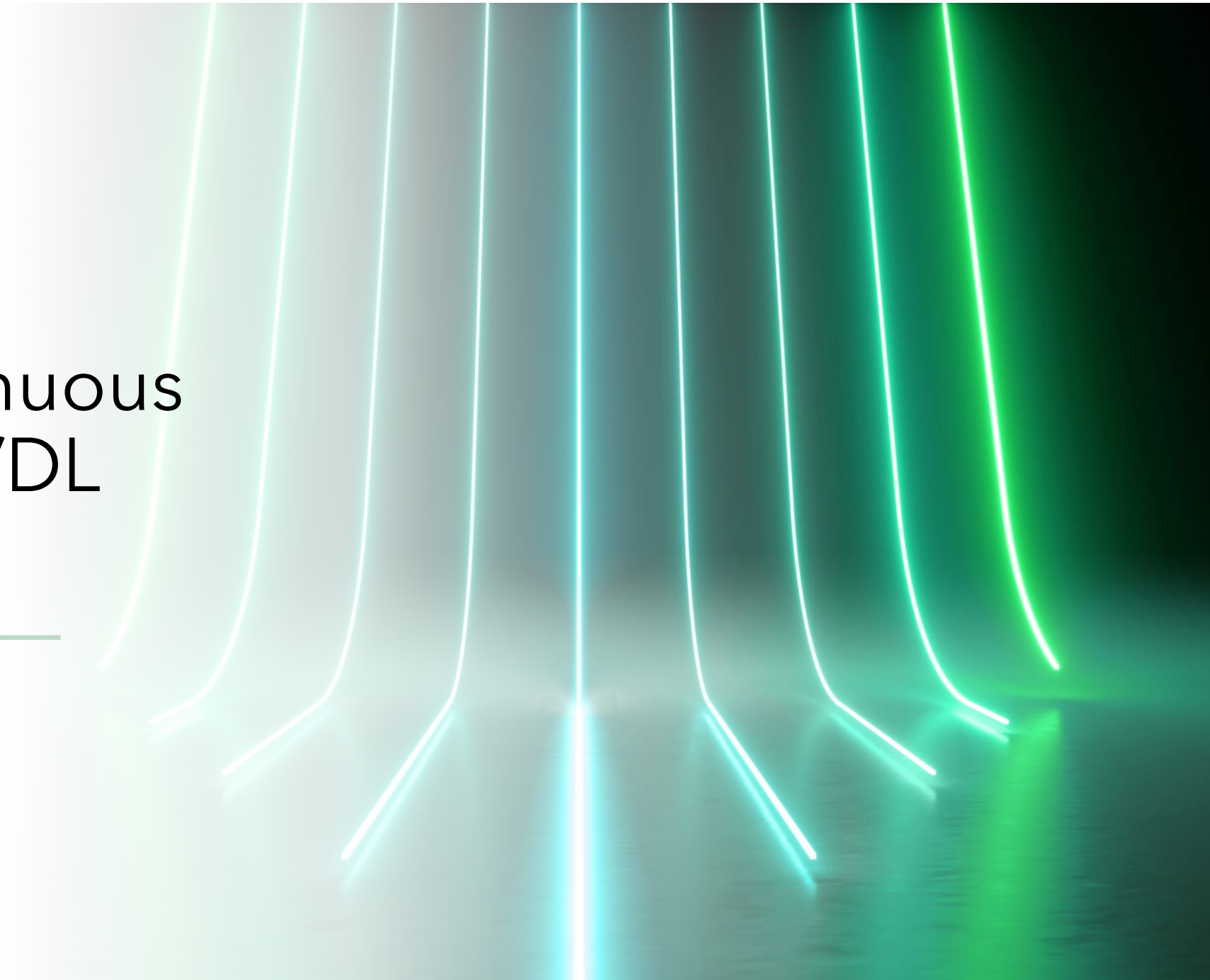
Towards Continuous Delivery of ML/DL Systems

Meenu Mary John



Background

- Artificial Intelligence (AI) is becoming increasingly popular in companies across domains due to the success of ML and DL technologies
- However, the end-to-end process of developing, deploying and successfully evolving industry-strength, production quality ML/DL models in systems introduces several challenges
- While software development is well supported in terms of methods, processes and techniques, this is not the case in relation to ML/DL model development where companies experience lack of systematic and structured method support

Research vision

- The overall vision is to help embedded systems companies address the complexity of developing, deploying and evolving ML/DL models as part of their systems
- To achieve this, we develop **a systematic and structured design method for the process of developing, deploying and successfully evolving ML/DL models**

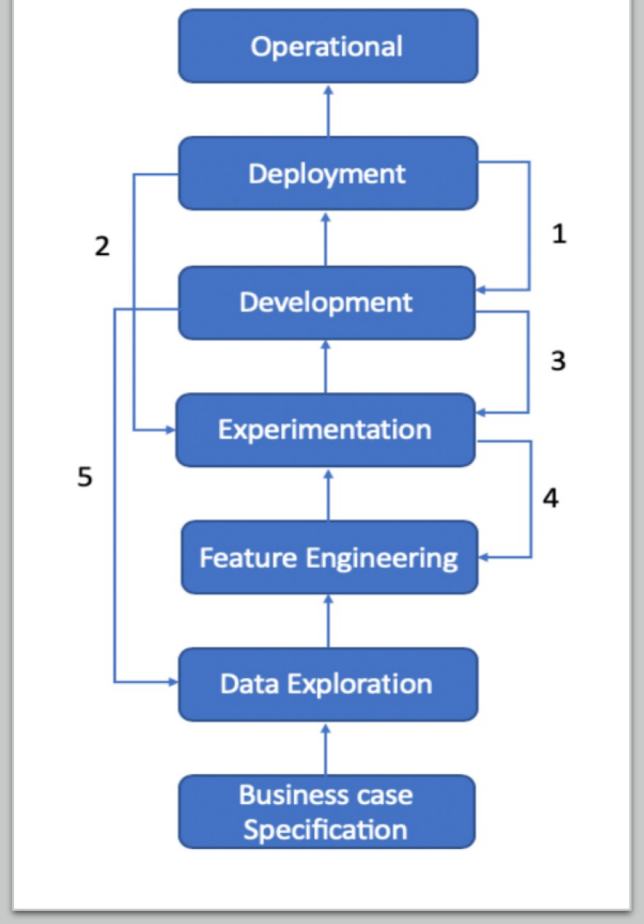
Scope and focus

- We study:
  - The **end-to-end MLOps process** involving methods and techniques for developing, deploying and evolving ML/DL models in production and as part of large complex embedded systems
  - How to **operationalize ML/DL models**, e.g. architectures for deployment and continuous optimization of the model in response to new requirements, new insights, changing user/system behaviors etc.
  - How to **continuously deliver and evolve Machine Learning Systems (MLOps)** to improve the key quality attributes of the larger system of which these are part?

Results (1/4)

Development phases for ML/DL models

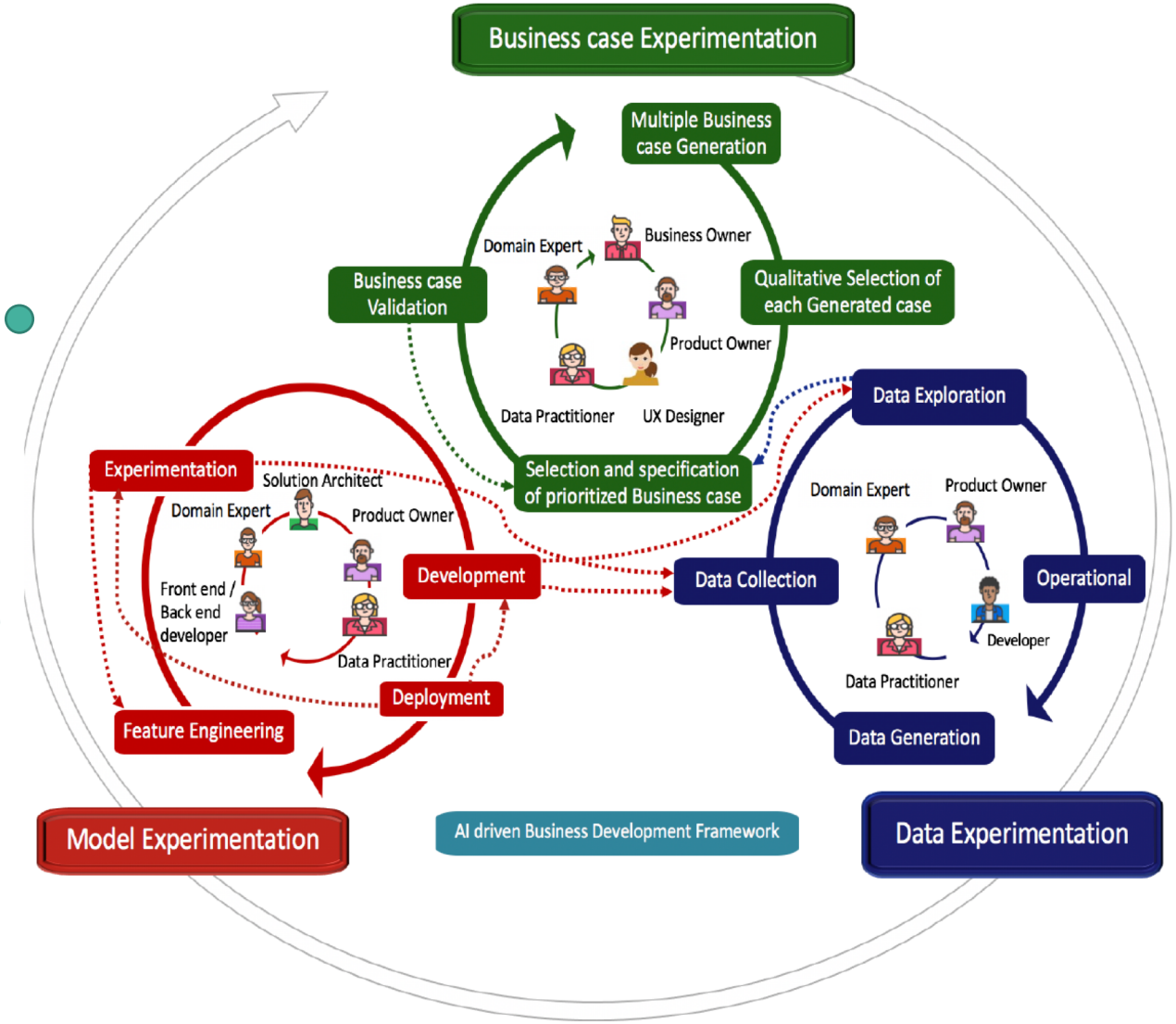
- Typical activities performed by developers
- Sequence of, and iterations between, activities
- Events that trigger iterations
- Challenges associated with each phase



Case study involving data scientists at Ericsson, CEVT, Grundfos, AB Volvo, Tetra Pak, BOSCH (and Peltarion)

AI-driven business development

- Includes phases
- Involves roles
- Requires collaboration and iteration
- Takes place in the context of a larger system



Development – Training – Deployment – Integration - Evolution

Continuous delivery of ML systems (MLOps)

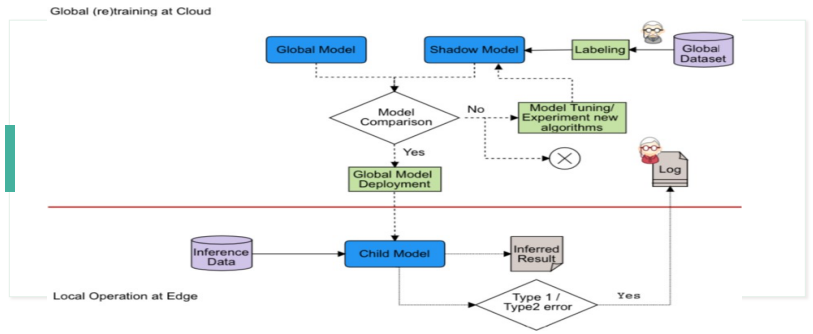
Results (2/4)

Architectural alternatives for training and re-training of ML/DL models

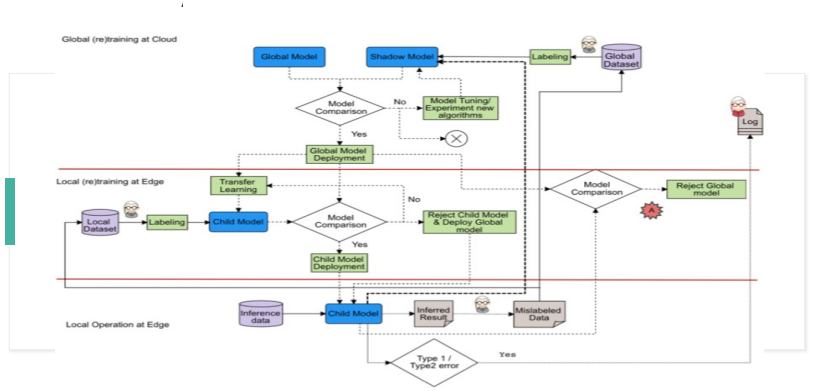
- Development of a generic framework consisting of five architectural alternatives
- The framework ranges from a centralized architecture where cloud (re)training is given more priority, to a decentralized architecture where the edge (re)training is instead given priority
- Validation of the framework in four companies
- Identification of the key challenges that experts face in selecting the optimal architectural alternative

Based on a six-months internship at Tetra pak as well as validation interviews at CEVT, BOSCH, Ericsson and Grundfos and follow-up study also includes AB Volvo and Siemens
On-going study -- Scania

Centralized

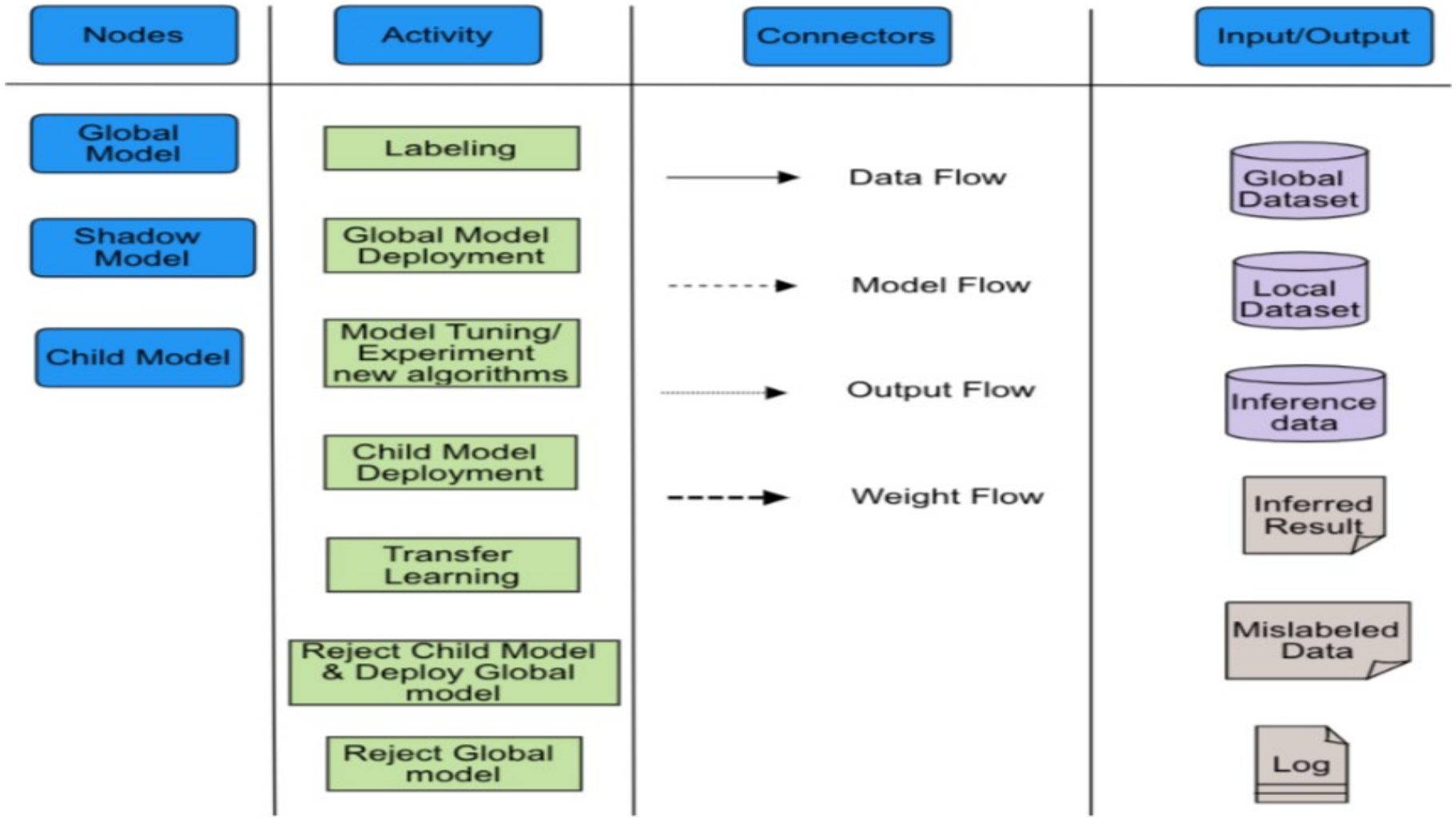


...

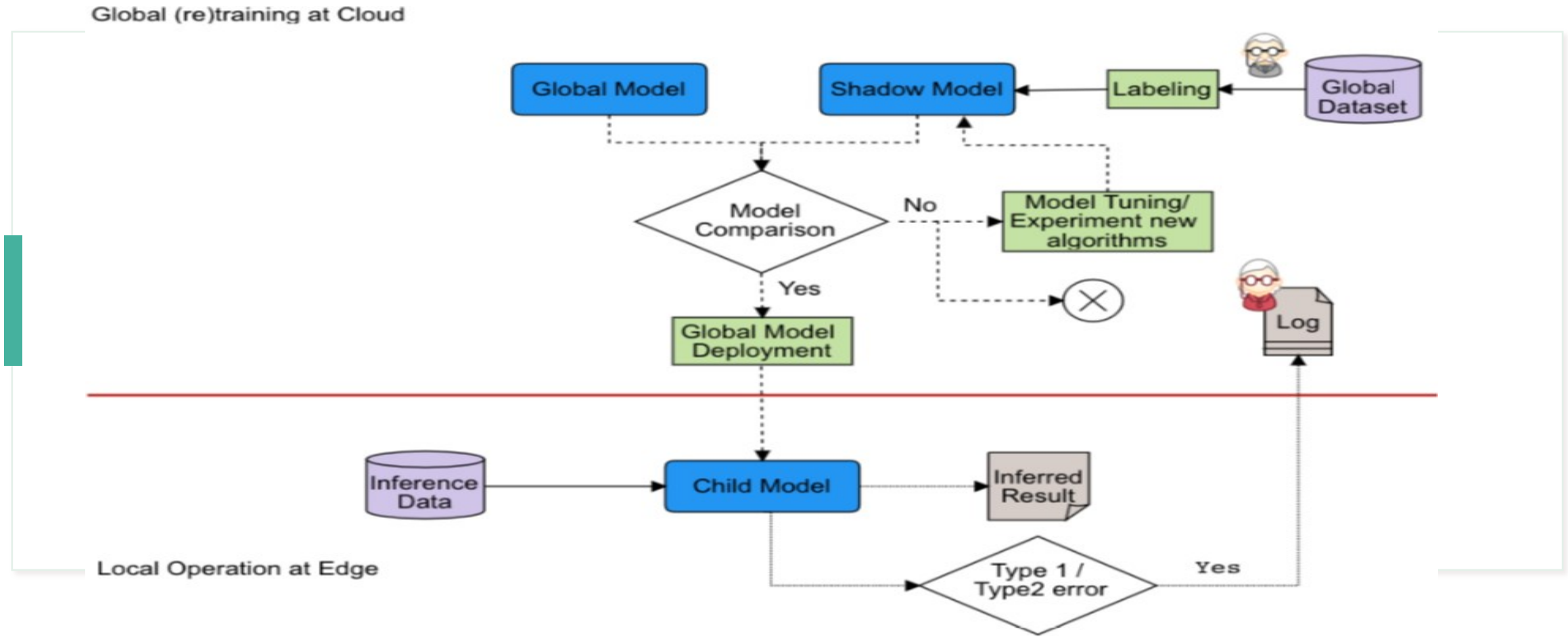


Decentralized

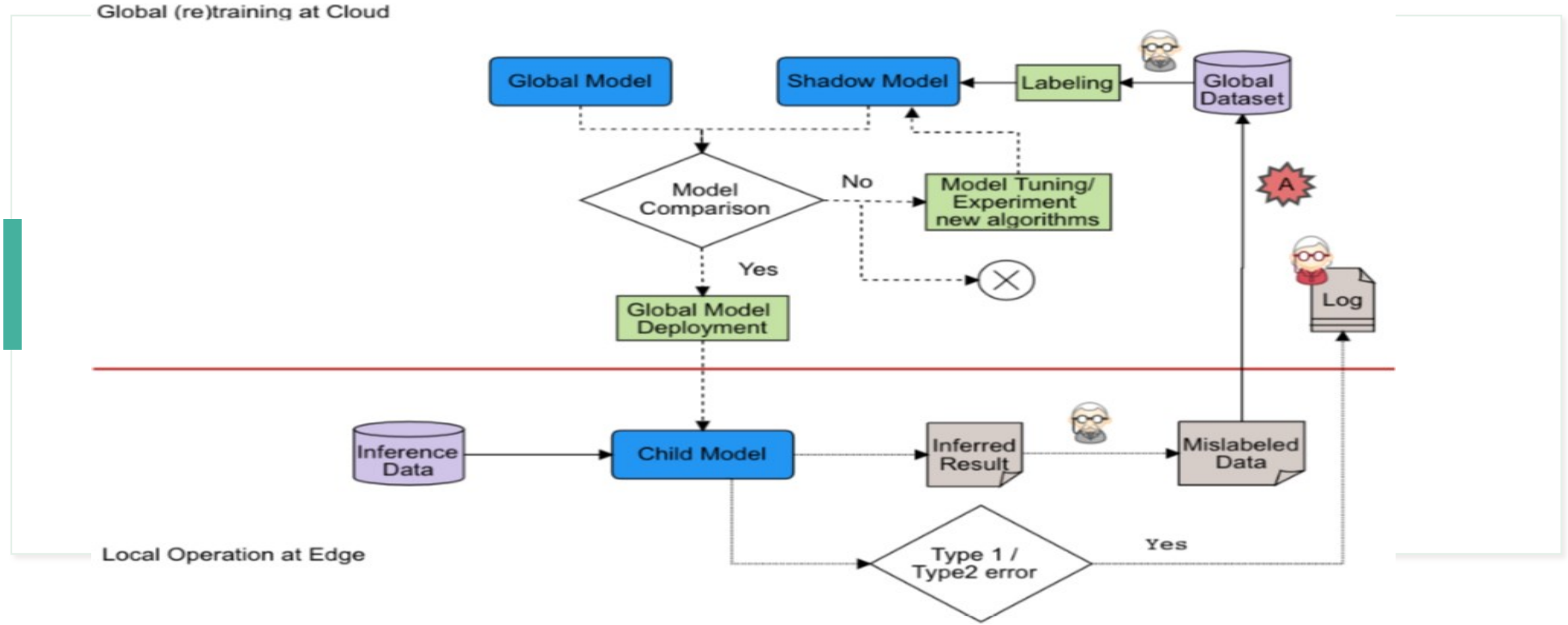
Meta Model (Set of Concepts used)



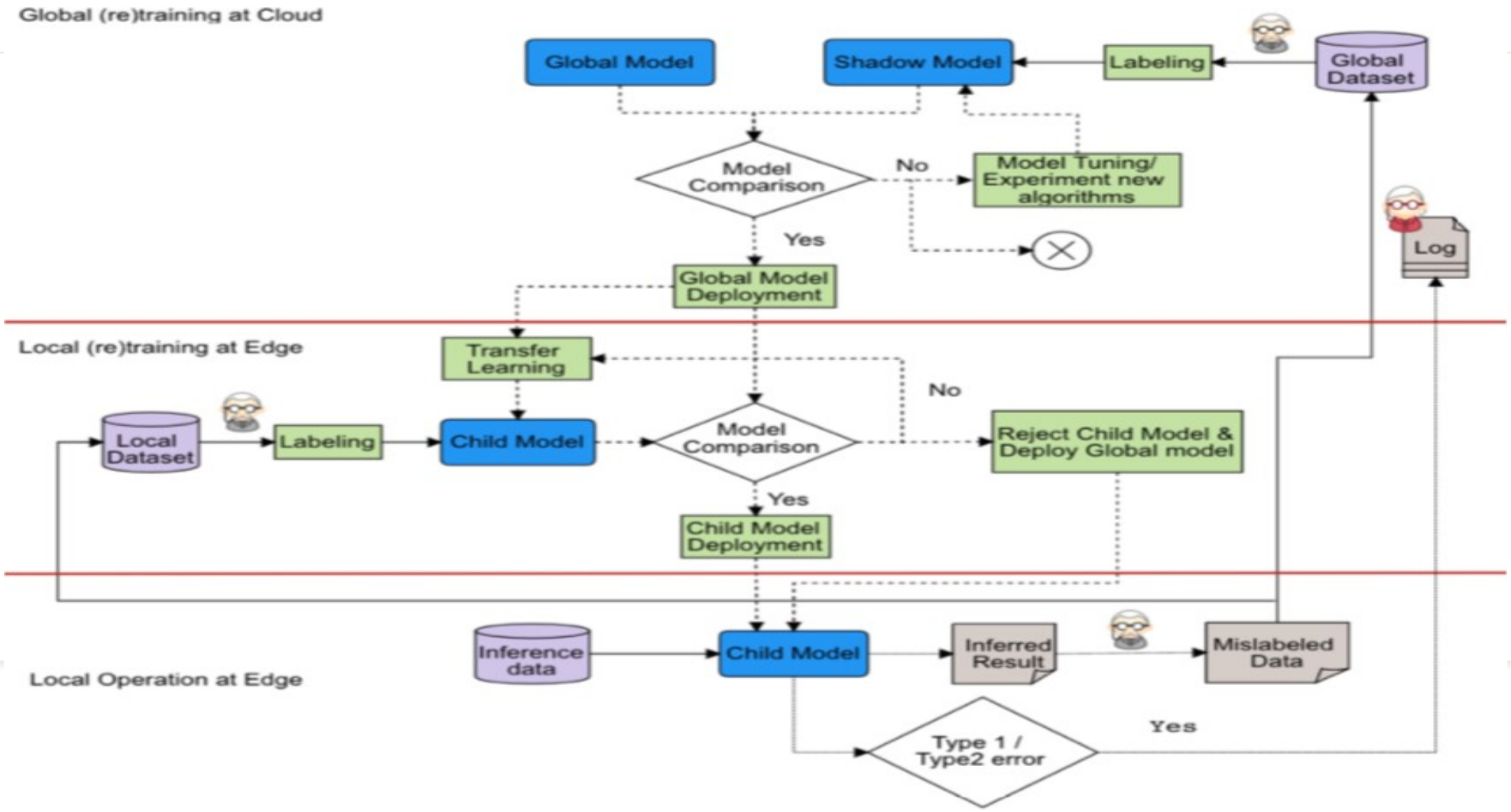
Alternative 1 Centralized approach



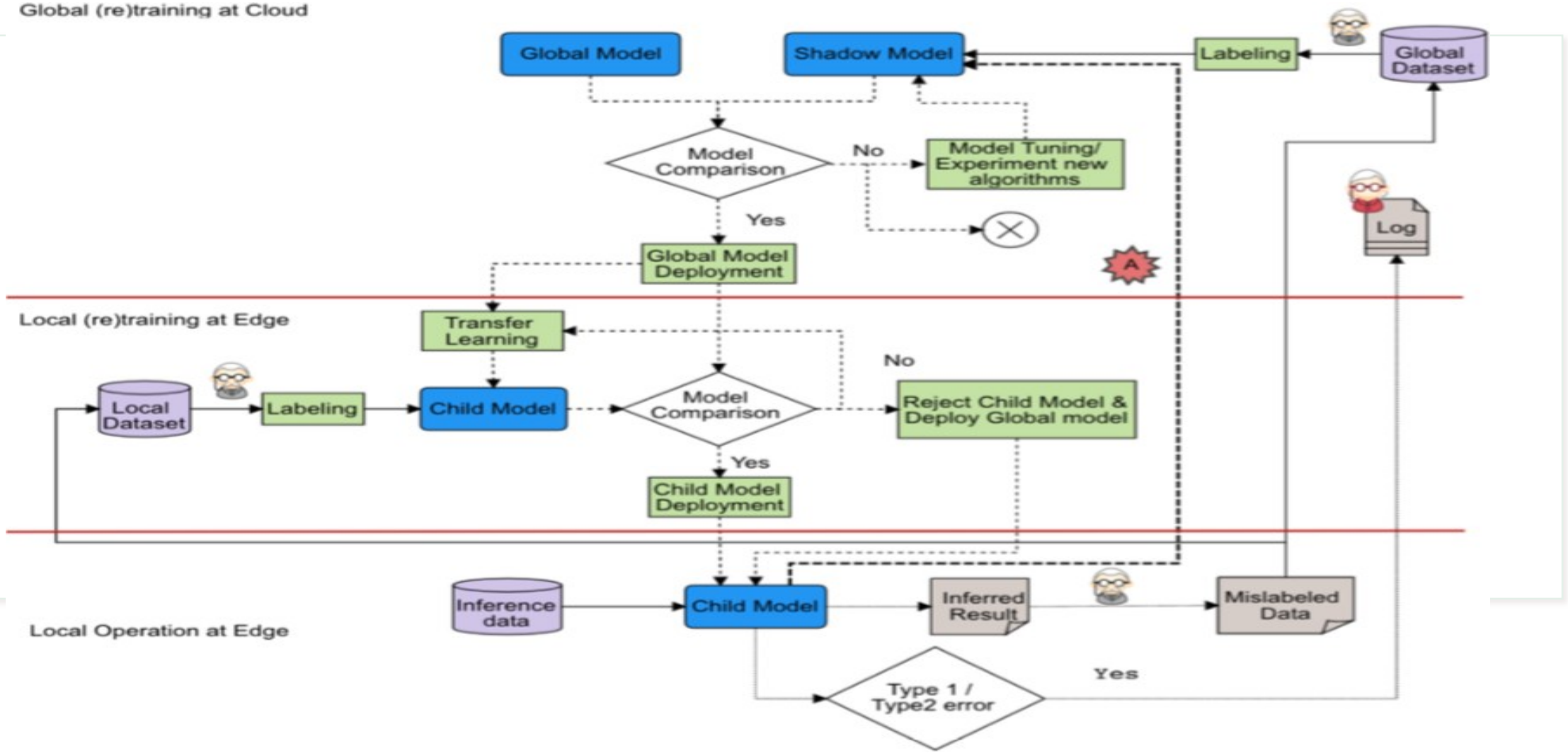
Alternative 2 More Centralized, Less Decentralised approach



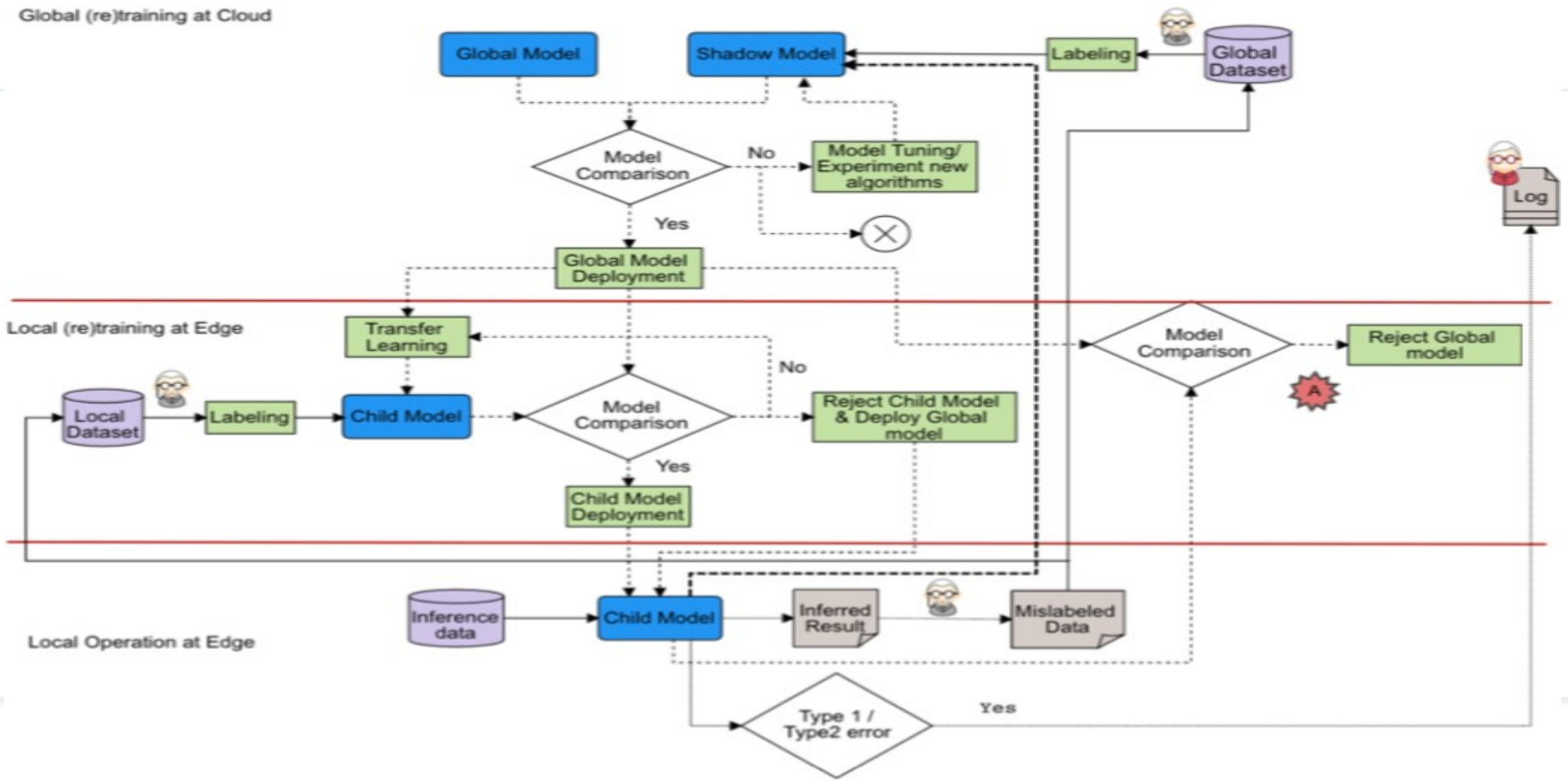
Alternative 3 - Mix of Centralized and Decentralised approach



Alternative 4 - Less Centralized, More Decentralised approach

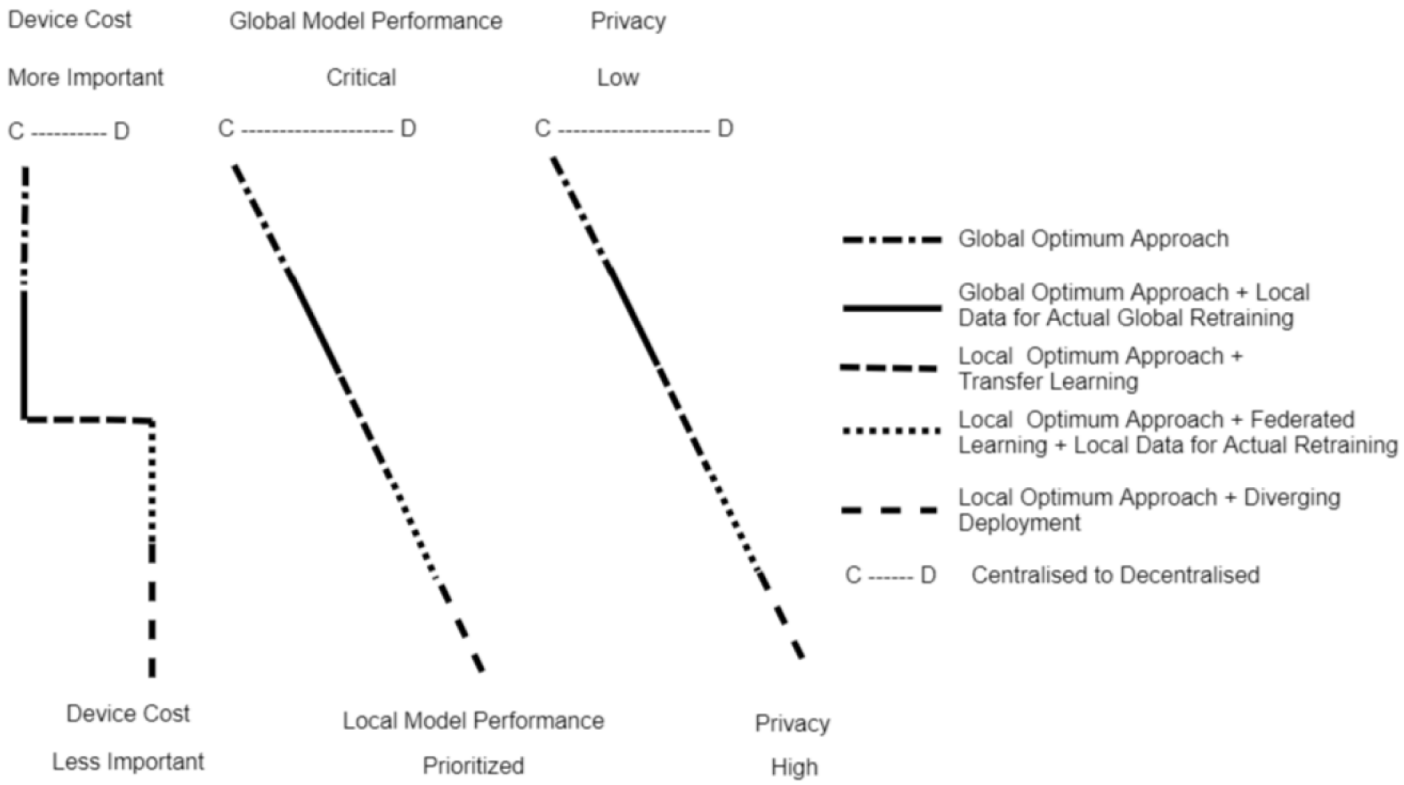


Alternative 5 - Decentralised approach



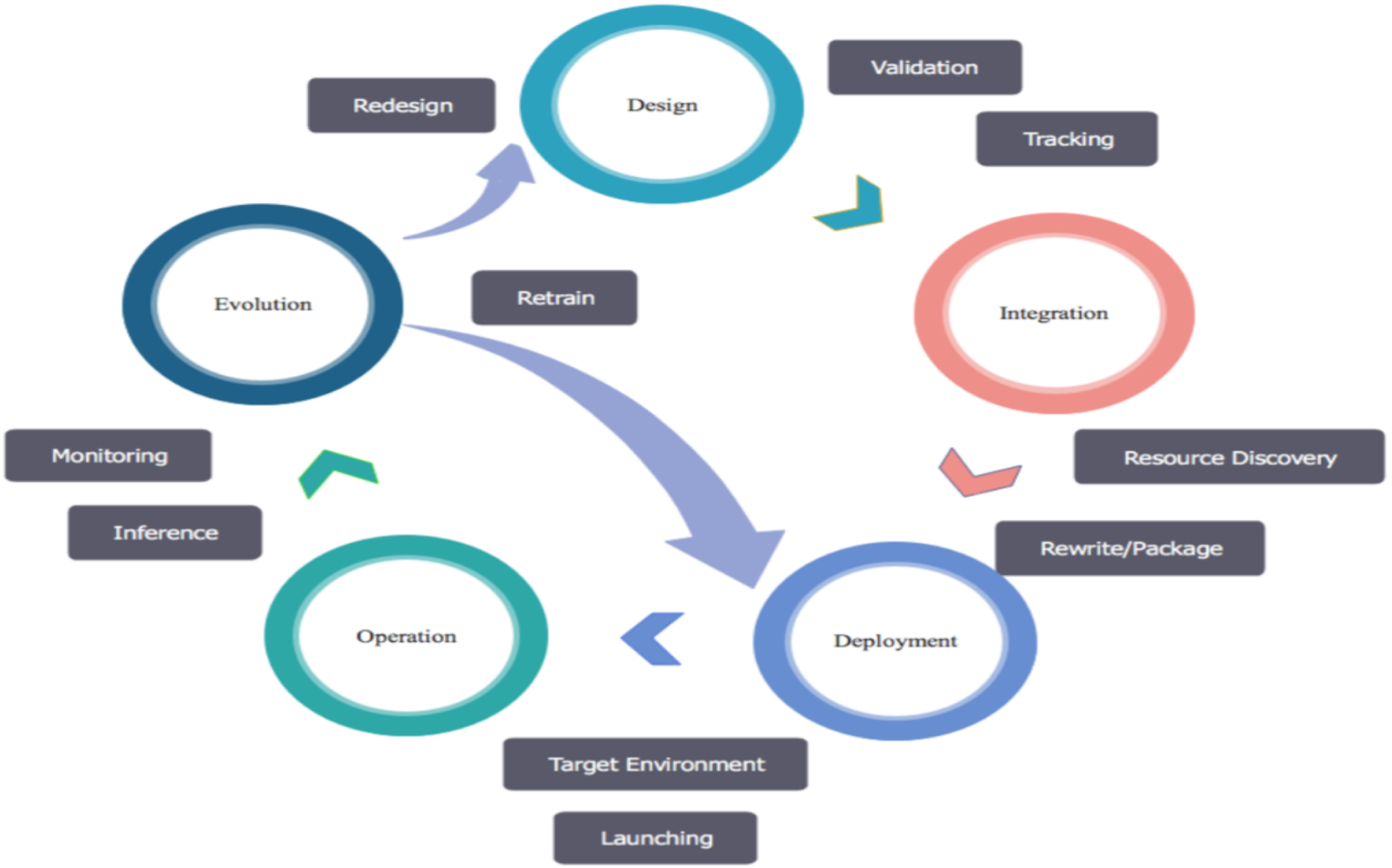
Results (3/4)

Architecture Selection Framework



John, M.M., Olsson, H.H. and Bosch, J., (2020, December). AI Deployment Architecture: Multi-case Study for Key Factor Identification. In Proceedings of the 27th Asia-Pacific Software Engineering Conference (APSEC).

Results (4/4) **End-to-end Deployment Process**



John, M.M., Olsson, H.H. and Bosch, J., (2020, November). Architecting AI Deployment: A Systematic Review of State-of-the-art and State-of-practice Literature. In 11th International Conference on Software Business (ICSOB).

Ongoing Research

- Integration of ML/DL models in software-intensive embedded systems
- Focusing on the impact on key quality attributes of the overall system
  - How to deploy and integrate ML components without negatively affecting key quality attributes (Scalability, Reliability and Maintainability) of the overall system?
  - How to, over time, establish and effectively evolve MLOps to improve the key quality attributes of the larger system of which these are part?

On-going study - Ericsson

Background and motivation

- A small fraction of real-world ML systems is made up of ML code and the surrounding infrastructure is complex
- Although there is extensive research on system quality attributes, there are few (if any) attempts to understand how these are affected when integrating ML/DL components into the system

Looking forward to... && Expected Deliverables...

Looking forward to...

- Explore ways of working in different companies to learn development and deployment of ML/DL models in large-scale embedded systems
- Understand how practitioners design the overall system by considering quality attributes and how they relate and link these attributes to the overall organization

Deliverables...

- **Case examples and best** practices for systematic development of ML/DL models in large-scale embedded systems
- **Design considerations** on multifaceted integration of ML/DL components in embedded systems without adversely impacting its quality attributes
- **A framework which outlines relationships, prioritization and trade-off** among selected quality attributes

Thank you!

- meenu-mary.john@mau.se
- helena.holmstrom.olsson@mau.se
- jan.bosch@chalmers.se