Kubernetes and MLOps for Scalable and Reproducible Generative AI Workflows

What will this session be about?

Who am I?

@AnnieTalvasto
CMO at VSHN

- CNCF Ambassador
- Azure MVP

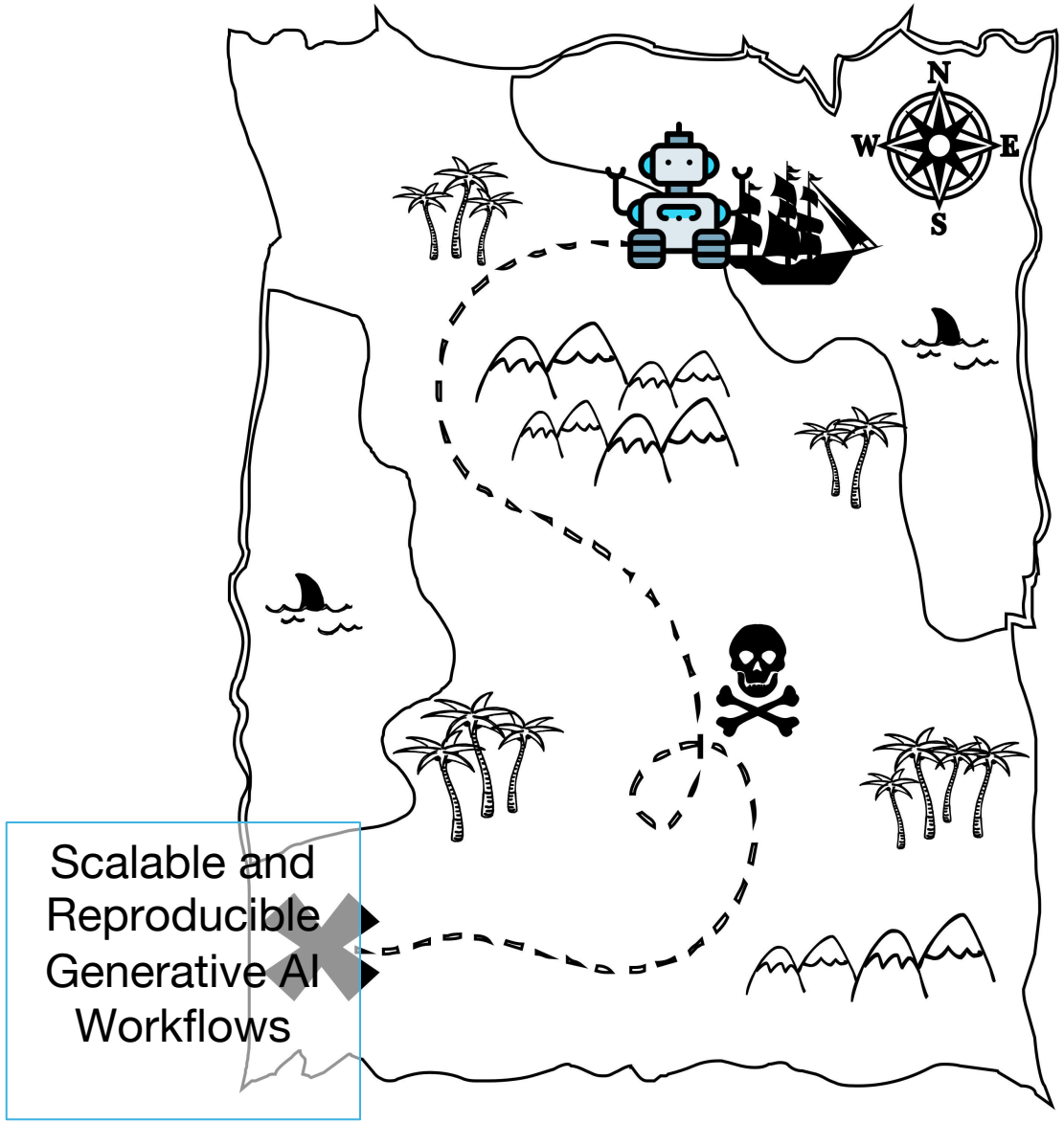
- Kubernetes & CNCF meetup co-organizer
- Startup-coach
- Co-host of Cloudgossip podcast - cloudgossip.net

Agenda

- Introduction
- Definitions
- Why is AI/ML different
- Kubernetes & MLOps
- Kubeflow
- Best practices
- Considerations
- Wrap up & Resources

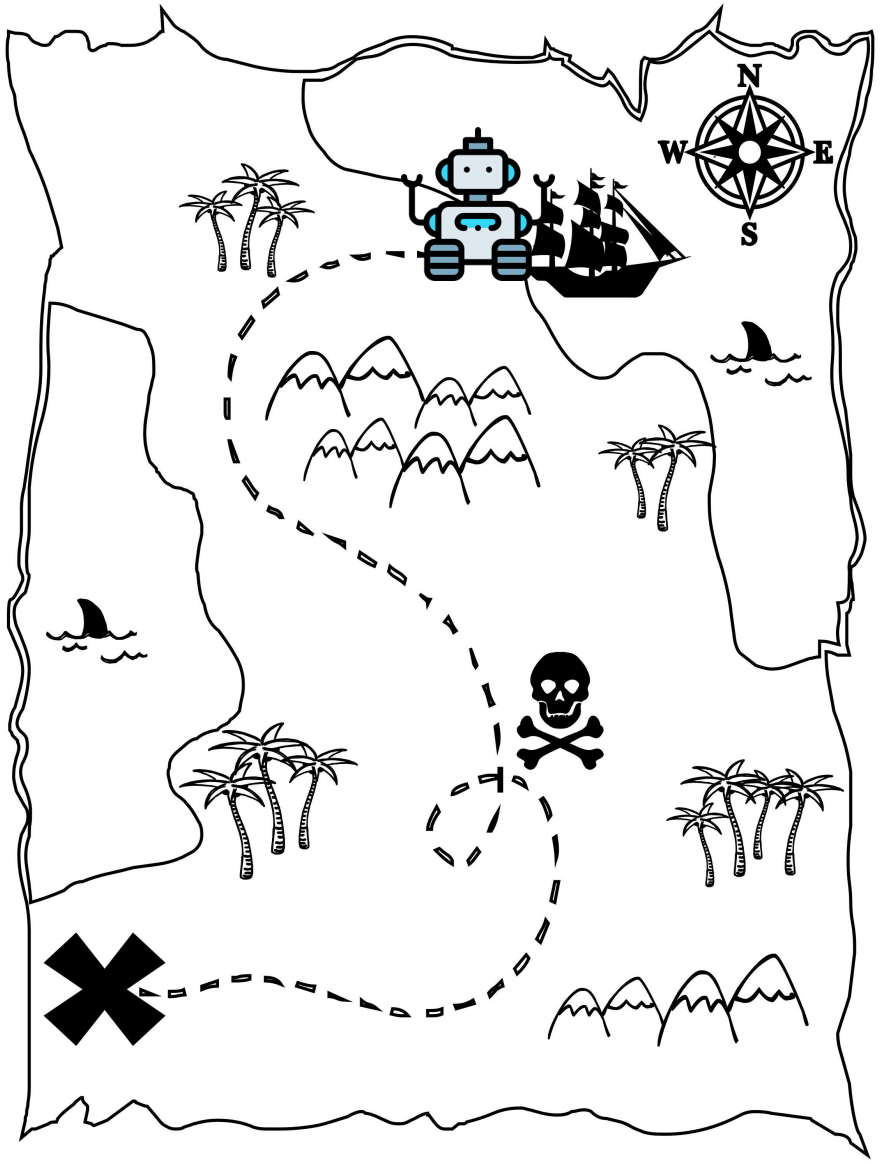
Agenda

- Introduction
- Definitions
- Why is AI/ML different
- Kubernetes & MLOps
- Kubeflow
- Best practices
- Considerations
- Wrap up & Resources



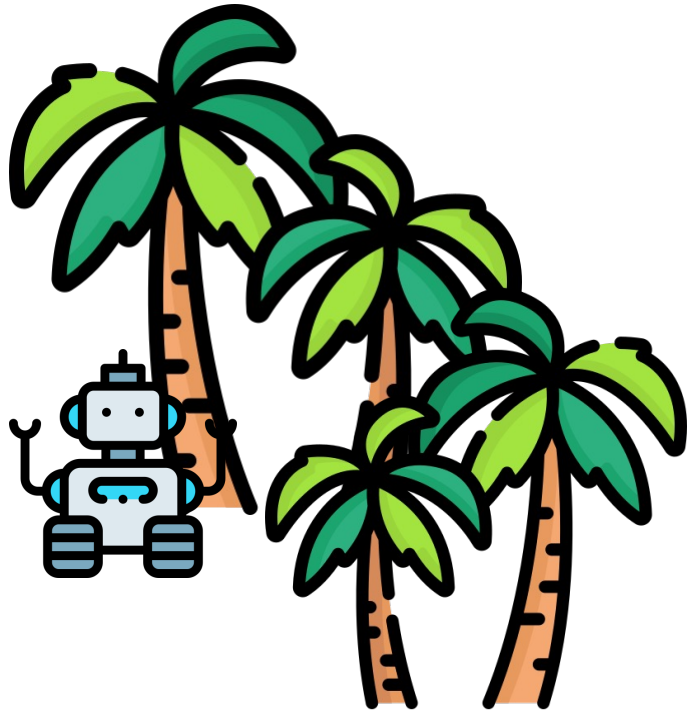
Agenda

- ~~Introduction~~
- Definitions
- Why is AI/ML different
- Kubernetes & MLOps
- Kubeflow
- Best practices
- Considerations
- Wrap up & Resources



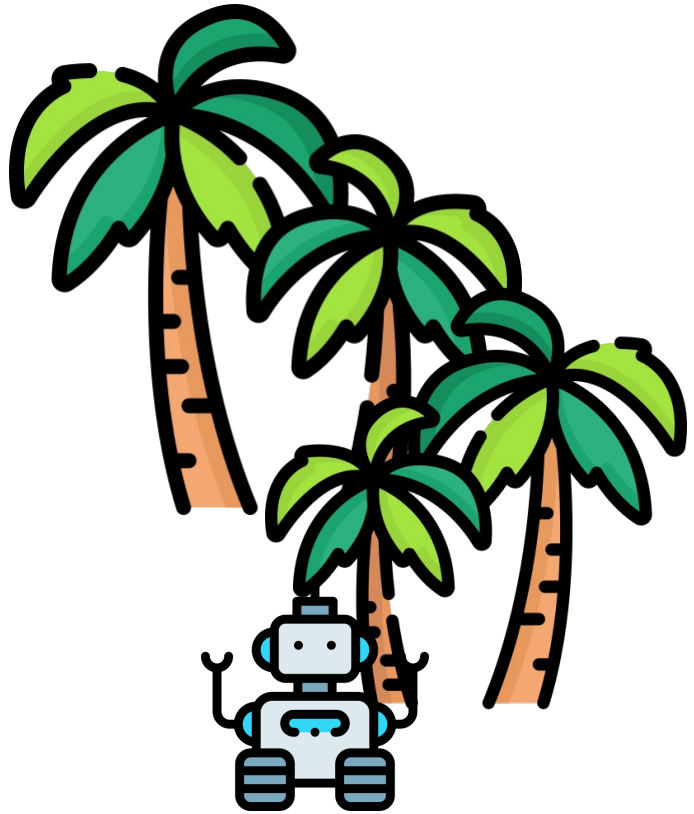
AI

- Artificial Intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think and learn like humans.



ML

- Machine Learning (ML) is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computer systems to improve their performance on a specific task through learning from data, without being explicitly programmed.

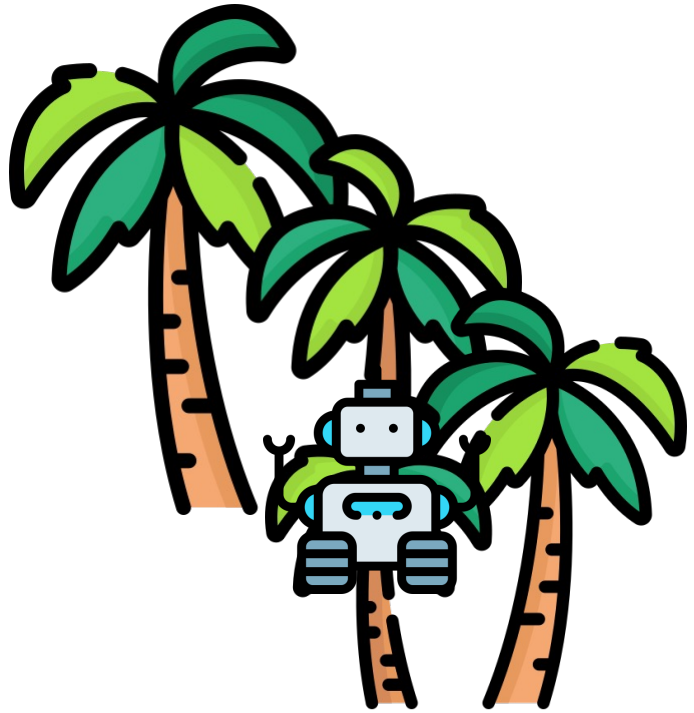


Data science

- Data science is a multidisciplinary field that uses various techniques, algorithms, processes, and systems to extract insights and knowledge from data.

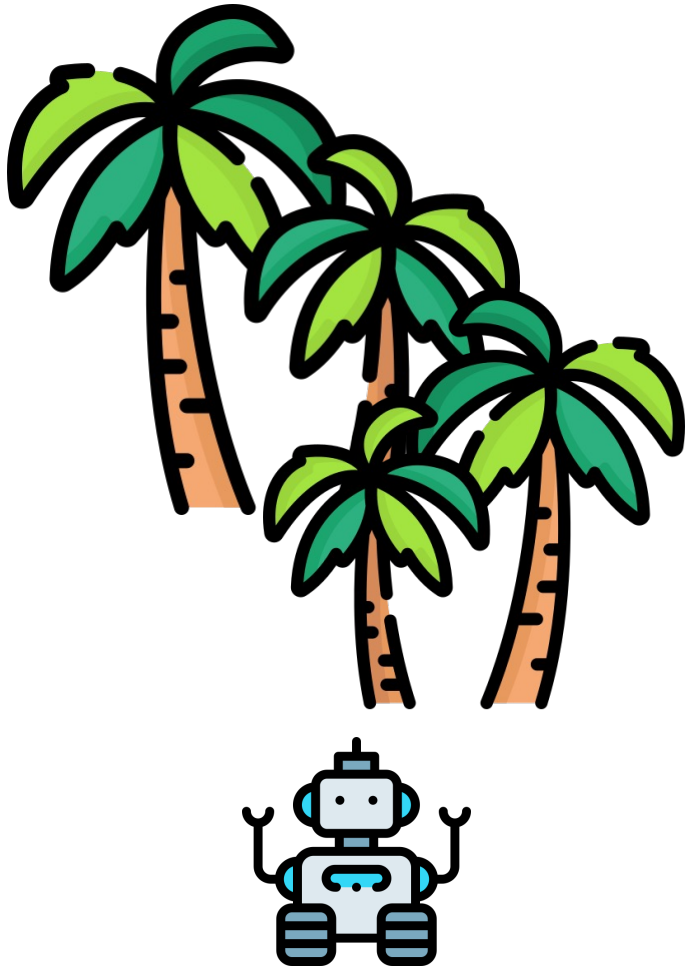
DevOps

- DevOps, short for "Development" and "Operations," is a set of practices, principles, and cultural philosophies that aim to improve and streamline collaboration between software development (Dev) and IT operations (Ops) teams.



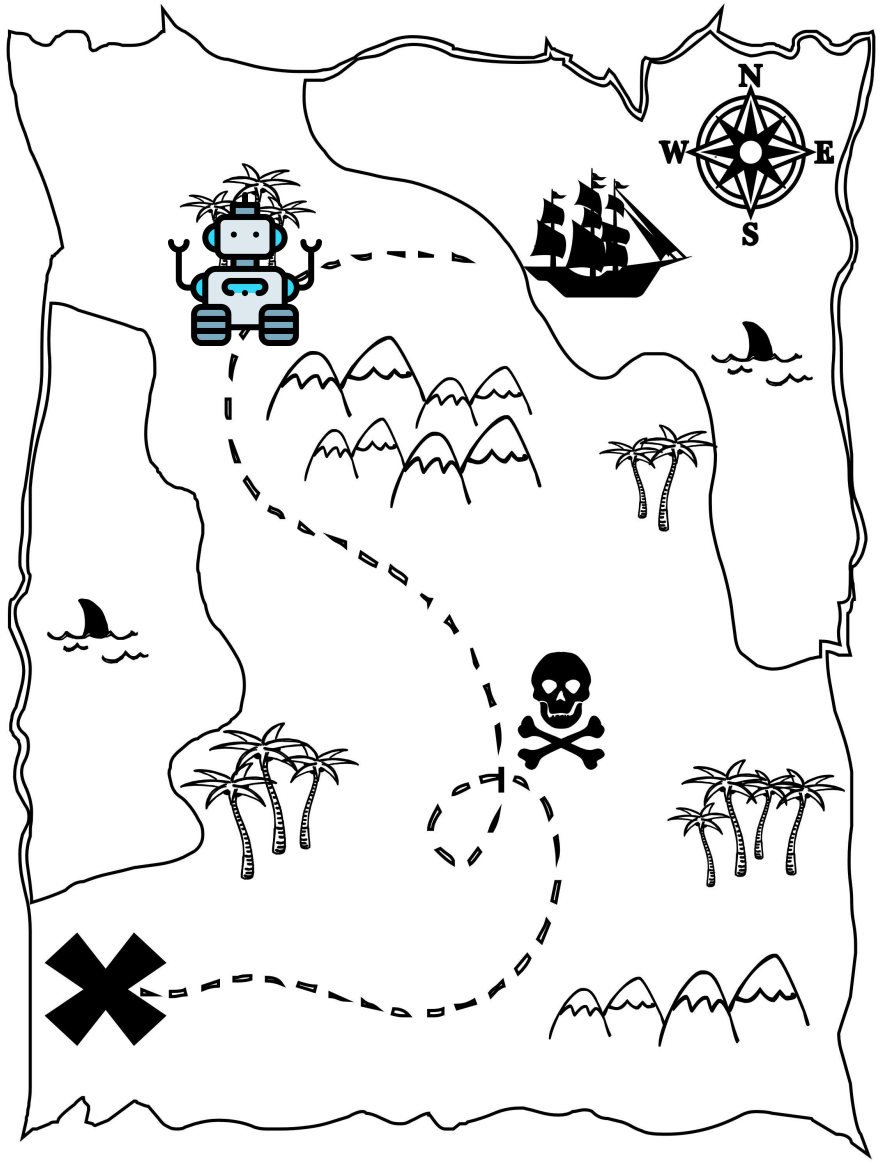
MLOps

- MLOps, short for "Machine Learning Operations," is a set of practices, principles, and tools that combine machine learning (ML) with the practices of DevOps.



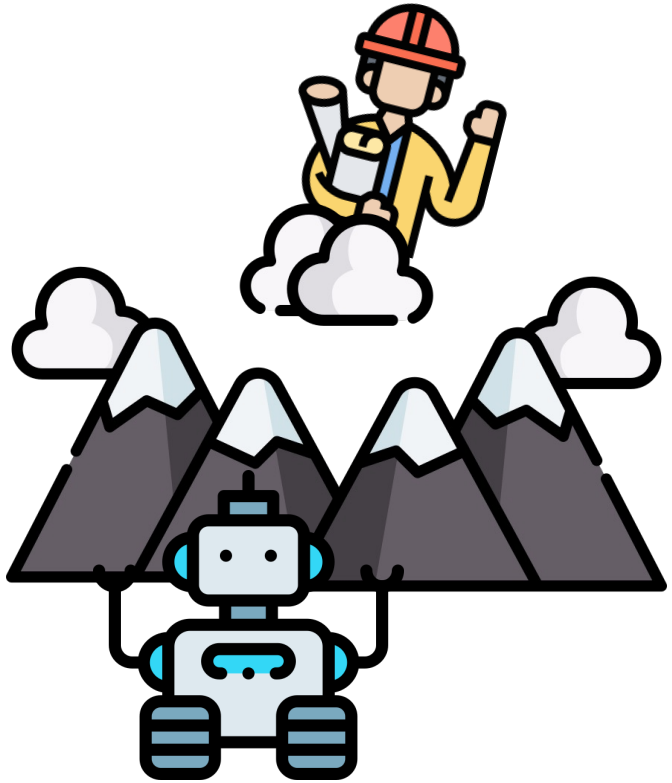
Agenda

- ~~Introduction~~
- ~~Definitions~~
- Why is AI/ML different
- Kubernetes & MLOps
- Kubeflow
- Best practices
- Considerations
- Wrap up & Resources

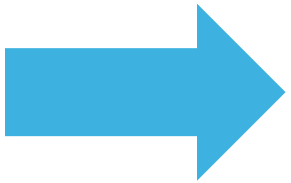


Differences with AI/ML

- Data-Centric Nature
- Model Complexity
- Iterative Development
- Scalability
- Monitoring and adaptation
- Testing and validation
- Explainabity and bias
- Rapid advancements
- Collaboration
- Deployment challenges
- Cost

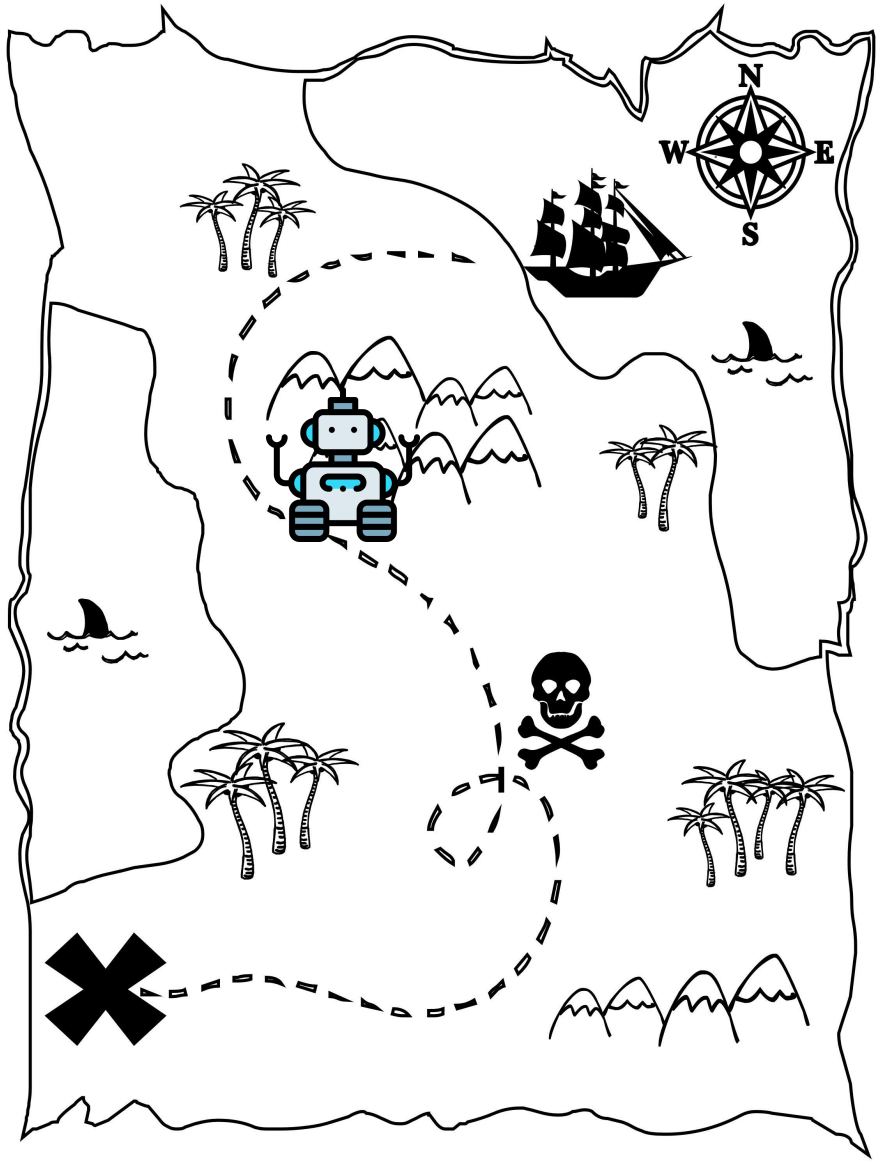


From Research to Production



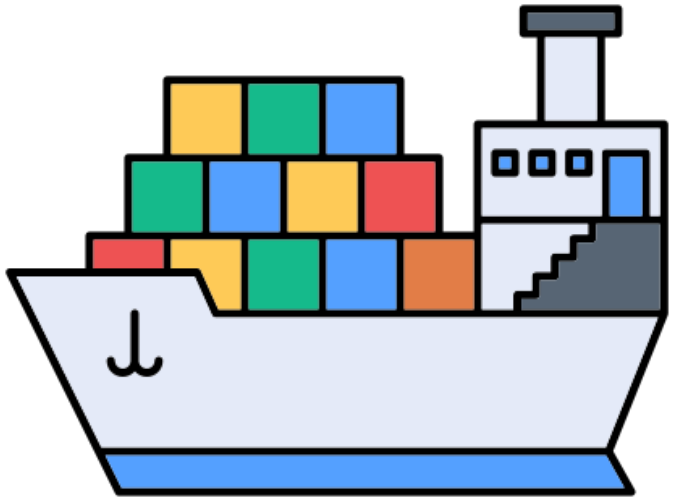
Agenda

- ~~Introduction~~
- ~~Definitions~~
- ~~Why is AI/ML different~~
- Kubernetes & MLOps
- Kubeflow
- Best practices
- Considerations
- Wrap up & Resources



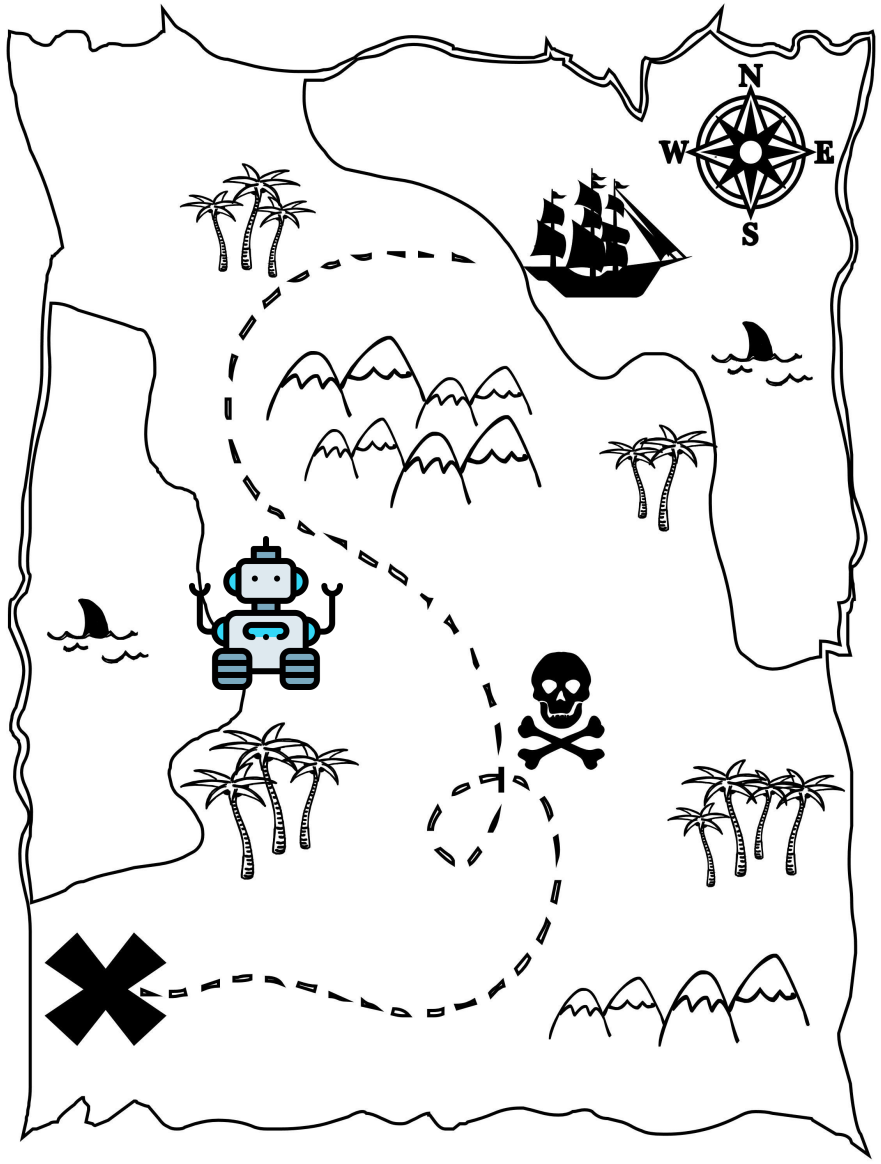
Kubernetes & MLOps

- Portability
- Customizability
- Performance
- Consistency
- Microservices
- Composability



Agenda

- ~~Introduction~~
- ~~Definitions~~
- ~~Why is AI/ML different~~
- ~~Kubernetes & MLOps~~
- Kubeflow
- Best practices
- Considerations
- Wrap up & Resources



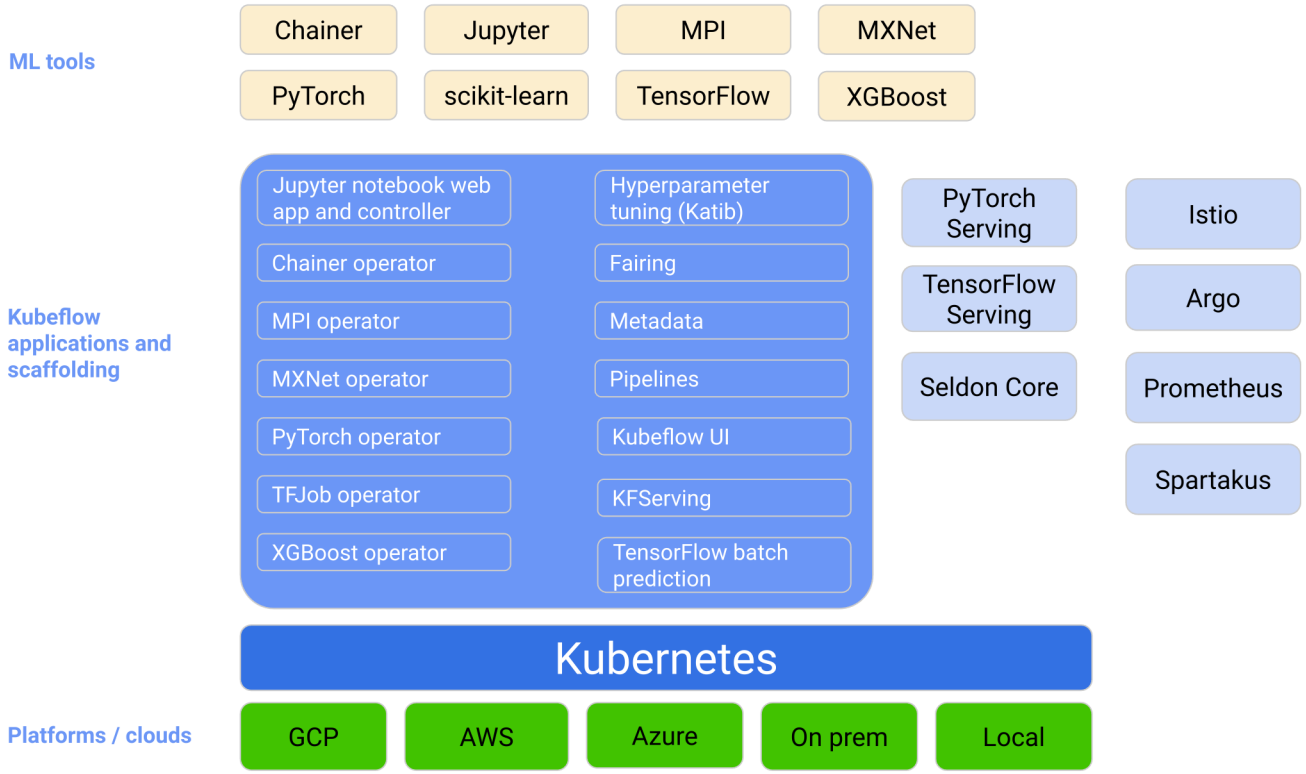
Kubeflow

The machine learning toolkit for Kubernetes.

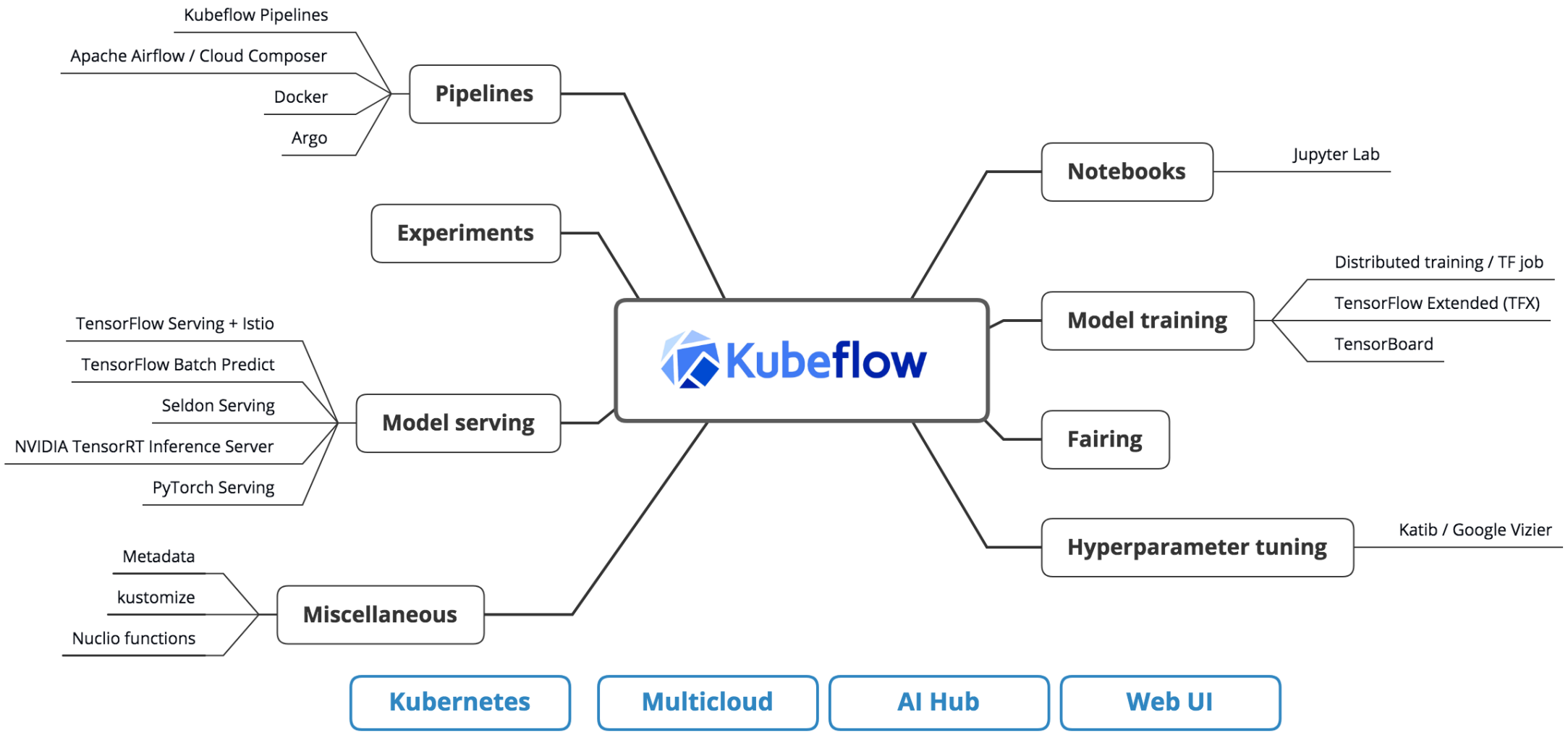
Kubeflow

What is Kubeflow?

- End-to-End ML workflow
- Scalability and resource management
- Reproducibility and collaboration

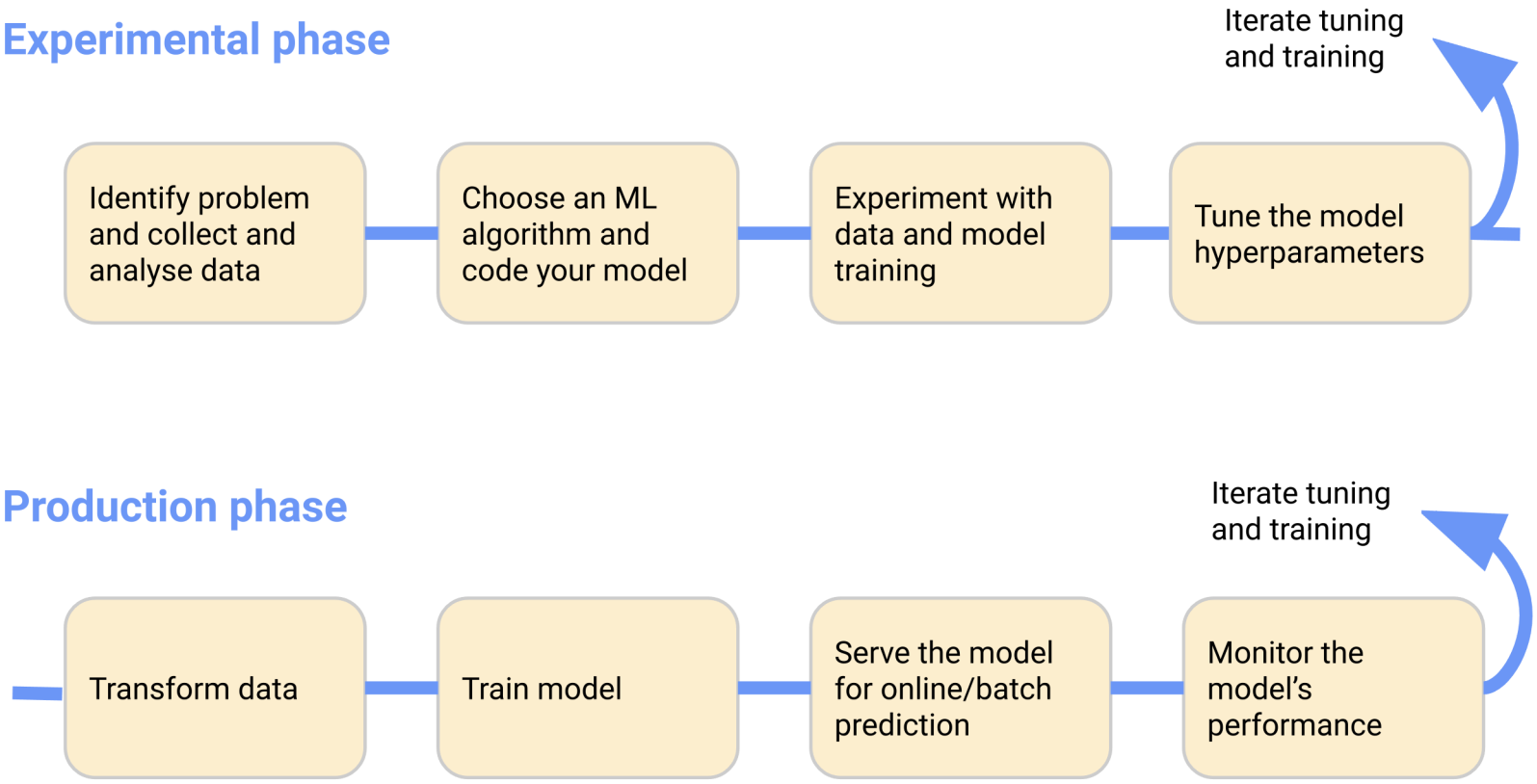


Source: https://www.kubeflow.org/docs/started/architecture/



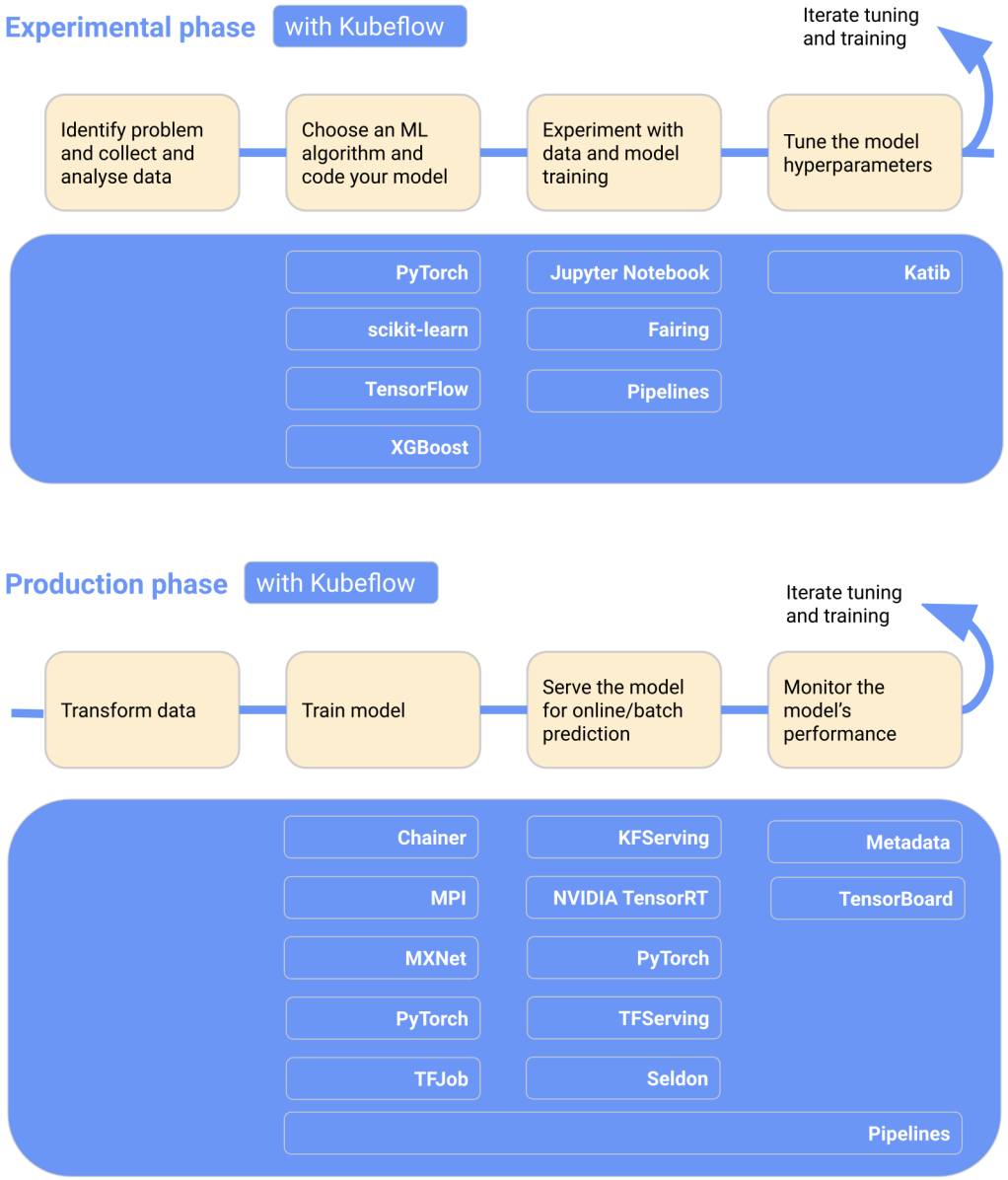
Source: https://www.analyticsvidhya.com/blog/2023/01/kubeflow-streamlining-mlops-with-efficient-ml-workflow-management//

Before Kubeflow



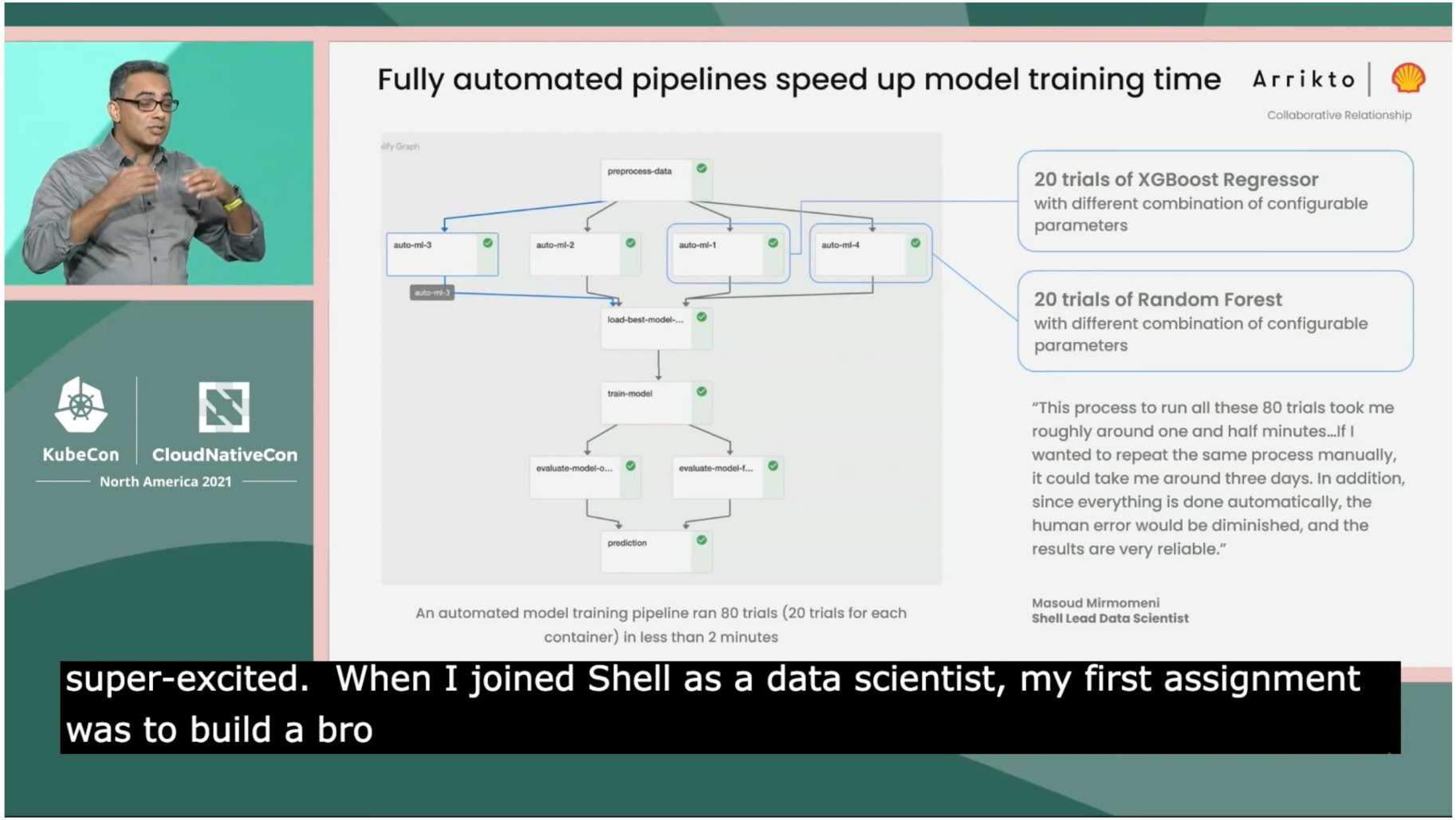
Source: https://www.kubeflow.org/docs/started/architecture/

With Kubeflow



Source: https://www.kubeflow.org/docs/started/architecture/

Example: Shell

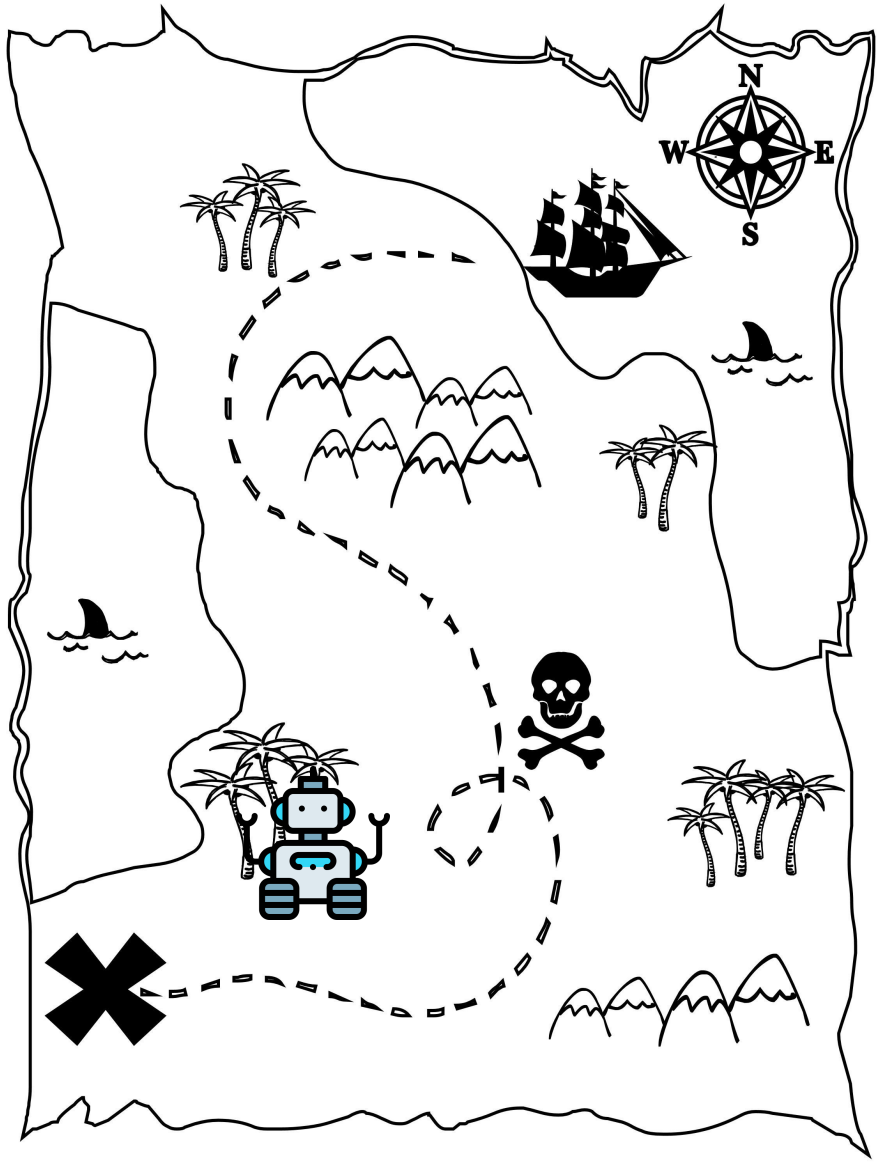


Keynote: Machine Learning on Kubernetes Made Easy With Kubeflow - Masoud Mirmomeni & Jimmy Guerrero

CNCF [Cloud Native Computing Foundation]
105K subscribers
Subscribe
5
Share
Clip
Save

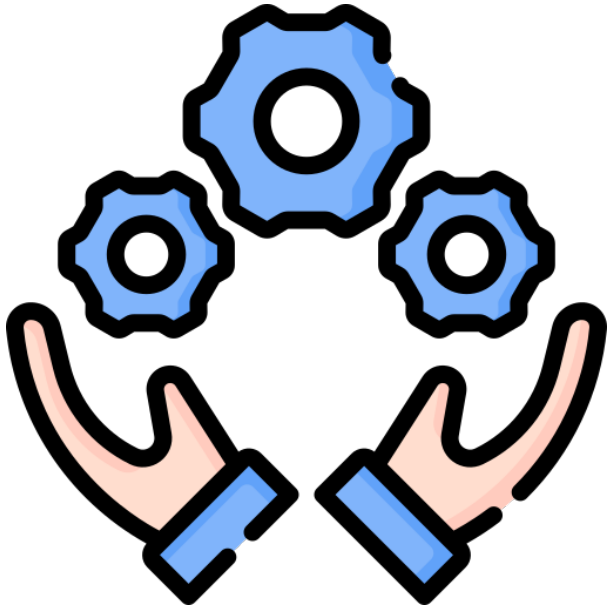
Agenda

- ~~Introduction~~
- ~~Definitions~~
- ~~Why is AI/ML different~~
- ~~Kubernetes & MLOps~~
- ~~Kubeflow~~
- Best practices
- Considerations
- Wrap up & Resources



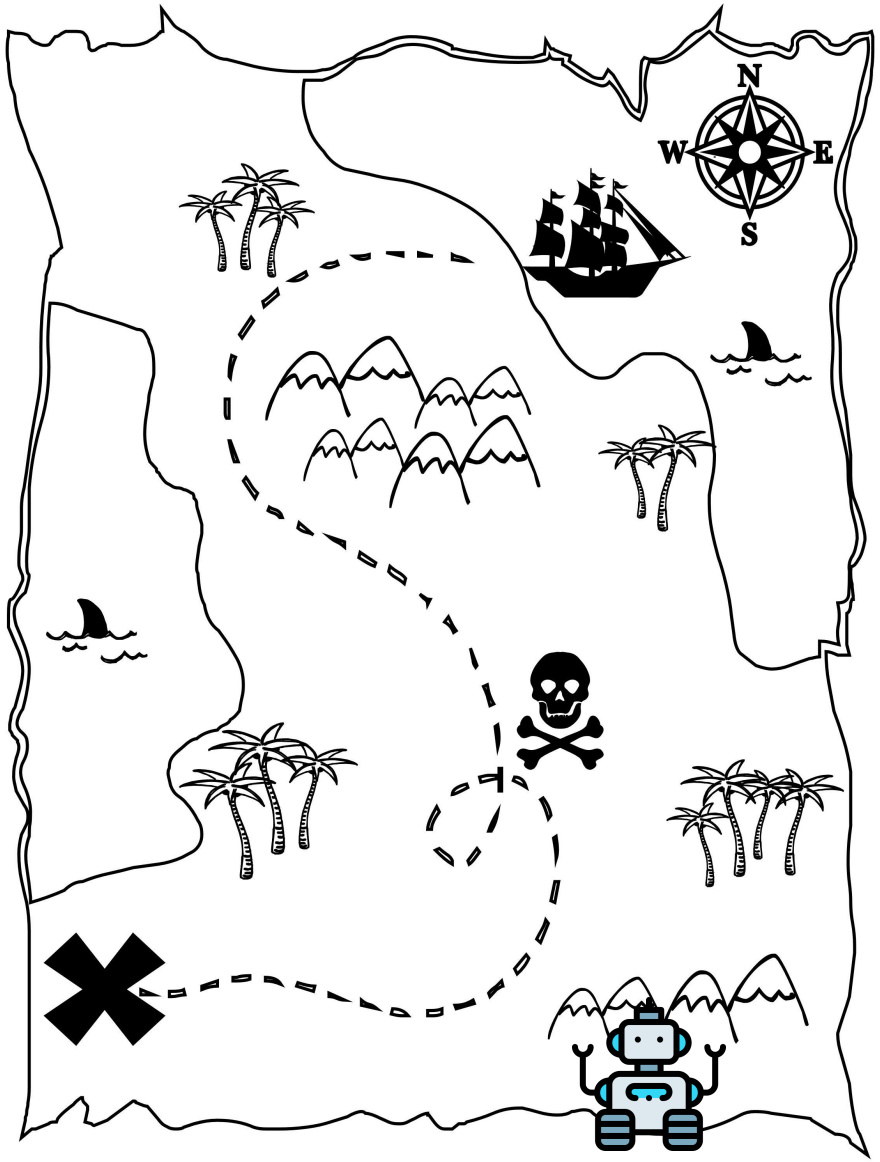
Best Practices

- It is always start with data
- Master all the necessary skills
- Know the tools you need and choose them wisely
- Understand the bottlenecks you face
- MLOps ❤️ DevOps
- Consider serverless
- AutoML, Kaizen



Agenda

- ~~Introduction~~
- ~~Definitions~~
- ~~Why is AI/ML different~~
- ~~Kubernetes & MLOps~~
- ~~Kubeflow~~
- ~~Best practices~~
- Considerations
- Wrap up & Resources



Ethical considerations

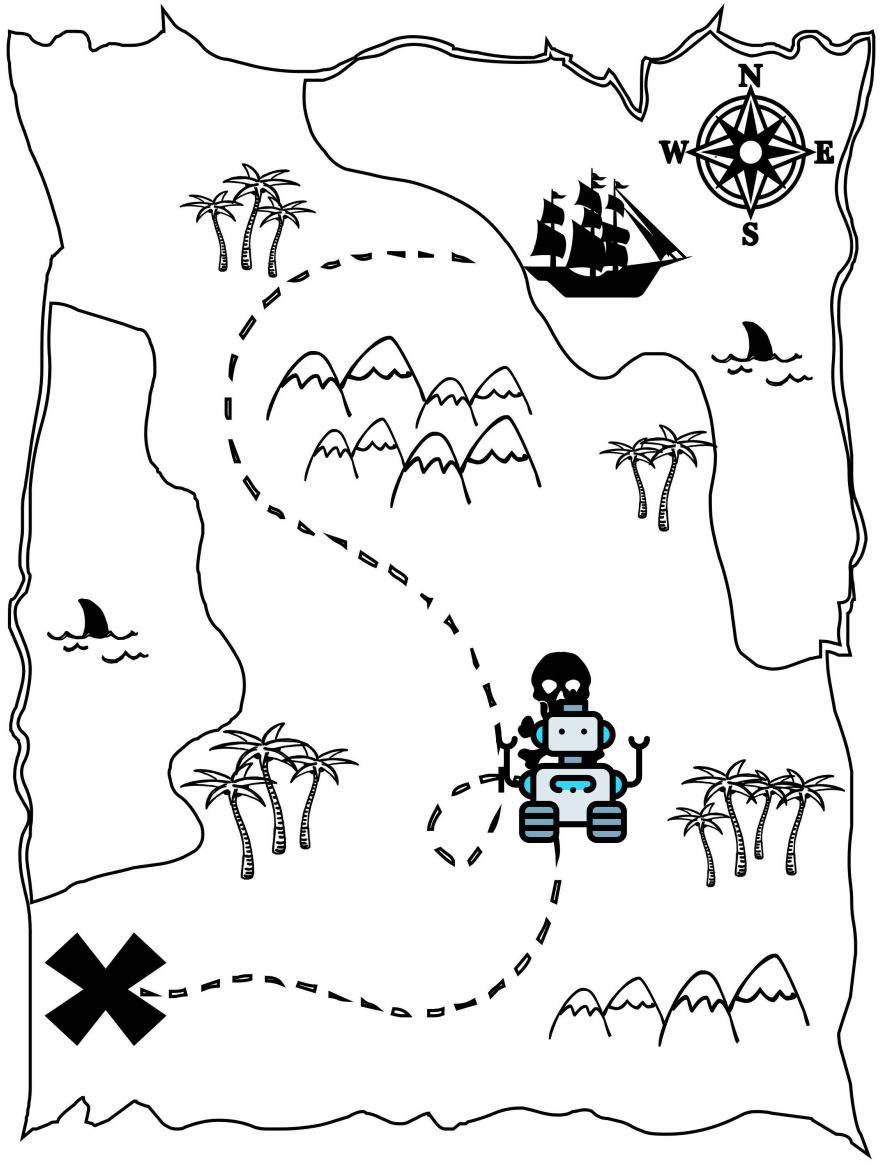
- Fairness & Bias
- Transparency
- Privacy
- Accountability
- Consent and user empowerment
- Security
- Human-Centric Design
- Data Governance
- Long-term consequences
- Regulatory Compliance

Learn more

- Links and slides: github.com/annietalvasto
- Practical MLOps by Noah Gift and Alfredo Deza (O'Reilly). Copyright 2021 Noah Gift and Alfredo Deza, 978-1-098-10301-9.
- Keynote: Machine Learning on Kubernetes Made Easy With Kubeflow - Masoud Mirmomeni & Jimmy Guerrero
  - https://www.youtube.com/watch?v=ick5hI5YI0k
- Kubeflow for Machine Learning • Holden Karau & Adi Polak • GOTO 2022
  - https://www.youtube.com/watch?v=dUZ5TrSReZc&t
- Webinar: MLOps automation with Git Based CI/CD for ML
  - https://www.youtube.com/watch?v=VCUDo9umKEQ
- All Kubernetes AI day Sessions
- Welcome + Opening Remarks: The State of Production MLOps in the Cloud Native... - Alejandro Saucedo
  - https://www.youtube.com/watch?v=xymbp8RWaCQ&list=PLj6h78yzYM2M9oVaU3amsqL5RXUwcu

Agenda

- ~~Introduction~~
- ~~Definitions~~
- ~~Why is AI/ML different~~
- ~~Kubernetes & MLOps~~
- ~~Kubeflow~~
- ~~Best practices~~
- ~~Considerations~~
- ~~Wrap up & Resources~~



Thank you!