AWS re:Invent

AIM326-R

# Implementing ML workflows with Kubernetes and Amazon SageMaker

**David Ping**
Principal ML Solutions Architect
Amazon Web Services

**Aditya Bindal**
Senior Product Manager
Amazon Web Services

**Suhas Guruprasad**
ML Lead
Zalando

AWS re:Invent

# Agenda

- Machine learning is a hard problem
- Fully managed machine learning with Amazon SageMaker
- Kubernetes architecture
- Build ML workflows with Amazon SageMaker and Kubernetes
- Machine learning at Zalando
- Demo

Your organization has adopted ☸ to modernize technology

Need to transform business with machine learning

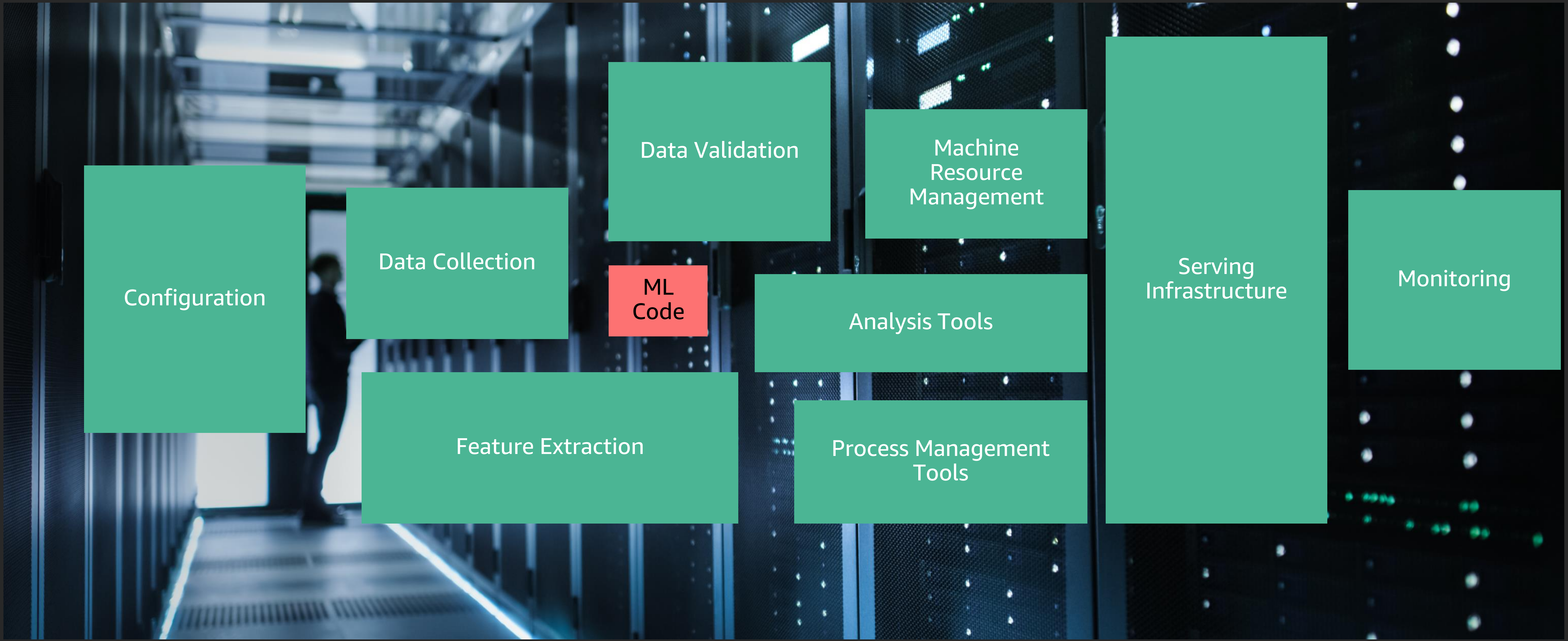
Data scientist

- Focuses on data science, business outcome and speed to market
- Wants minimum dependency on the DevOps team for experimentation and model development
- No or limited K8s and infrastructure knowledge

K8s DevOps engineer

- Wants to leverage existing K8s investment and best practices
- Wants to manage using familiar K8s construct and syntax
- Limited ML knowledge & engineering experience with ML workloads

# Machine learning is hard



Source: Sculley et al: Hidden Technical Debt in Machine Learning Systems

Do-it-yourself

+

Opportunity
for both?

Managed service

# Fully managed machine learning with Amazon SageMaker

# Amazon SageMaker is a fully managed service that covers the entire machine learning workflow

- Jupyter notebook instances
- High-performance algorithms
- Large-scale training
- Optimization
- One-click deployment
- Fully managed with auto-scaling

# Built on modern application architecture

### Containers

- Built-in algorithms for training and hosting
- Managed ML containers for training and hosting
- Bring-your-own containers for training and hosting

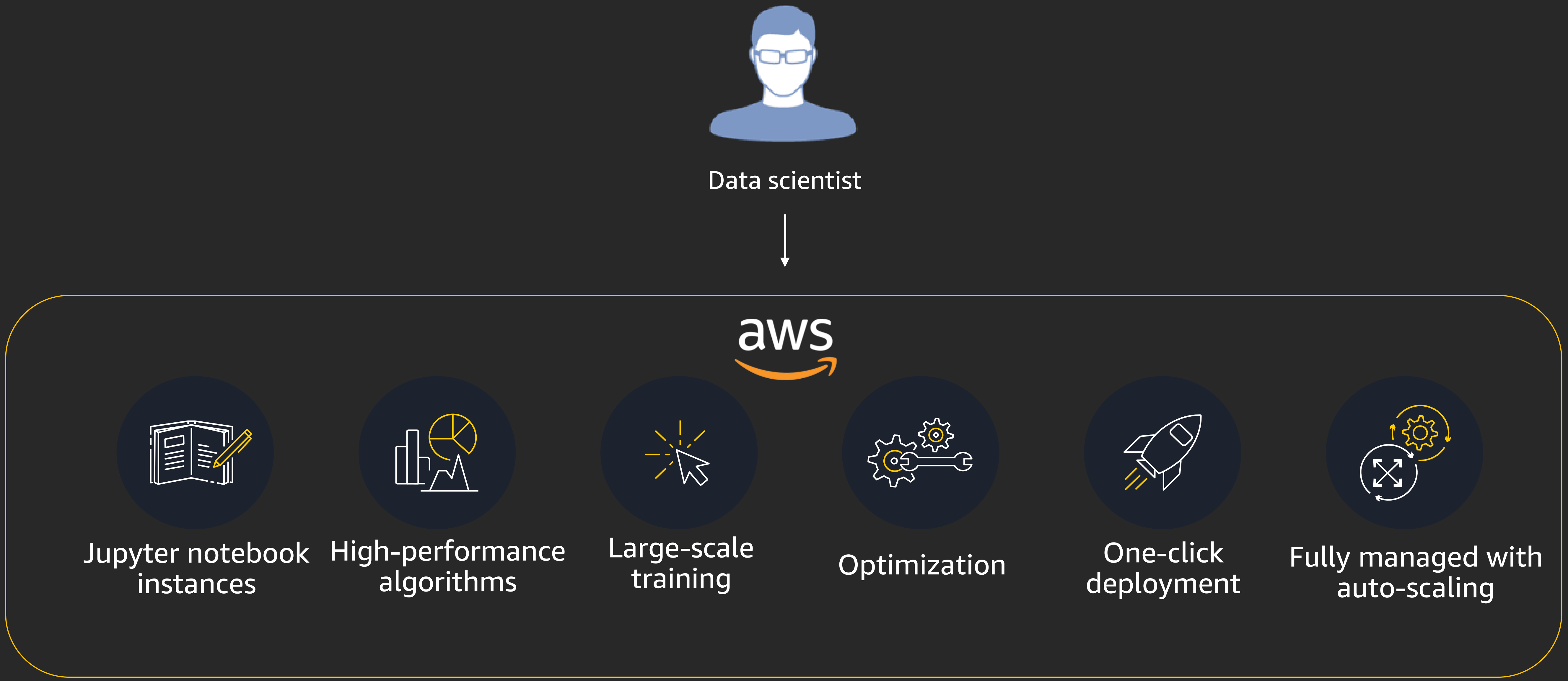
### Serverless

- No server to manage
- API driven
- Pay by usage
- Built-in monitoring and logging

### Cloud Native

- Multi-AZ deployment
- Scale with powerful compute resources
- IAM, VPC, Encryption
- Resource on-demand or SPOT, Amazon Elastic Inference, Multi-Model Endpoint

# Minimized dependency on DevOps engineering



Data scientist

Jupyter notebook instances

High-performance algorithms

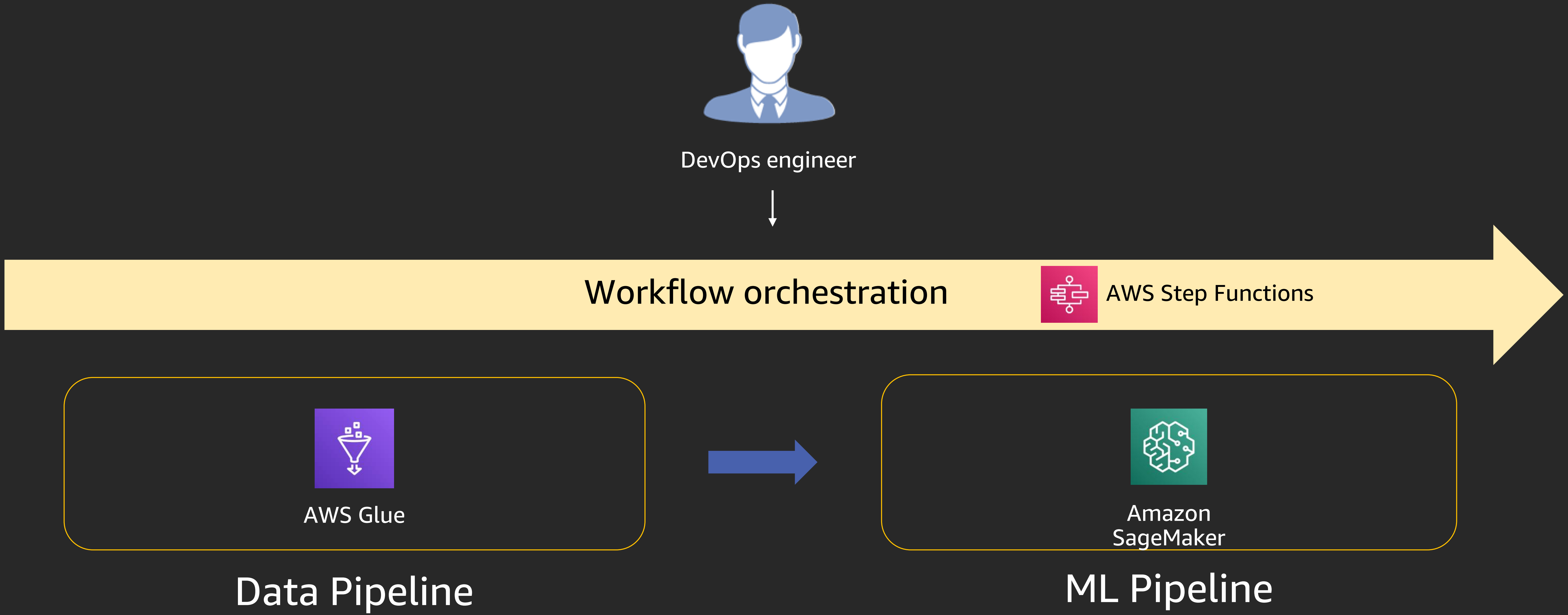
Large-scale training

Optimization

One-click deployment

Fully managed with auto-scaling

# Run Amazon SageMaker in a pipeline



# Kubernetes architecture

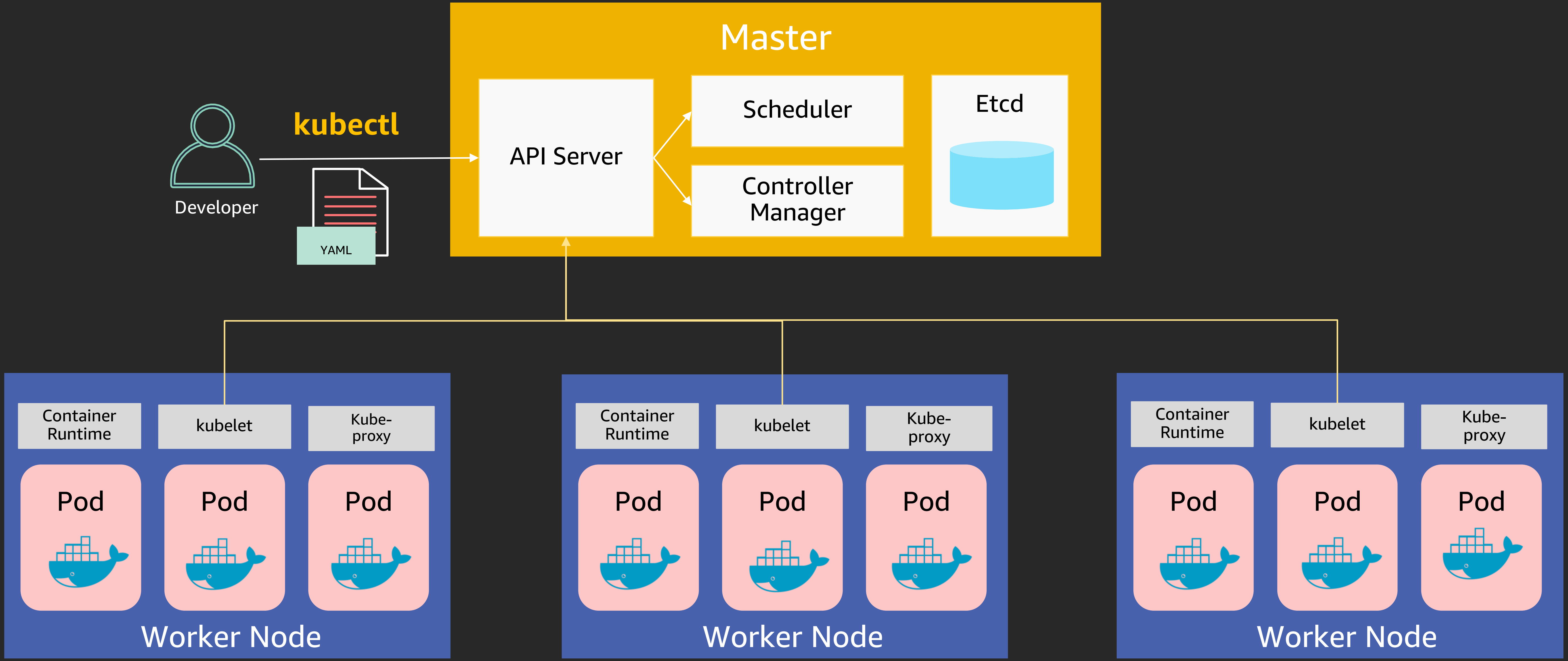
# What is Kubernetes (aka K8s)?

**Open-source container management platform**

**Helps you run containers at scale**

**Gives you primitives for building modern applications**

# Kubernetes architecture



# Building ML workflows with Kubernetes and Amazon SageMaker

AWS re:Invent

aws

# Using Kubernetes for ML is hard to manage and scale

Build and manage services within Kubernetes cluster for ML

+

Make disparate open-source libraries and frameworks work together in a secure and scalable way

+

Requires time and expertise from infrastructure, data science, and development teams

=

Need an easier way to use Kubernetes for ML

# Introducing Amazon SageMaker Operators for Kubernetes

## Kubernetes customers can now train, tune, and deploy models in Amazon SageMaker

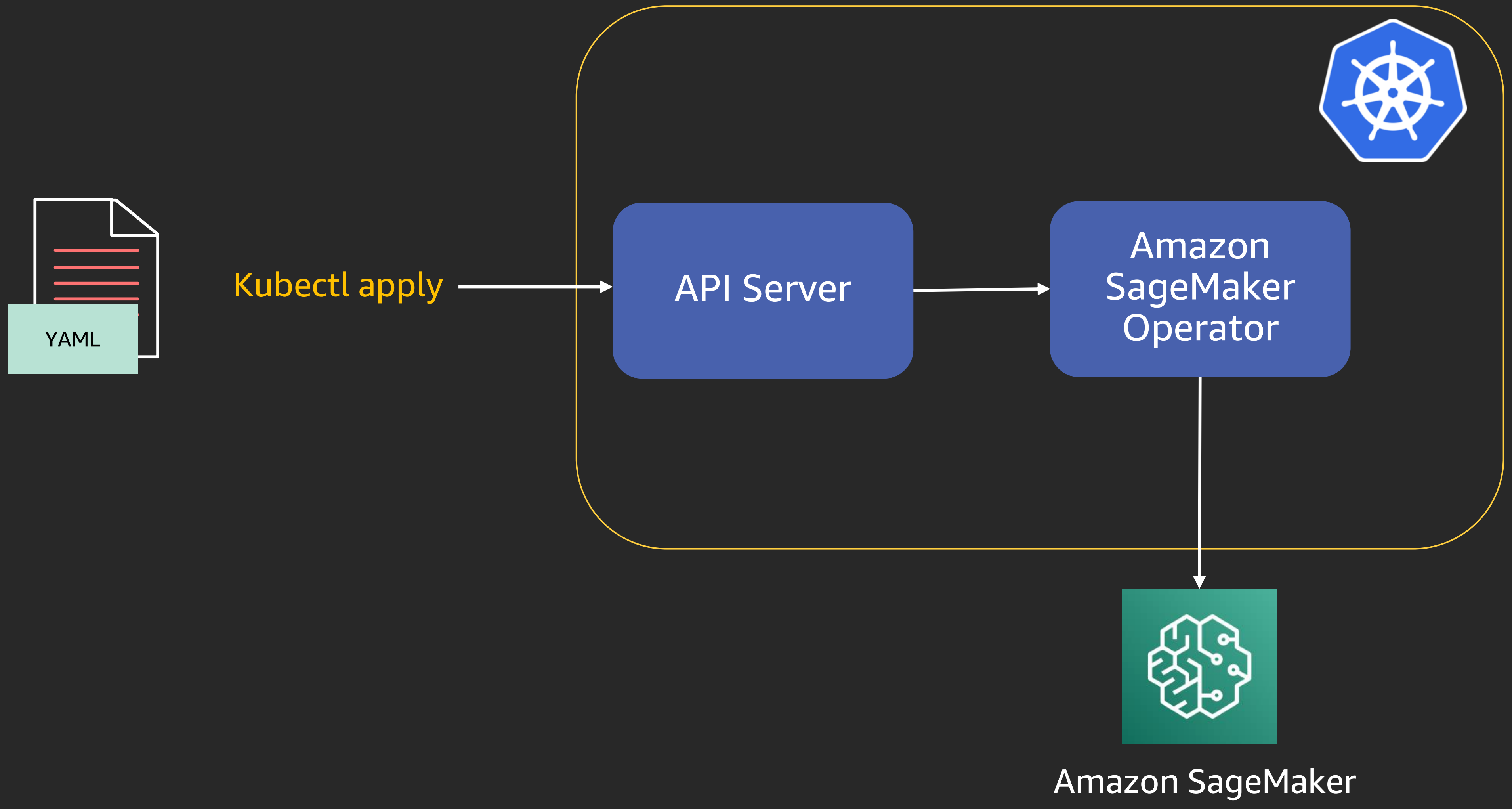
Train, tune, and deploy models in Amazon SageMaker

Orchestrate ML workloads from your Kubernetes environments

Create pipelines and workflows in Kubernetes

Fully managed infrastructure in Amazon SageMaker

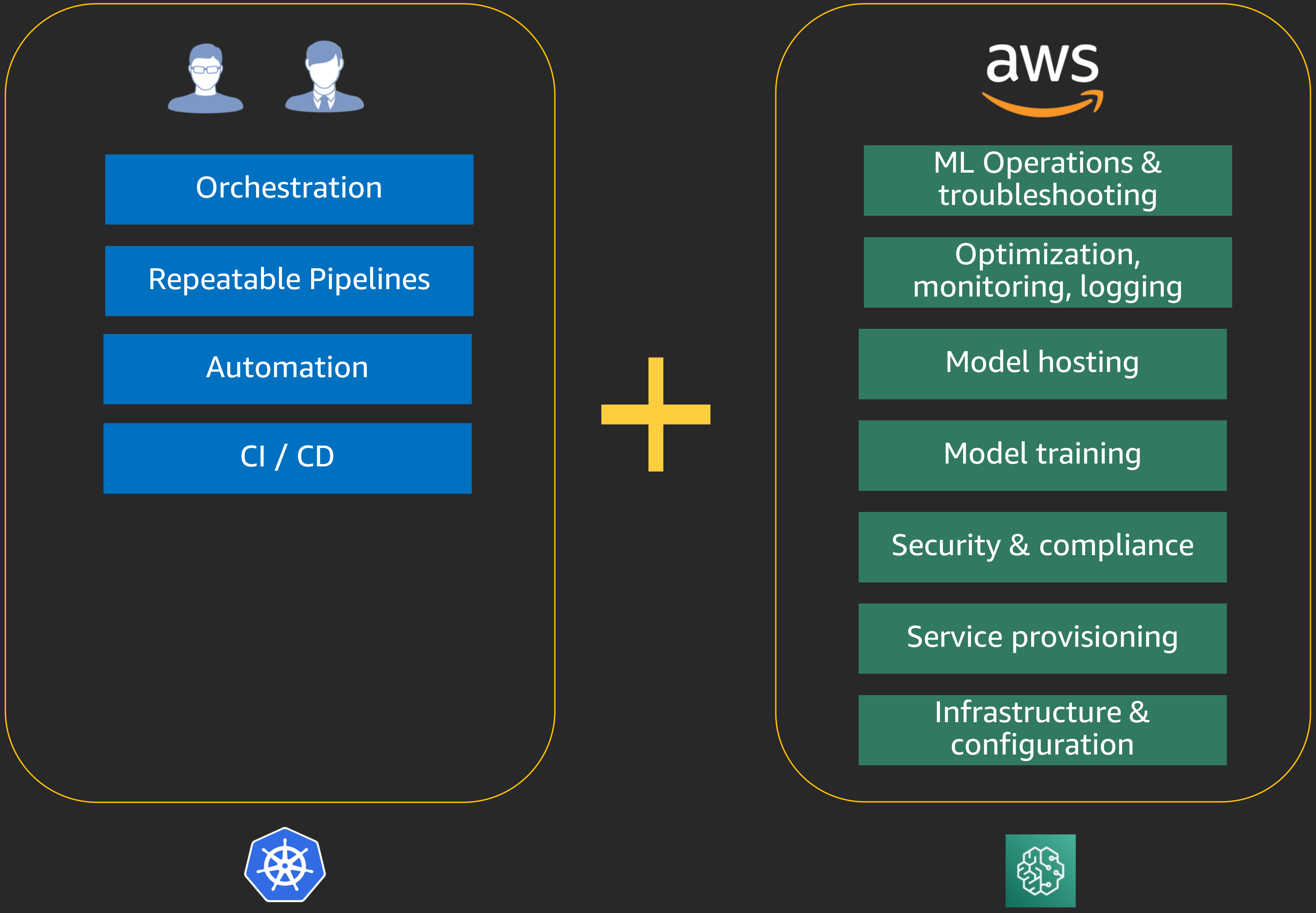
# Under the hood – Amazon SageMaker and Kubernetes



Key Features

- Amazon SageMaker Operators for training, tuning, inference
- Natively interact with Amazon SageMaker jobs using K8s tools (e.g., get pods, describe)
- Stream and view logs from Amazon SageMaker in K8s
- Helm Charts to assist with setup and spec creation

# Why together?



- Orchestration
- Repeatable Pipelines
- Automation
- CI / CD

+

- ML Operations & troubleshooting
- Optimization, monitoring, logging
- Model hosting
- Model training
- Security & compliance
- Service provisioning
- Infrastructure & configuration

# Machine learning at Zalando

# Zalando.

## The starting point for fashion.

# ZALANDO AT A GLANCE

~ 5.4 billion EUR

revenue 2018

> 300 million visits per month

> 29 million active customers

> 450,000 product choices

~ 14,000 employees in Europe

> 80% of visits via mobile devices

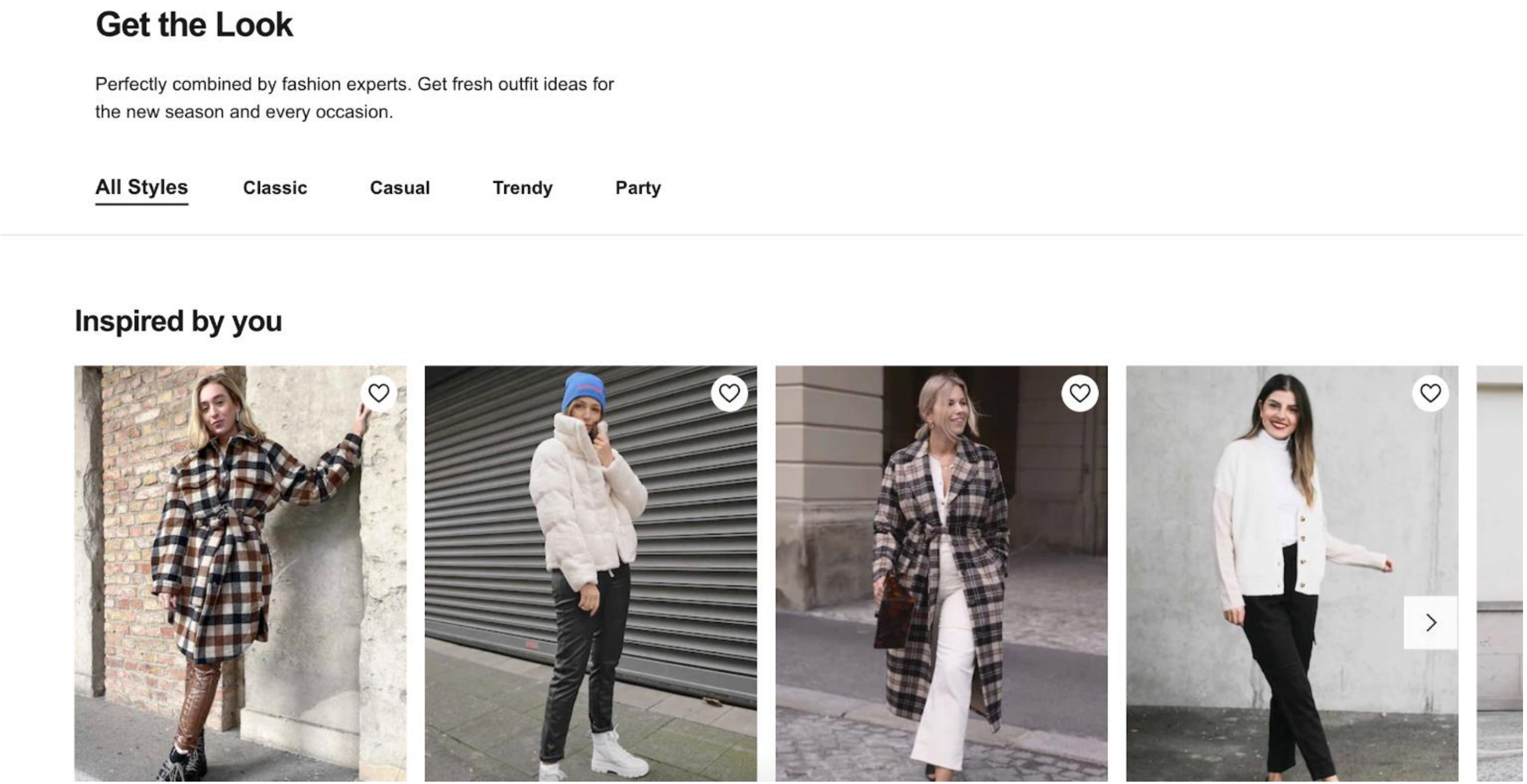
> 2,000 brands

17 countries

“A sustainable fashion platform with a net positive impact for people and planet”

# We leverage Machine Learning across the platform to deliver a better customer experience.

# MACHINE LEARNING AT ZALANDO



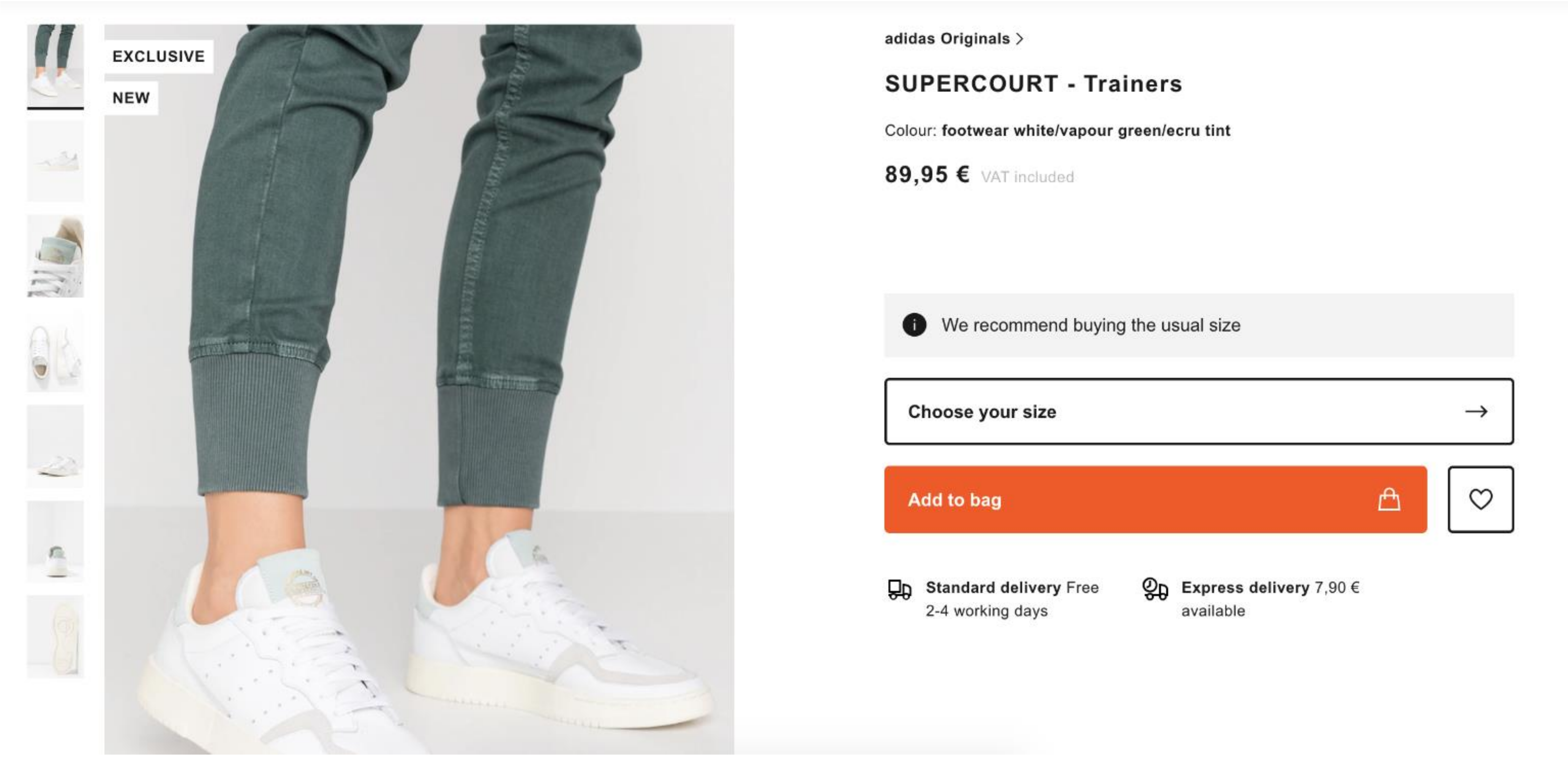
## Get the Look

Perfectly combined by fashion experts. Get fresh outfit ideas for the new season and every occasion.

**All Styles** Classic Casual Trendy Party

### Inspired by you

# MACHINE LEARNING AT ZALANDO



EXCLUSIVE

NEW

adidas Originals >

SUPERCOURT - Trainers

Colour: **footwear white/vapour green/ecru tint**

**89,95 €** VAT included

We recommend buying the usual size

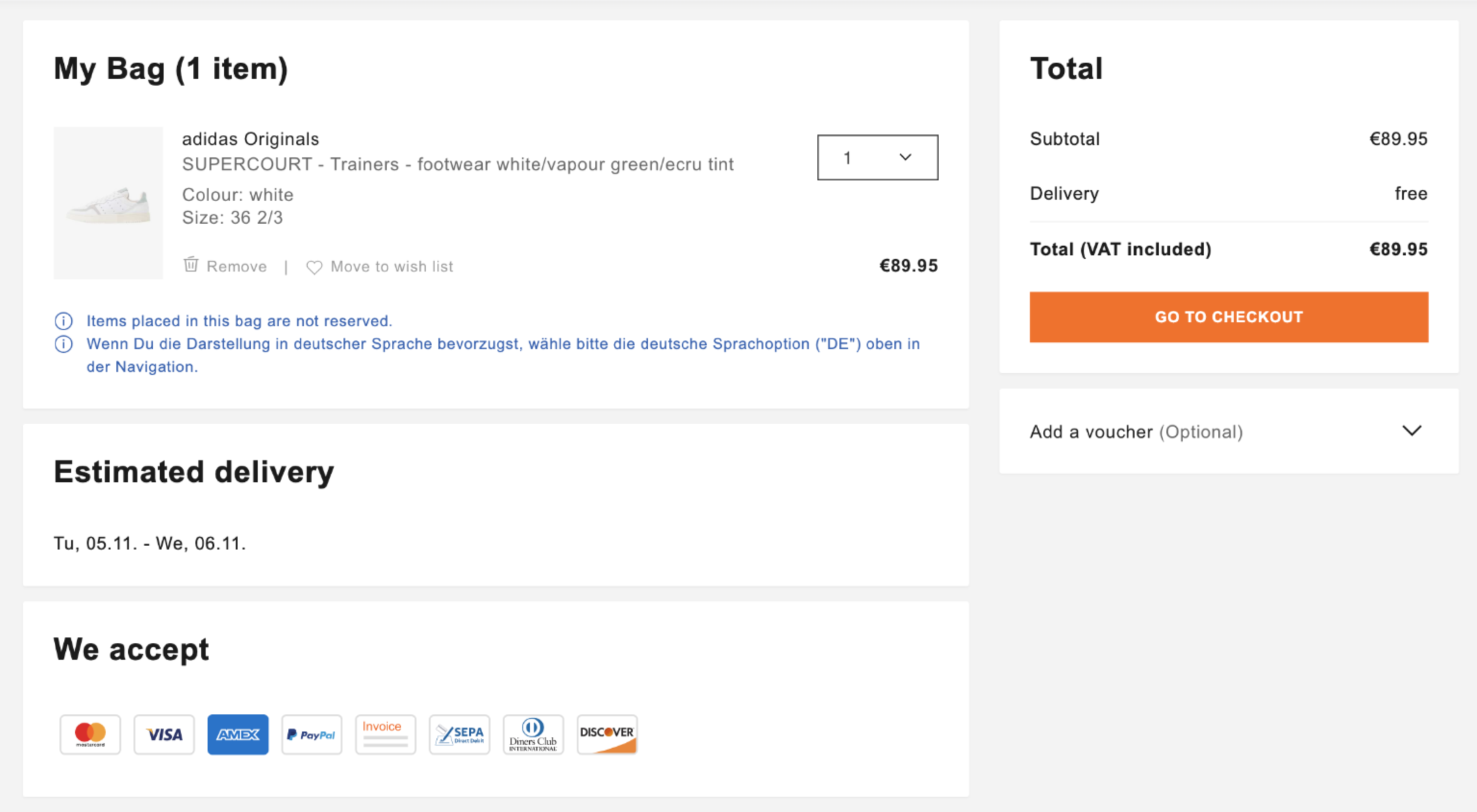
Choose your size →

Add to bag

**Standard delivery** Free
2-4 working days

**Express delivery** 7,90 €
available

# MACHINE LEARNING AT ZALANDO



## My Bag (1 item)

adidas Originals
SUPERCOURT - Trainers - footwear white/vapour green/ecru tint
Colour: white
Size: 36 2/3

1

Remove | Move to wish list

€89.95

ⓘ Items placed in this bag are not reserved.

ⓘ Wenn Du die Darstellung in deutscher Sprache bevorzugst, wähle bitte die deutsche Sprachoption ("DE") oben in der Navigation.

## Estimated delivery

Tu, 05.11. - We, 06.11.

## We accept

## Total

| | |
|---|---|
| Subtotal | €89.95 |
| Delivery | free |
| **Total (VAT included)** | **€89.95** |

GO TO CHECKOUT

Add a voucher (Optional)

# MACHINE LEARNING AT ZALANDO

# But what constitutes an ML project and how is it different from traditional software?

# MICROSERVICES AT ZALANDO

CODE

and commits to

GitHub

# MICROSERVICES AT ZALANDO

CODE

and commits to GitHub

CONTINUOUS DEPLOYMENT

builds triggered by pull requests

# MICROSERVICES AT ZALANDO

**CODE**

and commits to GitHub

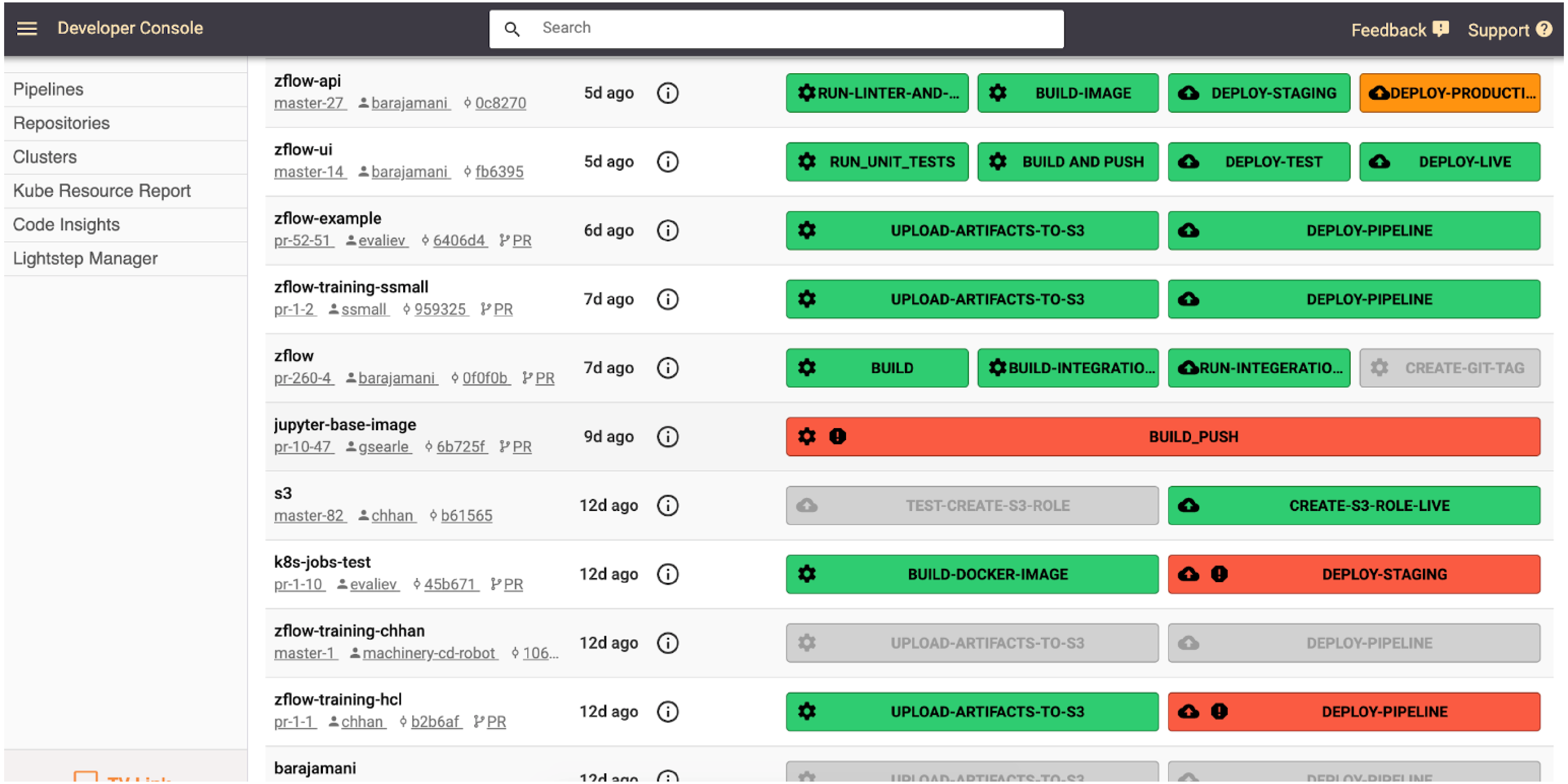
**CONTINUOUS DEPLOYMENT**

builds triggered by pull requests

**KUBERNETES**

Software runs on one of our >140 Kubernetes clusters on AWS

# CONTINUOUS DEPLOYMENT



| | | | | | | |
|---|---|---|---|---|---|---|
| zflow-api<br>master-27 barajamani 0c8270 | 5d ago | ⓘ | RUN-LINTER-AND-... | BUILD-IMAGE | DEPLOY-STAGING | DEPLOY-PRODUCTI... |
| zflow-ui<br>master-14 barajamani fb6395 | 5d ago | ⓘ | RUN_UNIT_TESTS | BUILD AND PUSH | DEPLOY-TEST | DEPLOY-LIVE |
| zflow-example<br>pr-52-51 evaliev 6406d4 PR | 6d ago | ⓘ | UPLOAD-ARTIFACTS-TO-S3 | | DEPLOY-PIPELINE | |
| zflow-training-ssmall<br>pr-1-2 ssmall 959325 PR | 7d ago | ⓘ | UPLOAD-ARTIFACTS-TO-S3 | | DEPLOY-PIPELINE | |
| zflow<br>pr-260-4 barajamani 0f0f0b PR | 7d ago | ⓘ | BUILD | BUILD-INTEGRATIO... | RUN-INTEGERATIO... | CREATE-GIT-TAG |
| jupyter-base-image<br>pr-10-47 gsearle 6b725f PR | 9d ago | ⓘ | BUILD_PUSH | | | |
| s3<br>master-82 chhan b61565 | 12d ago | ⓘ | TEST-CREATE-S3-ROLE | | CREATE-S3-ROLE-LIVE | |
| k8s-jobs-test<br>pr-1-10 evaliev 45b671 PR | 12d ago | ⓘ | BUILD-DOCKER-IMAGE | | DEPLOY-STAGING | |
| zflow-training-chhan<br>master-1 machinery-cd-robot 106... | 12d ago | ⓘ | UPLOAD-ARTIFACTS-TO-S3 | | DEPLOY-PIPELINE | |
| zflow-training-hcl<br>pr-1-1 chhan b2b6af PR | 12d ago | ⓘ | UPLOAD-ARTIFACTS-TO-S3 | | DEPLOY-PIPELINE | |
| barajamani | 12d ago | ⓘ | UPLOAD-ARTIFACTS-TO-S3 | | DEPLOY-PIPELINE | |

# THE MACHINE LEARNING JOURNEY

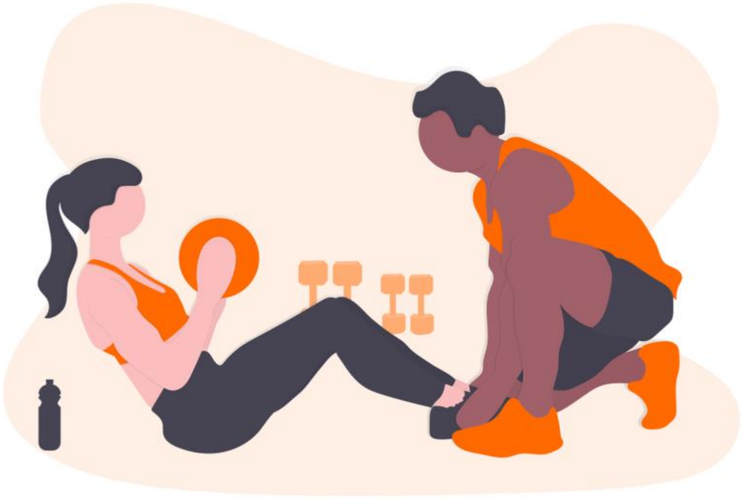
**ANALYSIS**

**Exploratory Data Analysis**

# THE MACHINE LEARNING JOURNEY

ANALYSIS

Exploratory Data Analysis



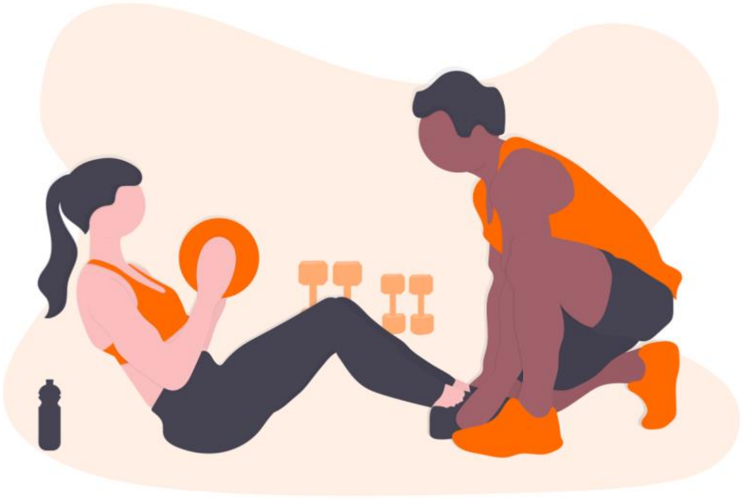
TRAINING

Features, Training & Evaluating models

# THE MACHINE LEARNING JOURNEY

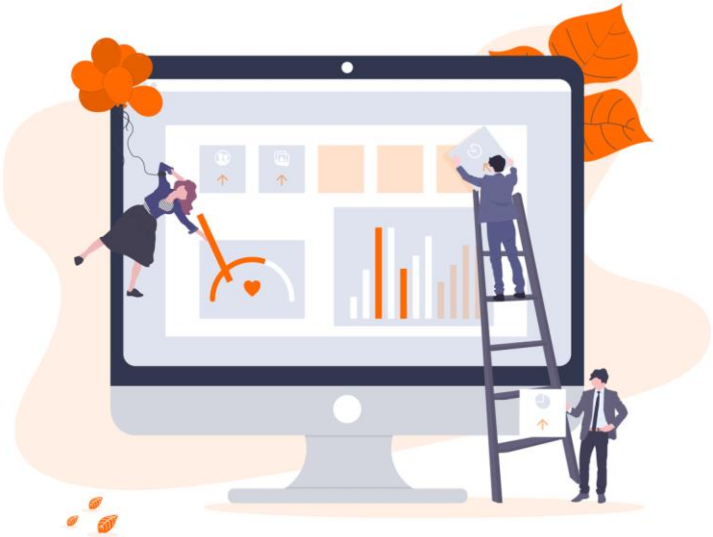
ANALYSIS

Exploratory Data Analysis



TRAINING

Features, Training & Evaluating models



OBSERVATIONS

Run A/B tests, see KPIs & adjust

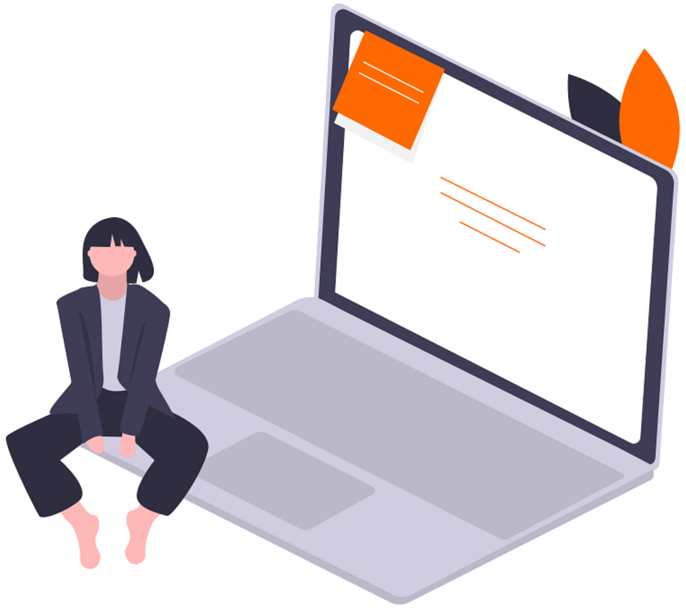
# Building a central ML platform: What are the challenges?

# CHALLENGES

**SPEED**

**Research to production time**

# CHALLENGES



**SPEED**

Research to production time

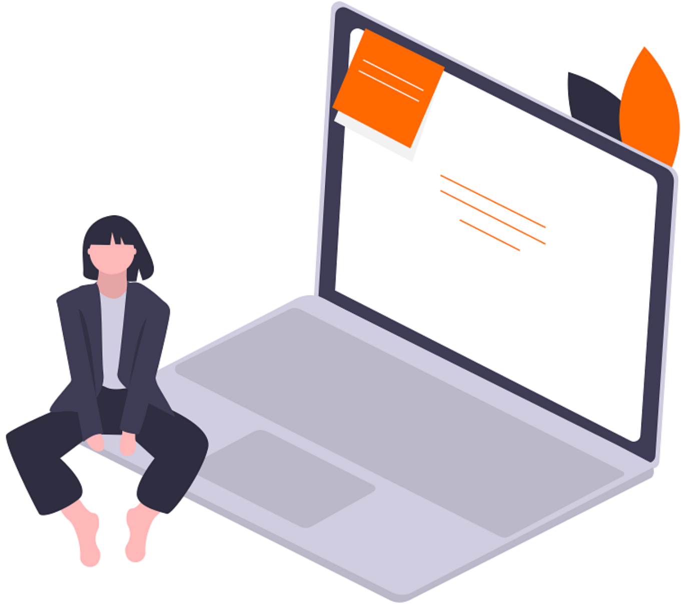
**SAFETY**

Understanding, testing, reproducibility, monitoring

# CHALLENGES

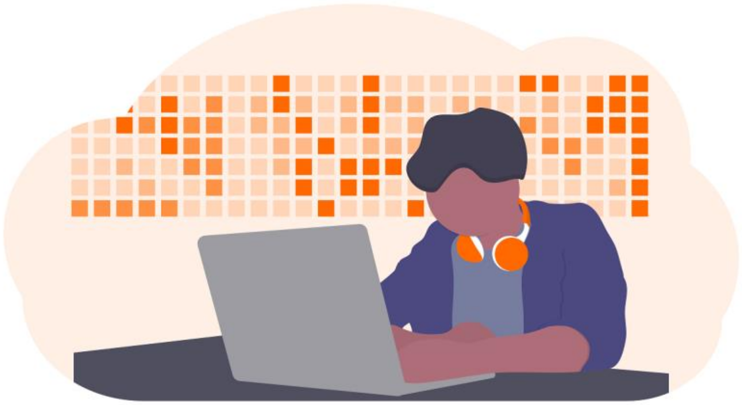
**SPEED**

Research to production time



**SAFETY**

Understanding, testing, reproducibility, monitoring



**SCALE**

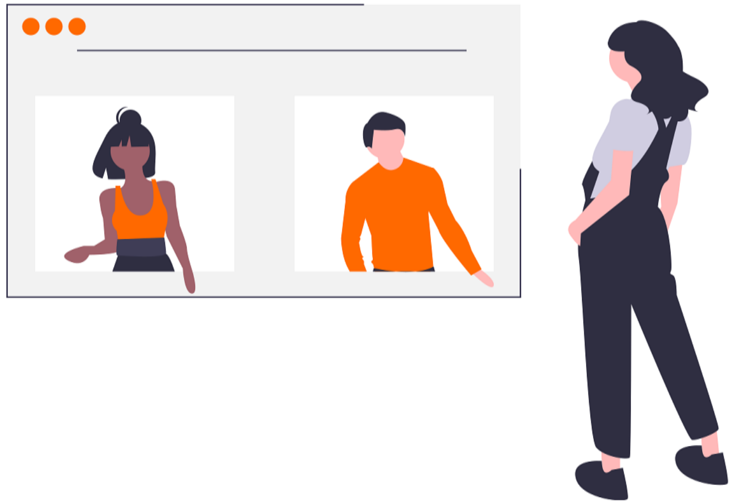
Run thousands of experiments

# CHALLENGES

**COST-EFFICIENCY**

**Experiments must be cheap at scale**

# CHALLENGES



## COST-EFFICIENCY

Experiments must be cheap at scale

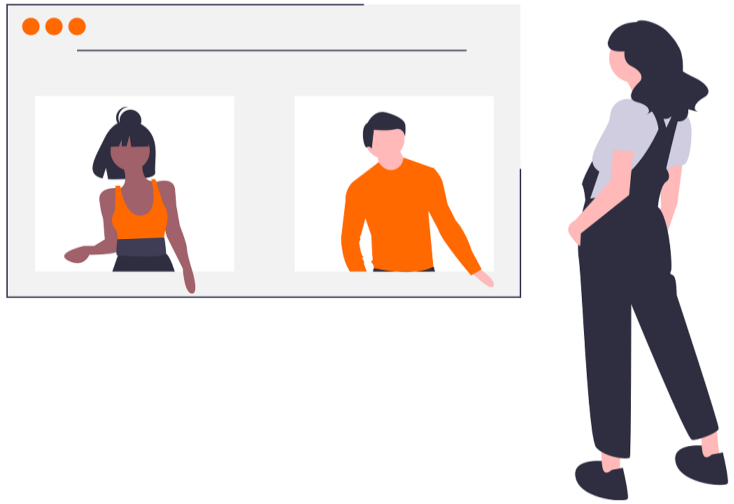
## COLLABORATION

Across multiple job families

# CHALLENGES

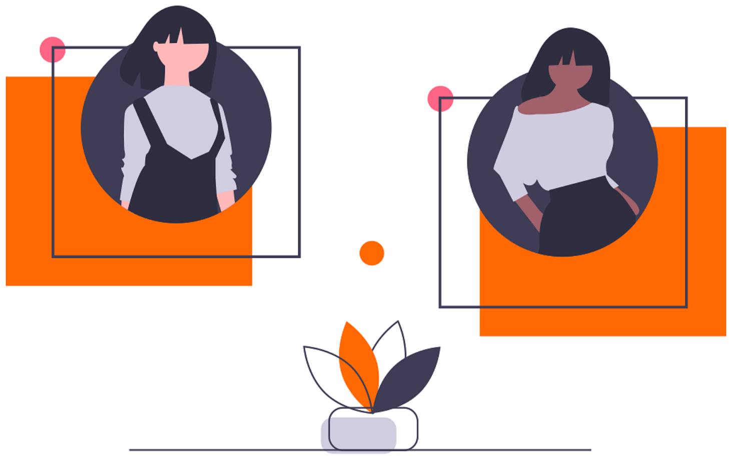
### COST-EFFICIENCY

Experiments must be cheap at scale



### COLLABORATION

Across multiple job families

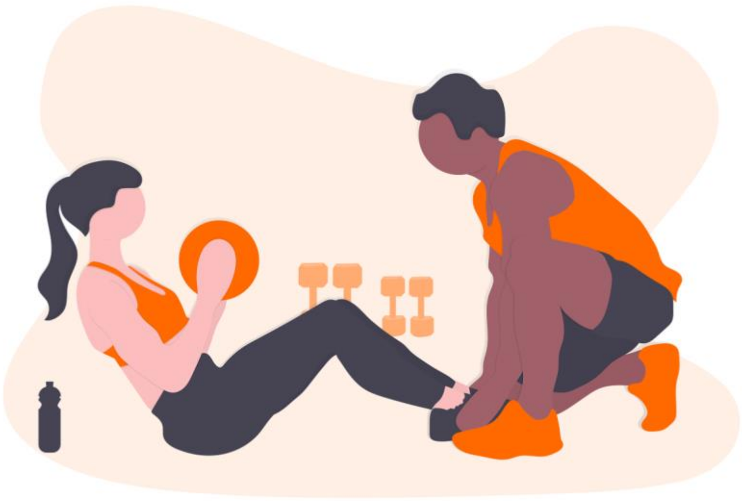


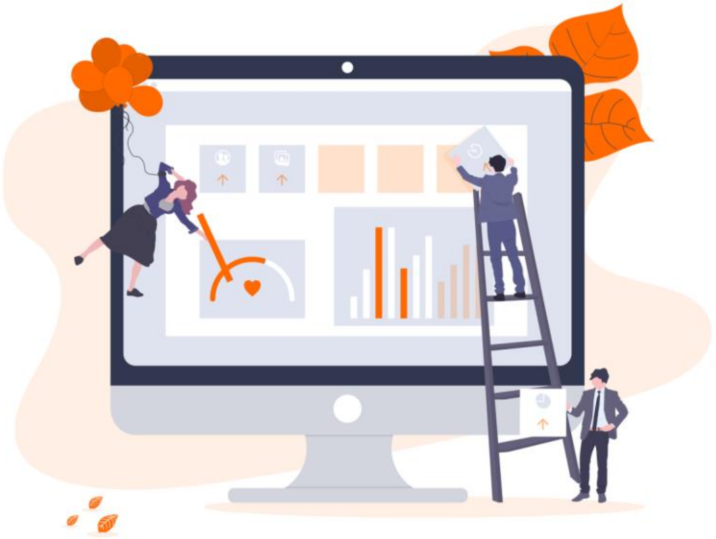
### SATISFACTION

Job satisfaction and happiness

# Relooking at the ML journey to define the ML pipeline for the ML platform

# THE MACHINE LEARNING JOURNEY





**ANALYSIS**

Exploratory Data Analysis

**TRAINING**

Features, Training & Evaluating models

**OBSERVATIONS**

Run A/B tests, see KPIs & adjust

# MACHINE LEARNING PIPELINES

**FUNDAMENTAL**

**ML should be about pipelines, not just data or models**

# Using Amazon SageMaker for ML pipelines at Zalando

# WHY AMAZON SAGEMAKER AT ZALANDO

## TRAINING JOBS

Training jobs repr.
core of a pipeline.

# WHY AMAZON SAGEMAKER AT ZALANDO

### TRAINING JOBS

Training jobs repr. core of a pipeline.

### ALGORITHMS

Built-in and bring-your-own docker images.

# WHY AMAZON SAGEMAKER AT ZALANDO

### TRAINING JOBS

Training jobs repr. core of a pipeline.

### ALGORITHMS

Built-in and bring-your-own docker images.

### INTEGRATED

With AWS offerings.

# WHY AMAZON SAGEMAKER AT ZALANDO

### TRAINING JOBS

Training jobs repr. core of a pipeline.

### ALGORITHMS

Built-in and bring-your-own docker images.

### INTEGRATED

With AWS offerings.

### “SERVERLESS”

Training instances are managed, jobs easy to distribute.

# WHY AMAZON SAGEMAKER AT ZALANDO

## TRAINING JOBS

Training jobs repr. core of a pipeline.

## ALGORITHMS

Built-in and bring-your-own docker images.

## INTEGRATED

With AWS offerings.

## “SERVERLESS”

Training instances are managed, jobs easy to distribute.

## BATCH & APIs

In the same package. Easy to reason and deploy.

# WHY AMAZON SAGEMAKER AT ZALANDO

### TRAINING JOBS

Training jobs repr. core of a pipeline.

### ALGORITHMS

Built-in and bring-your-own Docker images.

### INTEGRATED

With AWS offerings.

### “SERVERLESS”

Training instances are managed, jobs are easy to distribute.

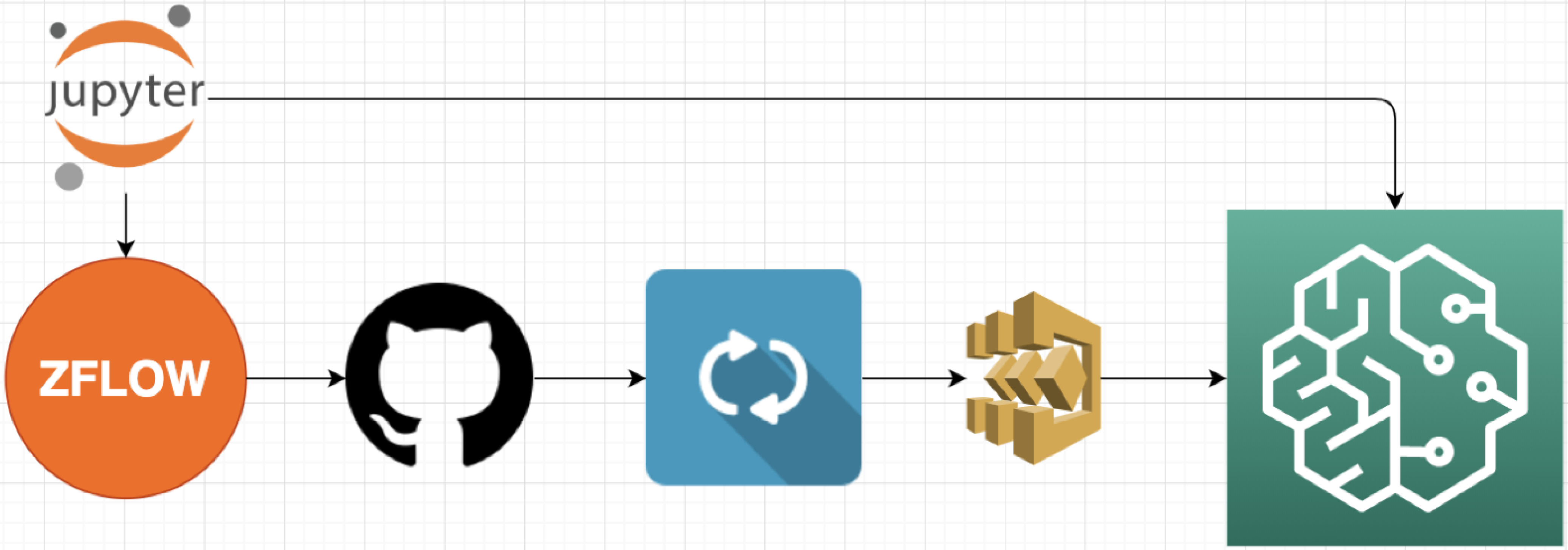
### BATCH & APIs

In the same package. Easy to reason and deploy.

### SCRIPT MODE

Easy to write quick TensorFlow scripts and tie in to pipelines

# ZALANDO MACHINE LEARNING PLATFORM



### ZFLOW

Zalando's pythonic interface to ML pipelines

### STEP FUNCTIONS

To orchestrate pipelines

### SAGEMAKER

The Core Engine

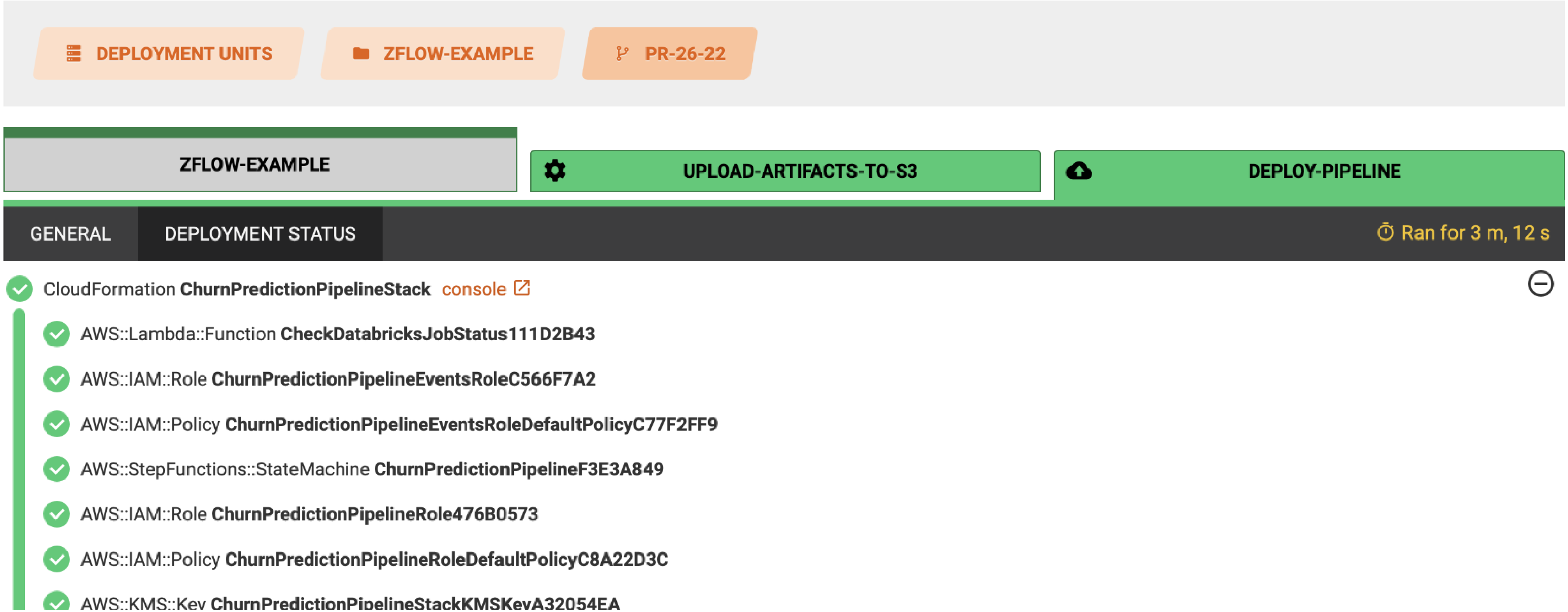
# ZALANDO MACHINE LEARNING PLATFORM

```
training_stage = training_job(
    name=”...”,
    input_data_configs=”...”,
    hyperparameters=”...”
)

batch_transform_stage =
batch_transform_job(
    name=”...”,
    s3_input_path=”...”,
    s3_output_path=”...”
)
```

# ZALANDO MACHINE LEARNING PLATFORM

```
pipeline = PipelineBuilder(“my_pipeline”)

pipeline
    .add_stage(data_processing_stage)
    .add_stage(training_stage)
    .add_stage(batch_transform_stage)
```

# ZALANDO MACHINE LEARNING PLATFORM



✓ CloudFormation **ChurnPredictionPipelineStack** console

✓ AWS::Lambda::Function **CheckDatabricksJobStatus111D2B43**

✓ AWS::IAM::Role **ChurnPredictionPipelineEventsRoleC566F7A2**

✓ AWS::IAM::Policy **ChurnPredictionPipelineEventsRoleDefaultPolicyC77F2FF9**

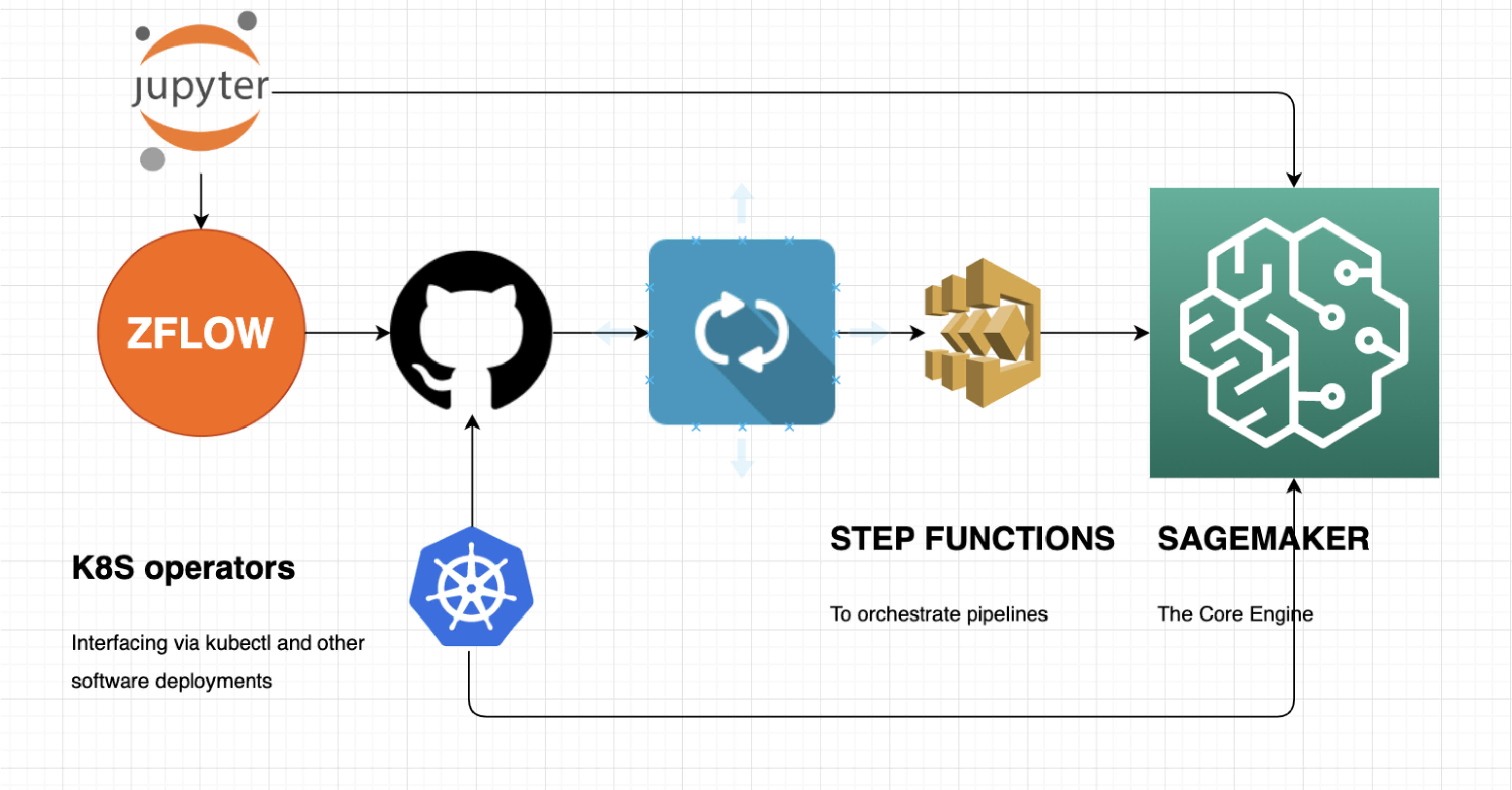
✓ AWS::StepFunctions::StateMachine **ChurnPredictionPipelineF3E3A849**

✓ AWS::IAM::Role **ChurnPredictionPipelineRole476B0573**

✓ AWS::IAM::Policy **ChurnPredictionPipelineRoleDefaultPolicyC8A22D3C**

✓ AWS::KMS::Key **ChurnPredictionPipelineStackKMSKeyA32054EA**

# UPCOMING



**K8S operators**

Interfacing via kubectl and other software deployments

**STEP FUNCTIONS**

To orchestrate pipelines

**SAGEMAKER**

The Core Engine

# ZALANDO MACHINE LEARNING PLATFORM

```
apiVersion: sagemaker.aws.amazon.com/v1
kind: TrainingJob
metadata
    name: kmeans-mnist
spec
    algorithmSpecification
        trainingImage: .../kmeans:1
        trainingInputMode: File
        hyperParameters: ...
        inputDataConfig:
    resourceConfig
        instanceType: ml.c4.8xlarge
    ...
```

# OVERCOMING THE CHALLENGES

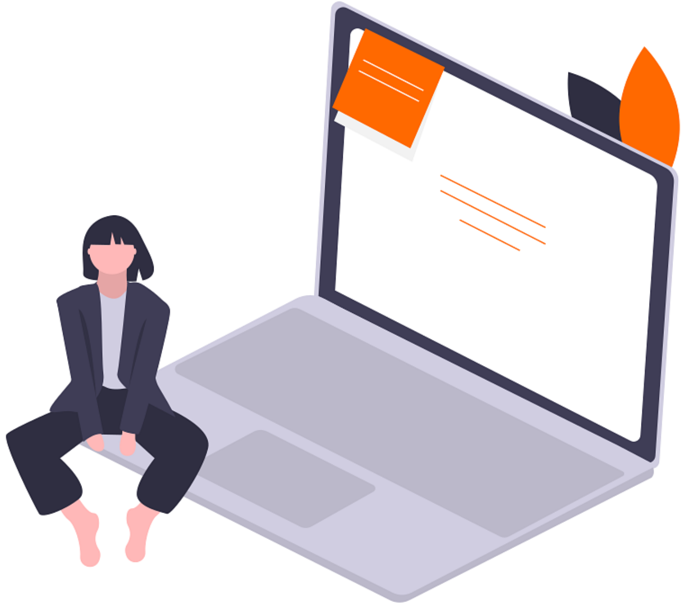
**SPEED**

**Notebooks + consistent interfaces**

# OVERCOMING THE CHALLENGES

SPEED

Notebooks + consistent interfaces



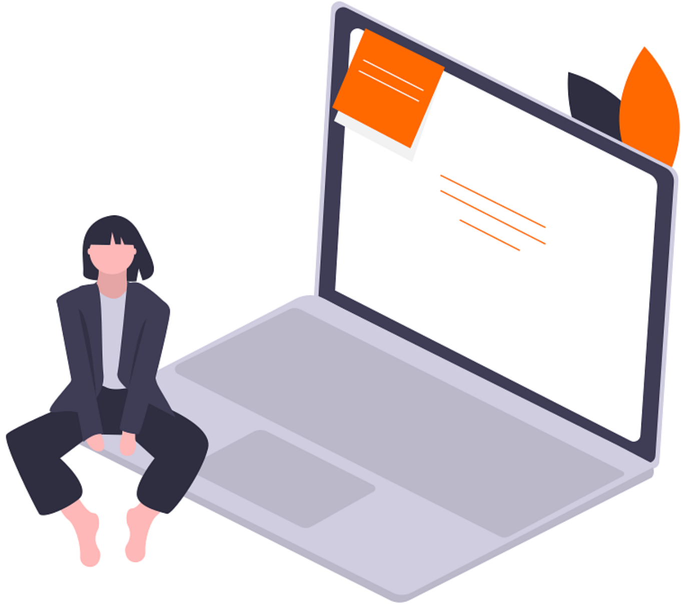
SAFETY

Amazon SageMaker training/scoring

# OVERCOMING THE CHALLENGES

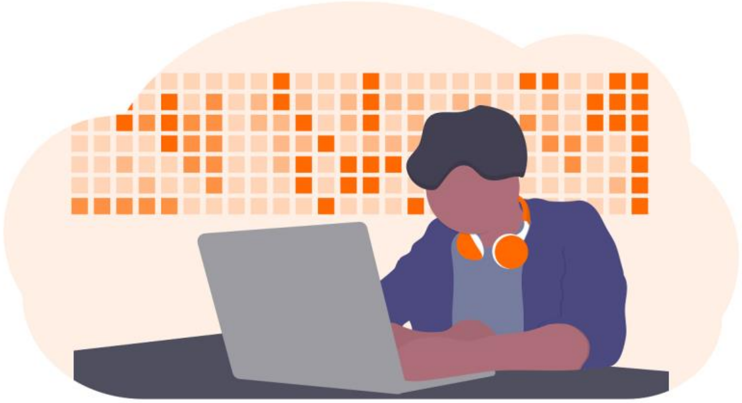
SPEED

Notebooks + consistent interfaces



SAFETY

Amazon SageMaker training/scoring



SCALE

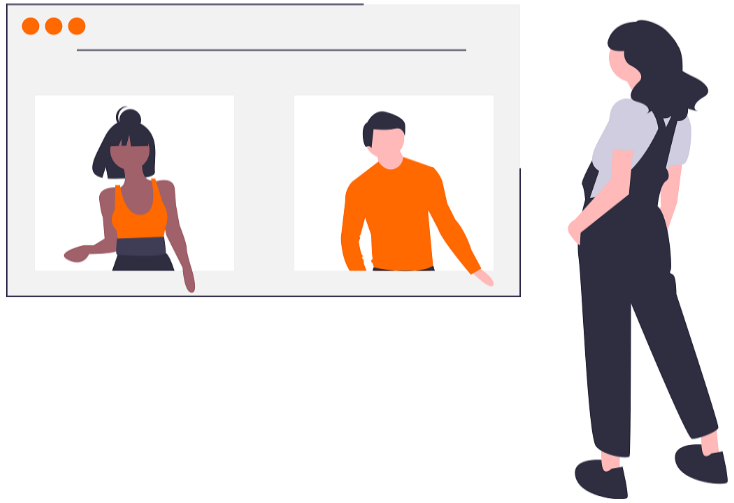
Amazon SageMaker training/scoring

# OVERCOMING THE CHALLENGES

**COST-EFFICIENCY**

**Amazon SageMaker training/scoring**

# OVERCOMING THE CHALLENGES



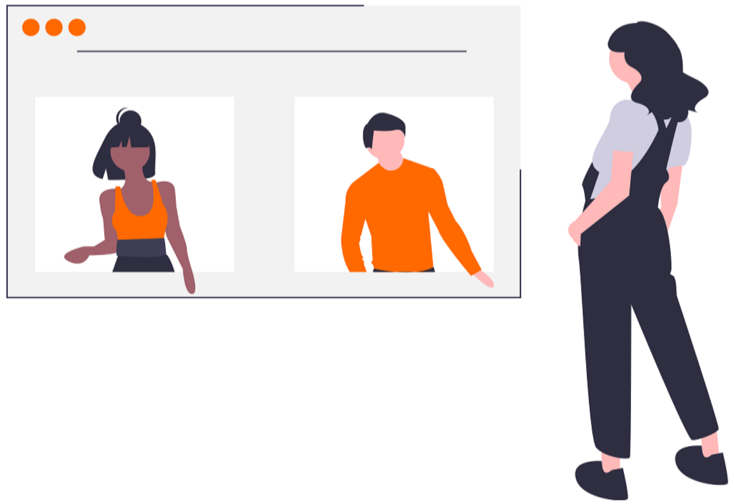
COST-EFFICIENCY

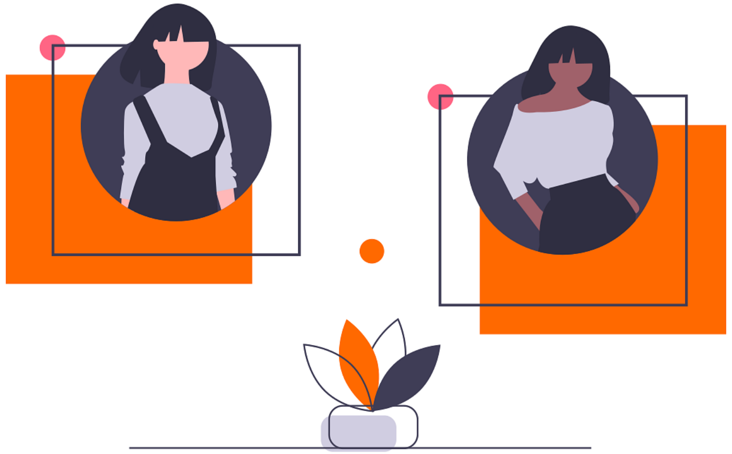
Amazon SageMaker training/scoring

COLLABORATION

Consistent interfaces

# OVERCOMING THE CHALLENGES





**COST-EFFICIENCY**

Amazon SageMaker training/scoring

**COLLABORATION**

Consistent interfaces

**SATISFACTION**

Central team support

# Machine learning at scale offers exciting new opportunities for us to serve our customers better.

# Demo

# In summary

- Machine learning is a hard problem
- Amazon SageMaker simplifies ML with modern application architecture
- Kubernetes is great for large-scale container orchestration
- Amazon SageMaker and Kubernetes can provide greater benefits when combined

# Thank you!

❗

Please complete the session survey in the mobile app.

AWS re:Invent

aws