Recommender Workshop

Part 1: Introduction

Customers Running Machine Learning On AWS Today

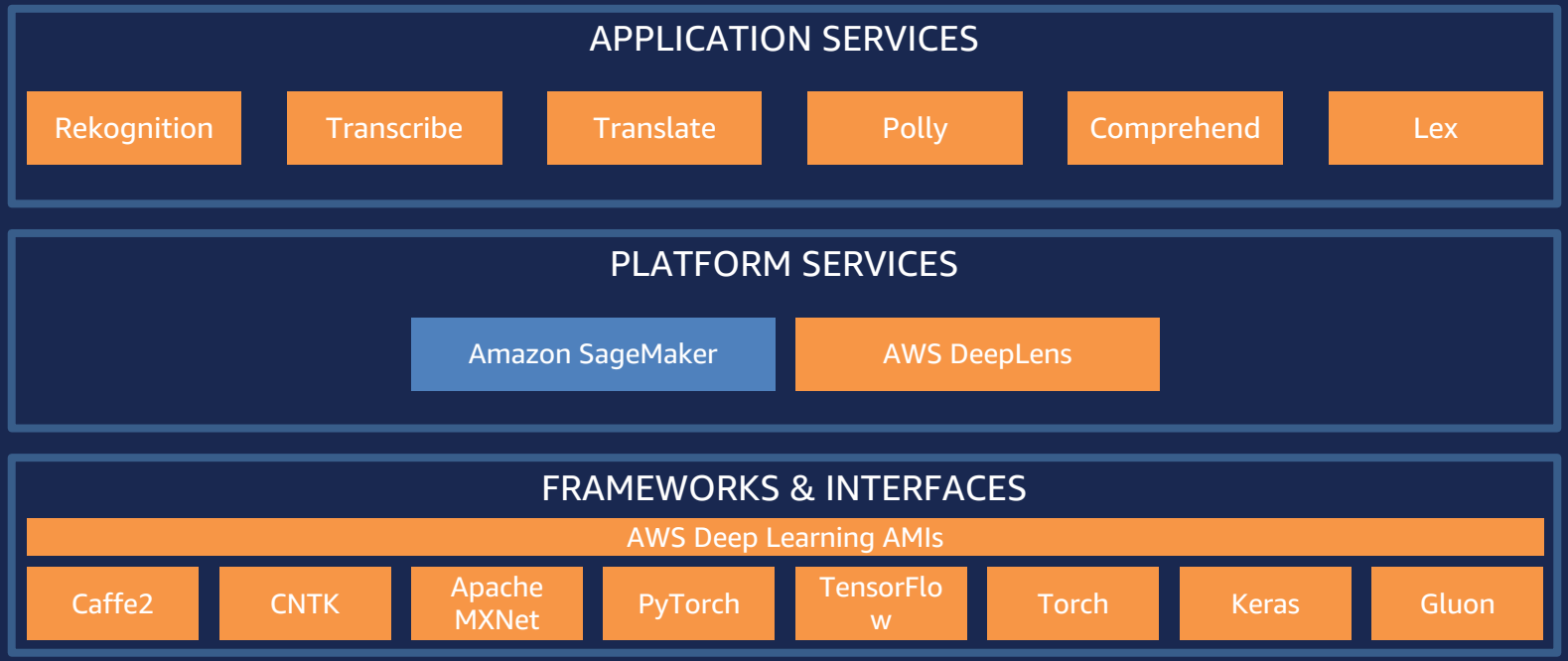
aws

Machine Learning at AWS

Our mission:

Put machine learning in the hands of every developer and data scientist

The Amazon Machine Learning Stack

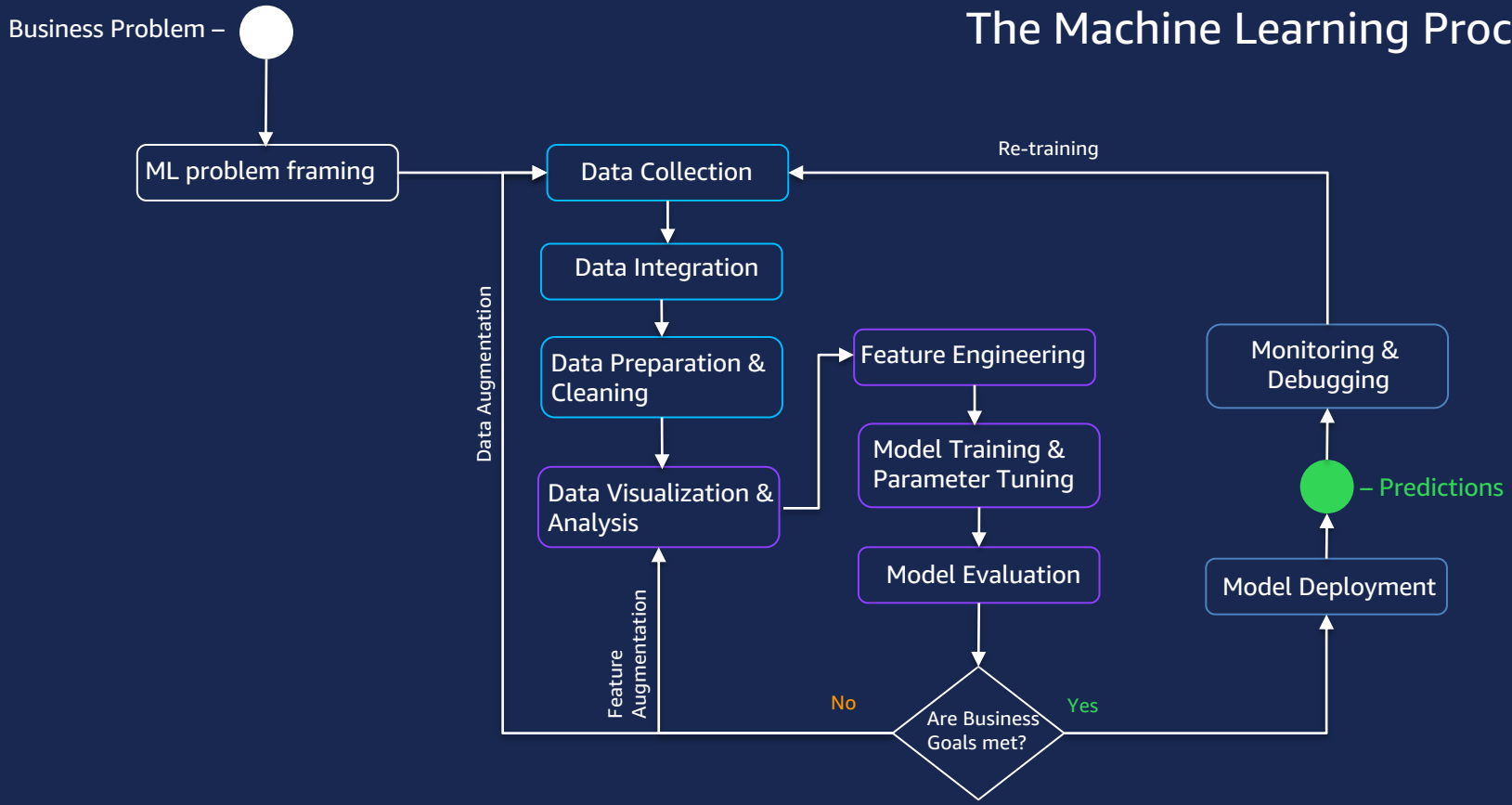


Recommender Workshop Agenda

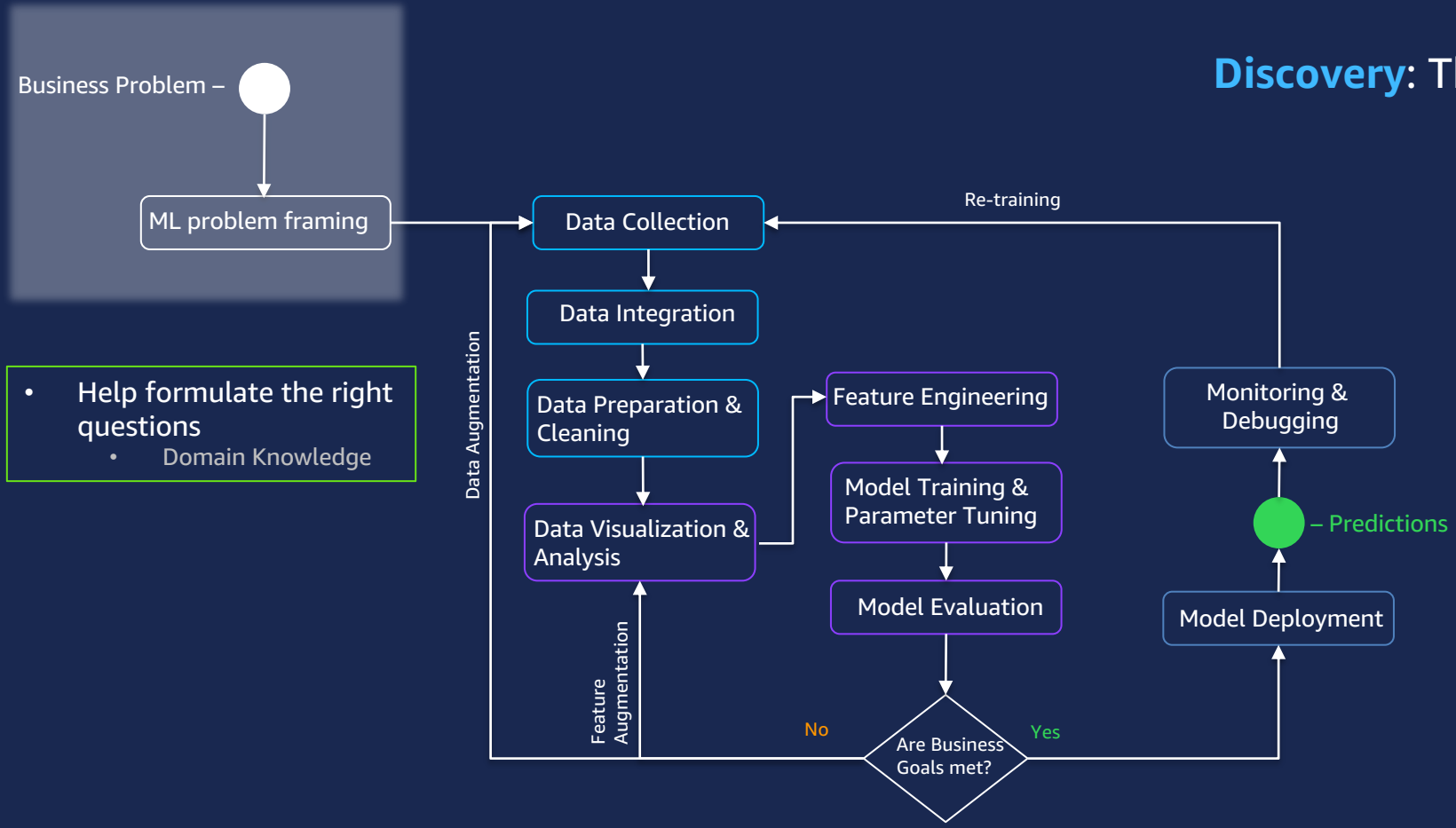
- Part 1: Introduction (You Are Here)
  - Overview of Machine Learning Process, Amazon SageMaker
  - Hands-on: Data Exploration
- Part 2: Collaborative Filtering
  - Core Concepts for Recommendations
  - Hands-on: K-Means Clustering
- Part 3: Matrix Factorization
  - Refining Recommendations
  - Hands-on: Factorization Machine
- Part 4: Hyperparameter Optimization

Let’s Review the ML Process

The Machine Learning Process

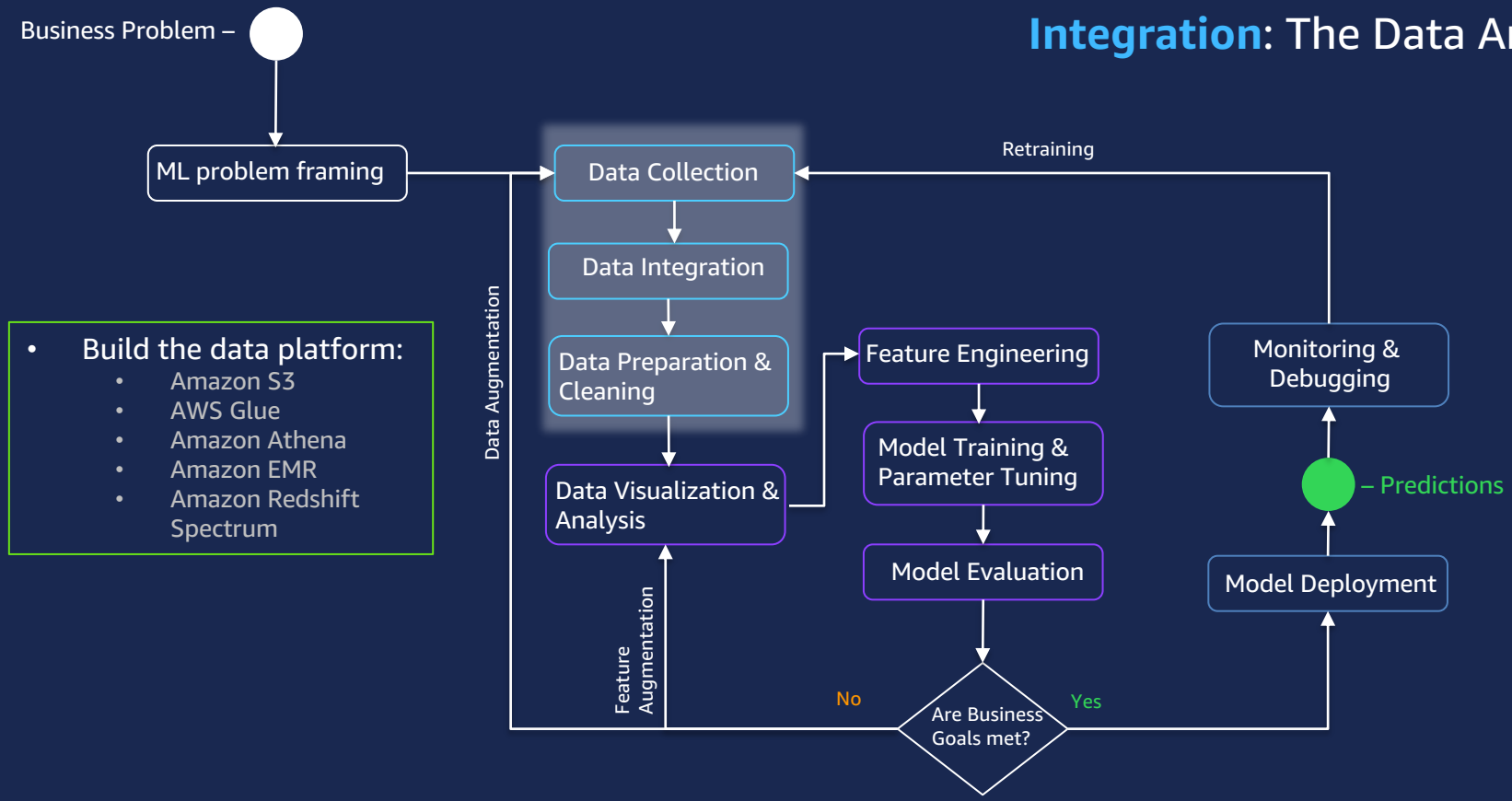


Discovery: The Analysts



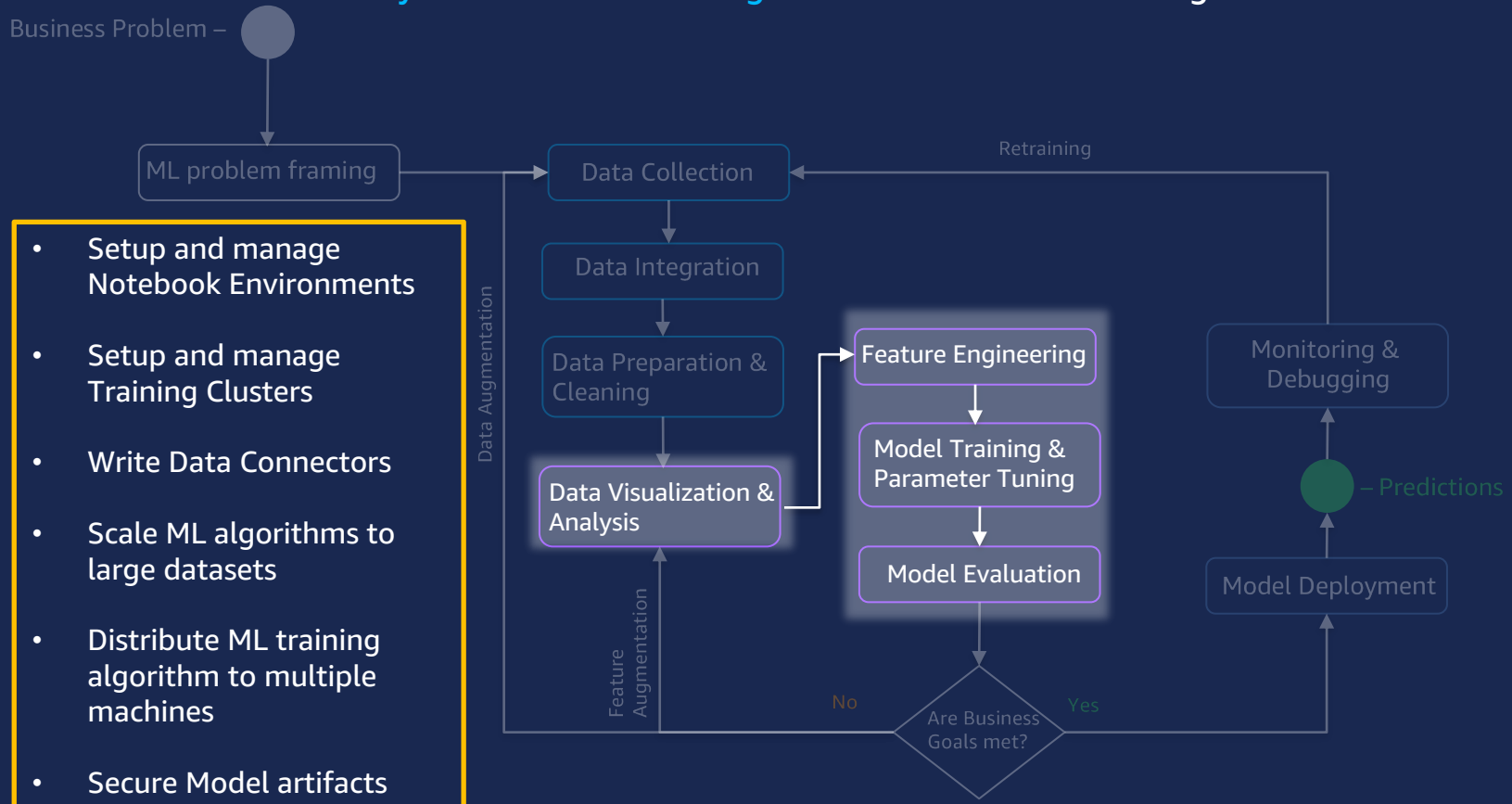
- Help formulate the right questions
  - Domain Knowledge

Integration: The Data Architecture



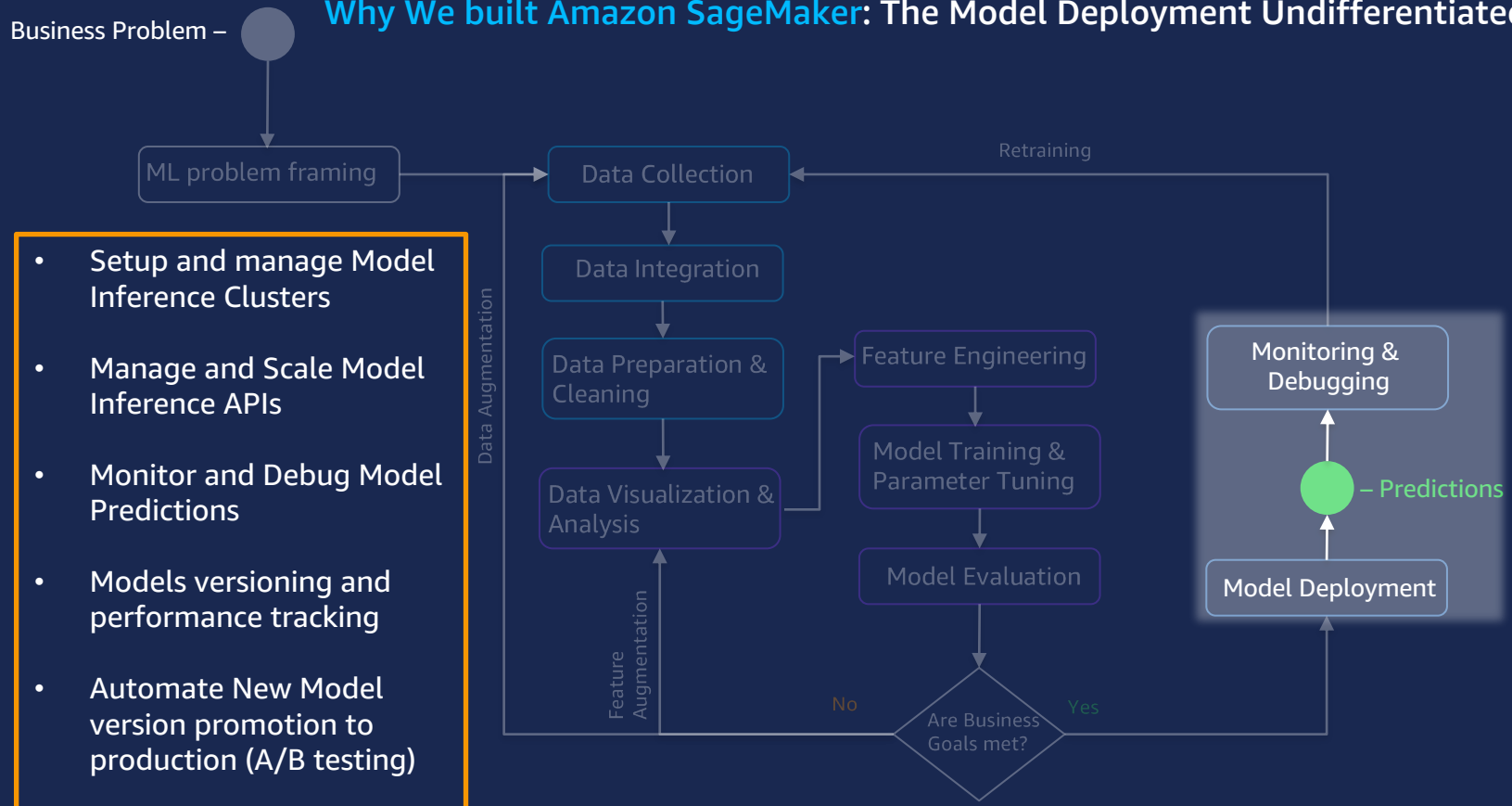
- Build the data platform:
  - Amazon S3
  - AWS Glue
  - Amazon Athena
  - Amazon EMR
  - Amazon Redshift Spectrum

Why We built Amazon SageMaker: The Model Training Undifferentiated Heavy Lifting



- Setup and manage Notebook Environments
- Setup and manage Training Clusters
- Write Data Connectors
- Scale ML algorithms to large datasets
- Distribute ML training algorithm to multiple machines
- Secure Model artifacts

Why We built Amazon SageMaker: The Model Deployment Undifferentiated Heavy Lifting



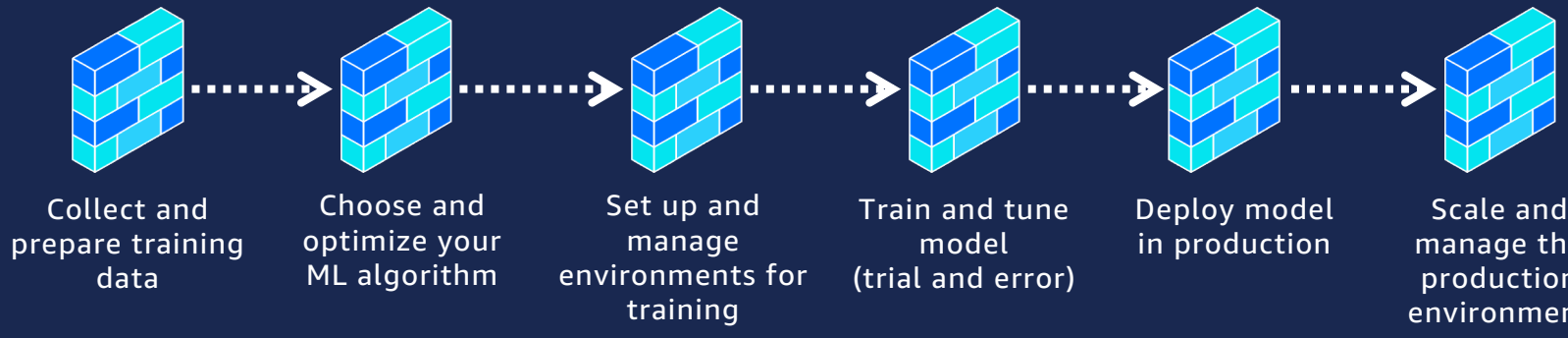
- Setup and manage Model Inference Clusters
- Manage and Scale Model Inference APIs
- Monitor and Debug Model Predictions
- Models versioning and performance tracking
- Automate New Model version promotion to production (A/B testing)

aws

Amazon SageMaker

Amazon SageMaker

Easily build, train, and deploy machine learning models



Amazon SageMaker

Easily build, train, and deploy machine learning models

Pre-built notebooks for common problems

Built-in, high performance algorithms

BUILD

Set up and manage environments for training

Train and tune model (trial and error)

Deploy model in production

Scale and manage the production environment

aws

Training ML Models Using Amazon SageMaker

SageMaker Built-in Algorithms

K-means Clustering
PCA
Neural Topic Modelling
Factorisation Machines
Linear Learner
XGBoost
Latent Dirichlet Allocation
Image Classification
Seq2Seq
DeepAR Forecasting
BlazingText (word2vec)
Random Cut Forest

aws

Training ML Models Using Amazon SageMaker

SageMaker Built-in Algorithms

K-means Clustering
PCA
Neural Topic Modelling
Factorisation Machines
Linear Learner – Regression
XGBoost
Latent Dirichlet Allocation
Image Classification
Seq2Seq
Linear Learner – Classification
DeepAR Forecasting

Bring Your Own Algorithms

ML Algorithms
R
MXNet
TensorFlow
Caffe
PyTorch
Keras
CNTK
…

aws

Training ML Models Using Amazon SageMaker

SageMaker Built-in Algorithms

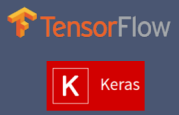
K-means Clustering
PCA
Neural Topic Modelling
Factorisation Machines
Linear Learner – Regression
XGBoost
Latent Dirichlet Allocation
Image Classification
Seq2Seq
Linear Learner – Classification
DeepAR Forecasting

Bring Your Own Algorithms

ML Algorithms
R
MXNet
TensorFlow
Caffe
PyTorch
Keras
CNTK
…

MXNet & TensorFlow SDK

TensorFlow SDK
MXNet (Gluon) SDK



aws

Training ML Models Using Amazon SageMaker

SageMaker Built-in Algorithms

K-means Clustering
PCA
Neural Topic Modelling
Factorisation Machines
Linear Learner – Regression
XGBoost
Latent Dirichlet Allocation
Image Classification
Seq2Seq
Linear Learner – Classification
DeepAR Forecasting

Bring Your Own Algorithms

ML Algorithms
R
MXNet
TensorFlow
Caffe
PyTorch
Keras
CNTK
…

MXNet & TensorFlow SDK

TensorFlow SDK
MXNet (Gluon) SDK

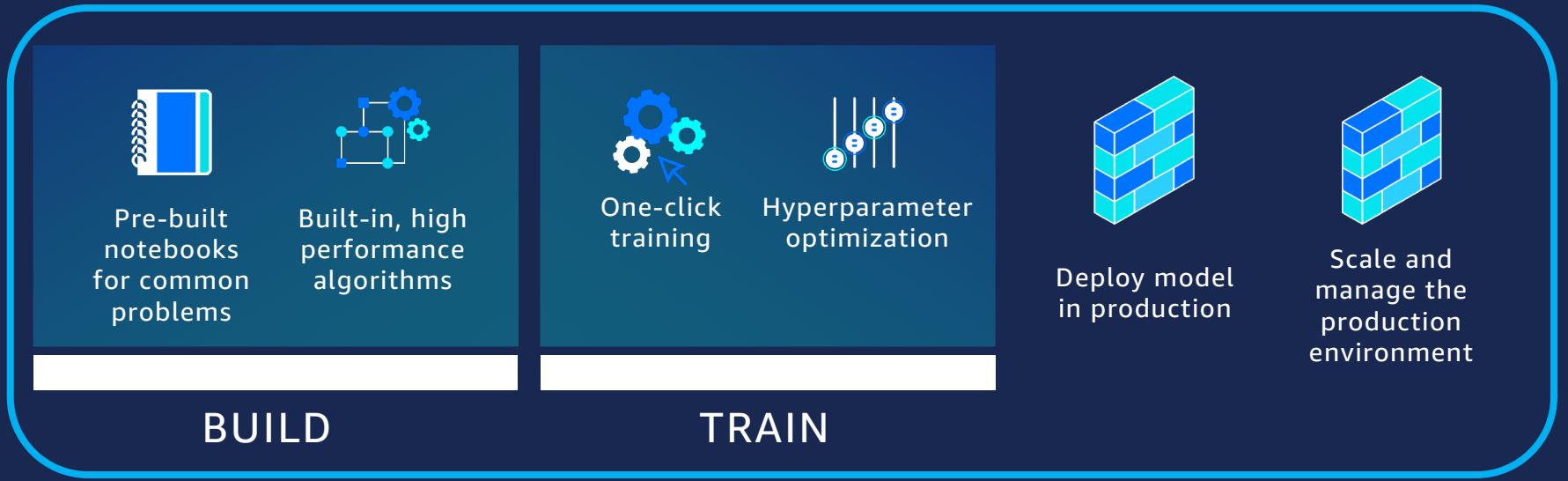
Apache Spark Estimator

Apache Spark Python library
Apache Spark Scala library

aws

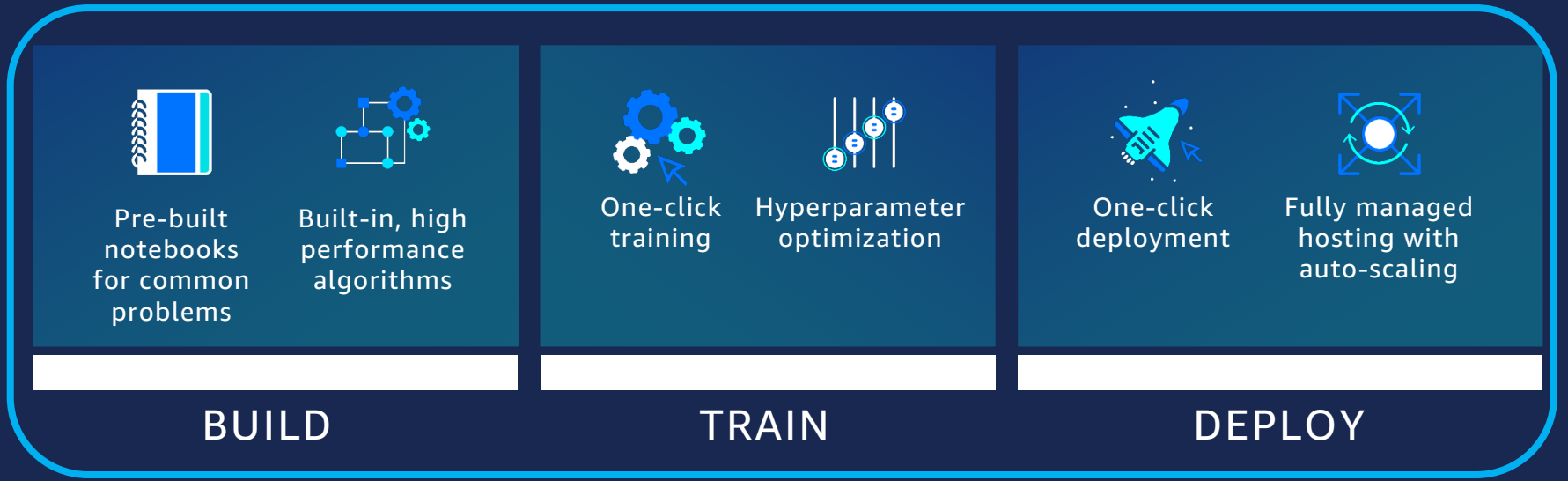
Amazon SageMaker

Easily build, train, and deploy machine learning models

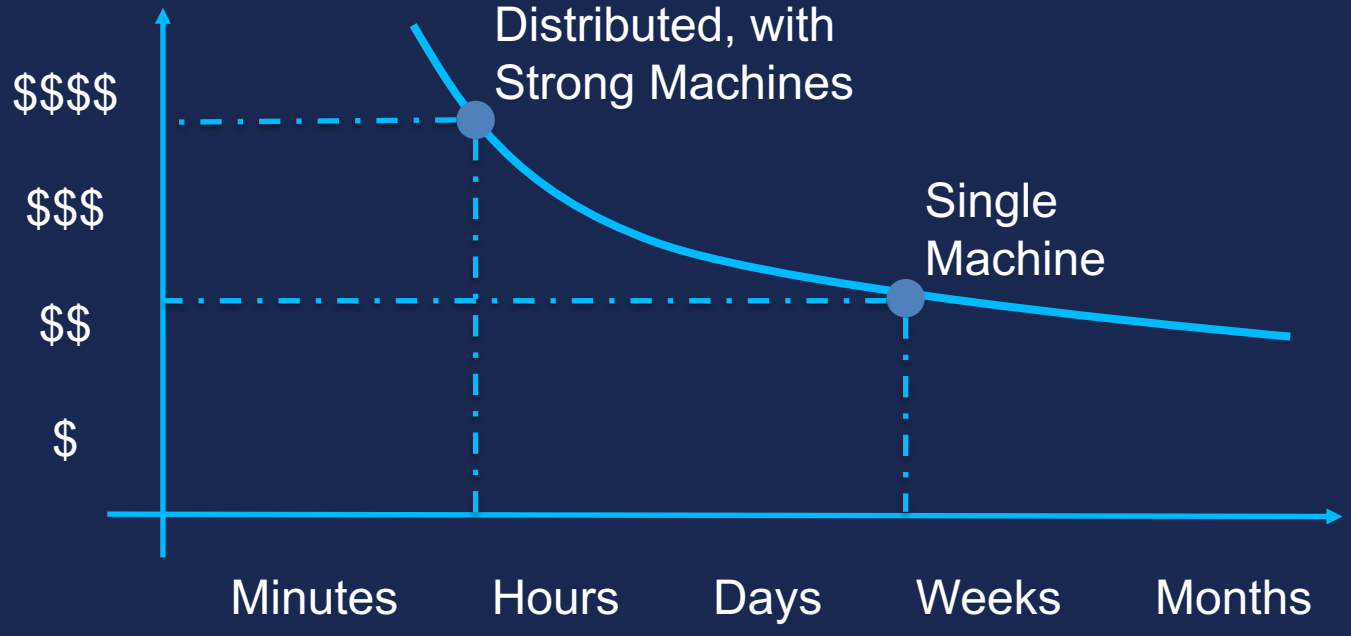


Amazon SageMaker

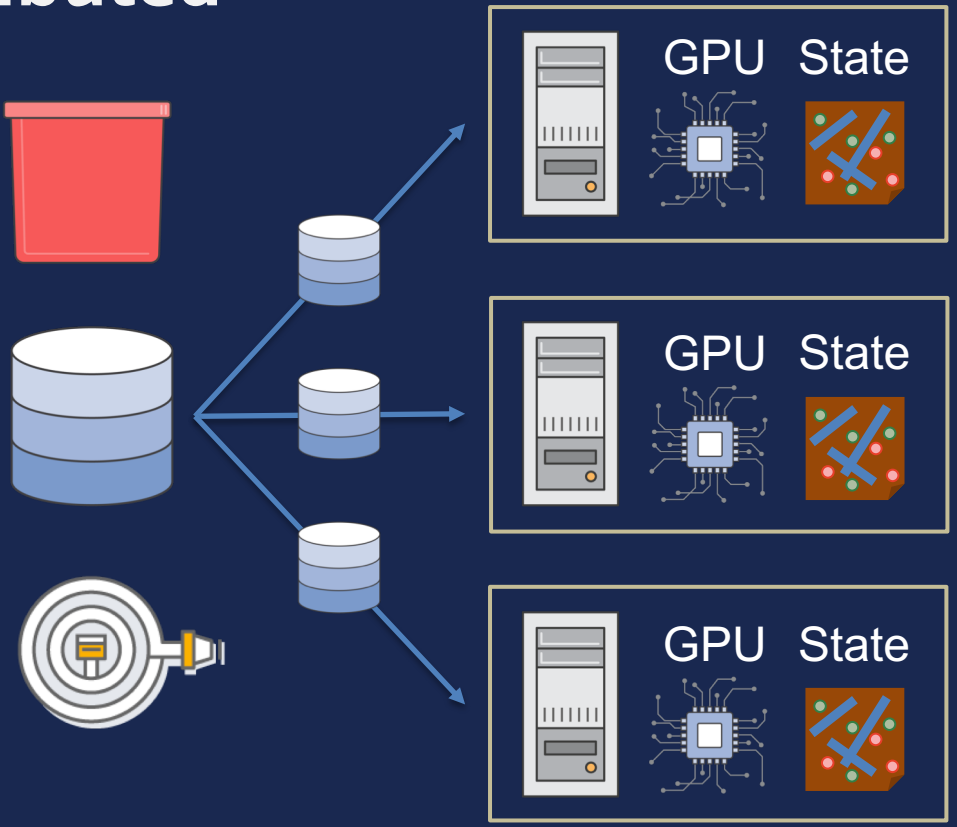
Easily build, train, and deploy machine learning models



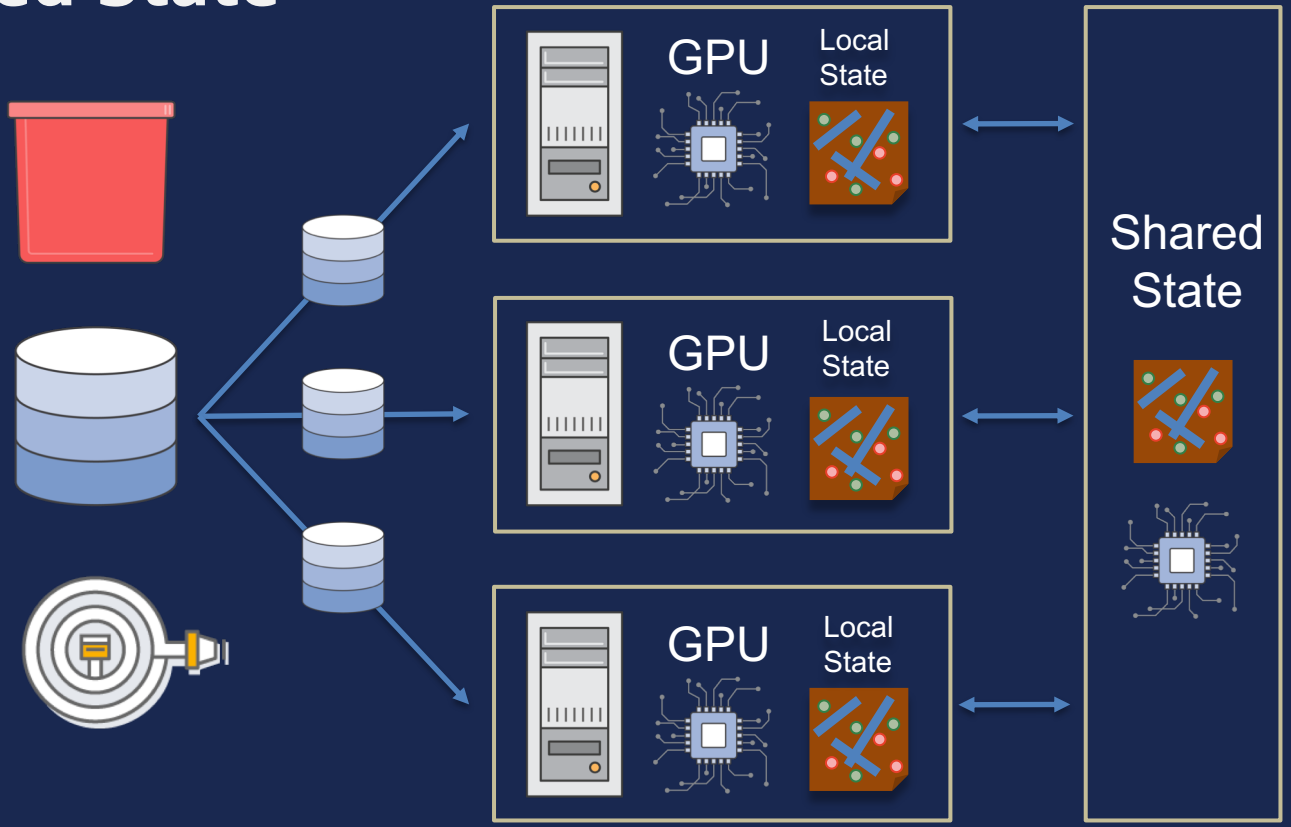
Cost vs. Time



Distributed



Shared State



Cost vs. Time

Recommender Workshop Repository

http://bit.ly/2wkaV0N

Our Data Set: Movielens

- Public Data Set produced by GroupLens Research
- https://grouplens.org/datasets/movielens/

```
In [15]: data = pd.read_csv("u.data", sep='\t', header=None,
             names=['userid', 'movieid', 'rating', 'timestamp'])
         data.head()
```

Out[15]:

	userid	movieid	rating	timestamp
0	196	242	3	881250949
1	186	302	3	891717742
2	22	377	1	878887116
3	244	51	2	880606923
4	166	346	1	886397596

Recommender Workshop Activity

- Log into the AWS console
- Change to us-east-1 region
- Find the Amazon SageMaker service
- Find Notebooks
- Spin up new notebook instance
    - Recommended: m1.m4 type
- Within notebook instance, open Terminal
    - `cd SageMaker`
    - `git clone https://github.com/shirkeyaws/sagemaker-recommender-workshop (aka http://bit.ly/2wkaV0N)`
- In Jupyter, within the repo path, find: 01_exploring_data.ipynb

aws

NEXT: Part 2