第二十八回「アップデート紹介と
ちょっぴり DIVEDEEP するAWS の時間」

SageMaker シャドウテストで、デプロイ前にMLモデルの性能を検証しよう

大前 遼 (Ryo OMAE)

Solutons Architect
Amazon Web Services Japan G.K.

大前 遼 (omaeryo@)
Solutions Architect

出身:
関西(京都市&神戸市)

得意な技術領域:
AI/ML (ベイズ、画像認識、グラフ等), IoT

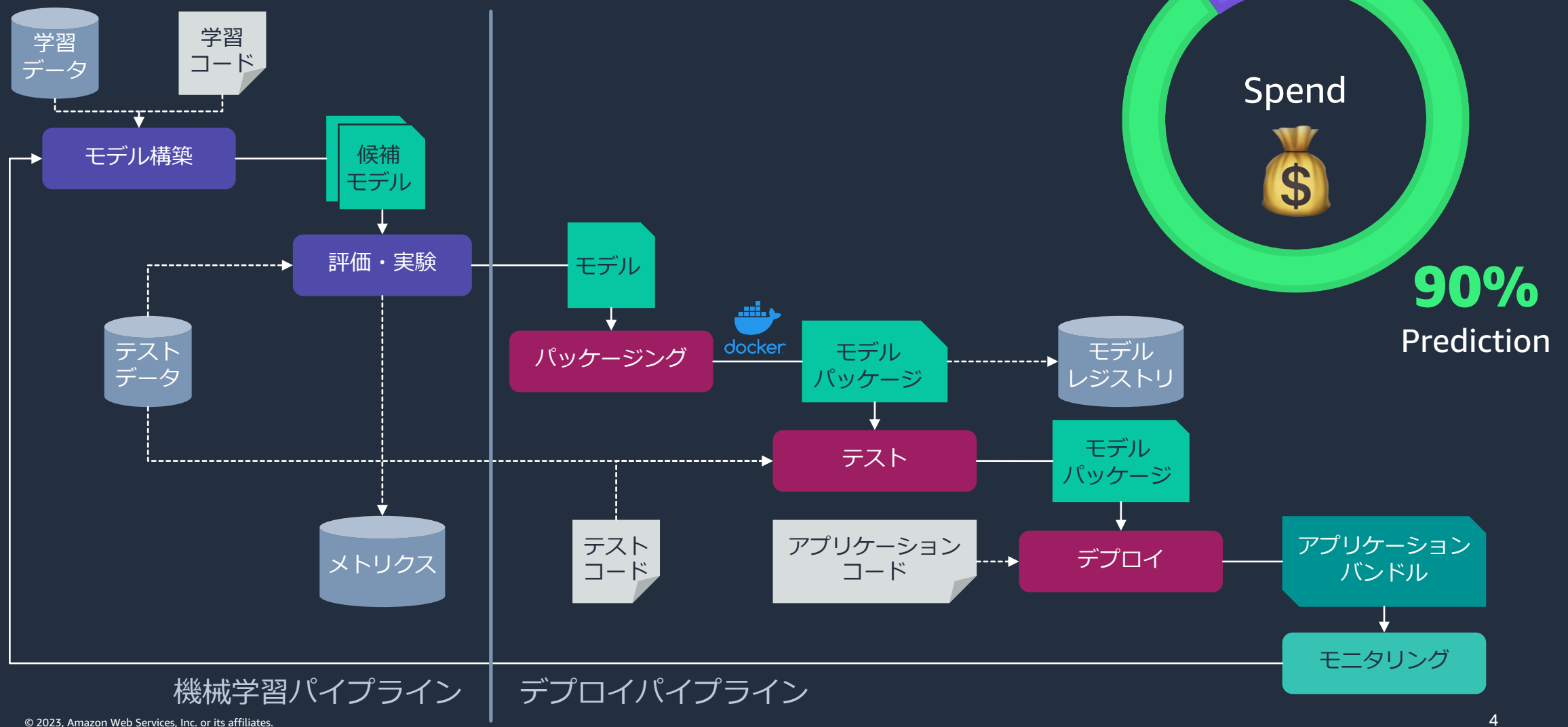
趣味:
バイク、ボルダリング、サウナ、AtCoder

好きなAWSサービス:
AWS CDK, Amazon Neptune, AWS SageMaker

Agenda

- 一般的な ML システム開発の流れ
- Shadow A/B Testing とは
- SageMaker Shadow Testing の概要
- デモ
- まとめ

機械学習モデルの開発とデプロイの流れ



MLモデルデプロイ時の課題

MLシステムにおいて、以下2点が重要

- MLモデルが所望の**精度を発揮できる**
- **処理性能が十分に確保**できる

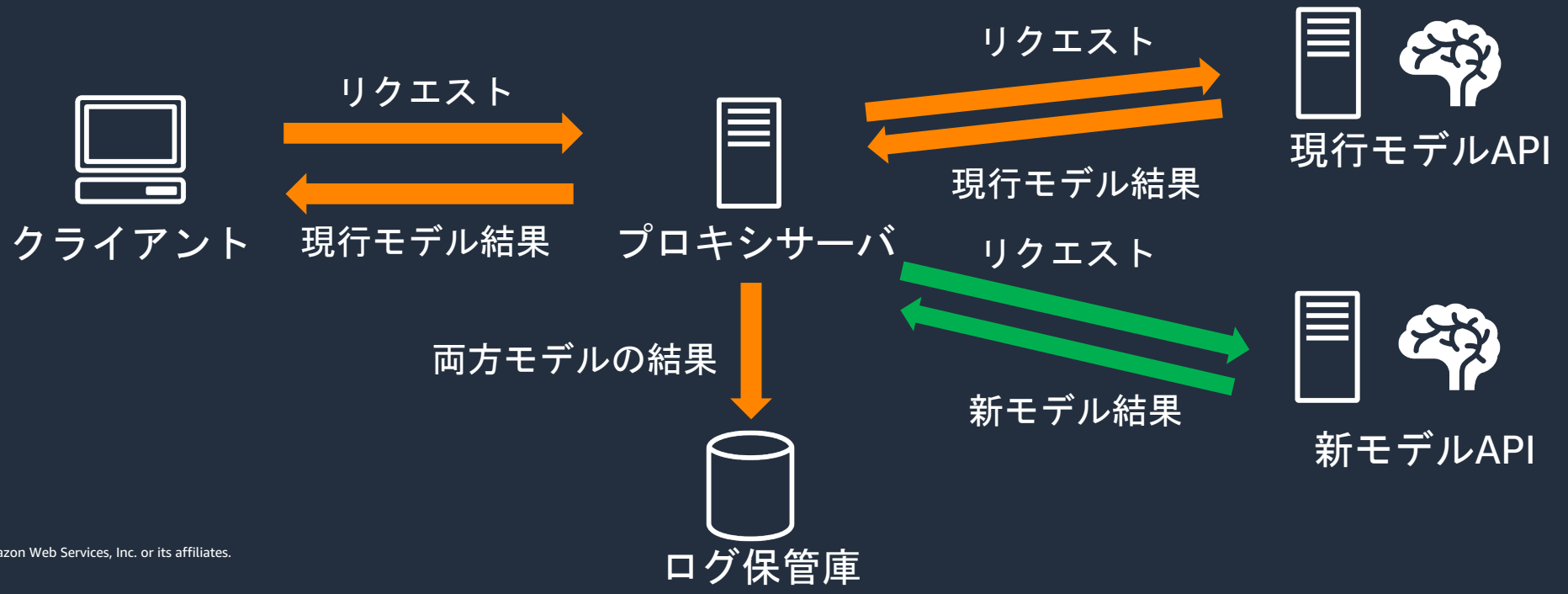
MLモデル変更時、デプロイ前後での検証が必要

Shadow A/B Testing とは

本番システムで稼働するリクエストを、新モデルへミラーリングし、推論だけ実行すること

クライアントへ推論結果を返すのは現行モデルのみ

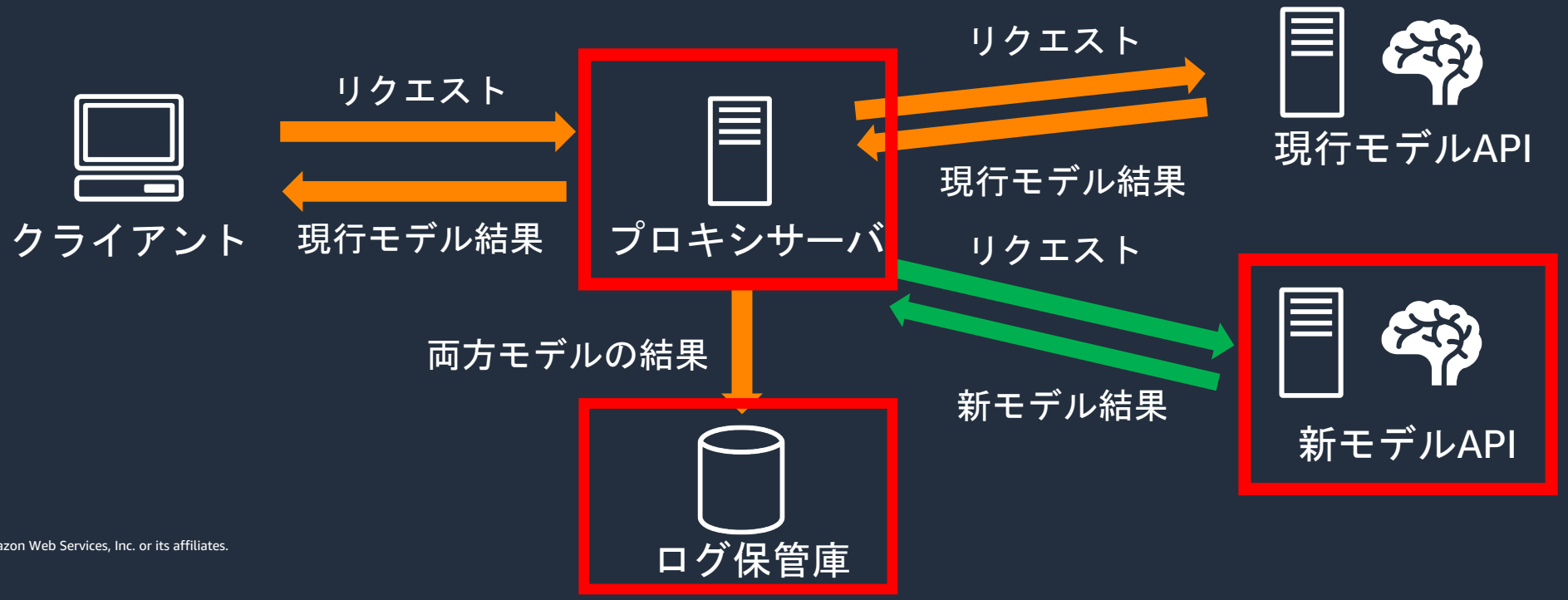
クライアントへの影響なしに、新モデル導入の可否を判断可能



Shadow A/B Testing の課題

テスト環境構築の手間がかかる

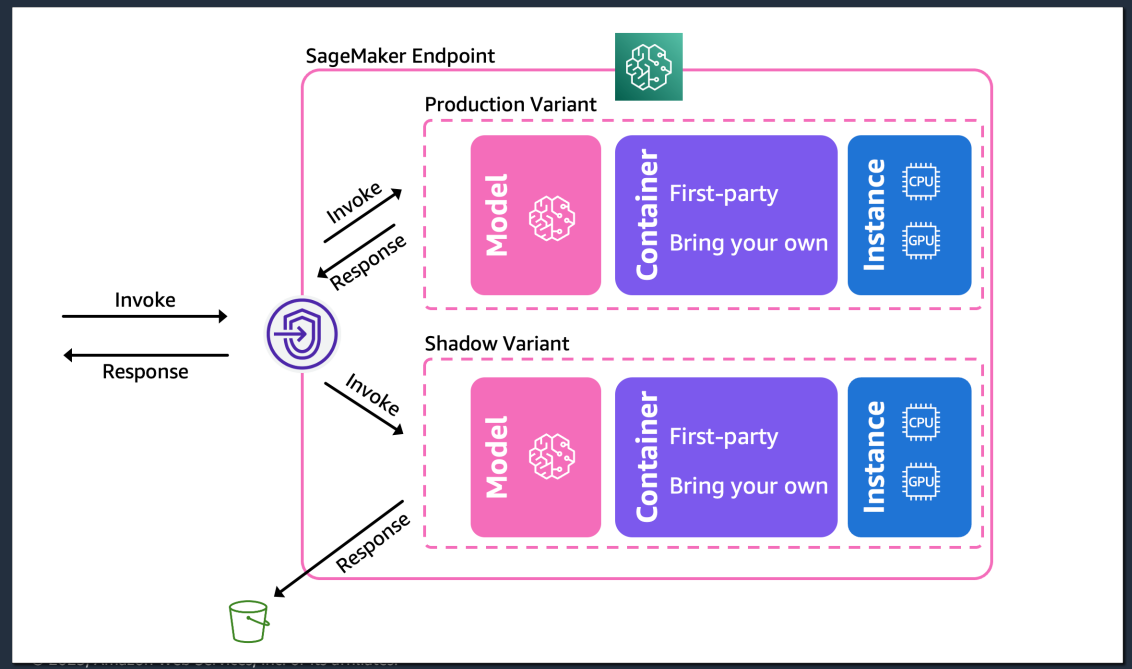
- プロキシサーバの新モデルへの転送設定が必要
- 新モデルAPI サーバの環境構築が必要
- 推論結果ログ保管の仕組みが必要

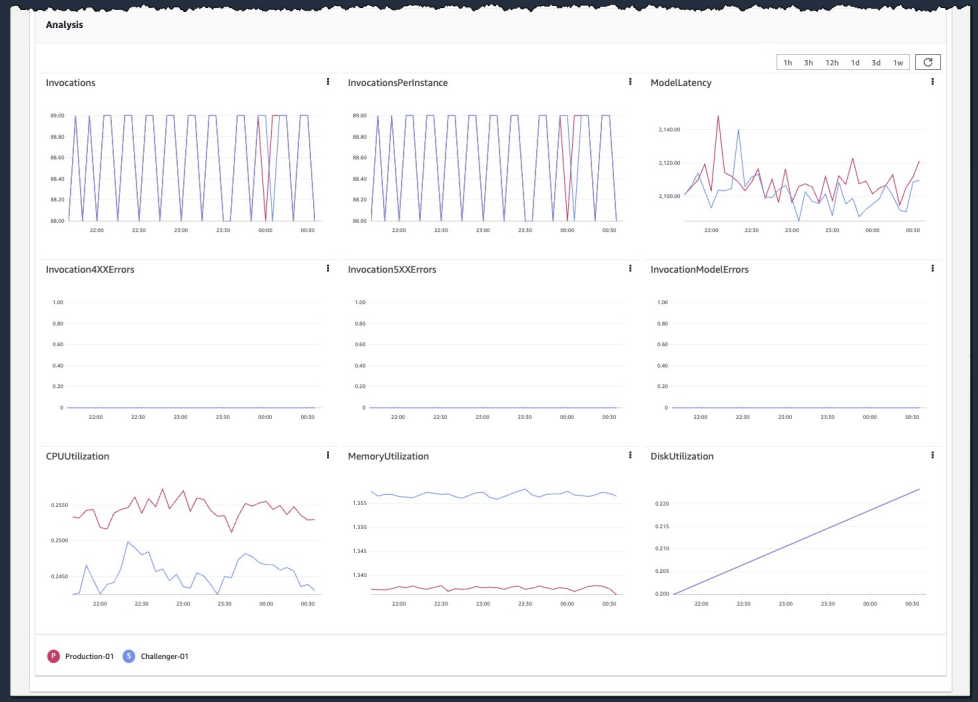


Amazon SageMaker Shadow Testing

シャドーテストの実行を容易にする機能

- 比較したい現行モデルと新しいモデルを選択すると、自動的に新しいモデルがデプロイされ、本番環境へのリクエストのコピーが生成されて新しいモデルに送られる
- レイテンシーやエラー率などのメトリクスをダッシュボードで参照し比較が可能
- シャドーテストの負荷を低減しリリースを加速

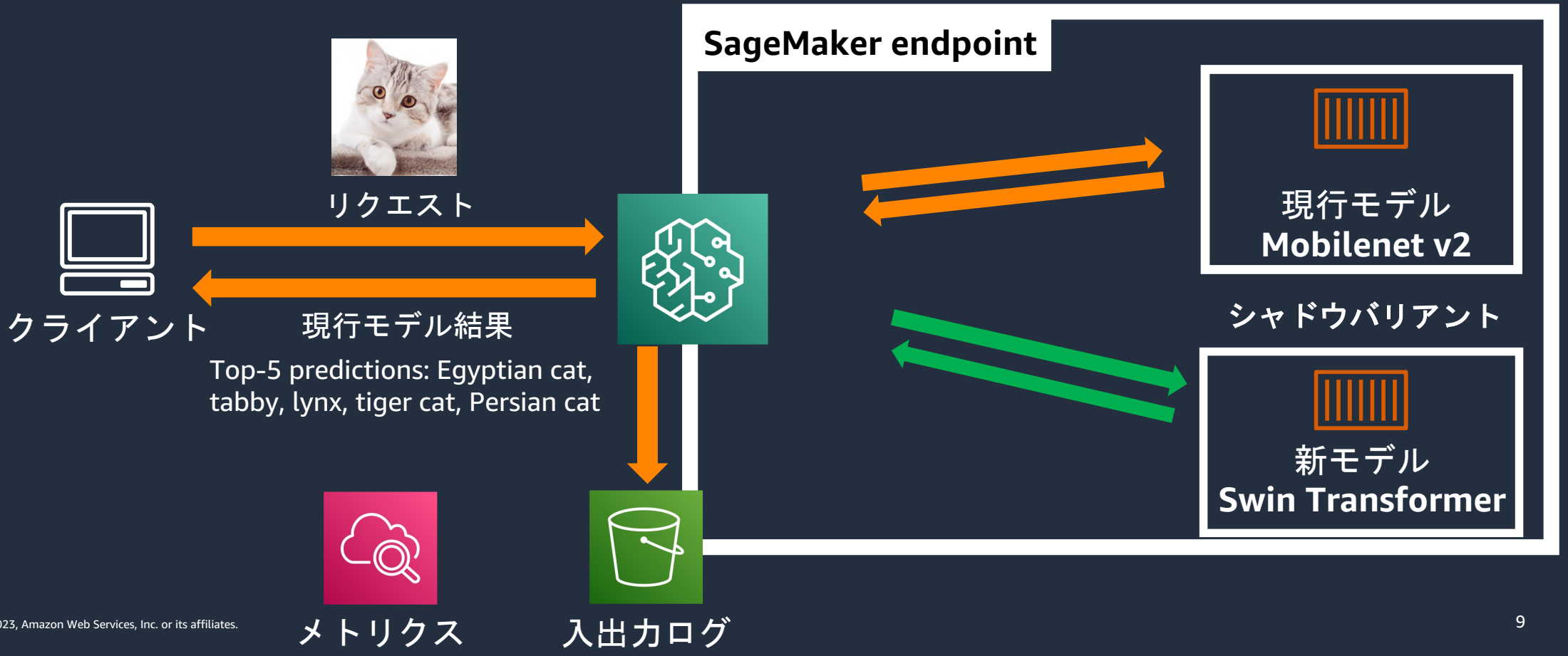




デモシナリオ

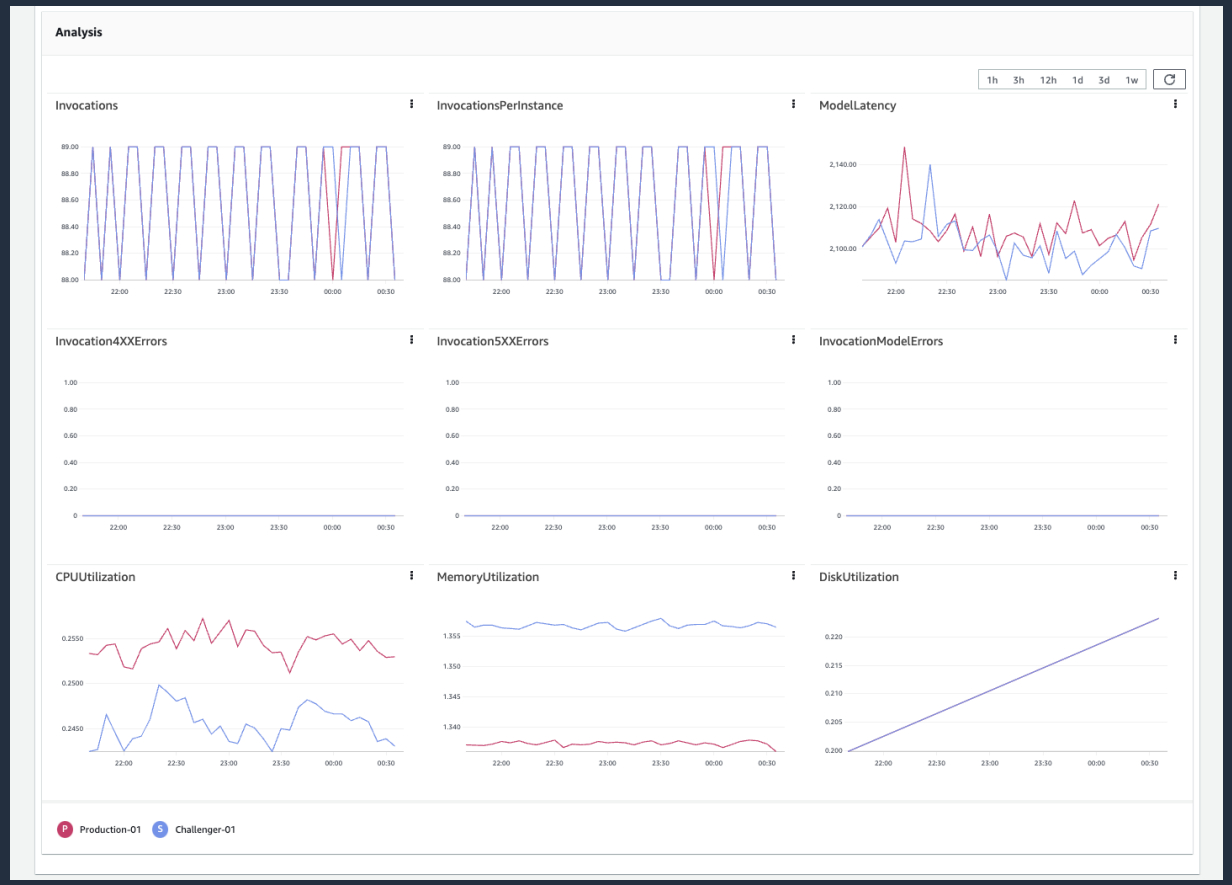
画像物体認識システム、上位5つの推論結果を返す

現行モデル(Mobilenet v2) と新モデル(Swin Transformer) を比較



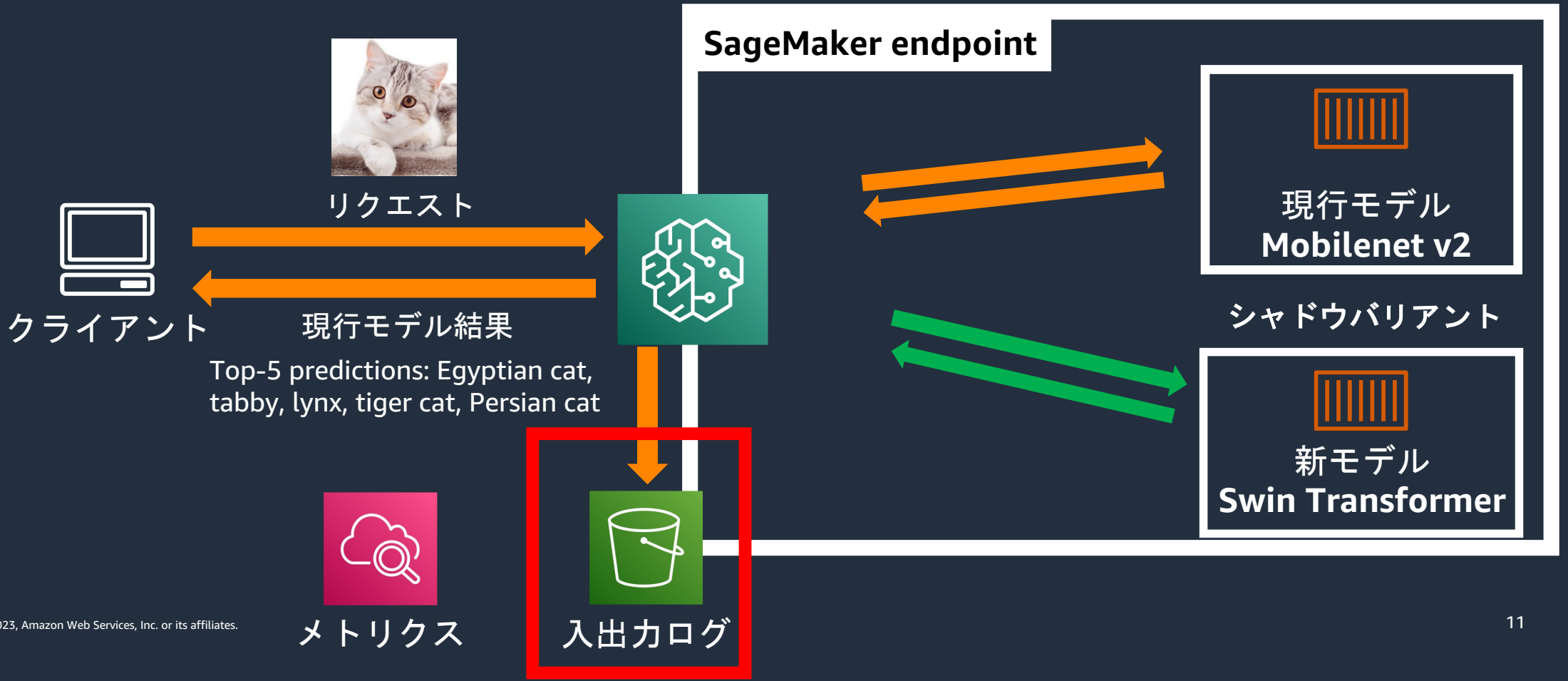
デモシナリオ: メトリクスの確認

処理性能比較のため、CPU利用率・メモリ利用率等メトリクス比較



デモシナリオ: 入出力ログファイルの確認

モデル精度検証のためのログファイルの検証



まとめ

- 機械学習システムにおいて、**モデル精度と処理性能は重要**
- MLモデルをデプロイする前にA/B Shadow Testingすることが有効
- Amazon SageMaker Shadow Testing機能を使うことで、簡単に、A/B Shadow Testingを始められる