Track 2|Session 5

利用 SageMaker 深度學習容器化在廣告推播之應用

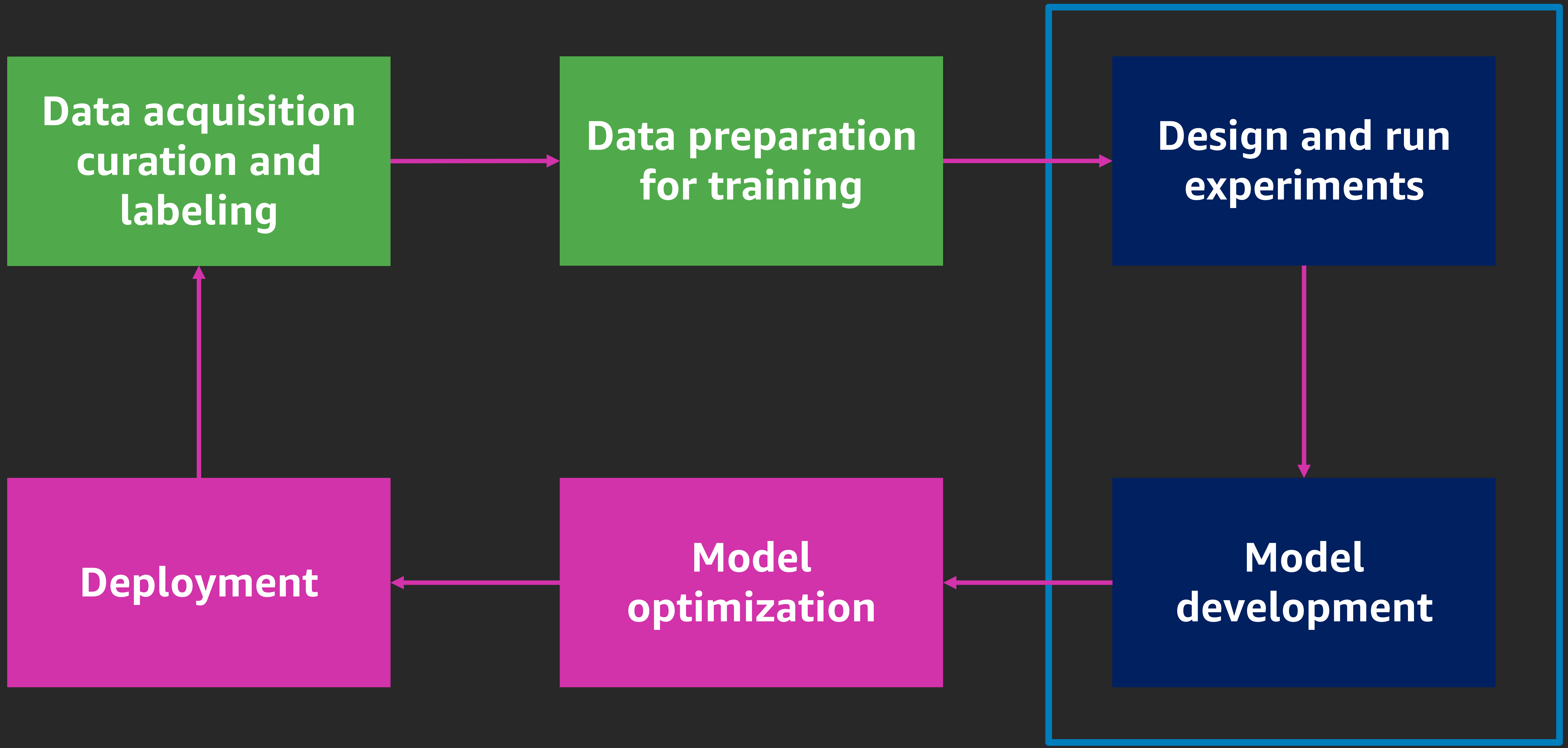
Young Yang
ML Specialist SA
Amazon Web Services

Hsuan Chiu
Senior Data Engineer
Data Science
VPON

AWS SUMMIT ONLINE

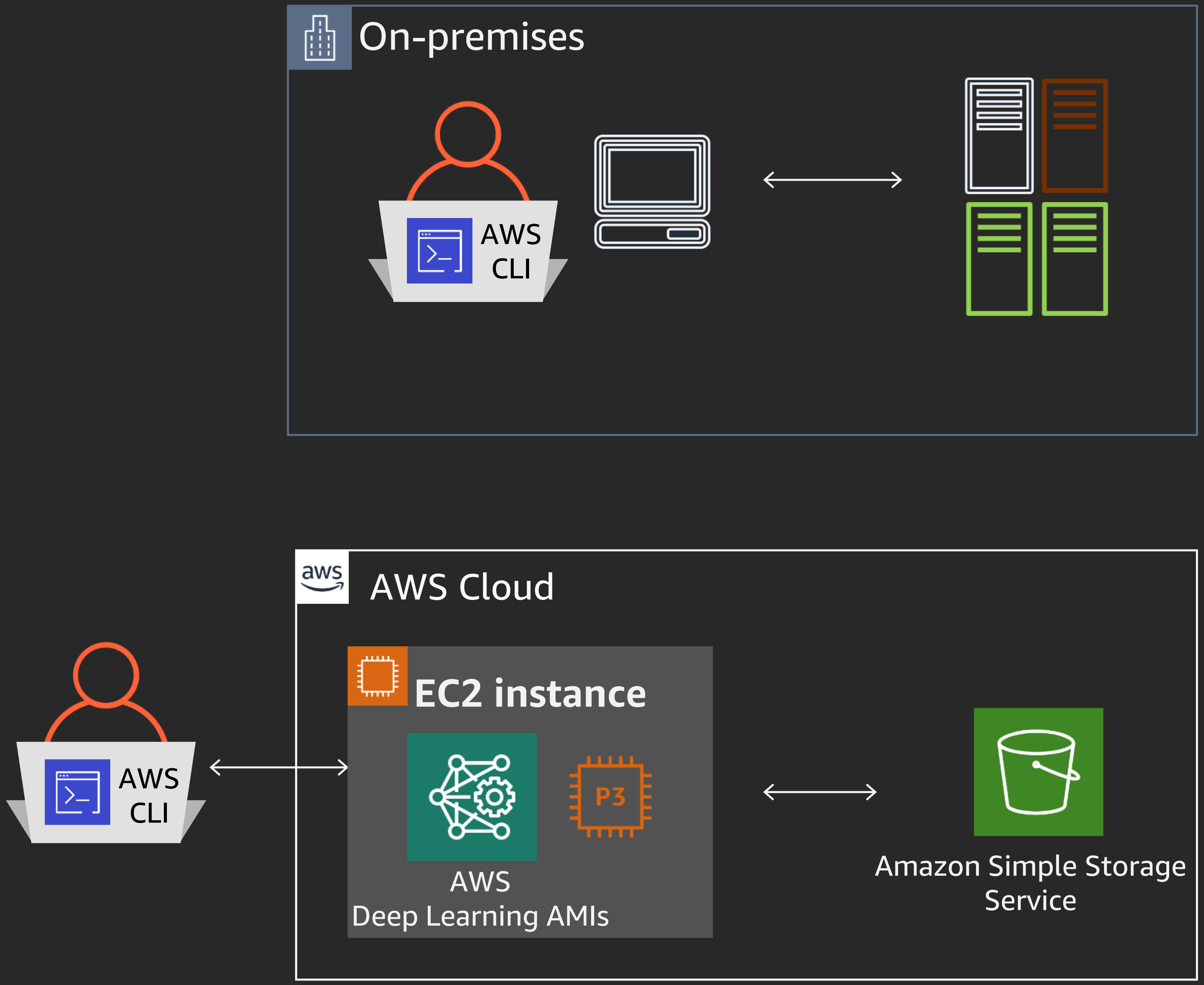
© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Machine learning workflow



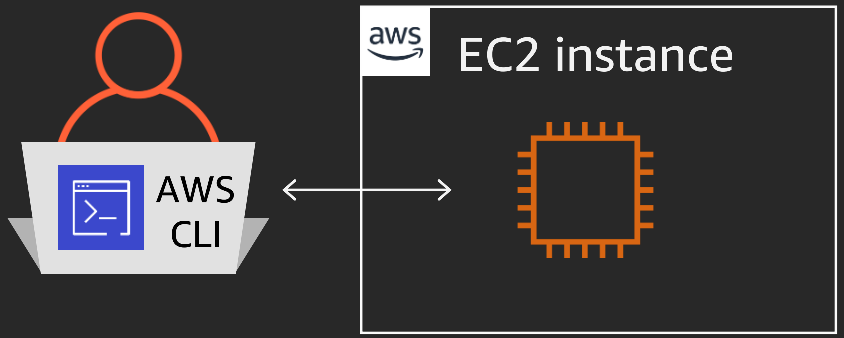
Common machine learning setups

1. Code & frameworks
2. Compute (CPUs, GPUs)
3. Storage

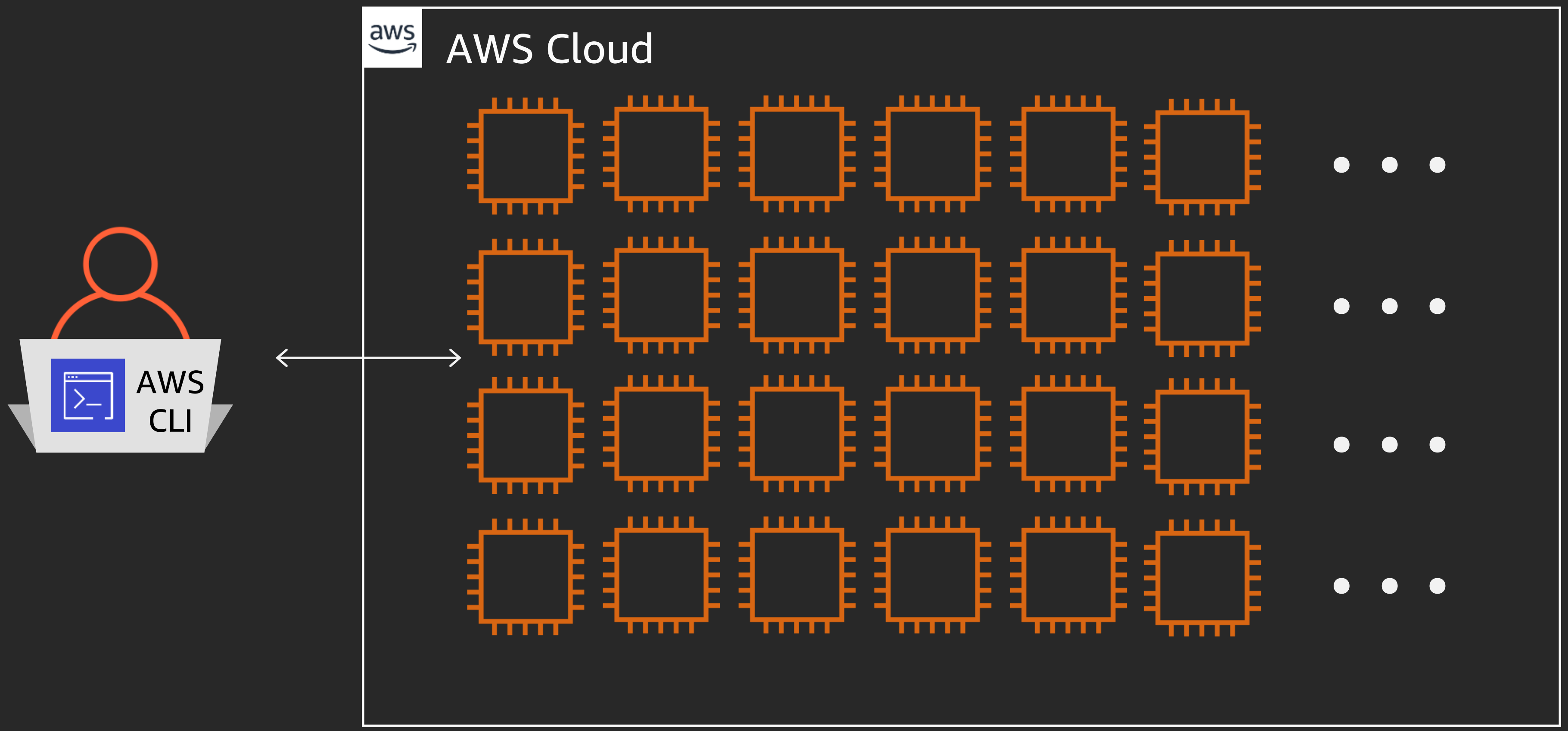


Deep learning is computationally expensive, but can be scaled-out

How do we go from



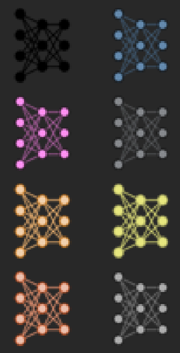
this,



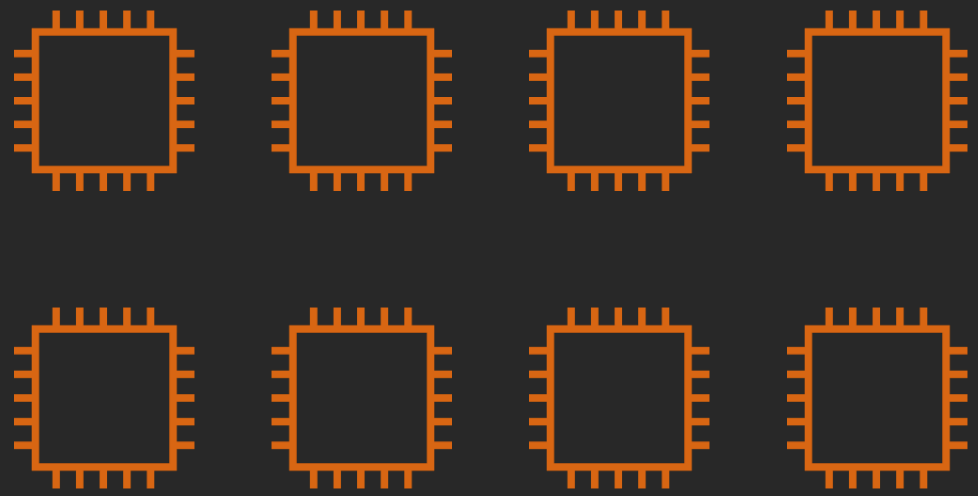
to this

Scaling-out deep learning training

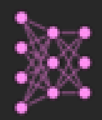
Parallel experiments



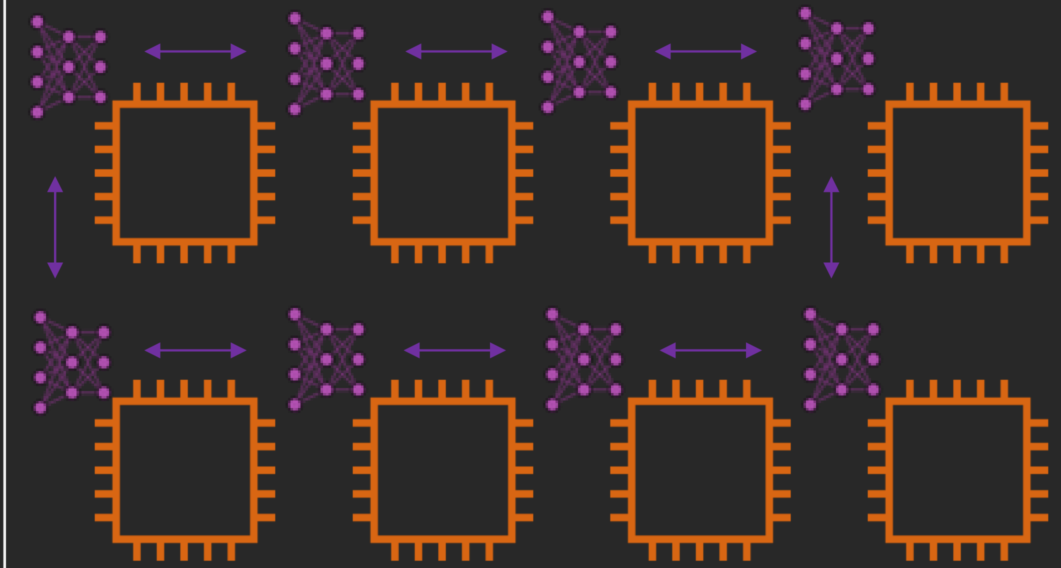
Different models running parallel to find the best model



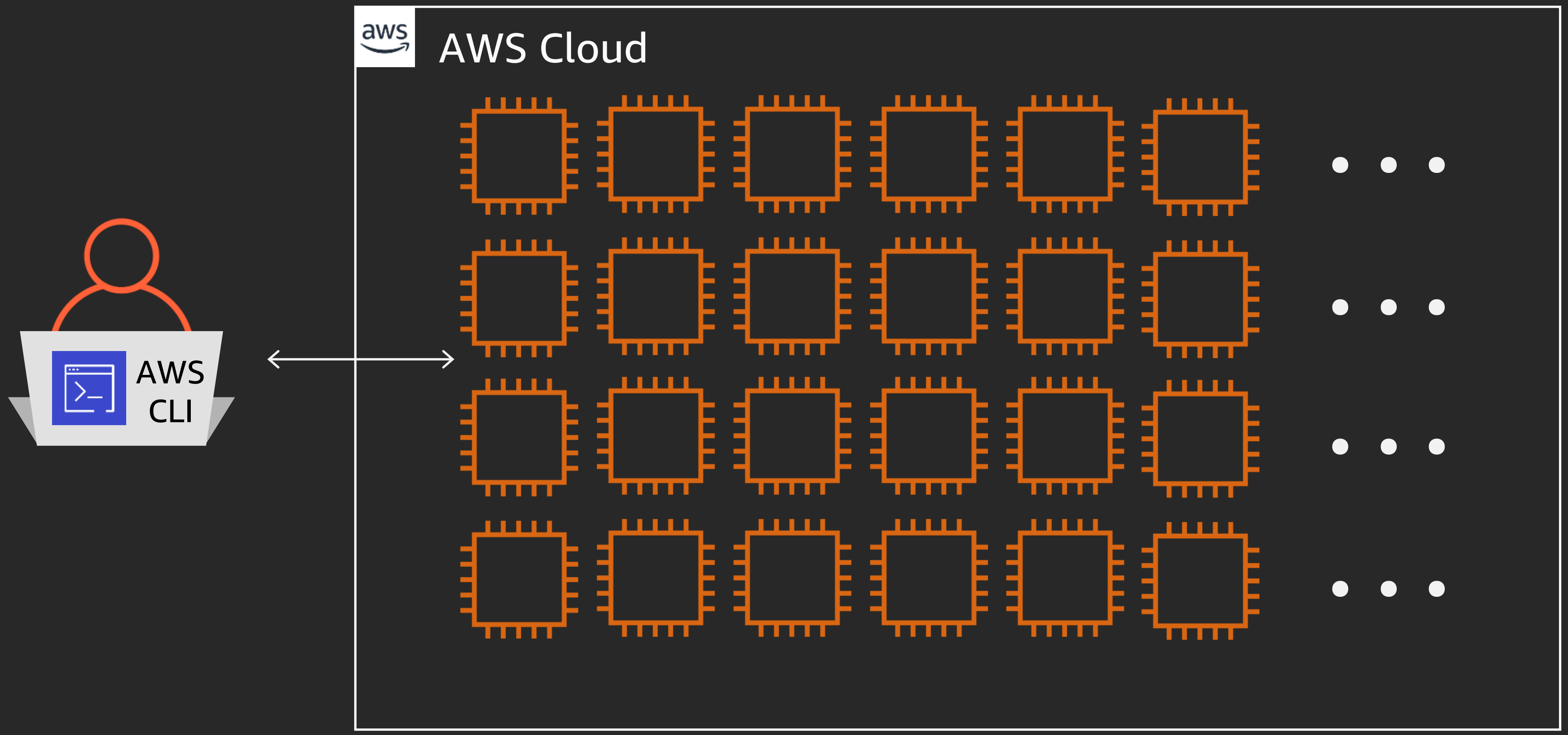
Distributed training



Distributing training of a single model to train faster



But there are challenges to scaling



Code and dependencies

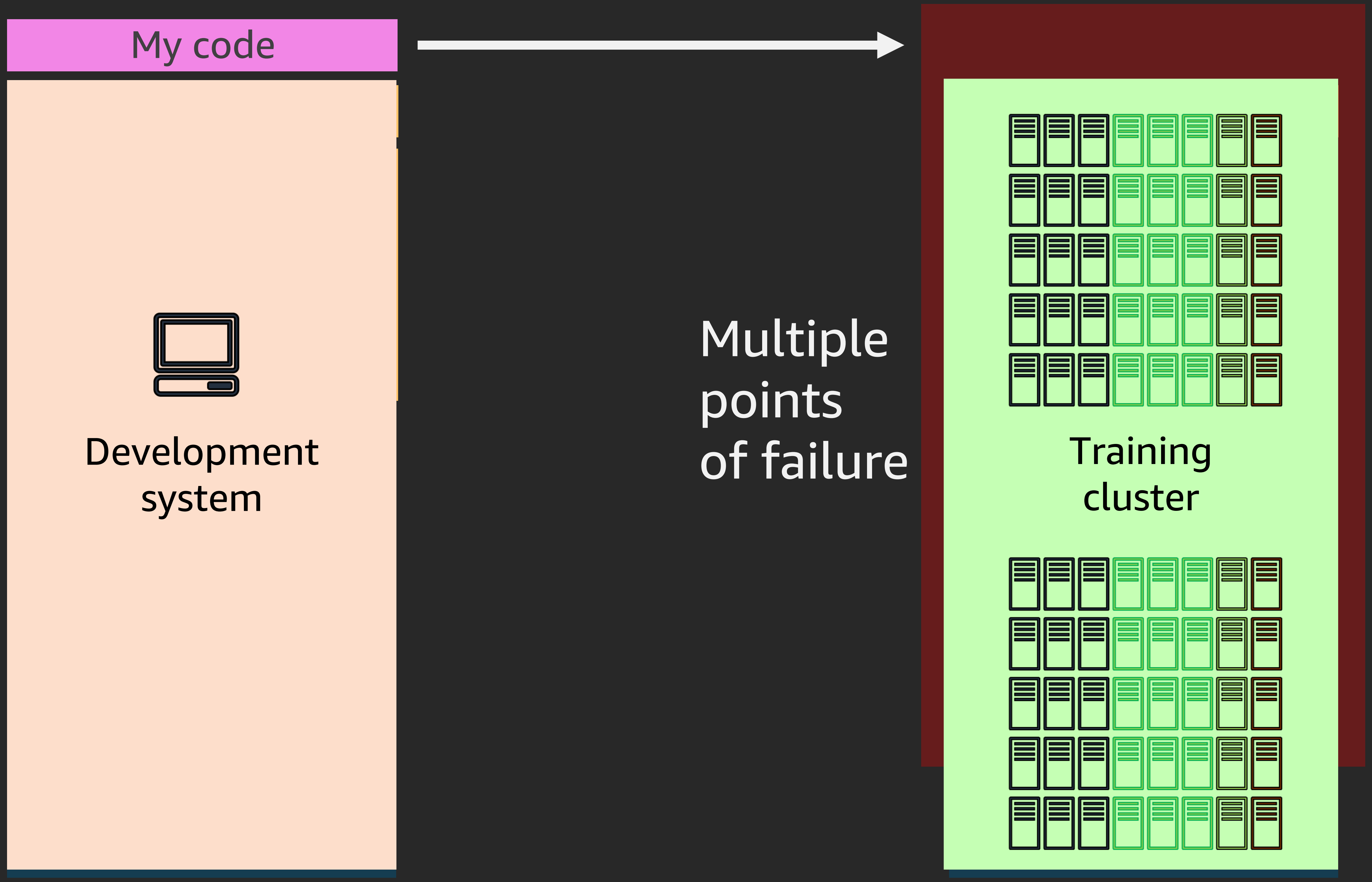
Cluster management

Infrastructure management

Machine learning stack is complex

- “My code requires building several dependencies from source”
- “My code isn’t taking advantage of the GPU/GPUs”
    - “Is cuDNN, NCCL installed? Is it the right version?”
- “My code is running slow on CPUs”
    - “Oh wait, is it taking advantage of AVX instruction set”
- “I updated my drivers and training is now slower/errors out”
- “My cluster runs a different version of framework/Linux distro”

Makes portability, collaboration, and scaling training really, really hard!


My code
Development system
Multiple points of failure
Training cluster

Containers for machine learning

TensorFlow

Keras
scikit-learn
horovod
pandas
numpy
openmpi
scipy
Python
others

CPU: mkl

GPU: cuDNN
cuBLAS
NCCL
CUDA toolkit

TensorFlow container image

+

Your training scripts

Container runtime

NVIDIA drivers

Host OS

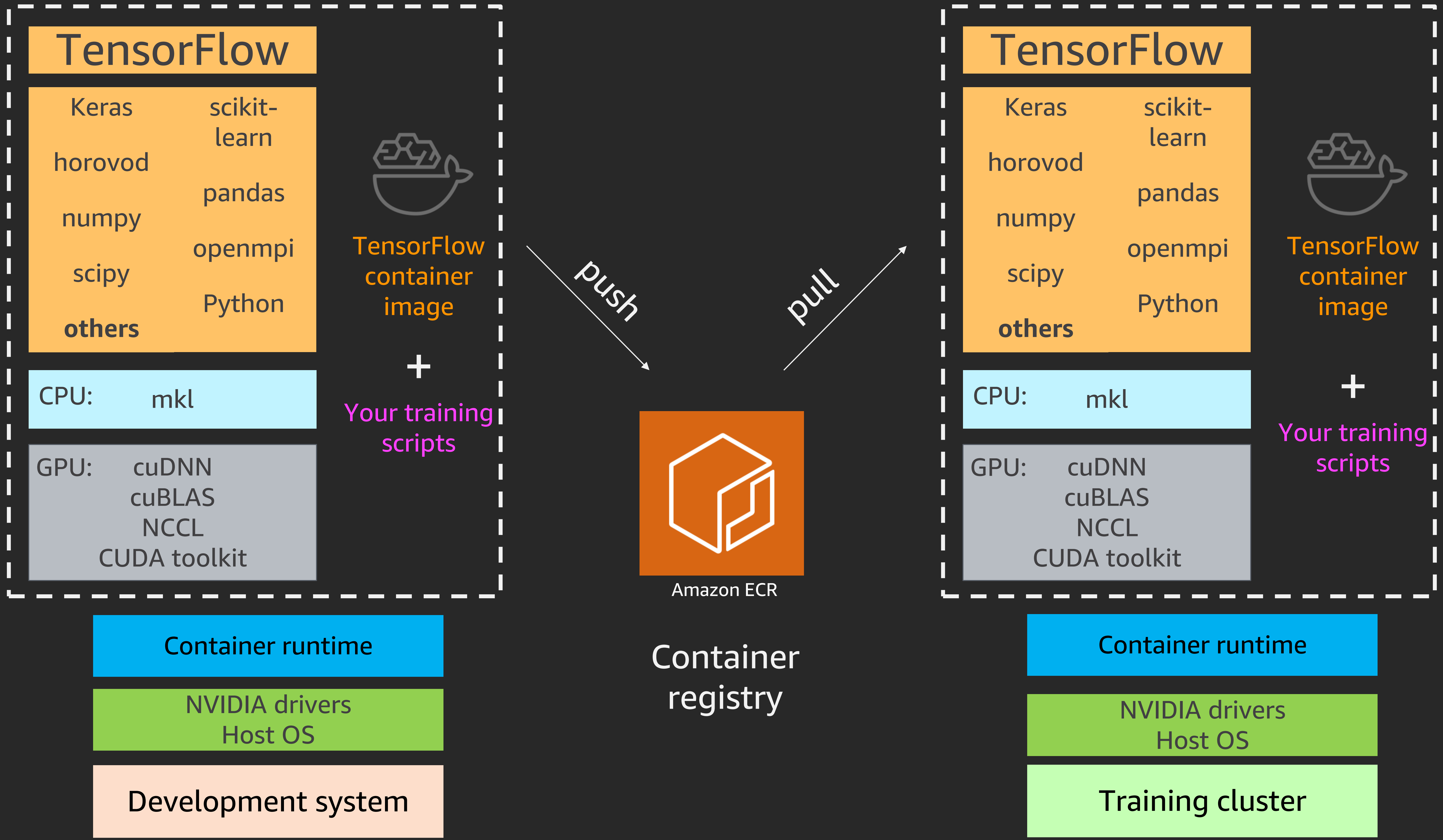
Infrastructure

Packages:

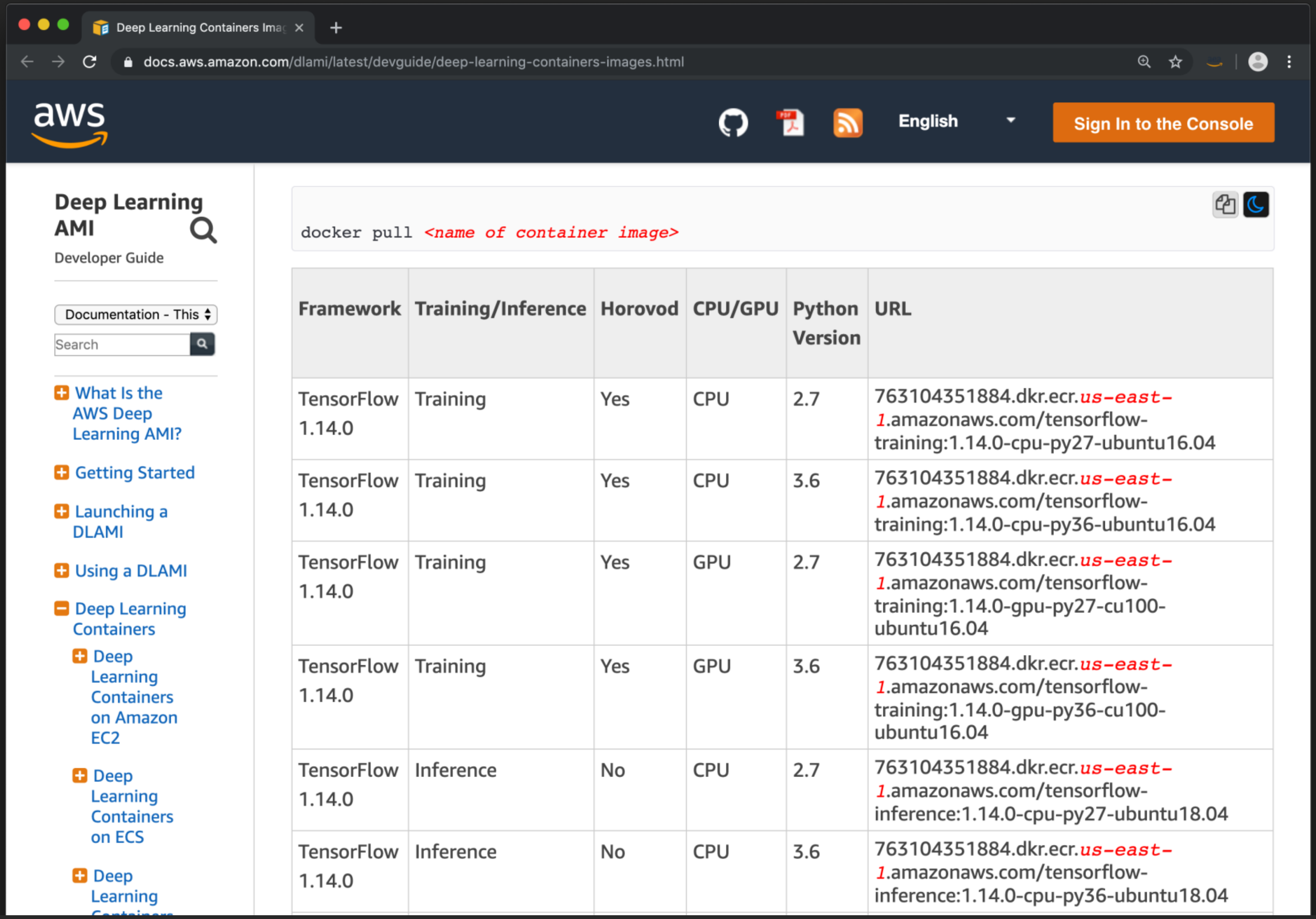
- Training code
- Dependencies
- Configurations

ML environments that are:

- Lightweight
- Portable
- Scalable
- Consistent


TensorFlow
Keras
scikit-learn
horovod
pandas
numpy
openmpi
scipy
Python
others
CPU: mkl
GPU: cuDNN
cuBLAS
NCCL
CUDA toolkit
TensorFlow container image
+
Your training scripts
Container runtime
NVIDIA drivers
Host OS
Development system
push
pull
Amazon ECR
Container registry
TensorFlow
Keras
scikit-learn
horovod
pandas
numpy
openmpi
scipy
Python
others
CPU: mkl
GPU: cuDNN
cuBLAS
NCCL
CUDA toolkit
TensorFlow container image
+
Your training scripts
Container runtime
NVIDIA drivers
Host OS
Training cluster

AWS Deep Learning Containers




```
docker pull <name of container image>
```

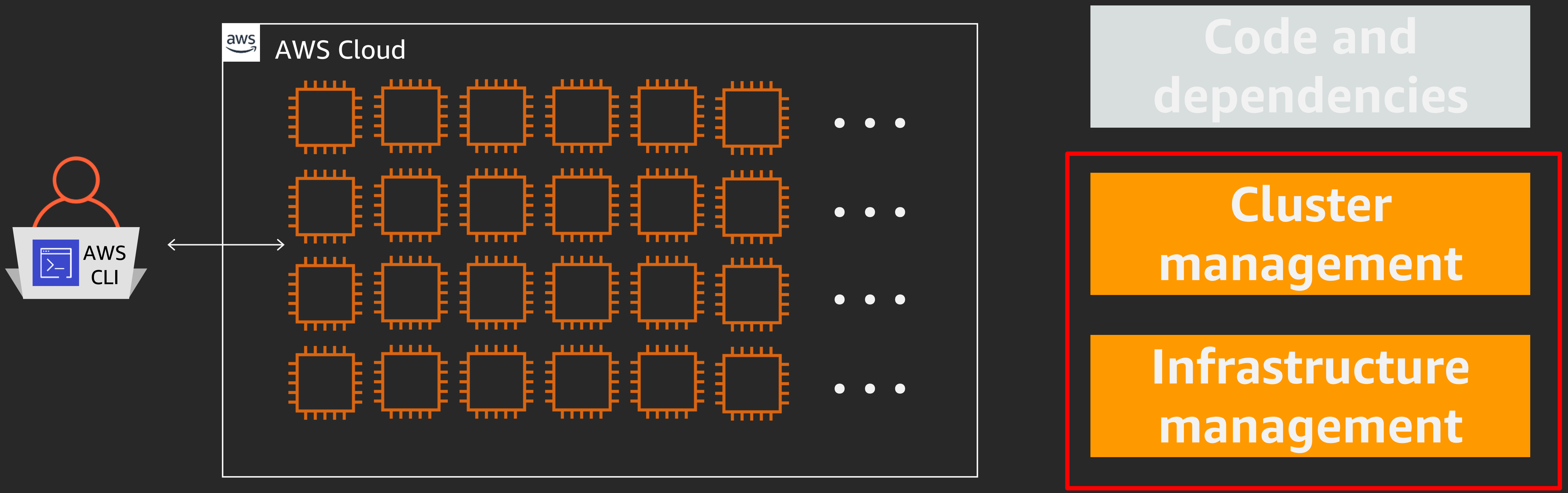
Framework	Training/Inference	Horovod	CPU/GPU	Python Version	URL
TensorFlow 1.14.0	Training	Yes	CPU	2.7	763104351884.dkr.ecr.*us-east-1*.amazonaws.com/tensorflow-training:1.14.0-cpu-py27-ubuntu16.04
TensorFlow 1.14.0	Training	Yes	CPU	3.6	763104351884.dkr.ecr.*us-east-1*.amazonaws.com/tensorflow-training:1.14.0-cpu-py36-ubuntu16.04
TensorFlow 1.14.0	Training	Yes	GPU	2.7	763104351884.dkr.ecr.*us-east-1*.amazonaws.com/tensorflow-training:1.14.0-gpu-py27-cu100-ubuntu16.04
TensorFlow 1.14.0	Training	Yes	GPU	3.6	763104351884.dkr.ecr.*us-east-1*.amazonaws.com/tensorflow-training:1.14.0-gpu-py36-cu100-ubuntu16.04
TensorFlow 1.14.0	Inference	No	CPU	2.7	763104351884.dkr.ecr.*us-east-1*.amazonaws.com/tensorflow-inference:1.14.0-cpu-py27-ubuntu18.04
TensorFlow 1.14.0	Inference	No	CPU	3.6	763104351884.dkr.ecr.*us-east-1*.amazonaws.com/tensorflow-inference:1.14.0-cpu-py36-ubuntu18.04

Prepackaged machine learning container images fully configured and validated

Optimized for performance with latest NVIDIA driver, CUDA libraries, and Intel libraries

https://docs.aws.amazon.com/dlami/latest/devguide/deep-learning-containers-images.html

Challenges with scaling deep learning



ML infrastructure and cluster management

ML services
Fully managed service that covers the entire machine learning workflow

Amazon SageMaker

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

- Easy, couple of LOC to scale
- Fully managed, no infrastructure effort
- Designed for machine learning
- Optimizing cost: on-demand / Spot

Management
Deployment, scheduling, scaling, and management of containerized applications

Amazon Elastic Container Service

Amazon Elastic Kubernetes Service

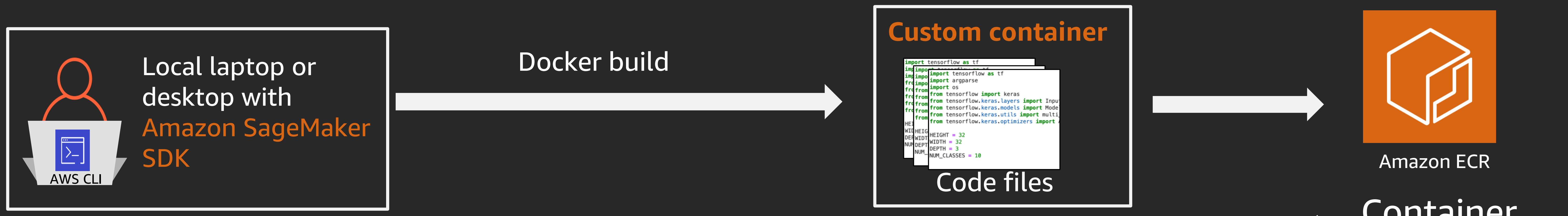
- Getting started hard, scaling easy
- Rely on IT/Ops for setup management
- DIY setup for ML use-cases
- Optimizing cost: DIY

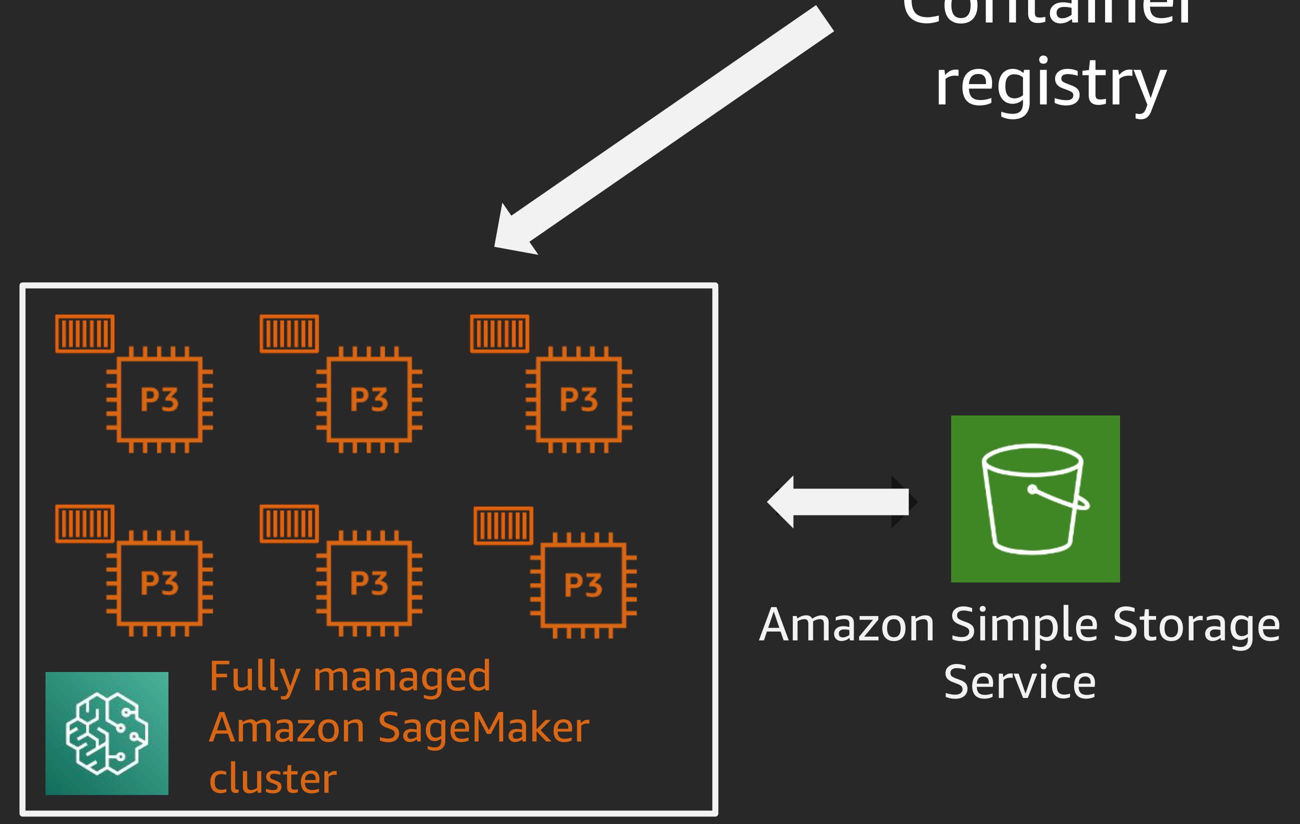
Compute
Where the containers run

Amazon EC2

- Getting started easy, scaling hard
- Rely on IT/Ops for setup management
- DIY setup for ML use-cases
- Optimizing cost: DIY

Hyperparameter search experiment using Amazon SageMaker





Approach:

1. Build a Docker image with your training scripts
2. Specify instance type (CPU, GPU)
3. Specify number of instances and hyperparameters to tune
4. Launch the tuning job

AWS如何加速
機器學習專案產品化

Hsuan Chiu

Senior Data Engineer
Data Science
VPON

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

About Vpon

The Critical Question In Digital Marketing

ML Case Study – Gender Prediction

Conclusion

Vpon

The Leading Big Data Company in Asia

Data Drives Transactions

Milestone

Vpon

2008
Founded in Taipei

2010
Establishment of Shanghai office

Launched 1st LBS mobile ad network in Asia

2011
Received $7M in Series A Funding

2014
Establishment of Hong Kong and Tokyo offices

Raised US$10M in Series B Funding

2015
Won 3rd for Forbes China's Top 100 Privately Held Small Businesses

- Won Agency & Advertiser Of The Year in 4 categories
- Won three Hong Kong Spark Awards

2016
Mob-ex Awards 2016
Won Bronze in Best In-app Advertising

Won Bronze for Campaign Greater China Specialist Agency of the Year

- Won Agency & Advertiser Of The Year in 3 categories: 2 Gold & 1 Sliver
- Festival of Media Global
- MARKies Awards
- ECI Awards
- Top Mobile Awards (TMA)

2017
Establishment of Singapore office

Mob-ex Awards 2017
Won Bronze in Best Mobile Advertising Platform

Won Bronze for Campaign Greater China Specialist Agency of the Year for two consecutive years

- Won Agency & Advertiser Of The Year in 4 categories: 3 Golds & 1 Silver
- Campaign Digital Media Awards
- MARKies Awards 2017
- Golden Mouse
- Tiger Roar

2018
Establishment of Osaka office

Won one Gold and two Bronze for Mob-Ex Awards 2018

Top 10 Big Data Solutions Providers in the APAC Region

- Won Gold for Best In-App Advertising
- Won Bronze for Best In-App Advertising
- Won Bronze for Best Mobile Advertising Platform

2019
Japan Office Expansion

Won Gold for Best Location-based at Mob-Ex Awards 2019

Won Big Data Solution award at the Capital Magazine's BOB Awards 2019

Won Silver award for digital transformation in eASIA Awards 2019

- **Won Gold winner in the Best Data-driven Marketing Campaign category at Brain Magazine Awards 2019**
- Mediazone's annual Most Valuable Services Awards in Hong Kong 2019
  - "Most Reliable Big Data Analytics and Application Leader"
  - "Best Cross-Border Marketing Services"
  - "Excellence in Corporate Governance Customer Services"
- Won Silver for Best Idea– Mobile at Markies Award 2019

Vpon Big Data Group

The Leading Big Data Company in Asia

21 Billion
daily biddable inventory

1500
renowned brands with collaborative experiences

12 years
of services across APAC

900 Million
unique devices per month

7
offices in Hong Kong, Shanghai, Singapore, Taipei, Bangkok, Tokyo and Osaka

Trata DMP - Largest Travel Audience Data Pool in Asia

100M+

Travel Intent Data in Asia

60M+

China Passport Holder

1000+

Traveler Tags

Vpon Available Tag Categories

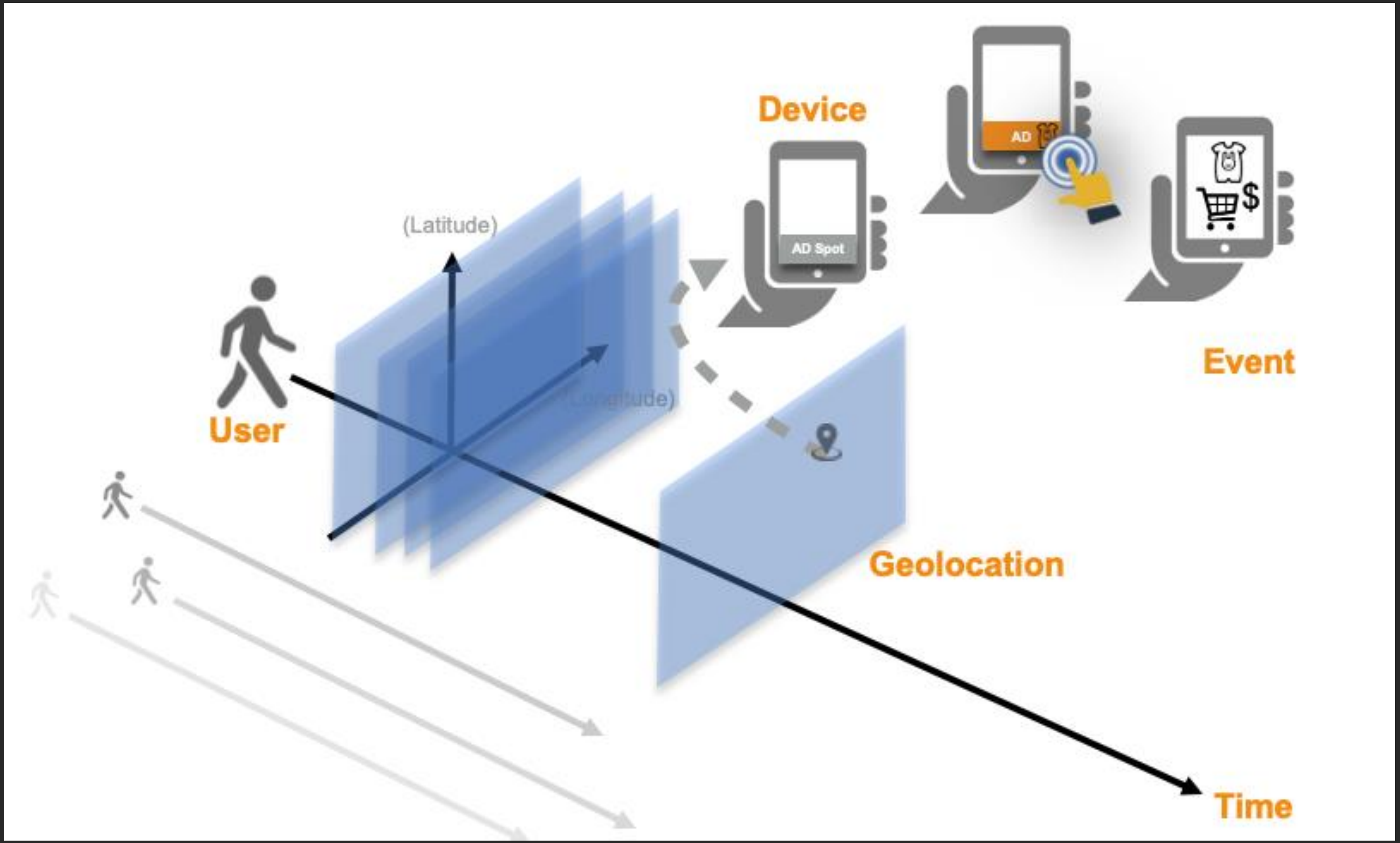
Locations	Demographics	Behaviors	Interests
Country	Gender	Ad Interests	Lifestyle
Province/ City	Age	Operation System	App Interests
	Age group of family members	Ad Format Preference	Fashion Style
	Income level	Travel Pattern	
	Device Language		
	Destination Country		

Based on multiple combinations of the tags, you can identify some of the hidden segment groups who may be your potential audiences with high chance.

The Critical Question In Digital Marketing

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The Critical Questions In Digital Marketing



Vpon AI Technology

More than 10 machine learning models with real case applications.

- Lookalike Modeling
- Recommendation
- Fraud Detection
- Traffic Accident Prediction
- Text Mining
- Behavior Prediction
- Gender Prediction Model
- Automated Location Discovery
- Intelligent Route Mining
- Traveling behaviors prediction

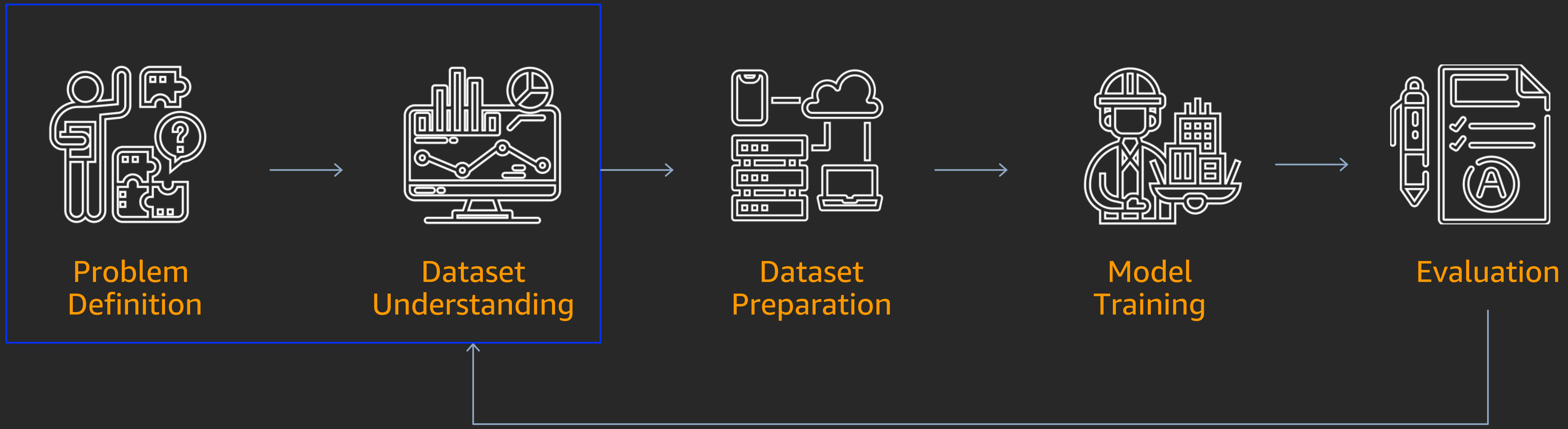
ML Case Study - Gender Prediction

aws SUMMIT ONLINE

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Research Stage

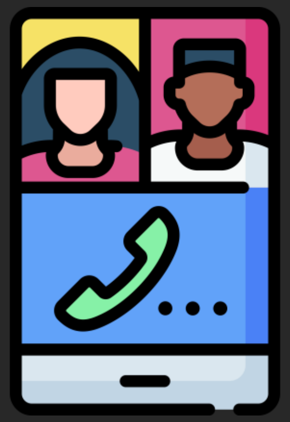
Research Workflow



Research Stage

Problem Definition

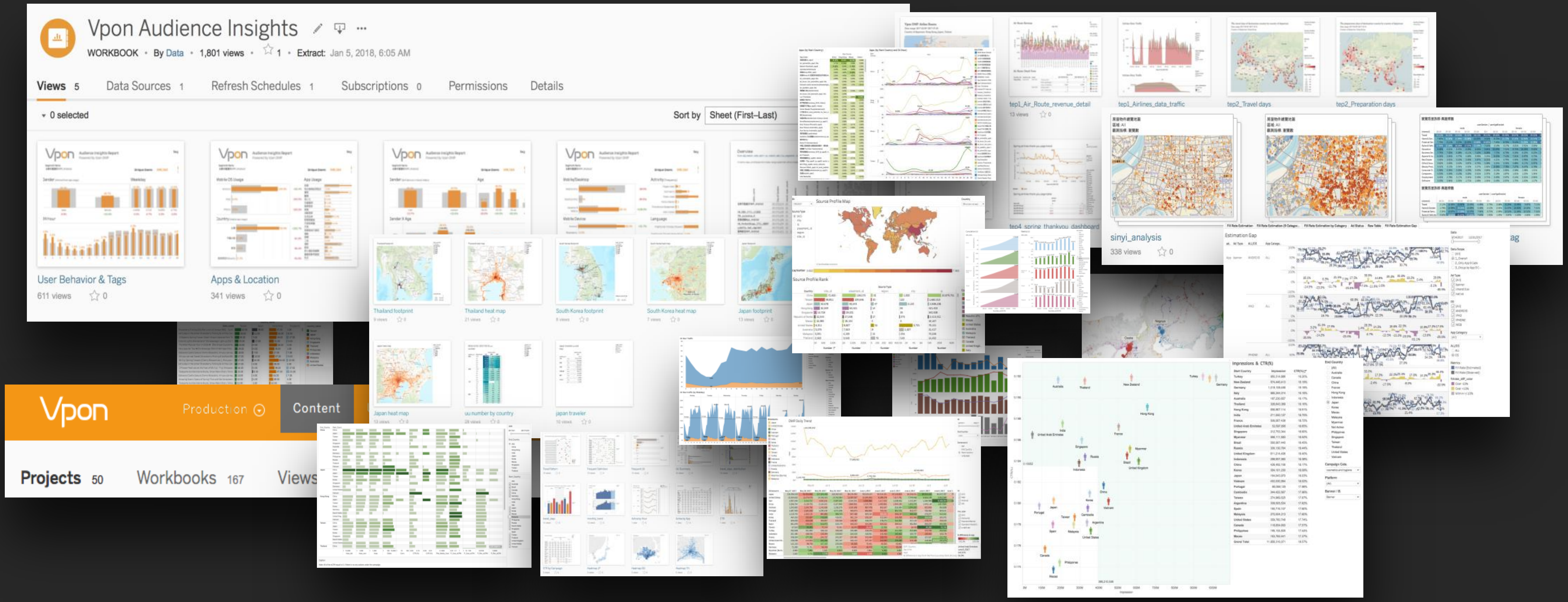
- Binary Classification.
  - Given a set of available mobile related features during a specific period, we want to predict whether the device owner behaves more like a man or woman.
- Features
  - Device geo info, app usage patterns, device info, active time range, etc.
- Goals
  - Accurately identify gender for new incoming devices.
  - Improve prediction accuracy for previous mis-labeled devices.



Research Stage

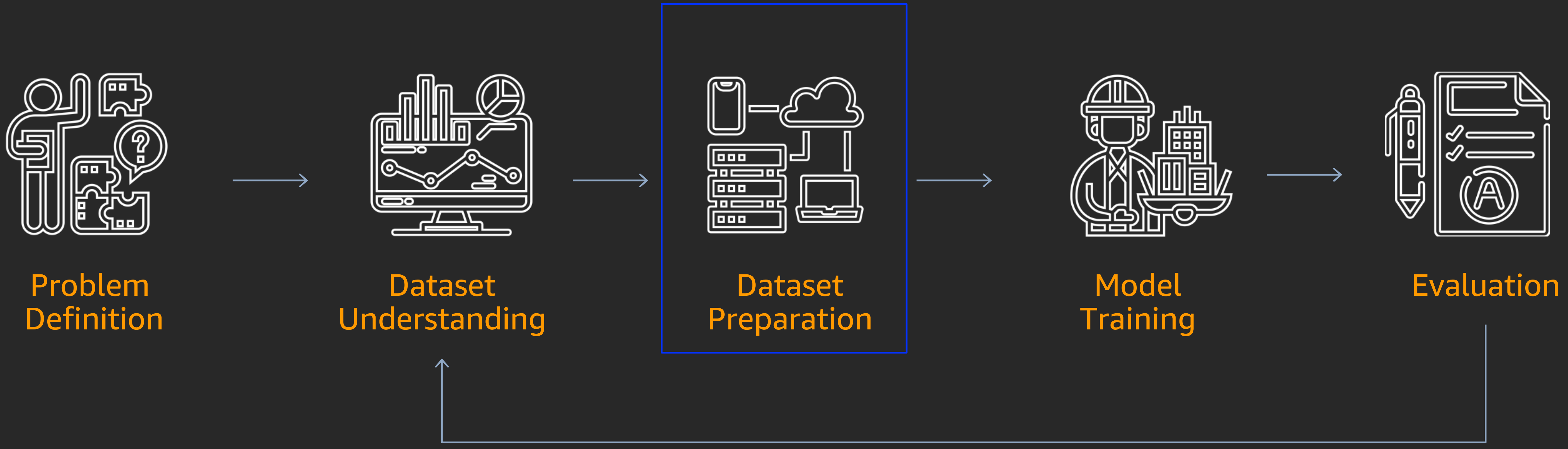
Dataset Understanding

- Ground-truth Data Analysis
  - Gender Label Distribution, Feature Importance, etc.
- Data visualization
  - Matplotlib, Tableau.



Research Stage

Research Workflow



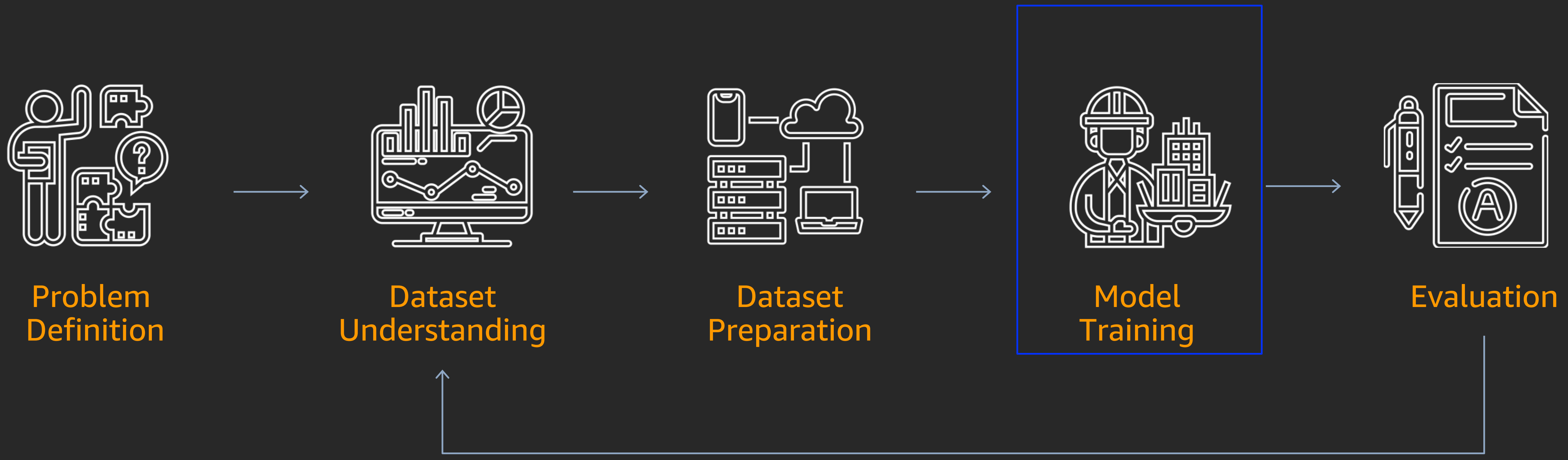
Research Stage

Dataset Preparation

- Data Collection
  - Collect and retrieve feature data.
- Data Integration
  - Enrich raw data with other useful info, e.g. google store metadata, poi.
- Data Cleaning
  - Remove null or abnormal values.

Research Stage

Research Workflow



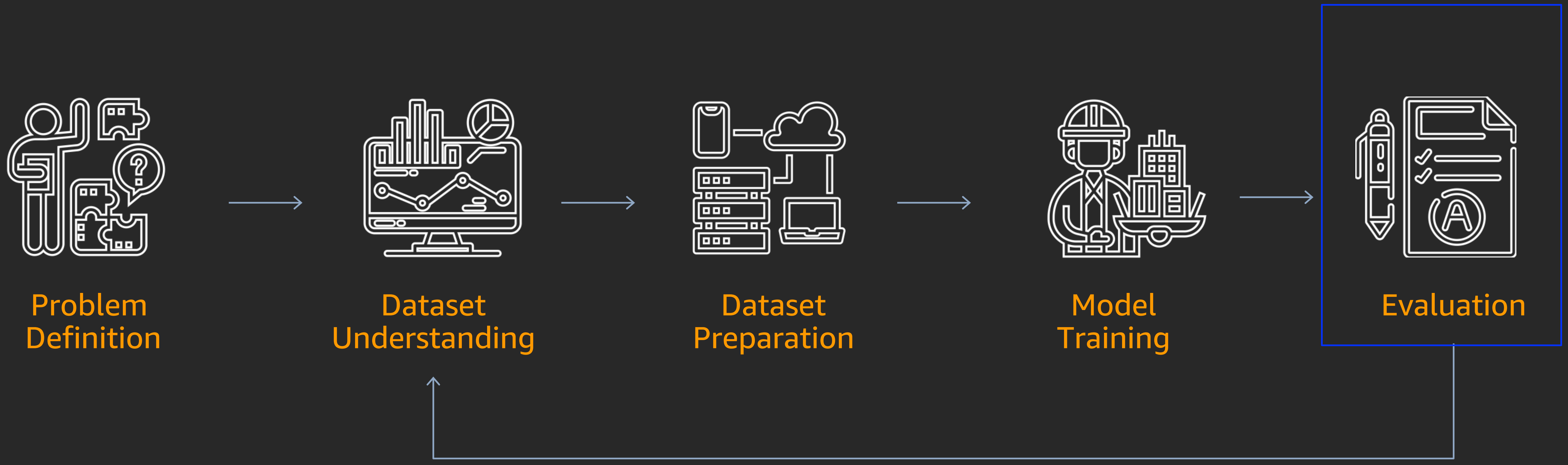
Research Stage

Model Training

- Feature Engineering
  - One-hot encoding, Feature combination, etc.
- Model Training and Parameter Tuning
  - Logistic Regression, XGboost, Deep Learning Framework, etc.
  - Feature Normalization.

Research Stage

Research Workflow



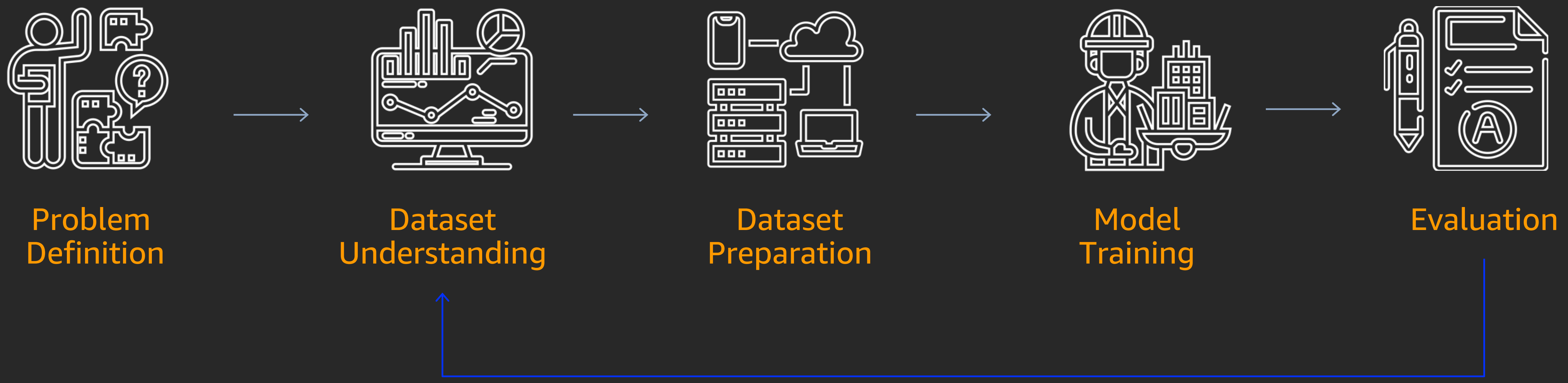
Research Stage

Evaluation

- Model Evaluation
  - PR, ROC, F1, etc.
- Prediction Result Evaluation
  - 3rd party verification: Google, FB.

Research Stage

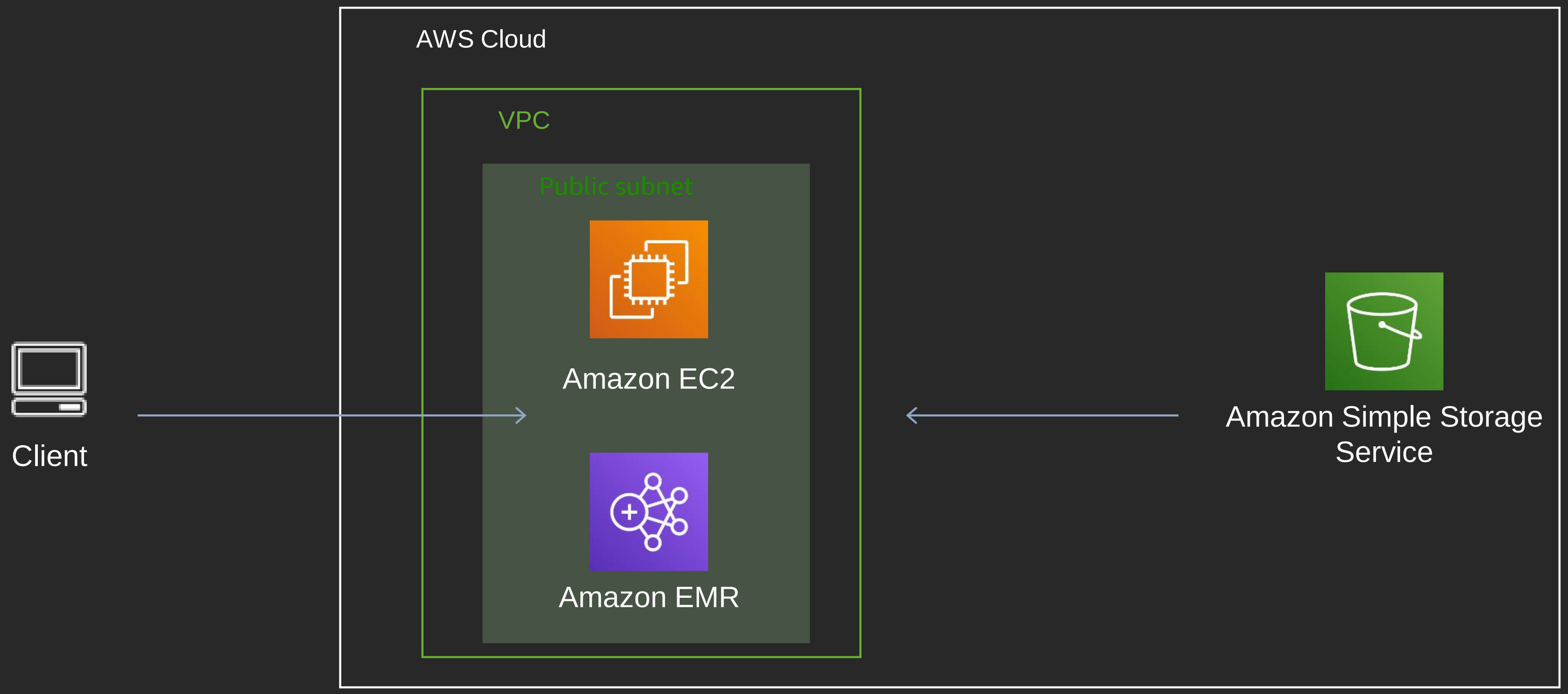
Research Workflow



Research Stage

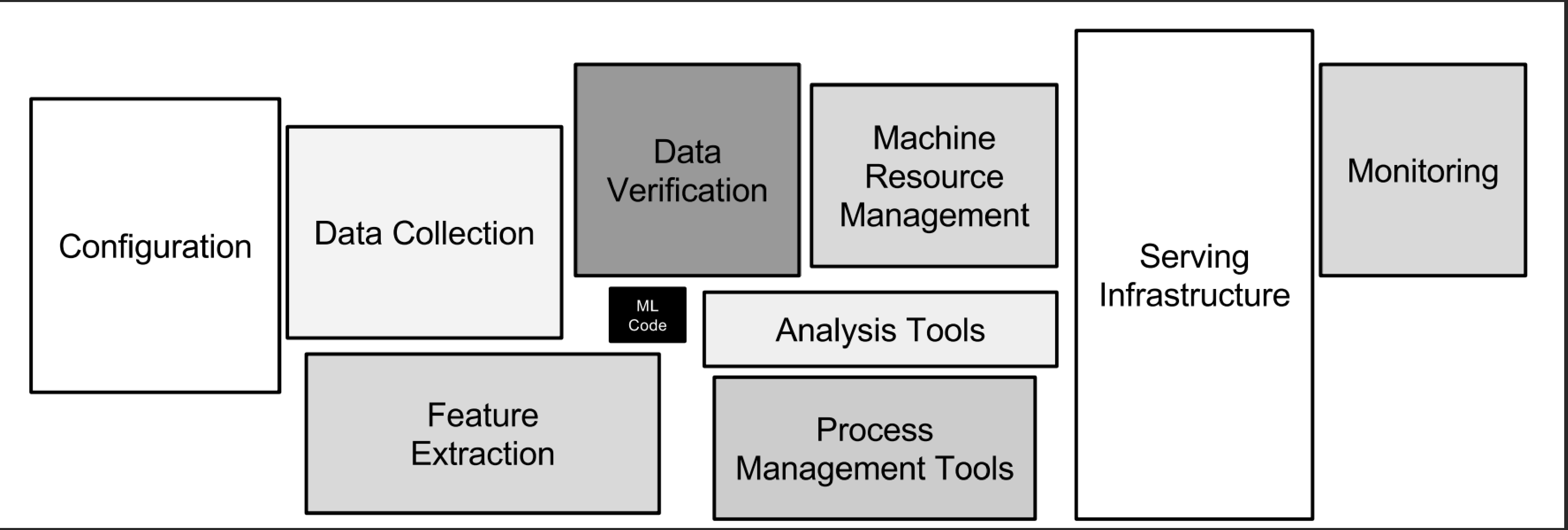
Research Tools and Environments.

- Research with local machine.
  - Pull data from S3 and repeat research work on local machine.
- Research with cloud env.
  - EC2.
  - EMR.
  - SageMaker.



From Research to Production

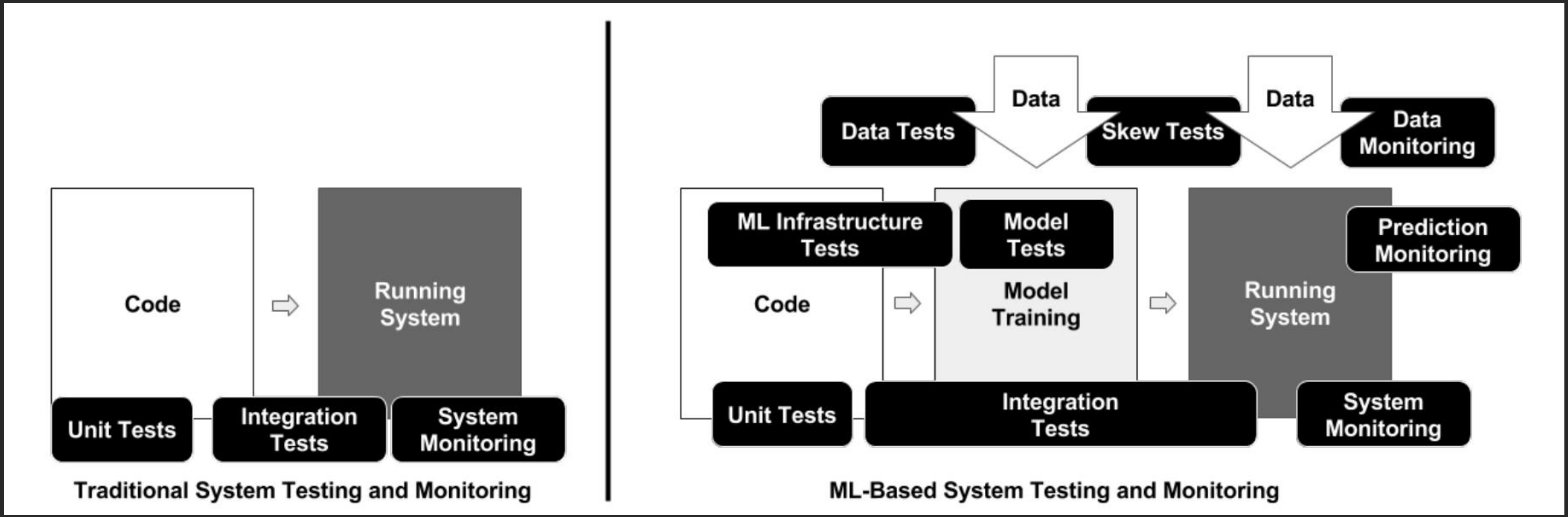
Hidden Technical Debt in Machine Learning Systems



Sculley et al., Hidden Technical Debt in Machine Learning Systems. NIPS 2015.

From Research to Production

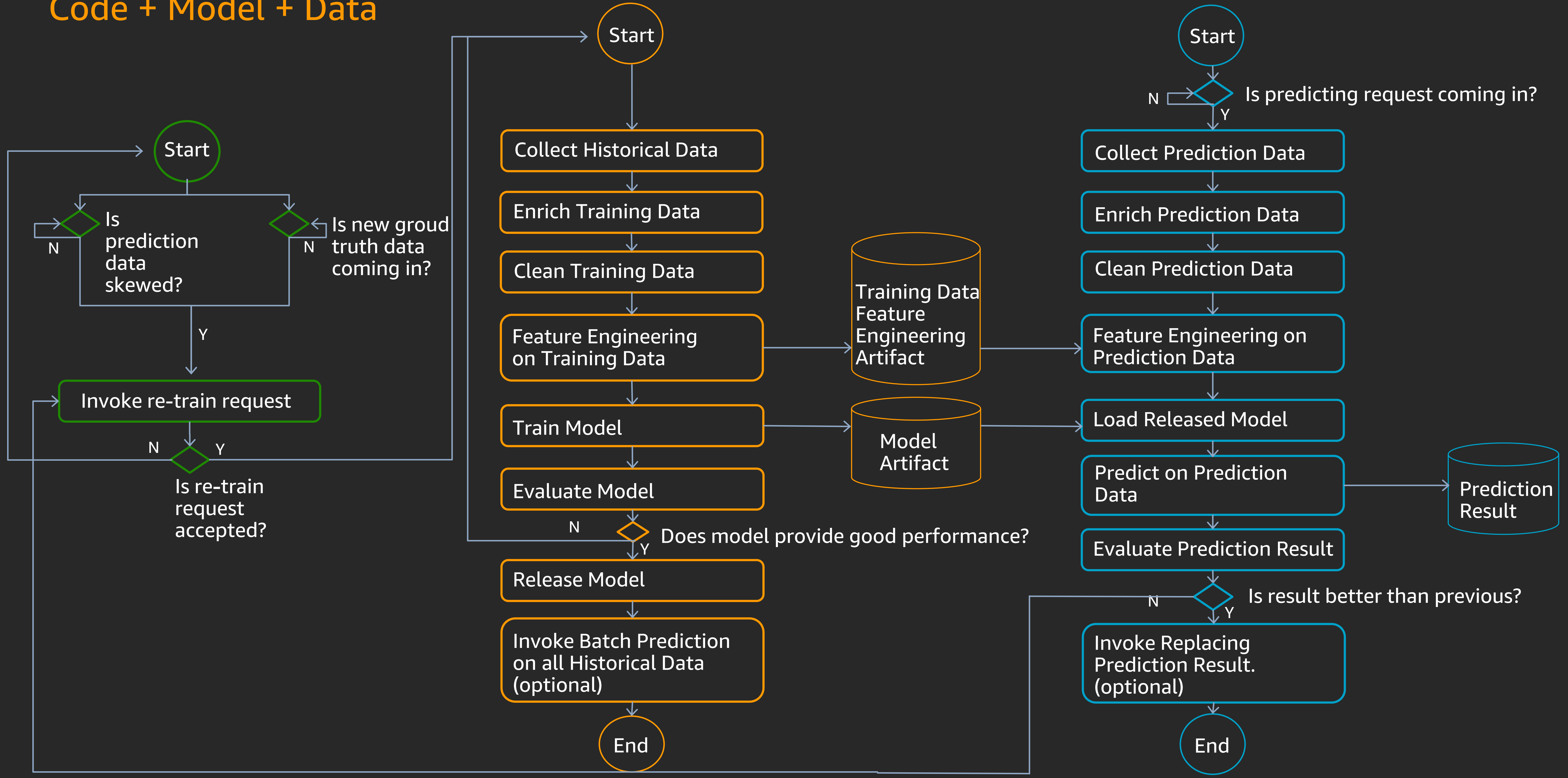
Code + Model + Data



Breck et al., The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction. IEEE Big Data 2017.

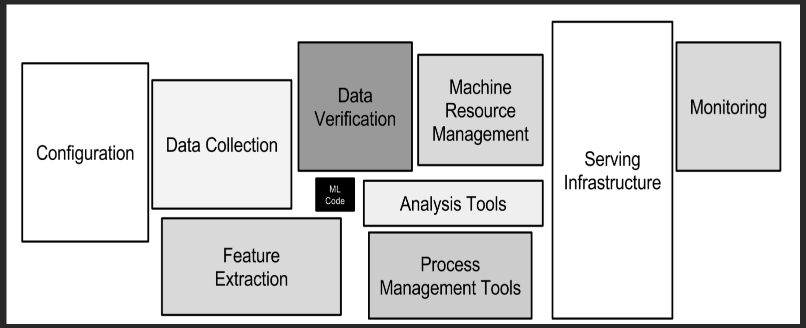
From Research to Production

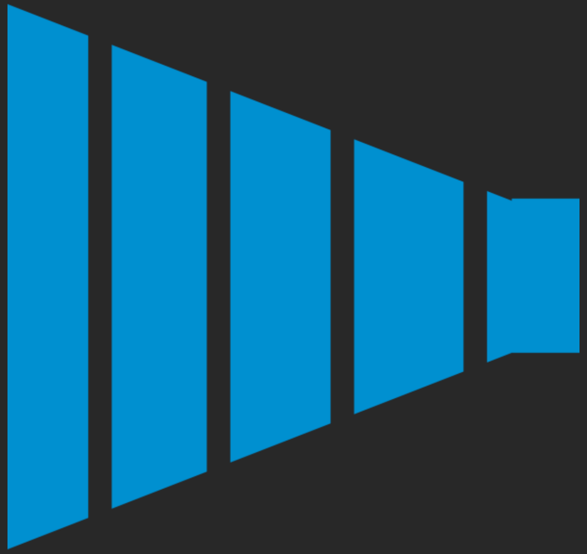
Code + Model + Data



From Research to Production

How to scale to production?

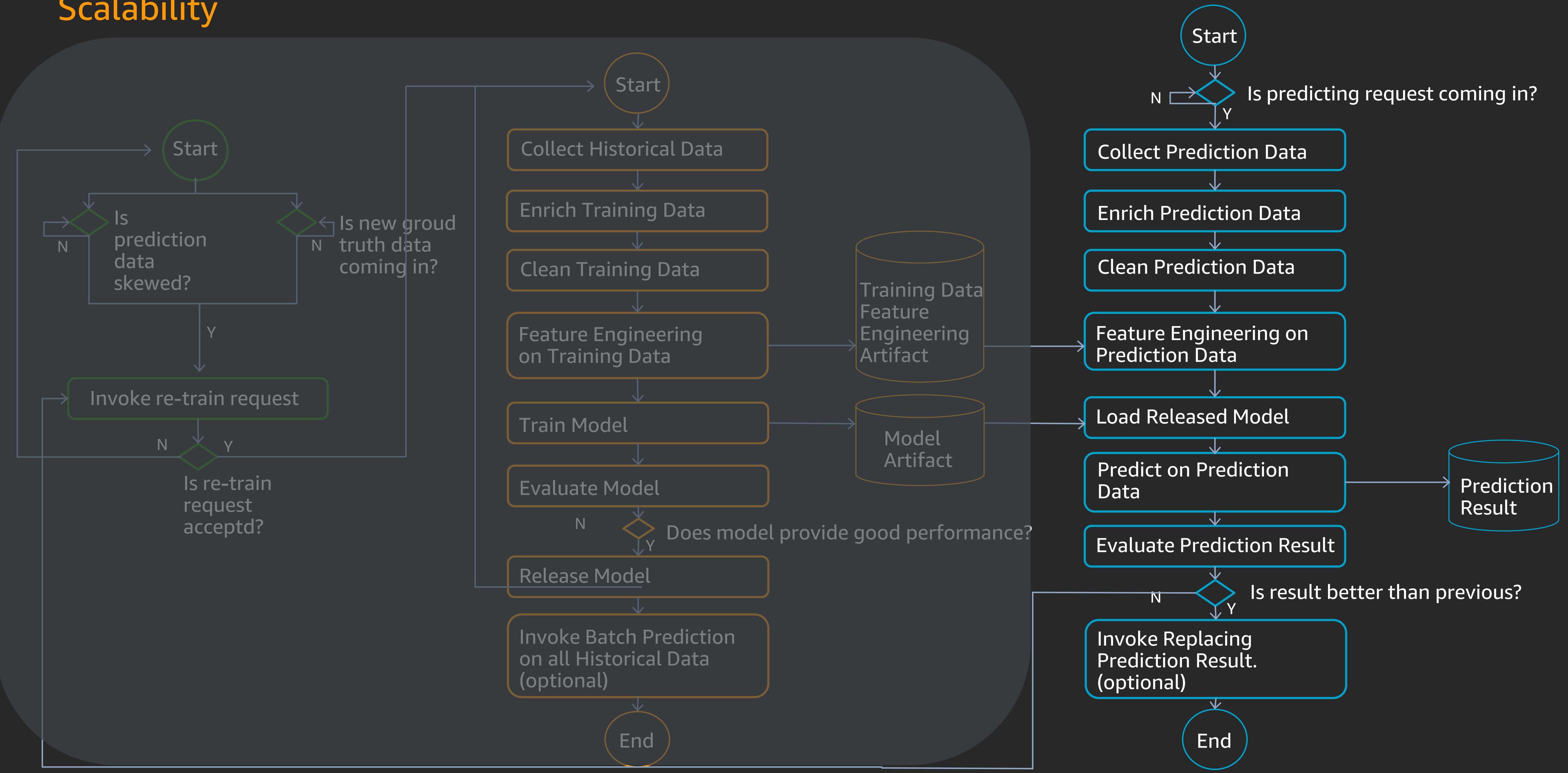




- Composability
- Scalability
- Portability

Production Stage

Scalability



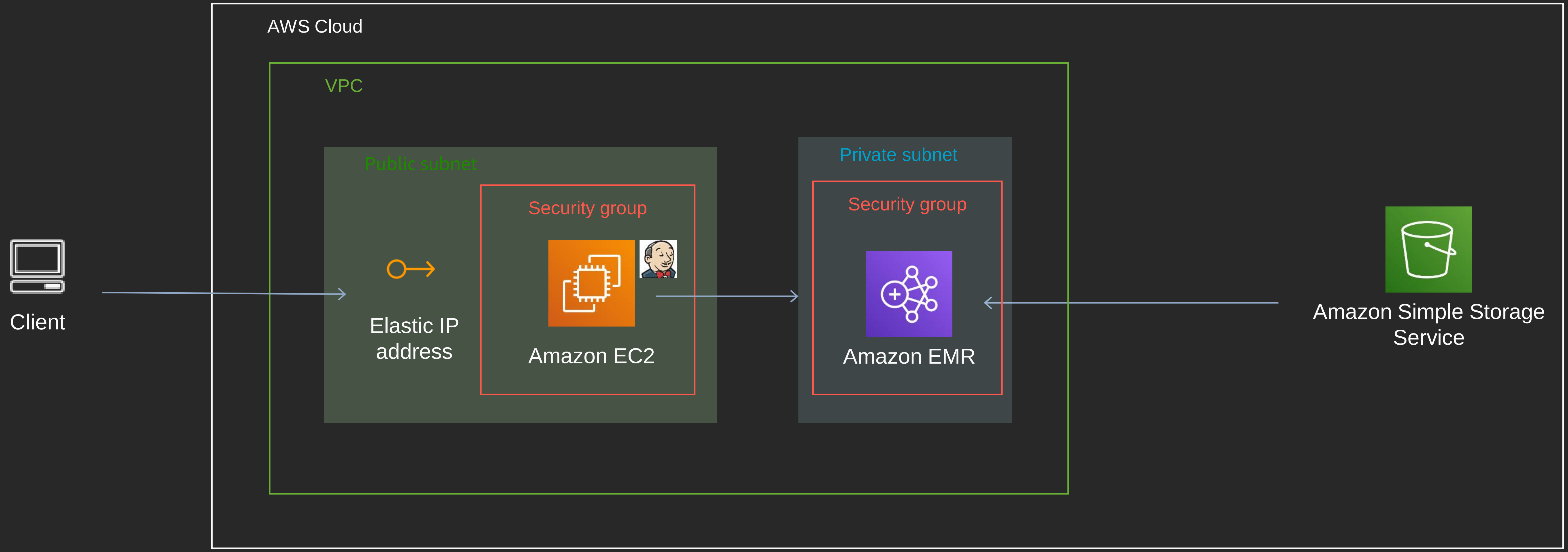
From Research to Production

Composability

- Pipeline Platform Requirements.
  - Reliable.
  - Visualization tool.
  - Pipeline scripting languages.

From Research to Production

Composability



From Research to Production

Scalability

- Prediction Service.
  - Prediction service is easily to be scaled out.
  - Prediction service can process batch or on-line requests interchangeably.

From Research to Production

Portability

- Multi-platform Model Deployment.
    - Model deployment should not be limited to a specific platform.
    - Model deployment should be easily to be integrated with other services, e.g. current existed microservices.
    - Model packaging is flexible so that adding self-made functions is achievable.

From Research to Production

Candidate Solutions

- Open Source ML Pipeline Platform.
  - Kubeflow, mlflow, airflow, TFX, etc.
- Prediction Service Framework.
  - Sagemaker, self-made restful api service, etc.

From Research to Production

Candidate Solutions

- AWS ML Experts
  - Organized 3 one-day offsite workshop together with AWS ML experts.
  - Hands-on packaging ML model into container and deploying to SageMaker.
  - Practice with cloud9 and SageMaker Notebook.
  - Consult with ML marketplace opportunity.

Production Stage

Tradeoff and Decision

	ML Pipeline Platform	Reason
Compatibility	Jenkins.	Lowest learning curve.
Portability	Jenkins.	Lowest platform transferring cost.
Scalability	Jenkins.	Jenkins can be deployed to k8s.

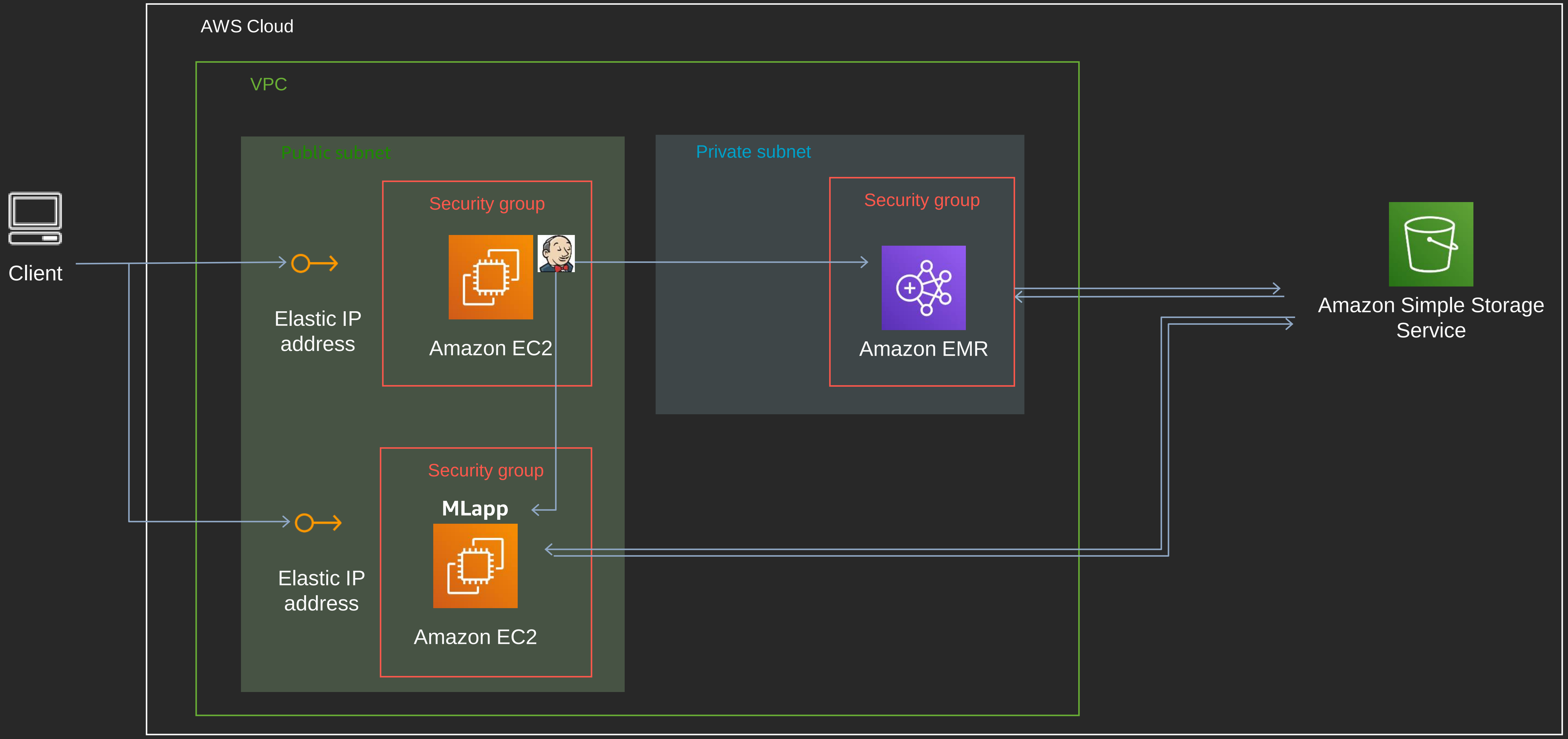
Production Stage

Tradeoff and Decision

	Prediction Service	Reason
Compatibility	Deploy customized SageMaker container on VM.	1. Easily to be integrated with Jenkins. 2. Easily to be deployed back to SageMaker.
Portability	Deploy customized SageMaker container on VM.	Easily migrate to other platforms.
Scalability	Replicate VM and add LB.	Easily scale out by LB.

Production Stage

Architecture



Production Stage

MLapp

- Definition
  - A containerized application for managing the prediction phase in machine learning product lifecycle.
- Functions
  - Official prediction requests API portal.
    - http://[IP]:[port]/api/providers/prediction/create
    - http://[IP]:[port]/api/providers/prediction/invoke
    - http://[IP]:[port]/api/providers/prediction/check
  - Monitoring
    - Tracking prediction status.
    - Recording and comparing prediction results.

Conclusion

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Conclusion

Current Status

- Project time allocation
  - 40% on research.
  - 30% on model and prediction results validation.
  - 30% on mlops and developments.
- Accuracy
  - Precision can achieve more than 80%.
- Total gender prediction pipeline execution time
  - Less than 30 minutes with monthly data.

Conclusion

Lessons Learned

- Ensure prediction quality at the top.
- Always understand current data distribution.
- Establish monitors to control data quality from the root.
- Keep production engineering work simple and reliable.

Thank you!

Young Yang
ML Specialist SA
Amazon Web Services

Hsuan Chiu
Senior Data Engineer
Data Science
VPON

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.