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

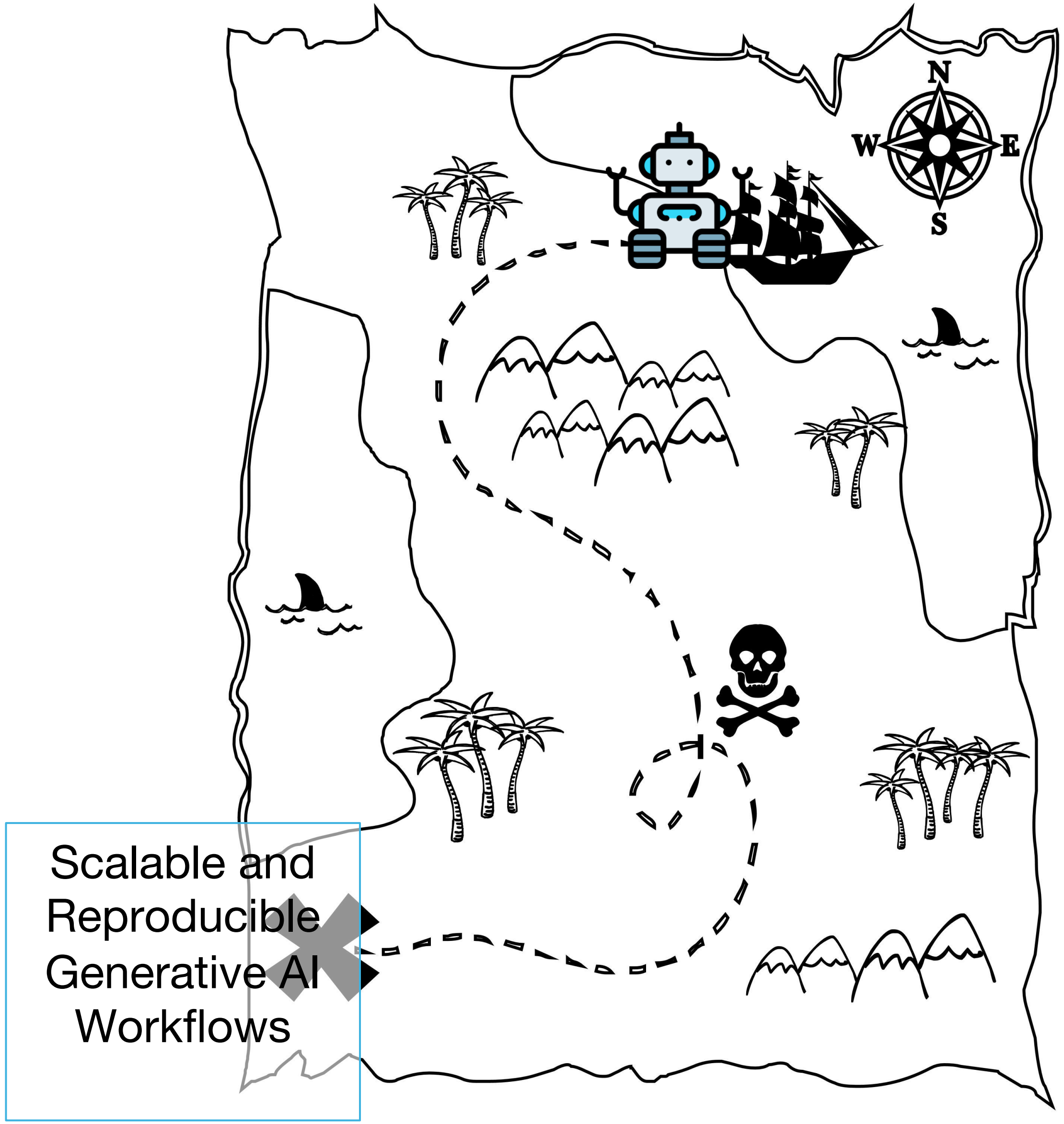
- Kubernetes & CNCF meetup co-organizer
- Startup-coach
- Created TechCraft Show – Tech interview show with Minecraft ⛏️💎🪨

Agenda

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

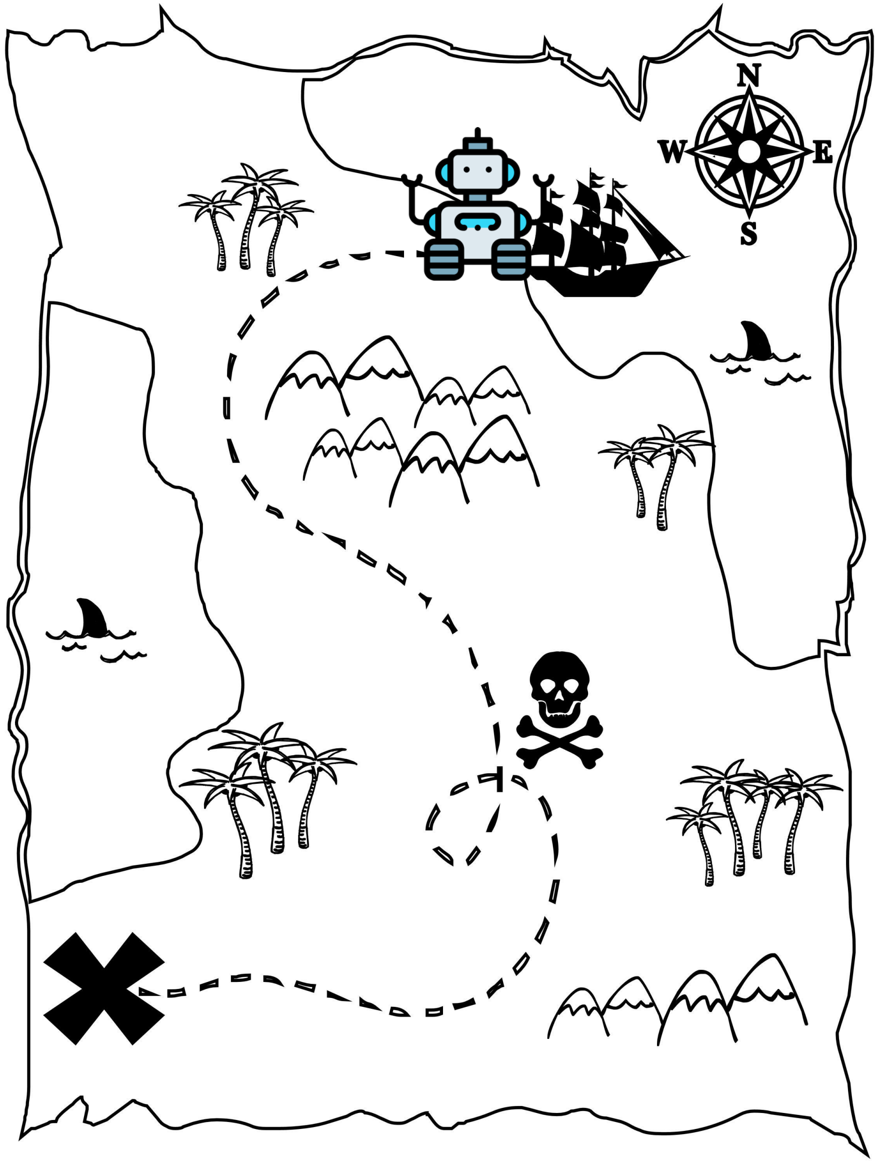
Agenda

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



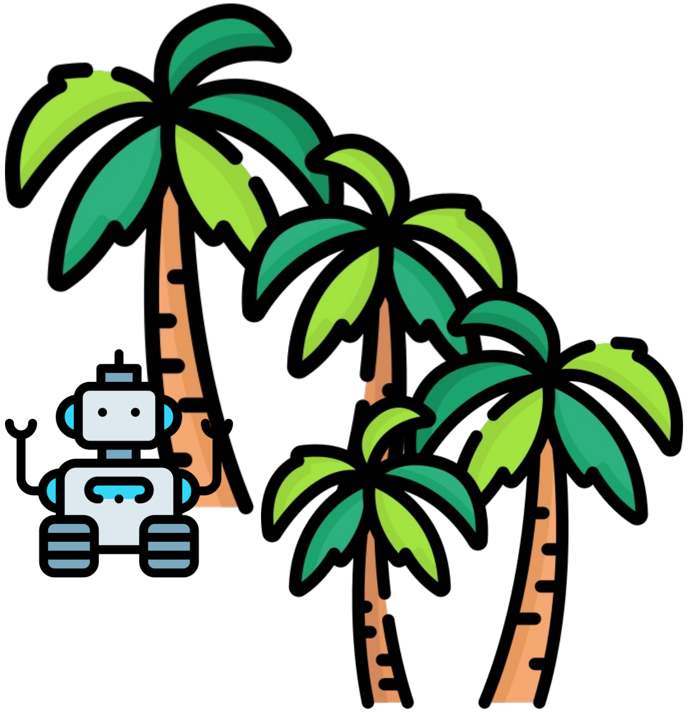
Agenda

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



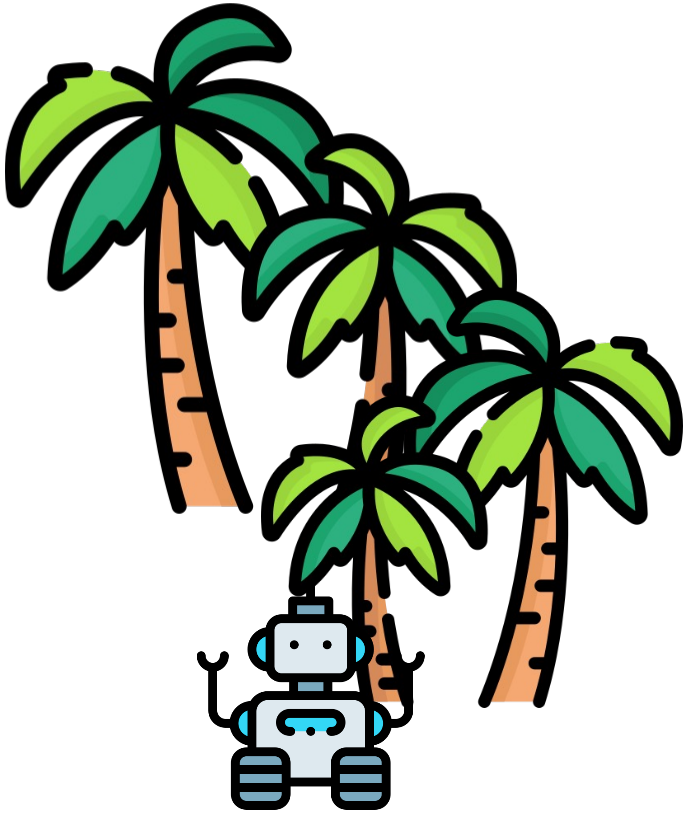
AI

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



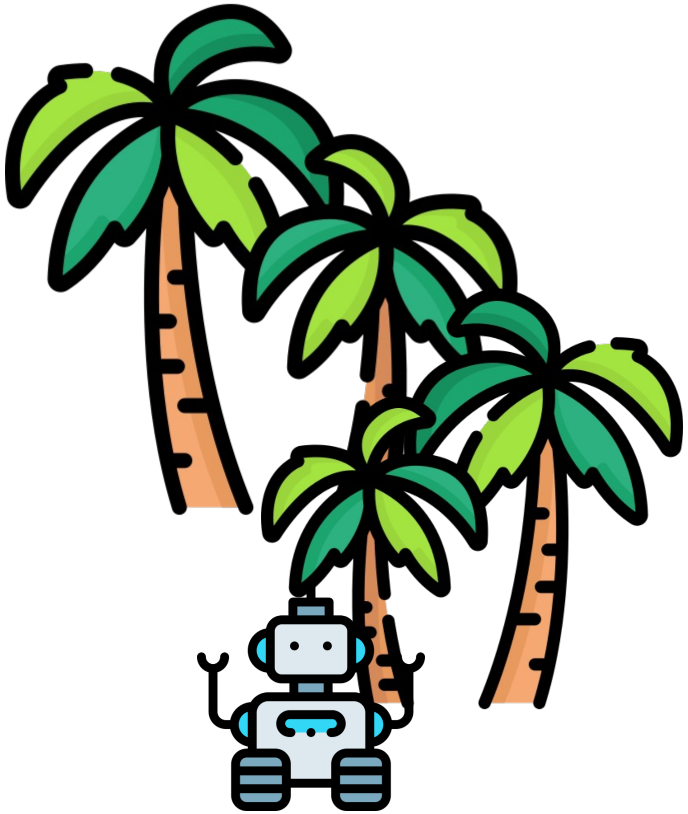
ML

- 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.



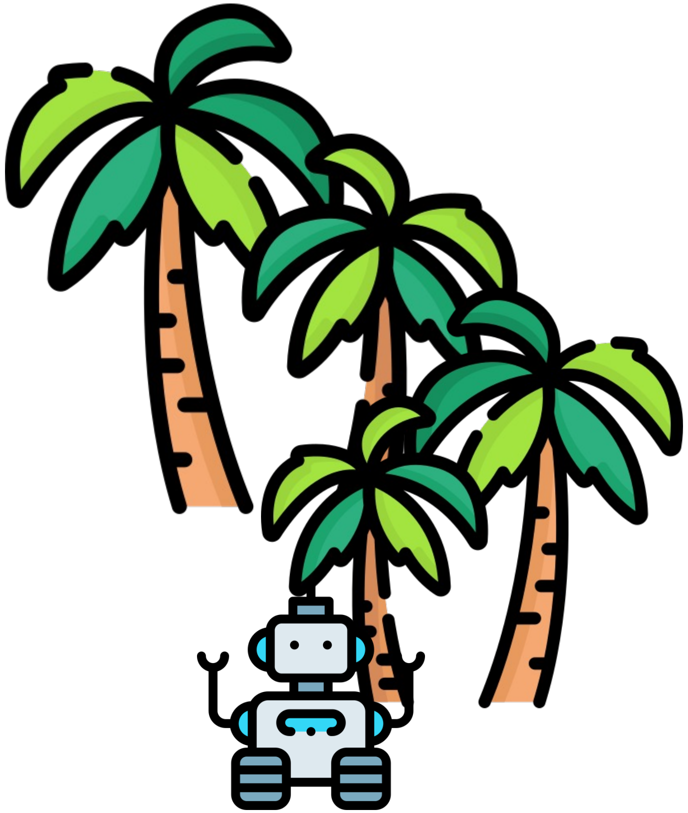
Data science

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



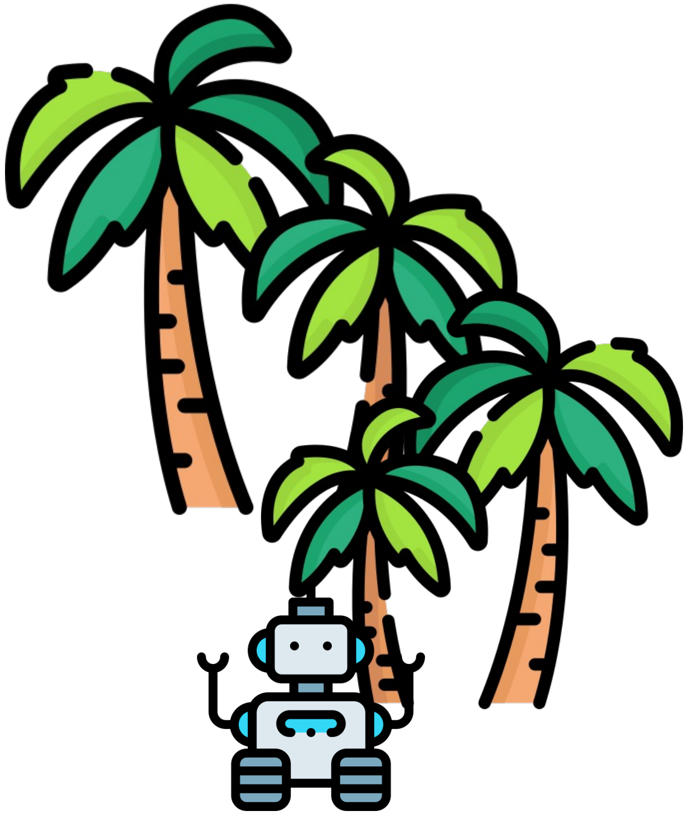
DevOps

- 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.



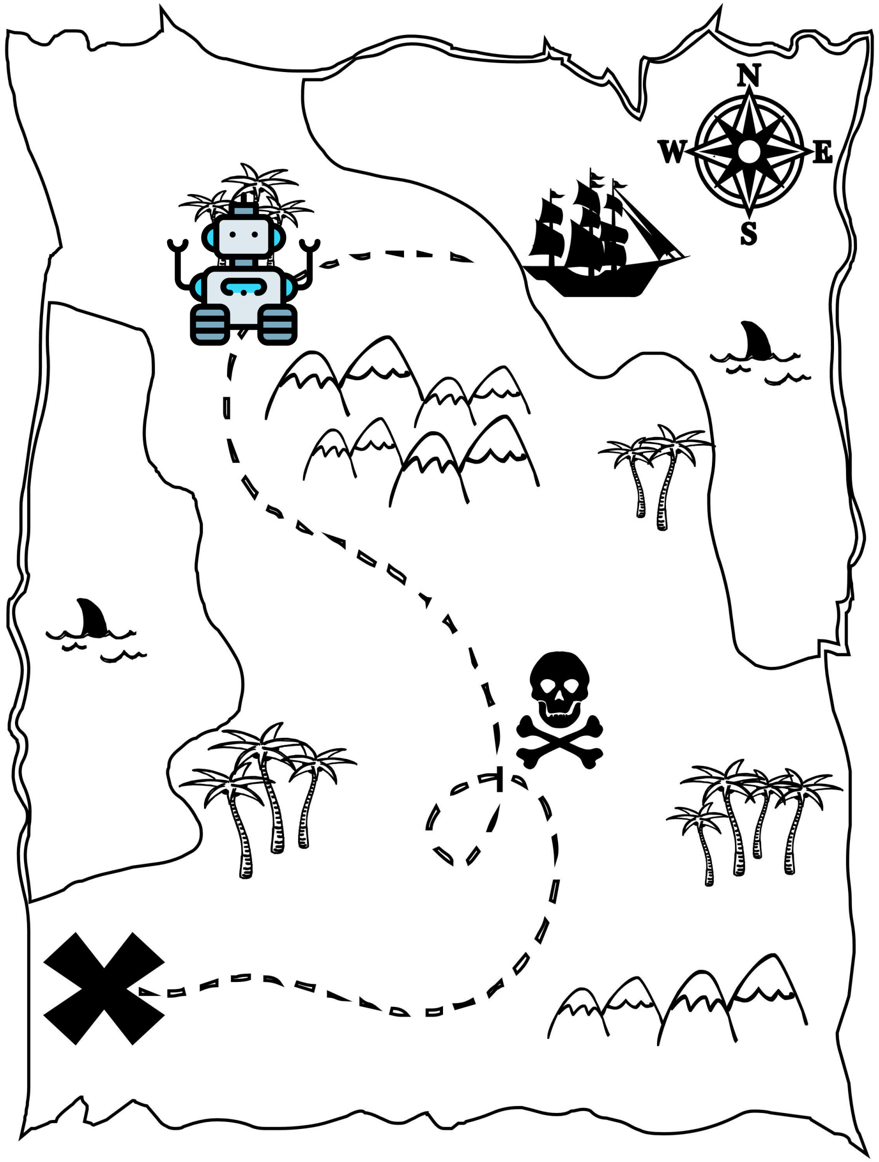
MLOps

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



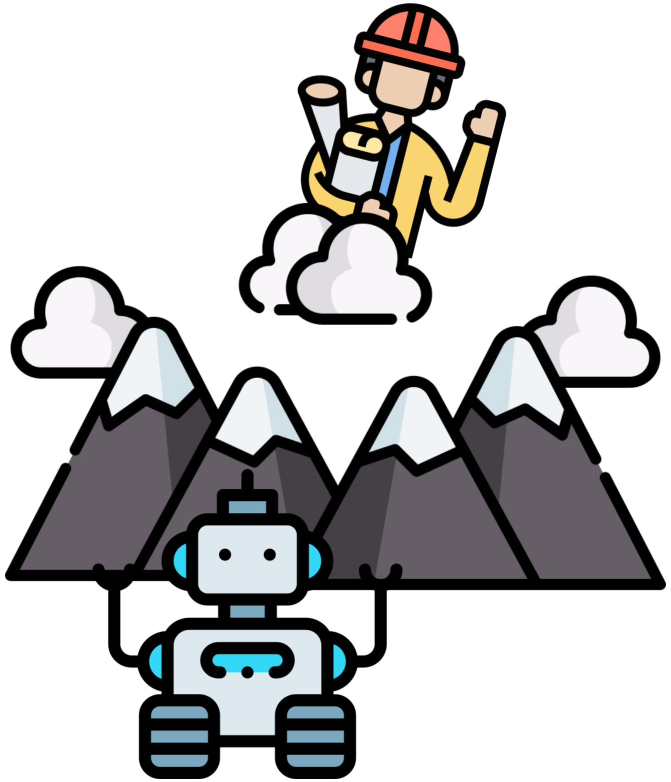
Agenda

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

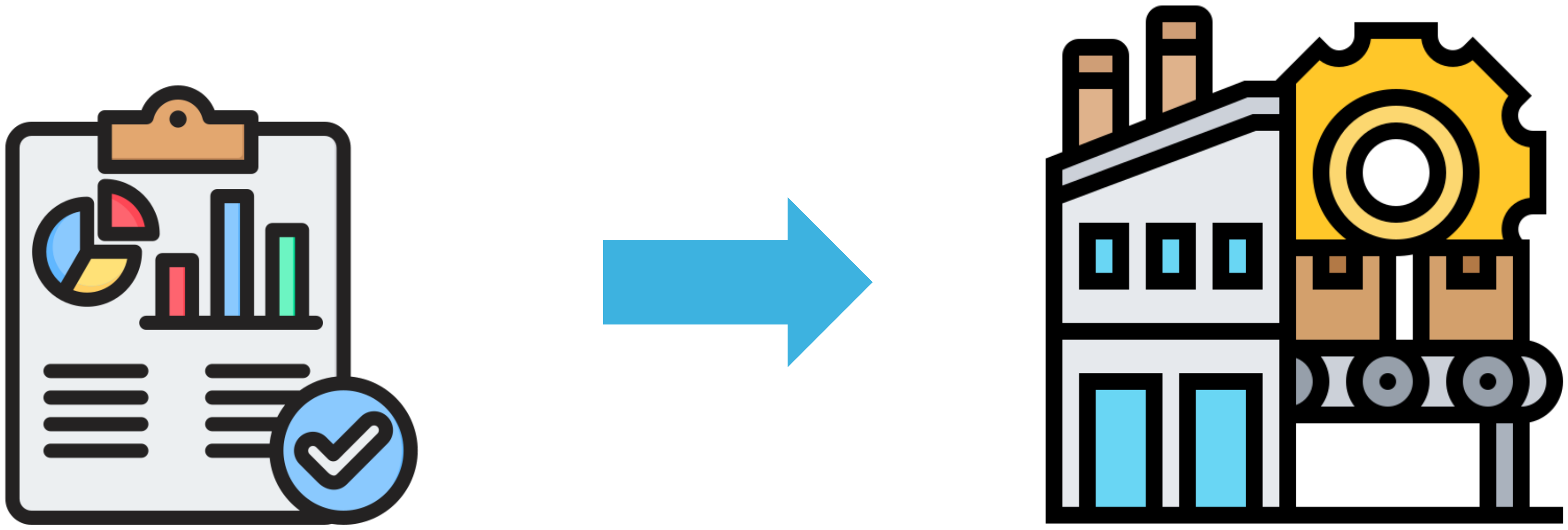


Differences with AI/ML

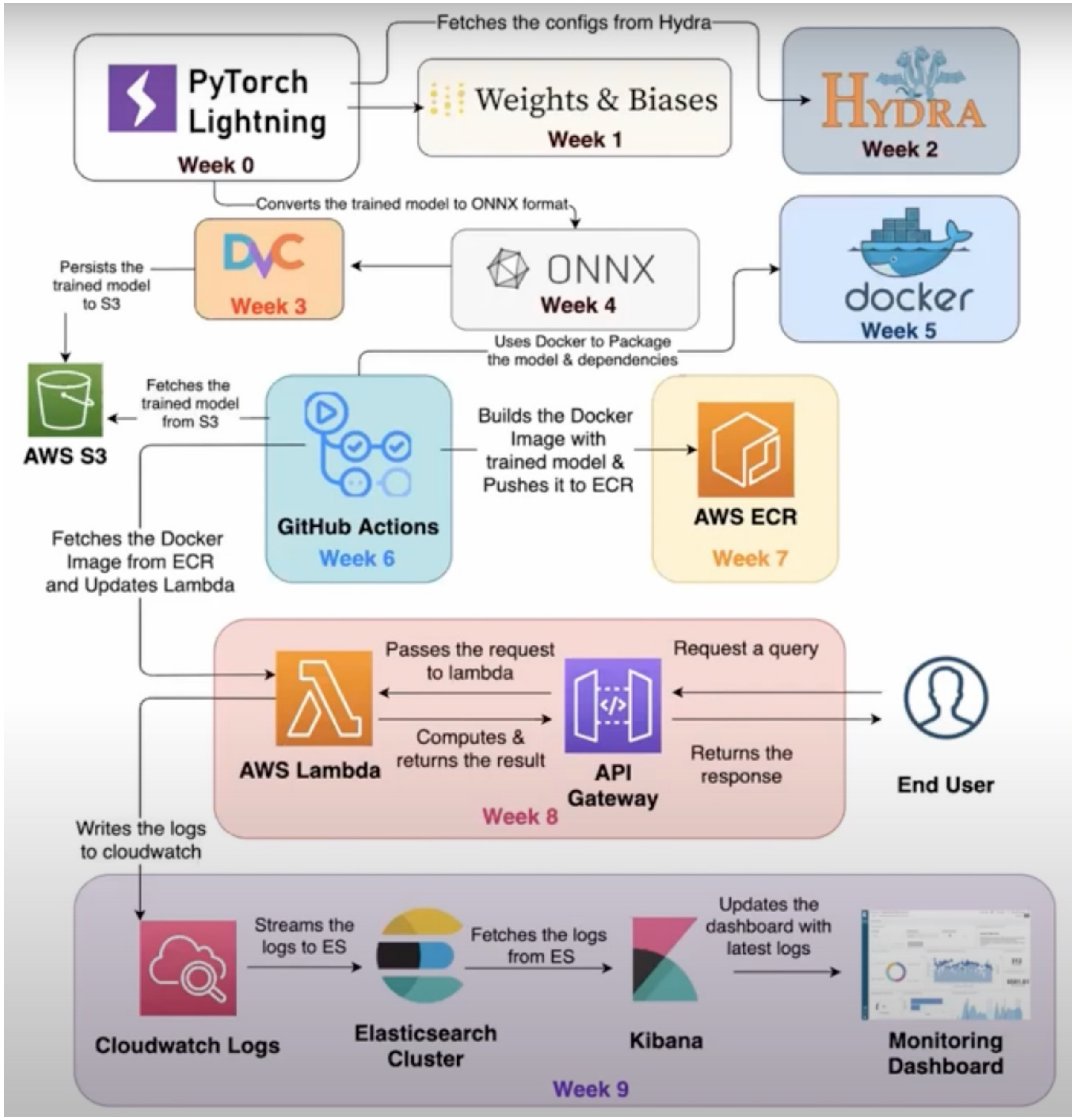
- Data-Centric Nature
- Model Complexity
- Iterative Development
- Scalability
- Monitoring and adaptation
- Testing and validation
- Explainabity and bias
- Rapid advancements
- Collaboration
- Deployment challenges
- Cost



From Research to Production

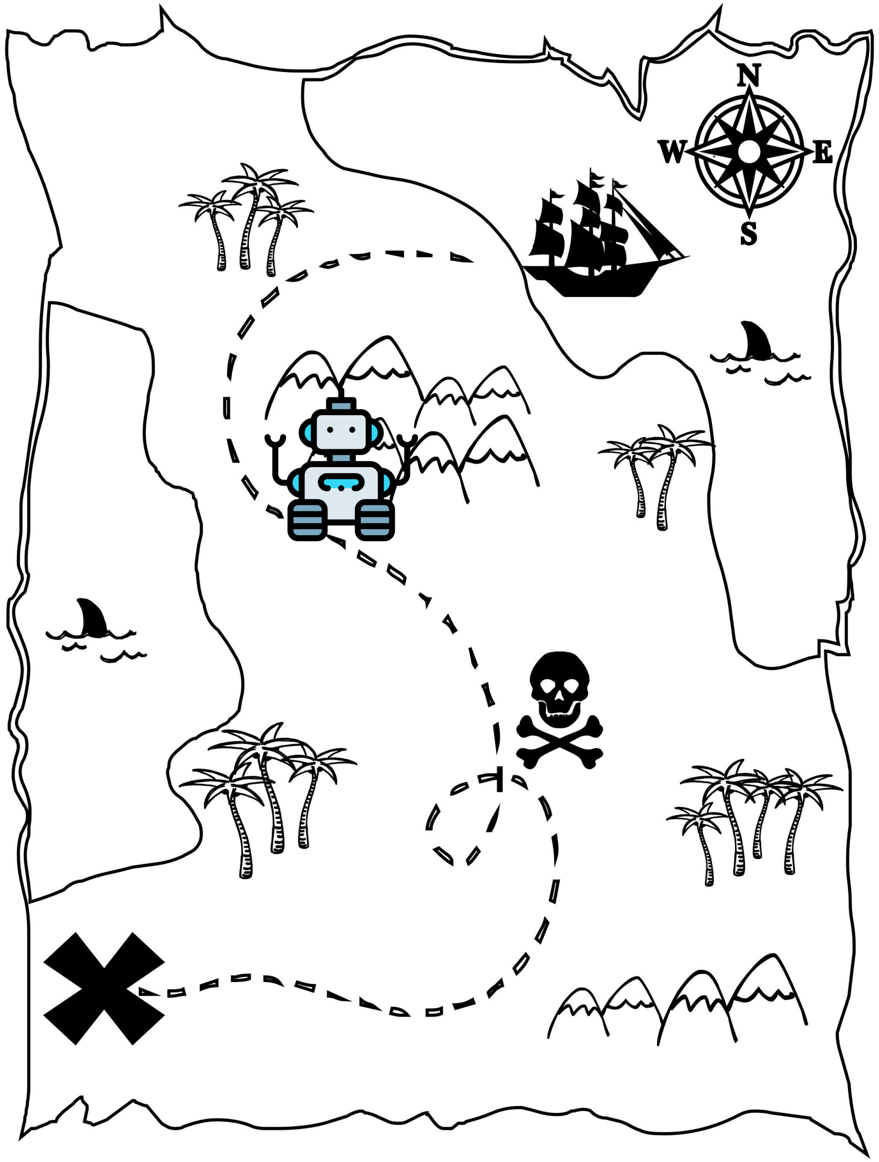


Example basic MLOps flow by Raviraja Ganta


Fetches the configs from Hydra
PyTorch Lightning
Week 0
Weights & Biases
Week 1
HYDRA
Week 2
Converts the trained model to ONNX format
Persists the trained model to S3
DVC
Week 3
ONNX
Week 4
docker
Week 5
Uses Docker to Package the model & dependencies
AWS S3
Fetches the trained model from S3
GitHub Actions
Week 6
Builds the Docker Image with trained model & Pushes it to ECR
AWS ECR
Week 7
Fetches the Docker Image from ECR and Updates Lambda
Passes the request to lambda
Request a query
AWS Lambda
Computes & returns the result
API Gateway
Returns the response
End User
Week 8
Writes the logs to cloudwatch
Streams the logs to ES
Fetches the logs from ES
Updates the dashboard with latest logs
Cloudwatch Logs
Elasticsearch Cluster
Kibana
Monitoring Dashboard
Week 9

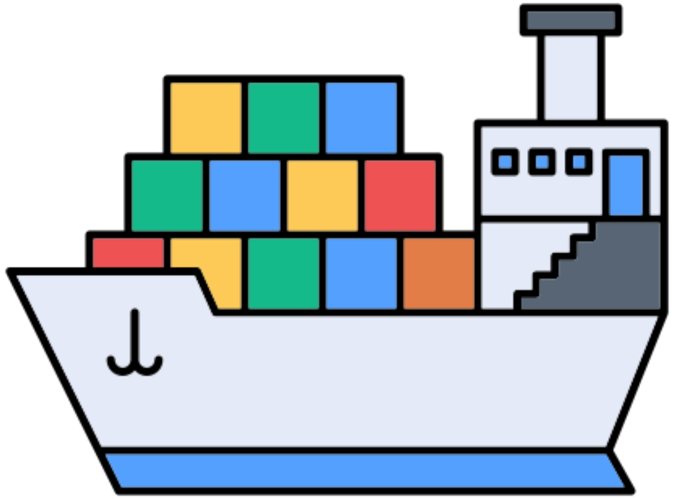
Agenda

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



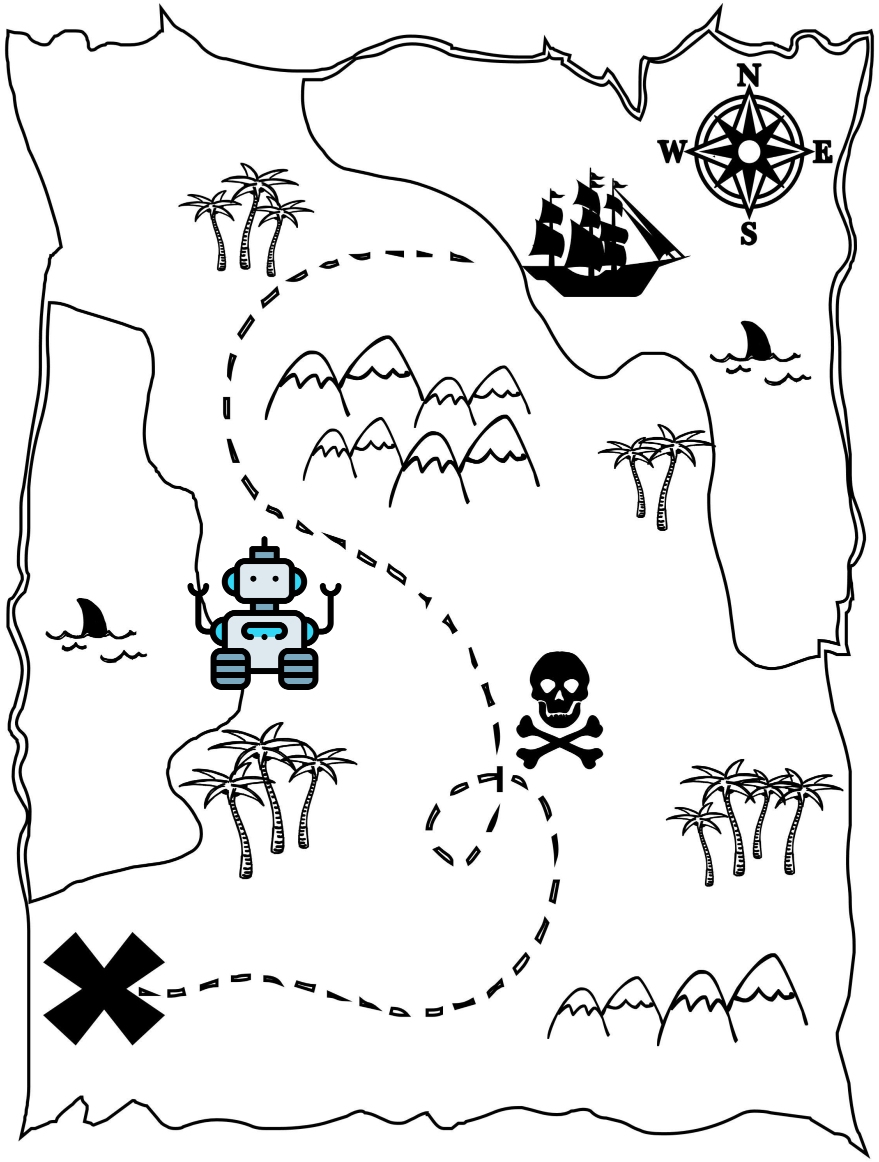
Kubernetes & MLOps

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



Agenda

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

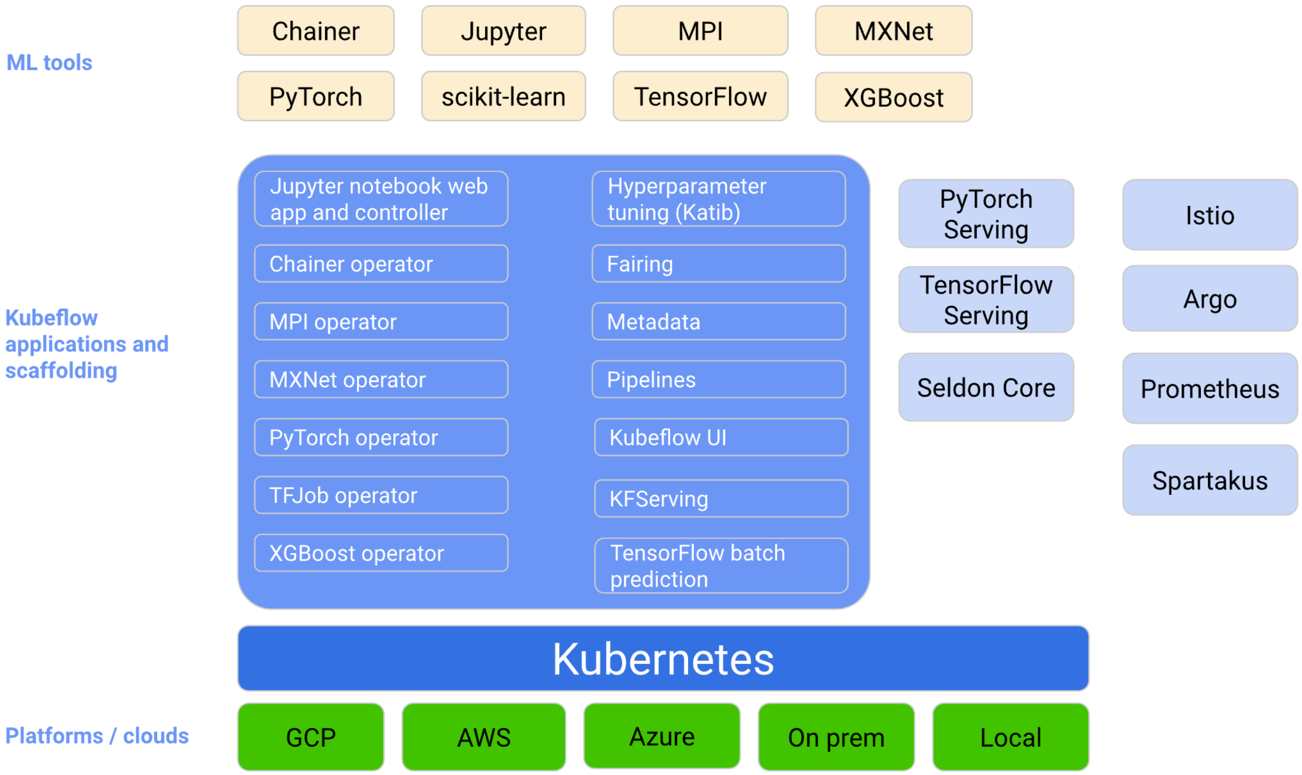


Kubeflow

The machine learning toolkit for Kubernetes.

What is Kubeflow?

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



ML tools

Chainer
Jupyter
MPI
MXNet
PyTorch
scikit-learn
TensorFlow
XGBoost

Kubeflow applications and scaffolding

Jupyter notebook web app and controller
Chainer operator
MPI operator
MXNet operator
PyTorch operator
TFJob operator
XGBoost operator
Hyperparameter tuning (Katib)
Fairing
Metadata
Pipelines
Kubeflow UI
KFServing
TensorFlow batch prediction

PyTorch Serving
TensorFlow Serving
Seldon Core

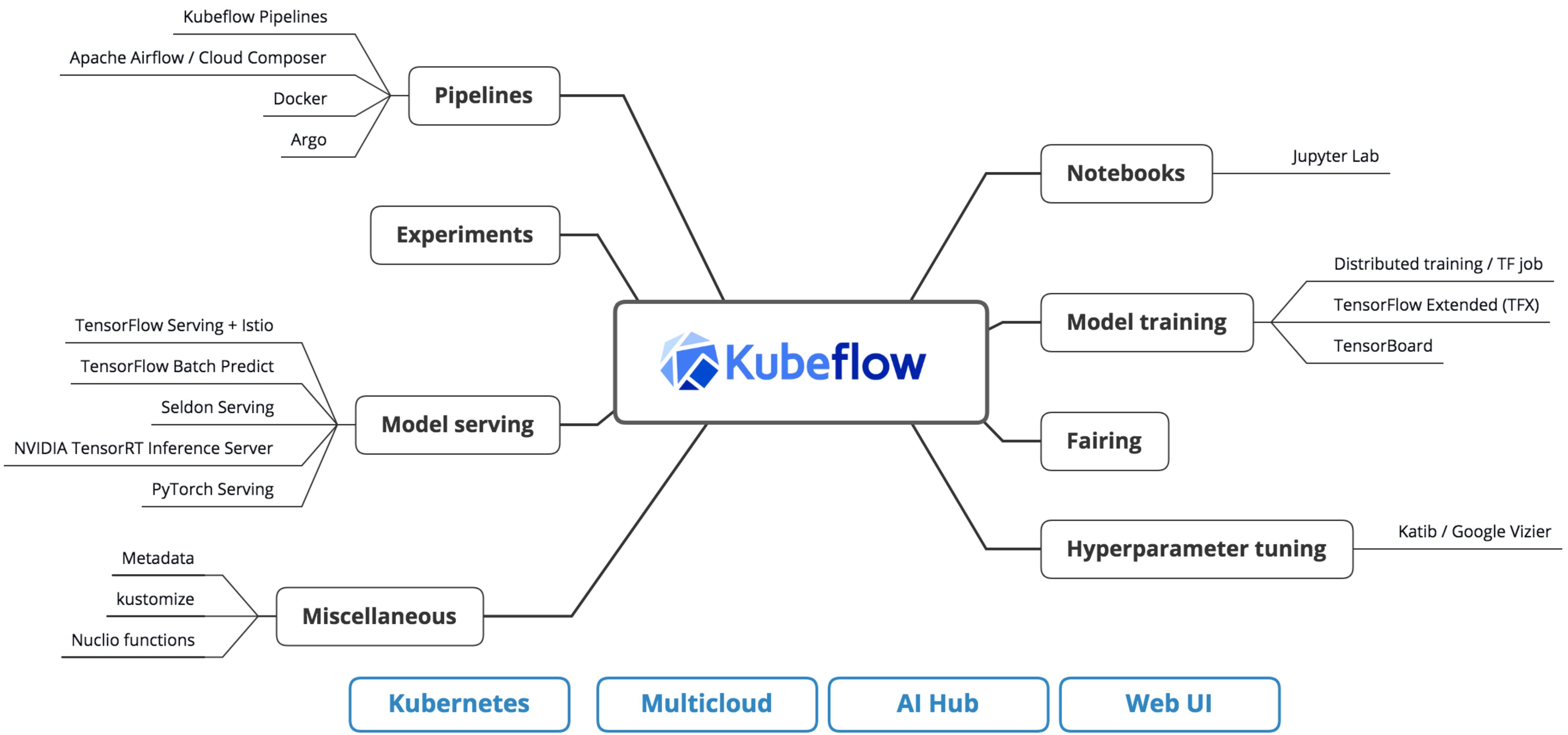
Istio
Argo
Prometheus
Spartakus

Kubernetes

Platforms / clouds

GCP
AWS
Azure
On prem
Local

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



Kubeflow

- Pipelines
  - Kubeflow Pipelines
  - Apache Airflow / Cloud Composer
  - Docker
  - Argo
- Experiments
- Model serving
  - TensorFlow Serving + Istio
  - TensorFlow Batch Predict
  - Seldon Serving
  - NVIDIA TensorRT Inference Server
  - PyTorch Serving
- Miscellaneous
  - Metadata
  - kustomize
  - Nuclio functions
- Notebooks
  - Jupyter Lab
- Model training
  - Distributed training / TF job
  - TensorFlow Extended (TFX)
  - TensorBoard
- Fairing
- Hyperparameter tuning
  - Katib / Google Vizier

Kubernetes

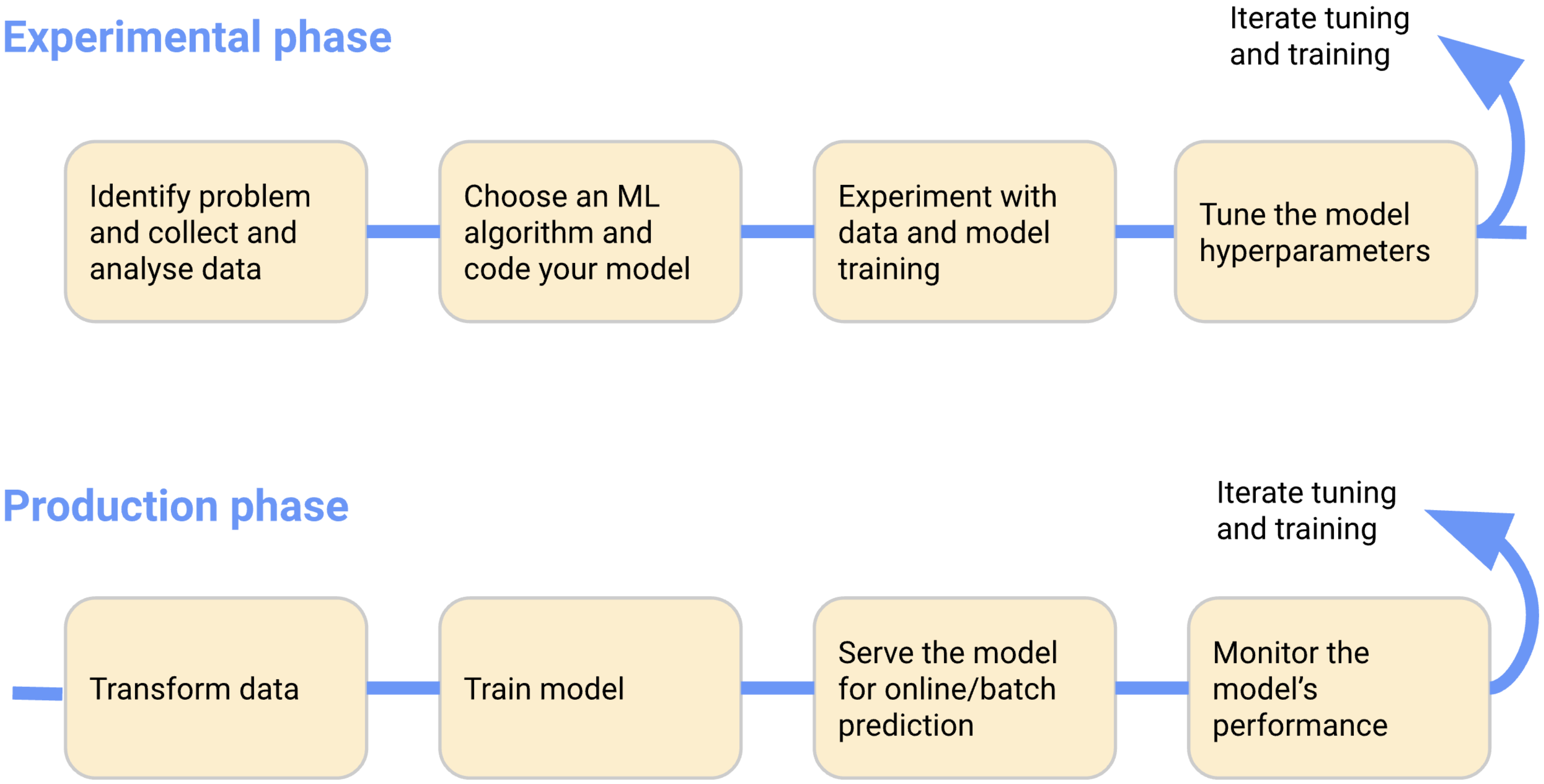
Multicloud

AI Hub

Web UI

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

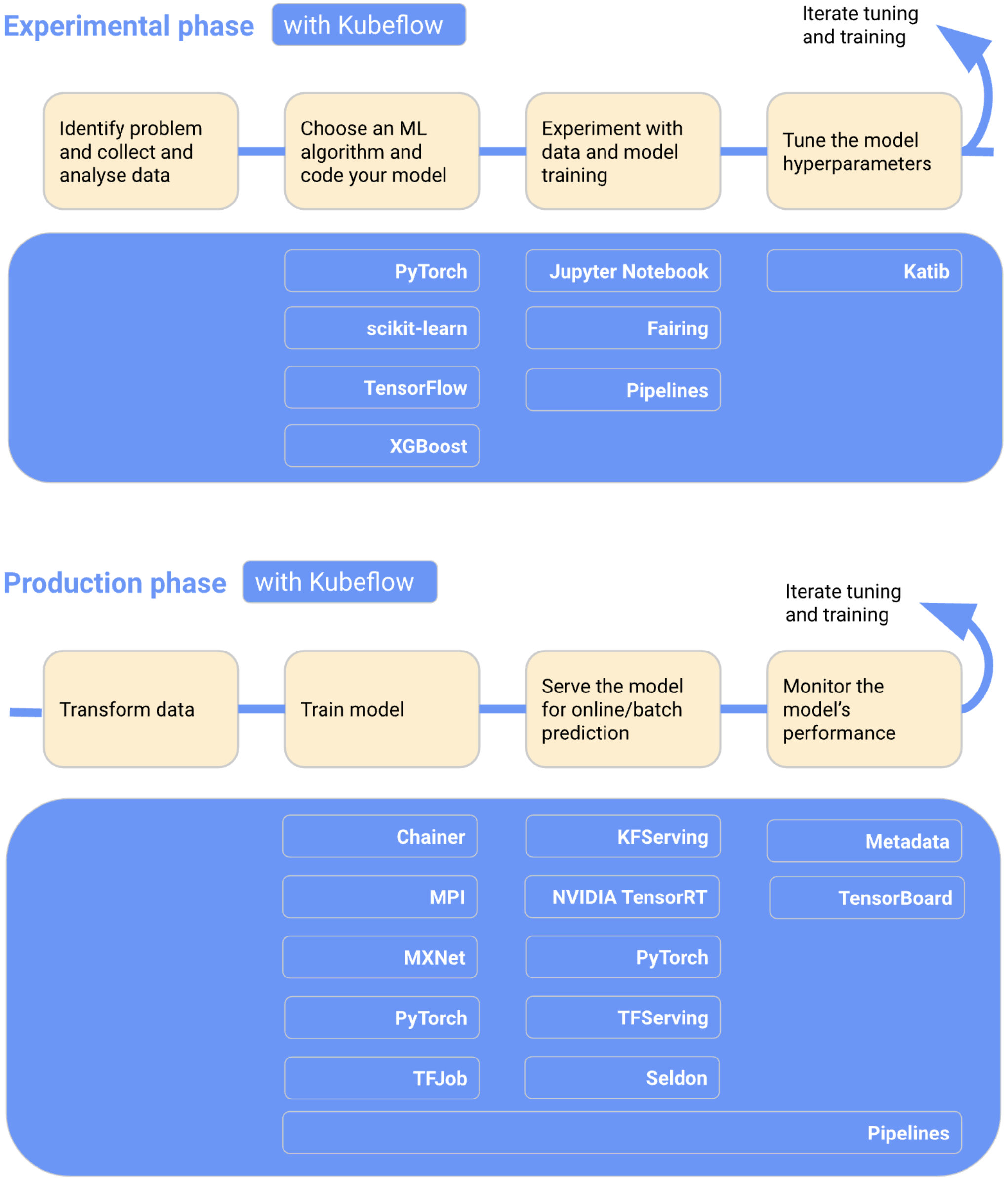
Before Kubeflow



Experimental phase
Iterate tuning and training
Identify problem and collect and analyse data
Choose an ML algorithm and code your model
Experiment with data and model training
Tune the model hyperparameters
Production phase
Iterate tuning and training
Transform data
Train model
Serve the model for online/batch prediction
Monitor the model's performance

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

With Kubeflow



Experimental phase with Kubeflow

Identify problem and collect and analyse data

Choose an ML algorithm and code your model

Experiment with data and model training

Tune the model hyperparameters

Iterate tuning and training

PyTorch

scikit-learn

TensorFlow

XGBoost

Jupyter Notebook

Fairing

Pipelines

Katib

Production phase with Kubeflow

Transform data

Train model

Serve the model for online/batch prediction

Monitor the model's performance

Iterate tuning and training

Chainer

MPI

MXNet

PyTorch

TFJob

KFServing

NVIDIA TensorRT

PyTorch

TFServing

Seldon

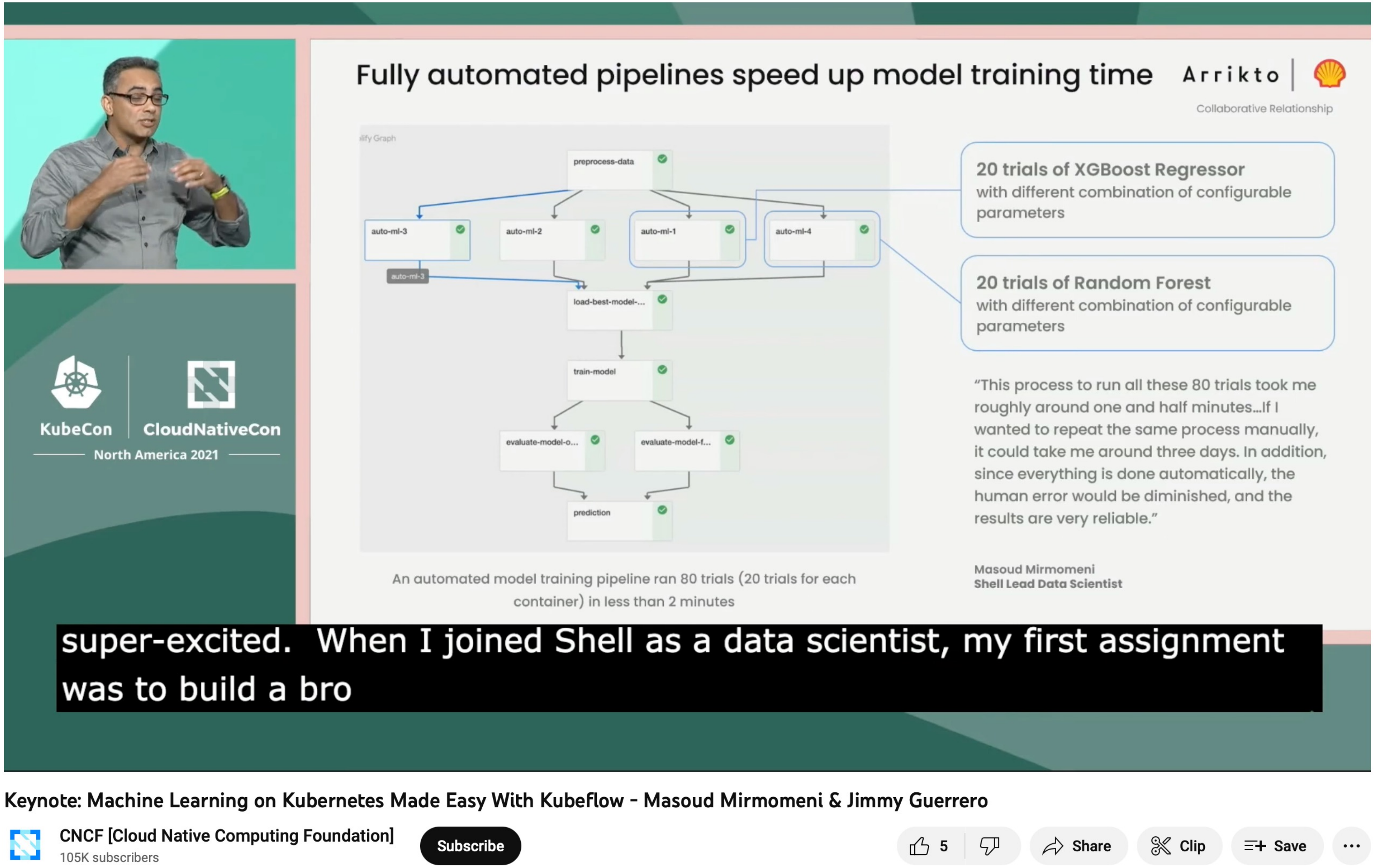
Metadata

TensorBoard

Pipelines

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

Example: Shell



Fully automated pipelines speed up model training time

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

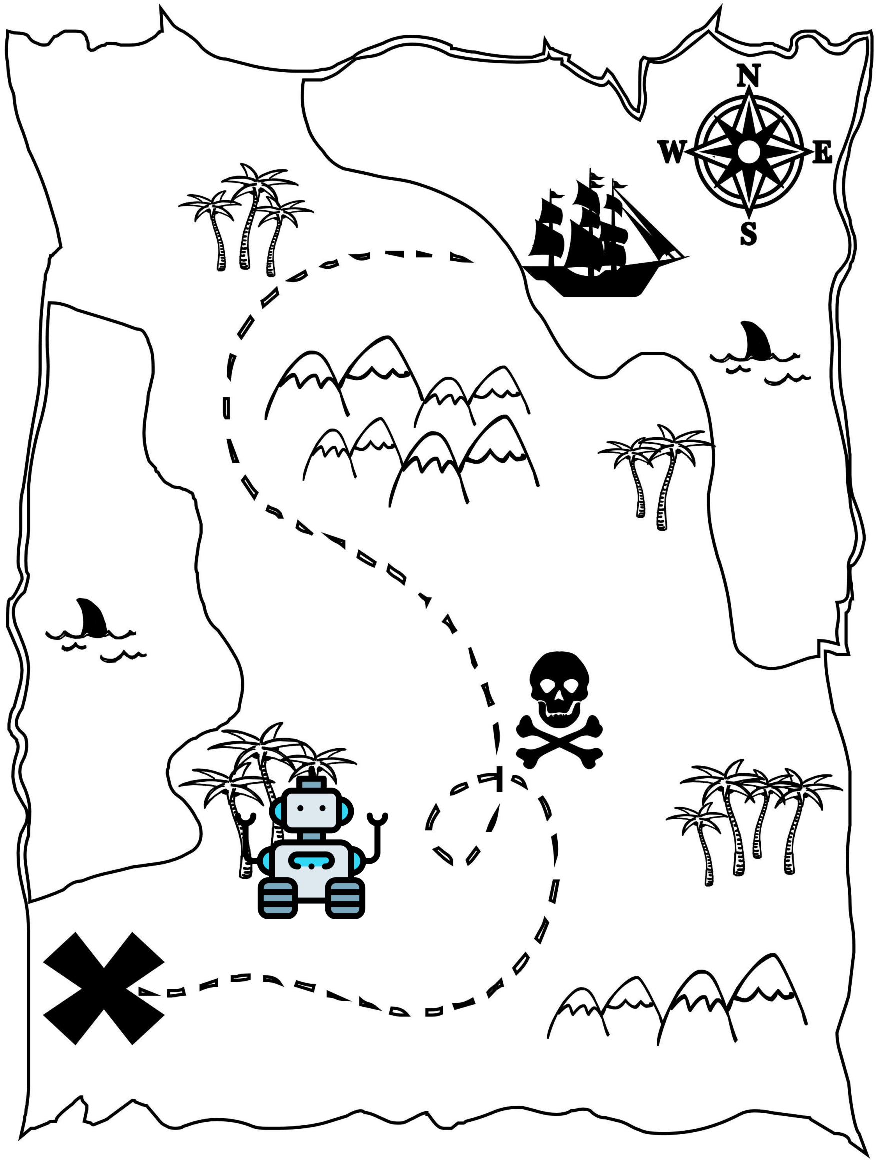
super-excited. When I joined Shell as a data scientist, my first assignment was to build a bro

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

CNCF [Cloud Native Computing Foundation]
105K subscribers

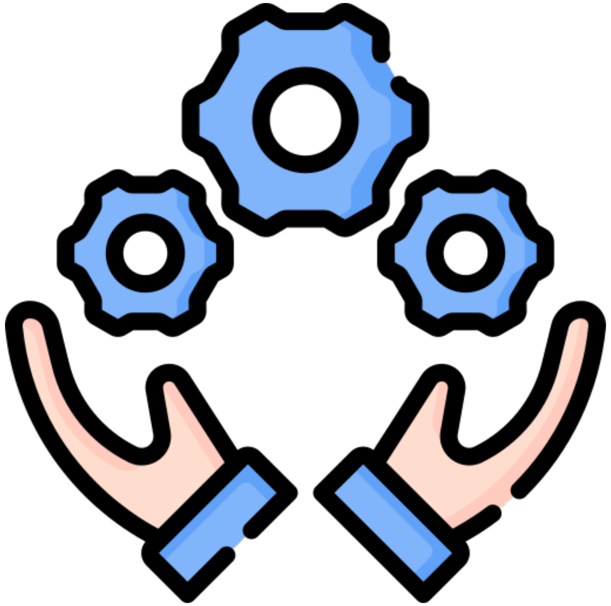
Agenda

- ~~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
- AutoML, Kaizen

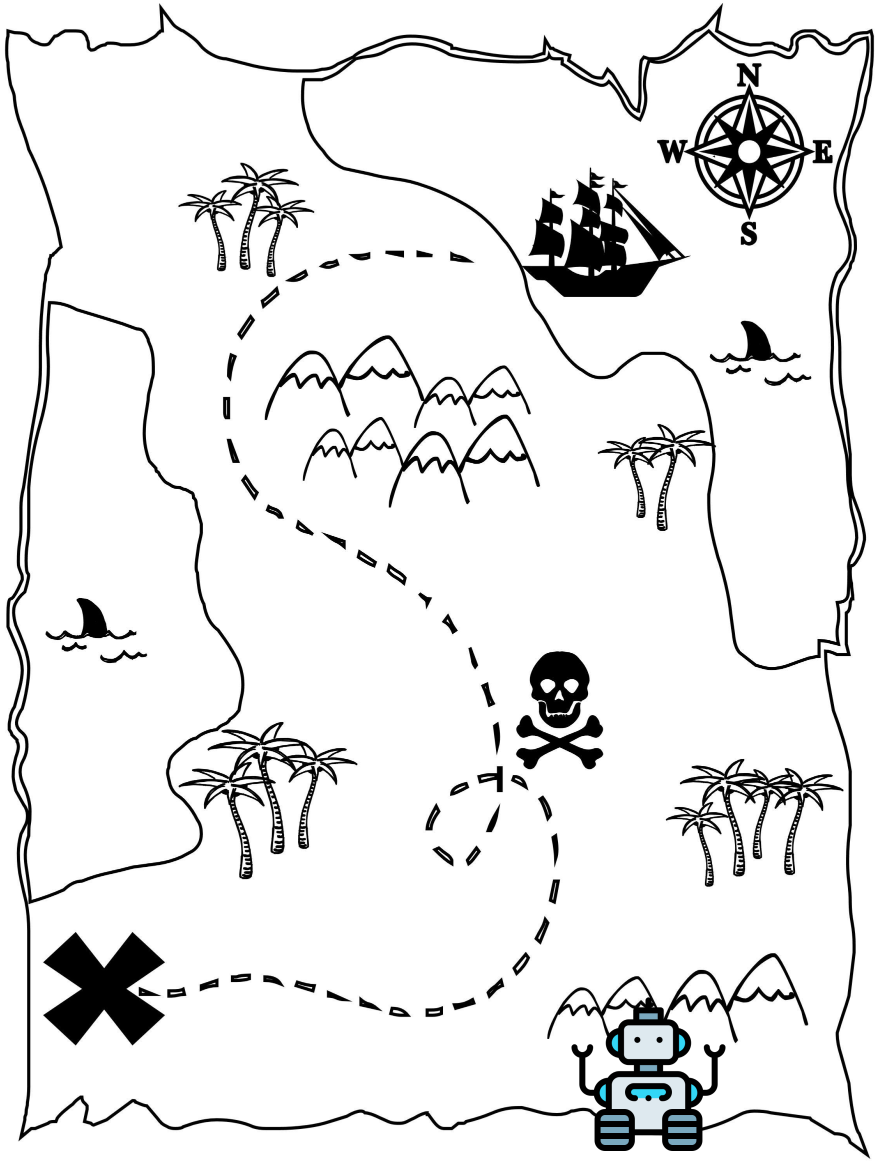


DevOps

- Automation
- CI/CD
- Infrastructure as Code
- Monitoring
- Culture, collaboration and comms
- Version Control
- Security

Agenda

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



Ethical considerations

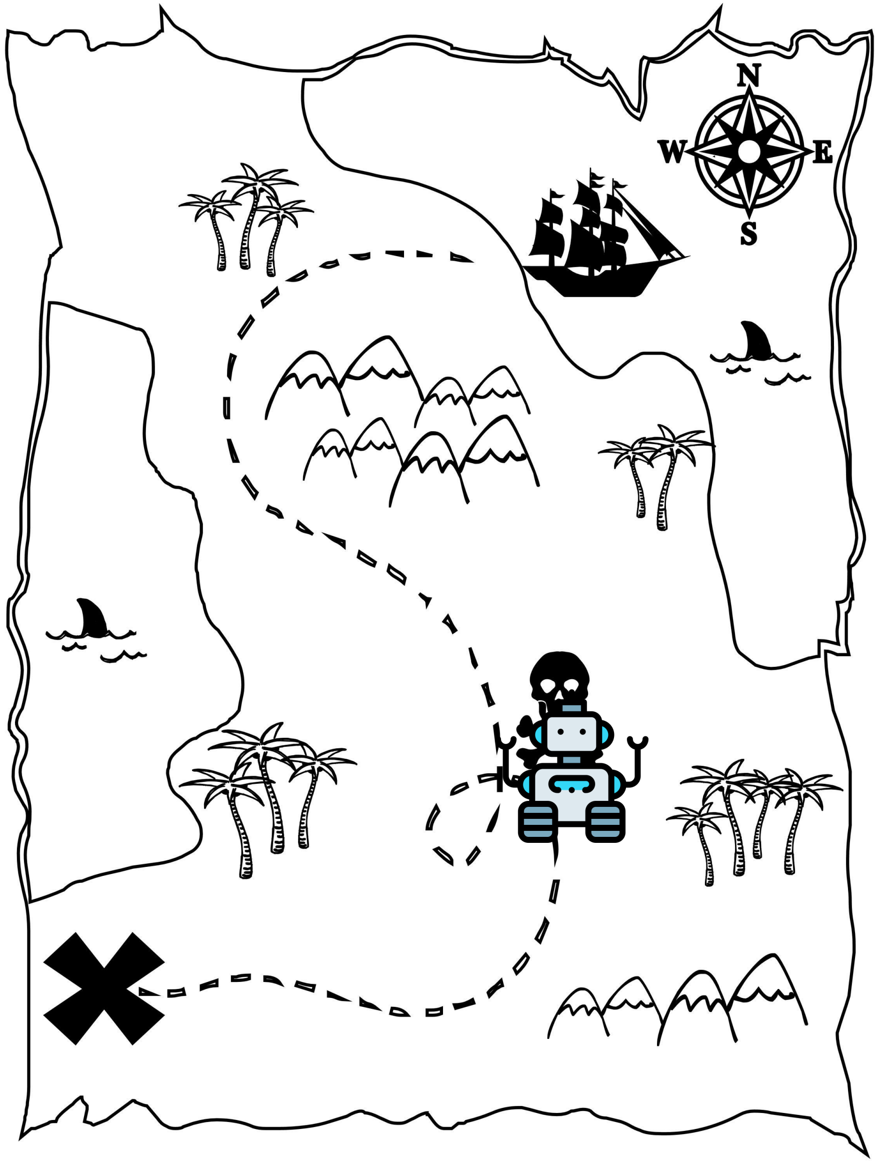
- Fairness & Bias
- Transparency
- Privacy
- Accountability
- Consent and user empowerment
- Security
- Human-Centric Design
- Data Governance
- Long-term consequences
- Regulatory Compliance

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
- End to End MLOps Basics // Raviraja Ganta // MLOps Meetup #82
  - https://youtu.be/B1t_Vb2MkRw?si=90TzpNKa-veq-ifK
- 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!