Deployment

02457 Machine Learning Operations

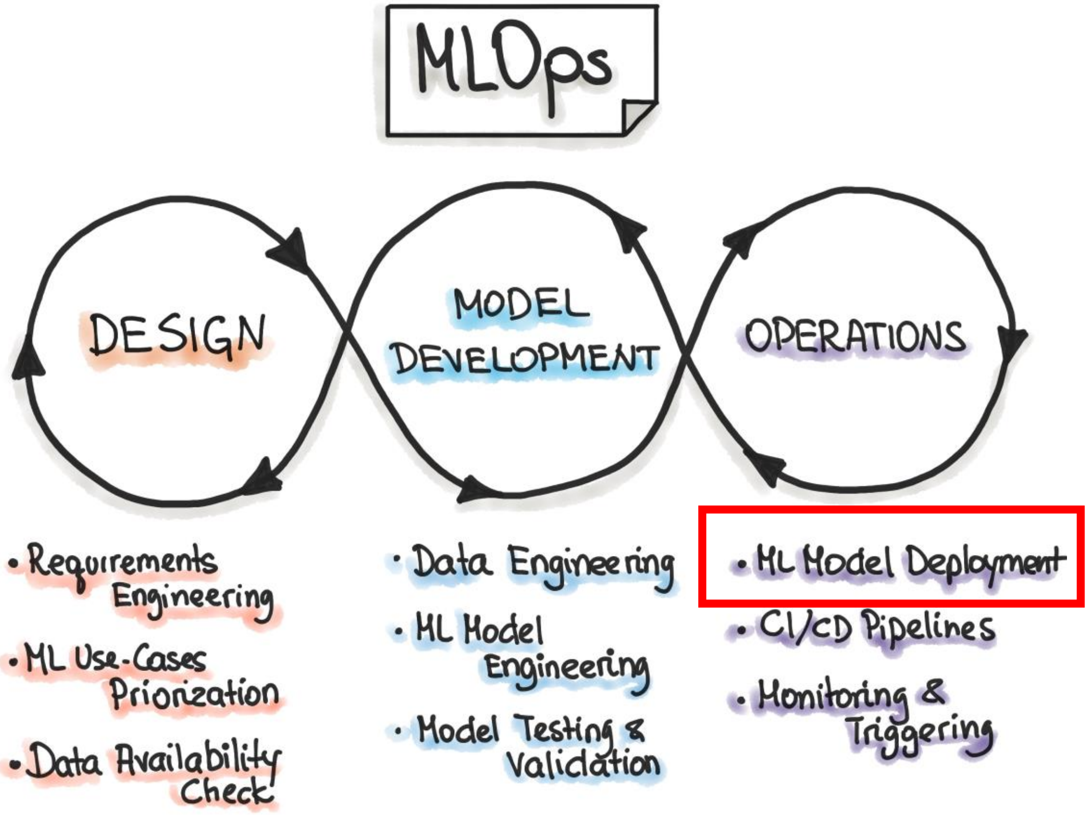
Nicki Skafte Detlefsen,

Postdoc

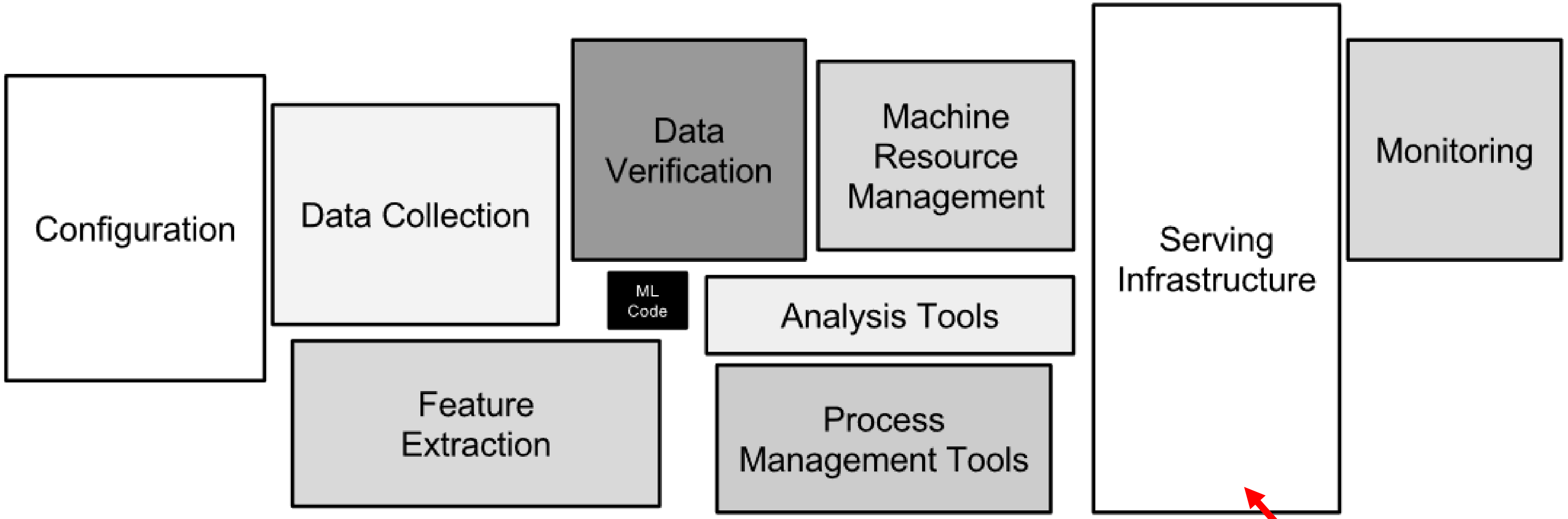
DTU Compute

Freeing the model

- Model deployment is part of the operations in MLOps
- In a nutshell: make the model available to others


MLOps
DESIGN
MODEL DEVELOPMENT
OPERATIONS
• Requirements Engineering
• ML Use-Cases Priorization
• Data Availability Check
• Data Engineering
• ML Model Engineering
• Model Testing & Validation
• ML Model Deployment
• CI/CD Pipelines
• Monitoring & Triggering

Remember this?



This is big for a reason

Many levels of deployment (within machine learning)

1. Github reposatory + link to model weights
   - Easy to ”deploy”
   - Pain in the *** to use
2. Deploy on local computer/cluster
   - Fairly easy getting up and running, just requires people can access from outside
   - Can be fairly easy to use
   - Does not scale at all
3. Deploy to cloud service
   - Can be a pain to setup
   - Easy to use and scales to ∞ (and beyond!)