MLOps with serverless architectures

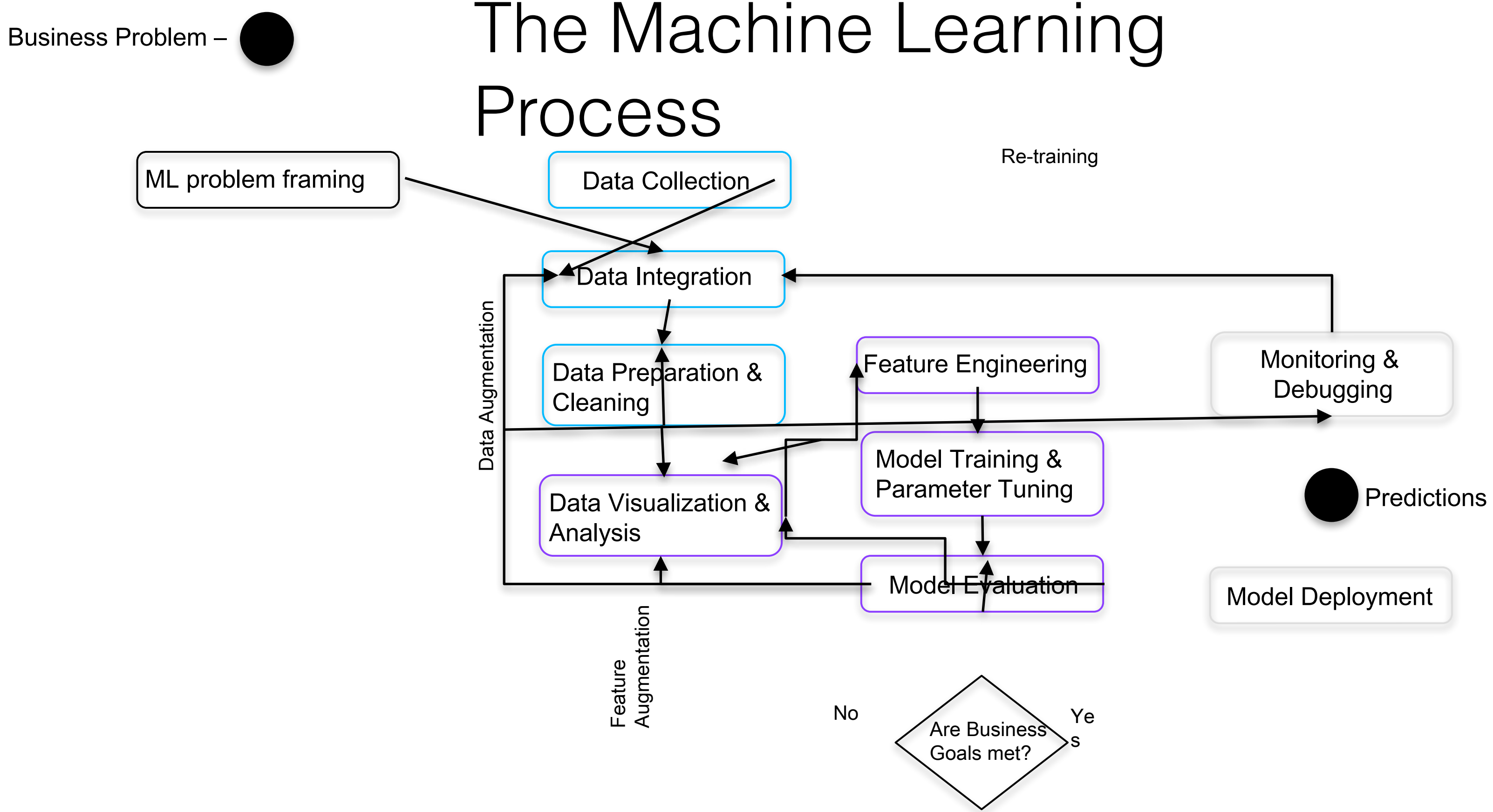
Julien Simon
Principal Technical Evangelist, AI and Machine Learning, AWS
@julsimon

Agenda

- But why?
- A quick recap on Amazon SageMaker
- A quick recap on serverless architectures
- Open Source tools: AWS Chalice, Serverless Framework
- Demos
- Resources

Rationale

The Machine Learning Process



Amazon SageMaker

Pre-built notebooks for common problems

Built-in, high-performance algorithms

Build

One-click training

Hyperparameter optimization

Train

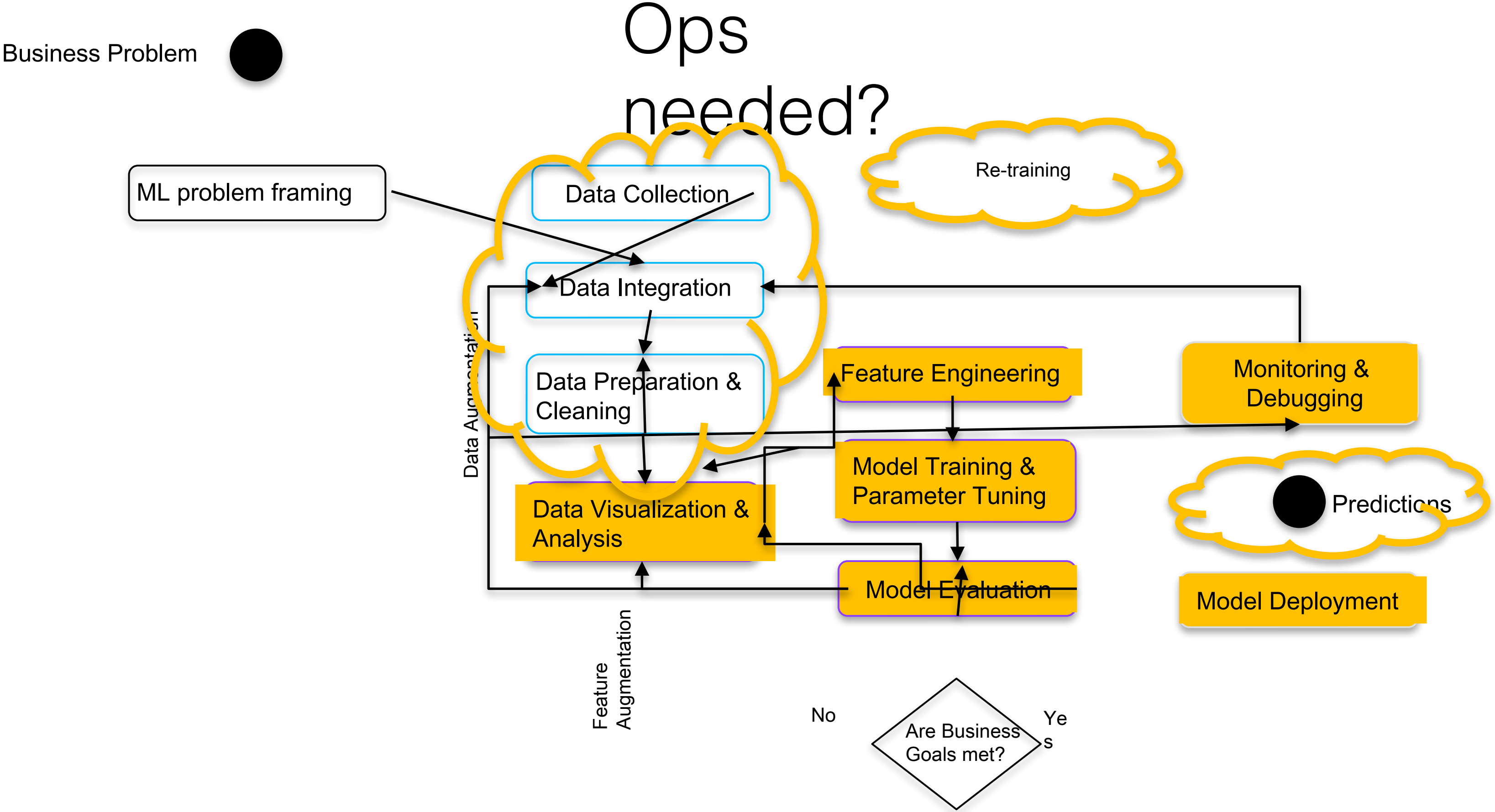
One-click deployment

Fully managed hosting with auto-scaling

Deploy

FREE TIER

Ops needed?



Business Problem

ML problem framing

Data Collection

Data Integration

Data Preparation & Cleaning

Data Visualization & Analysis

Feature Engineering

Model Training & Parameter Tuning

Model Evaluation

Are Business Goals met?

No

Ye s

Model Deployment

Monitoring & Debugging

Predictions

Re-training

Data Augmentation

Feature Augmentation

Serverless Architectures

Serverless architecture

=

Fully-managed services

+

AWS Lambda

Amazon API Gateway

Amazon Kinesis Streams

Amazon DynamoDB

Amazon S3

AWS Lambda

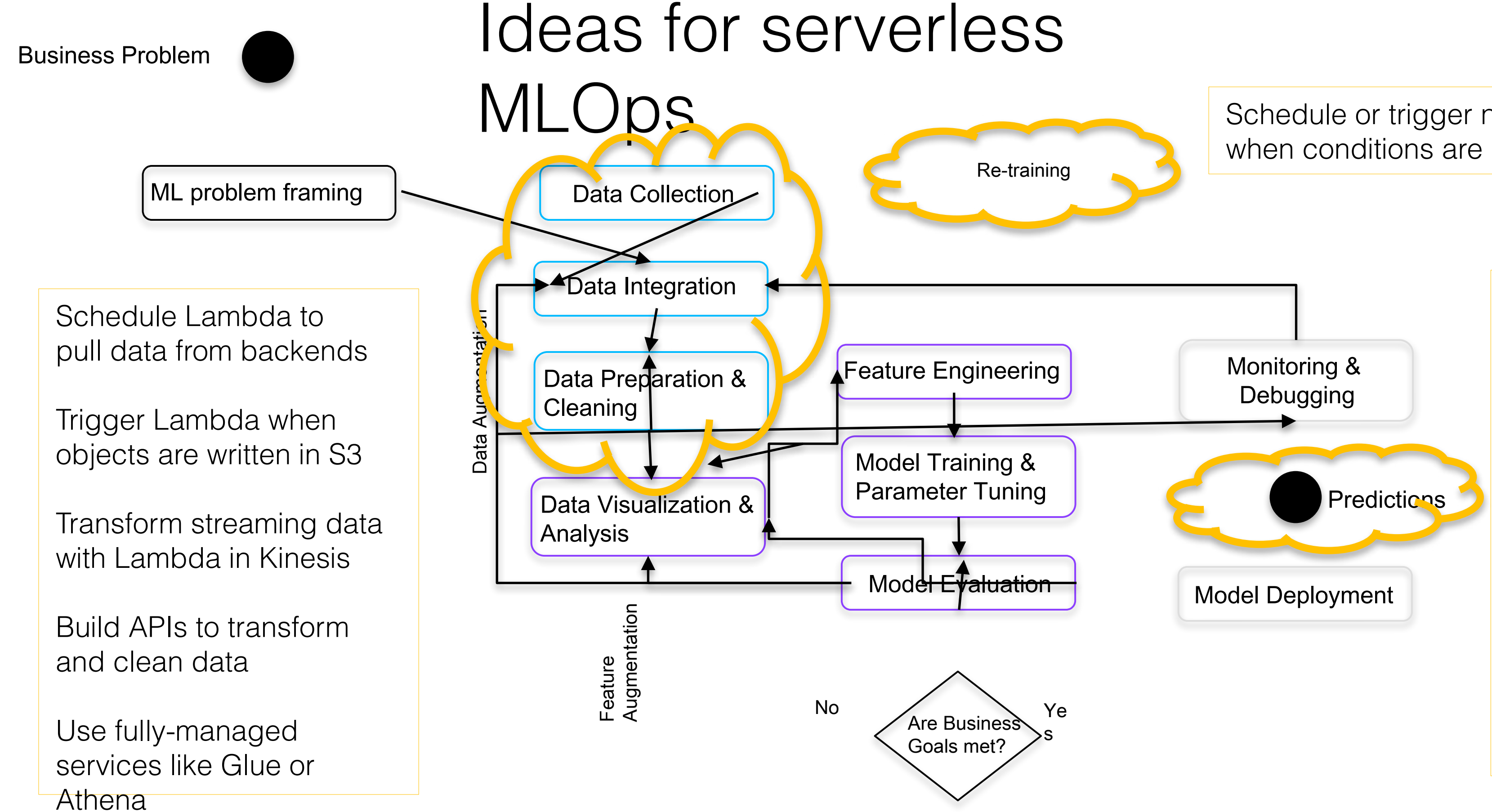
- Announced at re:Invent 2014
- Deploy functions in Java, Python, Node.js ,C# and Go.
- Just code, without the infrastructure drama
- Built-in scalability and high availability
- Integrated with many AWS services
- Pay as you go
  - Combination of execution time (100ms slots) & memory used.
  - Starts at $0.20 per million requests.
  - Free tier available: first 1 million requests per month are free.
- Orchestration with AWS Step Functions.

http://aws.amazon.com/lambda

What can you build with serverless architectures?

- Automate your AWS infrastructure
- Build event-driven applications
- Build APIs together with Amazon API Gateway

Ideas for serverless MLOps



Schedule or trigger new training when conditions are met

Schedule Lambda to pull data from backends

Trigger Lambda when objects are written in S3

Transform streaming data with Lambda in Kinesis

Build APIs to transform and clean data

Use fully-managed services like Glue or Athena

Build wrapper API for pre/post-processing

Schedule Lambda for batch predictions

Use Step Functions for Ensemble predictions

Open Source Frameworks

The Serverless framework

- Announced at re:Invent 2015
- Auto-deploys and runs Lambda functions, locally or remotely
- Auto-deploys Lambda event sources: API Gateway, S3, DynamoDB, etc.
- Creates all required infrastructure with CloudFormation
- Simple configuration in YML

http://github.com/serverless/serverless
https://serverless.com

AWS Chalice

- Open Source project released in July 2016
- “Flask for serverless”
- Just add your Python code
  - Deploy web services with a single CLI call and zero config
  - The API is created automatically
  - The IAM policy is auto-generated (crowd goes wild)
- Run and test APIs on local port 8000 (similar to Flask)

https://github.com/awslabs/chalice

Demo #1: resizing images with AWS Chalice

https://medium.com/@julsimon/using-chalice-to-serve-sagemaker-predictions-a2015c02b033

Demo #2: building a prediction wrapper with AWS Chalice

https://medium.com/@julsimon/using-chalice-to-serve-sagemaker-predictions-a2015c02b033

Demo #3: retraining models with the Serverless Framework

https://medium.com/@julsimon/retraining-sagemaker-models-with-chalice-and-serverless-71a585ddbc7d

Resources

https://ml.aws
https://aws.amazon.com/sagemaker
https://aws.amazon.com/lambda

http://github.com/serverless/serverless
https://github.com/awslabs/chalice

https://medium.com/@julsimon

Thank you!

Julien Simon

Principal Technical Evangelist, AI and Machine Learning

@julsimon