT have revolutionized AI applications, offering unprecedented potential for complex real-world scenarios. However, fully harnessing this potential comes with unique challenges such as model brittleness and the need for consistent, accurate outputs. These hurdles become more pronounced when developing production-grade applications that utilize LLMs as a software abstraction layer.

In this talk, we will tackle these challenges head-on. We introduce Guardrails AI, an open-source platform designed to mitigate risks and enhance the safety and efficiency of LLMs. We will delve into specific techniques and advanced control mechanisms that enable developers to optimize model performance effectively. Furthermore, we will explore how implementing these safeguards can significantly improve the development process of LLMs, ultimately leading to safer, more reliable, and robust real-world AI applications.

About the spaker: Shreya Rajpal is the creator and maintainer of Guardrails AI, an open source platform developed to ensure increased safety, reliability, and robustness of large language models in real-world applications. Her expertise spans a decade in the field of machine learning and AI. Most recently, she was the founding engineer at Predibase, where she led the ML infrastructure team. In earlier roles, she was part of the cross-functional ML team within Apple's Special Projects Group and developed computer vision models for autonomous driving perception systems at Drive.ai.

How technical is this talk?: 4/7

What You'll Learn: The audience will learn about the challenges and risks associated with Large Language Models (LLMs) and how the open-source platform, Guardrails AI, addresses these issues by providing specific techniques and advanced control mechanisms to optimize model performance, leading to safer, more reliable, and robust real-world AI applications.

Why is this takeaway(s) important?: It demonstrates how guardrails can provide developers with tools to optimize model performance, resulting in safer and more reliable real-world AI applications.

What is unique about this talk: This talk will cover implementing end-to-end guardrails and safety interventions for Large Language Models (LLMs). These measures will promote a comprehensive approach to evaluating safety-first LLM development.

3:45 PM - 4:45 PM EST

Build your own ChatGPT with open source tooling
Andreea Munteanu, MLOps product manager, **Canonical**

Workshop

Abstract: LLMs are gaining huge popularity with projects such ChatGPT, LLaMA or PaLM being open to everyone. Yet, enterprises feel overwhelmed by the large number of applications that requires, at first sight, a complete reshuffle of the existing infrastructure, in order to accommodate the needs for powerful cloud computing.

Depending on the industry, there are various use cases, as well as legacy architectures in place. Generative AI, however, can be deployed on any environment, whether it ia a public or private cloud. From the very beginning, when doing estimations of the cluster size, to upper layers in the stack where inference infrastructure is considered, there are a bunch of key factors such as GPU types, MLOps tooling or artefact types that influence the infrastructure of the project.

This talk will cover how you can build your infrastructure for a generative AI project, with a focus on building your own conversational assistant. It will go through the entire stack, including hardware and software applications, that cover the entire machine learning lifecycle, focusing on open source tooling and models. The session will feature a case study by a finserv company that build their own chatbot, using their data. After this presentation, you will understand how you can build and automate your infrastructure for a genAI project, using open source tooling.

About the spaker: I am a Product Manager at Canonical, leading the MLOps area. With a background in Data Science in various industries, such as retail or telecommunications, I used AI techniques to enable enterprises to benefit from their initiatives and make data-driven decisions. I am looking to help enterprises get started with their AI projects and then deploy them to production, using open-source, secure, stable solutions.

I am a driven professional, passionate about machine learning and open source. I always look for opportunities to improve, both myself and people within the teams that I am part of. I enjoy sharing my knowledge, mentoring young professionals and having an educational impact in the industry.

How technical is this talk?: 5/7

What You'll Learn: fine tuning LLMs using open source tooling

3:45 PM - 4:45 PM EST

Production ML Serving & Monitoring in Kubernetes
Andrew Willson, Head of Customer Success, Seldon

Workshop

Abstract: This talk offers a practical guide to building a state-of-the-art MLOps deployment platform using Kubernetes, with a focus on deploying deep learning models. Attendees will gain insights into the integration of key technologies including NVIDIA Triton Inference Server, Seldon Core v2, Kafka, Prometheus, and Grafana. The session covers an end-to-end workflow for serving complex models like transformers and CNNs and configuring monitoring on top. The knowledge shared will be valuable for those looking to enhance model performance, reduce costs, and unlock new use cases in machine learning.

Speakers Bio: Andrew is the Head of Customer Success at Seldon, enabling enterprise customers to get the most out of the Seldon ecosystem of tools. Andrew has spent the last 10 years in technical, client-facing roles with a focus on data, machine learning, and architecture. He loves transforming complex and ambiguous topics into clear actions that drive value within an organization. Outside of work, you might find Andrew riding his bike around the streets of London or snowboarding somewhere in The Alps.

How technical is this talk? 5/7

What You'll Learn: By attending this talk, the audience will learn how to set up a state-of-the art MLOps deployment platform on top of Kubernetes. Once it is up and running, they will learn how to serve deep learning models in a robust and scalable way. As inference is being performed, the audience will see how they can configure rich monitoring to observe their models in production.

Along the way, the audience will learn about each of the technologies being presented, including: Kubernetes, PyTorch, NVIDIA Triton Inference Server, Seldon Core v2, Kafka, Prometheus, and Grafana. More importantly, they will see how these tools integrate with one another to form a fully fledged MLOps deployment platform.

MLOps deployment platforms are very challenging systems to build, deploy, and operate in the real-world. As more and more business value is being driven by large, complex, deep learning (e.g. transformer, diffusion, CNN) models, it is crucial to be able to serve these in a scalable way. This type of a platform creates business value by: (1) reducing inference costs through efficient serving, (2) improving performance through continuous monitoring and deployment, (3) enhancing the user experience with low latency, (4) shortening the time to bring use cases to market, and (5) eliminating the up-front development costs of building a deployment platform.

Aside from benefitting the business, engineers can benefit knowing that they are using best-in-breed tools, that they have a consistent way of deploying ML models, and that they have monitoring and logging built in. Enabling engineers to build complex, data-centric pipelines opens up a new set of use cases, including advanced recommendation engines, explainable machine vision models, and even LLM question-and-answer applications.

3:45 PM - 4:15 PM EST

From Prototype to Product: Rapid iteration and ML model deployment at Dropbox
Richie Frost, Software Engineer, **Dropbox**
Yilin Ye, **Dropbox**

Which talk track does this best fit into?: Case Study

Speaker Bio: Richie Frost is a software engineer on the ML Foundations team at Dropbox, specializing in rapid iteration on ML inference platforms. Previously, he worked as a data scientist and software engineer at Microsoft, where he built and analyzed innovative AI-based solutions.

Abstract: As AI rapidly evolves, organizations must keep pace by iterating on ML models quickly. In the MLOps field, we face several challenges that make it hard to prototype and deploy quickly. In this presentation, we share how we leverage a mix of open-source tools and our own solutions at Dropbox to achieve faster iteration speeds, reducing prototyping and deployment times from weeks to under an hour.

The session will explore how to establish an end-to-end system to achieve rapid prototyping and deployment while maintaining security best practices. Attendees will gain insights into the integration of open source frameworks, such as TorchServe, Triton, KServe, and Kubernetes, along with internal tools to accelerate the process. The discussion will also delve into the challenges of iterating on ML models and their deployments and explore strategies for optimizing resource allocation, managing dependencies, and automating deployment processes.

How technical is this talk?: 4/7

What You'll Learn: How to develop a system for easily prototyping and deploying models in production, including LLMs

Why is this takeaway(s) important?: Many organizations struggle to implement a solid strategy for getting from prototype to production, especially with third-party models. Not only is this useful for iterating on in-house models, it enables teams to quickly spin up the latest and greatest publicly-available third-party LLMs so that they can stay ahead of the curve in a rapidly evolving AI ecosystem.

What is unique about this talk: This is an actual use case of how this has been done in production at a large scale company (700m users), providing for security implications as well as business and cross-team needs.

3:45 PM - 4:15 PM EST

Using Scouter Models to monitor Model Drift - A novel approach
Kumaran Ponnambalam, Principal Engineer - AI, **Cisco Systems Inc., Emerging Tech & Incubation**

Which talk track does this best fit into?: Advanced/Technical Research

Speaker Bio: Kumaran Ponnambalam is a technology leader with 20+ years of experience in AI, Big Data, Data Processing & Analytics. His focus is on creating robust, scalable AI models and services to drive effective business solutions. He is currently leading AI initiatives in the Emerging Technologies & Incubation Group in Cisco. In this role he is focused on building MLOps and Observability services to enable ML. In his previous roles, he has built data pipelines, analytics, integrations, and conversational bots around customer engagement. He has also authored several courses on the LinkedIn Learning Platform in AI and Big Data.

Abstract: Model Drift monitoring is a critical activity in the MLOps life cycle. Models in production may drift and decay due to several reasons. Identifying drift quickly and taking corrective action is critical for continued success of models. A severe limitation in drift monitoring in a number of cases, is the lack of true labels for production data. How do we monitor drift when true labels are not available? We built a technique called Scouter models, to identify concept drift, when true labels are not available. In this talk, we will talk about what a scouter model is, how to identify the right scouter feature and best practices in building and managing such models. This would be of great interest to data scientists and MLOps engineers.
How technical is this talk?: /7

What You'll Learn: Scouter model concepts, Building scouter models and best practices for managing them.

Why is this takeaway(s) important?: MLOps teams are severely limited by lack of true labels in production, to monitor model drift, as drift compares the predictions with corresponding true labels. This technique provides an interesting alternative in such cases.

What is unique about this talk: This is research/development that we worked on at Cisco Emerging Tech & Incubation

3:45 PM - 4:15 PM EST

Using Scouter Models to monitor Model Drift - A novel approach

Muller Mu, Solution Architect / Senior Scientist, **Roche & Scientific**
Naiel Samaan, Senior Product Owner, AI Platform, **Ford Motor Company**
Valmir Bucaj, AI/ML Platform Product Owner, **Ford Motor Company**

Business Strategy

Speaker Bio: Having amassed extensive experience in digitalization across different industries, Le (Muller) Mu is combining his IT experience with his knowledge in molecular biology to help accelerate the drug discovery process in Roche Pharma Research and Early Development (pRED). He is currently the tech lead on the Roche pRED MLOps Service team, driving the operationalization of many different machine learning models, which helps to bring better future medicines faster into the hands of the patients.

-

As a seasoned product, people, and project manager, I thrive on tackling complex problems with a combination of data, analysis, leadership, and creativity. Currently, I lead a team of data scientists, tech leads, and software developers in building an enterprise machine learning operations platform for Ford Motor Company.

My expertise lies in leveraging data and machine learning, along with innovative ideas, to solve business challenges. I've successfully managed every aspect of the process, from evaluating customer needs to informing product development to leading cross-functional teams to success.

My passion for developing teams, combined with my entrepreneurial drive and expertise in data and technology, make me the ideal candidate for any company looking to solve complex challenges and achieve growth.

-

"Valmir Bucaj is a technical leader and product owner of Ford's AI/ML Platform. He has 3+ years of experienced in leading cross-functional teams, who specialize in building multi-cloud enterprise MLOps frameworks to help ML Engineers and Data Scientists scale and productionize their machine learning projects faster and more efficiently. Valmir is passionate in building AI/ML products that customers both need and love. He previously used to work as a machine learning engineer, focusing on graph neural networks for social recommendations.

Valmir used to also work as an assistant professor of mathematics at West Point, where among other things, he designed and taught the Academy's first Data Science course for their newly established major.

Valmir holds a PhD in Mathematics, from Rice University."

Abstract: Model Drift monitoring is a critical activity in the MLOps life cycle. Models in production may drift and decay due to several reasons. Identifying drift quickly and taking corrective action is critical for continued success of models. A severe limitation in drift monitoring in a number of cases, is the lack of true labels for production data. How do we monitor drift when true labels are not available? We built a technique called Scouter models, to identify concept drift, when true labels are not available. In this talk, we will talk about what a scouter model is, how to identify the right scouter feature and best practices in building and managing such models. This would be of great interest to data scientists and MLOps engineers.

How technical is this talk?: 6/7

What You'll Learn: Scouter model concepts, Building scouter models and best practices for managing them.

OCT 26TH

9:45 AM - 10:25 AM EST

Is it too much to ask for a stable baseline?
D. Sculley, CEO, **Kaggle**

Advanced Technical/Research

Abstract: Evaluation and monitoring are the heart of any reliable machine learning system. But finding a stable reference point, a reliable comparison baseline, or even a decent performance metric can be surprisingly difficult in a world that is beset by changing conditions, feedback loops, and shifting distributions. In this talk, we will look at some of the ways that these conditions show up in more traditional settings like click through prediction, and then see how they might reappear in the emerging world of productionized LLMs and generative models.

Speaker Bio: D. is currently CEO of Kaggle and GM of 3P ML Ecosystems at Google. Prior to this role, he was a director of engineering in the Google Brain team, leading research teams working on robust, responsible, reliable and efficient ML and AI. During his 15 years at Google, he has worked on nearly every aspect of machine learning, and have led both product and research teams including those on some of the most challenging business problems. His work on machine learning and technical debt helped lay the foundation for the field of MLOps, and the book Reliable Machine Learning was named Best MLOps Book of 2022.

How technical is this talk? 3/7

What You'll Learn: Evaluation is hard, but not impossible, and with enough care we can probably say something useful about our models.

10:55 AM - 11:25 AM EST

Getting higher ROI on MLOps initiatives: five lessons learned while building out the MLOps Platform for 100+ Data Scientists
Stefan Krawczyk, CEO & Co-founder, **DAGWorks Inc.**

Business Strategy

Abstract: MLOps is hard, because there's so many "things" that you might want to integrate and connect with: A/B testing, feature stores, model registries, data catalogs, lineage systems, python dependencies, machine learning libraries, LLM APIs, orchestration systems, online vs offline systems, speculative business ideas, etc. In this talk I'll cover five lessons that I learned while building out the self-service MLOps platform for over 100 data scientists at Stitch Fix. This talk is for anyone building their own, or buying it all off the shelf. Either way you're still going to want everything to fit cohesively together, i.e. as a platform, and learning what to avoid/focus on will increase your ROI on MLOps initiatives.

Speaker Bio: A hands-on leader and Silicon Valley veteran, Stefan has spent over 15 years thinking about data and machine learning systems, building product applications and infrastructure at places like Stanford, Honda Research, LinkedIn, Nextdoor, Idibon, and Stitch Fix. A regular conference speaker, Stefan has guest lectured at Stanford's Machine Learning Systems Design course and is an author of a popular open source framework called Hamilton. Stefan is currently CEO of DAGWorks, an open source startup that is enabling teams a standardized way to build and maintain data, ML and LLM pipelines without the coding nightmares.

How technical is this talk? 4/7

What You'll Learn: Five lessons that will help them with MLOps initiatives:
1. Build for immediate adoption.
2. Don't build for every user equally. Let those with stronger SWE skills do more themselves.
3. Don't give users direct access to vendor/cloud APIs.
4. Take time to ensure you can live in the shoes of your users.
5. Provide two layers of APIs to keep your development nimble: a foundational layer, and then an opinionated higher level layer.

10:55 AM - 11:25 AM EST

Gen AI and Lightning: Accelerating AI Innovation while Ensuring Scalability and Security
William Falcon, Founder and CEO, **Lighting AI**

Track: Advanced Technical/Research

Abstract: "As artificial intelligence (AI) continues to reshape industries and drive transformative advancements, the need for accelerated AI innovation has become paramount. Organizations across the globe are seeking novel ways to leverage AI technologies to gain a competitive edge, optimize processes, and deliver enhanced customer experiences. However, as AI applications grow in complexity and scale, so do the challenges related to scalability and security.

This talk aims to address the critical aspects of accelerating Enterprise AI innovation and will delve into the following key topics:

- Harnessing the Power of PyTorch and PyTorch Lightning: We will explore the benefits of modern AI frameworks, such as PyTorch, PyTorch Lightning, and Fabric, and the techniques that facilitate rapid development and deployment of AI solutions.
- Scalability: The exponential growth of data and increasing demands for AI applications necessitate scalable architectures. We will discuss strategies for designing AI systems that can effortlessly handle massive datasets and adapt to the evolving requirements of AI-driven applications.
- Security Considerations in AI: The integration of AI technologies introduces new security risks. This talk will highlight best practices for safeguarding AI systems against potential threats, ensuring data privacy, and maintaining compliance.
- Collaborative Ecosystems for Innovation: Accelerating AI innovation requires collaboration between diverse stakeholders, including researchers, developers, policymakers, and businesses. We will explore successful collaborative models that foster innovation.

Speaker Bio: William Falcon is the creator of PyTorch Lightning, a deep learning framework. He is also the founder and CEO of Lightning AI, and was previously a co-founder and CTO of NextGenVest. He began working on these projects while completing a Ph.D. at NYU, which was funded by Google DeepMind and the NSF. Additionally, he worked as a researcher at Facebook AI and at Goldman Sachs.

How technical is this talk? 5/7

11:30 AM - 12:00 PM EST

Creating the World’s Premier Biological Foundation Model
Jess Leung, Staff ML Engineer, Recursion

Case Study

Abstract: Recursion has built the world's most expansive biological foundation model, boasting billions of parameters and trained on hundreds of millions of high-resolution images. Dive into an exploration of Recursion's groundbreaking approach that is reshaping and revolutionizing the drug discovery process. This talk will provide an in-depth look at the infrastructure and tooling necessary for building such models. We'll share insights into the intricate process of efficient data management, large-scale model training, scaling inference, and effective use of our in-house supercomputer and public cloud environments.

Speaker Bio: Jess Leung has been shipping machine learning to production throughout their career. They are currently a Staff Machine Learning Engineer at Recursion, where they lead the ML Platform team. Prior to Recursion, Jess has held technical leadership where they have shipped products in a wide variety of domains including: internet-scale platforms, e-commerce, life science solutions, public transportation services, and financial systems. Jess holds a B.Sci in Electrical Engineering from Queen's University.

How technical is this talk? 5/7

What You'll Learn: What it takes (infrastructure, techniques, talent, culture and practices) that is involved in training large deep learning models

11:30 AM - 12:00 PM EST

Lessons Learned from Implementing GenAI at Large Enterprises
Dr. Ilyas Iyoob, *Faculty*, ***University of Texas***

Business Strategy

Abstract: In the rapidly evolving landscape of GenAI, large US enterprises face unique challenges when considering its implementation. Beyond the well-acknowledged concerns of data privacy, security, bias, and regulatory compliance, our journey in executing GenAI within mission-critical applications has revealed additional complexities. In this session we will walk through a number of real examples of failed implementations and the lessons learned from them.

Speakers Bio: Dr. Ilyas Iyoob is chief data scientist and global head of Research at Kyndryl. He pioneered the seamless interaction between machine learning and operations research in the fields of autonomous computing, fintech, and blockchain. As a successful entrepreneur at Gravitant, a start-up focused on optimizing the cloud journey, he helped build and sell the company to IBM in 2016. Dr. Iyoob currently advises over a dozen venture funded companies and serves on the faculty of the Cockrell School of Engineering at the University of Texas at Austin. He has earned a number of patents and industry recognition for cloud intelligence and was awarded the prestigious World Mechanics prize by the University of London.

How technical is this talk? 3/7

What You'll Learn: Learn from mistakes; new best practices for successful implementation of GenAI

11:30 AM - 12:00 PM EST

End-to-end Production Machine Learning Workflows
Goku Mohandas, *ML Lead*, ***Anyscale***

Which talk track does this best fit into?: Advanced Technical/Research

Abstract: We'll start by breaking down the machine learning development lifecycle into experimentation (design + develop) and production (deploy + iterate). We'll walk through the best practices for developing and executing ML workloads and iteratively build a production workflow around it (manual to CI/CD to continual learning). We'll also take a look at the biggest obstacles in the way of taking machine learning in production (standardization, integrations, scaling workloads in Python, dev to prod transition, reliability, etc.) While our specific use case will be fine-tuning an LLM for a supervised NLP use case, all the content will easily extend to any algorithm (regression to LLMs), application (NLP, CV, tabular, etc.), tooling stacks and scales.

Speakers Bio: "I've spent my career developing ML applications across all scales and industries. Specifically over the last four years (through Made With ML), I’ve had the opportunity to help dozens of F500 companies + startups build out their ML platforms and launch high-impact ML applications on top of them. I started Made With ML to address the gaps in education and share the best practices on how to deliver value with ML in production.
While this was an amazing experience, it was also a humbling one because there were obstacles around scale, integrations and productionization that I that I didn’t have great solutions for. So, I decided to join a team that has been addressing these precise obstacles with some of the best ML teams in the world and has an even bigger vision I could stand behind. So I'm excited to announce that Made With ML is now part of Anyscale to accelerate the path towards production ML."

How technical is this talk? 4/7

What are some of the infrastructure you plan to discuss?: Easily executing ML workloads in a distributed fashion that scales beyond the constraints of a single machine.

What are some of the languages you plan to discuss? Python

What kind of Dev ops tools you plan to discuss? Open source? GitHub Actions

What You'll Learn: "- Learn the best practices for developing ML workloads (data, train, tune, serve, etc.)
- Learn how to incorporate MLOps concepts to our data science work (experiment tracking, testing code, data + models, monitoring, etc.)
- Learn how to fine-tune an LLM for a supervised use case
- Learn how to iteratively put ML into production (manual deployment, CI/CD, continual learning)
- Learn how to execute ML workloads across multiple machines easily"

12:05 PM - 12:35 PM EST

Building Computers for AI
*David Bennett, Chief Customer Officer, **Tenstorrent***

Abstract: Tenstorrent is an innovative hardware architecture designed to enhance the efficiency and scalability of artificial intelligence (AI) workloads. Addressing the compute-intensive nature of modern AI applications, Tenstorrent leverages a unique grid-based architecture to enable efficient execution of both sparse and dense computations. Its dynamic code generation and execution capabilities allow flexibility across various AI models and algorithms. Furthermore, Tenstorrent emphasizes scalability, with the potential to deploy individual units in massive, interconnected networks for larger AI tasks. The system is designed to optimize power efficiency and performance, making it an ideal solution for advanced machine-learning jobs. Architectures like Tenstorrent will be crucial in harnessing its full potential as AI drives technological progress.

Speakers Bio: I am currently the Chief Customer Officer (CCO) for Tenstorrent Inc., a Toronto based unicorn developing software, silicon and systems to run AI and ML faster than anyone else. We are also developing a full line up of RISC-V CPUs.
I recently left Tokyo, Japan where I ran Lenovo's Japan operations including both the Lenovo brand and market leading NEC Personal Computer; the leader in Consumer, Commercial and Enterprise solutions, Lenovo is the largest PC manufacturer in Japan.
Previously, I ran AMD's Asia based Global MNC accounts and was the Asia Pacific and Japan Mega Region Vice President where the team has achieved 12 quarters of consecutive YoY revenue and share growth. I also lead a global team driving >$600M a year in Commercial Client, Thin Client and Server CPU and Graphics processor revenue. We focus on an end-to-sale beginning with our OEM partners, right through to our end customers.
I am known for having exceptionally high bandwidth, a sales methodology based on creativity, passion and reliability, and for delivering results. I am constantly ranked top of our management surveys, and I believe that to survive and thrive we must know more than the competition about our products, the industry and our customers.
Fascinated by the future of technology and computing, I leverage my knowledge of hardware and software design and programming to predict trends and to shape our future product development and regional deployment strategies.

How technical is this talk? 3/7

What You'll Learn: The audience will learn about hardware for AI

Why is this takeaway(s) important? We hop to underscore the multifaceted significance of reducing compute costs in the context of AI growth. Firstly, we're promoting broader participation by making AI more accessible. Secondly, we advocate for exploring alternatives to Nvidia, fostering competition, and diversifying the AI hardware landscape. Lastly, we will highlight the advantages of a tightly integrated RISCV and AI hardware solution, showcasing the potential benefits of a cohesive and efficient technological framework. These aspects contribute to a more inclusive, competitive, and optimized AI ecosystem.

What is unique about this talk: Open sourced software and hardware

12:05 PM - 12:35 PM EST

Supercharging Search with LLMs: The Instacart Journey
*Prakash Putta, Staff Software Engineer at **Instacart***

Which talk track does this best fit into?: Applied Case Studies

Abstract: Discover how Instacart's search journey has been revolutionized through the implementation of Language Models (LLMs). By leveraging the power of LLMs, we have achieved significant enhancements, transforming the search experience for Instacart users. Join us to discover real-world use cases, gain insights into our seamless integration strategies, and witness how LLMs have empowered us to overcome challenges, deliver personalized recommendations, and elevate the overall search experience at Instacart.
[An example - https://techcrunch.com/2023/05/31/instacart-in-app-ai-search-tool-powered-by-chatgpt/]

Speaker Bio: Prakash Putta, Staff Software Engineer at Instacart
https://www.linkedin.com/in/prakashreddyputta/

How technical is this talk? 4/7

What You'll Learn: The audience will gain valuable insights into the practical implementation of LLMs in real production use cases. While LLMs have garnered significant attention, their effective integration into live environments remains a challenge. By attending this talk, participants will learn firsthand about the successful utilization of LLMs to enhance Instacart's search journey. Through real-world examples, they will discover how LLMs can be harnessed to supercharge search capabilities and derive actionable knowledge for their own production scenarios.

Why is this takeaway(s) important? The takeaway from this talk is important because it provides a deeper understanding of the practical application of LLMs in real-world production environments. While LLMs have gained popularity, many organizations still face challenges in effectively utilizing them. By learning about the specific use cases at Instacart, attendees can gain insights into the potential of LLMs to revolutionize search journeys. This knowledge can be invaluable for industry professionals seeking to leverage LLMs in their own projects, enabling them to make informed decisions and drive successful implementation strategies. Ultimately, this takeaway empowers the audience with practical knowledge and inspires them to explore the transformative possibilities of LLMs in their own organizations.

What is unique about this talk: What sets this talk apart is the unique firsthand perspective it offers on the implementation of LLMs in a real production setting. While there is ample information available online about LLMs, the specific challenges, strategies, and successes encountered during the integration process are often not readily accessible. By sharing our experiences at Instacart, including real use cases and lessons learned, this talk provides exclusive insights that cannot be easily found online. Attendees will gain a deeper understanding of the practical considerations, trade-offs, and best practices involved in harnessing LLMs to supercharge the search journey. This unique perspective will equip them with valuable knowledge that extends beyond theoretical concepts, enabling them to navigate the complexities of integrating LLMs into their own production environments more effectively

1:35 PM - 3:15 PM EST

MLOps for Production-ready LLM — Putting LLMs into production through iterative training, fine-tuning, and serving
Jay Chun, Co-founder & CTO, VESSL AI

Workshop

Abstract: Despite the onset of commercially viable open-source Large Language Models and generative AI, companies are struggling to leverage cutting-edge models like Llama2 and Stable Diffusion for production-ready applications. Creating a simple demo page on a personal laptop and training, fine-tuning, and serving multi-billion parameter LLMs on HPC-scale infrastructure - with proprietary enterprise data - involves entirely different engineering challenges.

In this session, Jihwan, who co-foudned and now leads the product and engineering team at VESSL AI, explores the common data, security, and cost challenges of Enterprise LLMs,
and share how companies like Hyundai Motors, TMAP Mobility, and Scatter Lab leverage MLOps infrastructure to go from a simple model playground to deploying enterprise AI services in weeks through iterative training, fine-tuning, and serving.

Speaker Bio: Jihwan co-founded VESSL AI and currently serves as the CTO, leveraging his decade-long expertise in DevOps in AI to help customers scale their machine learning workloads on VESSL AI. Before founding VESSL AI, Jay was at Google as a Site Reliability Engineer where he helped build out a distributed SQL database management & storage service. Prior to Google, Jay worked at PUBG, the gaming studio behind the blockbuster multiplayer shooting game, PlayerUnknown's Battlegrounds, handling the cloud and data infrastructure for 3M+ peak concurrent users. He was also an early DevOps engineer at DEVSISTERS where he oversaw the Kubernetes-backends for a popular mobile game with 20M+ downloads.

How technical is this talk? 5/7

What You'll Learn: Learn how companies can go from LLM playgrounds to production services through iterative training, fine-tuning, and serving.

1:35 PM - 2:05 PM EST

Stable Diffusion for Your Images: Custom Dream
Sandeep Singh, Head of Applied AI, Beans.AI

Which talk track does this best fit into?: Advanced Technical/Research

Abstract: "Welcome to the ""Stable Diffusion for Your Images: Custom Dream"" workshop! In this one-hour session, we will dive into the fascinating world of stable diffusion techniques for creating custom images.

Dreambooth is an innovative platform that allows users to unleash their creativity and transform ordinary images into extraordinary works of art. Through stable diffusion, we will explore how to manipulate and enhance images in a visually stunning and captivating way.

During the workshop, participants will learn the fundamentals of stable diffusion and its applications in image editing. We will cover various techniques, including color manipulation, texture enhancement, and image blending, to create visually striking and unique compositions.

Additionally, we will delve into the intricacies of Dreambooth's user-friendly interface, providing hands-on demonstrations and step-by-step guidance. Participants will have the opportunity to experiment with different filters, effects, and settings, unleashing their artistic potential and transforming their photos into mesmerizing masterpieces.

Whether you are a professional photographer looking to add an extra flair to your work or an amateur enthusiast eager to explore new creative avenues, this workshop is designed to inspire and empower you. Join us for an hour of exploration, experimentation, and artistic expression as we unlock the potential of stable diffusion with Dreambooth."

Speaker Bio: "Sandeep Singh is a leader in applied AI and computer vision in Silicon Valley's mapping industry, and he is at the forefront of developing cutting-edge technology to capture, analyze and understand satellite imagery, visual and location data.

With a deep expertise in computer vision algorithms, machine learning and image processing and applied ethics, Sandeep is responsible for creating innovative solutions that enable mapping and navigation software to accurately and efficiently identify and interpret features to remove inefficiencies of logistics and mapping solutions.

His work includes developing sophisticated image recognition systems, building 3D mapping models, and optimizing visual data processing pipelines for use in logistics, telecommunications and autonomous vehicles and other mapping applications.

With a keen eye for detail and a passion for pushing the boundaries of what's possible with AI and computer vision, Sandeep's leadership is driving the future of applied AI forward."

How technical is this talk? 6/7

What You'll Learn: Stable Diffusion is very easy for anybody to use and customize for the purpose at hand.

1:35 PM - 2:05 PM EST

Amumu brain; How League of Legends uses machine learning an applied data science
Ian Schweer, Staff Software Engineer, **Riot Games**

Case Study

Abstract: League of legends faces lots of interesting problems in the data space that are unique due to the video game aspect. How do you deploy and train models in a binary video game? What is the fundamental data model? How has the data and ML stack changed since the league's inception in 2009? How do you do player-facing ML (Lane detection, feeding detection, etc.) and decision science at this scale?

Speaker Bio: Ian is a staff software engineer at Riot Games, working on the League Data Central team. Along with his team, Ian ships Machine Learning and Data products to millions of league of legends and tft players including in game recommendations, player behavior models, and internal decision science to help make the game a better place for all. Ian has worked on large data systems at Adobe and Doordash before coming to Riot Games. In his free time, he plays in metal bands and hangs out with his 2 year old daughter.

How technical is this talk? 4/7

What You'll Learn: My goal is to teach a very pragmatic approach to shipping machine learning in a more constrained environment. This will include problems around network saturation, how to measure at player scale, and the challenges of interpretability for game design.

1:35 PM - 1:40 PM EST

How NOT to get ML Models into production
Ryan Turner, ML Solutions Engineer, **DVC**

Abstract: Ryan Turner will be presenting on the various pitfalls of productionizing ML models. Common issues include dependency management, correct testing processes, and over reliance on Python notebooks. There will be no shortage of satire and metaphor. The talk will draw upon the collective experience of several ML engineers at DVC.

Speaker Bio: Ryan has worked as an ML engineer at companies like Uber and Twitter. He is now developing the platform at Iterative.AI. He grew up in Santa Cruz, CA. After spending a few years in the UK and then Canada, he now lives in Reno, NV.

1:45 PM - 1:50 PM EST

Feature Store are NOT about Storing Features
Simba Khadder, Founder & CEO, **Featureform**

Abstract: Feature Store is a misnomer. Their objective is not to simply store feature values but rather to facilitate the organization, collaboration, versioning, discovery, and serving of feature definitions. In this Ignite talk, we'll break down the actual goal of feature stores, their value, and where they fit into the MLOps stack.

Speaker Bio: Simba Khadder is the founder & CEO of Featureform. He started his ML career in recommender systems where he architected a multi-modal personalization engine that powered 100s of millions of user's experiences. He later open-sourced and built a company around their feature store. Featureform is the virtual feature store. It enables data scientists to define, manage, and serve model features using a Python API. Simba is also a published astrophysicist, an avid surfer, and ran a marathon in basketball shoes.

1:55 PM - 2:00 PM EST

Generative AI: the open source way
Andreea Munteanu, MLOps Product Manager, **Canonical**

Abstract: Generative AI is probably the topic of the year. Leaders across the world feel the pressure of missed opportunities related to the latest technology. Professionals are also left worried about the impact that latest technology will have on their roles. Data scientists and machine learning engineers are challenge and need to quickly upskill. With such a stretched picture in mind, will generative AI deliver up to the great promises that it has right now?

This lighting talk will depict the opportunities that open source gives to generative AI to accelerate innovation. Between anxiety and enthusiasms, the latest technologies bring a new angle to the market. Let’s learn together about genAI with open source: from models to tooling to applications

Speaker Bio: Andreea is a Product Manager at Canonical, leading the MLOps area. With a background in Data Science in various industries, such as retail or telecommunications, Andreea used AI techniques to enable enterprises to benefit from their initiatives and make data-driven decisions. Andreea is looking to help enterprises get started with their AI projects and then deploy them to production, using open-source, secure, stable solutions.

2:00 PM - 2:05 PM EST

Building an end-to-end MLOps Pipeline
Aurimas Griciūnas, Head of Product, **Neptune.ai**

Abstract: The talk will be about MLOps and the lifecycle of ML projects. I will go through stages involved in the ML project lifecycle and some key highlights from each of them. I will also explain how CI/CD is different in Machine Learning project when compared to regular software and highlight how it evolves with the maturity of MLOps processes in the organisation. I will also ground the explanations with real life examples of building out MLOps capabilities, successes and failures.

Speaker Bio: Aurimas has over a decade of work experience in various data-related fields: Data Analytics, Data Science, Machine Learning, Data Engineering, and Cloud Engineering. For a few years he also led teams working with Data and Infrastructure. Today, Aurimas is Head of Product at neptune.ai.

2:10 PM - 2:40 PM EST

Evolved Structures: Using AI and Robots to build spaceflight structures at NASA
*Ryan McClelland, Research Engineer, **NASA Goddard Space Flight Center***

Which talk track does this best fit into?: Case Study

Abstract: Come get a first hand account of how NASA is leveraging Generative Design to reduce the cost and increase the performance of spaceflight missions. Also, how the concept of AI Prompt Engineering can be practically applied to diverse fields, such as structures development.

Speaker Bio: From a young age, Ryan McClelland has been captivated by futurism and technology, aspiring to contribute to a brighter future. As a Research Engineer in NASA GSFC's Instrument Systems and Technology Division, he pursues the development and implementation of digital engineering technologies for space-flight mission. Ryan is particularly excited about the potential of Artificial Intelligence, Virtual Reality, Generative Design, and Digital Manufacturing to accelerate space systems development.
With a diverse background in technology development, Ryan's previous research encompasses lightweight X-ray optics, aluminum foam core optical systems, and the investigation of non-linear effects in kinematic mechanisms. In addition to his research, Ryan has played a significant role in various flight missions, including designs currently on orbit aboard the Hubble Space Telescope and International Space Station. Recently, he served as the Roman Space Telescope Instrument Carrier Manager. Ryan holds a B.S. in Mechanical Engineering, summa cum laude, from the University of Maryland.

How technical is this talk? 4/7

Why is this takeaway(s) important? How the concept of AI Prompt Engineering can be practically applied to diverse and critical fields, such as structures development for spaceflight.

What is unique about this talk: NASA's application of Generative D
esign and Digital Manufacturing.