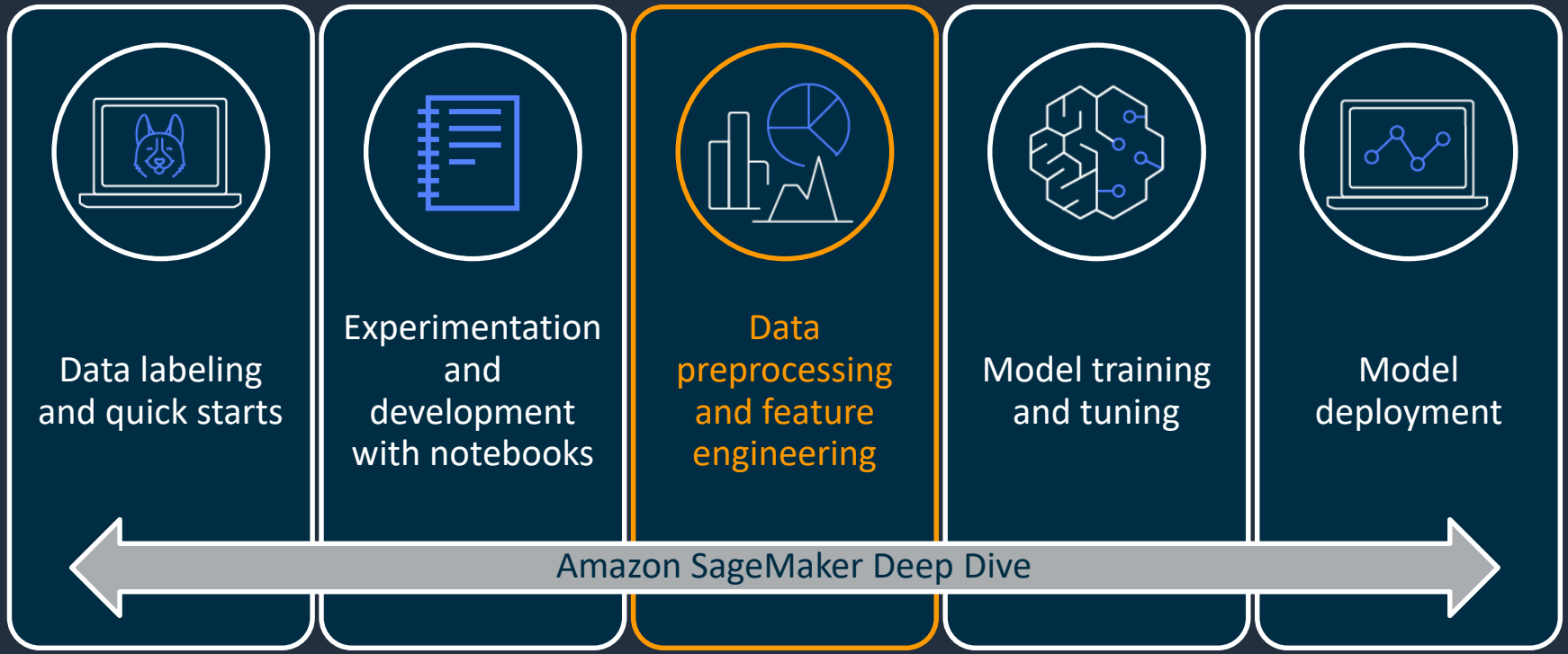
aws partnercast

Data preprocessing and feature engineering

Amazon SageMaker Deep Dive Series

Amazon SageMaker Deep Dive Series



Amazon SageMaker key benefits

Most complete, end-to-end ML service

Accelerate ML development

20+ tools covering the entire ML development lifecycle

Boost data scientist productivity

The world's first integrated development environment (IDE)

Reduce cost

Eliminate costs of writing custom integration code with integrated functionality optimized for ML

Amazon SageMaker overview

Features covered in this session

Amazon SageMaker

PREPARE

SageMaker Ground Truth
Label training data for machine learning

SageMaker Data Wrangler
Aggregate and prepare data for machine learning

SageMaker Processing
Built-in Python, BYO R/Spark

SageMaker Feature Store
Store, update, retrieve, and share features

SageMaker Clarify
Detect bias and understand model predictions

BUILD

SageMaker Studio Notebooks
Jupyter notebooks with elastic compute and sharing

Built-in and Bring your-own Algorithms
Dozens of optimized algorithms or bring your own

SageMaker Autopilot
Automatically create machine learning models with full visibility

SageMaker JumpStart
Pre-built solutions for common use cases

SageMaker Canvas
Generate accurate machine learning predictions—no code required

SageMaker Studio Lab
Learn and experiment with ML using a no-setup, free development environment

RStudio
Fully integrated development environment for machine learning

TRAIN & TUNE

Managed Training
Distributed infrastructure management

SageMaker Experiments
Capture, organize, and compare every step

Automatic Model Tuning
Hyperparameter optimization

Distributed Training Libraries
Training for large datasets and models

SageMaker Debugger
Debug and profile training runs

Managed Spot Training
Reduce training cost by 90%

Managed Training Compiler
Accelerate training of deep learning models by up to 50%

DEPLOY & MANAGE

Managed Deployment
Fully managed, ultra low latency, high throughput

Kubernetes & Kubeflow Integration
Simplify Kubernetes-based machine learning

Multi-Model Endpoints
Reduce cost by hosting multiple models per instance

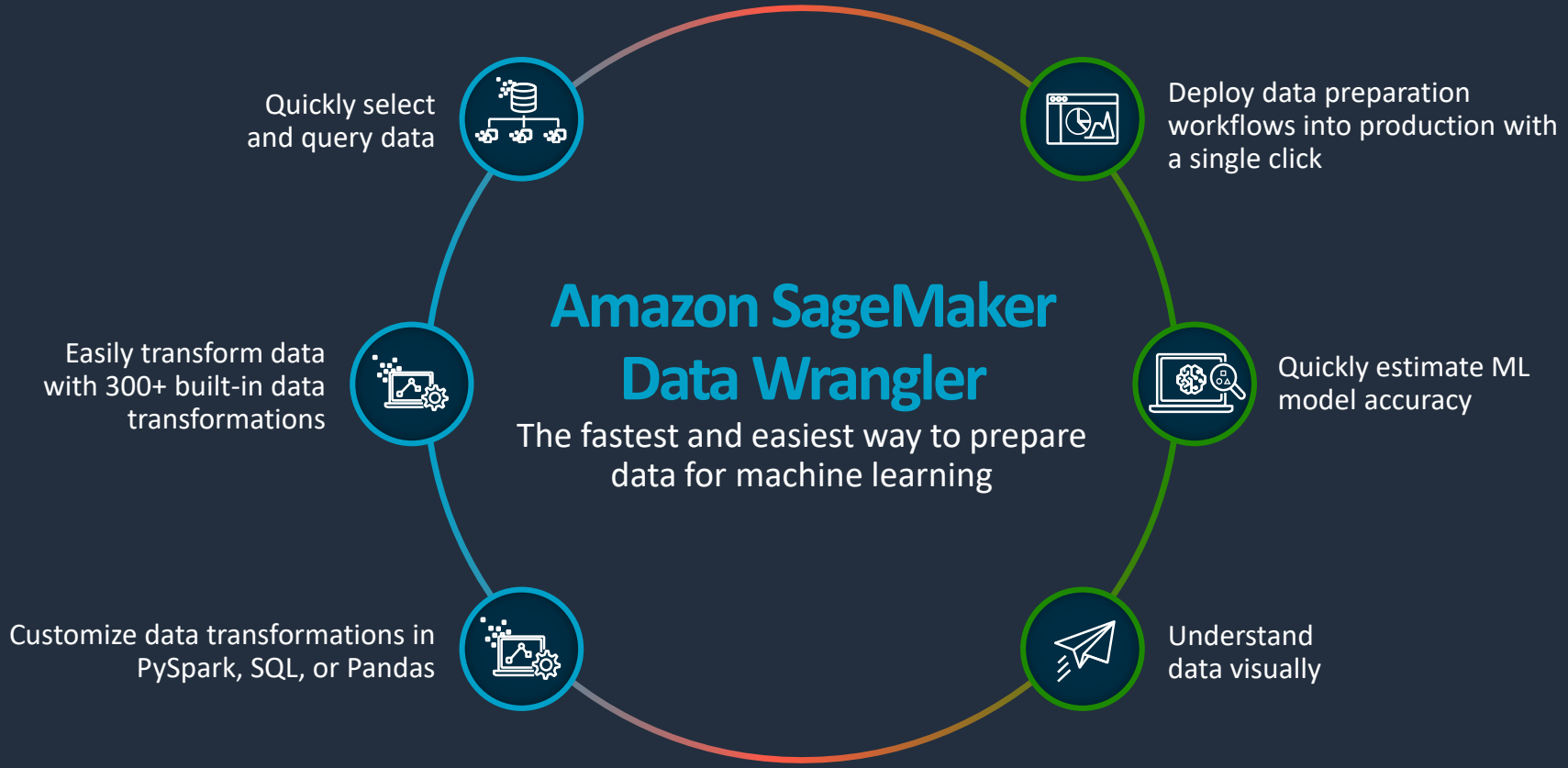
SageMaker Model Monitor
Maintain accuracy of deployed models

SageMaker Edge Manager
Manage and monitor models on edge devices

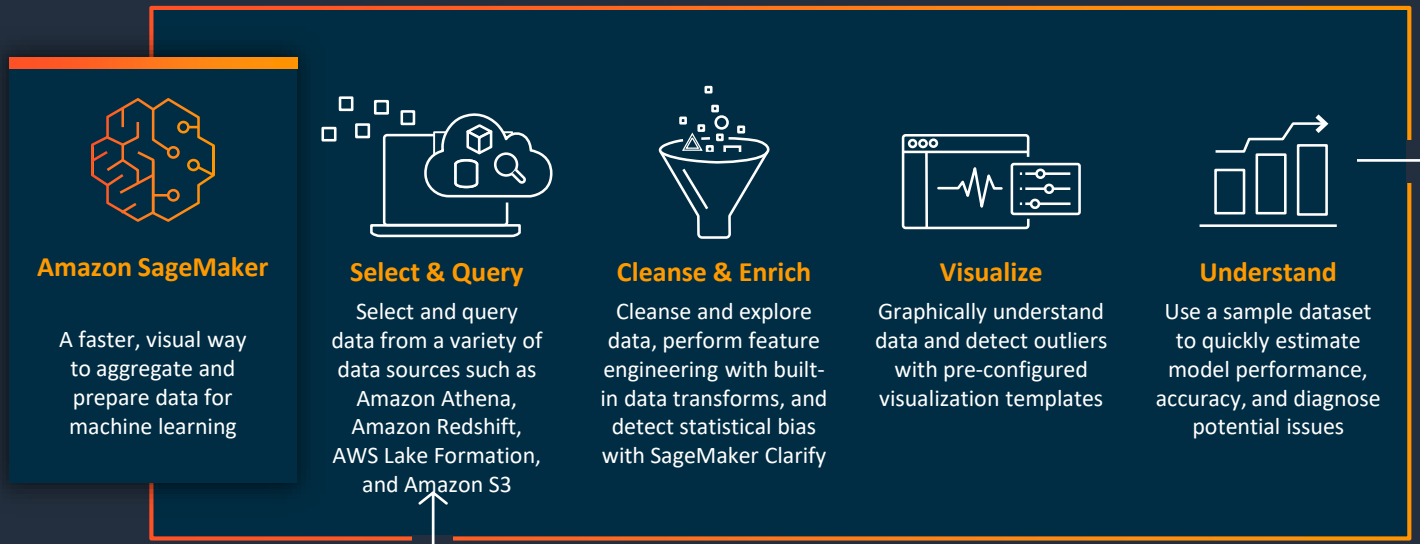
SageMaker Pipelines
Workflow orchestration and automation

SageMaker Inference Recommender
Automate load testing and optimize model performance across ML instances

SageMaker Studio
Integrated development environment (IDE) for ML



How SageMaker Data Wrangler works



Amazon SageMaker

A faster, visual way to aggregate and prepare data for machine learning

Select & Query

Select and query data from a variety of data sources such as Amazon Athena, Amazon Redshift, AWS Lake Formation, and Amazon S3

Cleanse & Enrich

Cleanse and explore data, perform feature engineering with built-in data transforms, and detect statistical bias with SageMaker Clarify

Visualize

Graphically understand data and detect outliers with pre-configured visualization templates

Understand

Use a sample dataset to quickly estimate model performance, accuracy, and diagnose potential issues

Operationalize

Export data preparation workflow to a notebook or code script to bring the workflow into production

Import data from a feature store such as Amazon SageMaker Feature Store

Integrate data preparation workflow with Amazon SageMaker Pipelines

Export prepared data into Amazon SageMaker Feature Store

Demo

SageMaker Data Wrangler

Amazon SageMaker Processing

Managed solution for data processing and model evaluation jobs

Fully managed
Achieve distributed processing for clusters

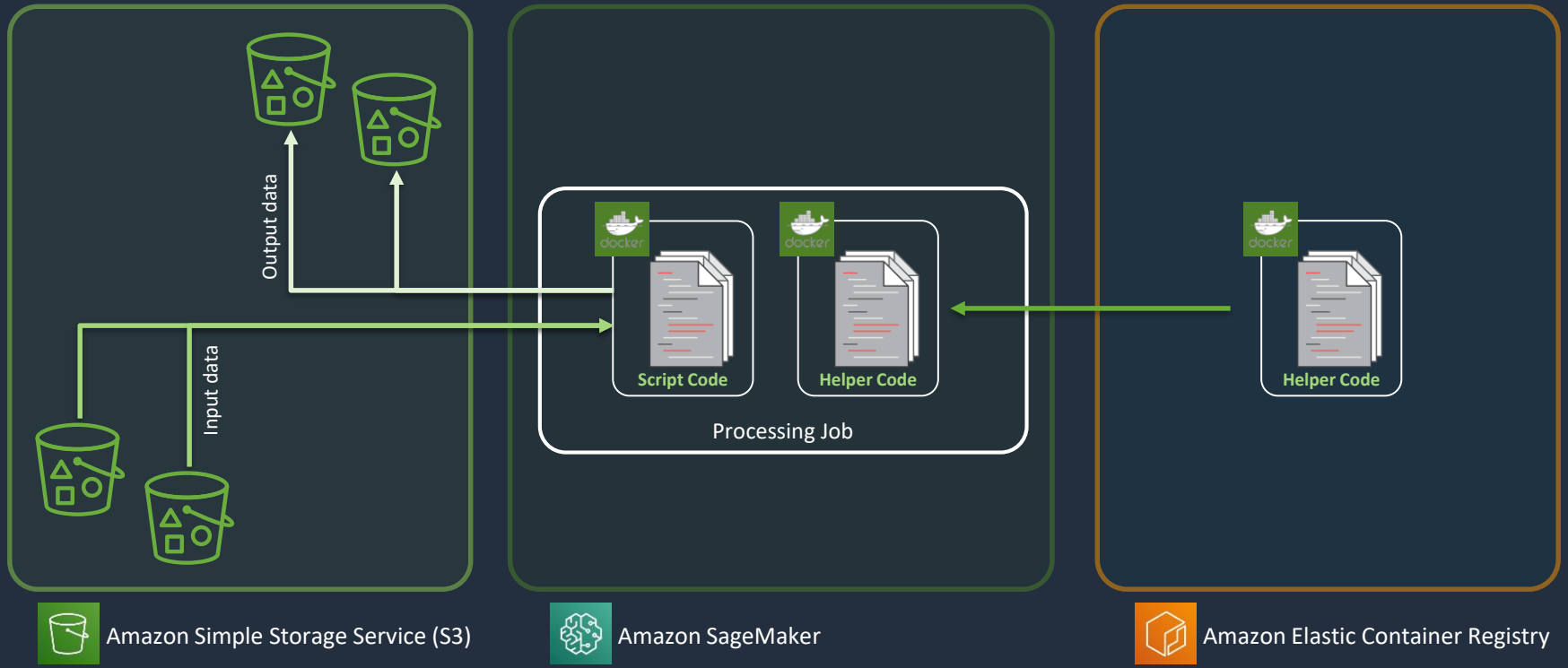
Custom processing
Bring your own script for feature engineering

Container support
Use SageMaker's built-in containers or bring your own

Security and compliance
Leverage SageMaker's security and compliance features

Automatic creation and termination
Your resources are created, configured, and terminated automatically

SageMaker Processing – Under the hood



SageMaker Processing details

Data sources

- S3
- Athena
- Redshift

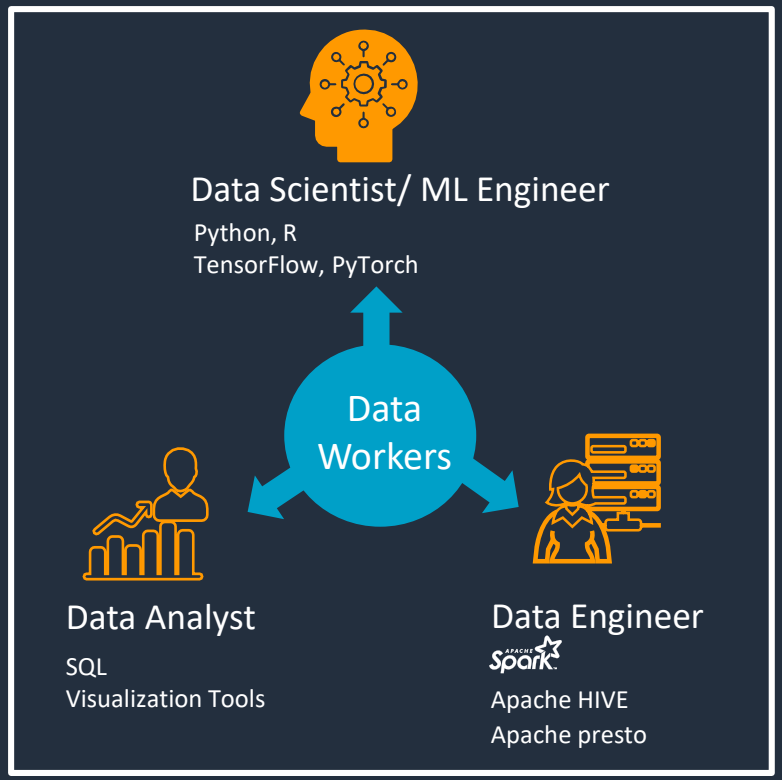
Environments

- Apache Spark
- scikit-learn
- Hugging Face
- MXNet
- PyTorch
- TensorFlow
- XGBoost
- Custom

Used for

- Data preprocessing
- Model evaluation
- Clarify (bias and explainability)
- Debugger
- Model Monitor

Unifying data platforms



- Collaboration and productivity across data analytics and ML personas
- ML Platform Engineers and admins need to build/manage resources across analytics + ML
- Security and access control needs to be consistent and transparent across analytics + ML

EMR and SageMaker notebooks

Discover, connect to, create and terminate EMR clusters (Hive, Spark and Presto)

Collaborate using Scala-based Spark and PySpark notebook kernels

Interactive analysis and processing jobs (PyHive, Spark on EMR & Local)

Bring your own image and customize notebook lifecycle configuration

Enforce fine-grained data access

Automate EMR, Glue and ML pipelines in production

SageMaker Studio Universal Notebook

EMR

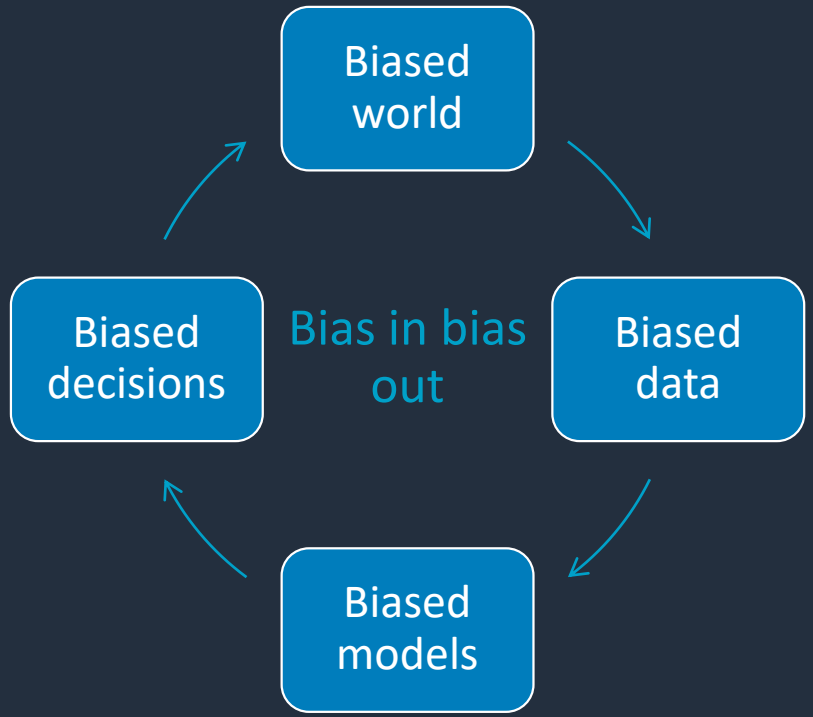
SageMaker

Demo

EMR and SageMaker

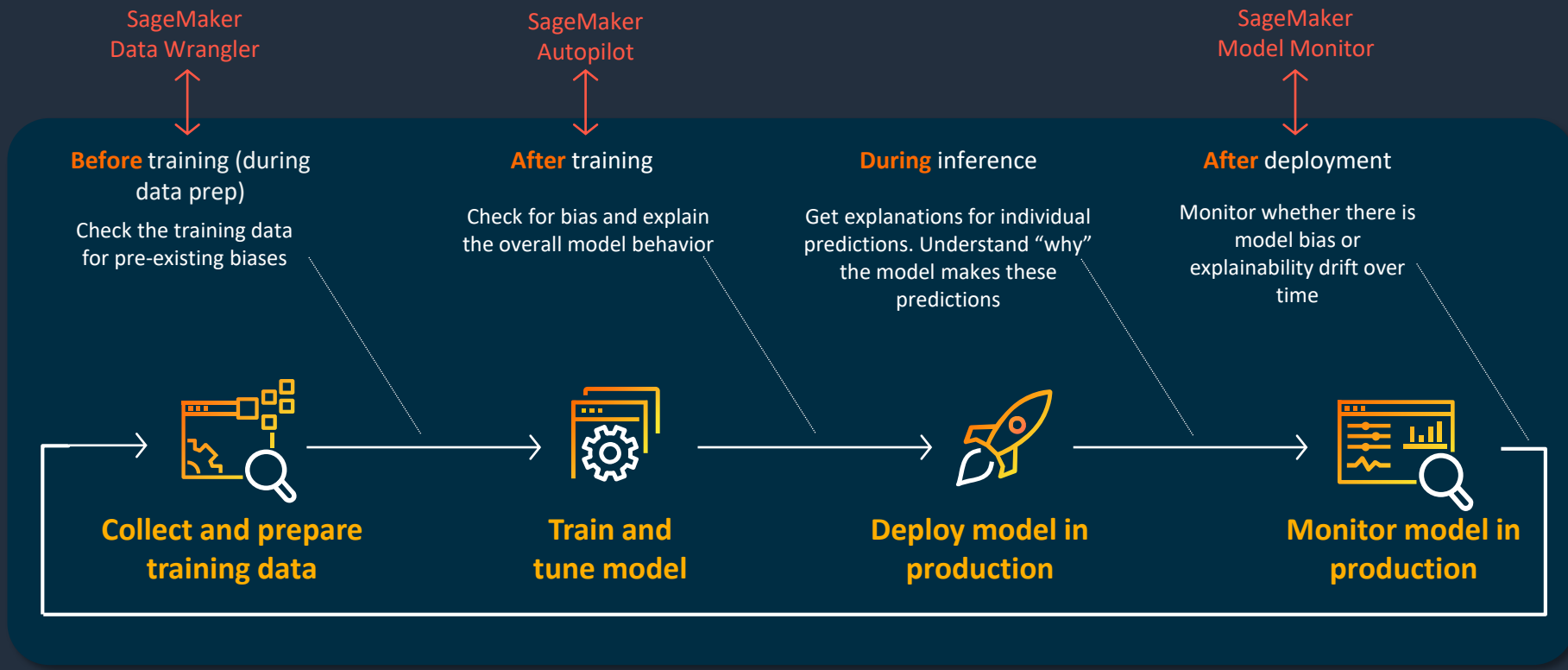
Bias is everywhere

- You cannot avoid bias
- You just have to account for it

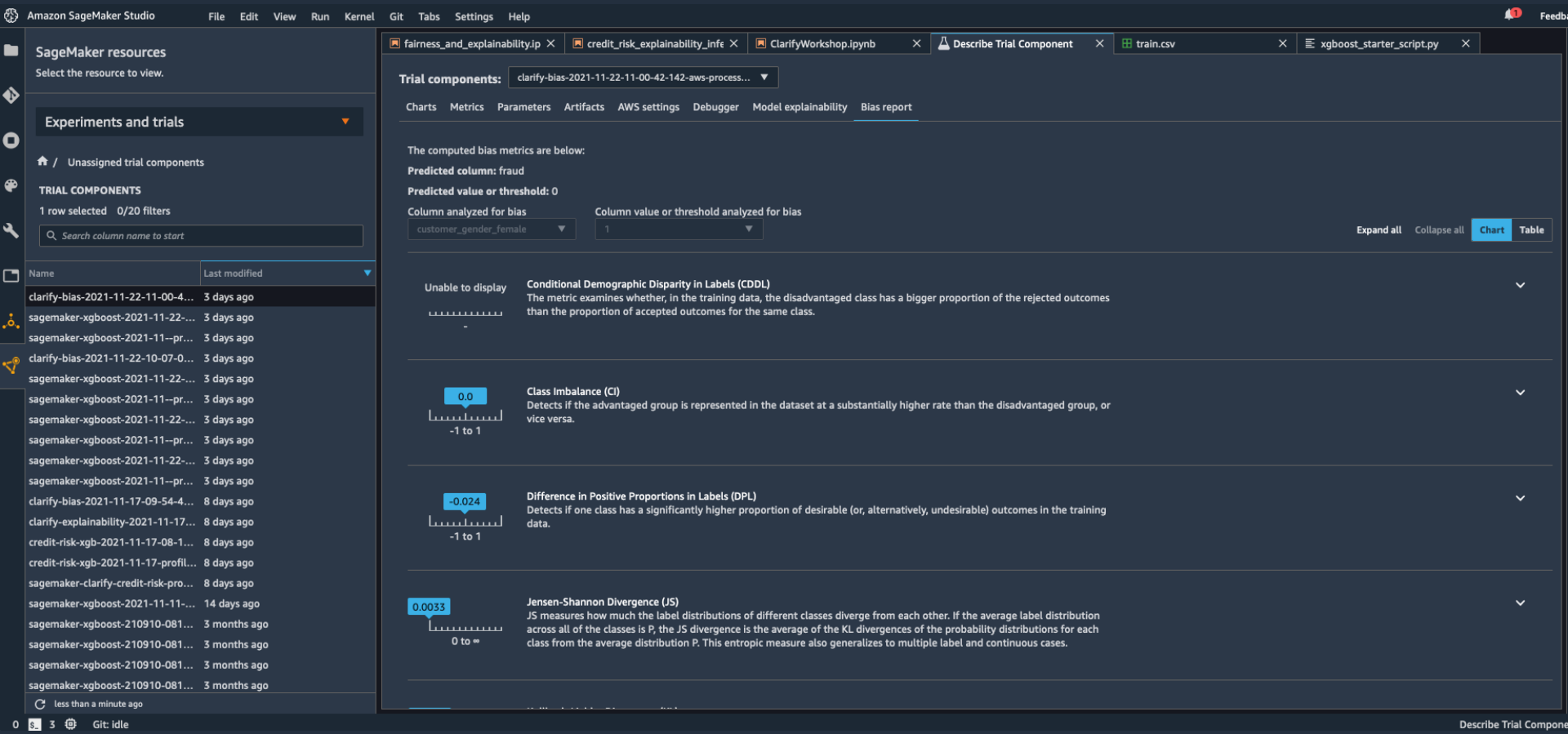


ML models can become part of **self-reinforcing feedback loops**, amplifying the existing biases of the society

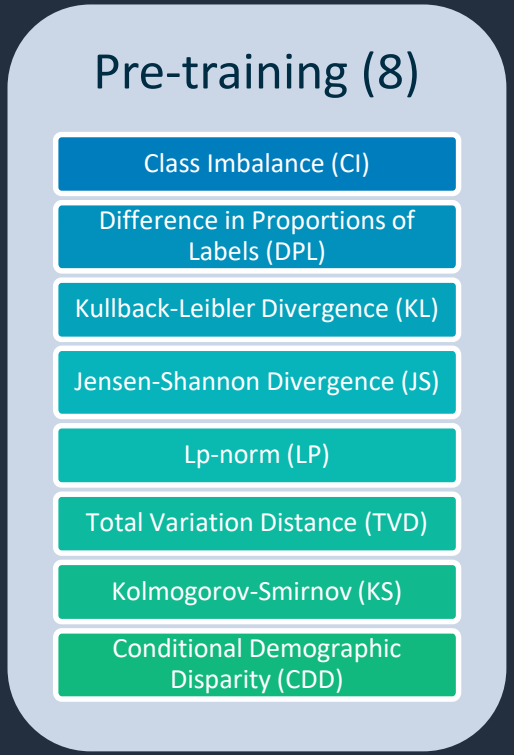
SageMaker Clarify



SageMaker Clarify bias reports



SageMaker Clarify bias metrics



- Many different concepts of fairness
- Not all fairness concepts can be satisfied simultaneously
- Human judgment is required to understand and choose which metrics are relevant for a use case

Demo

Clarify in Data Wrangler

Features covered in this session

Amazon SageMaker

PREPARE

SageMaker Ground Truth
Label training data for machine learning

SageMaker Data Wrangler
Aggregate and prepare data for machine learning

SageMaker Processing
Built-in Python, BYO R/Spark

SageMaker Feature Store
Store, update, retrieve, and share features

SageMaker Clarify
Detect bias and understand model predictions

BUILD

SageMaker Studio Notebooks
Jupyter notebooks with elastic compute and sharing

Built-in and Bring your-own Algorithms
Dozens of optimized algorithms or bring your own

SageMaker Autopilot
Automatically create machine learning models with full visibility

SageMaker JumpStart
Pre-built solutions for common use cases

SageMaker Canvas
Generate accurate machine learning predictions—no code required

SageMaker Studio Lab
Learn and experiment with ML using a no-setup, free development environment

RStudio
Fully integrated development environment for machine learning

TRAIN & TUNE

Managed Training
Distributed infrastructure management

SageMaker Experiments
Capture, organize, and compare every step

Automatic Model Tuning
Hyperparameter optimization

Distributed Training Libraries
Training for large datasets and models

SageMaker Debugger
Debug and profile training runs

Managed Spot Training
Reduce training cost by 90%

Managed Training Compiler
Accelerate training of deep learning models by up to 50%

DEPLOY & MANAGE

Managed Deployment
Fully managed, ultra low latency, high throughput

Kubernetes & Kubeflow Integration
Simplify Kubernetes-based machine learning

Multi-Model Endpoints
Reduce cost by hosting multiple models per instance

SageMaker Model Monitor
Maintain accuracy of deployed models

SageMaker Edge Manager
Manage and monitor models on edge devices

SageMaker Pipelines
Workflow orchestration and automation

SageMaker Inference Recommender
Automate load testing and optimize model performance across ML instances

SageMaker Studio
Integrated development environment (IDE) for ML

Resources

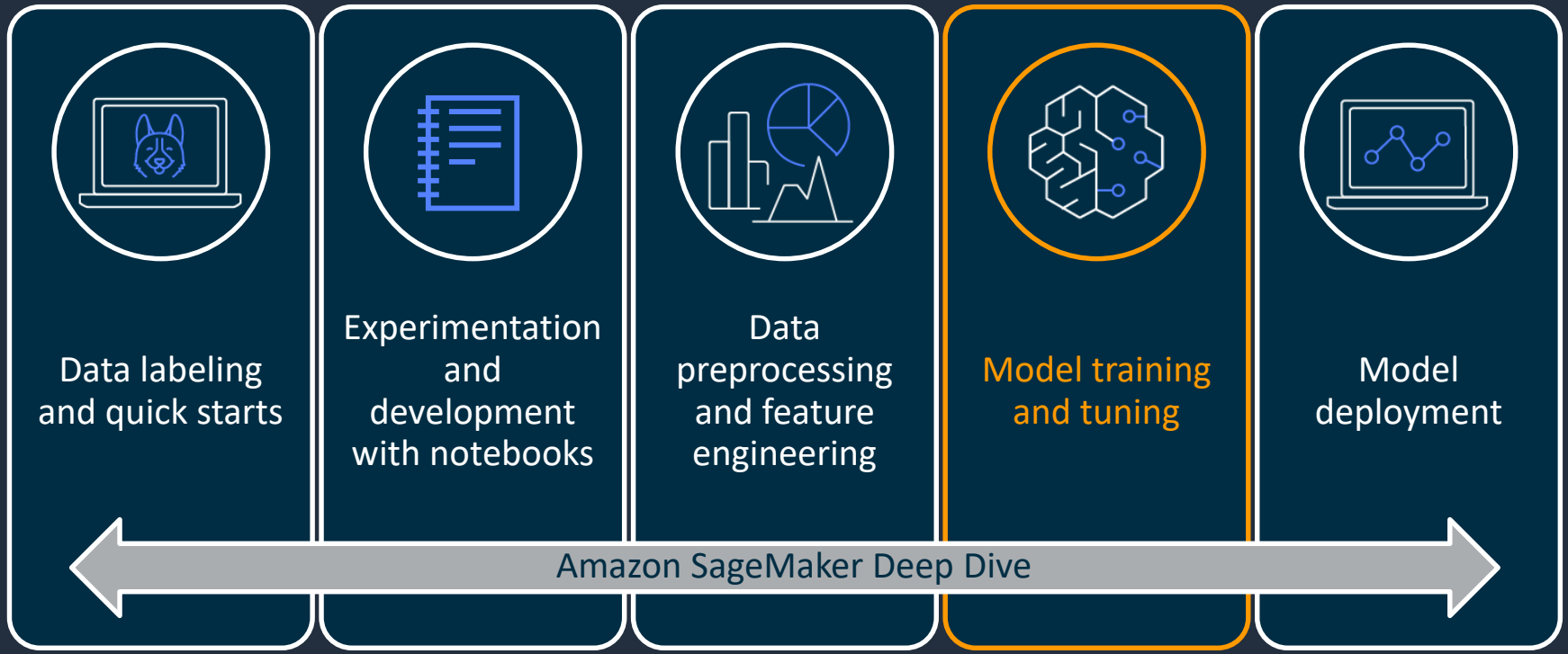
[Documentation] Amazon SageMaker Data Wrangler
[Workshop] Amazon SageMaker Data Wrangler

[Documentation] Amazon SageMaker processing jobs
[Code Samples] Amazon SageMaker processing
[Tutorials] Amazon SageMaker processing
[Workshop] Amazon SageMaker processing

[Documentation] Prepare Data at Scale with Studio Notebooks
[Blog] Create and manage Amazon EMR Clusters from SageMaker Studio to run interactive Spark and ML workloads

[Documentation] Detect pretraining data bias
[Workshop] Bias and explainability

Join us in the next session



aws partnercast

Thank you!

Australia & New Zealand
Brad Ryan
AI/ML Partner Development Specialist
brdryn@amazon.com

Sara van de Moosdijk
AI/ML Partner Solutions Architect
sarmoosd@amazon.com

ASEAN
Ling Chang
AI/ML Partner Development Specialist
linglych@amazon.com

Vasileios Vonikakis
AI/ML Partner Solutions Architect
vonikakv@amazon.com

India
Ankit Kandoi
Data Analytics & ML Partner Development Specialist
akkandoi@amazon.com

aws training and certification