



# 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](https://cloudgossip.net)

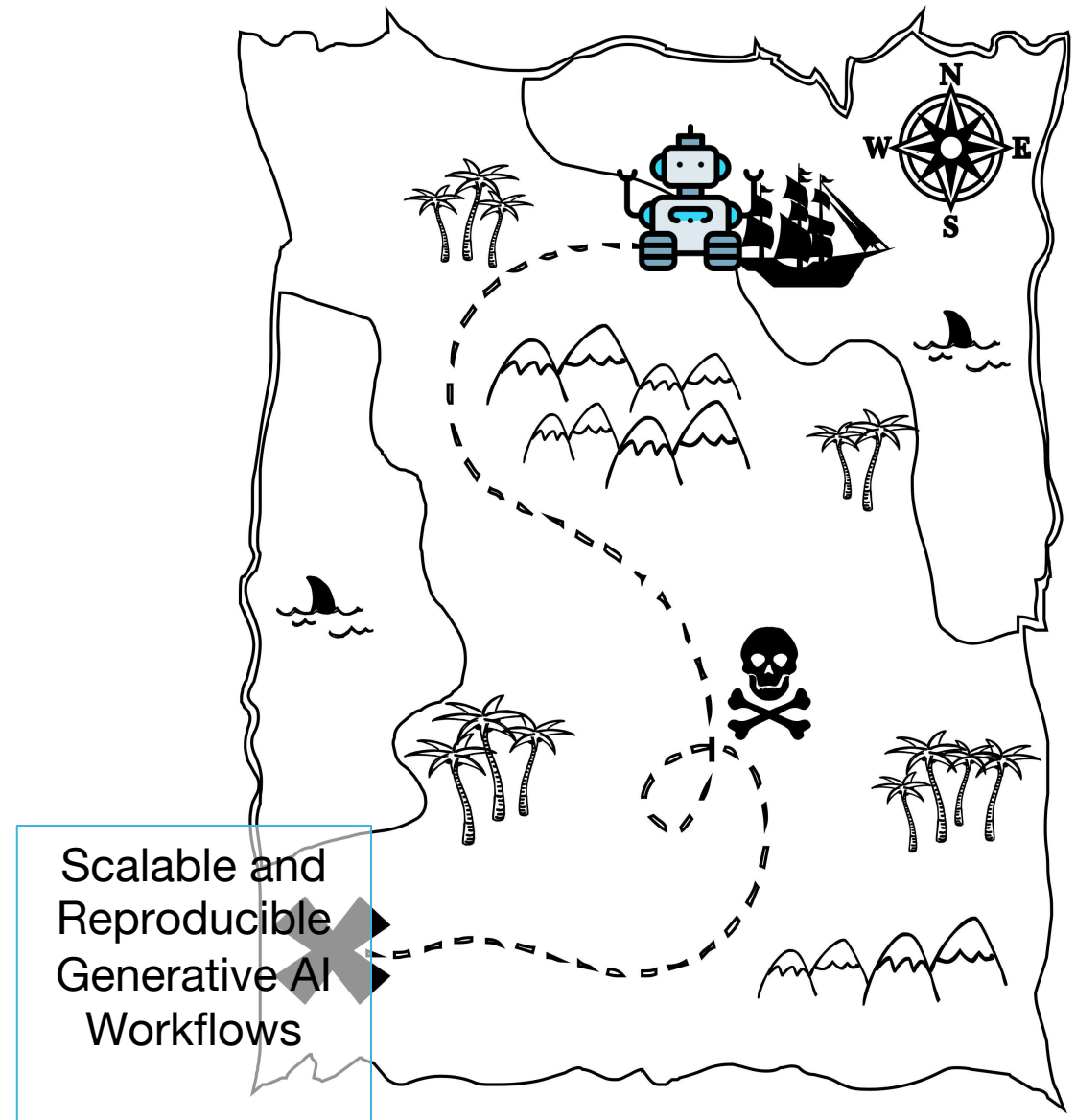


# Agenda

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

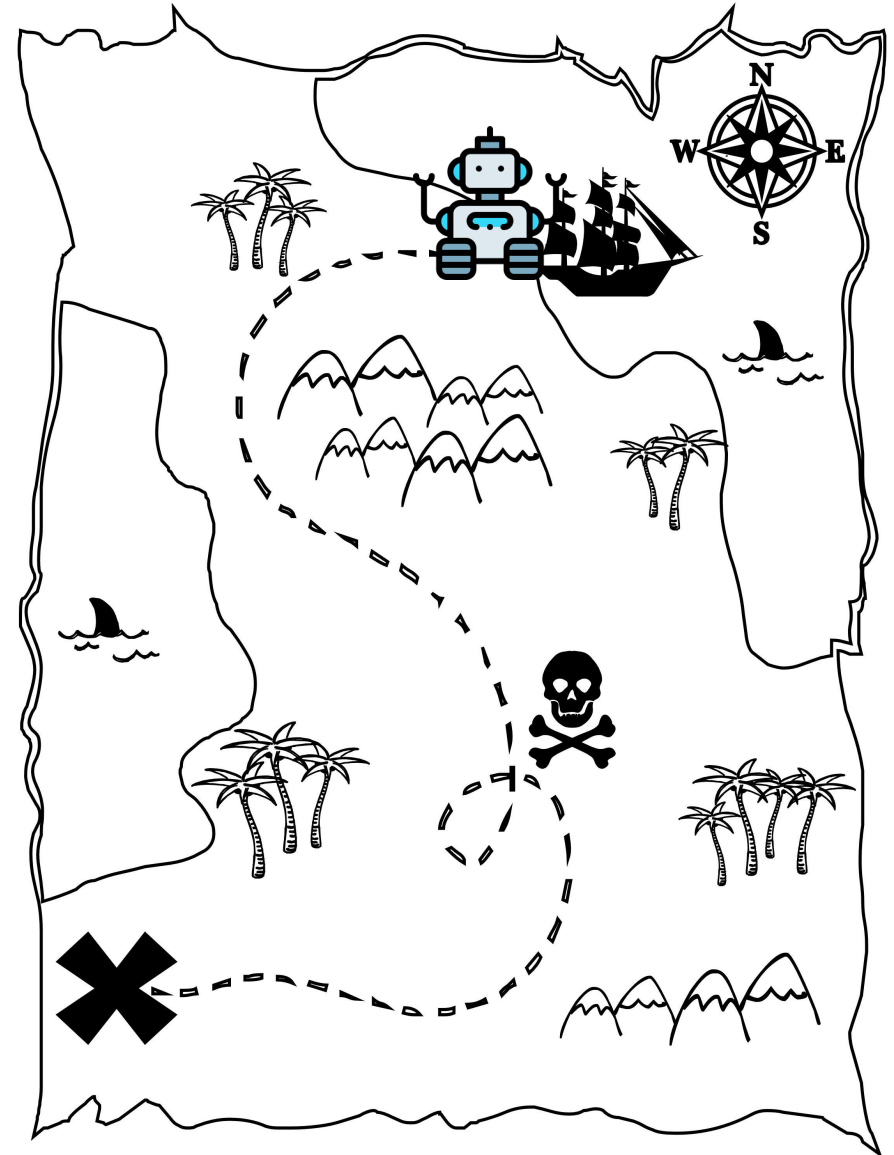
# Agenda

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



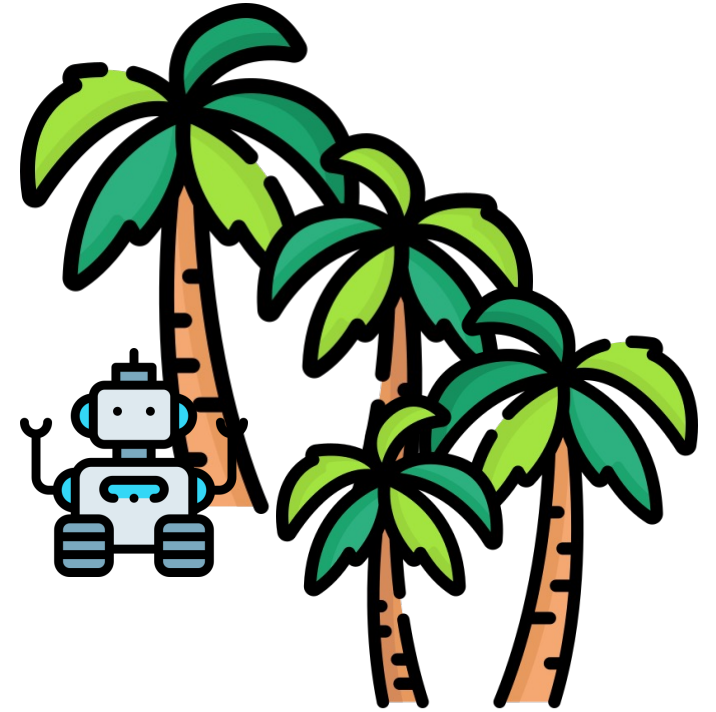
# Agenda

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



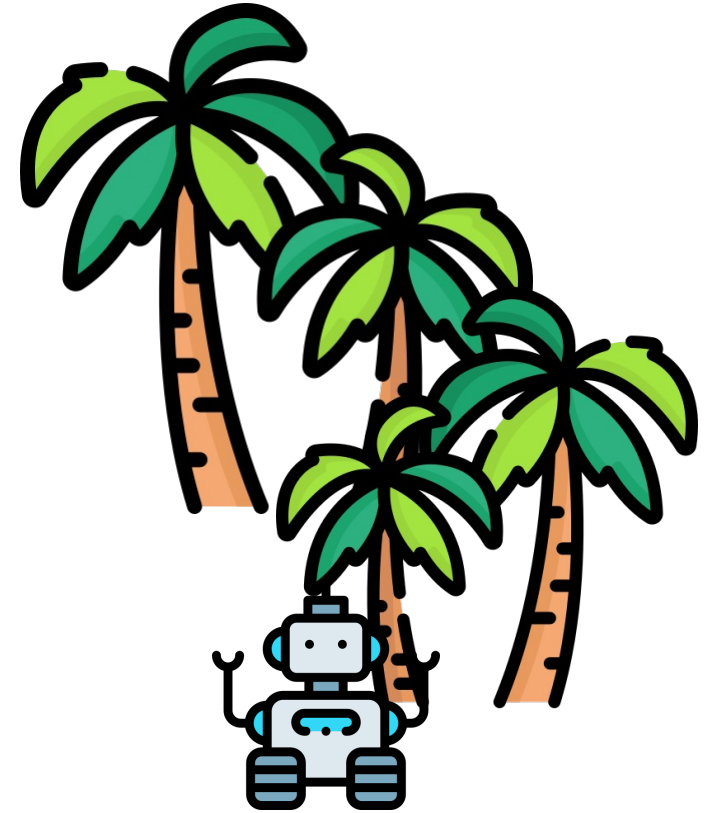
# AI

- 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.





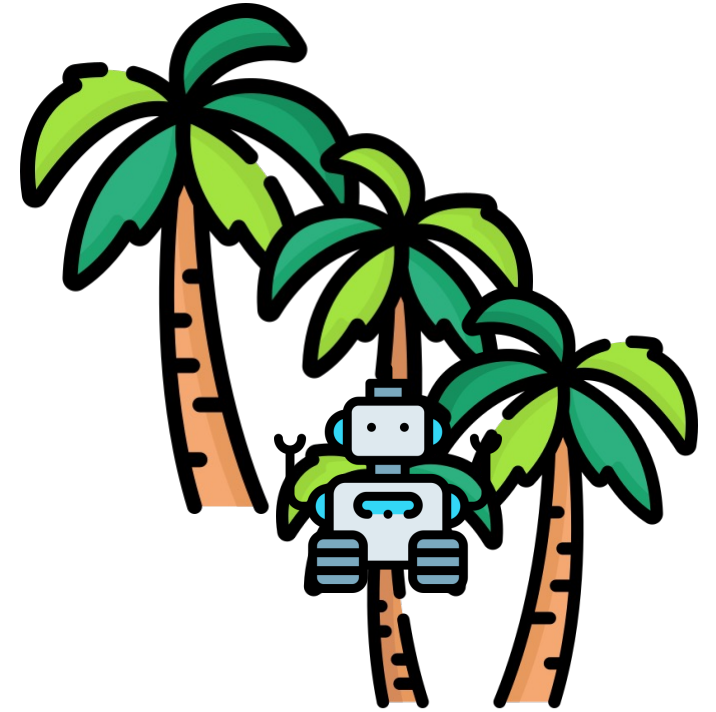
# Data science

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



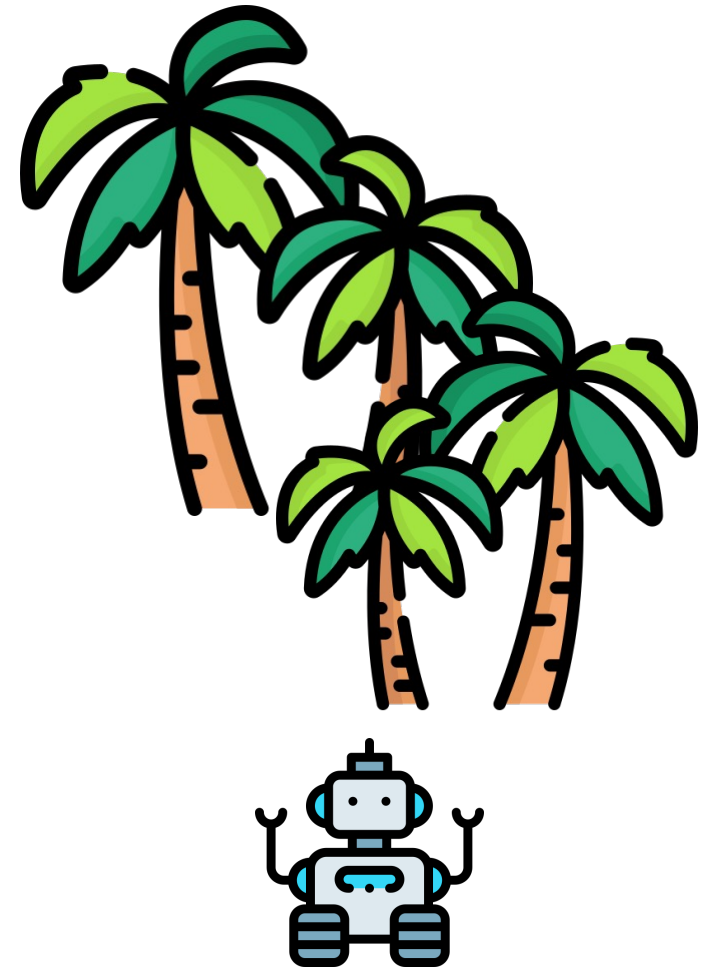
# DevOps

- 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.



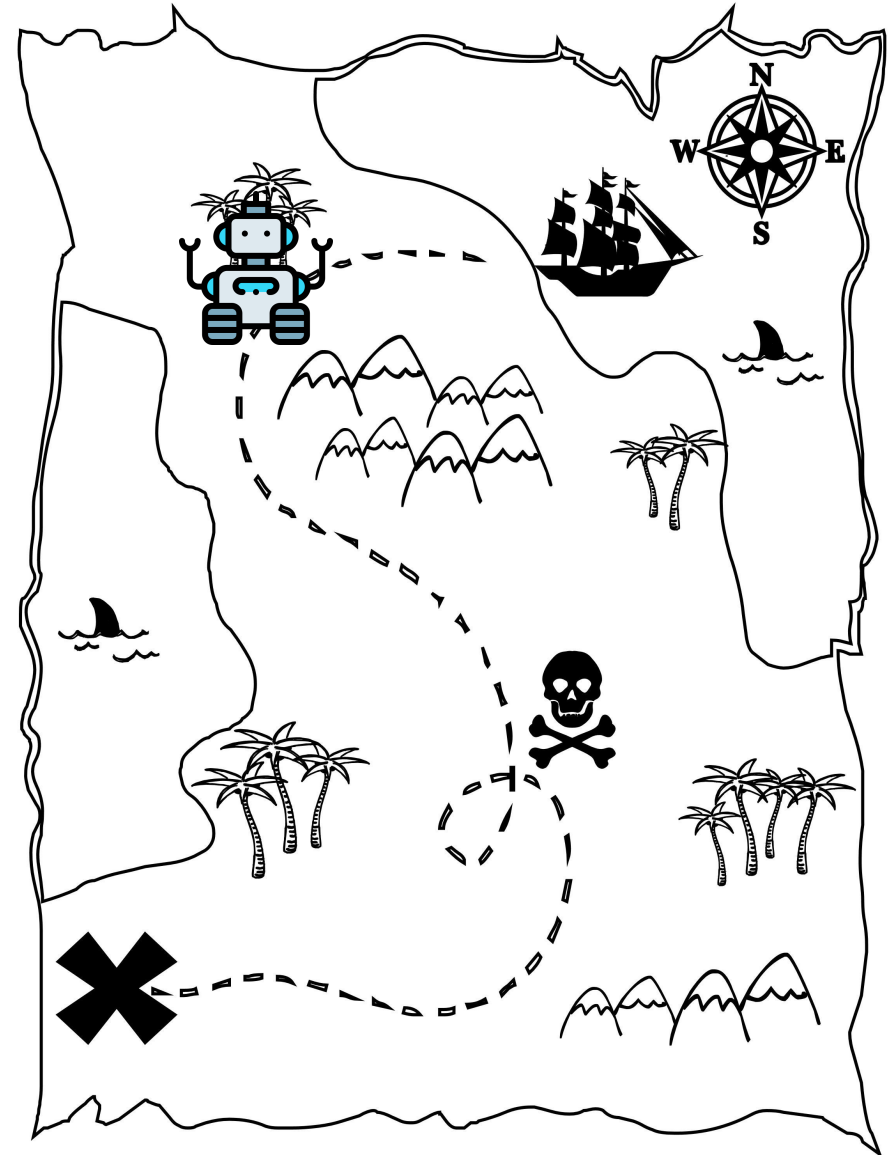
# MLOps

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



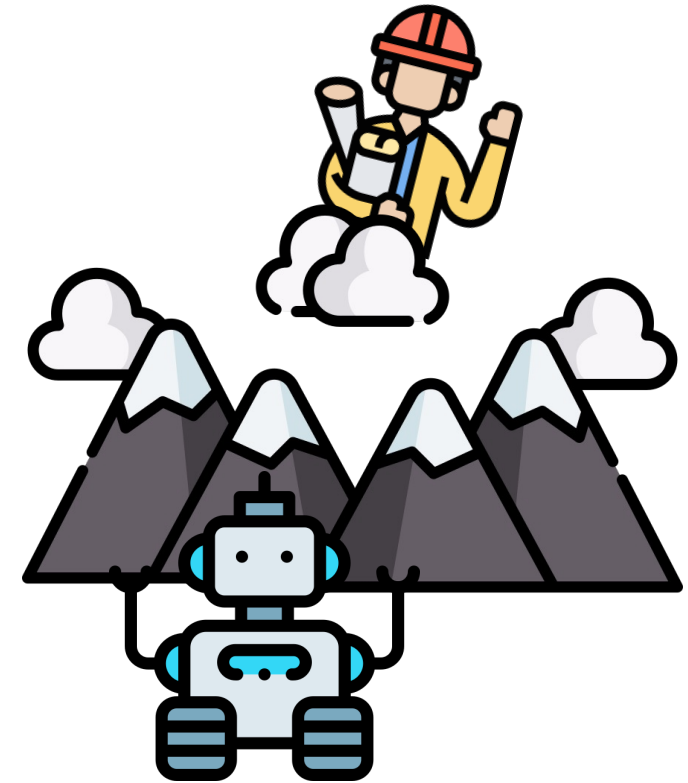
# Agenda

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

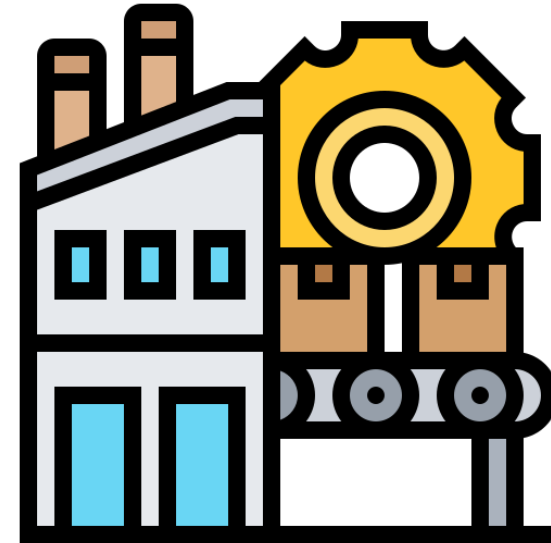
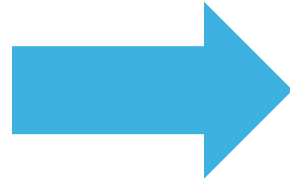


# Differences with AI/ML

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

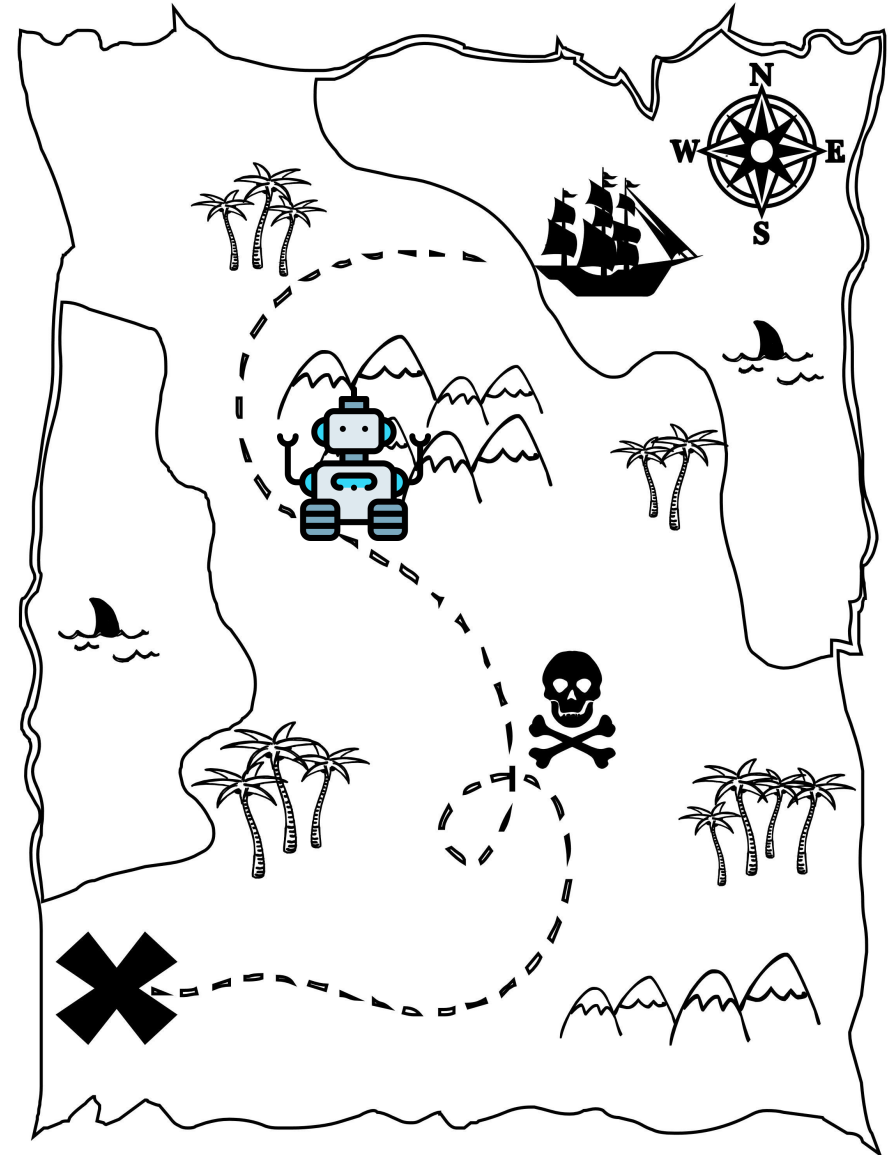


# From Research to Production



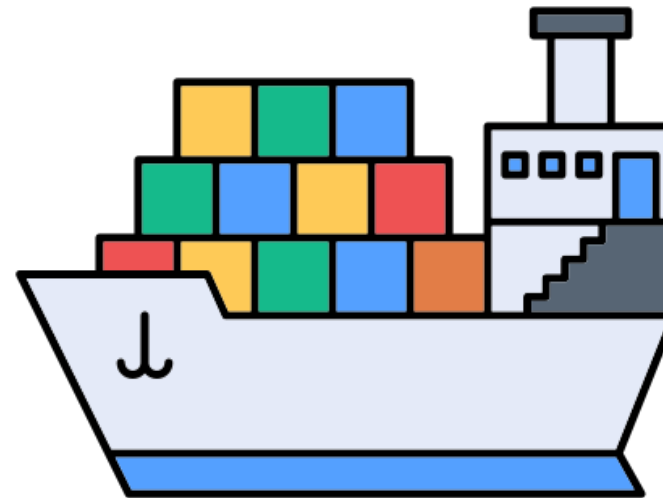
# Agenda

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



# Kubernetes & MLOps

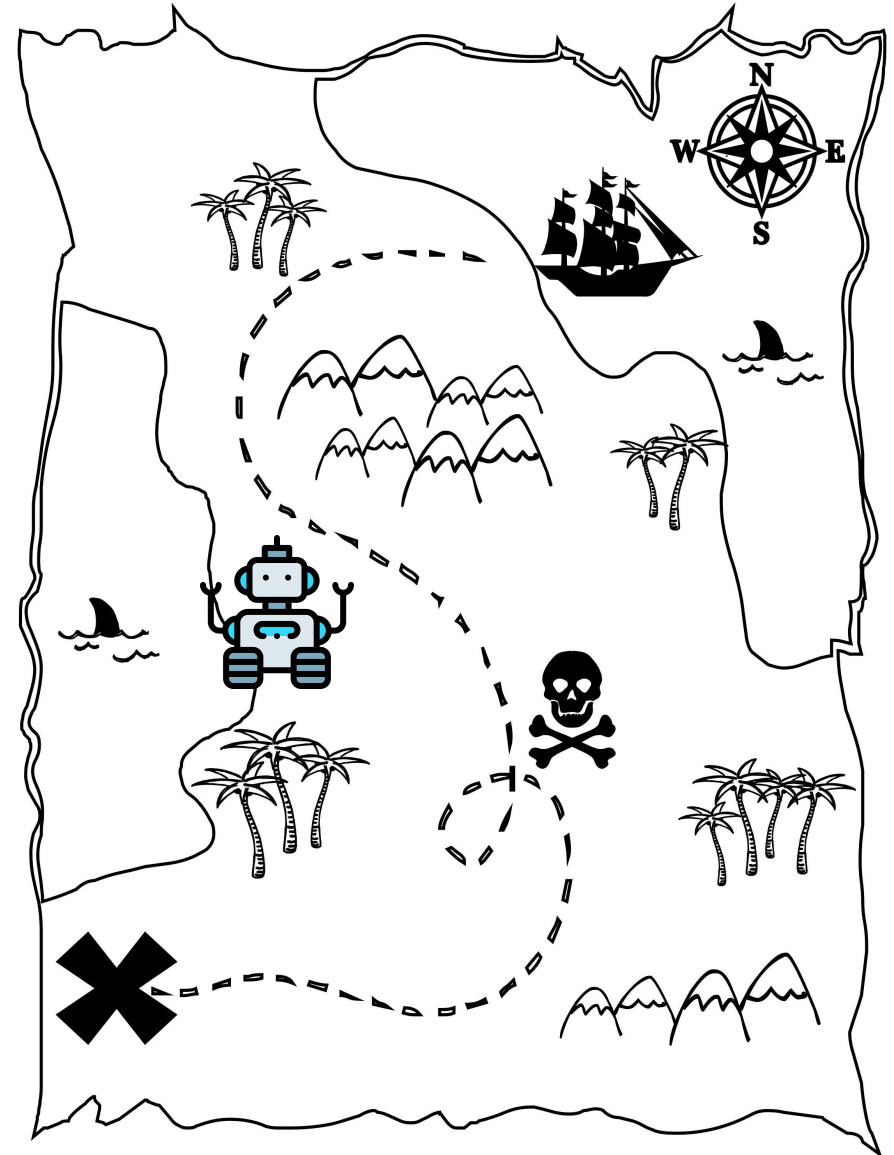
- Portability
- Customizability
- Performance
- Consistency
- Microservices
- Composability





# Agenda

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



# Kubeflow

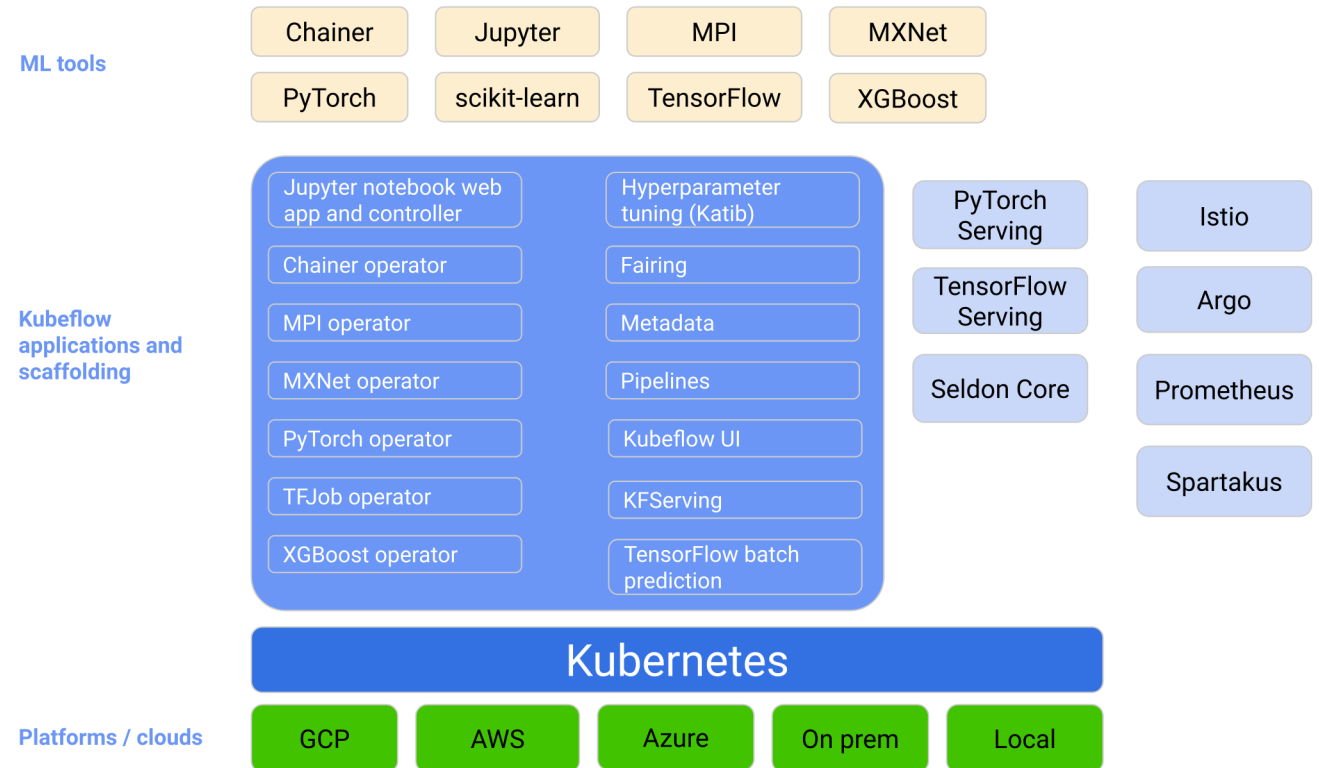
The machine learning  
toolkit for Kubernetes.



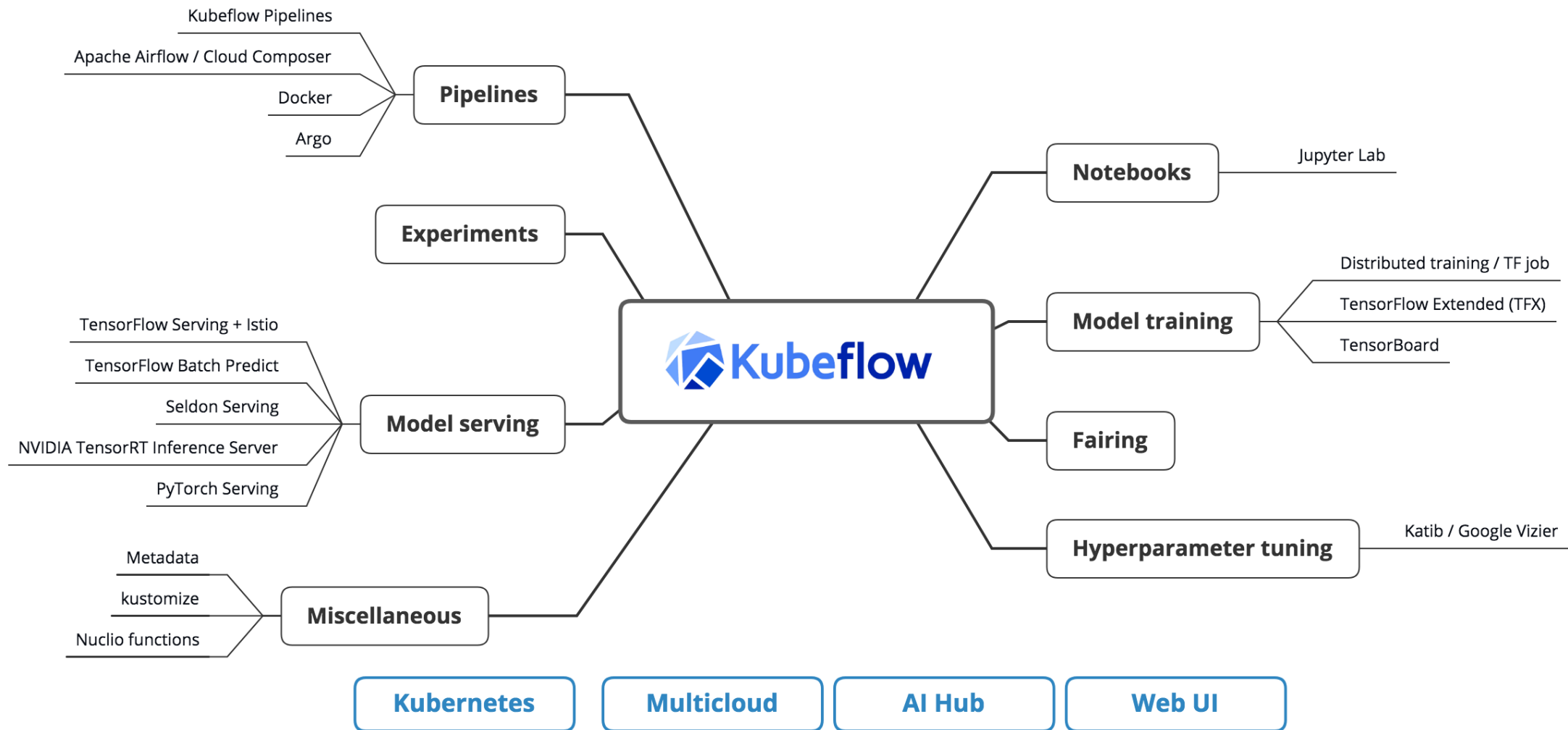
**Kubeflow**

# What is Kubeflow?

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



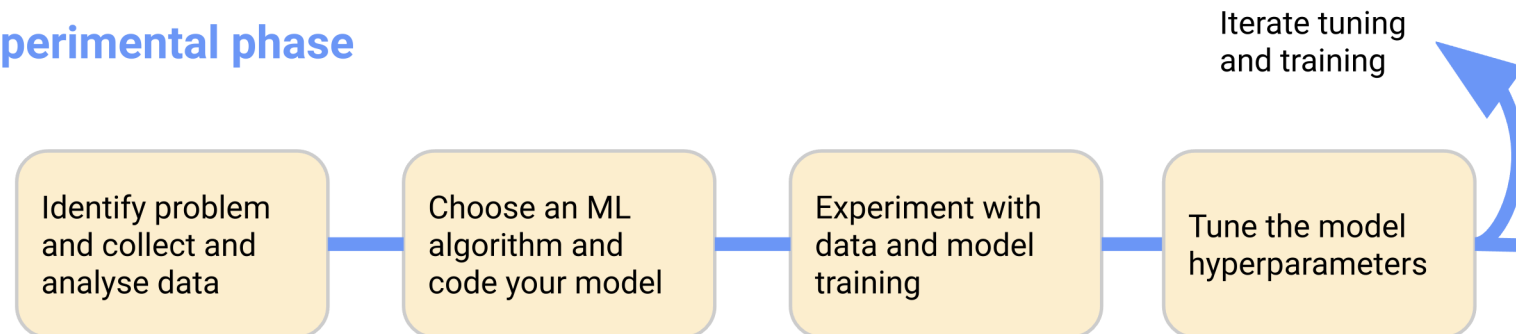
Source: <https://www.kubeflow.org/docs/started/architecture/>



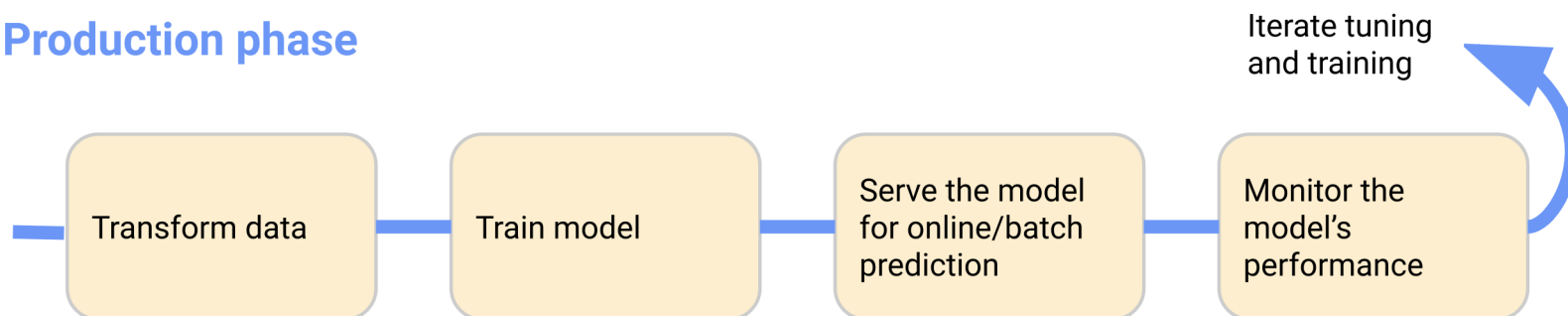
Source: <https://www.analyticsvidhya.com/blog/2023/01/kubeflow-streamlining-mlops-with-efficient-ml-workflow-management/>

# Before Kubeflow

## Experimental phase

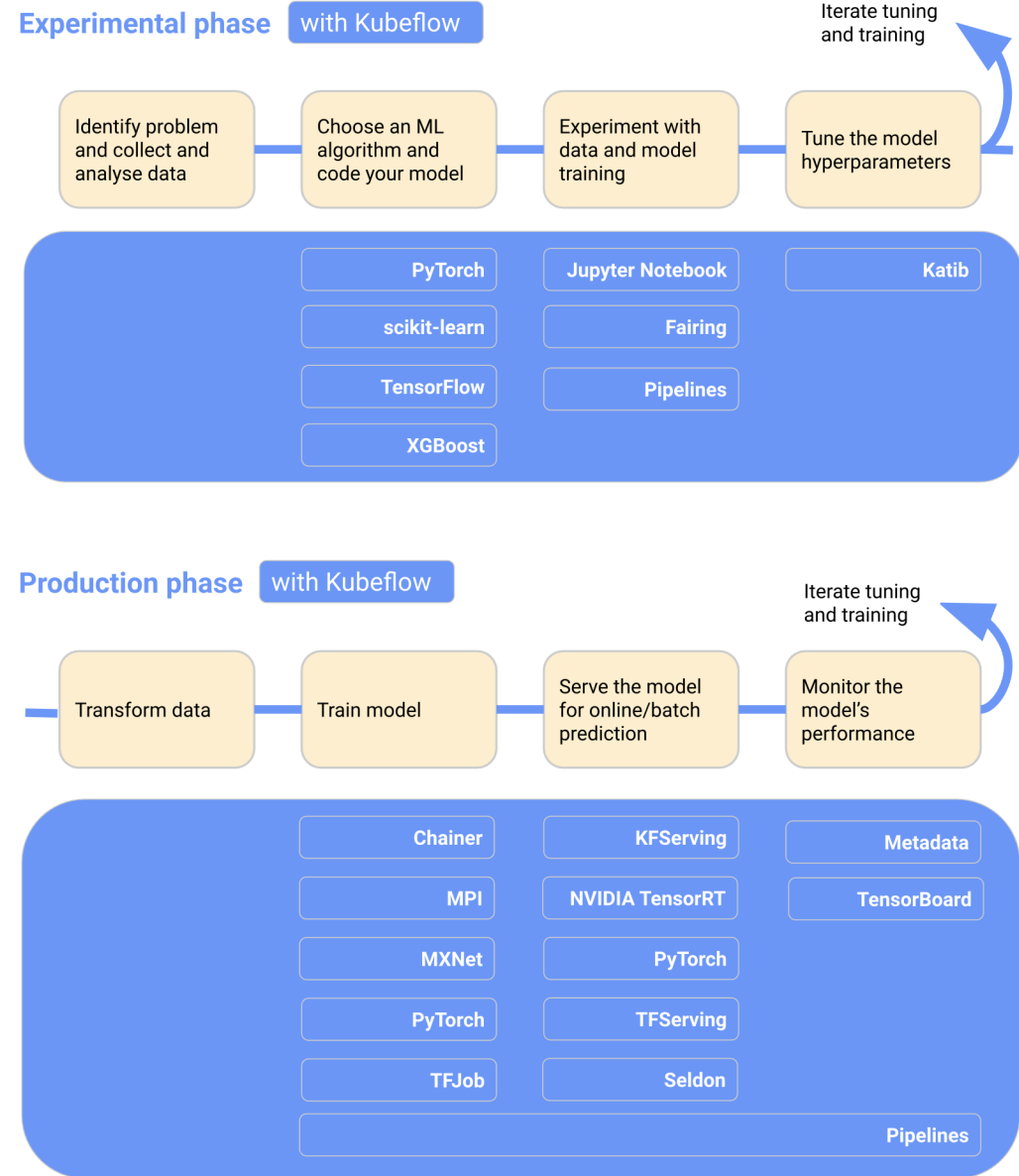


## Production phase





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

# With Kubeflow




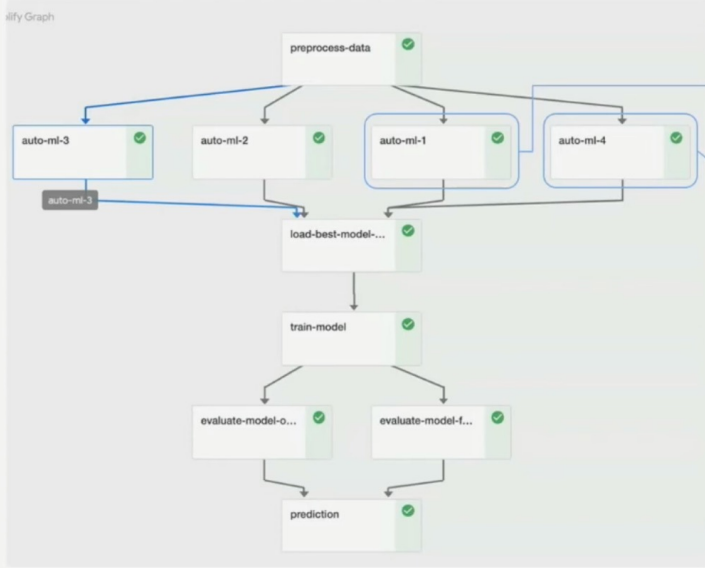
Source: <https://www.kubeflow.org/docs/started/architecture/>

# Example: Shell



## Fully automated pipelines speed up model training time

Arrikto |   
Collaborative Relationship



20 trials of XGBoost Regressor with different combination of configurable parameters

20 trials of Random Forest with different combination of configurable parameters

"This process to run all these 80 trials took me roughly around one and half minutes...If I wanted to repeat the same process manually, it could take me around three days. In addition, since everything is done automatically, the human error would be diminished, and the results are very reliable."

Masoud Mirmomeni  
Shell Lead Data Scientist

An automated model training pipeline ran 80 trials (20 trials for each container) in less than 2 minutes

super-excited. When I joined Shell as a data scientist, my first assignment was to build a bro

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



CNCF [Cloud Native Computing Foundation]  
105K subscribers

Subscribe



5



Share



Clip



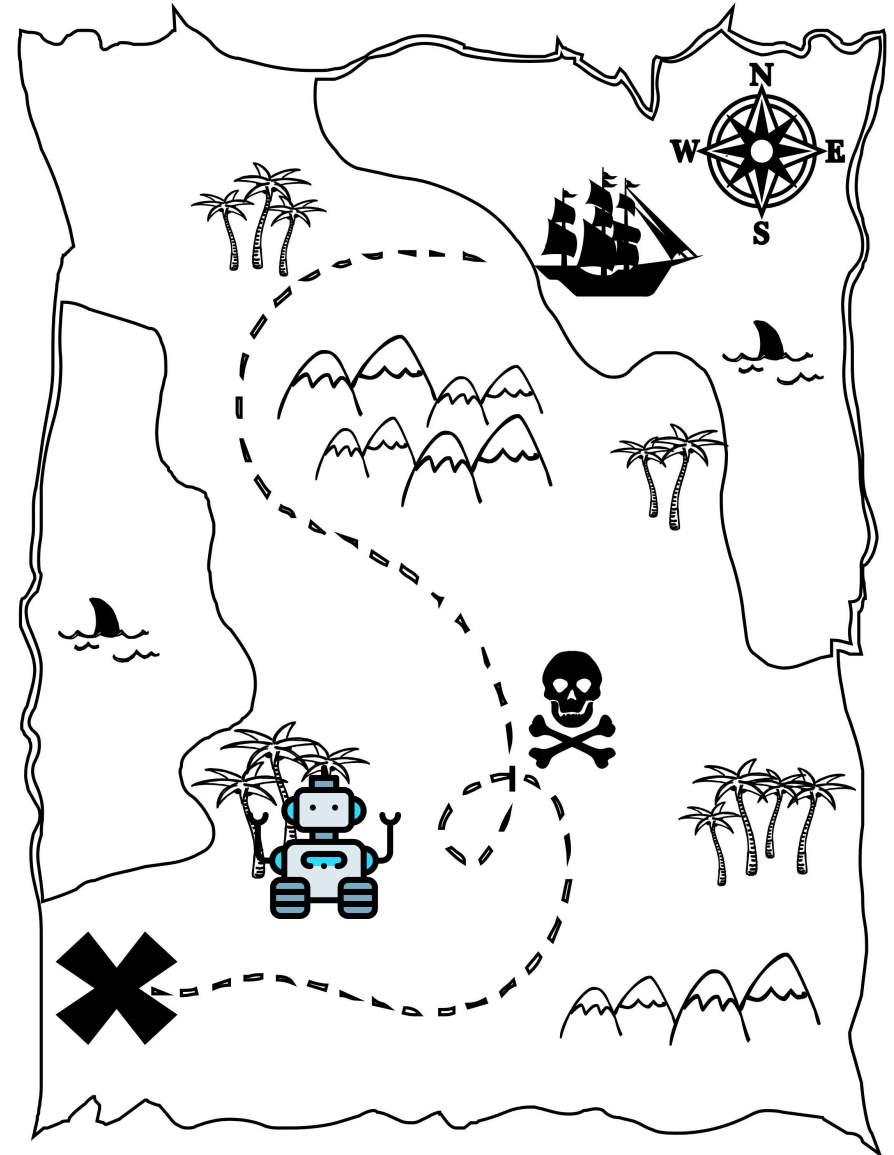
Save



@AnnieTalvasto

# Agenda

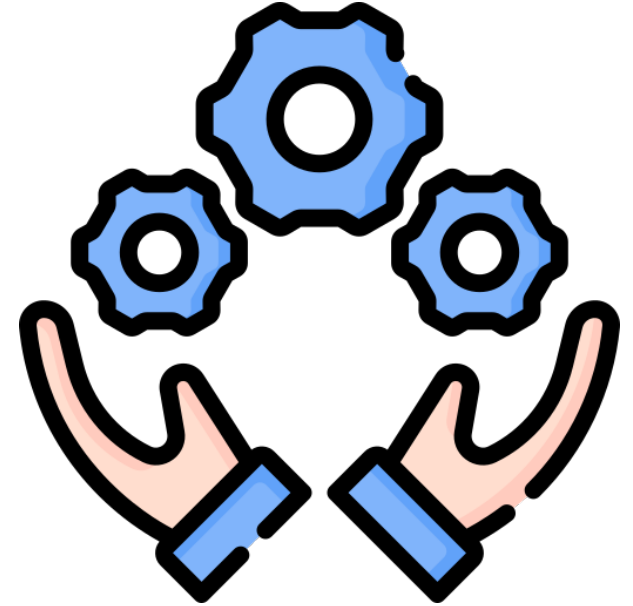
- Introduction
- Definitions
- ~~Why is AI/ML different~~
- ~~Kubernetes & MLOps~~
- ~~Kubeflow~~
- Best practices
- Considerations
- Wrap up & Resources





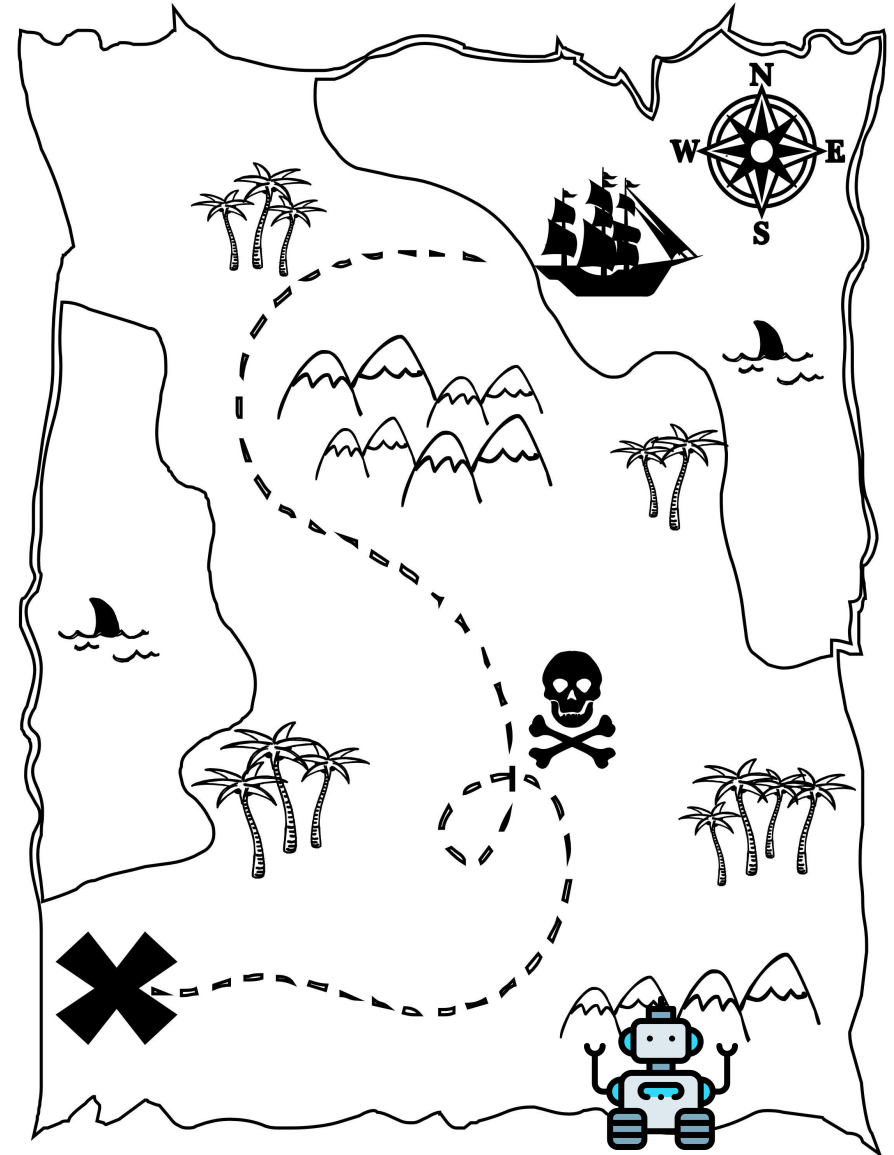
# Best Practices

- 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

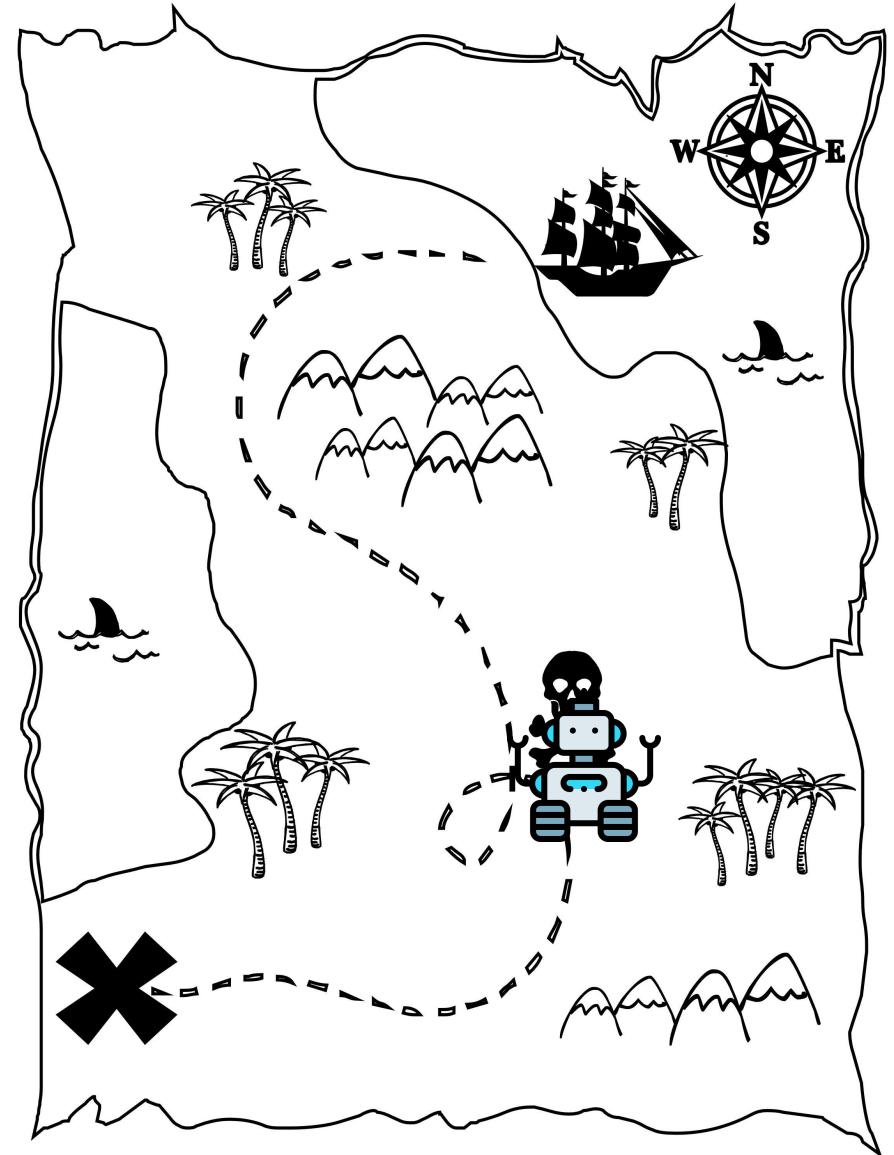


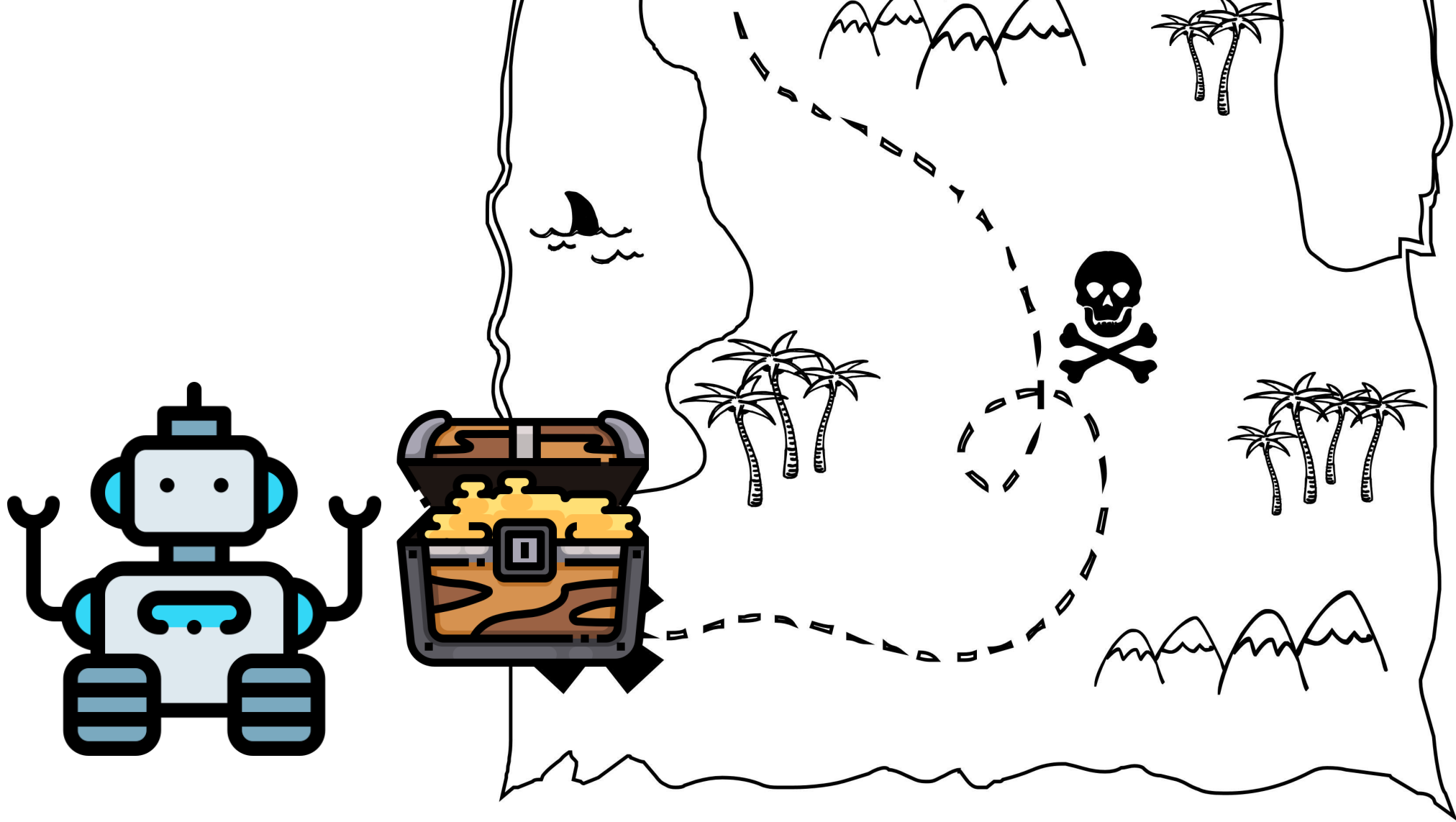
# Learn more

- Links and slides: [github.com/annietalvasto](https://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=PLj6h78yzYM2M9oVaU3amsqL5RXUwc\\_u](https://www.youtube.com/watch?v=xymbp8RWaCQ&list=PLj6h78yzYM2M9oVaU3amsqL5RXUwc_u)

# Agenda

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





Thank you!

