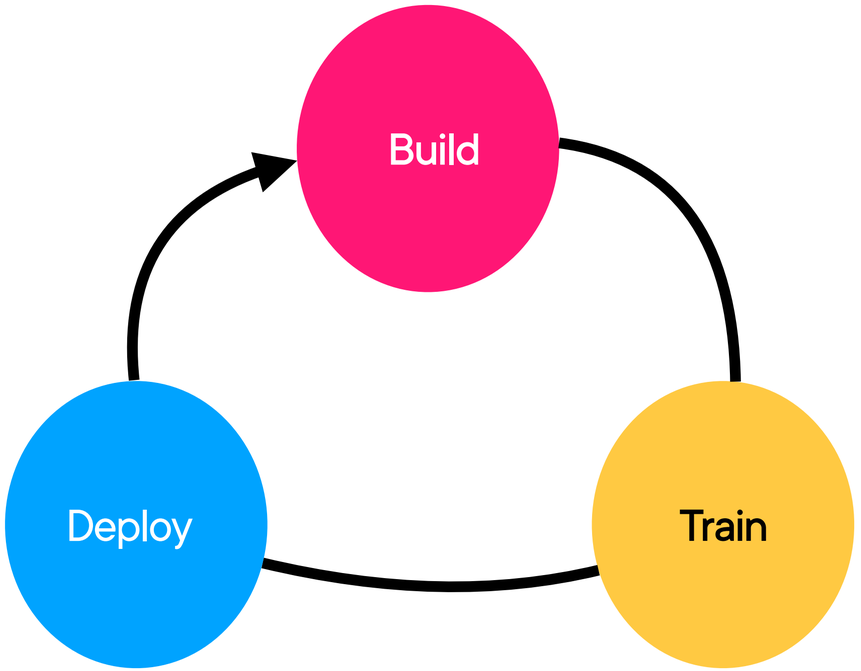
Unpacking Amazon SageMaker

Noreen Hasan

Cloud Author at Pluralsight

Amazon SageMaker

- A fully managed service for developers and data scientists
- Enables users to prepare, build, train, tune, and deploy ML models from scratch
- It supports version control
- It supports supervised, unsupervised, reinforcement, and deep learning



SageMaker Studio

Comprehensive Toolbox for ML

- It offers an IDE
- A one-stop shop for building, training, and deploying ML models
  - Notebooks
  - Canvas
  - Data preparation and visualization
  - Collaboration tools

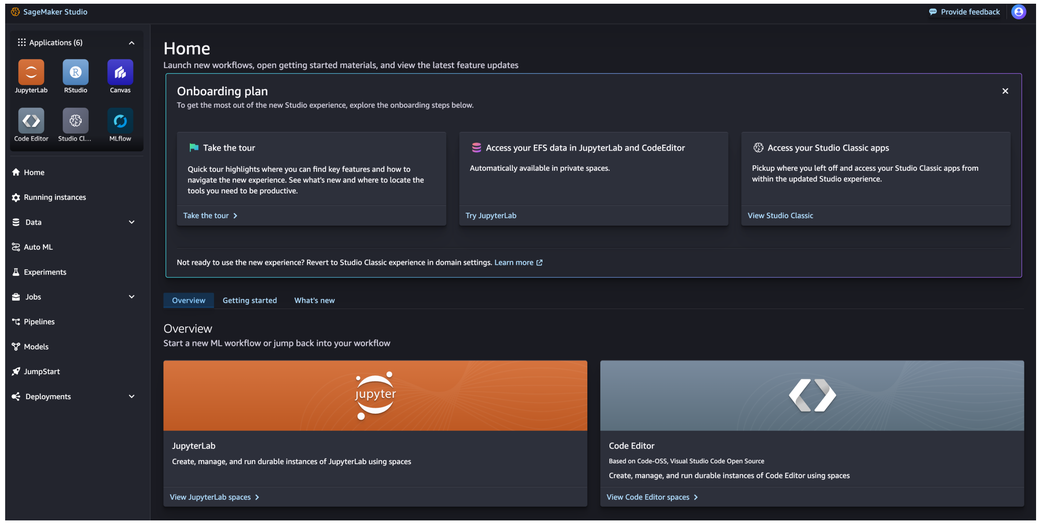
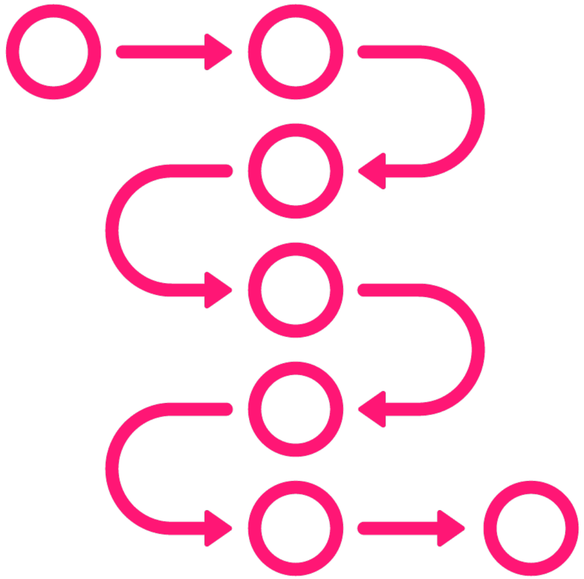


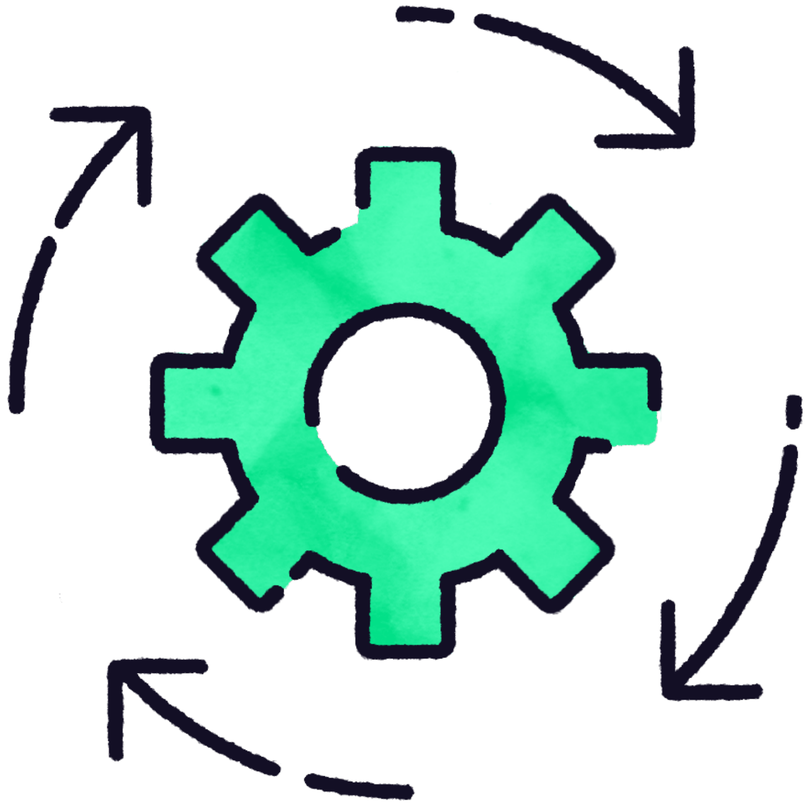
Image: Personal Screenshot from amazon.com

SageMaker Pipelines

ML Workflows

- Designed to orchestrate and automate the entire machine learning workflow, from data preparation to model deployment
- Supports versioning and tracking
- Allows you to define and manage the steps involved in building, training, evaluation, and deployment
- You can integrate AutoML as a step





AutoML (Autopilot)

- Automates the model selection and hyperparameter tuning process
- Allows you to quickly generate a model with minimal manual intervention
  - You upload your data
  - Specify the target variable (e.g., Is_Spam)
  - AutoML handles the rest

Amazon SageMaker Data Wrangler

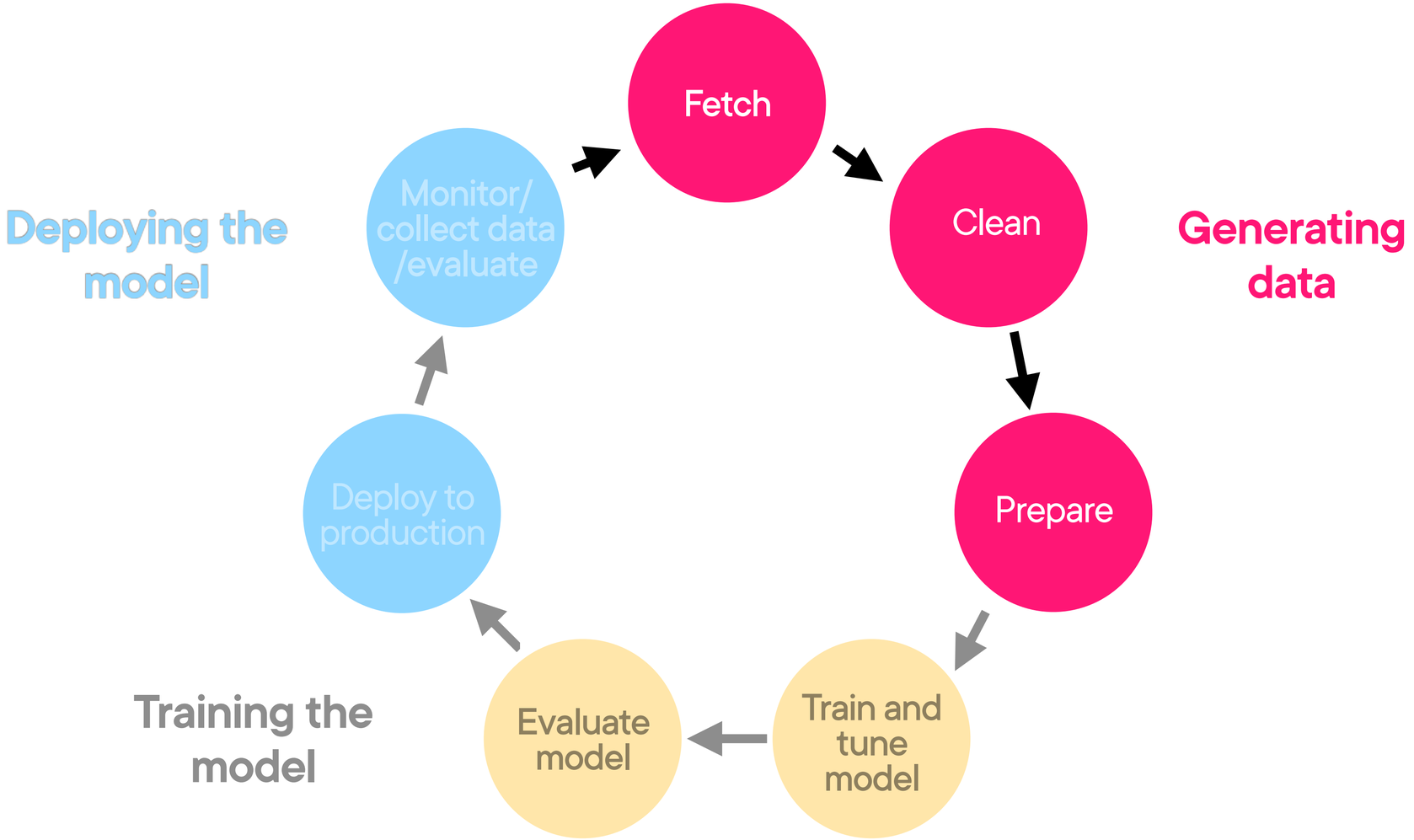
Data Is Messy!

ZIP	House Size	NumOfBedrooms	NumOfBathrooms	Price
84103-0437	1,500	Medium	Yes	$11326
9873	1,500	Medium	Yes	€8984
—	1,800	Small	Yes	N/A
		Large	NaN	$316
43789	2,000	Medium	Yes	€7290

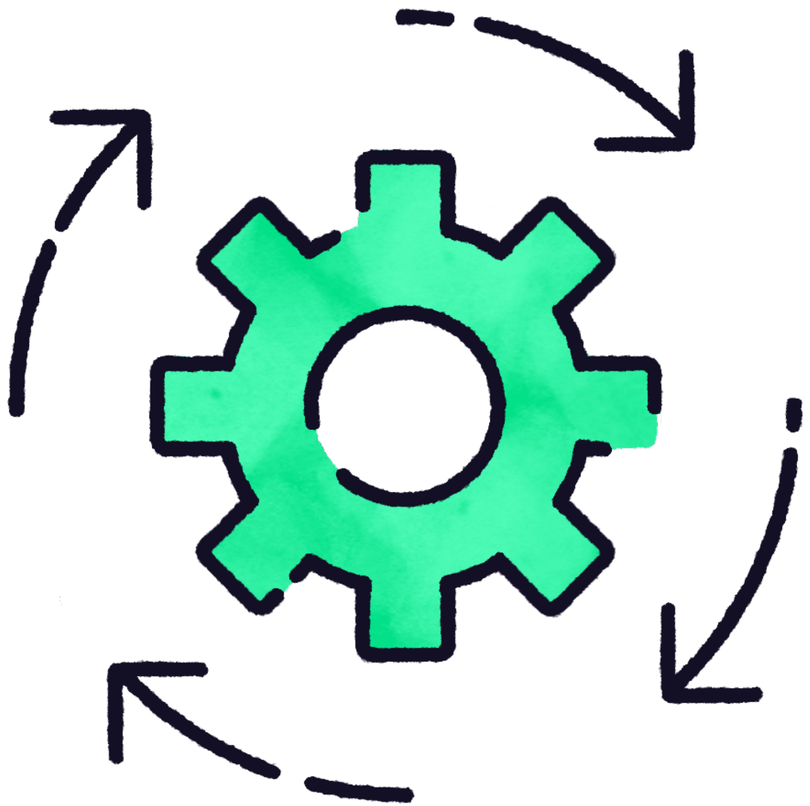
Raw messy data

Data Wrangling

Clean, transformed, and organized data



The Machine Learning Lifecycle



Data Wrangler

- Part of the SageMaker Studio
- A one-stop interface that simplifies the process of preparing and transforming data
- It can work with various forms of data
  - Tabular
  - Image
  - Text
  - Time series
- You can use SQL to query and manipulate data directly

How Can Data Wrangler Prepare Data?

Import

Easily import data from various sources

S3

Redshift

Athena

Clean

Identify missing values, duplicates, and outliers

Feature Engineer

Create new features from existing data using built-in transformations

Visualize (EDA)

Visualize distributions, summary statistics, and relationships between variables

Prepare Data

Anytime, you need to clean, prepare, or transform data, consider Data Wrangler

Amazon SageMaker Feature Store

Insufficient raw data

Feature Engineering

ML Features

Transform the data into a format that is easier for a ML model to learn from.

Feature Engineering: Auto Loan Model

Type	Year	Size	Used	Miles	Price	Credit Score	Loan Approved
Truck	2018	Medium	Yes	11326	36498	718	Yes
SUV	2019	Medium	Yes	8984	32099	785	Yes
Sedan	2016	Small	Yes	58446	9650	690	Yes

Feature Extraction:

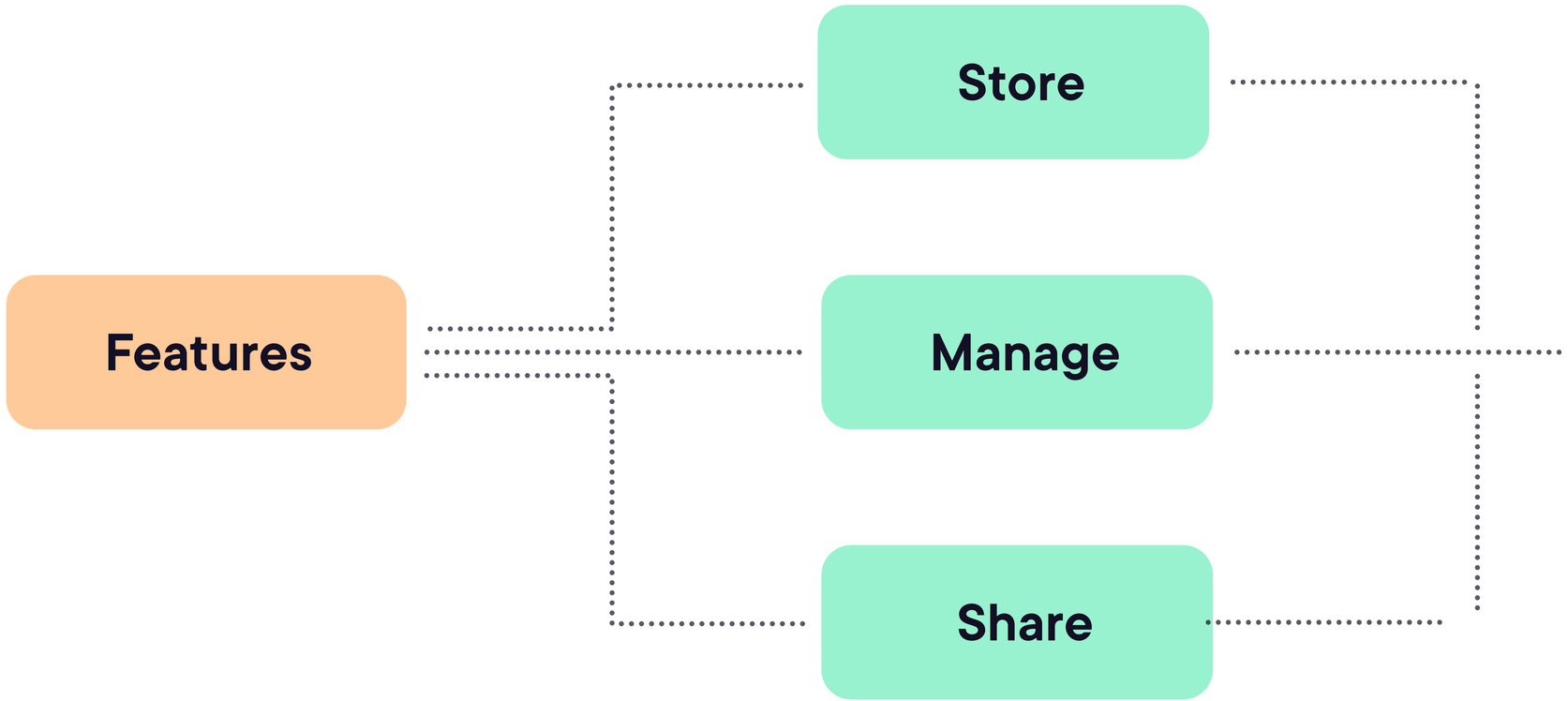
Age of the car:

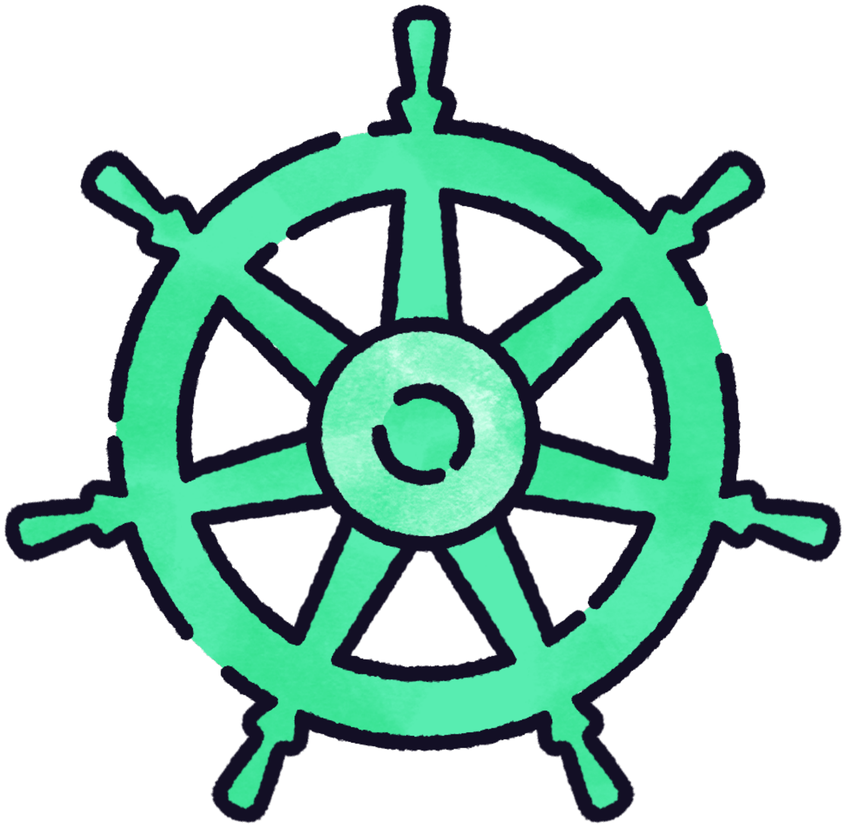
- Current year - year

Price per mile:

- Price of the car / miles driven

To reduce redundancy of features and save time



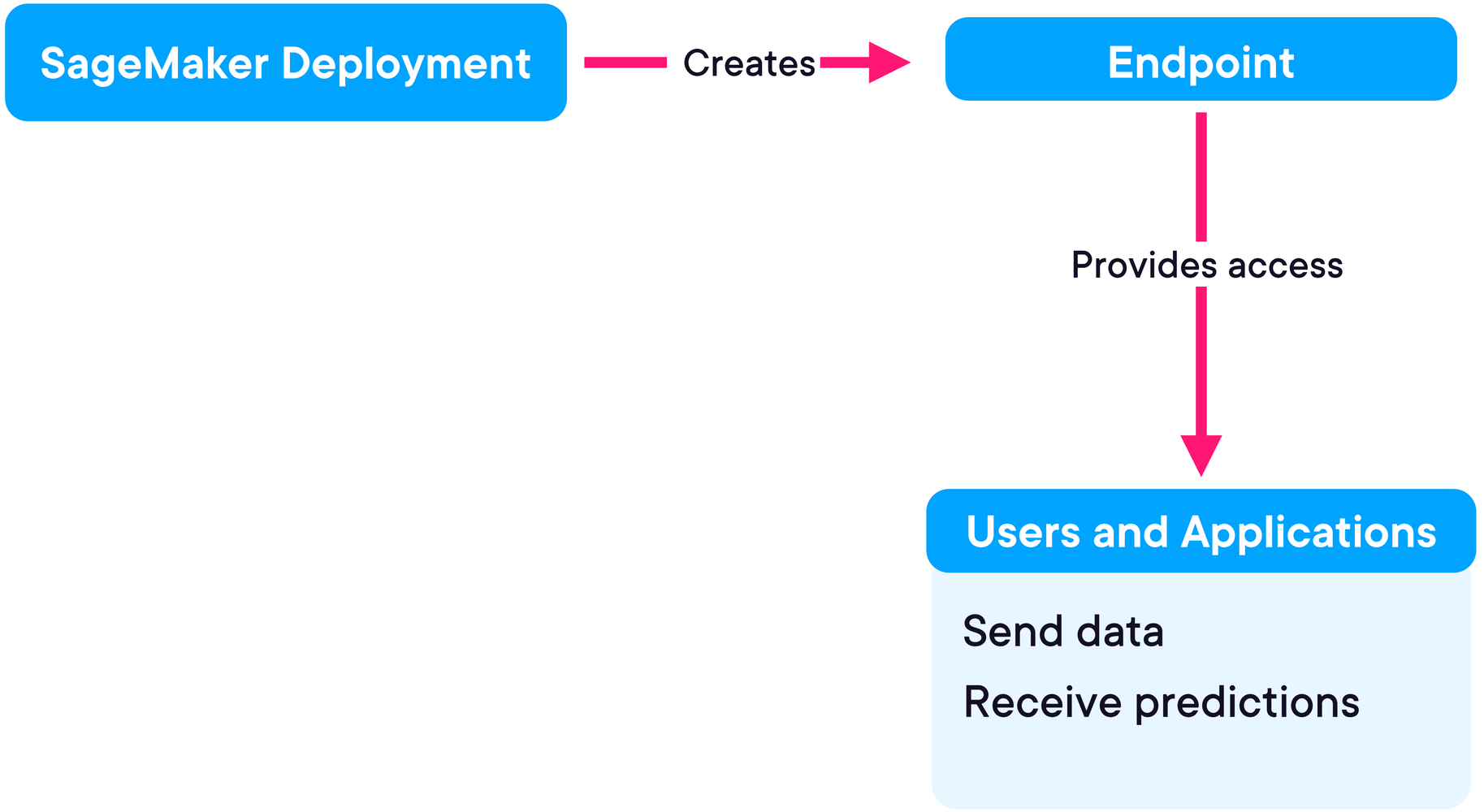


- **You can create features:**
  - In SageMaker DataWrangler, then publish to SageMaker Feature Store
  - Directly in SageMaker Feature Store

Demo: Amazon SageMaker Console Walkthrough

Explore the features of the Amazon SageMaker Studio

Amazon SageMaker Deployments



Real-time (Synchronous Inference)

Serve predictions in real-time

- Provides instant responses to your requests as soon as the model process the input
- Ideal for applications requiring immediate predictions
    - Chatbots
    - Recommendation systems
    - Fraud detection

When to use?

- Require low-latency responses
- Have a predictable traffic pattern

Asynchronous Inference

Serve predictions in an async fashion

- You submit the request, then check back later for the results once the model is done processing in the background
- Response time can vary significantly
    - Processing large files
    - Complex models
    - Image or video processing

When to use?

- Unpredictable workloads
- Your model takes longer to respond

Batch Transform (Batch Inference)

Processes large batches

- It's executed asynchronously
- Ideal for applications where you have a large volume of data to process
    - Generating predictions for a dataset overnight

When to use?

- You have large datasets to predict
- Latency is not a concern

Serverless

Automatically scales to accommodate varying loads

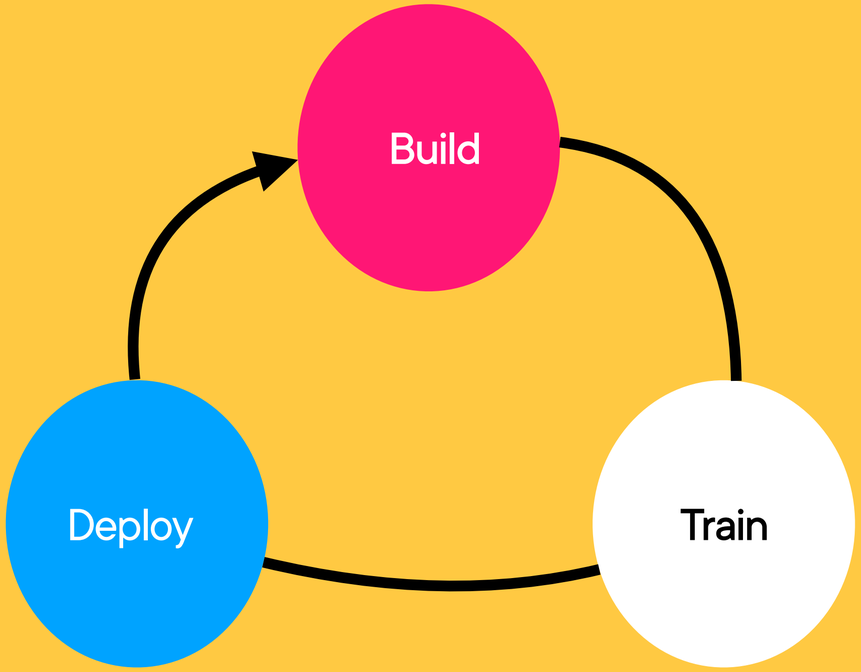
- Allows you to run your models without managing the underlying infrastructure
- Ideal for applications with sporadic traffic
    - Prototyping and development phases

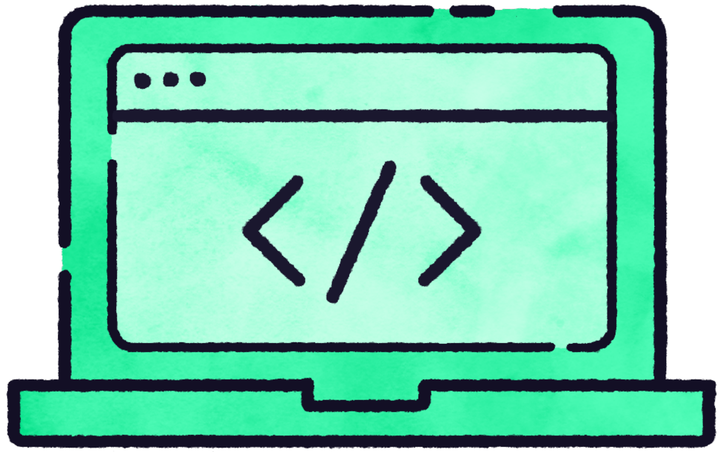
When to use?

- Expect variable workloads
- Applications with unpredictable usage patterns

Exam Tips

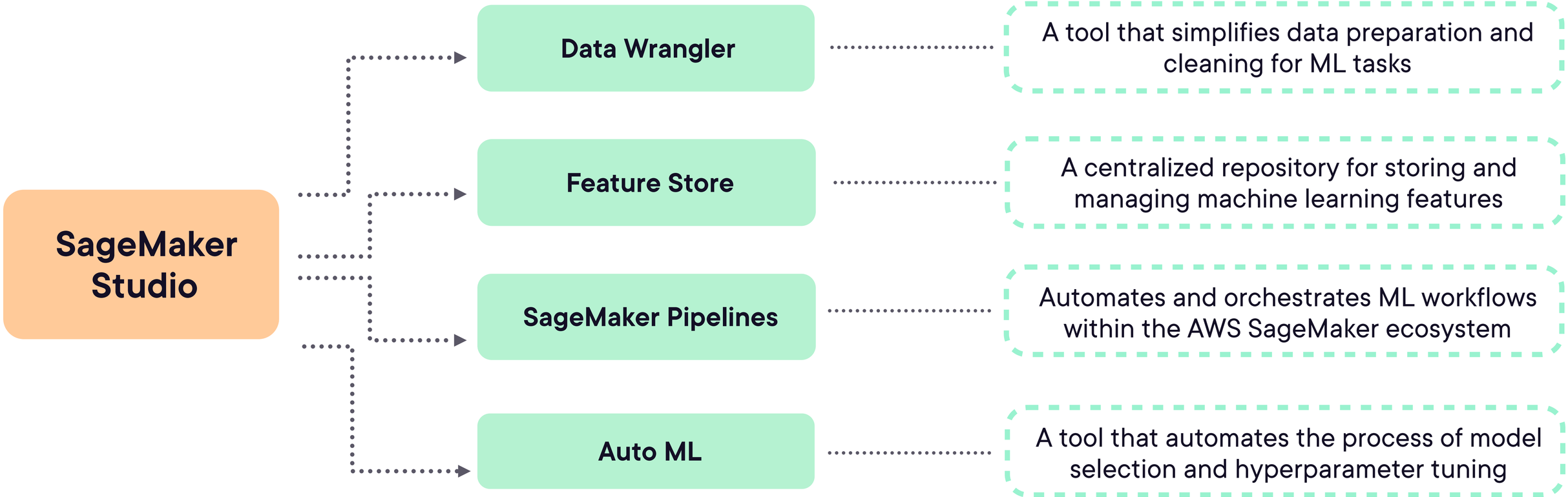
Amazon SageMaker is a fully managed service that helps you build, train, and deploy machine learning models quickly



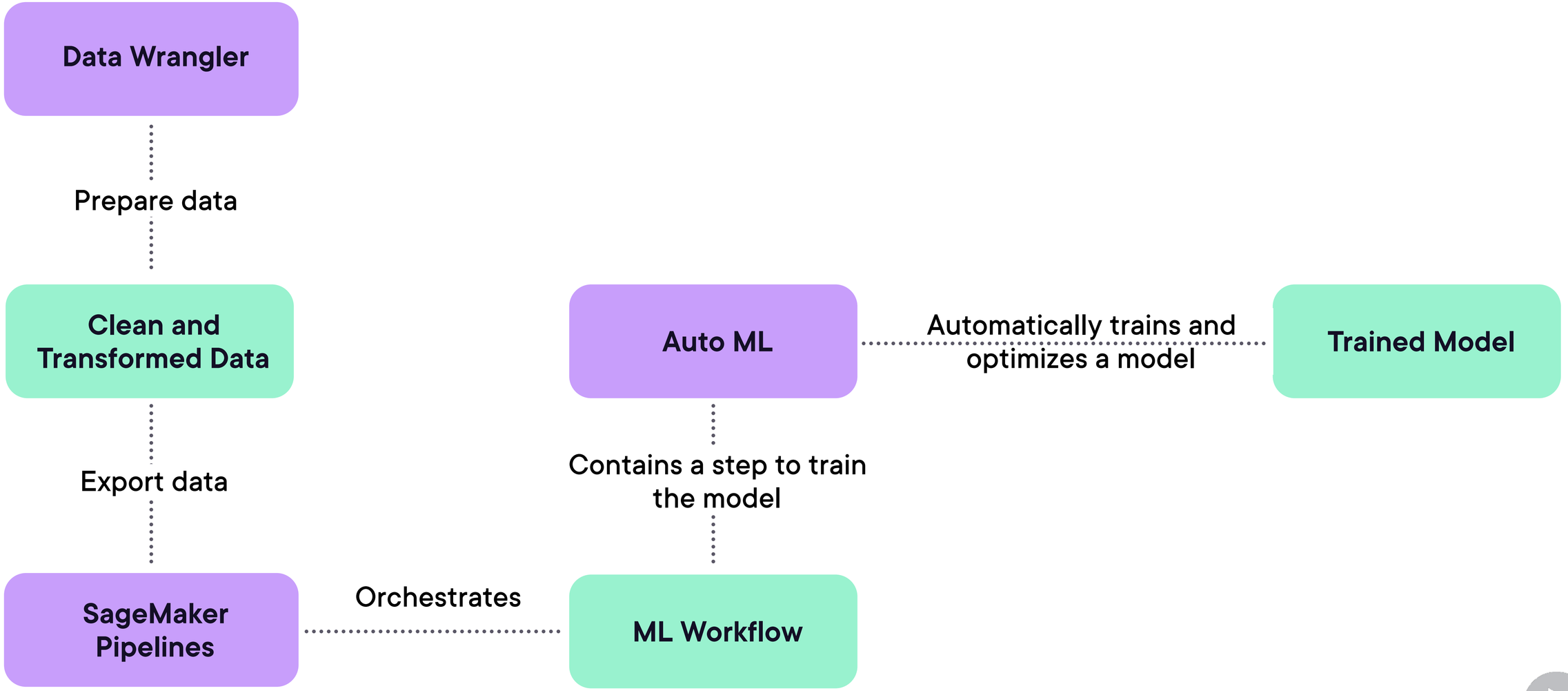


SageMaker Studio

- An integrated development environment for machine learning, making it easier to manage the entire ML workflow



SageMaker Studio



SageMaker Deployments

Inference Type	Latency	Use Case
Real-time (Synchronous Inference)	Low (sub-second to few seconds)	Chatbots
Asynchronous Inference	Moderate to High	Image and video analysis
Batch Transform (Batch Inference)	High (minutes to hours)	Processing large datasets
Serverless	Low to Moderate	Scalable applications with variable workloads