DATA+AI
SUMMIT 2022

Detecting financial crime using an Azure advanced analytics platform and MLOps approach


Lars Haringa, Data Scientist, ABN AMRO Bank

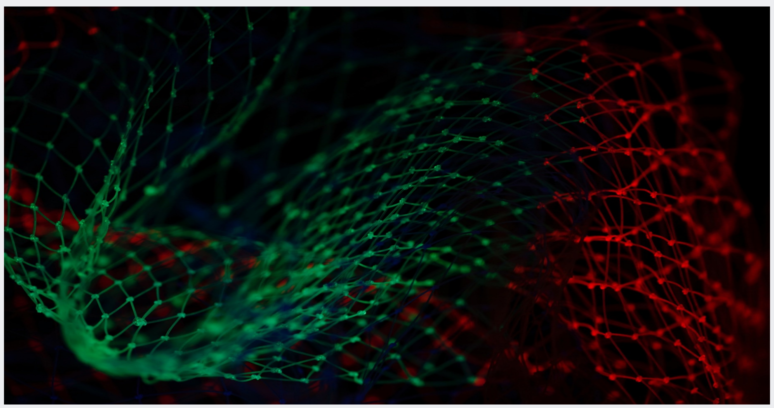

Saman Amini, ML Engineer, ABN AMRO Bank

ORGANIZED BY databricks

Outline

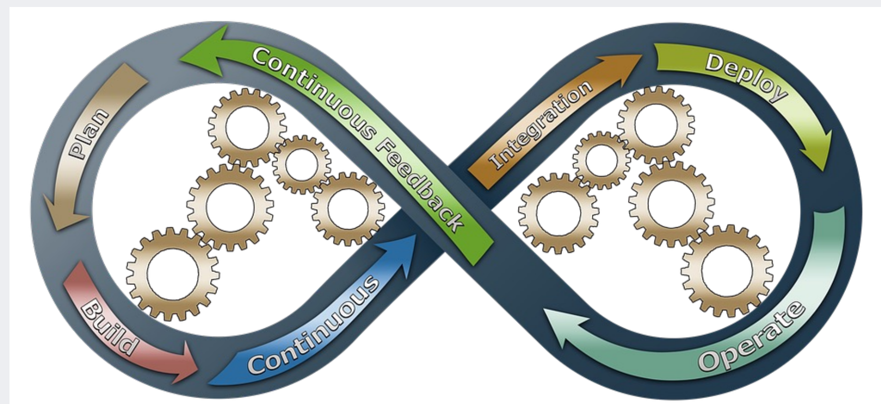
Transaction Monitoring with Machine Learning

Lars Haringa



MLOps

Saman Amini



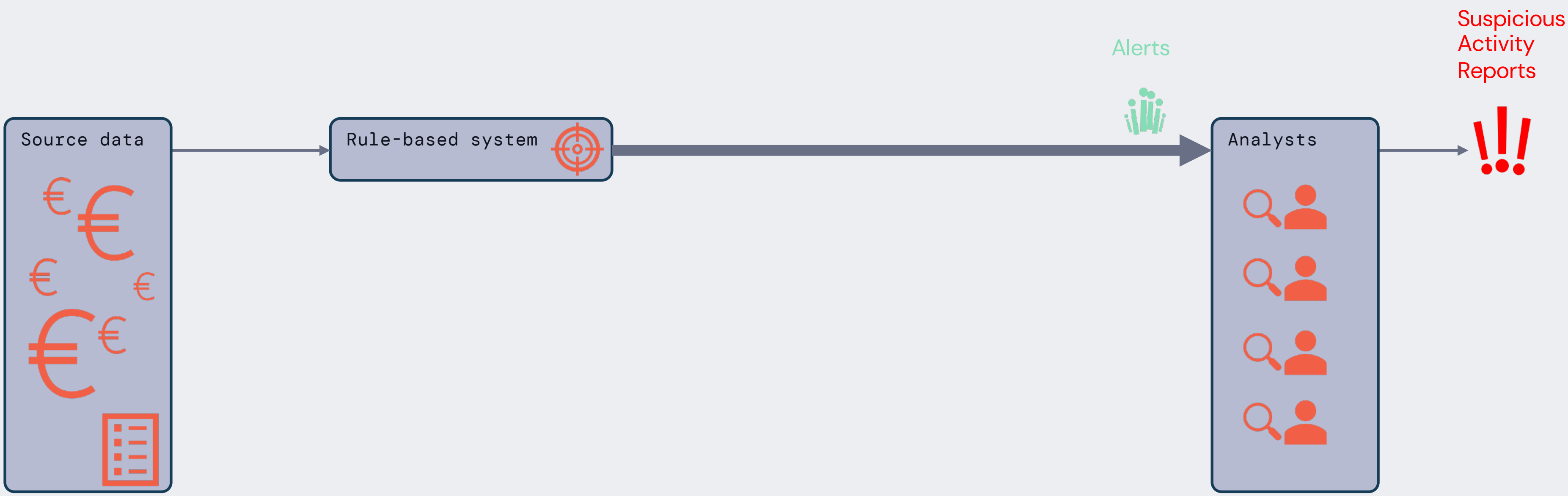
Gatekeepers of the financial system

ABN·AMRO

- 1/4 Systemically Important Banks banks in NL
- Many millions of transactions per day
- Strictly regulated

- Legal obligation for Transaction Monitoring (TM) ← profits of organised crime
- Unique position for detecting financial crime (DFC)
- Costly operation

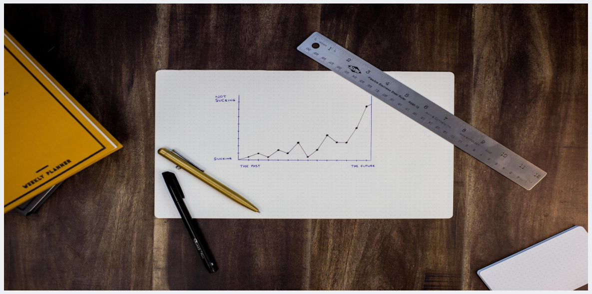
Rule-based TM

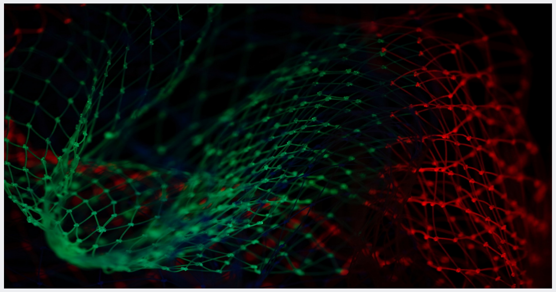


From rule-based TM to TMML

- Hand-built (a priori)
- Simple
- Many false positives

- *E.g. if deposit > threshold, investigate*





- Data-driven Transaction Monitoring Machine Learning (TMML)
- Complex: 100s simultaneous decisions
- Efficient

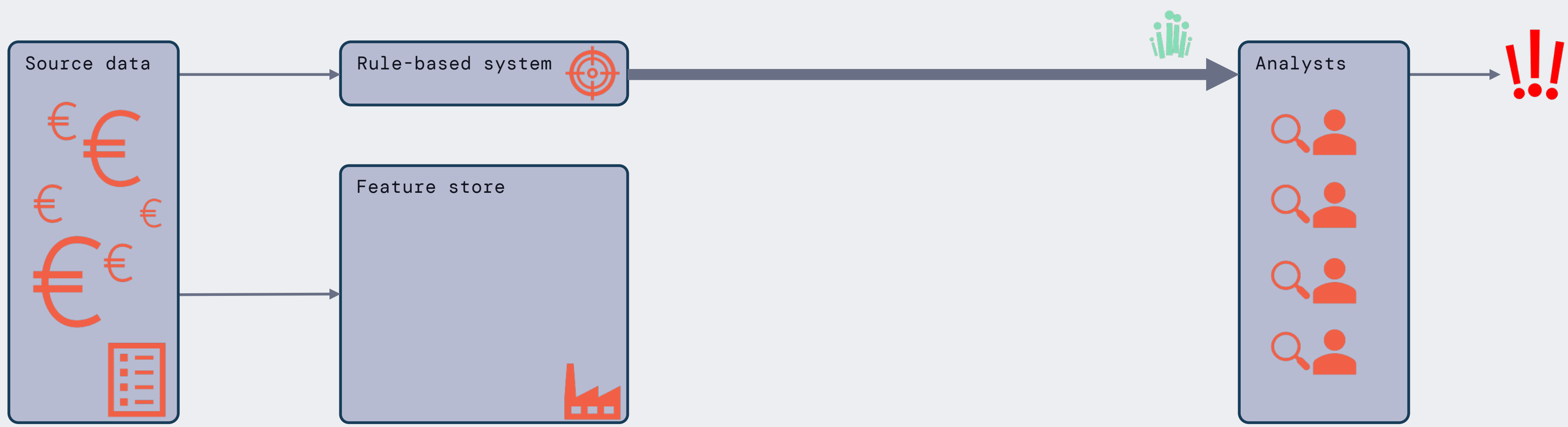
Feature store

- Transaction data is rich (e.g. sensitive)

<u>Obvious possibilities</u>
- Categories
- Periodicity
- Expected behaviour
- Network

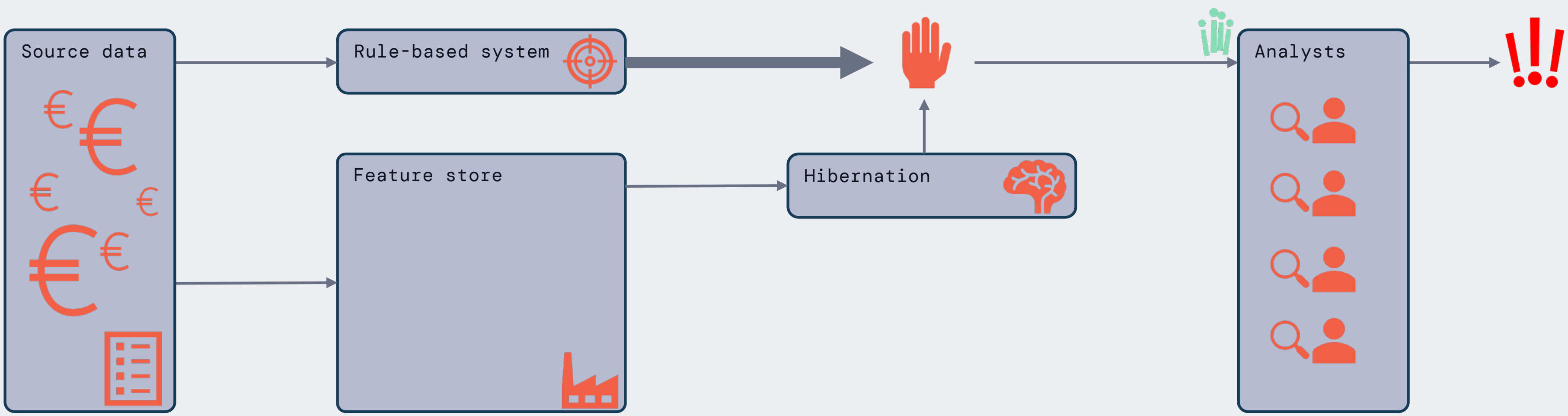
- **IT to combine and transform various sources continuously**

Feature store

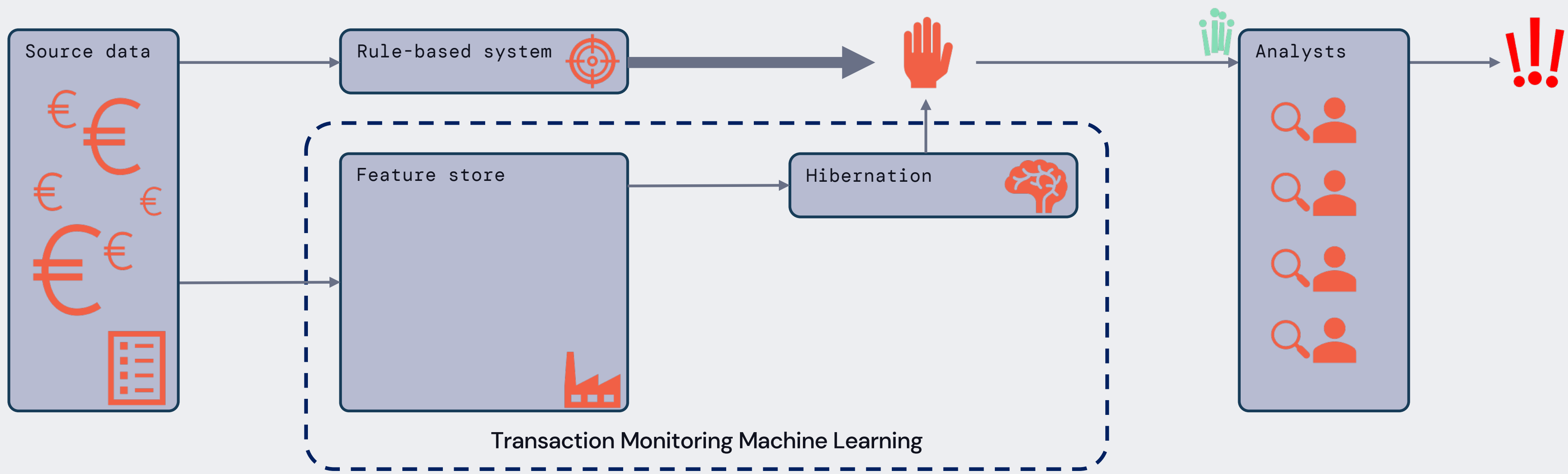


DATA+AI
SUMMIT 2022

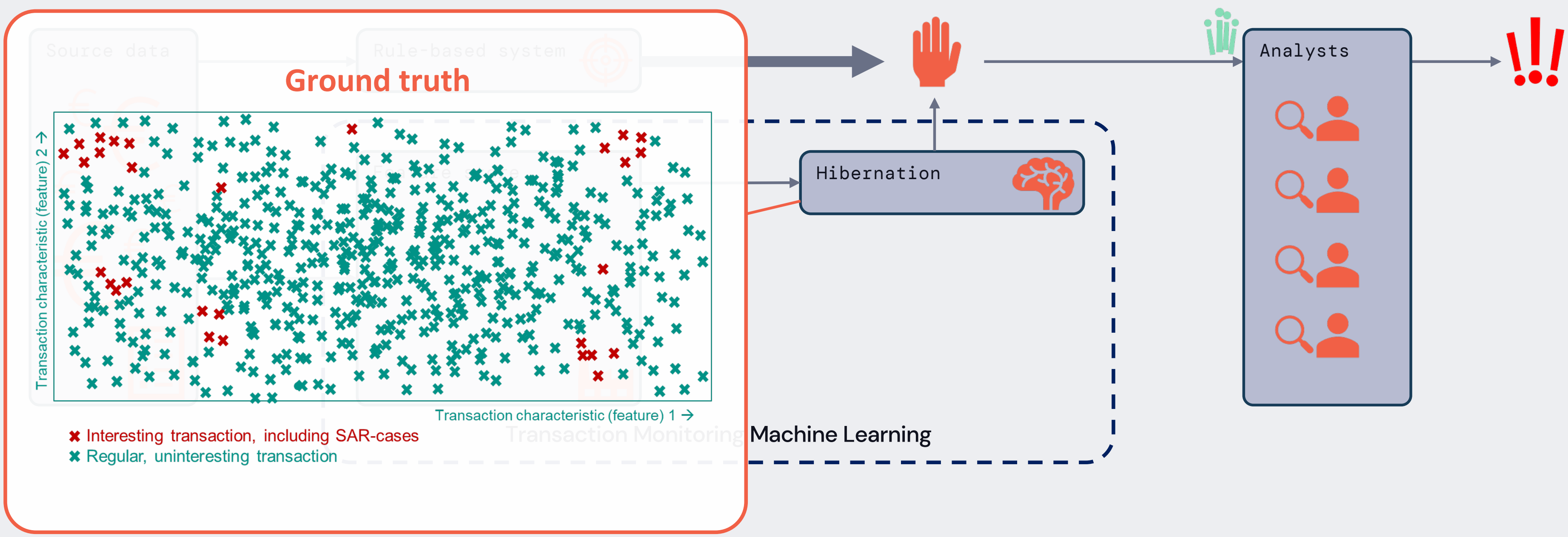
Supervised model – reducing false positives



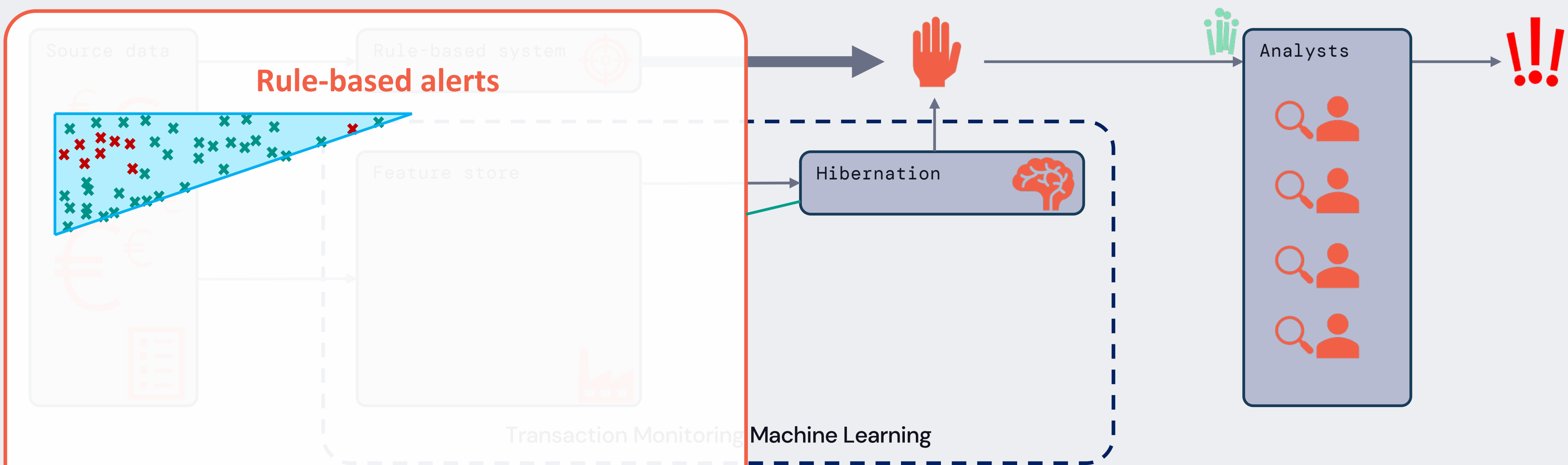
Supervised model – reducing false positives



Supervised model – reducing false positives

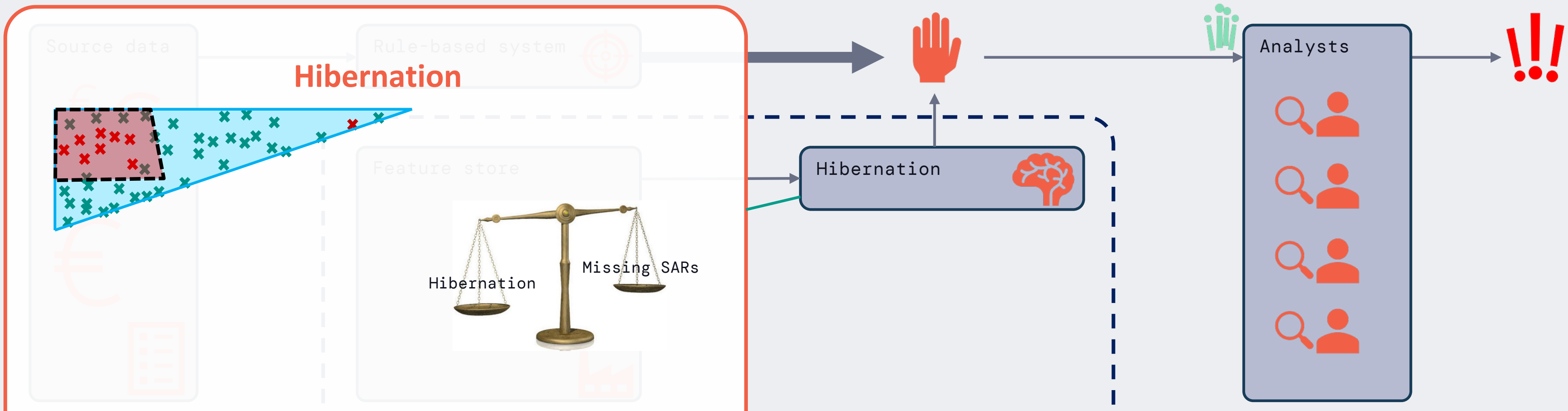


Supervised model – reducing false positives



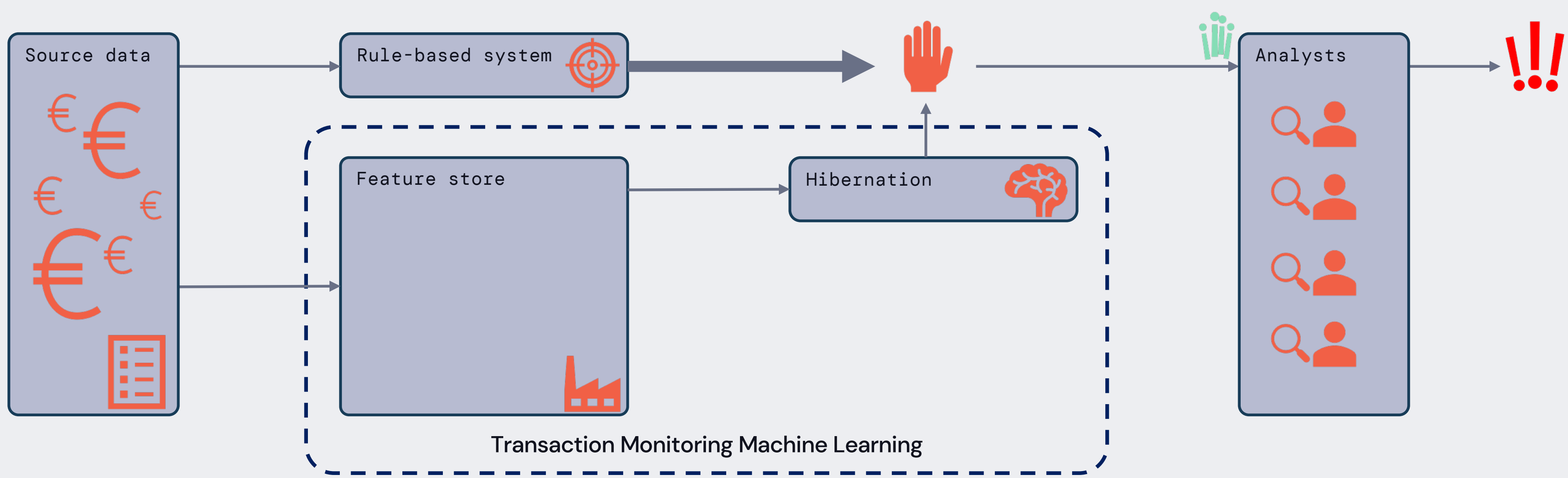
- Crude decision boundaries

Supervised model – reducing false positives

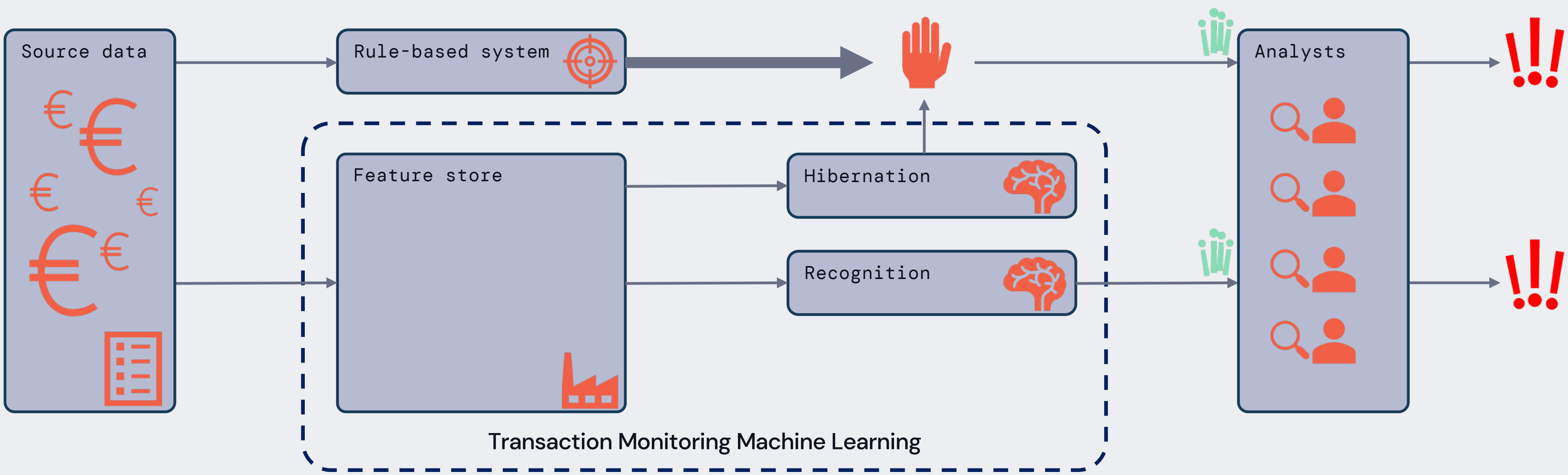


- Detailed decision boundaries
- XGBoost, 100+ features
- Cloud GPU training & inference

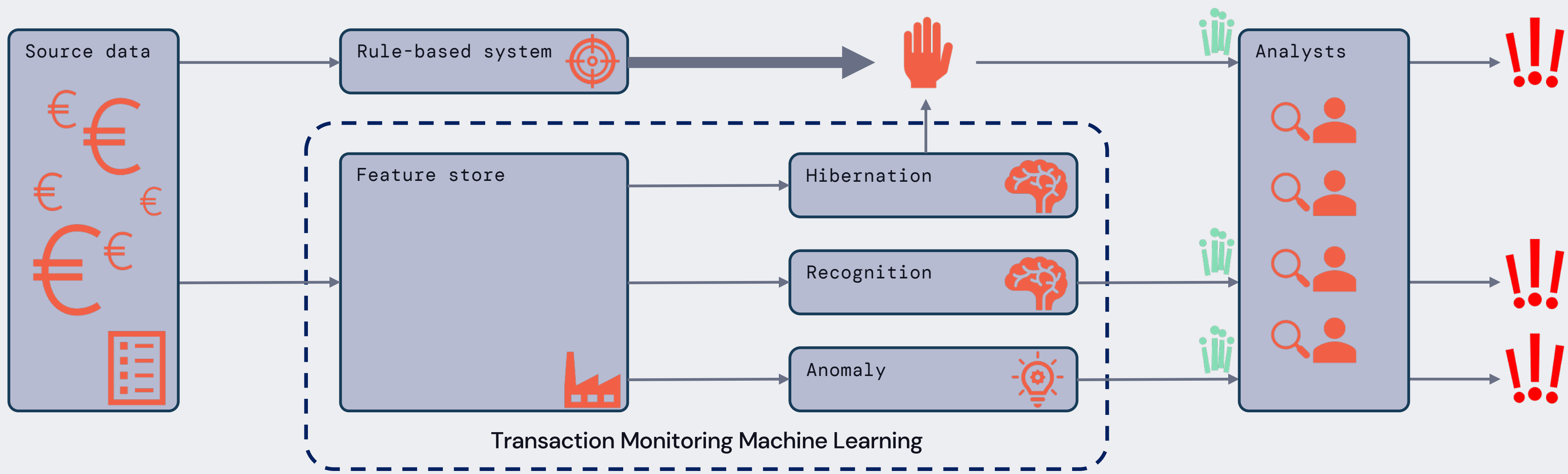
Supervised model – reducing false positives



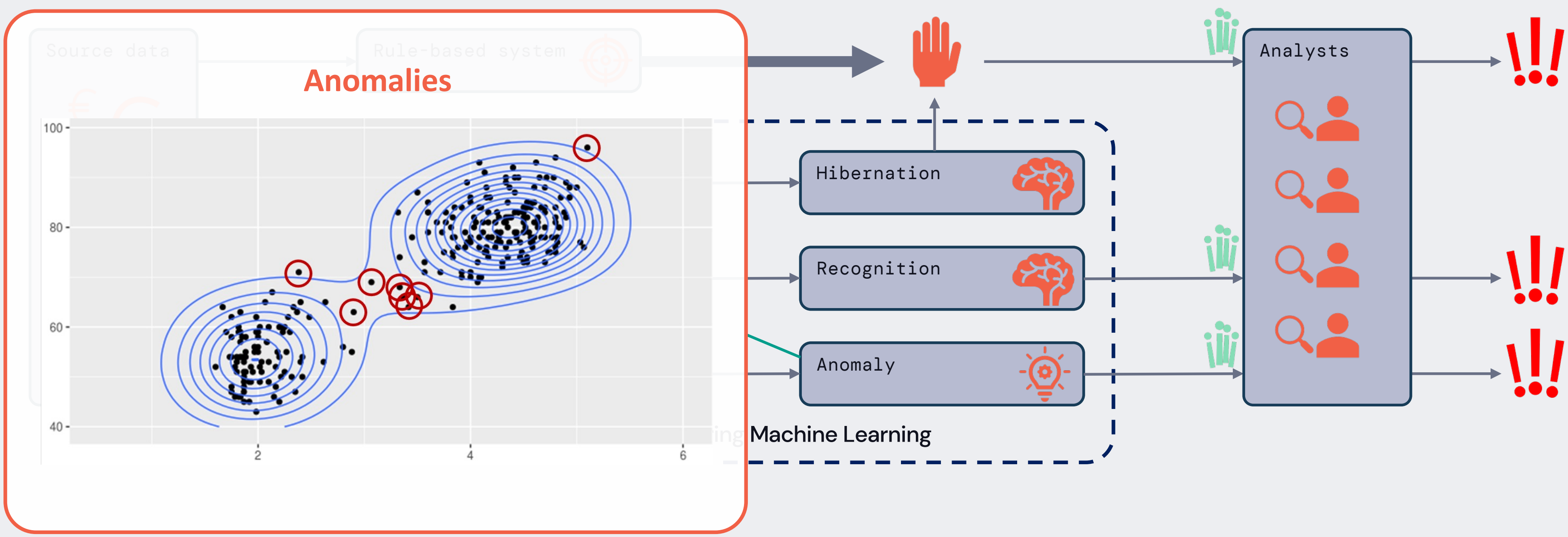
Supervised model – detecting similar cases



Unsupervised model – anomaly detection



Unsupervised model – anomaly detection



Unsupervised model – anomaly detection

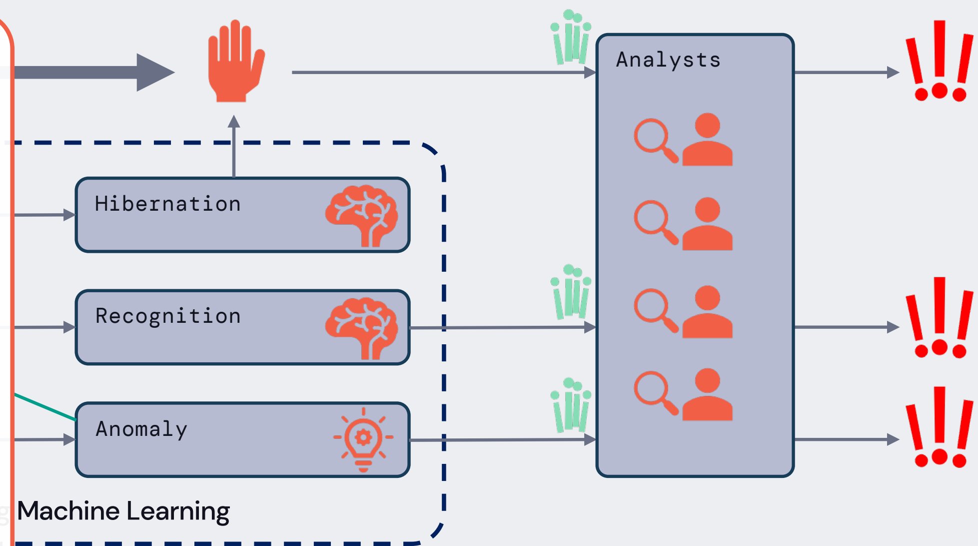
Anomaly methodology

<u>Features determine anomalies</u>

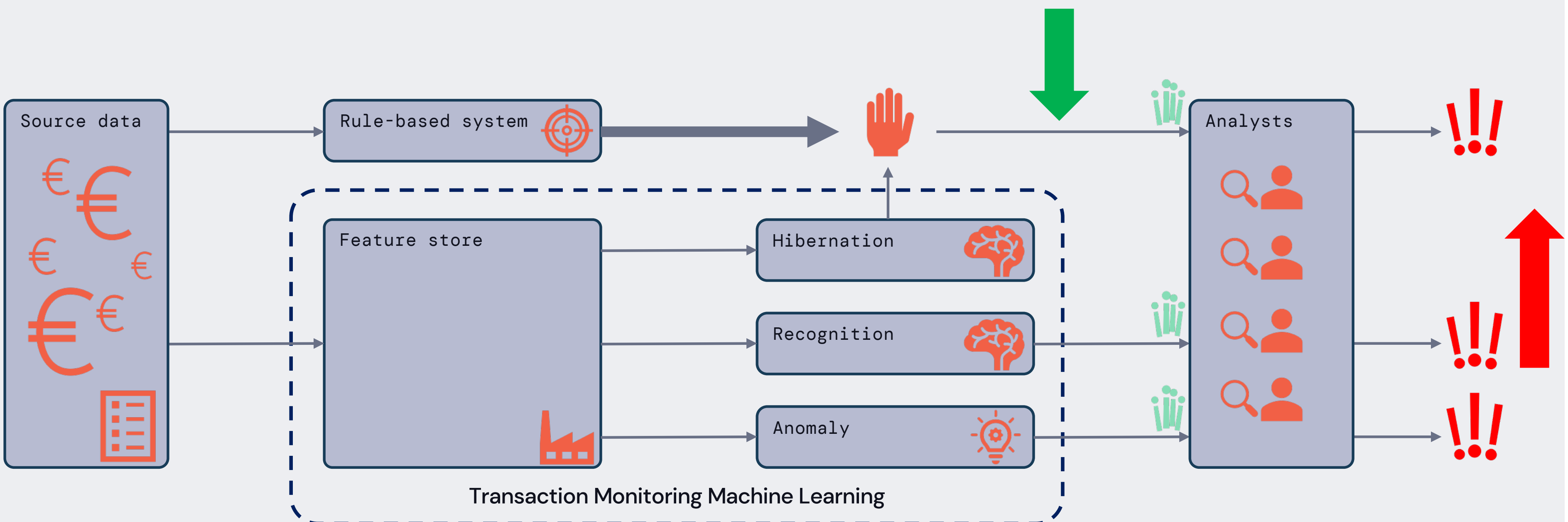
- ‘Normal’ vs. suspicious
- Informed by laundering patterns and other risks

<u>Soundness for validation without labels</u>

- Ensemble: VAE NN, isolation forest, cluster-based
- Empirical verification



Results



- Less false positives, more Suspicious Activity Reports

Results – business value

False positive reduction: refocus work

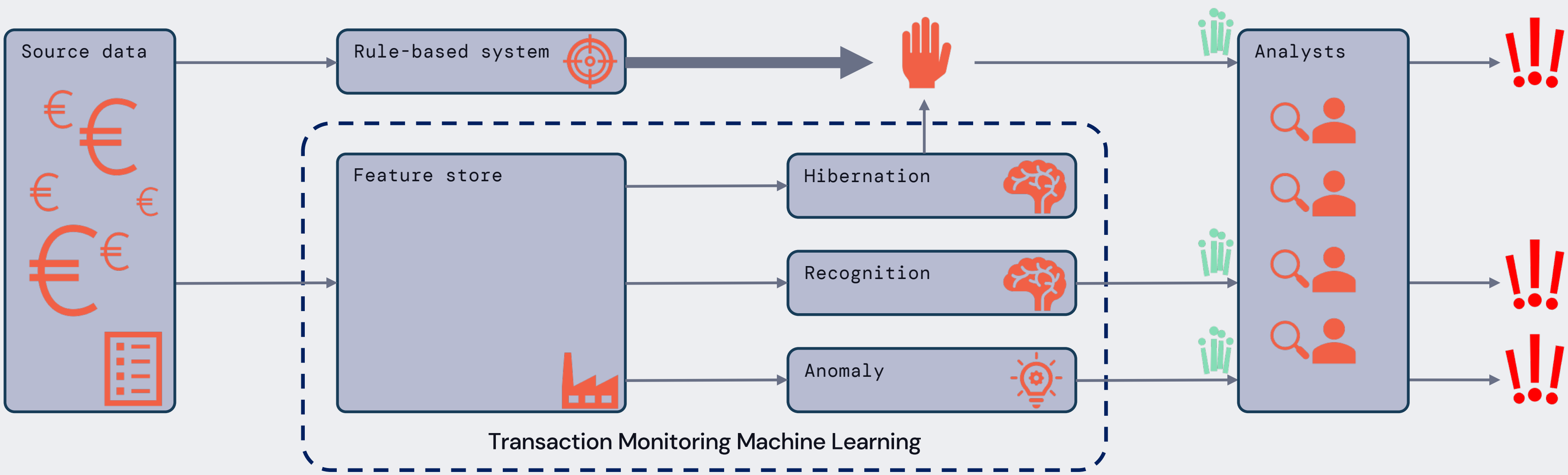
- Analyst capacity is needed elsewhere (e.g. anomalies, new models)
- Hibernating false positives saves analysts unrewarding, repetitive work
- Analysts need exposure to actual suspicious behaviour

Increased vigilance

- Recognition model extrapolates existing knowledge of SAR-filings
- Anomaly model expands existing knowledge, searching for ‘unknown unknown’ suspicious behaviour

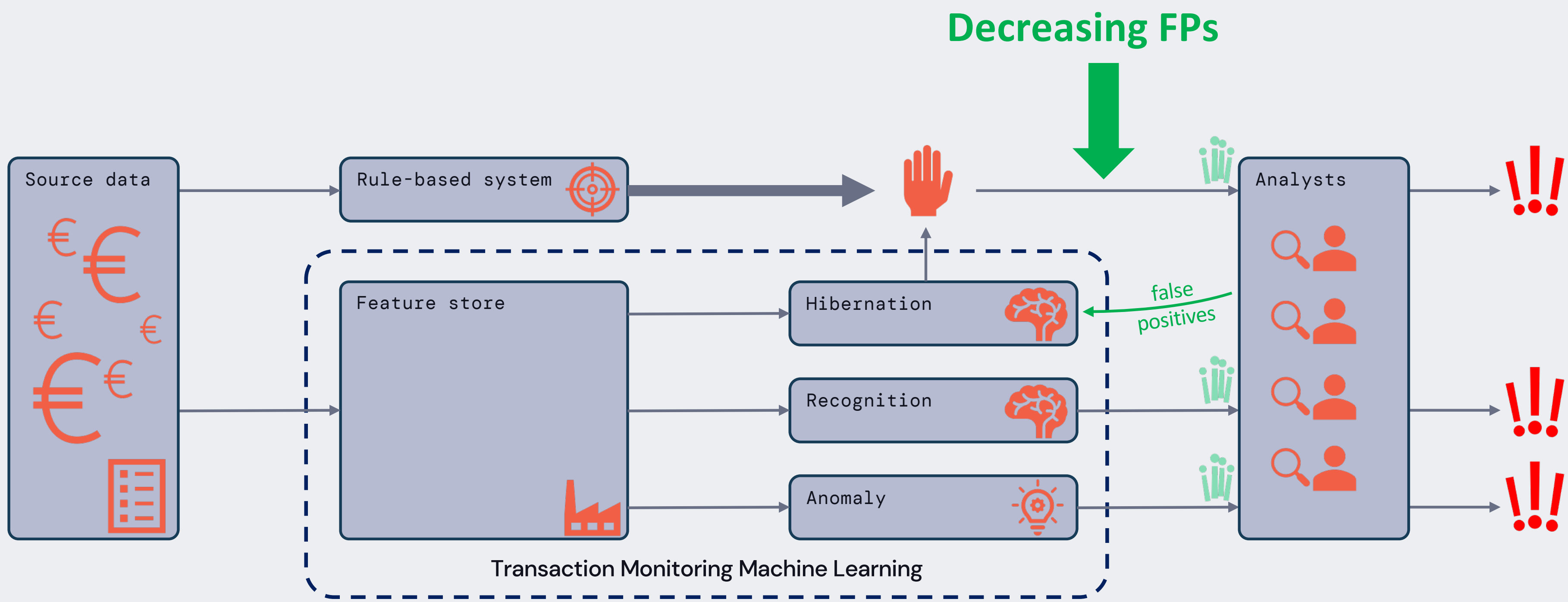
Business value supports data science + developer teams

Continuous retraining



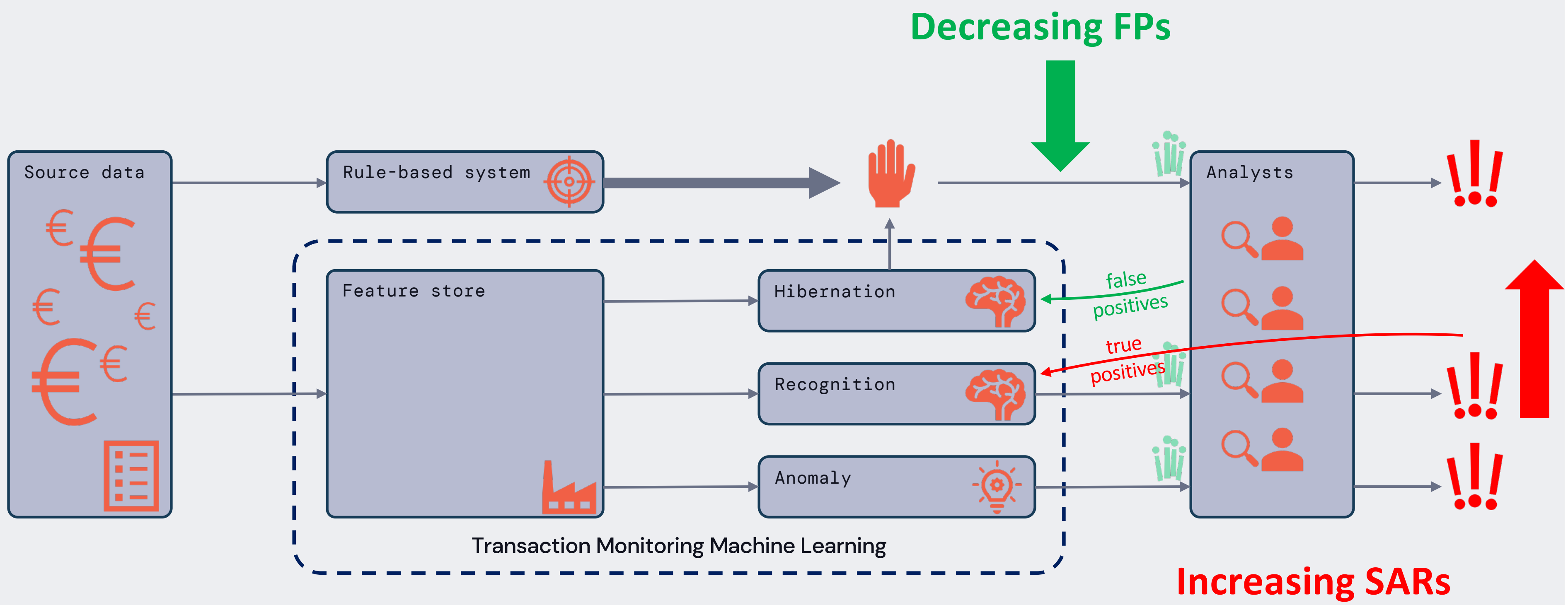
DATA+AI
SUMMIT 2022

Continuous retraining – human-in-the-loop



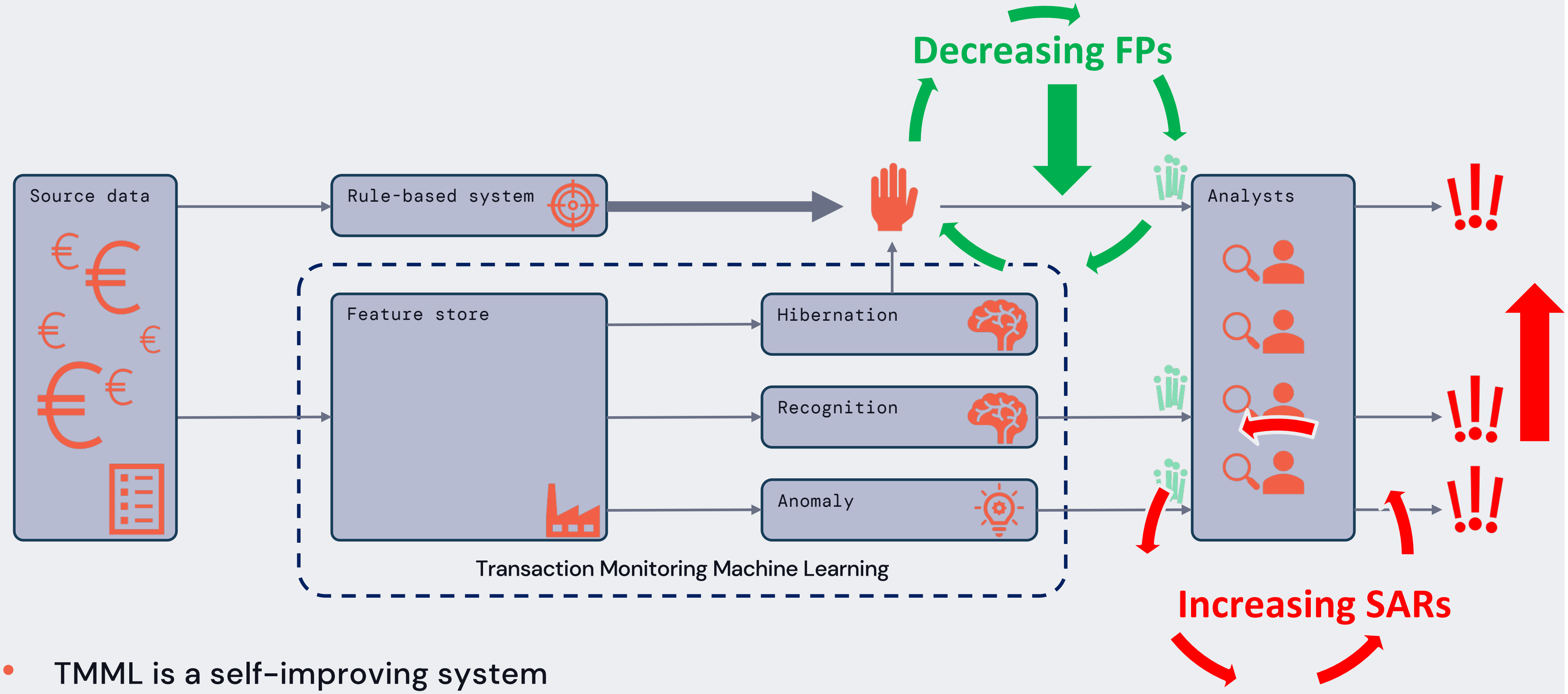
- Rule-based system + hibernation is self-correcting

Continuous retraining – human-in-the-loop



- TMML memory grows

Continuous retraining – human-in-the-loop

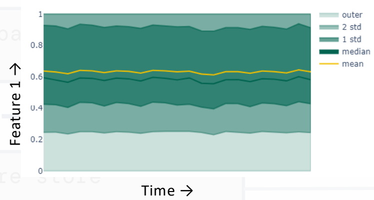


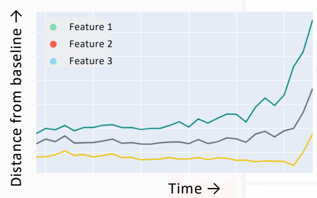
- TMML is a self-improving system

Continuous retraining – monitoring

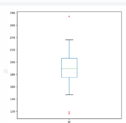
Feature drift
- Distribution percentile drifting
- Statistical distance from training distributions

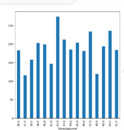
Wasserstein
Jensen-Shannon





Model behaviour
- Output expectations
- Fairness





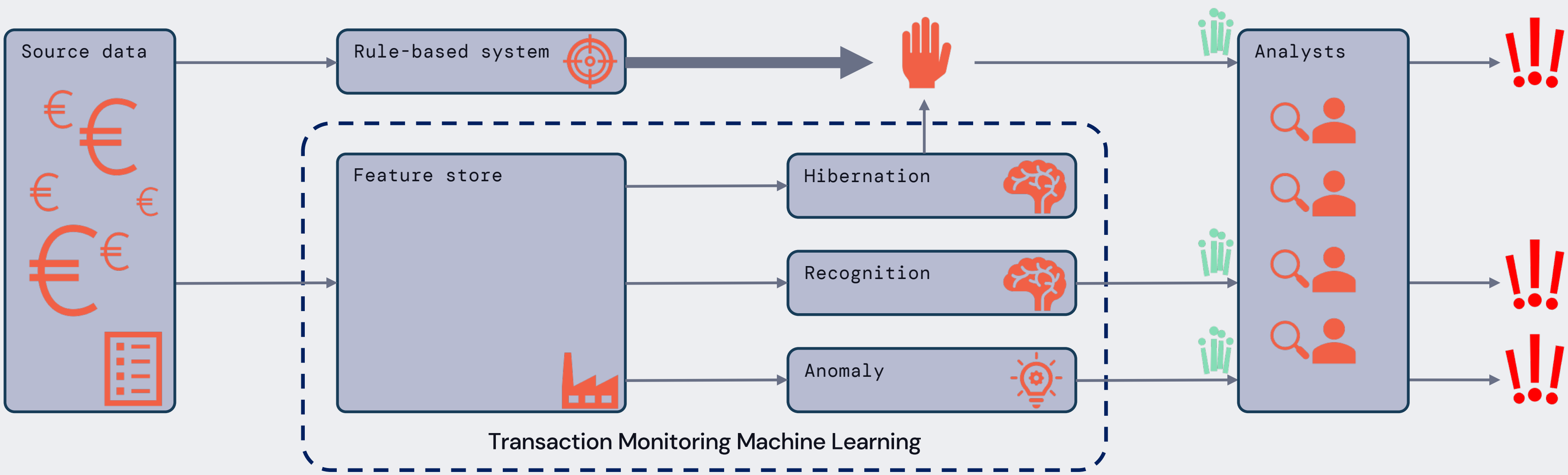
Hibernation

Recognition

Anomaly

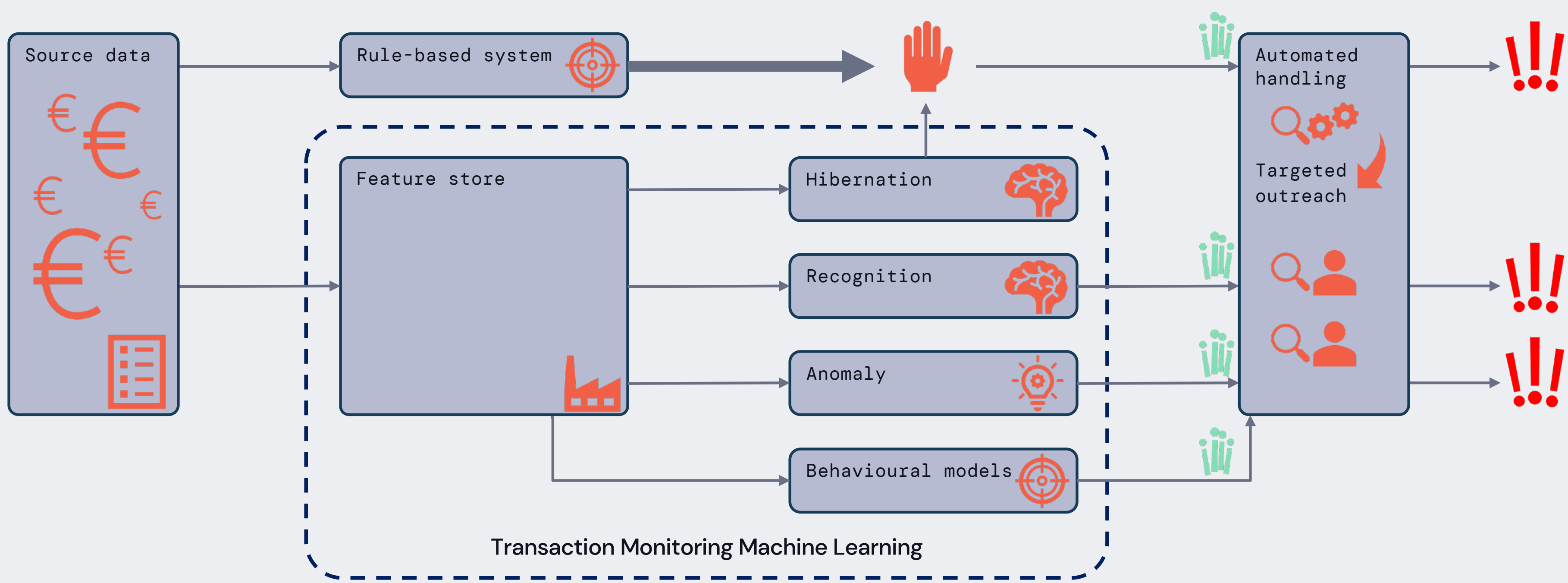
Retraining integrity
- Overlap between model versions
- Performance comparison
- Hyperparameter validity
- Stability
- Overfit

- **Model validation**

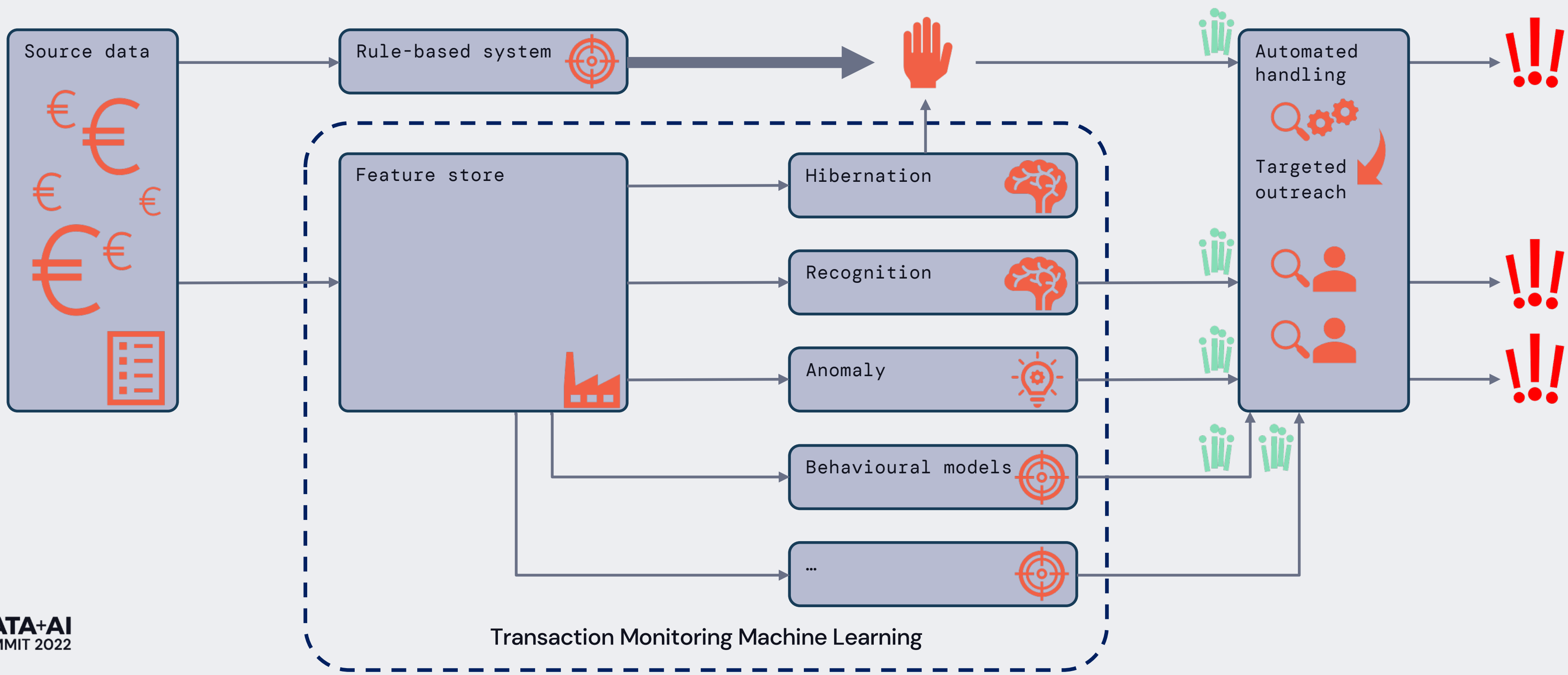


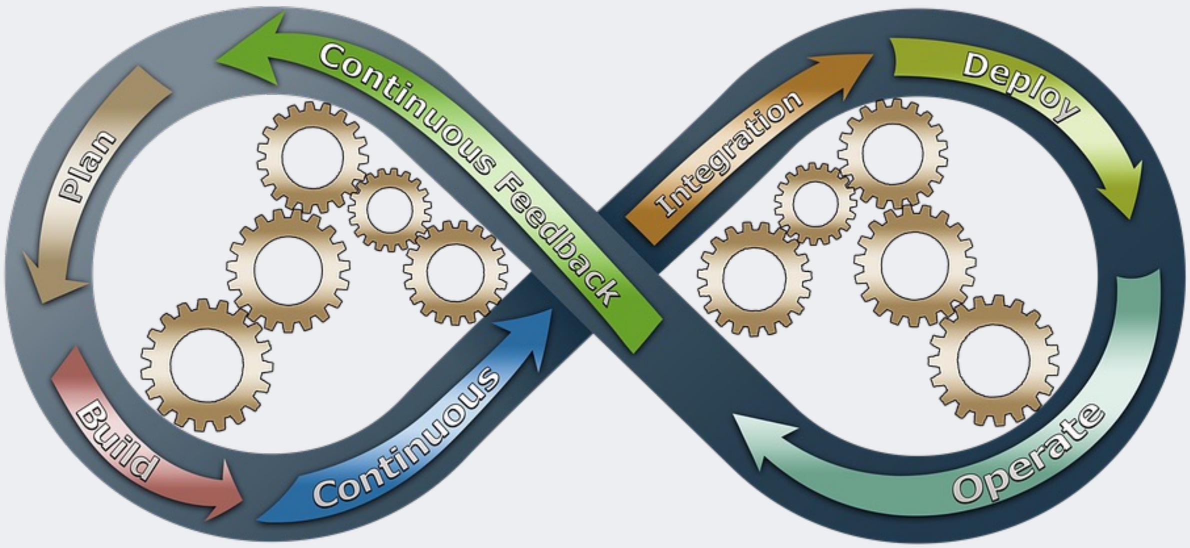
DATA+AI
SUMMIT 2022

Flexible setup – expandability



Flexible setup – expandability




Plan
Build
Continuous
Continuous Feedback
Integration
Deploy
Operate

Delivering business value by data science

A Common Story

ginablaber
@ginablaber

The story of enterprise Machine Learning: “It took me 3 weeks to develop the model. It’s been >11 months, and it’s still not deployed.” @DineshNirmalIBM #StrataData #strataconf

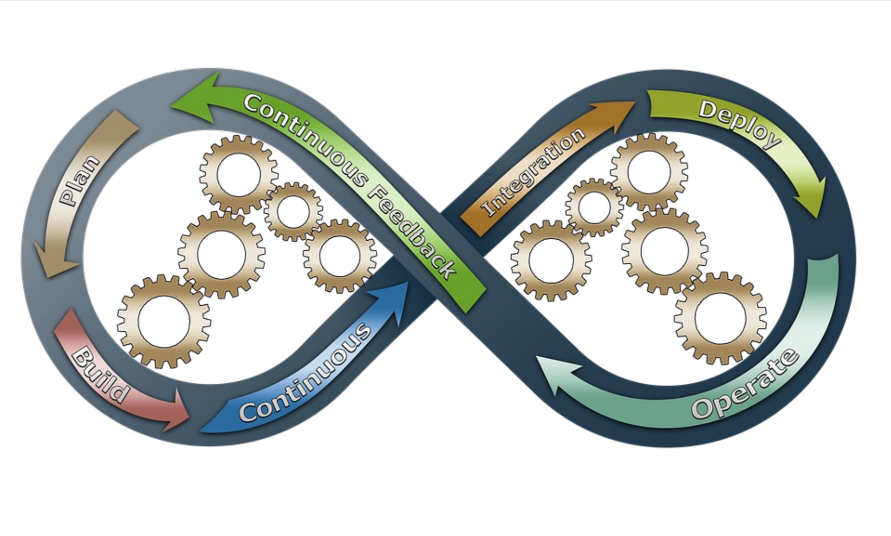
Tweet vertalen

Challenges associated with model productionalization

- Lack of suitable infrastructure
- Lack of a central model registry
- Alignment between IT engineers and data scientists
- Lack of multidisciplinary team
- Versioning and reproducibility
- Feedback and iteration
- Sharing and reusing features

DevOps principles

- Fast flow from Development to Operations
- Shorten and amplify feedback between teams
- Foster a culture of continual experimentation and learning



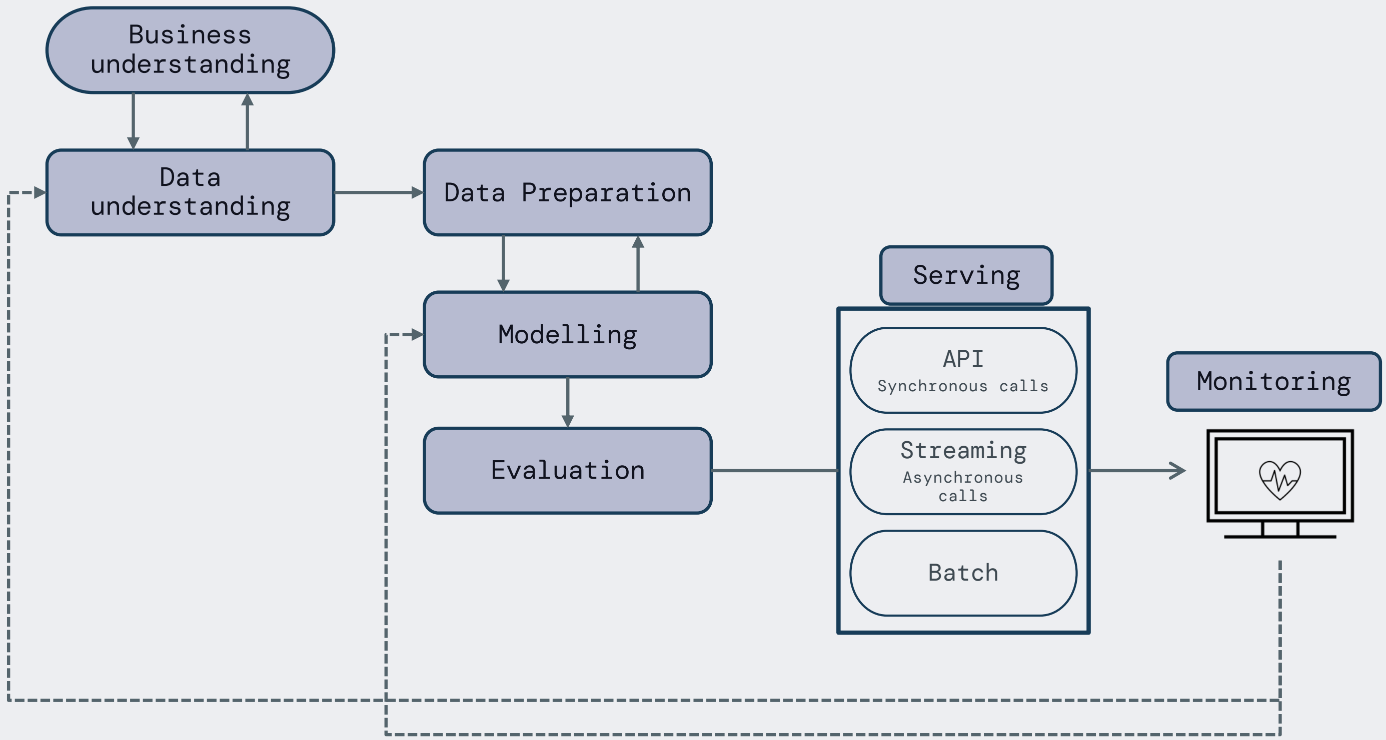
What is MLOps

- Agile ML engineering approach inspired by DevOps
- An approach in which a multi-disciplinary team develops and operationalizes machine learning solutions based on code, data, and models in small increments
- Fully automated deployment of ML model into production
- Reproducible and reliable workflows

Why did we adopt MLOps

- Productionalizing an increasing number of models without having a standardized framework was challenging.
- Retraining and monitoring of existing models, became an increasing bottleneck for data scientists.
- Lack of a framework to implement organizational quality gates.

A typical flow of model development and serving



Our tech stack

Databricks: Notebooks, clusters, Repos, spark, ...

mlflow **MLflow:** tracking experiments, central model registry

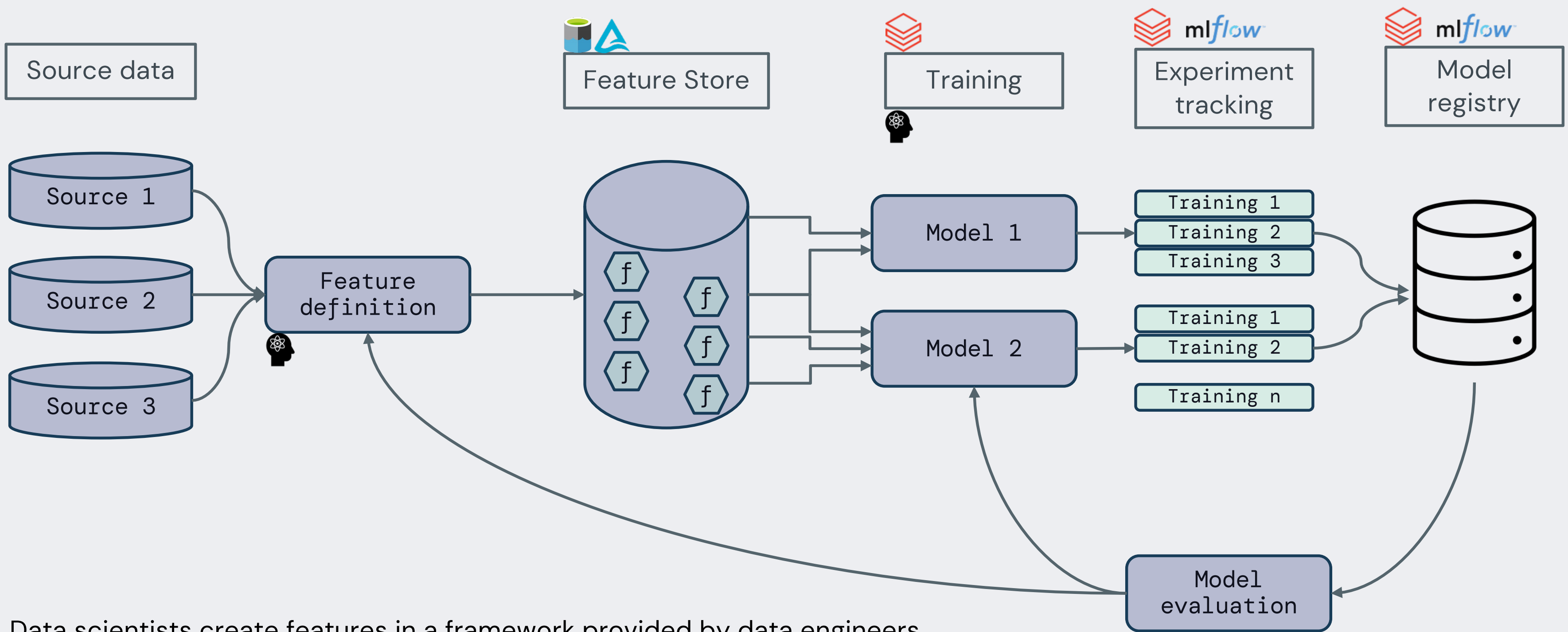
Azure repos: version control codebase

Azure pipelines: automated build and deploy pipelines

Azure Data Lake storage Gen2 and Delta Lake: feature store implementation

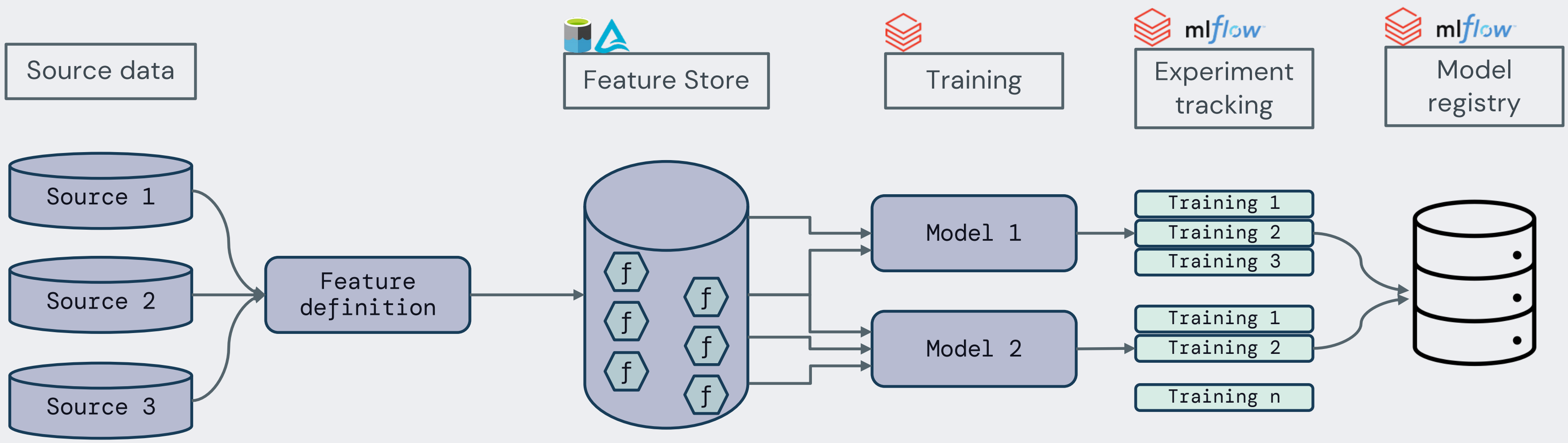
Azure Data Factory: orchestrating data movement and transforming data at scale

Exploration environment

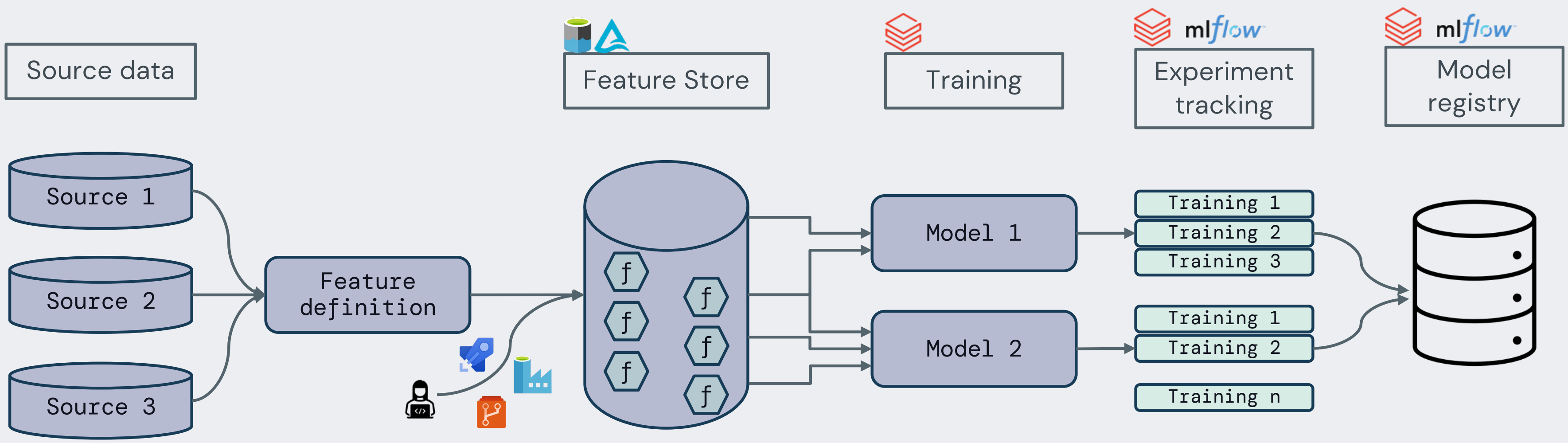


- Data scientists create features in a framework provided by data engineers
- Data scientists develop models in a framework provided by ML engineers

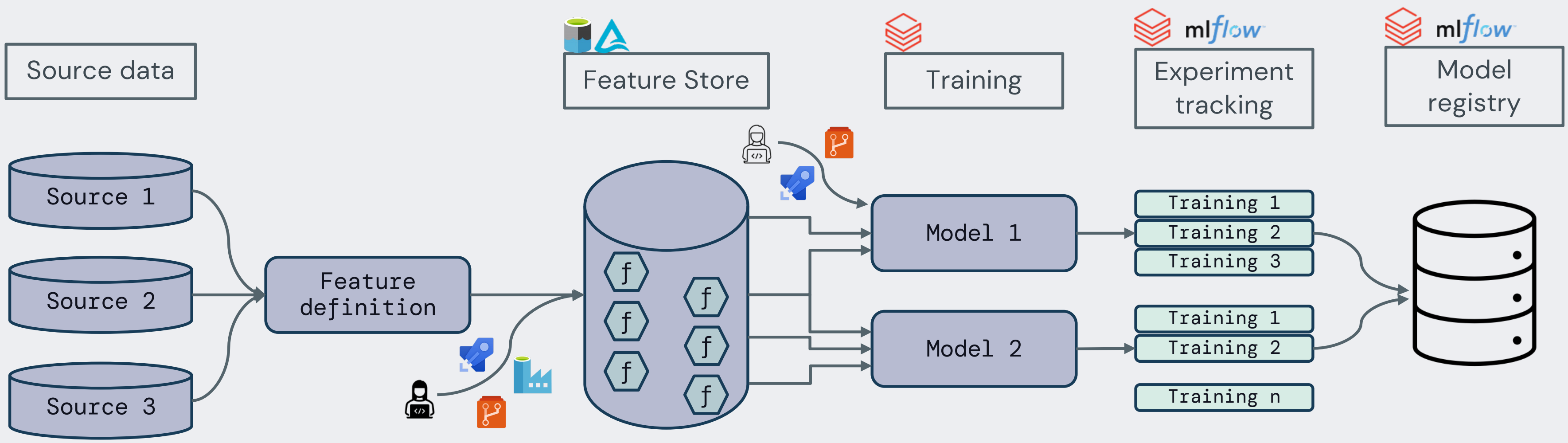
DTAP environment



DTAP environment

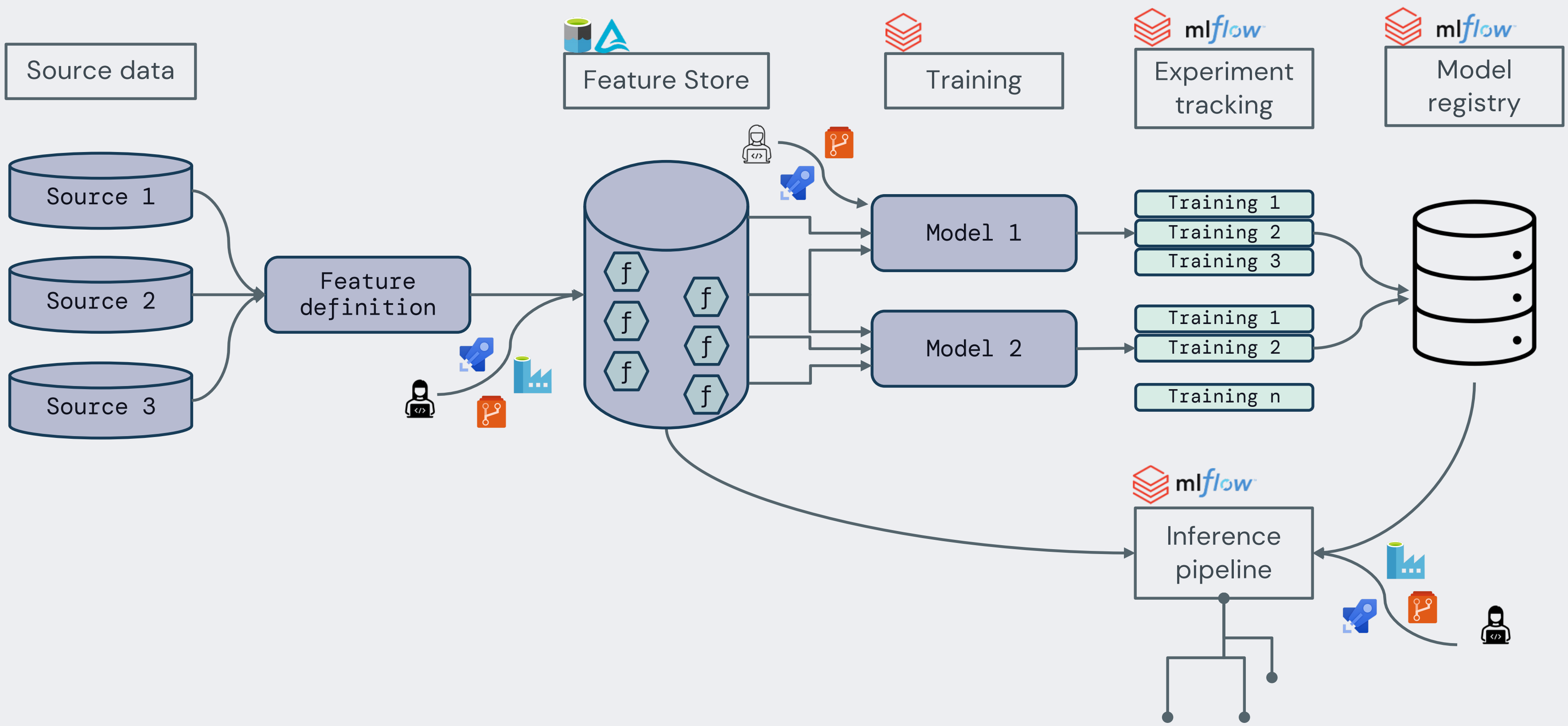


DTAP environment

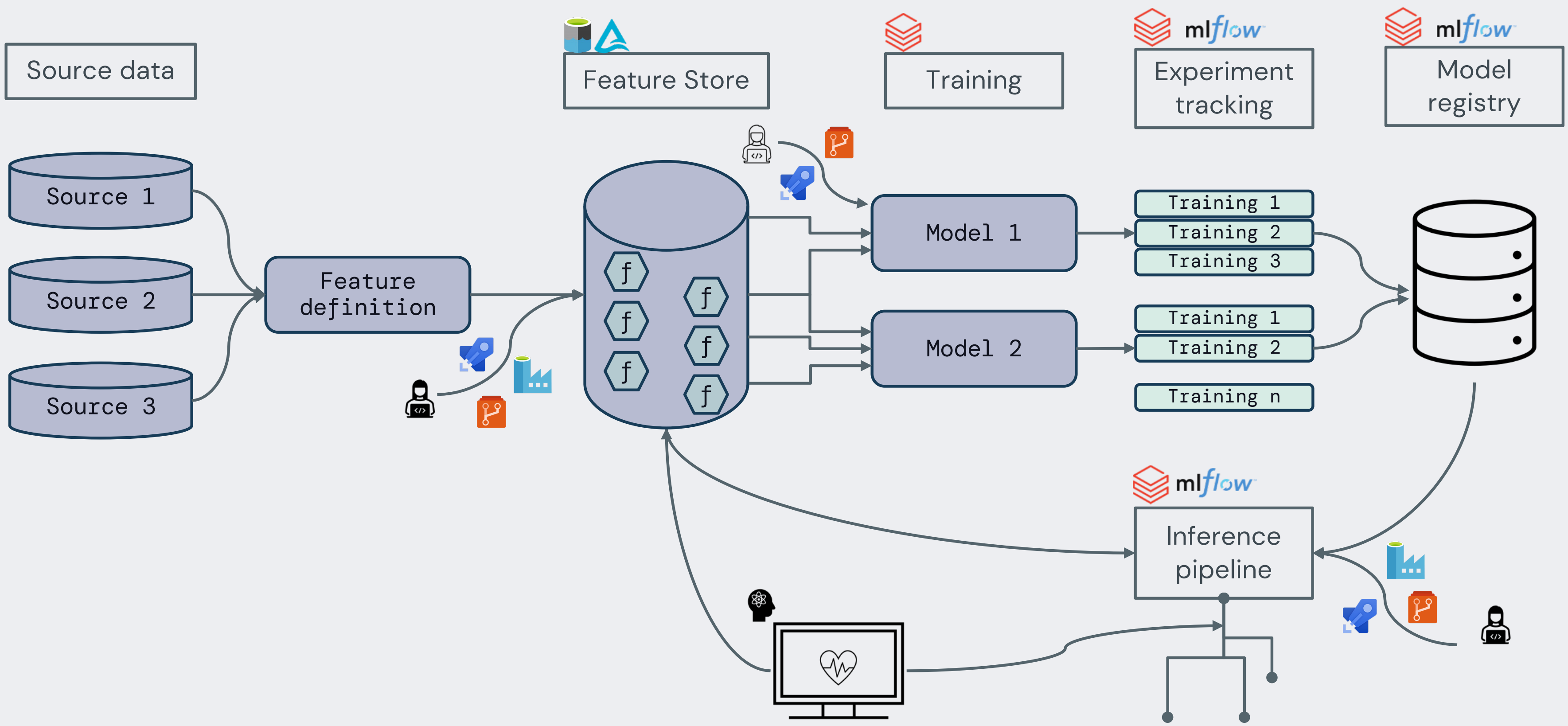


- New model implementations are shared with MLEs via pull requests

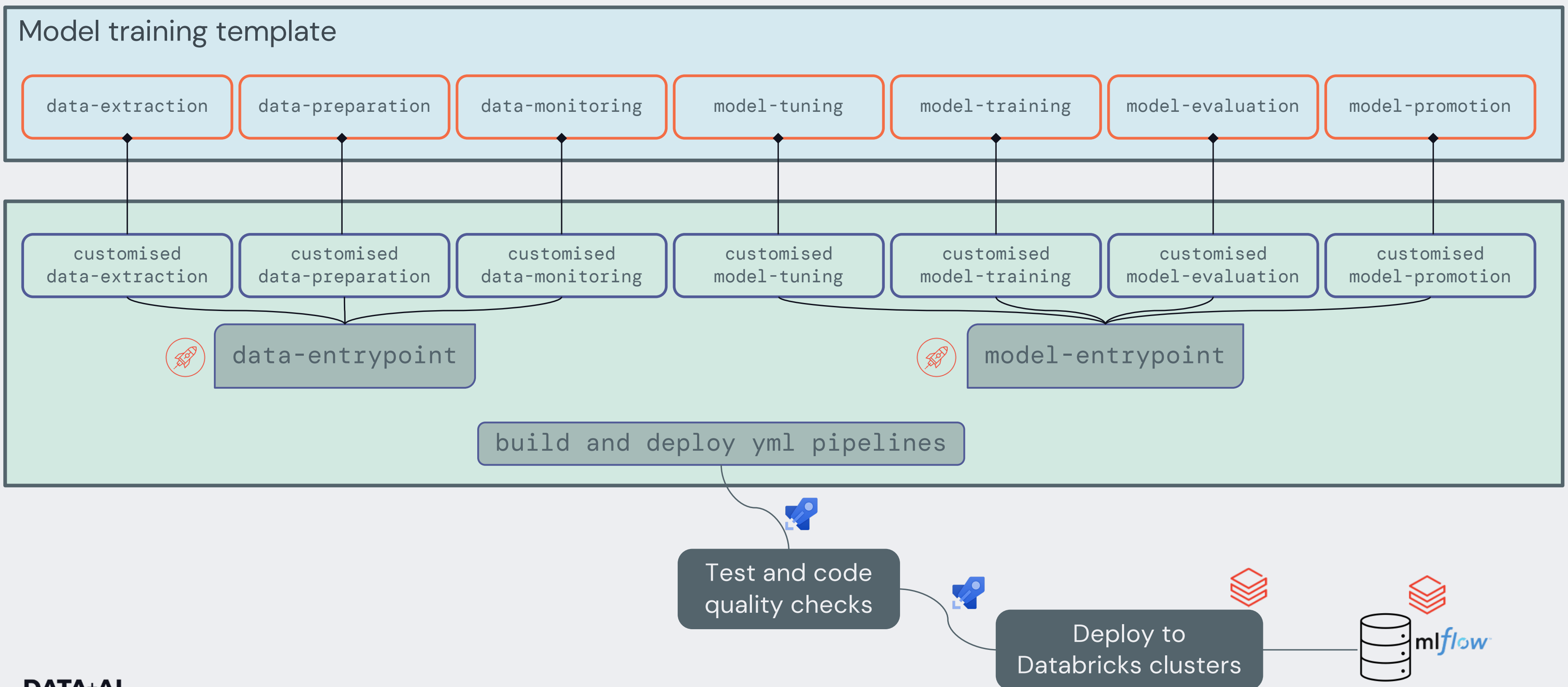
DTAP environment

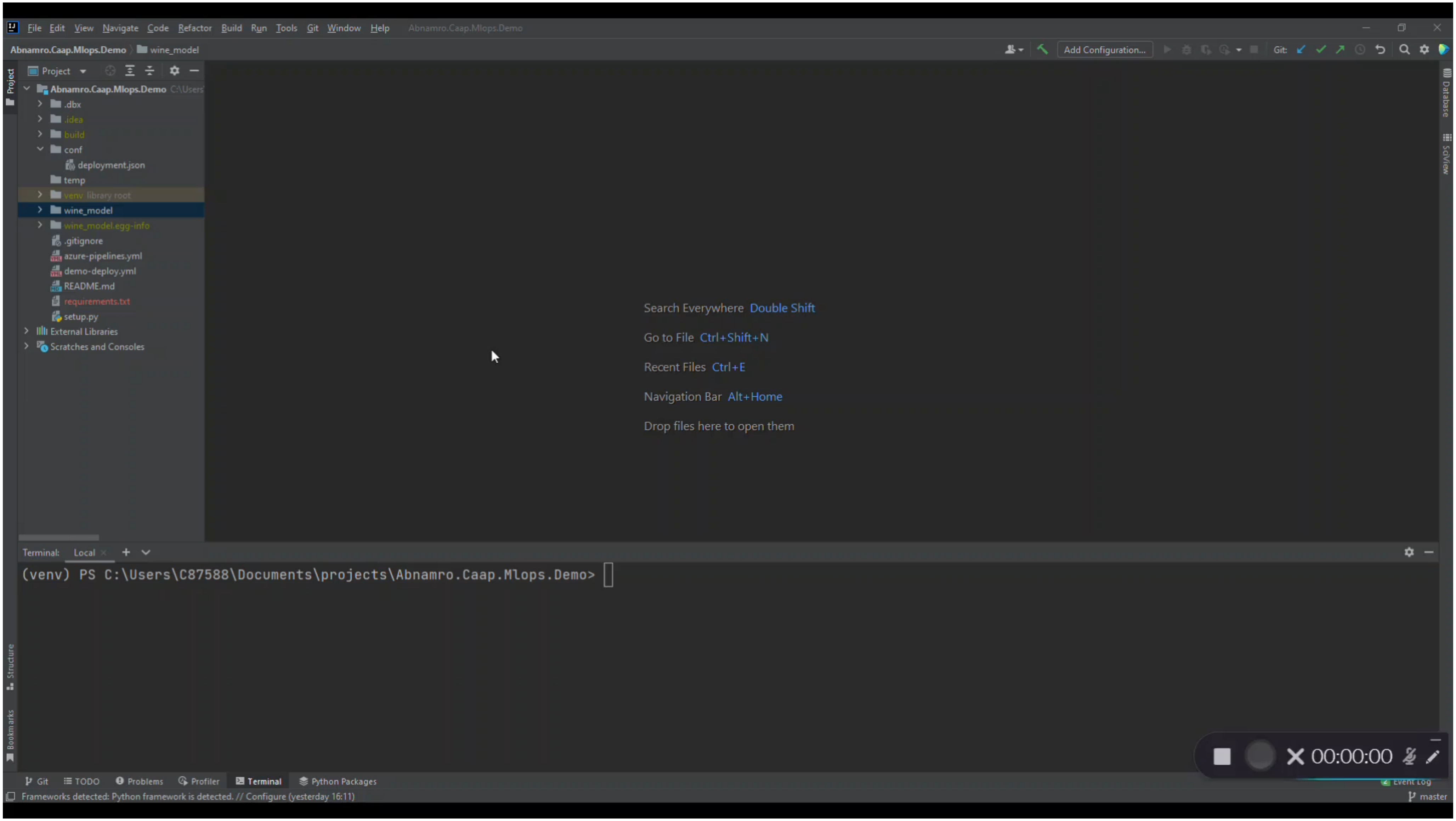


DTAP environment



Standardized training template




File Edit View Navigate Code Refactor Build Run Tools Git Window Help Abnamro.Caap.Mlops.Demo
Abnamro.Caap.Mlops.Demo › wine_model
Add Configuration...
Project
Abnamro.Caap.Mlops.Demo C:\Users
.dbx
.idea
build
conf
deployment.json
temp
venv library root
wine_model
wine_model.egg-info
.gitignore
azure-pipelines.yml
demo-deploy.yml
README.md
requirements.txt
setup.py
External Libraries
Scratches and Consoles
Search Everywhere Double Shift
Go to File Ctrl+Shift+N
Recent Files Ctrl+E
Navigation Bar Alt+Home
Drop files here to open them
Terminal: Local
(venv) PS C:\Users\C87588\Documents\projects\Abnamro.Caap.Mlops.Demo>
00:00:00
Git TODO Problems Profiler Terminal Python Packages
Frameworks detected: Python framework is detected. // Configure (yesterday 16:11)
master

Values delivered by applying MLOps

standardization of model development

built-in reproducibility

automated retraining

built-in quality assurance

standardization of quality checks

built-in traceability

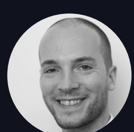
semi-automated monitoring

centralized model management

faster productionalization

DATA+AI SUMMIT 2022

Thank you



Lars Haringa, Data Scientist, ABN AMRO Bank

Saman Amini, ML Engineer, ABN AMRO Bank