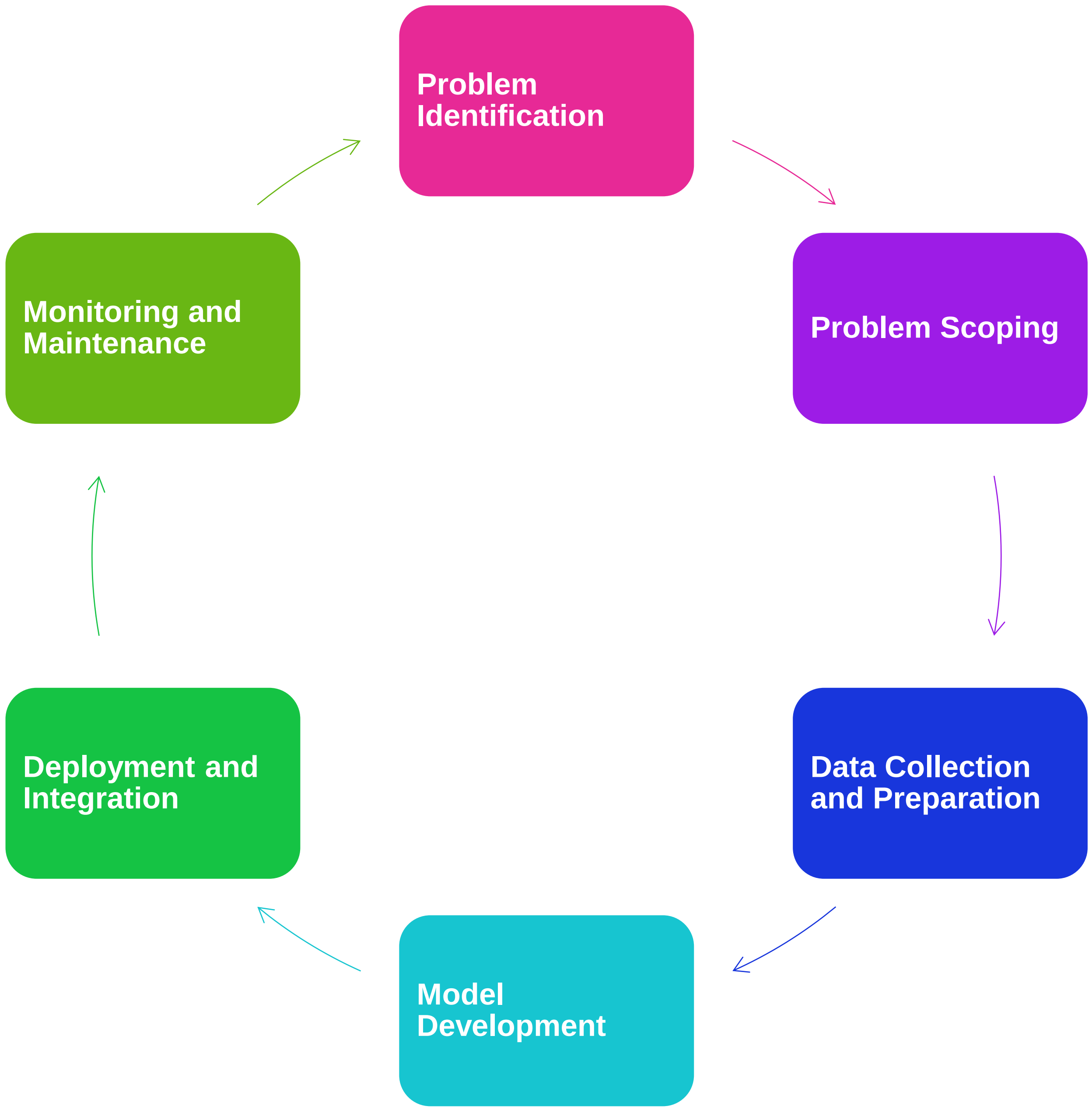
DEPLOYMENT AND INTEGRATION: DEPLOYING AI MODELS IN REAL-WORLD SETTINGS. INTEGRATION WITH EXISTING SYSTEMS AND APPLICATIONS. CONTINUOUS MONITORING AND UPDATES

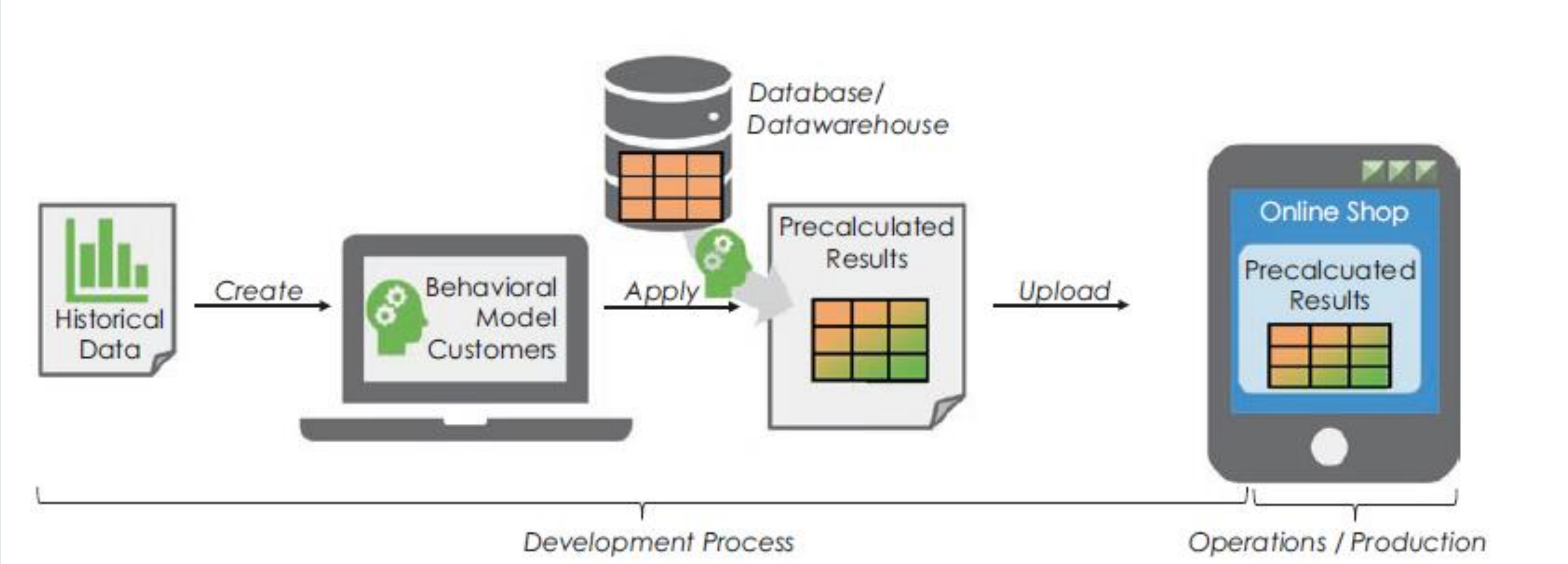
STAGES OF THE AI PROJECT LIFECYCLE



AI MODEL DEPLOYMENT AND INTEGRATION

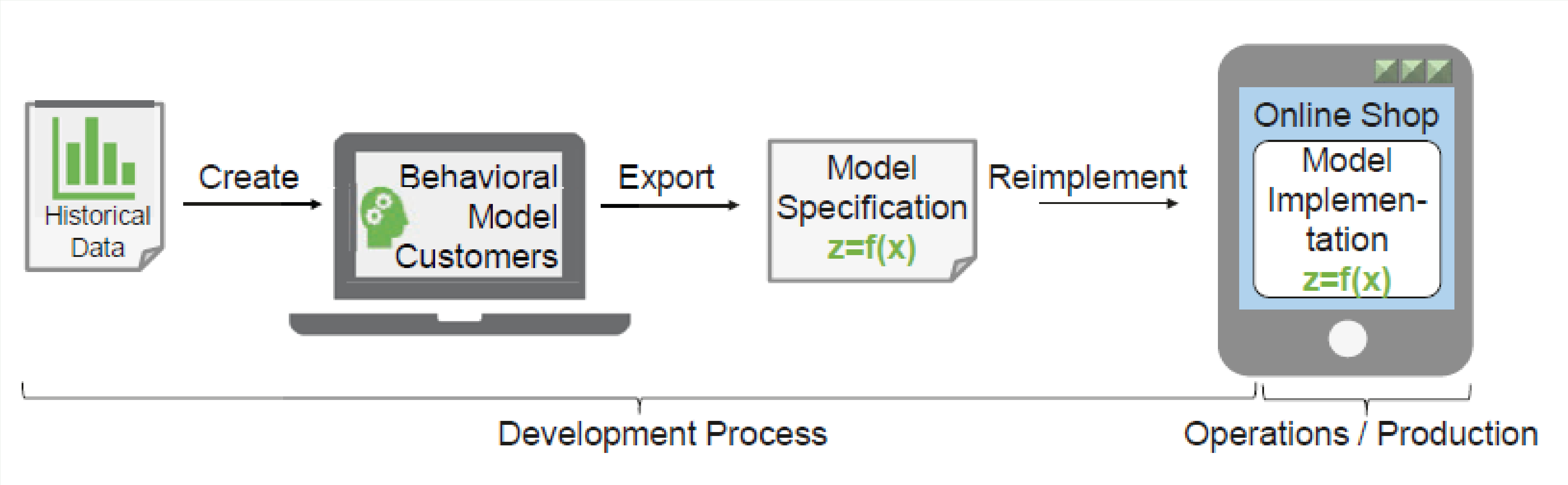
- Model deployment – moving/transitioning machine learning models into production, enabling their predictions to be accessible to users, developers, or systems. This allows for data-driven business decisions, application interactions (such as facial recognition), and more.
- 3 patterns of deployment and integration [1]
  - Precalculation
  - Reimplementation
  - Encapsulated AI component
- https://www.ibm.com/blog/ai-model-lifecycle-management-deploy-phase/

PRECALCULATION



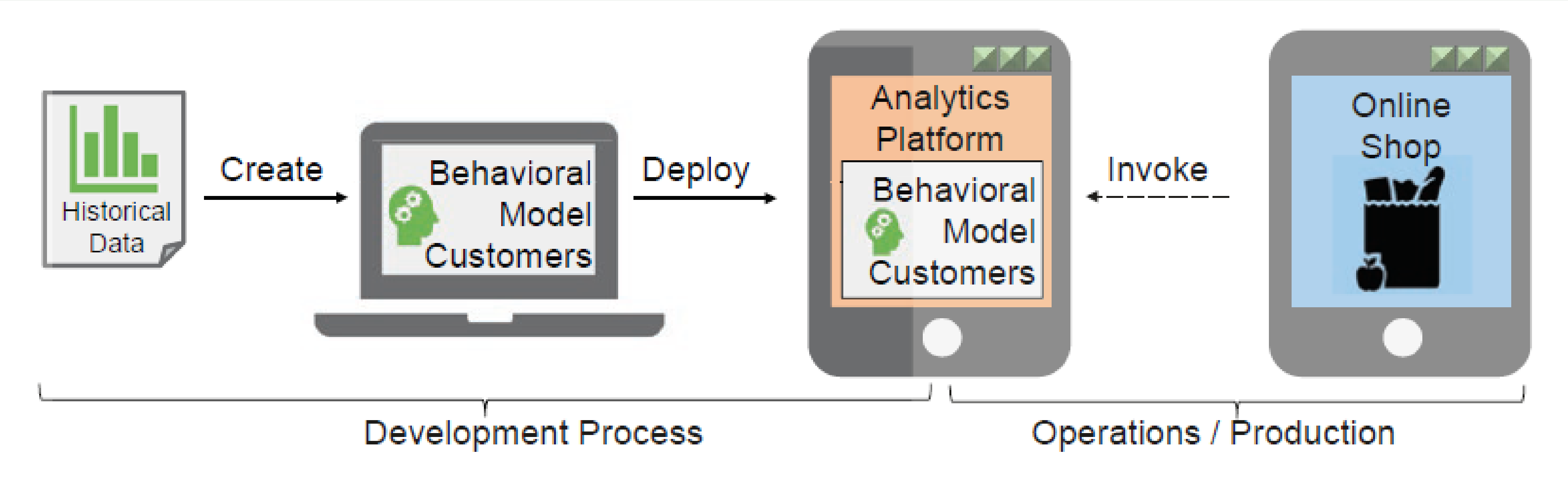
Klaus Haller. Managing AI in the Enterprise

REIMPLEMENTATION



Klaus Haller. Managing AI in the Enterprise

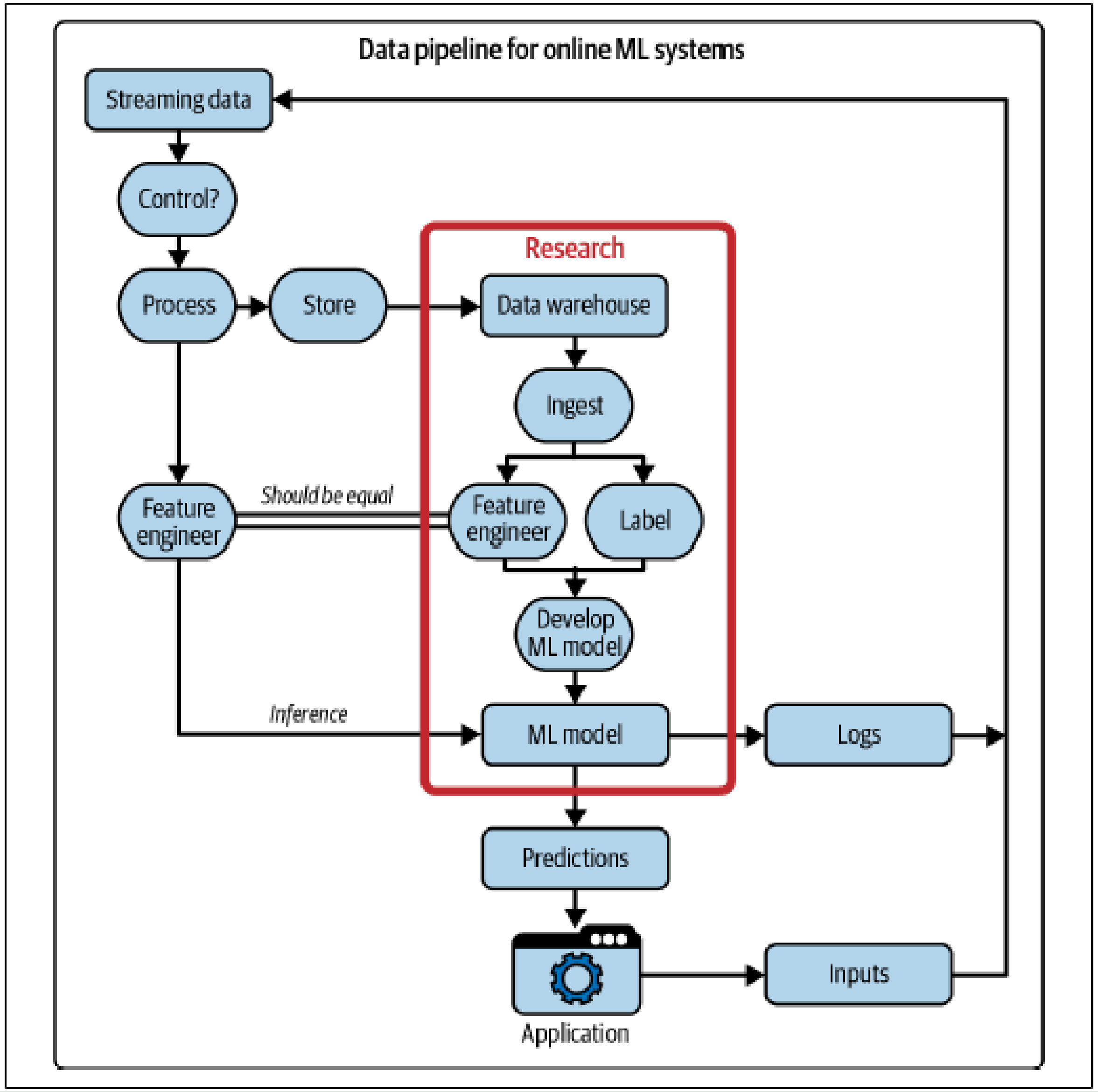
ENCAPSULATED ANALYTICS



Klaus Haller. Managing AI in the Enterprise

DEPLOYMENT STRATEGIES

- Answer a question - how does your model generates and gives its predictions to end users?
- Batch Prediction vs Online Prediction/Real-time prediction [2]
  - Batch prediction – only batch features
  - Online prediction with only batch features
  - Online prediction – batch + streaming features (Streaming prediction)
- https://medium.com/mlops-republic/mlops-batch-vs-online-ml-system-cbacee834837
- https://developers.google.com/machine-learning/crash-course/production-ml-systems/static-vs-dynamic-inference



Huyen, C. Designing Machine Learning Systems

DEPLOYMENT STRATEGIES

- On-Premises, Cloud, and Edge Deployment
  - On-premises – computations are done within the company / using own resources
  - On the cloud – a large amount of computations are done on the cloud (public clouds or private clouds)
  - Edge deployment - a large amount ofcomputations are done on consumer devices

- https://medium.com/@cprasenjit32/deployment-of-machine-learning-models-on-premises-and-in-the-cloud-39b021efba97
- https://www.trek10.com/blog/ml-on-premise-vs-ml-cloud
- https://www.vector8.com/en/articles/mlops-in-on-prem-environments

COMPILING AND OPTIMIZING MODELS FOR EDGE DEVICES

- Framework has to be supported by the hardware vendor
- Model optimization
    - Vectorization
    - Parallelization
    - Loop tiling
    - Operator fusion
    - Using ML to optimize ML models
- ML in Browsers
    - TensorFlow.js
    - Synaptic,
    - Brain.js
    - But JS is slow
    - WebAssembly (WASM).

INTEGRATION

- API-based Integration
- Database and Middleware
- Compatibility.Interfacing with Front-End Applications
- Legacy System Challenges.

DEPLOYMENT PIPELINES AND CI/CD FOR AI

- AI project always has an artifact called an ML pipeline
- The ML pipeline is a software artifact that addresses several considerations in your AI system [3]
- AI system [3]:
  - AI algorithms operate on the data (Storage questions)
  - Data quality
  - Different sources
  - Results must be presented to end user or ensure that the AI system follows some business and safety rules that are specific to domain

DEPLOYMENT PIPELINES AND CI/CD FOR AI (CHALLENGES)

- Real-world ML pipelines are complex
- No universal solution
- Codification of the technical AND business decisions
- The cost of maintenance of the ML pipeline

WHY USE CI/CD

- Automation of the pipeline
- Catching errors early
- Reproducibility
- Testing and monitoring
- Faster iteration
- Scalability

TOOLS AND TECHNOLOGIES

- Deployment tools:
  - AWS SageMaker
  - Azure ML
  - TensorFlow
  - Serving
- Monitoring tools:
  - Prometheus (Open source)
  - Grafana
- Version Control:
  - DVC
  - MLflow