FEATURE STORE SUMMIT

12-13 OCTOBER | 08:30 AM - 4:00 PM PT

ORGANIZED BY HOPSWORKS

FIXING MODELS BY FIXING DATASETS

// BEING DATA-CENTRIC IS THE FUTURE OF AI

Atindriyo Sanyal
Co-Founder
Galileo Technologies

FIXING MODELS BY FIXING DATASETS

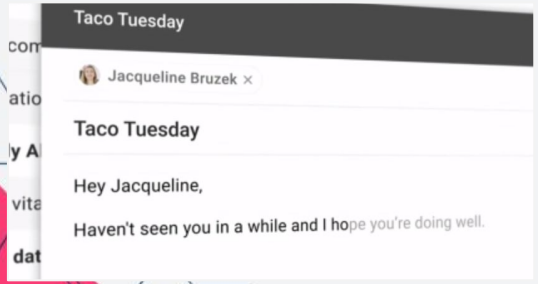
// BEING DATA-CENTRIC IS THE FUTURE OF AI

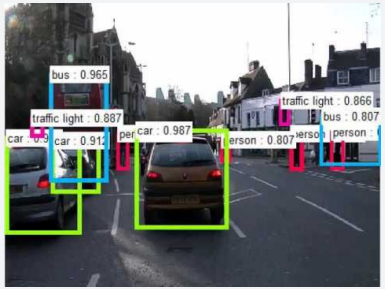
Feature Stores for ML

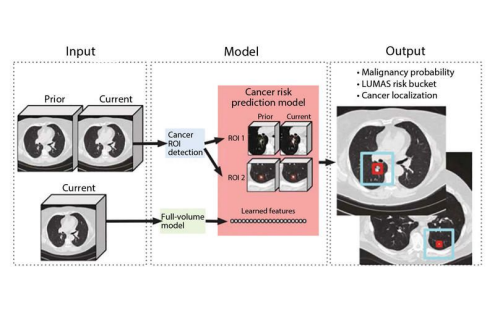
INTRODUCTION

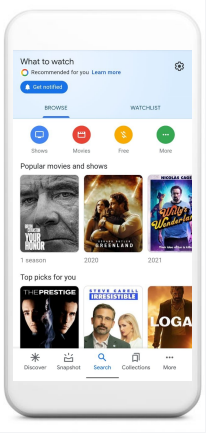
- ML is everywhere
  - Software applications → 2010s
  - ML applications → 2020s

- Early Innings and critical for high stakes decisions









Feature Stores for ML

MODEL ARCHITECTURES ARE COMMODITIZED

- Models are increasingly commoditized 🤗 Transformers IMAGENET
- Hyperparameter tuning is standardized
- ML Infrastructure maturing
  - Feature Stores / Embedding Stores
  - Scalable Deep Learning Frameworks

EXPLOSION OF ML MODELS

BAD DATA QUALITY

WELL TRAINED MODEL

POOR PREDICTIONS

REDUCED CHURN
HIGHER NPS
INCREASED LTV
HIGHER REVENUE
... AND MORE

ENSURING **CONTINUOUS HIGH QUALITY DATA** POWERS **HIGH QUALITY PREDICTIONS**

Feature Stores for ML

THE ISSUE LIES IN THE DATA

- Features are hard to find
- When found, often noise laden, insufficiently well-labelled and low quality
- Kitchen sink of data thrown at the model
  - Models trained on noisy data
  - Lack of real time observability leading to training-serving skew

Feature Stores for ML

DATA QUALITY ISSUES ACROSS THE WORKFLOW

Feature Discovery

Feature Preparation

Model Training

Model Evaluation

Model Inference

Feature Stores for ML

DATA QUALITY ISSUES ACROSS THE WORKFLOW

Feature Discovery

Feature Preparation

Model Training

Model Evaluation

Model Inference

Finding the ideal dataset.

→ Curate & Find the most representative data

→ Get maximum lift with minimum data

Feature Stores for ML

DATA QUALITY ISSUES ACROSS THE WORKFLOW

Feature Discovery

Feature Preparation

Model Training

Model Evaluation

Model Inference

Trusting your dataset.

→ Identify regions of Model Underperformance
→ Model robustness across sub-populations
→ Identify similar and dissimilar examples
→ Identify and fix noisy data and labels

Feature Stores for ML

DATA QUALITY ISSUES ACROSS THE WORKFLOW

Feature Discovery

Feature Preparation

Model Training

Model Evaluation

Model Inference

Evaluating your Test Data

→ Identify regions of Model Underperformance
→ Model robustness across sub-populations
→ Detecting clusters of noise

Feature Stores for ML

DATA QUALITY ISSUES ACROSS THE WORKFLOW

Feature Discovery

Feature Preparation

Model Training

Model Evaluation

Model Inference

Close the training<>serving loop.

→ Training<>Serving skew detection

→ Drift Detection

→ Real time data quality monitoring

Feature Stores for ML

DATA QUALITY ISSUES ACROSS THE WORKFLOW

Feature Discovery	Feature Preparation	Training & Evaluation	Inference
Finding the right Features Curate the most representative data	Inspect & Curate your dataset Identify and fix 'noisy' examples Identify anomalies	Optimize ML developer time ML data error analysis Model robustness across regions	Adapting to the changing world Training-Serving skew detection Data Quality Monitoring

Feature Stores for ML

CURATE THE MOST REPRESENTATIVE DATA

- Data Annotation budgeting via Active Learning
  - Identifying high value data via embeddings
    - K distributed cores
  - Identifying high value data via model uncertainty
    - sampling based on proximity to decision boundary
  - Clustering similar samples
  - Clustering dissimilar samples

- Optimal Feature Discovery
  - Filter by Feature Redundancy and Relevance to Labels
  - Entropy and Mutual Information

Feature Stores for ML

IDENTIFYING NOISY DATA AND LABELS

- Model based **confidence-uncertainty** metric
  - Margin sampling to identify potentially mislabelled examples
  - Monitoring high-low confidence-certainty regions

- Model based **certainty** in datasets
  - Class conditional joint distributions
  - Estimating Confident Joints
  - Data-independent estimation of incorrect classes

FINDING DATASET VULNERABILITIES

- Detecting regions of model underperformance
  - Model based confidence-certainty
    - Detecting confusion
    - Detecting noise
    - Separating easy data from hard data
- Instrumenting sub-populations of underperformance
  - Programmatic Assertions on Datasets
  - Patterns / Clusters within underperforming regions
    - Leveraging embeddings
- Optimizing ML development time
  - Systematic ML Data Error Analysis
  - Test model robustness across subpopulations

Feature Stores for ML

ADAPTING TO THE CHANGING WORLD

- Real-time detection of **Training<>Serving skew**
- **Monitoring subpopulations** of interest
- **Augmenting new data** for auto-retraining
  - Synthetic Data Generation
  - Alternative Knowledge Bases or Pre-trained models

Feature Stores for ML

DATA QUALITY ACROSS THE WORKFLOW

- **Evaluating** Data Quality for Train & Test Datasets
- **Gating** Pipelines on Data Quality
- **Storing** high value feature sets from unlabelled data
- Model Evaluation using **historically similar Feature Sets**
- **A/B Testing** across training, evaluation and serving data

Feature Stores for ML

Thank you!

Do you have any questions?

Atindriyo Sanyal
Co-Founder, Galileo

https://www.linkedin.com/in/atinsanyal/

https://twitter.com/atinsanyal