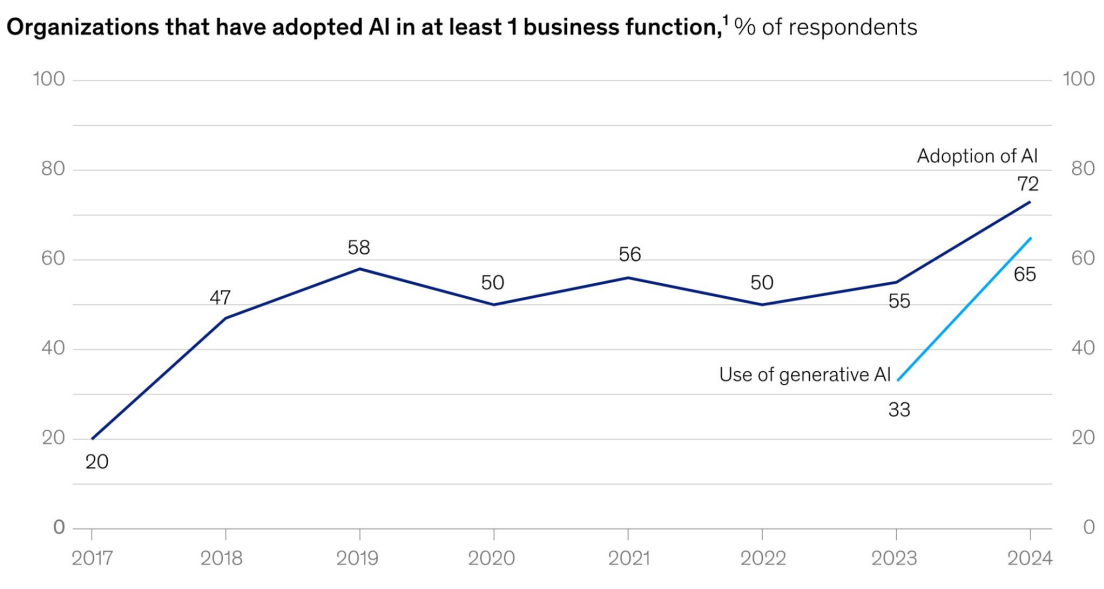
Scaling AI Workloads with Apache Airflow

Shubham Mehta
Sr. Product Manager, AWS

Rajesh Bishundeo
Software Development Manager, AWS

The AI Revolution

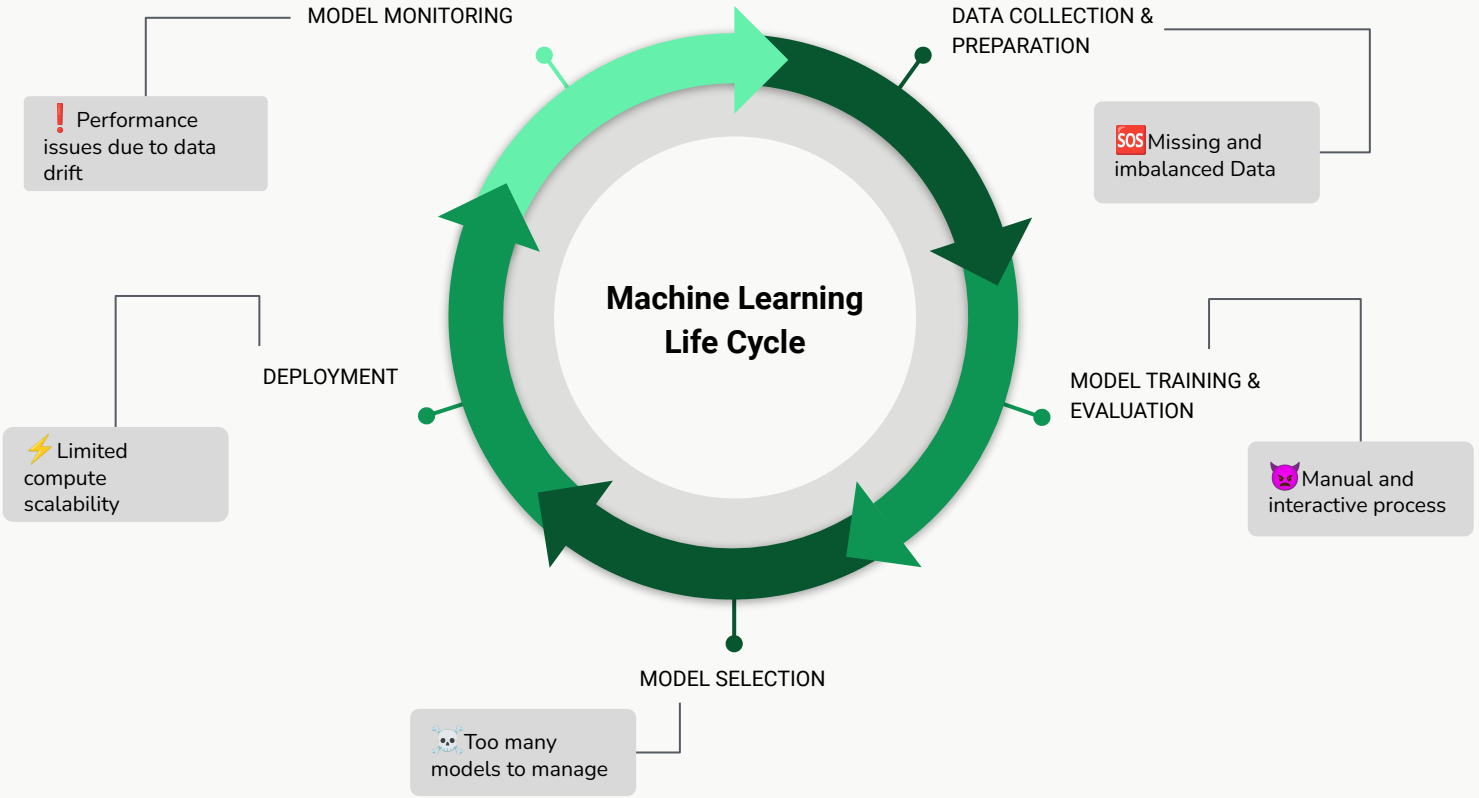
AI Adoption worldwide has increased dramatically in the past year, after slow growth for years.



Source: McKinsey The State of AI report

The AI Revolution

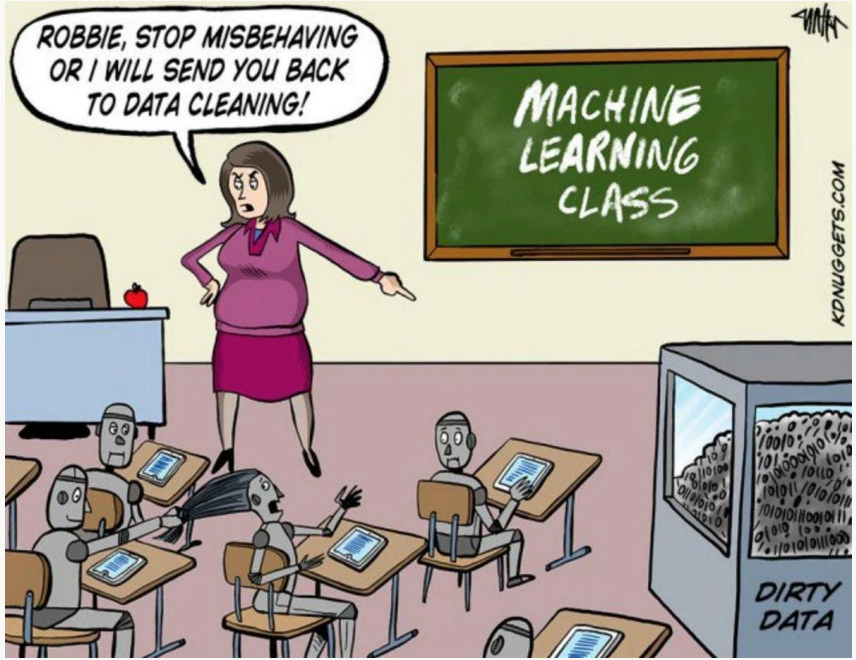
The ML Lifecycle



Common Pitfalls in AI Operations

Data Collection and Preparation Woes

Issues with data quality, consistency, and imbalance.

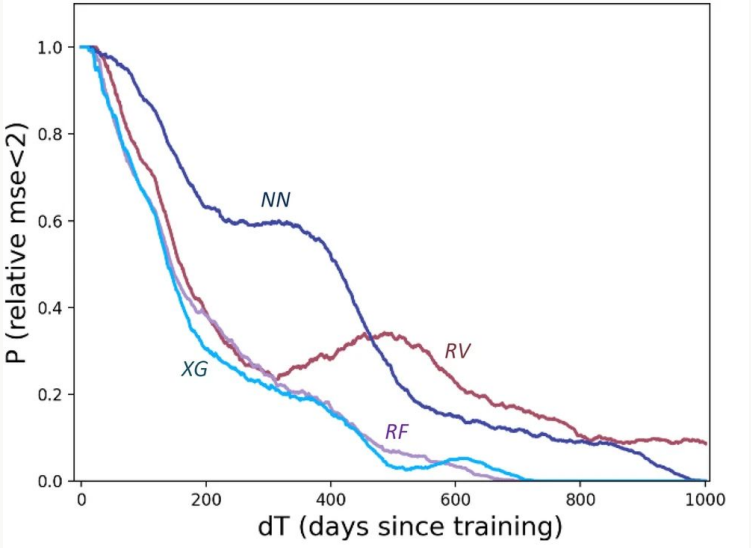


The Manual Model Training Grind

Issues with manual and interactive model training processes

Performance Degradation

Challenges in detecting and addressing data drift



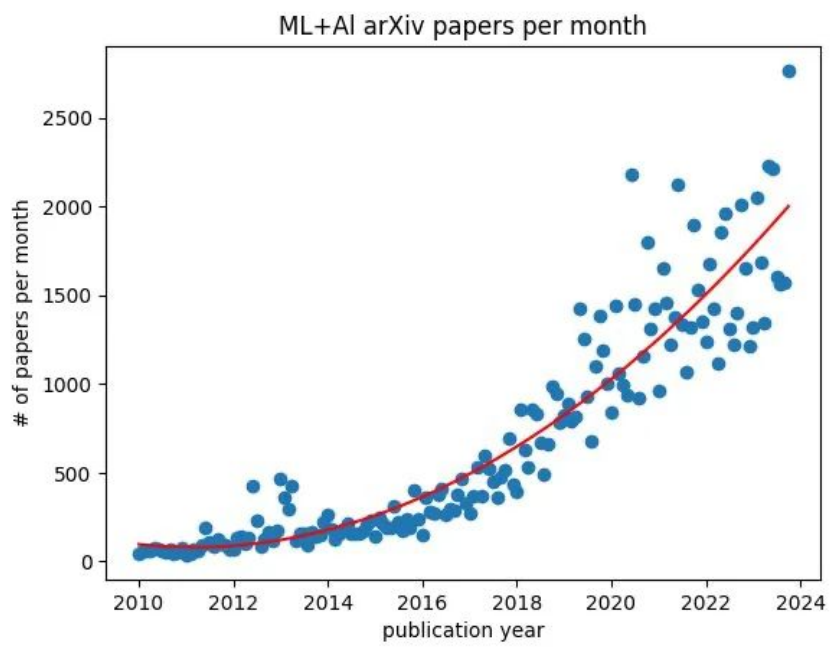
Probability of Models performing accurately

Penalized linear regression RidgeRegressor model (RV)
Random forest RandomForestRegressor model (RF)
Gradient boosting XGBoost model (XG)
Neural network MLPerceptronRegressor model (NN)

Source: Vela, D., Sharp, A., Zhang, R. et al. Temporal quality degradation in AI models. Sci Rep 12, 11654 (2022)

Model Proliferation Problems

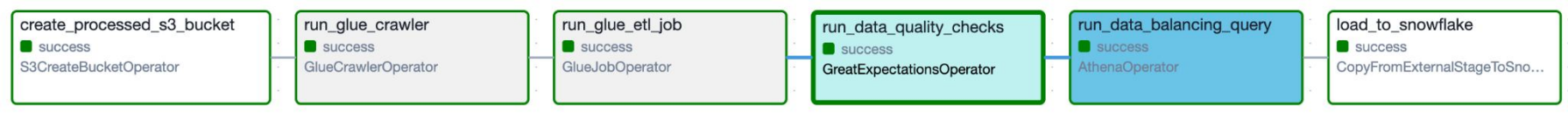
Challenges in managing multiple models



Enter Apache Airflow

Automating Data Preparation

Using Airflow for data quality checks and balancing



Data Wrangling

- Amazon SageMaker Data Wrangler
- Databricks (Spark)
- Amazon EMR & Glue (Spark)

Data Quality

- Great Expectations
- AWS Glue DQ (PyDeequ)
- Soda

Automating Data Preparation

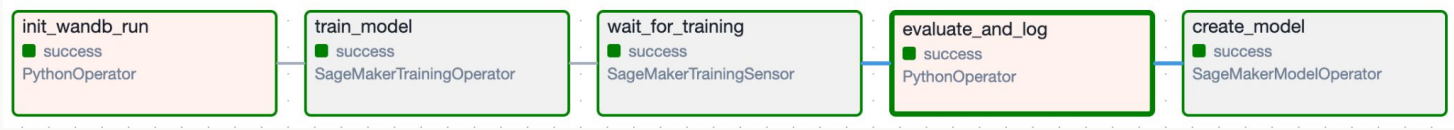
Rule Sets

- Record Counts
  RowCount > 100
- Validness
  “A” in [‘N’, ‘Y’]
- Completeness (non-null)
  IsComplete “A”
- Standard Deviation
  StdDev “A” between 3 and 5

```
create_rule_set = GlueDataQualityOperator(
    task_id="create_rule_set",
    name=rule_set_name,
    ruleset=RULE_SET,
    data_quality_ruleset_kwargs={
        "TargetTable": {
            "TableName": athena_table,
            "DatabaseName": athena_database,
        }
    },
)
start_evaluation_run = GlueDataQualityRuleSetEvaluationRunOperator(
    task_id="start_evaluation_run",
    datasource={
        "GlueTable": {
            "TableName": athena_table,
            "DatabaseName": athena_database,
        }
    },
    rule_set_names=[rule_set_name],
)
```

Integration with Experimentation Platform

Using Airflow for model versioning, deployment, and monitoring



Model Training

- Amazon SageMaker
- Databricks
- Amazon Bedrock

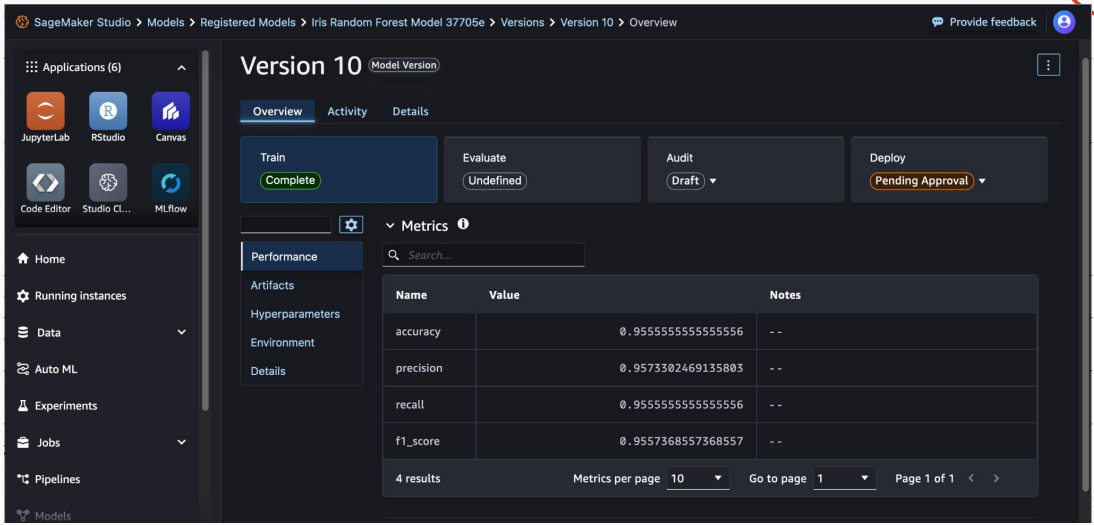
Model Cataloging

- Weights & Biases
- MLFlow (SageMaker & Databricks)
- Neptune

Integration with Experimentation Platform

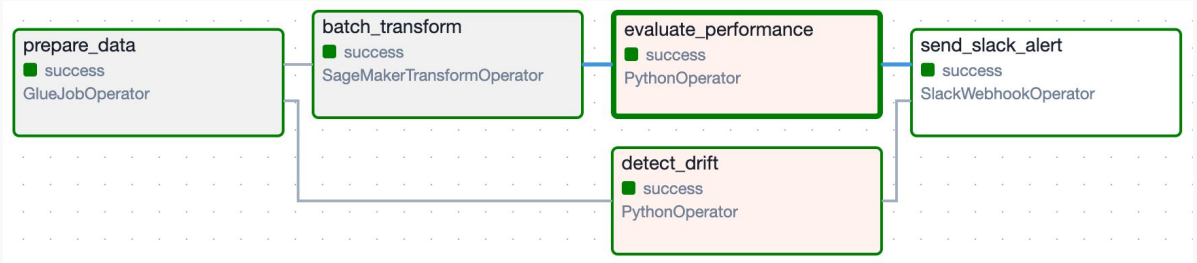
SageMaker MLFlow

- Experiment Tracking
- Model Evaluation
- Model Registry



Automated Evaluation and Monitoring

Using Airflow for continuous monitoring



Evaluation Frameworks

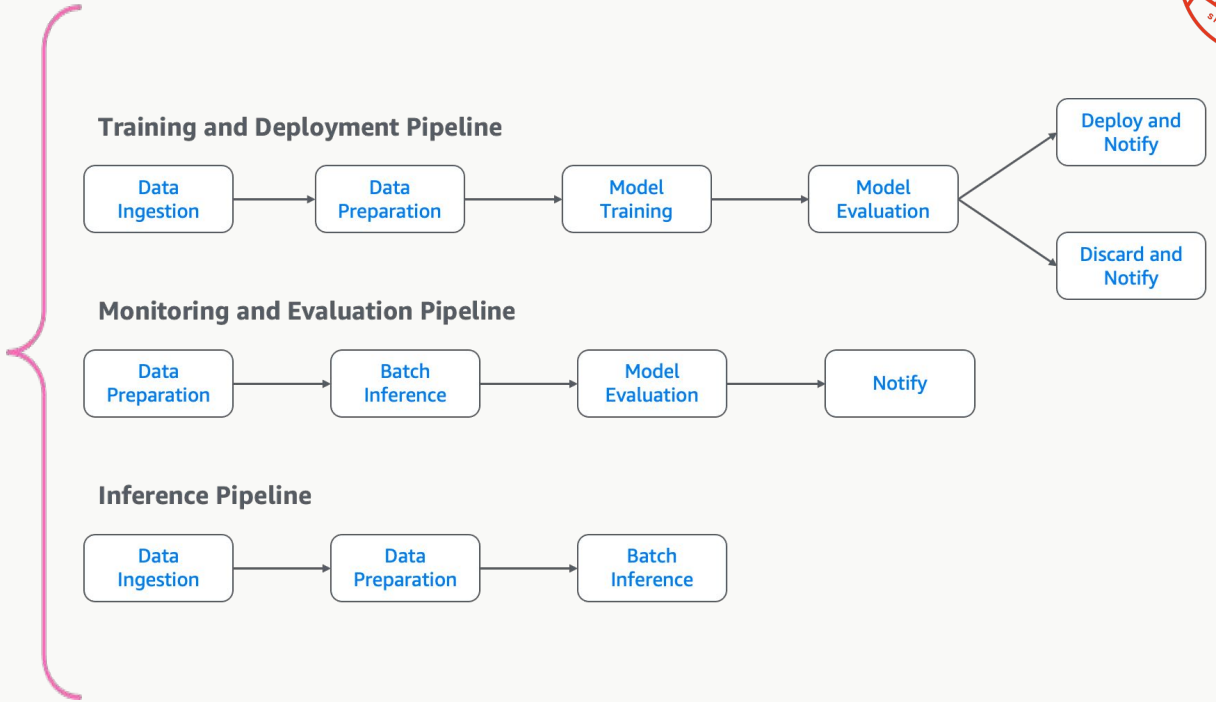
- Amazon SageMaker
- Evidently (Data drift Evaluation)
- Ragas (LLM RAG Evaluation)

Notification Systems

- Slack Notification
- Amazon Simple Notification Service
- Email

Scaling AI Adoption with Airflow

- Dag Factory
- Dag Templating
- Data-driven Scheduling



Real-World Success Stories

Airflow Success Story

What do they do?

- Identity Verification - Financial, Healthcare, Gambling (including biometric solutions)

Challenges

- People
- Tech
- Governance

Solutions

- Airflow - Data prep & ETL
- SageMaker - model training
- Templates
- Datahub

Airflow Success Story

What do they do?

- Industry leader in GenAI for contact centers: GenerativeAgent, AutoTranscribe, and AutoSummary

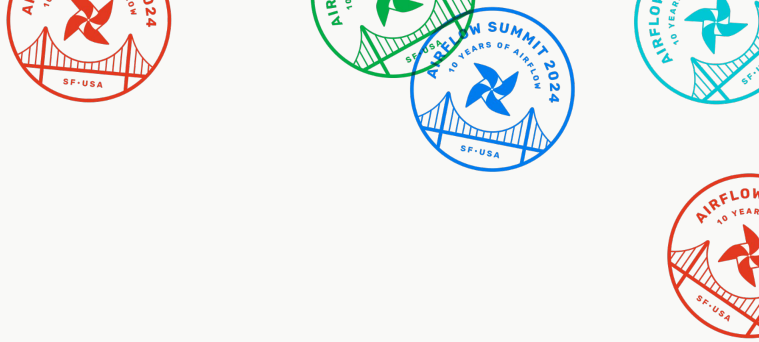
Challenges

- Workflow management: ML Lifecycle, Tech

Solutions

- Airflow: Data ingestion & preparation; model training/tuning/evaluation; parallel workloads; scalable AI workload orchestration

For more details: Airflow, Spark, and LLMs: Turbocharging MLOps at ASAPP by Udit Saxena (Thursday @11am)



Questions?

Shubham Mehta
LinkedIn: linkedin.com/in/shubh-mehta
Email: shubhx@amazon.com

Rajesh Bishundeo
LinkedIn: linkedin.com/in/rajeshwar-bishundeo/
Email: rbishund@amazon.com