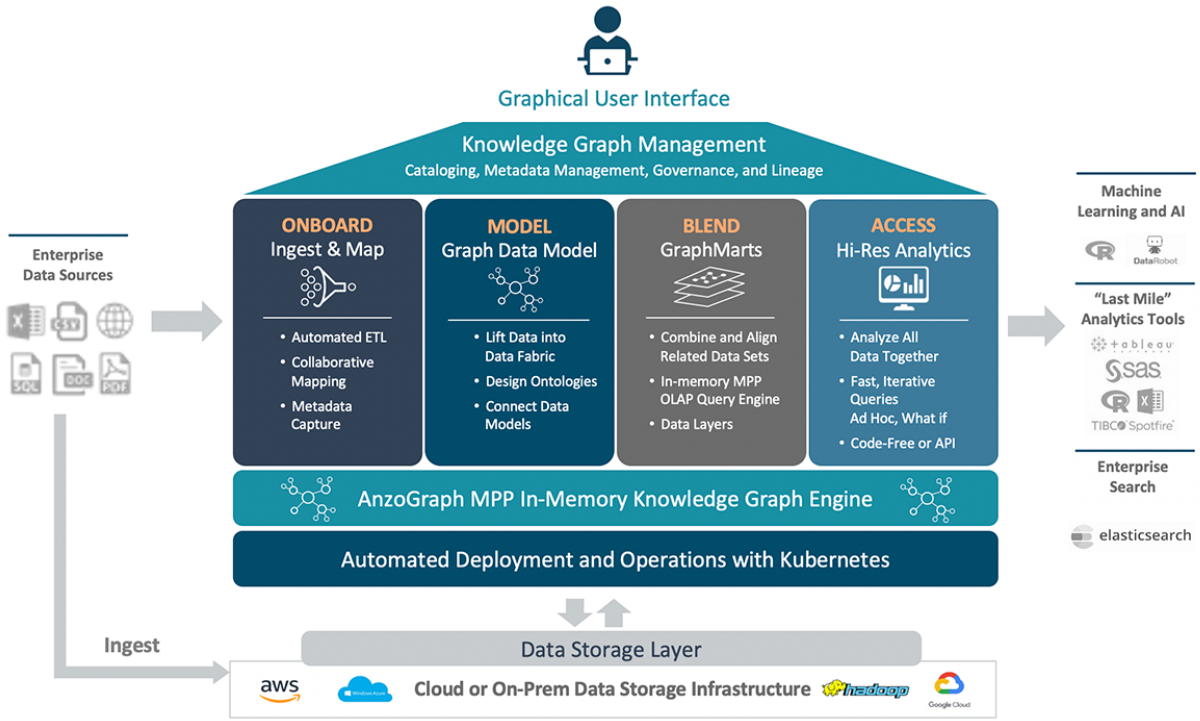
The conference included a mix of virtual and in-person talks, workshops and demos; except for one or two sessions, there were different sets of sessions in the virtual and in-person option.

**Underlying themes from the sessions I attended:**
- **Effective feature engineering**
- **Interpretability of ML**
- **Knowledge graphs**

**Cons**
- **sales pitch**
- **Advertised their virtual option - Some virtual sessions were watched after the conference**

**MAY 9**
**(Demo) Accelerating AI/ML initiatives with Knowledge Graph**
Greg West | Cambridge Semantics | Principal Presales Engineer

I attended this session to understand the utility of knowledge graphs in AI and ML workflows. This was a demonstration one how to apply Anzo's knowledge graph platform created by Cambridge semantics to optimize knowledge graph generation and visualization.




**(Talk) Fundamentals Statistical and Machine Learning Models for Time Series Analysis**
Jeffrey Yau | Fanatics Collectibles | Chief Data & A.I. Officer

This workshop was indeed a baseline understanding of time series modeling using the classics (e.g., ARIMA). I was hoping that the speaker would expand the workshop to include other

types of modeling but that was not the case. Overall, my biggest takeaway is that forecasting does not mean predicting and that you must make that distinction when creating and deploying time series models.

**(Virtual | Talk) Train and sustain: Why data leaders need to pay attention to human-in-the-loop (HITL)**
Matt Beale | Cloudfactory | Senior Solutions Consultant

The speaker about how Cloudfactory connects people technology and how humans have made an impact in the AI/ML space from gathering the right data to creating training data and sustaining data models. He also mentioned how training data affects how we create, train, build, and deploy models. He talked about how messy most open source datasets and why it's important for people using these datasets to clean them before applying to workflow; he mentioned how cleaning 7000 labels in one of the PASCAL Visual Object Classes datasets increased model precision by 13%.

**(Talk) Unlock Hidden Signals in Your Data with Graph Data Science**
Katie Roberts PhD | Neo4j, Inc. | Data Science Solution Architect

The talk was focused on how neo4j graphs enrich all phases of the AI ecosystem. Graphs enrich **data** in the form of knowledge graphs (Neo4j), **modeling** in the form of graph feature engineering and graph ML (Neo4j GDS), **analysis** in the form of graph analytics, investigations (Neo4j Bloom) and counterfactuals and. Integrations in heuristic AI (Neo4j Connectors). Graphs can be used to represent structured and unstructured data to find the patterns in connected data, identify associations, anomalies, and trends using unsupervised machine learning, and learn features that you don't know are important yet. Insights from graph algorithms are derived from various metrics including centrality, pathfinding, community detection, similarity, embeddings, and link prediction. I liked that the speaker shared different use cases for knowledge graphs; for example graph algorithms for entity link analysis detect first party and synthetic identity fraud across channels in financial sector and other industries.

**(Virtual | Tutorial) Graph Technology and Data Science Workshop**
Alison Cossette | Neo4j, Inc. | GDS Developer Advocate - Data Scientist

This was a deeper dive into the application of graphs in the data science workflow. By applying graphs, you're able to bring context and unlock unattainable predictions. The speaker talked about how there needs to be a paradigm shift on how data is visualized from columns and rows to connections. I liked that the speaker dived deeper into various graph algorithms and the application of the Neo4j capability in graph data science. The speaker walked through a sandbox tutorial using a data on airline location and airline flights and another tutorial on exploring a contact tracing database. In the tutorial the speaker showed the data model on node types and relationship types; queries applied to load and explore the data, and metrics used to test connections.

**(Virtual | Talk) If We Want AI to be Interpretable, We Need to Measure Interpretability**
Jordan Boyd-Graber, PhD | University of Maryland | Associate Professor

The talk was centered around why AI should be interpretable and not a black box and measures for interpretability in supervised and unsupervised methods. He shared his work in

applying Interpretability to topic models, moving from model evaluation using held out likelihood testing to more human centric evaluations (human-in-the-loop).

**(Workshop) Advanced Gradient Boosting (I): Fundamentals, Interpretability, and Categorical Structure**
Brian Lucena | Numeristical | Principal

This workshop was centered around applying gradient boosting methods to transform a week model by sequentially building models to reduce errors of the previous models. The workshop was mostly done in a Jupyter notebook with examples of how to apply gradient boosting with catboost regressors, XGboost regressors, LightGBM, and sklearn in hyperparameter optimization; interpreting models through metrics like root mean square, mean squared error, mean absolute error, Individual Conditional Expectation plots, and SHAP values. The speaker also walked through the usage of a boosting method he developed called structure boost in categorical and multi-classification model workflow.

**MAY 10**
**(Virtual | Workshop) Generative AI**
Leonardo De Marchi | Thomson Reuters | VP of Labs

This was a hands-on workshop on the usage of AI models to generate creative outputs. The speaker discussed how generative ai is applied to generate poetry using NLP models like LSTM and Transformer, create digital art using computer vision models like Deep Dream and StyleGAN, and generate music using GANs and other AI models. He presented on the history and additional applications of generative AI; for example, describing images, creating images from text (Mansimov et al, 2015), creating better resolution images (super resolution by Google AI). He also described different generative approaches; probabilistic models, autoregressive models, rule-based models, adversarial networks, hybrid approaches and evolutionary algorithms. Exercises ranged from text generation with GPT-2 to the use of GPT4all Langchain for chatbots, generative question and answering, and summarization.

**(Workshop) Machine Learning with XGBoost**
Matt Harrison | MetaSnake | Python & Data Science Corporate Trainer

This tutorial was centered around how to apply XGboost in predictive models with structured data. Time was spent explaining XGboost and walking through how to apply the library in predictions. The speaker covered how to explore the data, ensure the model isn't overfitting using tools like SHAP, ICE plots, monotonic constraints, and learning curves. I will also teach you how to deploy your model, how to tune and evaluate model created and how to derive insights from the predictions.

**(Talk) Uncovering Behavioral Segments by Applying Unsupervised Learning to Location Data**
Ali Rossi | Foursquare | Data Science Tech Lead

The speaker talked about how she applied unsupervised learning for behavioral segmentation on foursquare data from stop-detection technology that snaps people to places with precision, collecting data from multiple sensors at various location. This method can identify patterns in data without guidance and labelling of target variables, group individuals based on shared

characteristics or behavior patterns to create segments to help marketers gain insight into the behavior of different customer segments and better target marketing efforts. She walked through how her process of deriving meaningful features from the data, bucketing and standardizing features, condensing initial features using PCA for dimensionality reduction, then implementing K-means clustering with cluster names for interpretability. Her analysis identified 6 holiday shopper audiences with behaviors associated with people's shopping activities which enabled the targeting team within foursquare to test the data-driven segments with a beta client.

**(Virtual | Workshop) The Data Cards Playbook: A Toolkit for Transparency in Dataset Documentation**
Andrew Zaldivar, PhD | Google Research | Senior Developer Relations Engineer
Mahima Pushkarna | Google | Senior User Experience Designer

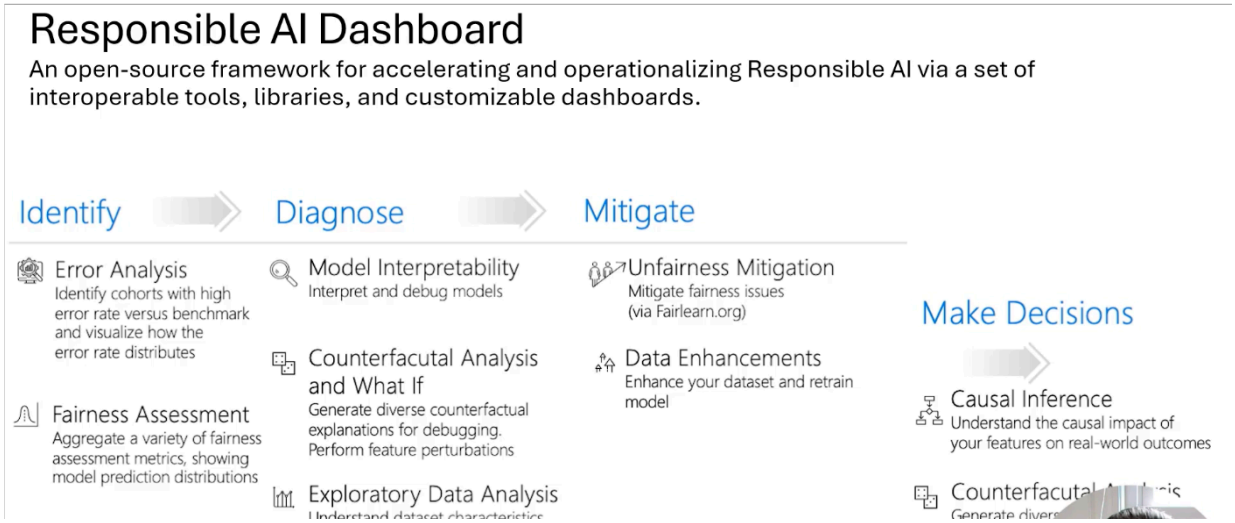
The speaker talked about how transparency in documentation provides clear and easily understandable explanation of datasets, models and products. Data cards is a structured framework that summarizes important information about datasets and models, enabling end-users and stakeholders to make better decisions across the project lifecycle. The tutorial was centered around using free and accessible resources associated with the toolkit to create new data card templates, modify existing documentation, and apply data cards to explain example projects; the resource can be accessed here.

**(Virtual | Talk) Responsible AI In Practice**
Minsoo Thigpen | Microsoft | Senior Product Manager
Mehrnoosh Sameki, PhD | Microsoft | Principal PM Manager
Besmira Nushi | Microsoft | Principal Researcher

In the talk, Minsoo showcased 6 core ethical recommendations in the future computing that represents Microsoft's view on AI. The principles are fairness, reliability & safety, privacy & security, inclusiveness, transparency, and accountability. To enact these principles at scale, at the foundation there should be a governance structure to enable progress and accountability, then rules to standardize our responsibility AI requirements, then create and/or improve training and organization practices to promote a human centered mindset and finally the tool and processes for implementation. Microsoft open source tools like Fairlearn, InterpretML, Error Analysis, and the Responsible AI dashboard (all aforementioned tools + more) can help with identifying, diagnosing, mitigating issues with AI to aid decision making; these tools have been integrated into Azure machine learning. Mehrnoosh/Besmira demo'd the use of the responsible AI dashboard in data analysis, feature and model evaluation and shared other tools in the responsible AI toolbox that can be utilized; Responsible AI Mitigations, a Python library for implementing and exploring mitigations of Responsible AI on tabular data; Responsible AI Tracker, a Jupyterlab extension for tracking, managing and comparing Responding AI mitigation and experiments on tabular data.

**(Talk) How to build and operationalize time series models**
Philip Wauters | Tangent Works | Customer Success Manager and value engineer

This talk was centered around how to use the "Tangent Works' Tangent Information Modeler (TIM) in time series modeling. I was not expecting the talk to be so centered around TIM; in fact the speaker did not mention this in the summary of the talk. My takeaway from this talk was on the importance of effective feature engineering in automation, a step between preprocessing




and modeling; although the speaker mostly walked through how to create and evaluate features with TIM.

**MAY 11**

**(Virtual | Workshop) When Privacy Meets AI - Your Kick-Start Guide to Machine Learning with Synthetic Data**

Alexandra Ebert , Chief Trust Officer | Chair of the IEEE Synthetic Data IC Expert Group | AI, Privacy & GDPR Expert | MOSTLY AI | EEE Standards Association

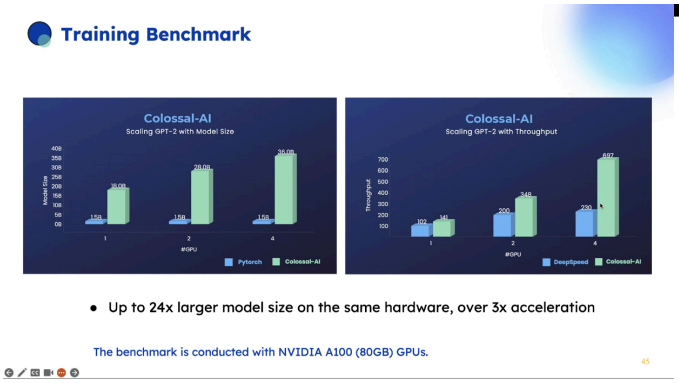
This was a great talk on the benefits of synthetic data and the need for privacy mechanisms when producing synthetic data. The speaker walked through a Jupyter notebook of metrics she's applied to evaluate the fidelity and privacy of synthetic data. Synthetic data can anonymize customer data, reduce time to data and time to market of data projects, augment and diversify data, increase data accuracy if the synthetic data retains similar structure, correlation and time dependencies, can allow data to be shared widely and freely within and across organizations, can be cost saving and can streamline legal compliance and risk management procedures. Not all synthetic data are anonymous and must have privacy mechanisms to produce and evaluate the data. Automated privacy mechanisms should prevent re-identification (there should not be a 1 to 1 link between the original and synthetic data), not introduce noise during synthesis, overfitting, and protect data value.

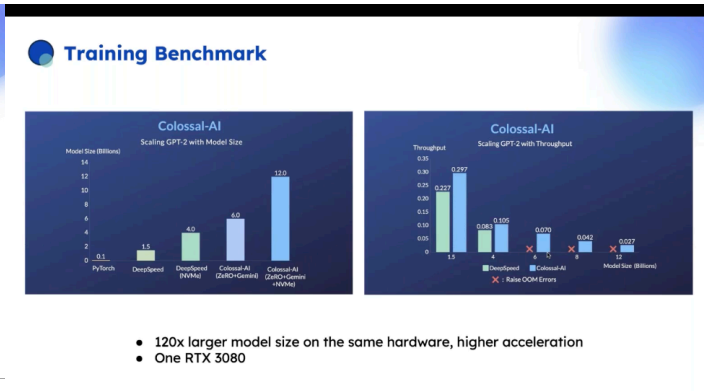
**(Virtual | Talk) Colossal-AI: A Unified Deep Learning System For Large-Scale Parallel Training**

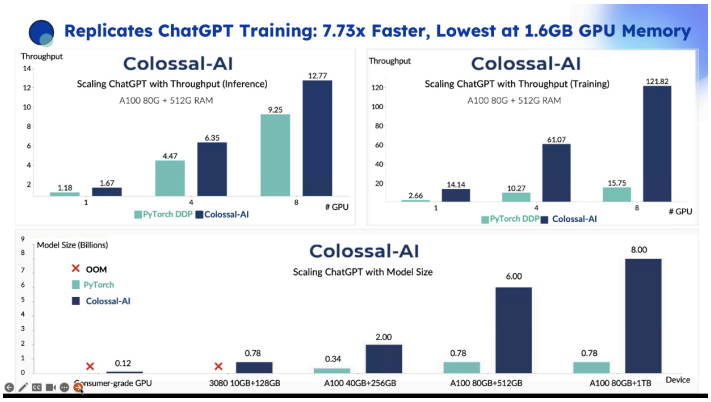
James Demmel, PhD | UC Berkeley | Professor of Mathematics and Computer Science
Yang You, PhD | National University of Singapore | Presidential Young Professor

This was a demo on the colossal AI, a platform for training large models. The challenges of using large AI models include the need to re-train on new data repeatedly, single GPU servers running out of memory and the need for a cluster of GPUs to load data and make predictions, expensive infrastructure and systems to deploy models. According to the speaker, Colossal AI maximizes computational efficiency, minimizes system running time, minimize communication, minimizes code refactoring, offers dynamic adaptive scaling, and reduces memory footprint. The capability serves as the in-between for hardware (e.g., CPU, GPU, TPU, FPGA) and

common frameworks (e.g., PyTorch, Keras, Transformers, Hugging Face, PyTorch Lightning). The speaker walked through the pros and cons of parallel data processing, model parallelism, tensor parallelism, and sequence parallelism of commonly used systems and how Colossal AI's parallelism system includes all types of parallelism with improved performance. It looks like colossal AI outperforms in model training when compared to PyTorch and deepseed, even when competing resources are limited.




**(Talk) Incremental Adoption of Spark, Dask, and Ray**
Han Wang | Lyft | Senior Staff ML Engineer

This talk focused on using the Fugue API, a unified interface for distributed computing, a capability intended for users to migrate algorithms and processes from pandas, python, arrow, and polars into Apache spark, Dask and Ray. From the title of the talk to the initial sides shared by the speaker, I thought that the talk will be focused on techniques to increase adoption of Spark, Dask and Ray because these popular frameworks for distributed computing improve scalability and efficiency of algorithms ran on pandas, python and some SQL based platforms. It seems to me that the Fugue API can scale existing Python and Pandas code to be used in Spark, Dask, or Ray; I am not sure if this capability has been fully developed.

**(Lightning Talks) Can You Forecast the Next Two Weeks? How about the Next 20 Years?: Digital Transformation in Market Forecasting at GE Aerospace**
Alexander Antony PhD | GE Aerospace | Senior Staff Data Scientist

This was a 15 minute talk on the application of market forecasting for multiple market demand indicators at GE Aerospace. Some market demand indicators include aircraft departures, flight

hours, air travel demand (revenue passenger kilometers/RPK), and capacity (available seat kilometers/ASK). To successfully build forecasting models, the speaker talked about the importance of data availability and storage, the right tools and infrastructure, the right models that understands business needs and estimate uncertainties, and subject matter expertise to balance judgements and quantitative forecasts. The speaker stated that “successful forecasters are good business owners” and shared some lessons learned which includes embracing and modeling uncertainty where it exists, expecting business requirements to chase, recognizing that subject matter expertise can be just as critical as statistical validation.

**(Lightning Talks) Why Orchestration and Airflow is the secret ingredient in MLOps**
Viraj Parekh | Astronomer | Cofounder

This was also a 15 minute talk on Apache Airflow, an open-source tool for programmatically authoring, scheduling, and monitoring data pipelines. The building blocks of Airflow are operators that act as wrappers around a task to define what the task does; common operators include PythonOperator, SimpleHttpOperator, ExternalTaskSensor, databricks-submitrunoperator, s3tosnowflakeoperator, aws-gluejobSensor, googlecloud-bigqueryoperator and Microsoft azure-Azurebatchoperator. The speaker mentioned that airflow can be used for feature engineering and shared examples of how airflow has been used in fraud detection and sports betting models.