

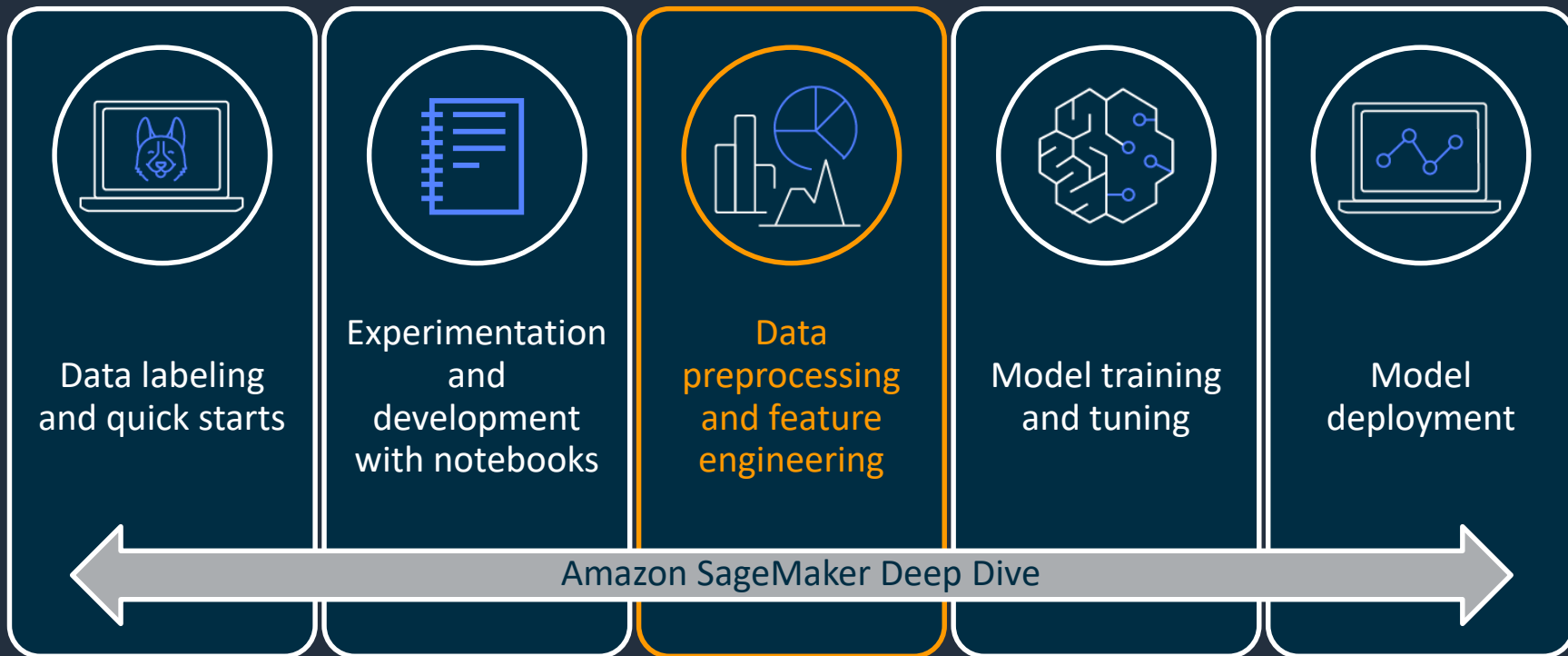


Data preprocessing and feature engineering

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

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

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

Amazon SageMaker Data Wrangler

The fastest and easiest way to prepare
data for machine learning

Quickly select
and query data



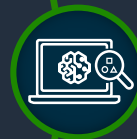
Deploy data preparation
workflows into production with
a single click



Easily transform data
with 300+ built-in data
transformations



Quickly estimate ML
model accuracy



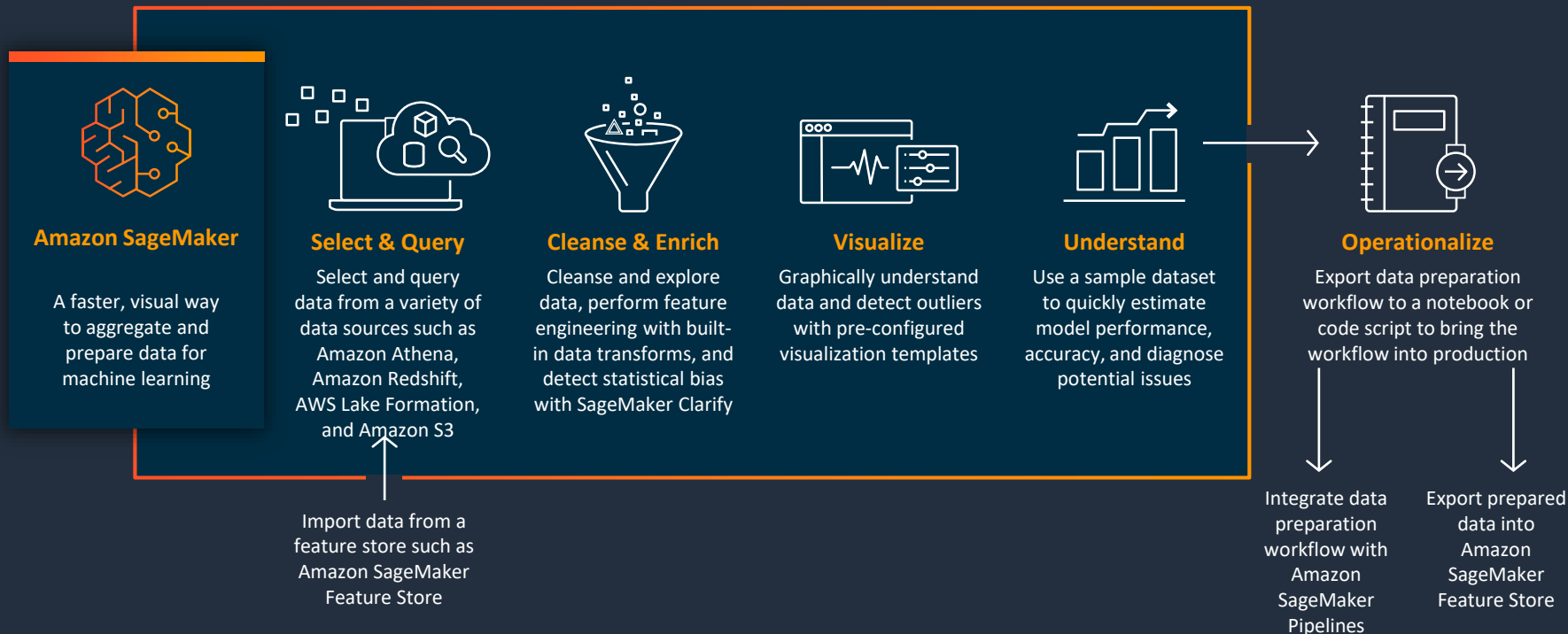
Customize data transformations in
PySpark, SQL, or Pandas



Understand
data visually



How SageMaker Data Wrangler works



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

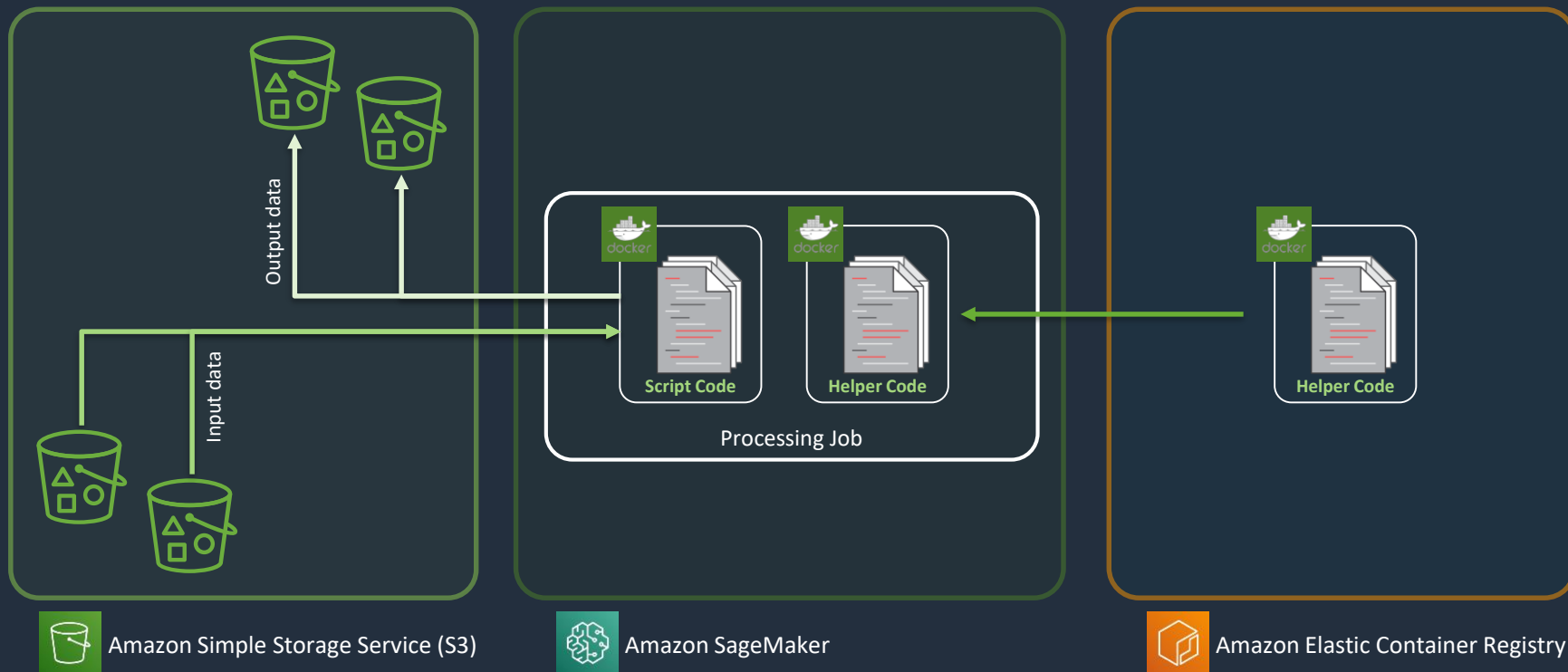
Leverage SageMaker's security and compliance features



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

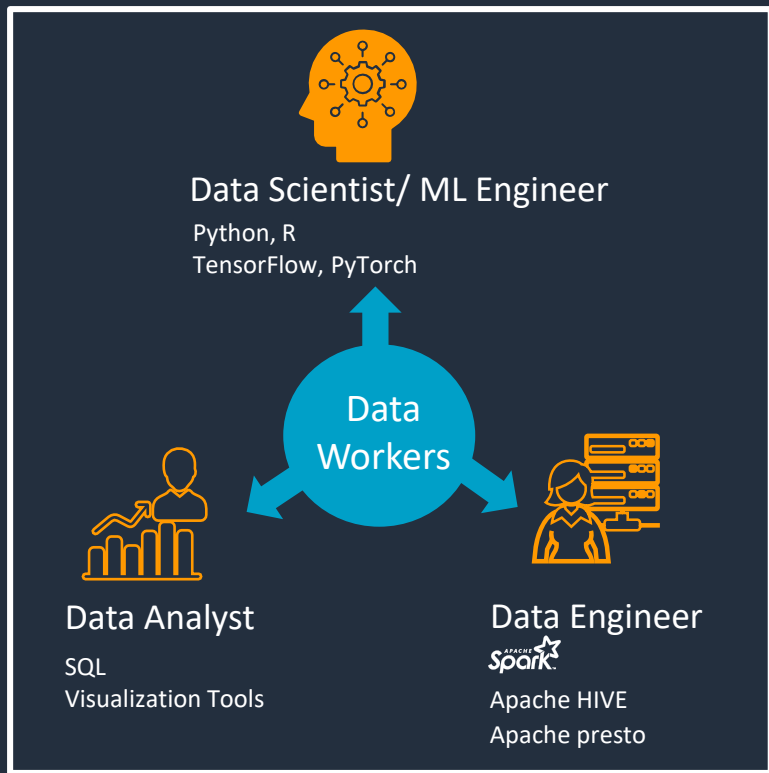
- Apache Spark
- scikit-learn
- Hugging Face
- MXNet
- PyTorch
- TensorFlow
- XGBoost
- Custom



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)



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



Enforce fine-grained data access



Collaborate using Scala-based Spark and PySpark notebook kernels



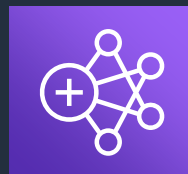
Bring your own image and customize notebook lifecycle configuration



Automate EMR, Glue and ML pipelines in production



SageMaker Studio
Universal Notebook



EMR



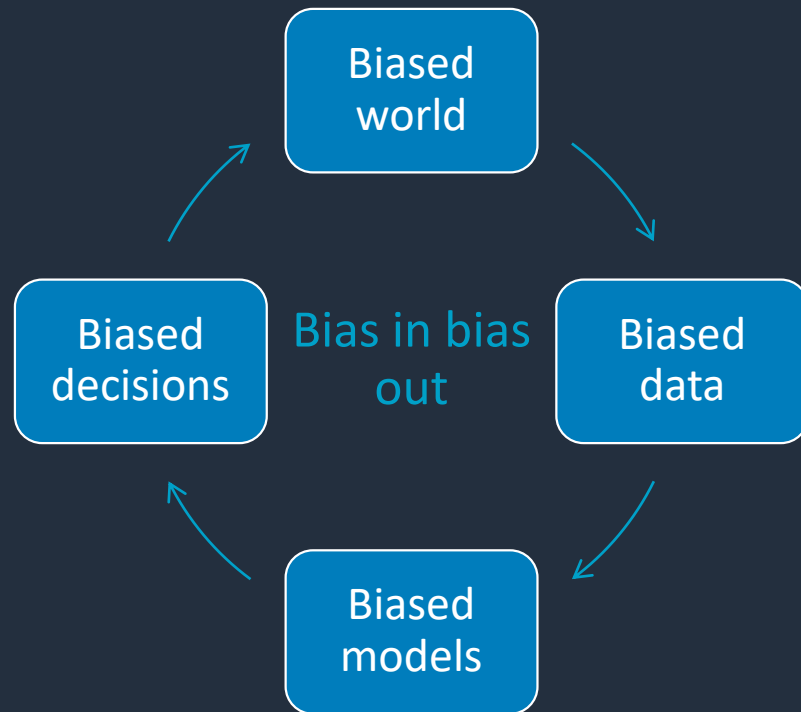
SageMaker

Demo

EMR and SageMaker

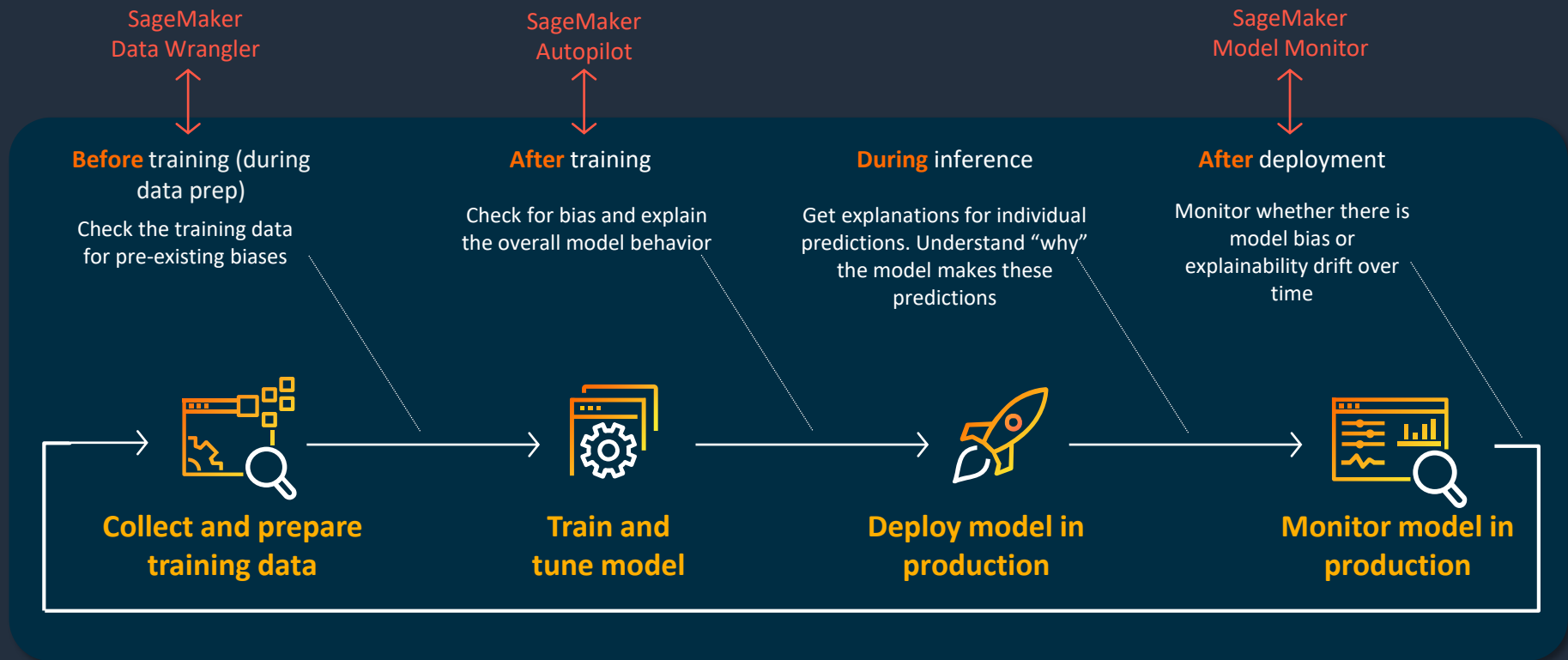
Bias is everywhere

- 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

Amazon SageMaker Studio

File Edit View Run Kernel Git Tabs Settings Help

SageMaker resources

Select the resource to view.

Experiments and trials

▼

🏠 / Unassigned trial components

TRIAL COMPONENTS

1 row selected 0/20 filters

Name	Last modified
clarify-bias-2021-11-22-11-00-4...	3 days ago
sagemaker-xgboost-2021-11-22-...	3 days ago
sagemaker-xgboost-2021-11--pr...	3 days ago
clarify-bias-2021-11-22-10-07-0...	3 days ago
sagemaker-xgboost-2021-11-22-...	3 days ago
sagemaker-xgboost-2021-11--pr...	3 days ago
sagemaker-xgboost-2021-11-22-...	3 days ago
sagemaker-xgboost-2021-11-22-...	3 days ago
sagemaker-xgboost-2021-11--pr...	3 days ago
sagemaker-xgboost-2021-11-22-...	3 days ago
sagemaker-xgboost-2021-11--pr...	3 days ago
clarify-bias-2021-11-17-09-54-4...	8 days ago
clarify-explainability-2021-11-17...	8 days ago
credit-risk-xgb-2021-11-17-08-1...	8 days ago
credit-risk-xgb-2021-11-17-profil...	8 days ago
sagemaker-clarify-credit-risk-pro...	8 days ago
sagemaker-xgboost-2021-11-11-...	14 days ago
sagemaker-xgboost-210910-081...	3 months ago
sagemaker-xgboost-210910-081...	3 months ago
sagemaker-xgboost-210910-081...	3 months ago
sagemaker-xgboost-210910-081...	3 months ago
🕒 less than a minute ago	

Trial components: clarify-bias-2021-11-22-11-00-42-142-aws-process...

Charts Metrics Parameters Artifacts AWS settings Debugger Model explainability Bias report

The computed bias metrics are below:

Predicted column: fraud

Predicted value or threshold: 0

Column analyzed for bias: customer_gender_female

Column value or threshold analyzed for bias: 1

Expand all Collapse all Chart Table

Unable to display

Conditional Demographic Disparity in Labels (CDDL)

The metric examines whether, in the training data, the disadvantaged class has a bigger proportion of the rejected outcomes than the proportion of accepted outcomes for the same class.

0.0

Class Imbalance (CI)

Detects if the advantaged group is represented in the dataset at a substantially higher rate than the disadvantaged group, or vice versa.

-0.024

Difference in Positive Proportions in Labels (DPL)

Detects if one class has a significantly higher proportion of desirable (or, alternatively, undesirable) outcomes in the training data.

0.0033

Jensen-Shannon Divergence (JS)

JS measures how much the label distributions of different classes diverge from each other. If the average label distribution across all of the classes is P, the JS divergence is the average of the KL divergences of the probability distributions for each class from the average distribution P. This entropic measure also generalizes to multiple label and continuous cases.

0

3

Git: Idle

Describe Trial Component

SageMaker Clarify bias metrics

Pre-training (8)

Class Imbalance (CI)

Difference in Proportions of
Labels (DPL)

Kullback-Leibler Divergence (KL)

Jensen-Shannon Divergence (JS)

Lp-norm (LP)

Total Variation Distance (TVD)

Kolmogorov-Smirnov (KS)

Conditional Demographic
Disparity (CDD)

- 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](#)

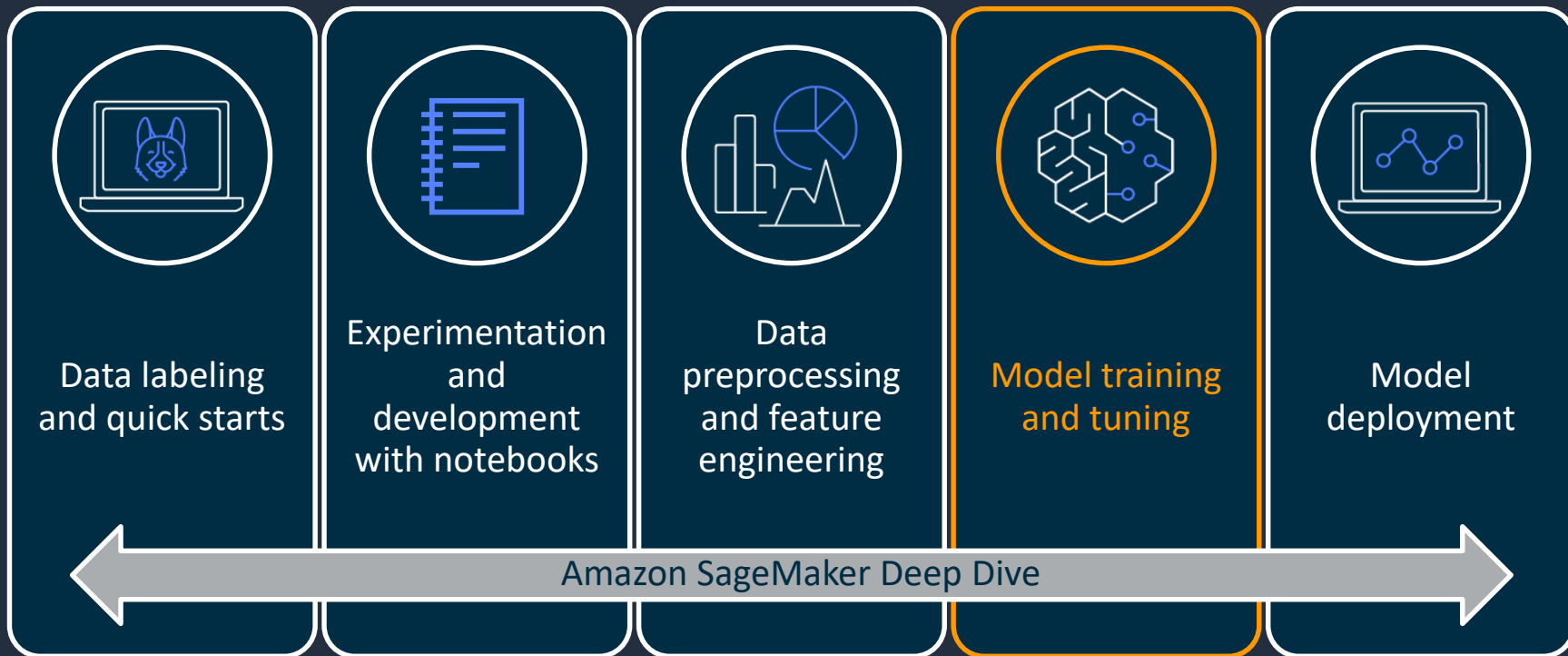
[Documentation] [Prepare Data at Scale with Studio Notebooks](#)

[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





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