Mastering Amazon SageMaker

Data Collection & Preparation with Amazon SageMaker

Nithin Reddy Cheruku — Sr AI/ML Partner Solutions Architect

Mani Khanuja — Sr AI/ML Specialist Solutions Architect

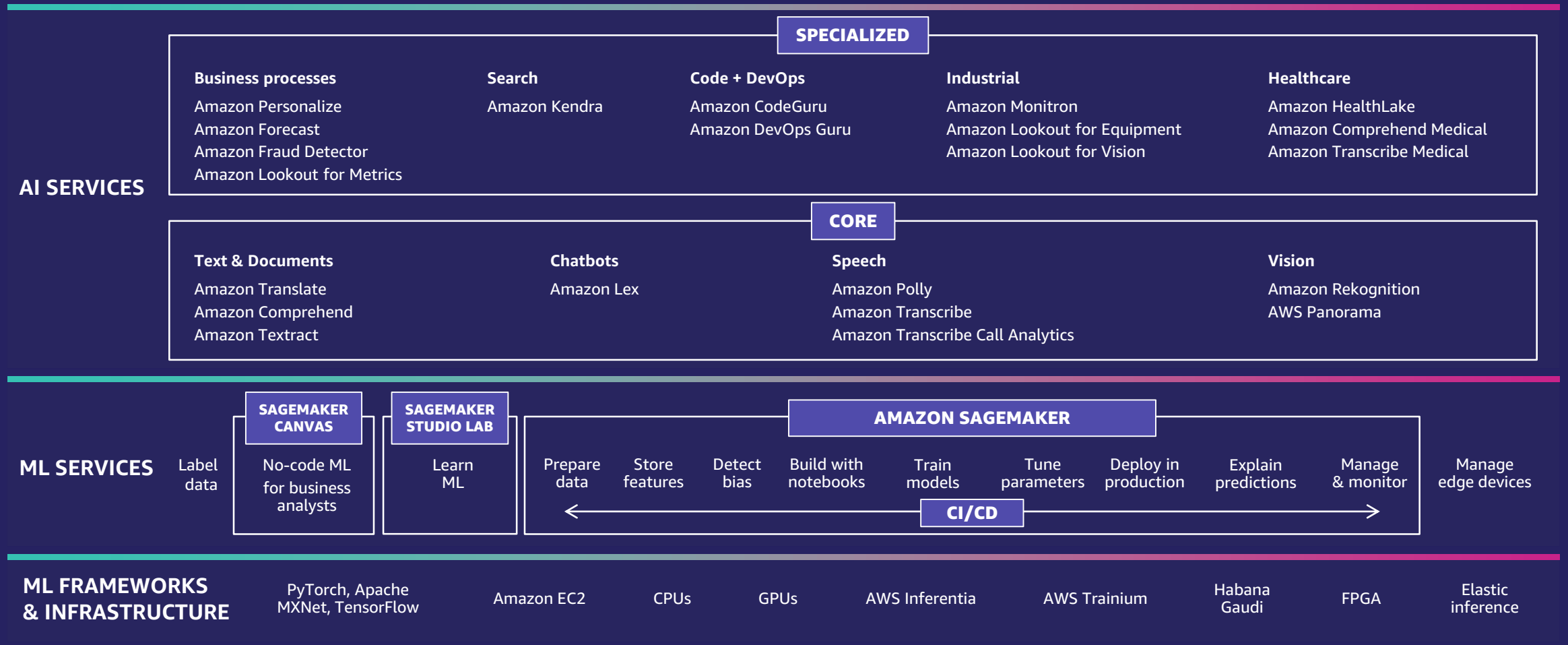
Agenda

- Overview of Amazon SageMaker
- Integrated ML tools in a single interface
- Data Labeling using Amazon Sagemaker GroundTruth
- Module 1 – SageMaker for Feature Engineering
  - SageMaker Data Wrangler
  - SageMaker Processing
  - SageMaker Feature Store
  - Demo
- Q&A
- Survey

aws machine learning

The AWS AI/ML stack

Broadest and most complete set of machine learning capabilities

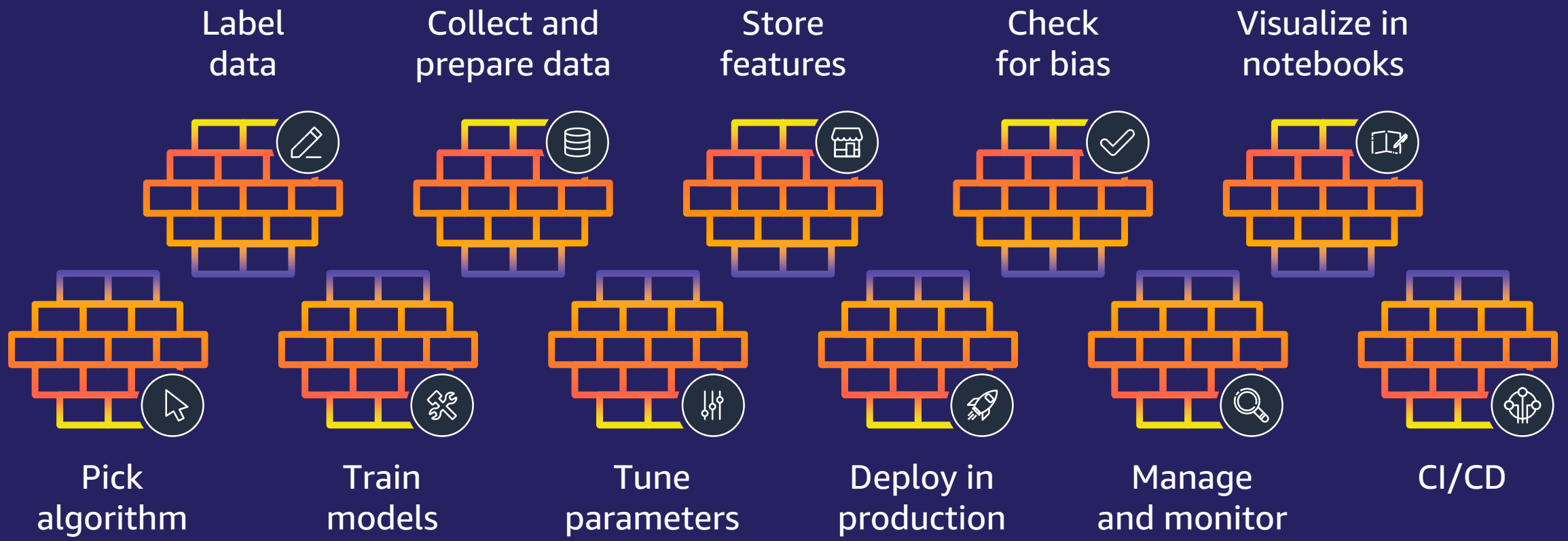


aws machine learning

- # Overview of Amazon SageMaker

aws machine learning

Machine learning development is complex and costly



Amazon SageMaker: Built to make ML more accessible

MODEL MANAGEMENT FOR EDGE DEVICES

Amazon SageMaker overview

Prepare

SageMaker Ground Truth
Label training data for machine learning

SageMaker Data Wrangler NEW
Aggregate and prepare data for machine learning

SageMaker Processing
Built-in Python, BYO R/Spark

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

SageMaker Clarify NEW
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

Local mode
Test and prototype on your local machine

SageMaker Autopilot
Automatically create machine learning models with full visibility

SageMaker JumpStart NEW
Prebuilt solutions for common use cases

Train and tune

Managed training
Distributed infrastructure management

SageMaker Experiments
Capture, organize, and compare every step

Automatic model tuning
Hyperparameter optimization

Distributed training libraries NEW
Training for large datasets and models

SageMaker Debugger NEW
Debug and profile training runs

Managed spot training
Reduce training cost by 90%

Deploy and 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 NEW
Manage and monitor models on edge devices

SageMaker Pipelines NEW
Workflow orchestration and automation

SageMaker Studio | RStudio
Integrated development environment (IDE) for ML

MLOps: Pipelines | Projects | Model Registry

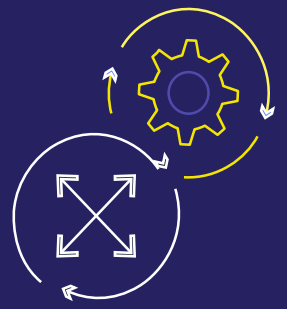
Canvas

aws machine learning

Data labelling using SageMaker GroundTruth

aws machine learning

Data labeling challenges



Massive scale

ML models need large, labeled datasets

High accuracy

ML models depend on accurately labeled data

Time consuming

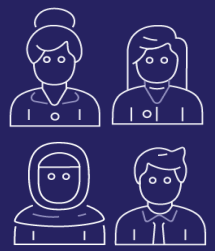
Building training datasets takes up to 80% of a data scientist's time

Amazon SageMaker Ground Truth capabilities

Fully managed and secure

30+ labeling workflows

Assistive tooling, Auto-labeling, Consensus

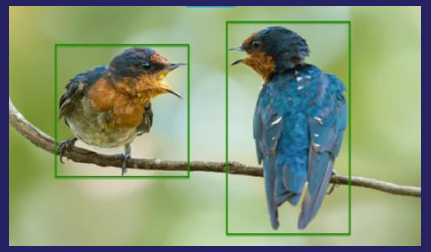


Your Choice of workforce

Pay as you go, Volume discounts

30+ data labeling workflows

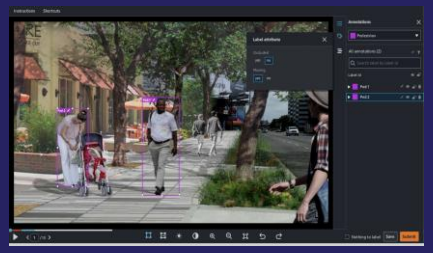
Image classification



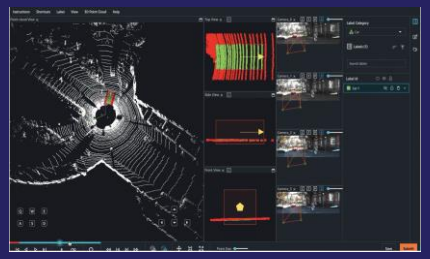
Bounding boxes

Semantic segmentation

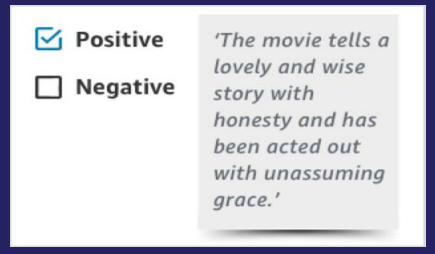
Label Verification



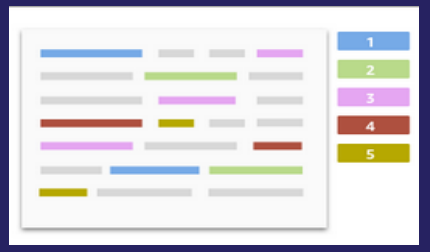
Video



LIDAR 3D Point Cloud



Text classification



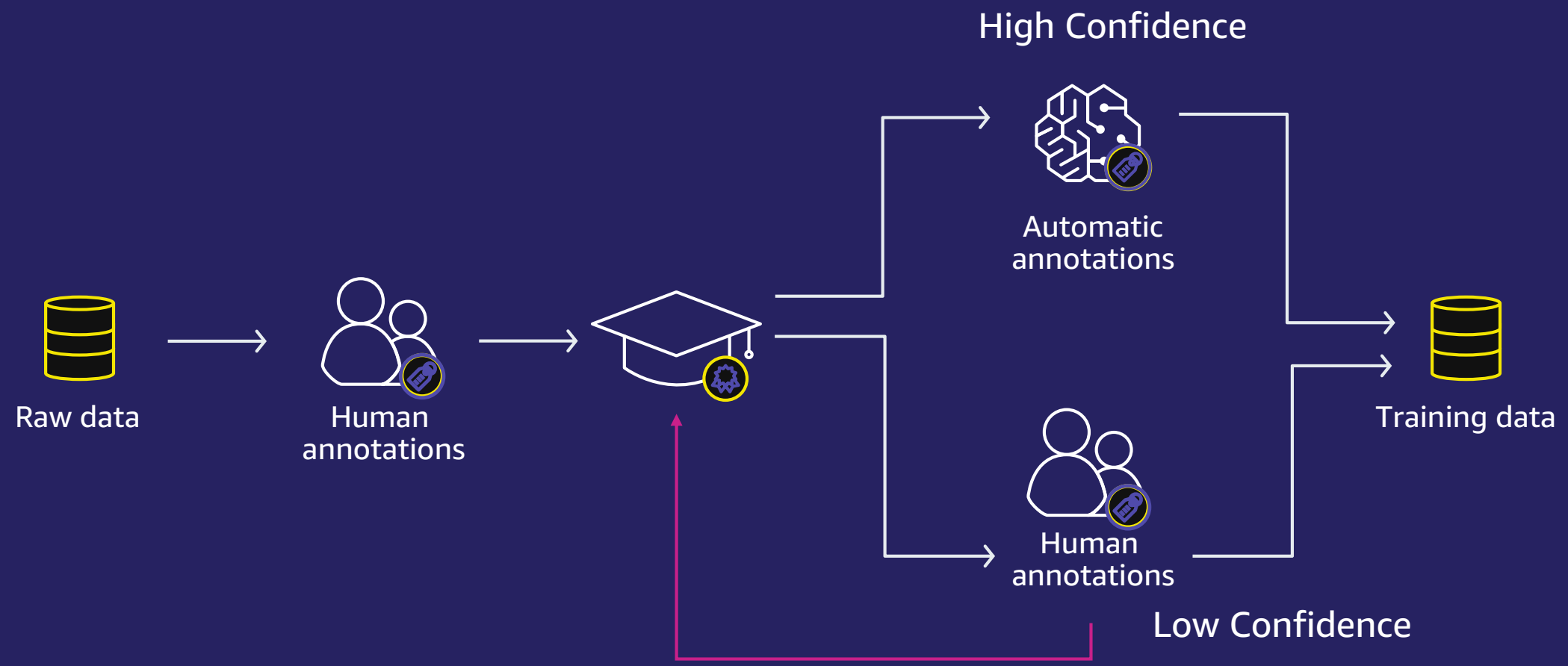
Named entity recognition

Custom(25 Templates)

- Key point
- Line
- Instance Segmentation

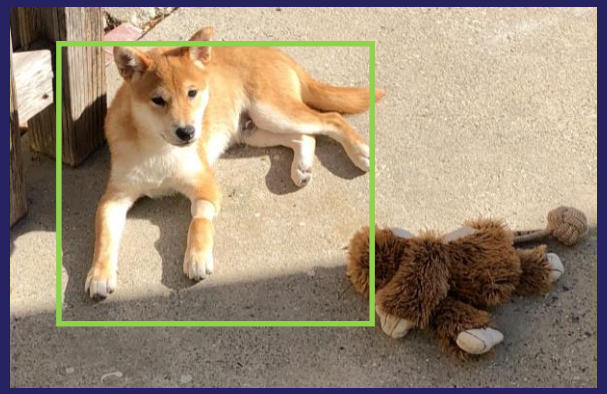
aws machine learning

Auto-labeling to reduce labeling cost by up to 70%



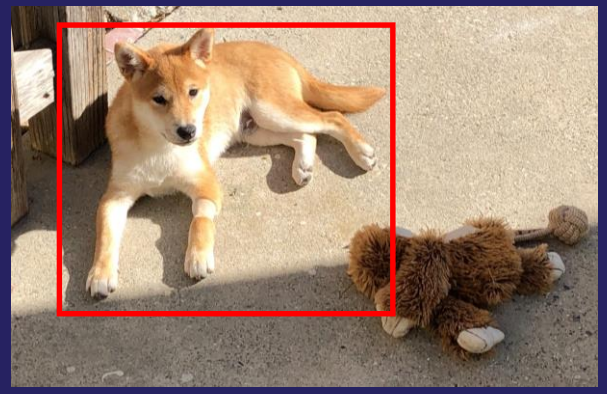
Annotation consolidation to improve label quality by 20%

Bob



Shirley

Dan



Amazon SageMaker Ground Truth uses annotation consolidation to output a single high-accuracy label

aws machine learning

Your choice of workforce options

Mechanical Turk

An on-demand 24x7 workforce of over 500,000 independent contractors worldwide, powered by Amazon Mechanical Turk

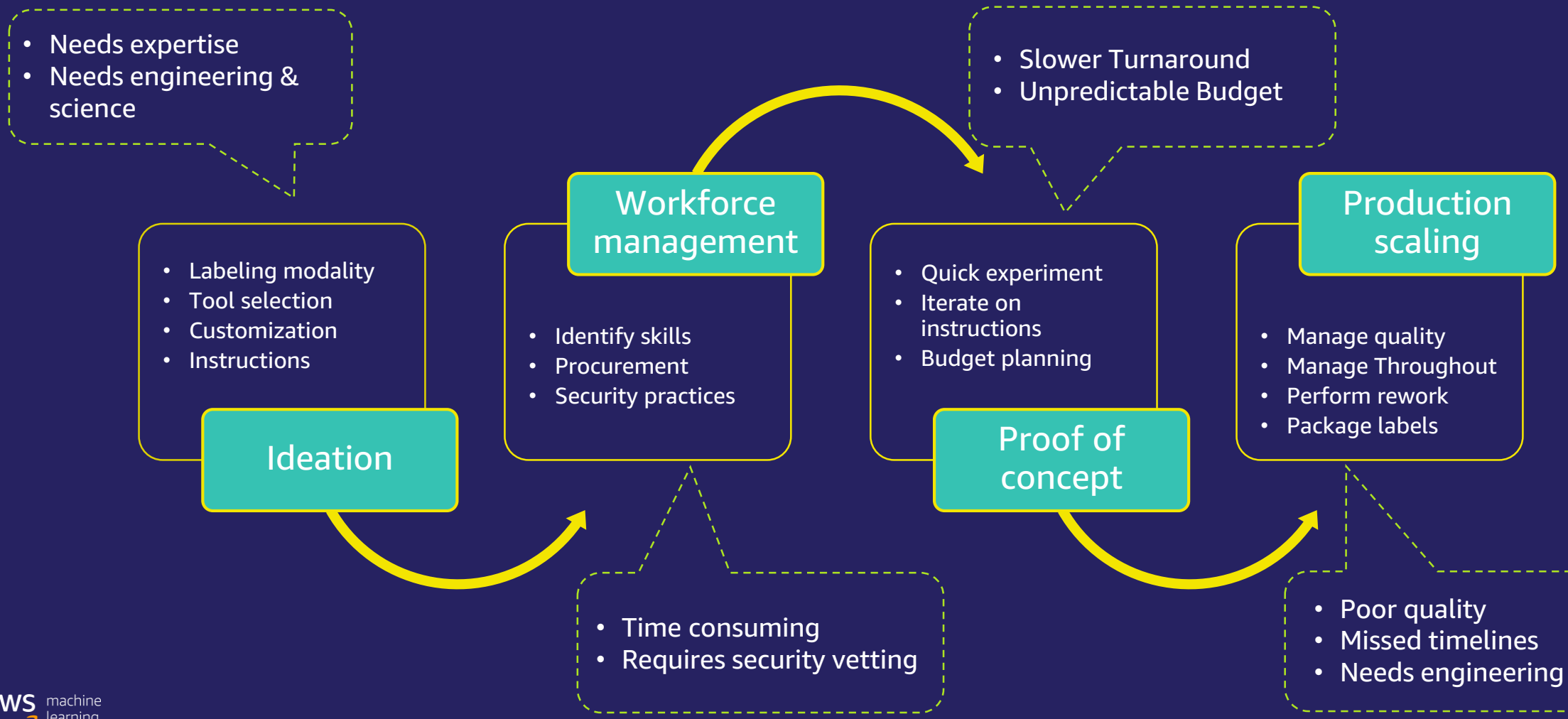
Private

A team of workers that you have sourced yourself, including your own employees or contractors, for handling data that needs to stay within your organization

Vendor

A curated list of third-party vendors that specialize in providing data labeling services, available via the AWS Marketplace

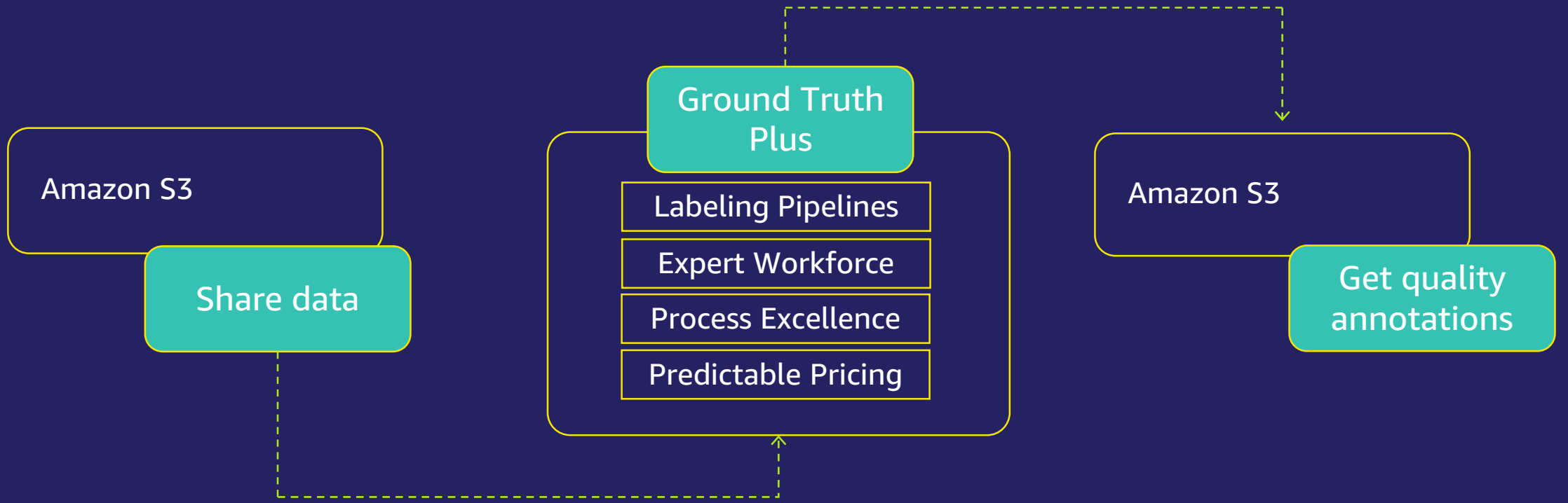
Challenges: operating data labeling projects



aws machine learning

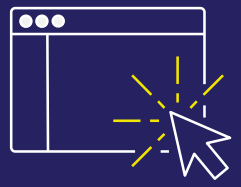
SageMaker Ground Truth Plus

Turn key Data labeling as a service



aws machine learning

SageMaker Ground Truth Plus benefits



Turnkey solution

Guaranteed throughput

Higher quality and feedback loop

Predictable and lower cost

Tracking & visibility

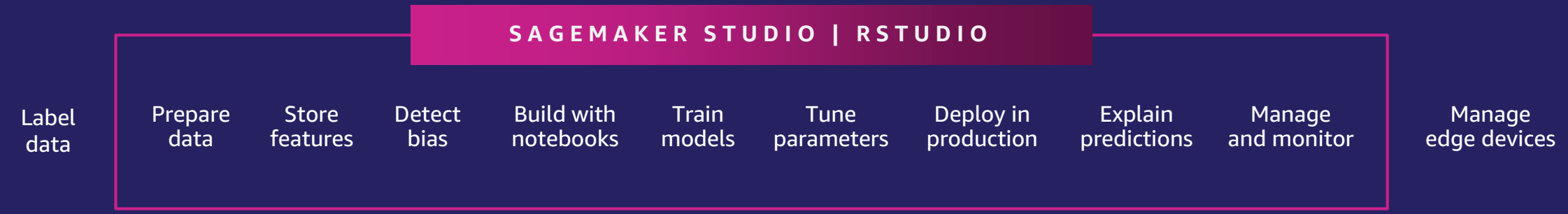
Data security and operational excellence

aws machine learning

Integrated ML tools in a single interface

aws machine learning

Amazon SageMaker brings tools for every step of the ML lifecycle under one unified visual user interface

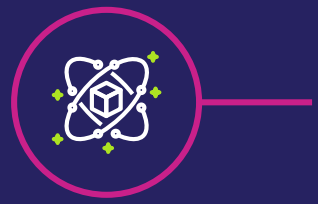


aws machine learning

Amazon SageMaker Studio Notebook

Perform data engineering, analytics, and ML workflows in one notebook

Connect with Amazon EMR, Amazon S3, and more



Interactively access, transform, and analyze a wide range of data

Build, train, and deploy models using your preferred framework

aws machine learning

Amazon SageMaker Notebooks
Fast-start sharable notebooks

Easy access with Single Sign-On (SSO)

Access your notebooks in seconds

Fully managed and secure

Administrators manage access and permissions

Fast setup

Start your notebooks without spinning up compute resources

Easy collaboration

Share notebooks with a single click

Flexible

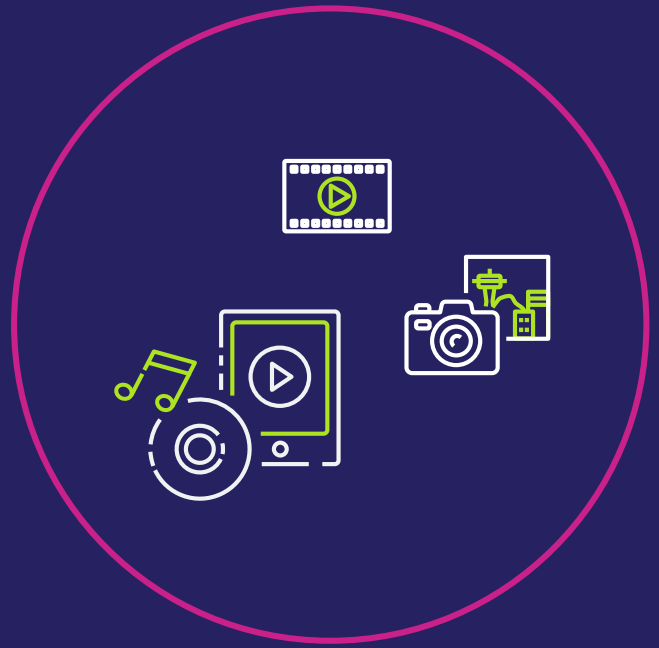
Dial up or down compute resources (coming soon)

Session I

Amazon SageMaker for Feature Engineering

aws machine learning

Structured



Unstructured

aws machine learning

Amazon SageMaker Data Wrangler

No-code data preparation

Single visual interface for common data prep techniques

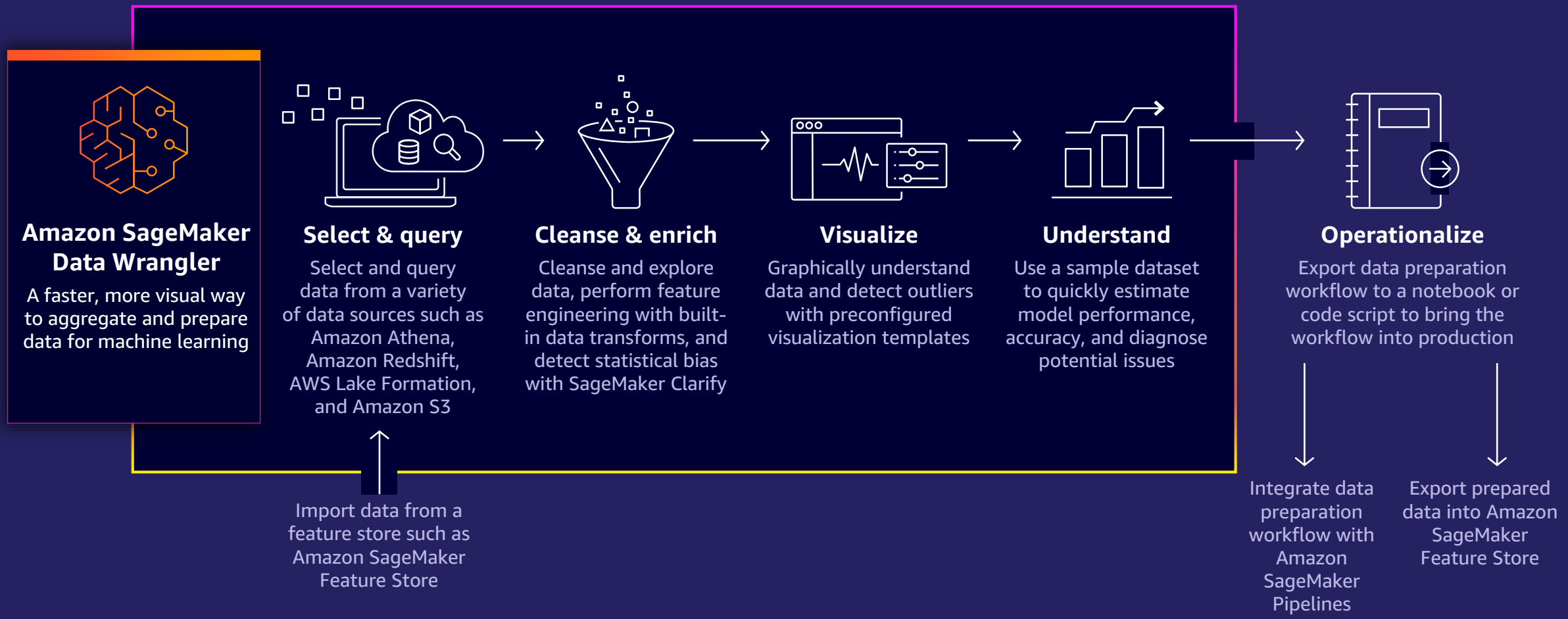
Select data from multiple sources

300+ built-in transformations to prepare data without writing code

aws machine learning

Amazon SageMaker Data Wrangler

- How it works



Amazon SageMaker Processing Job

Managed solution for data processing and model evaluation jobs

Achieve distributed processing for clusters

Bring your own script for feature engineering

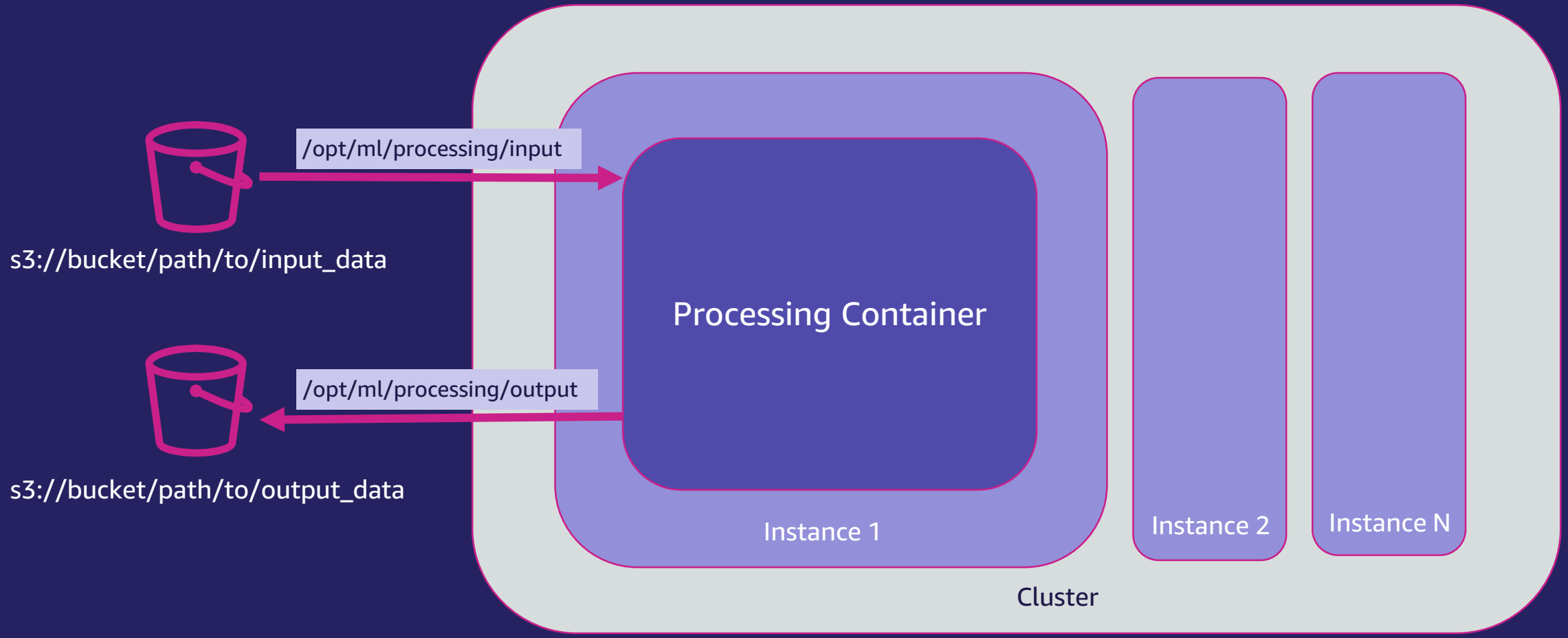
Use SageMaker's built-in containers or bring your own

Leverage SageMaker's security and compliance features

Your resources are created, configured, and terminated automatically

aws machine learning

Data Processing on Amazon SageMaker

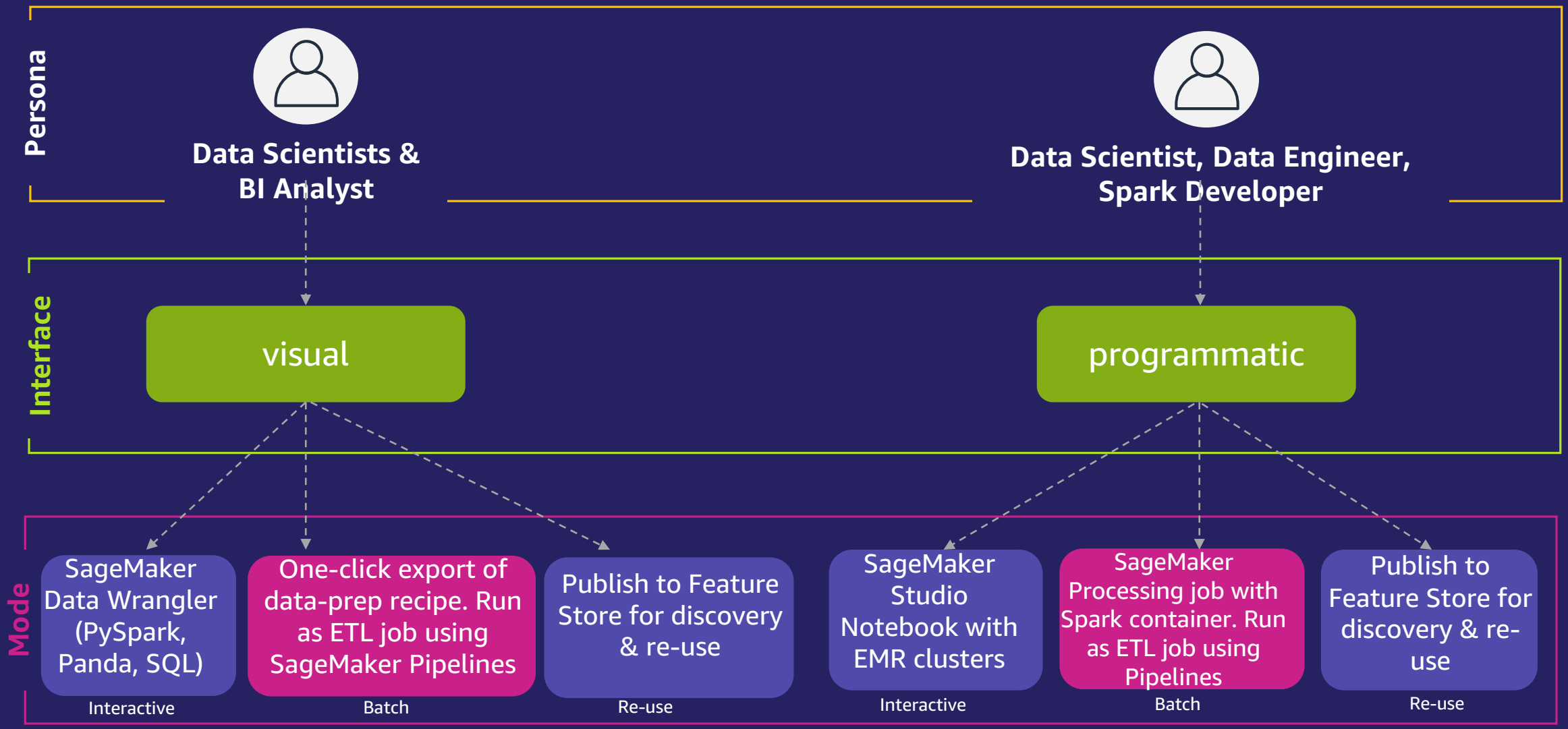


Data Processing on Amazon SageMaker

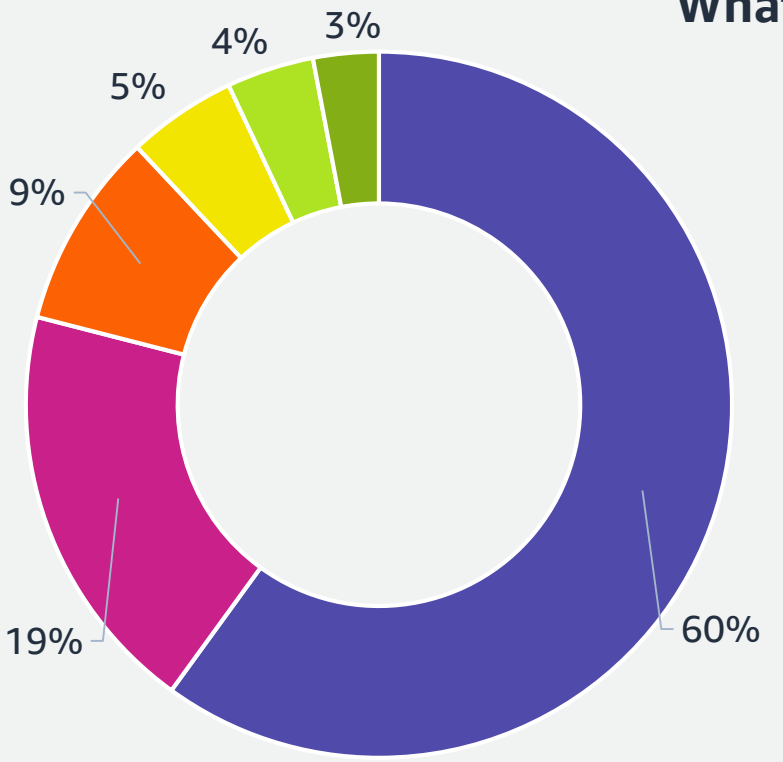
```
import boto3
import sagemaker
from sagemaker import get_execution_role
from sagemaker.sklearn.processing import SKLearnProcessor

region = boto3.session.Session().region_name

role = get_execution_role()
sklearn_processor = SKLearnProcessor(
    framework_version="0.20.0", role=role, instance_type="ml.m5.xlarge", instance_count=1
)
```

```
from sagemaker.processing import ProcessingInput, ProcessingOutput

sklearn_processor.run(
    code="preprocessing.py",
    # arguments = ['arg1', 'arg2'],
    inputs=[ProcessingInput(source="dataset.csv", destination="/opt/ml/processing/input")],
    outputs=[
        ProcessingOutput(source="/opt/ml/processing/output/train"),
        ProcessingOutput(source="/opt/ml/processing/output/validation"),
        ProcessingOutput(source="/opt/ml/processing/output/test"),
    ],
)
```

SageMaker Studio Data Prep Portfolio



80% of time spent on data prep



What data scientists spend the most time doing

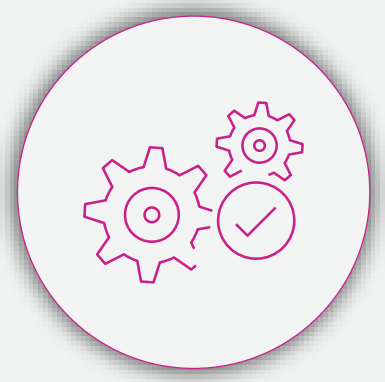
- Cleaning and organizing data
- Collecting data sets
- Mining data for patterns
- Other
- Refining algorithms
- Building training sets

Source: Forbes survey of 80 data scientists, March 2016

aws machine learning

Challenges of separate feature stores

Feature drift



Feature duplication

Slow model development/deployment

Amazon SageMaker Feature Store

Securely store, discover, and share features for ML

Online and off-line

Millisecond latency

Consistent features

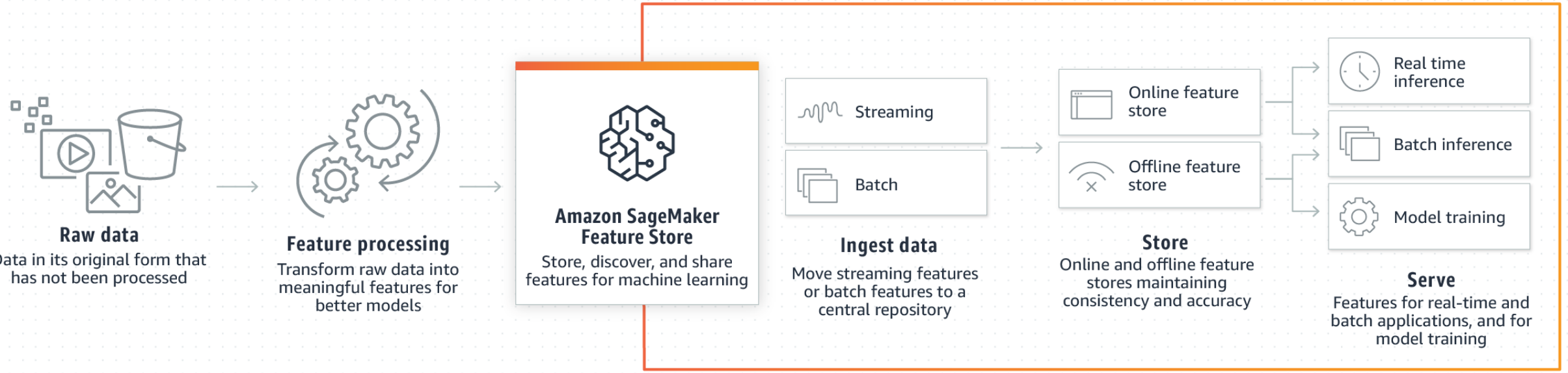
Visual search

Sharing and collaboration

aws machine learning

Amazon SageMaker Feature Store

How it works



Demo

Amazon SageMaker

Next Steps

Onboarding & Processing

- https://docs.aws.amazon.com/sagemaker/latest/dg/gs-studio-onboard.html
- https://docs.aws.amazon.com/sagemaker/latest/dg/processing-job.html

Training

- https://docs.aws.amazon.com/sagemaker/latest/dg/train-model.html
- https://docs.aws.amazon.com/sagemaker/latest/dg/distributed-training.html
- https://aws.amazon.com/sagemaker/debugger

Deployment

- https://docs.aws.amazon.com/sagemaker/latest/dg/realtime-endpoints.html
- https://docs.aws.amazon.com/sagemaker/latest/dg/serverless-endpoints.html
- https://docs.aws.amazon.com/sagemaker/latest/dg/async-inference.html
- https://docs.aws.amazon.com/sagemaker/latest/dg/batch-transform.html

https://github.com/aws/amazon-sagemaker-examples

https://sagemaker.readthedocs.io/en/stable/index.html

aws machine learning

Q & A

Please Complete the session Survey

Thank you!