European Commission | HORIZON EUROPE

UCD DUBLIN

MLSysOps

MLaaS – Decoupling Application Intelligence from Application Logic

John Byabazaire

Dimitris Chatzopoulos

Outline

- Background & Motivation
- System design
- Model registration
- Deployment and monitoring
- Drift monitoring and retraining
- Q & A

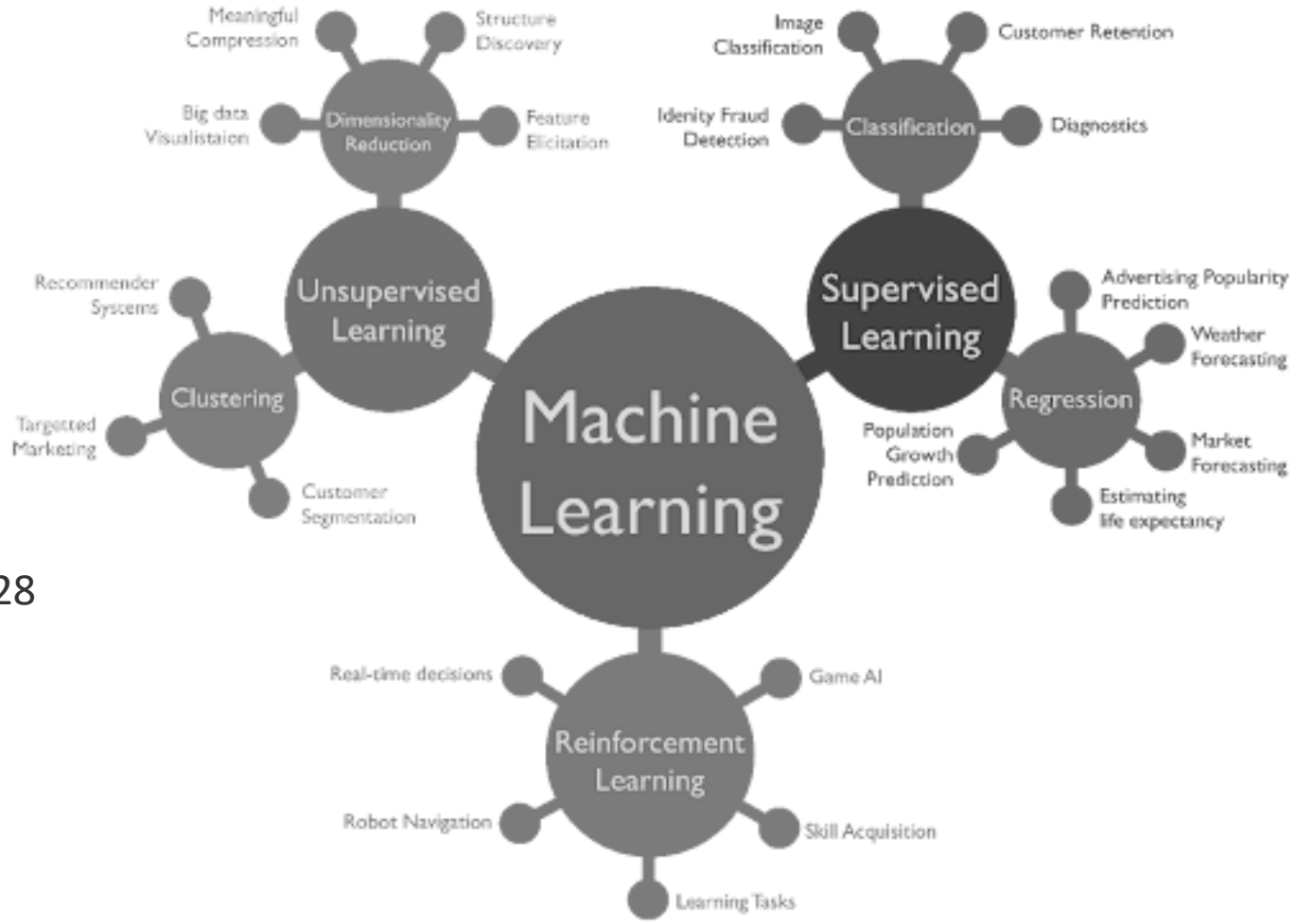
Background & Motivation

- The Internet is host to 55.7 billion connected devices, generating a staggering 73.1 zettabytes (ZB) of data
  - ✓ Clickstream logs
  - ✓ User data
  - ✓ Server logs
  - ✓ Transactions
  - ✓ Social media
  - ✓ Sensor data

A. Kishor and C. Chakarbarty, "Task Offloading in Fog Computing for Using Smart Ant Colony Optimization," Wireless Personal Communications, vol. 127, pp. 1683–1704, 11 2022

Background & Motivation

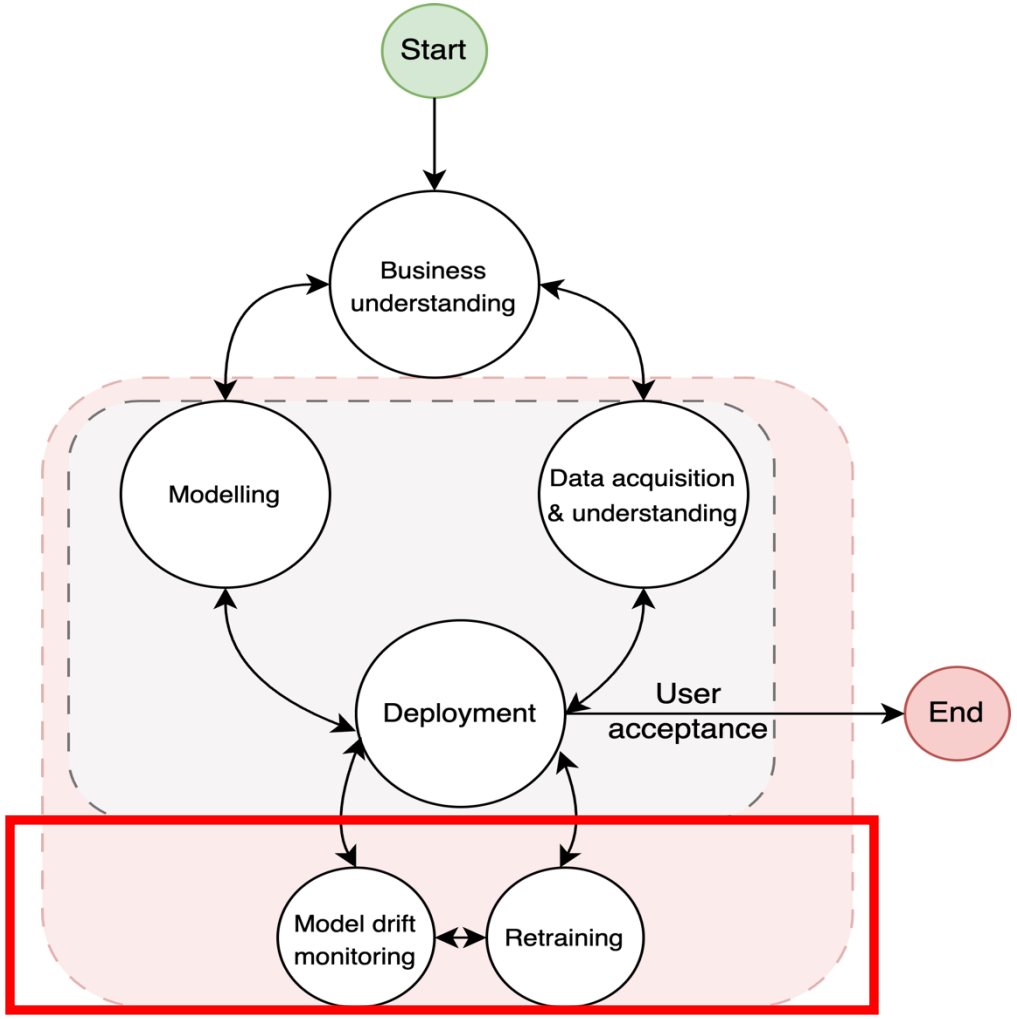
- Sales and marketing
  - 49% of organizations utilize ML and AI to identify sales prospects (Harvard Business Review Analytic Services)
- Healthcare
  - The global AI in healthcare market is projected to reach $187.95B by 2030 (Precedence Research, Statista)
- Banking
  - Estimated global market for AI in the banking, financial services, and insurance sector is expected to reach $15.32B by 2028 (Business Insider)
- Retail
  - The global AI in the retail market is expected to grow to $45.74B by 2032 (Precedence Research)



https://www.intuition.com/machine-learning-by-the-numbers-its-impact-on-business/

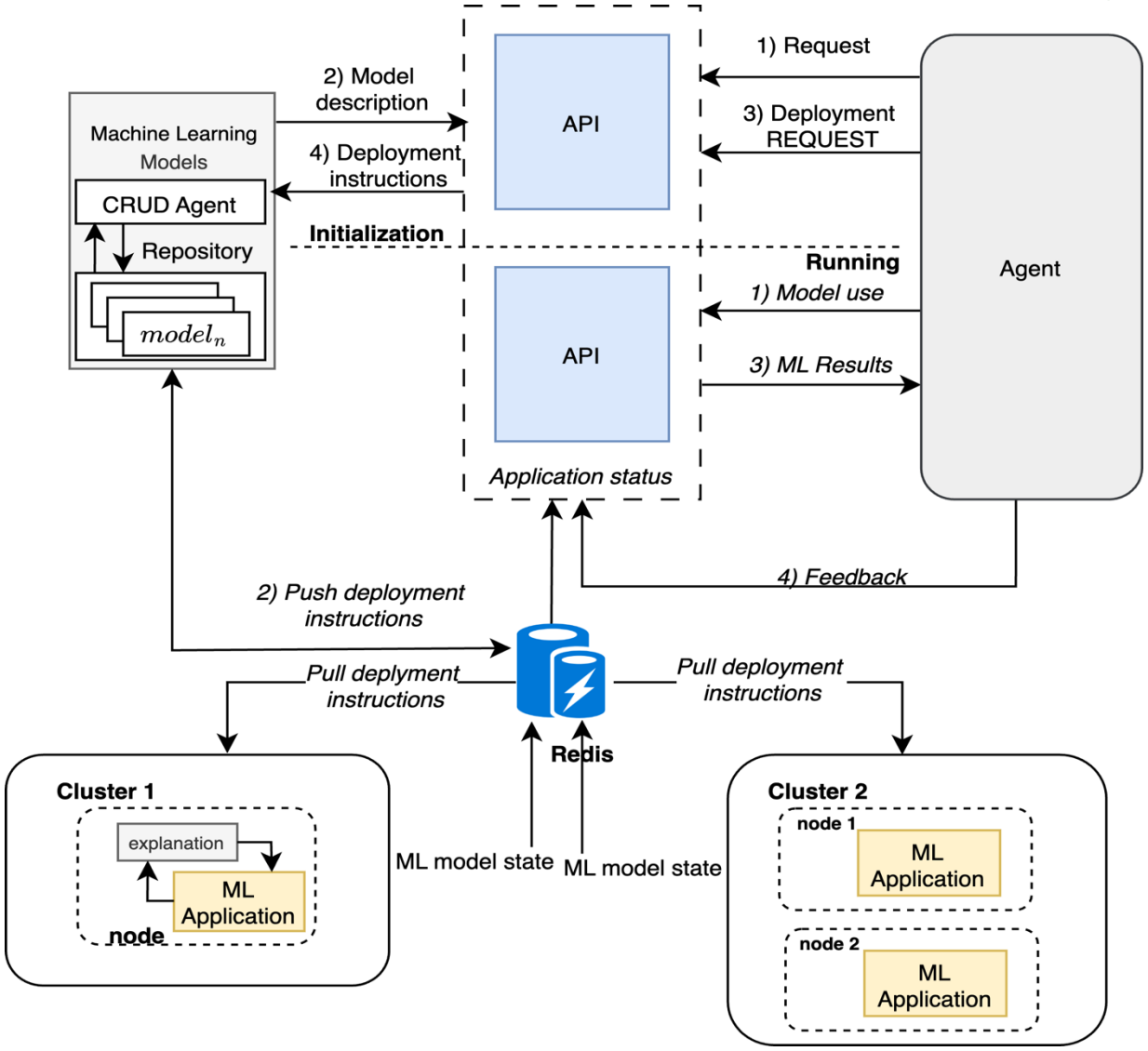
Background & Motivation

- High cost and expertise to setup scalable ML solutions
  - Big companies can invest
- Machine learning as a service (MLaaS)
  - Small and medium sized companies
  - Cost effective
  - High scalability
- ML life cycle
  - Problem definition
  - Data acquisition
  - Model development
  - Deployment
  - **Drift monitoring**
  - **Retraining**



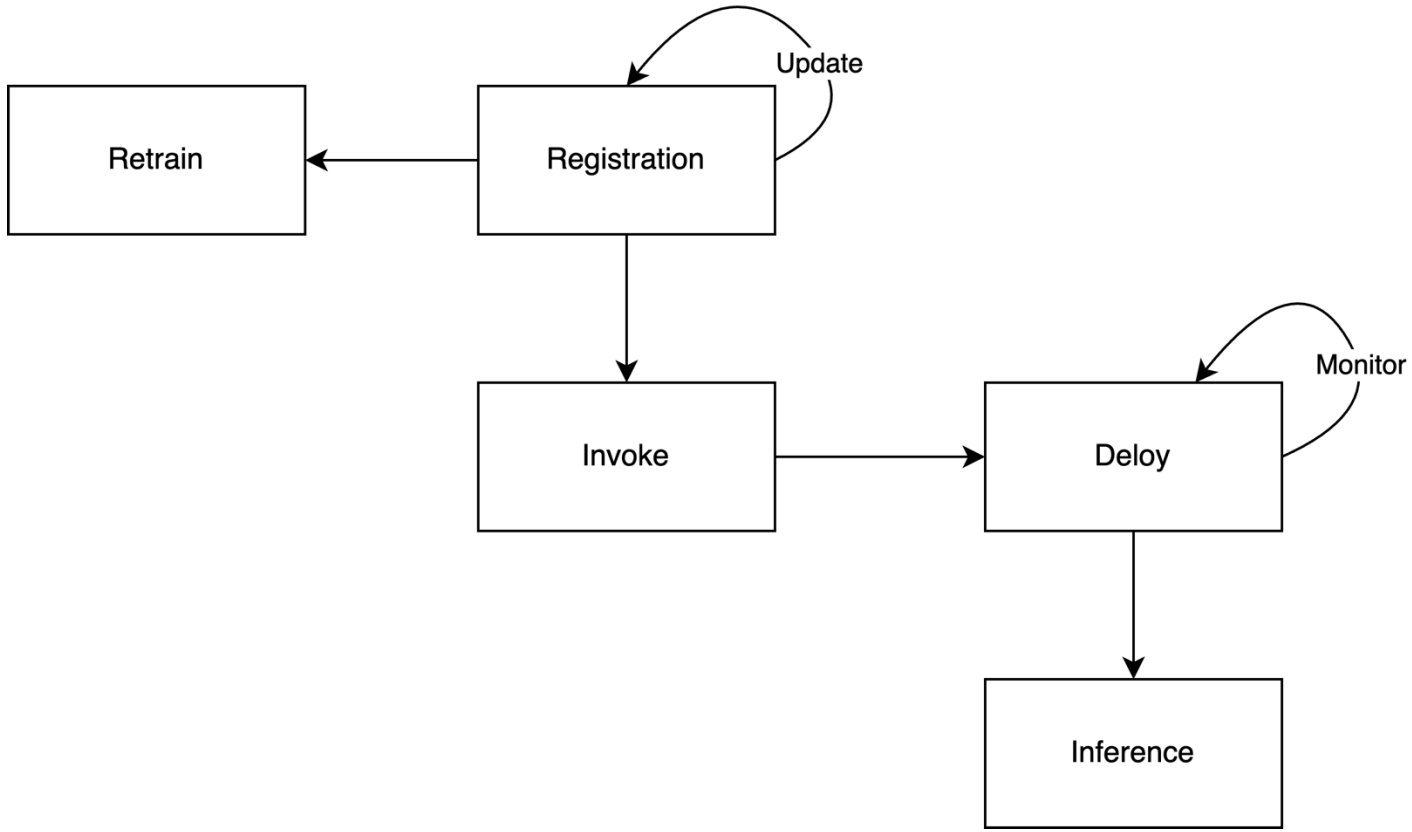
System design

- Based on micro-service architecture
- Two decoupled REST applications
  - API for registration and discovery
  - API for deployment, inference and drift monitoring



Model registration

- Model details
  - Eg classification
  - Training data
  - Hyperparameters
- Performance
  - Eg F1
- Deployment
  - Same node
  - Cluster
- Inference
  - Data format and input

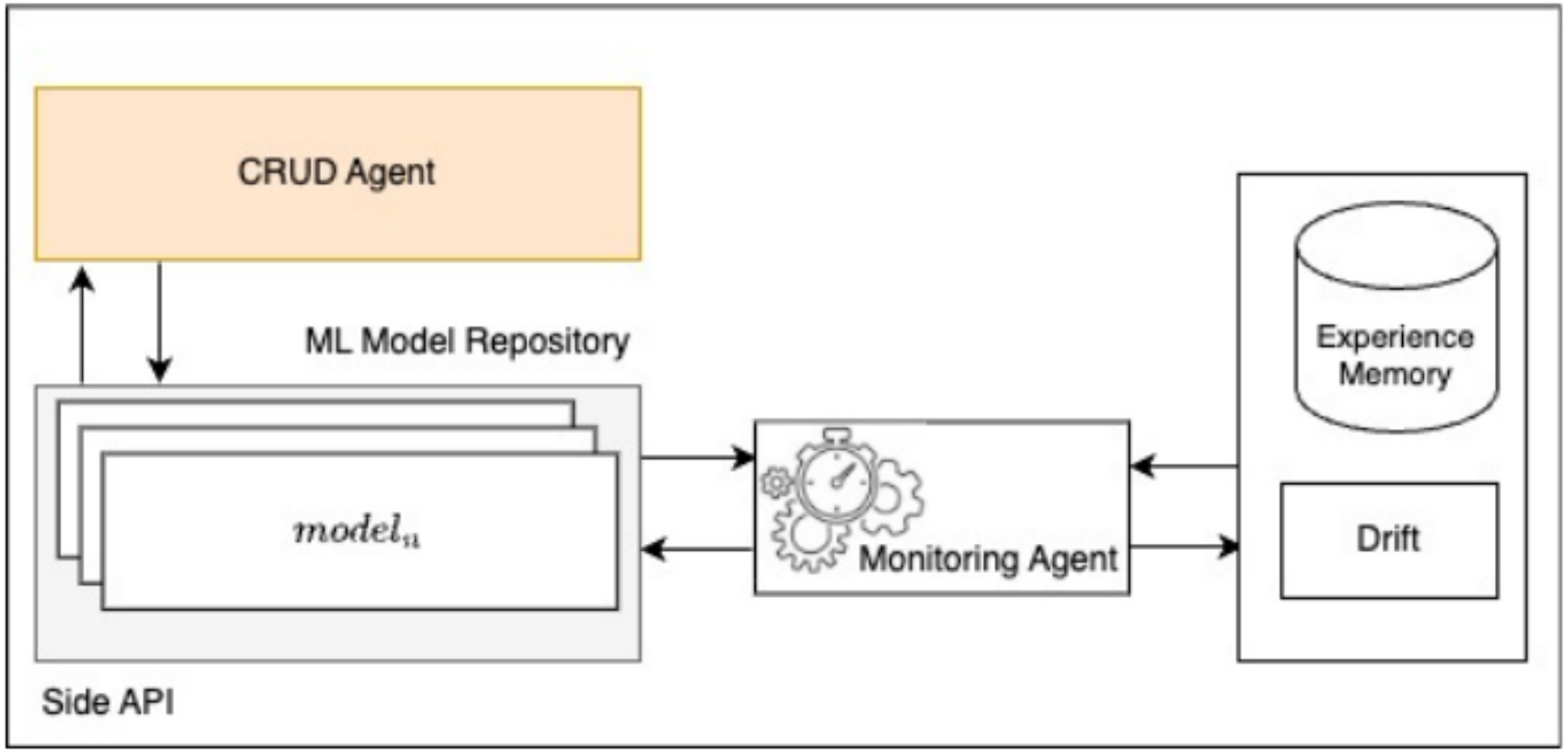


Model deployment, and monitoring

- Search and discovery
  - The agent queries a very specific ML model
  - Category of ML models
  - Complex search and discovery??
- Docker container with an inference endpoint
  - Generic classification model
- Can take static data, or reference to data
- Inject inference results back to side-api via telemetry

Drift monitoring and retraining

- Uses feedback loop to inject inference results back to side-api via telemetry



HORIZON
EUROPE

Thank you!

john.byabazaire@ucd.ie