aws machine learning

Architecting for ML, On AWS

Mark Roy
Machine Learning Solutions Architect

 |

My role

Machine Learning Specialist

- Talking algorithms
- Proposing ML solutions
- Analyzing model behavior

Solutions Architect

- Breaking down a problem
- Helping you navigate AWS
- Facilitating your outcomes

Agenda

Day 1	Day 2	Day 3
AI/ML on AWS Intro lab	Feature engineering Model evaluation	Moving to production Build
Team up Define problem	Build	Present
Write-up	Working model	Solution architecture

aws machine learning

Centerpiece for digital transformation

Customer experience

Business operations

Decision making

Innovation

Competitive advantage

40% of digital transformation initiatives supported by AI in 2019

—IDC 2018

 |

aws machine learning

Our mission at AWS

Put machine learning in the hands of every developer

aws machine learning

Why AWS for ML

Broadest and deepest set of AI and ML services

200 new features and services launched this last year alone

Unmatched flexibility

Accelerate your adoption of ML with SageMaker

70% cost reduction in data-labeling

10x faster performance

75% lower inference cost

Built on the most comprehensive cloud platform optimized for ML

AWS holds the top spots on Stanford's benchmark, for fastest training time, lowest cost, lowest inference latency

More machine learning happens on AWS than anywhere else

10,000+ customers | 2x the customer references | 85% of TensorFlow projects in the cloud happen on AWS

 |

aws machine learning

Amazon ML stack: Broadest & deepest set of capabilities

AI SERVICES

Easily add intelligence to applications without machine learning skills

Vision | Documents | Speech | Language | Chatbots | Forecasting | Recommendations

aws machine learning

Amazon ML stack: Broadest & deepest set of capabilities

AI SERVICES

Easily add intelligence to applications without machine learning skills

Vision | Documents | Speech | Language | Chatbots | Forecasting | Recommendations

ML SERVICES

Build, train, and deploy machine learning models fast

Data labeling | Pre-built algorithms & notebooks | One-click training and deployment

aws machine learning

Amazon ML stack: Broadest & deepest set of capabilities

AI SERVICES

Easily add intelligence to applications without machine learning skills

Vision | Documents | Speech | Language | Chatbots | Forecasting | Recommendations

ML SERVICES

Build, train, and deploy machine learning models fast

Data labeling | Pre-built algorithms & notebooks | One-click training and deployment

ML FRAMEWORKS & INFRASTRUCTURE

Flexibility & choice, highest-performing infrastructure

Support for ML frameworks | Compute options purpose-built for ML

aws machine learning

Amazon ML stack: Broadest & deepest set of capabilities

AI Services

VISION			SPEECH		LANGUAGE		CHATBOTS	FORECASTING	RECOMMENDATIONS
REKOGNITION IMAGE	REKOGNITION VIDEO	TEXTRACT	POLLY	TRANSCRIBE	TRANSLATE	COMPREHEND & COMPREHEND MEDICAL	LEX	FORECAST	PERSONALIZE

ML Services

Amazon SageMaker

Ground Truth	Notebooks	Algorithms + Marketplace	Reinforcement Learning	Training	Optimization	Deployment	Hosting

ML Frameworks + Infrastructure

FRAMEWORKS	INTERFACES
TensorFlow, mxnet, PYTORCH	GLUON, Keras

INFRASTRUCTURE							
EC2 P3 & P3DN	EC2 G4	EC2 C5	FPGAS	DL CONTAINERS & AMIs	GREENGRASS	ELASTIC INFERENCE	INFERENTIA

aws machine learning

Amazon SageMaker

Bringing machine learning to all developers

Collect and prepare training data

Choose and optimize your ML algorithm

Set up and manage environments for training

Train and tune model (trial and error)

Deploy model in production

Scale and manage the production environment

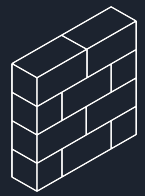
Amazon SageMaker

Bringing machine learning to all developers

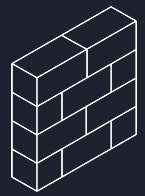
Pre-built notebooks for common problems

Collect and prepare training data

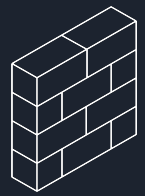
Choose and optimize your ML algorithm



Set up and manage environments for training



Train and tune model (trial and error)



Deploy model in production

Scale and manage the production environment

Amazon SageMaker

Bringing machine learning to all developers

Pre-built notebooks for common problems

Collect and prepare training data

Built-in, high performance algorithms

Choose and optimize your ML algorithm

- K-Means Clustering
- Principal Component Analysis
- Neural Topic Modelling
- Factorization Machines
- Linear Learner (Regression)
- BlazingText
- Reinforcement learning
- XGBoost
- Topic Modeling (LDA)
- Image Classification
- Seq2Seq
- Linear Learner (Classification)
- DeepAR Forecasting

aws machine learning

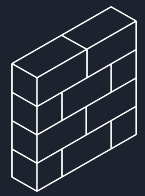
Amazon SageMaker

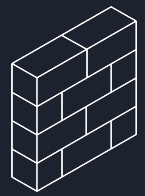
Bringing machine learning to all developers

Pre-built notebooks for common problems

Built-in, high performance algorithms

One-click training





Collect and prepare training data

Choose and optimize your ML algorithm

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 machine learning

Amazon SageMaker

Bringing machine learning to all developers

Pre-built notebooks for common problems

Built-in, high performance algorithms

One-click training

Optimization

Collect and prepare training data

Choose and optimize your ML algorithm

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 machine learning

Amazon SageMaker

Bringing machine learning to all developers

Pre-built notebooks for common problems

Collect and prepare training data

Built-in, high performance algorithms

Choose and optimize your ML algorithm

One-click training

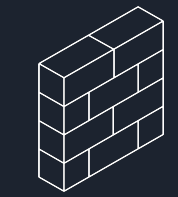
Set up and manage environments for training

Optimization

Train and tune model (trial and error)

One-click deployment

Deploy model in production



Scale and manage the production environment

aws machine learning

 |

Amazon SageMaker

Bringing machine learning to all developers

Pre-built notebooks for common problems	Built-in, high performance algorithms	One-click training	Optimization	One-click deployment	Fully managed with auto-scaling, health checks, automatic handling of node failures, and security checks
					
Collect and prepare training data	Choose and optimize your ML algorithm	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 machine learning

Custom machine learning for your business

AMAZON SAGEMAKER

REDUCE COSTS

70%
cost reduction for data labeling using Ground Truth

75%
cost reduction for inference with Elastic Inference

INCREASE PERFORMANCE

10x
better algorithm performance

2x
performance increases from model optimization with Neo

EASE-OF-USE

One-click
model training and deployment

Train once
run anywhere

The best place to run TensorFlow

Amazon SageMaker is the best place to run TensorFlow in the cloud

- Fully-managed training and hosting
- Near-linear scaling across 100s of GPU
- 75% lower inference costs with Amazon Elastic Inference
- 3x faster network throughput with EC2 P3

65% Stock TensorFlow

90% AWS-optimized TensorFlow

Scaling efficiency with 256 GPUs

aws machine learning

Advancing pharma research and discovery

Celgene uses Apache MXNet on Amazon SageMaker for toxicology prediction to virtually analyze biological impacts of potential drugs without putting patients at risk. A model that once took 2 months to train can now be trained in 4 hours.

Fueling product innovation

Using Amazon SageMaker, Intuit developed ML models that can pull a year's worth of bank transactions to find deductible business expenses for customers. Using SageMaker, Intuit reduced machine learning deployment time by 90%, from 6 months to 1 week.

Enhancing the fan experience

One week of NFL games now creates 3 TB of data. NFL uses Amazon SageMaker to analyze telemetry data to predict plays. Computations that could take months to refine now take only weeks or days.

WATCH VIDEO >>

 |

aws machine learning

Driving better healthcare outcomes

Using Amazon SageMaker, GE Healthcare developed an ML model that can learn from thousands of medical scans to detect anomalies more accurately and efficiently, allowing radiologists to prioritize patients needing immediate attention.

Optimizing supply chain operations

Using Amazon SageMaker, Convoy builds and trains machine learning models to optimize driver schedules and to ensure load balancing, capacity planning, pricing and payments.

Better decisions through predictions

MLB uses Amazon SageMaker to predict stolen-base success using game data. SageMaker eliminates manual, time-intensive processes like scorekeeping, capturing game notes, and classifying pitches.

WATCH VIDEO >>

Accelerating financial analysis

Using TensorFlow on Amazon SageMaker, Siemens Financial Services developed an NLP model to extract critical information to accelerate investment due diligence, reducing time to summarize diligence documents from 12 hours down to 30 seconds.

SIEMENS

aws machine learning

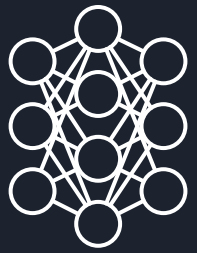
Personalizing the gaming experience

Using Amazon SageMaker, Sony Interactive Entertainment modernized the PlayStation Store, using predictive ML to drive highly personalized customer experiences, improve enterprise data reporting, and drive product feature innovation.

Amazon EC2 P3dn instance
The largest P3 instance, optimized for distributed training

Reduce machine learning training time

Better GPU utilization

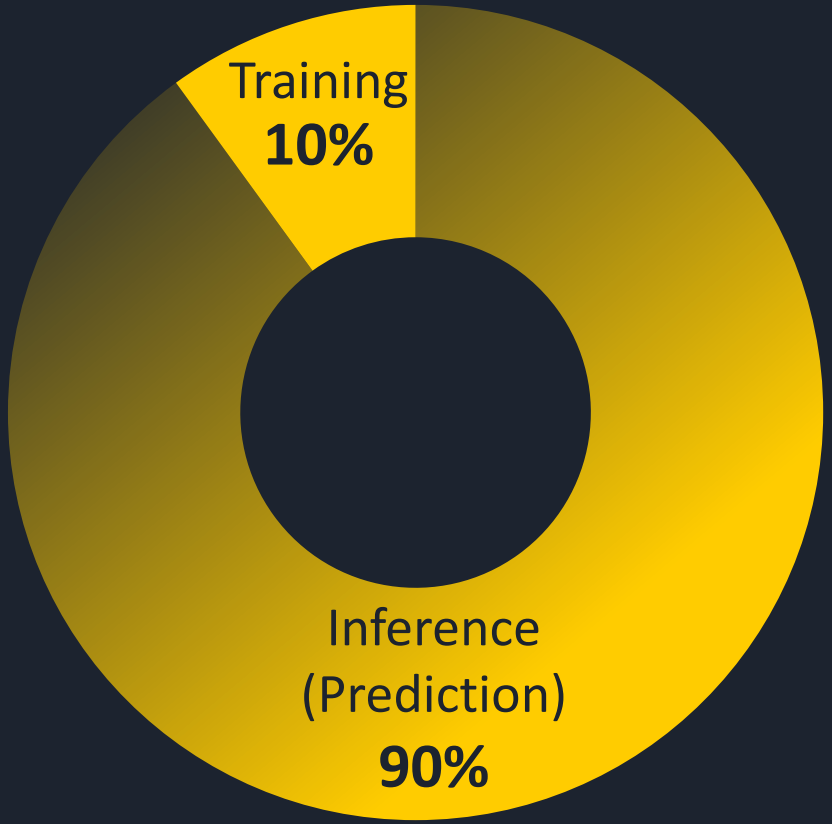


Support larger, more complex models

KEY FEATURES

100Gbps of networking bandwidth (4x > P3)	8 NVIDIA Tesla V100 GPUs	32GB of memory per GPU (2x > P3)	96 Intel Skylake vCPUs (50% more than P3) with AVX-512

Predictions drive complexity and cost in production



aws machine learning

Amazon Elastic Inference
Reduce deep learning inference costs up to 75%

Lower inference costs

Match capacity to demand

Available between 1 to 32 TFLOPS

KEY FEATURES

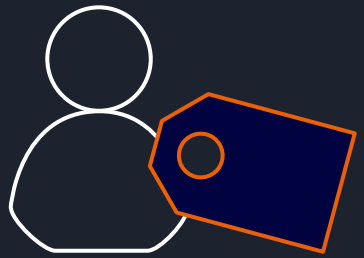
Integrated with Amazon EC2, Amazon SageMaker, and Amazon DL AMIs

Support for TensorFlow, Apache MXNet, and ONNX with PyTorch coming soon

Single and mixed-precision operations

Amazon SageMaker Ground Truth
Label machine learning training data easily and accurately

Quickly label training data



Easily integrate human labelers

Get accurate results

KEY FEATURES

Automatic labeling via machine learning

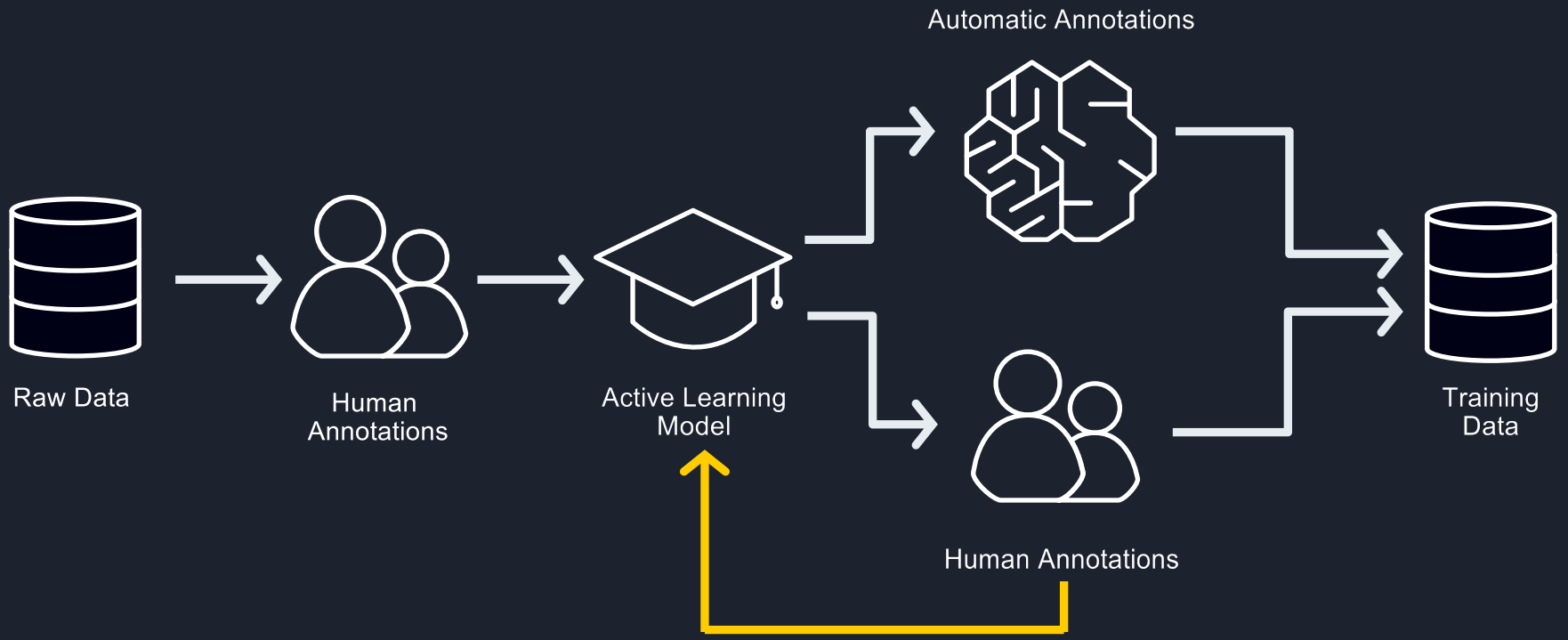
Ready-made and custom workflows for image bounding box, segmentation, and text

Private and public human workforce

Label management

aws machine learning

SageMaker Ground Truth: How it works



aws machine learning

AWS Marketplace for machine learning
ML algorithms and models available instantly

Browse or search
AWS Marketplace

Subscribe in a
single click

Available in
Amazon SageMaker

KEY FEATURES

SELLERS

- Automatic labeling via machine learning
- IP protection
- Automated billing and metering

BUYERS

- Broad selection of paid, free, and open-source algorithms and models
- Data protection

aws machine learning

Amazon SageMaker Neo
Train once, run anywhere with 2x the performance

Get accuracy and performance

Automatic optimization

mxnet
TensorFlow
PYTORCH

Broad framework support

intel
NVIDIA
arm

Broad hardware support

KEY FEATURES

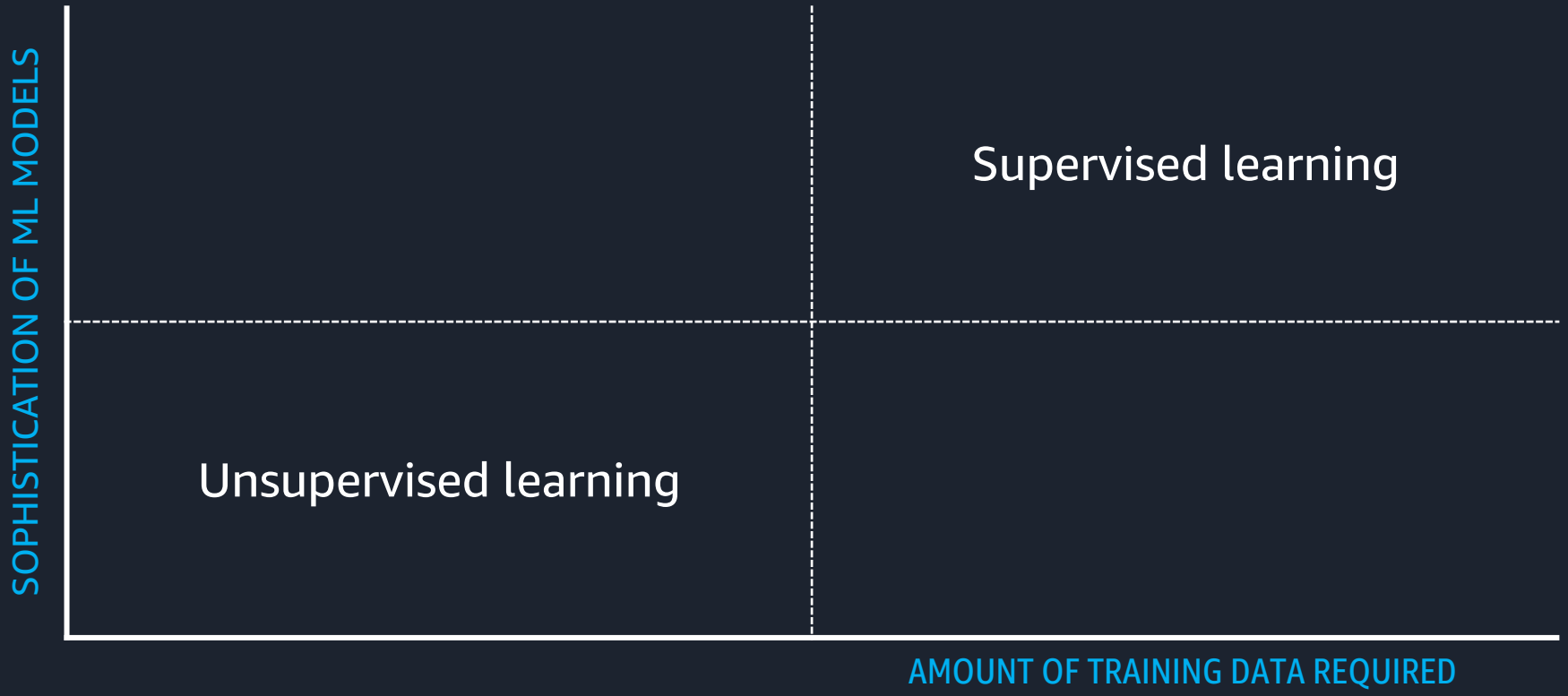
Compiler & run-time are open source

1/10th the size of original models

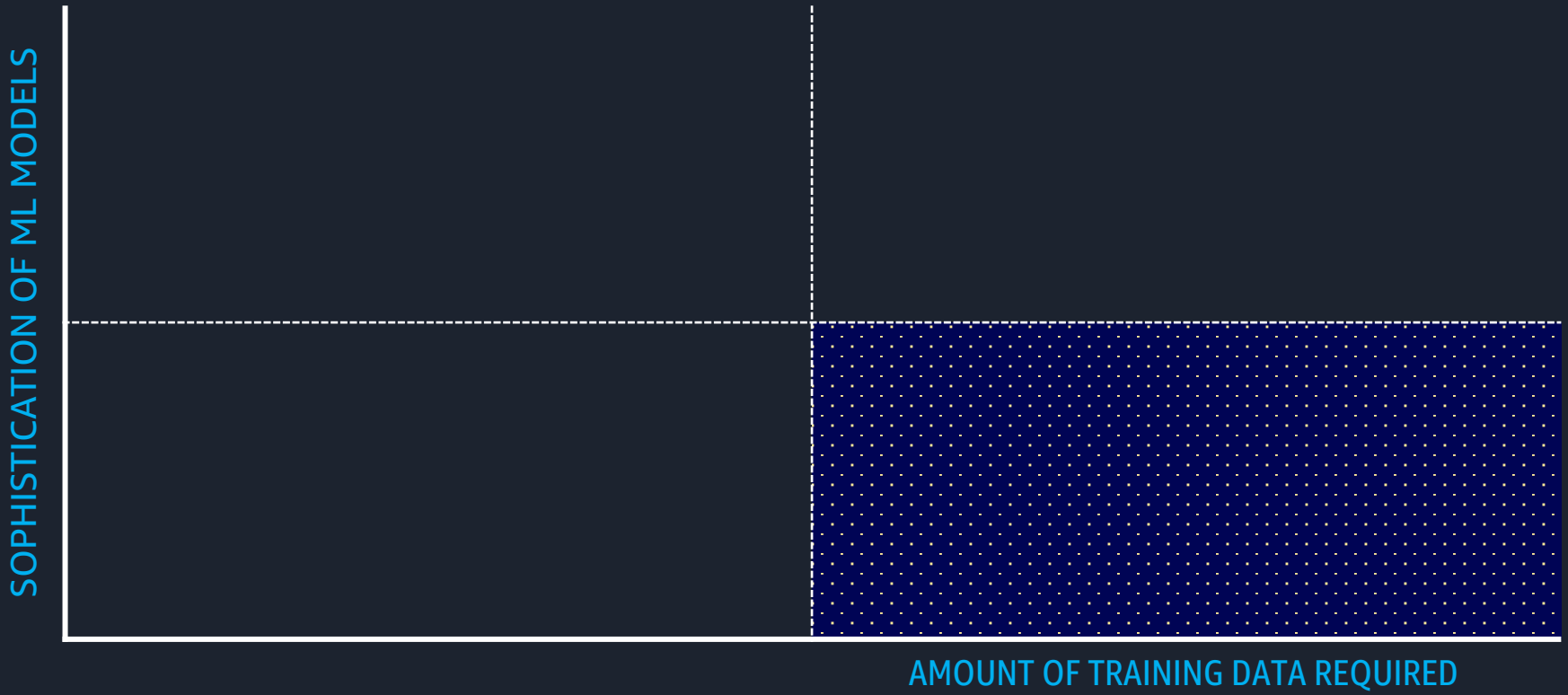
What's next for machine learning?

aws machine learning

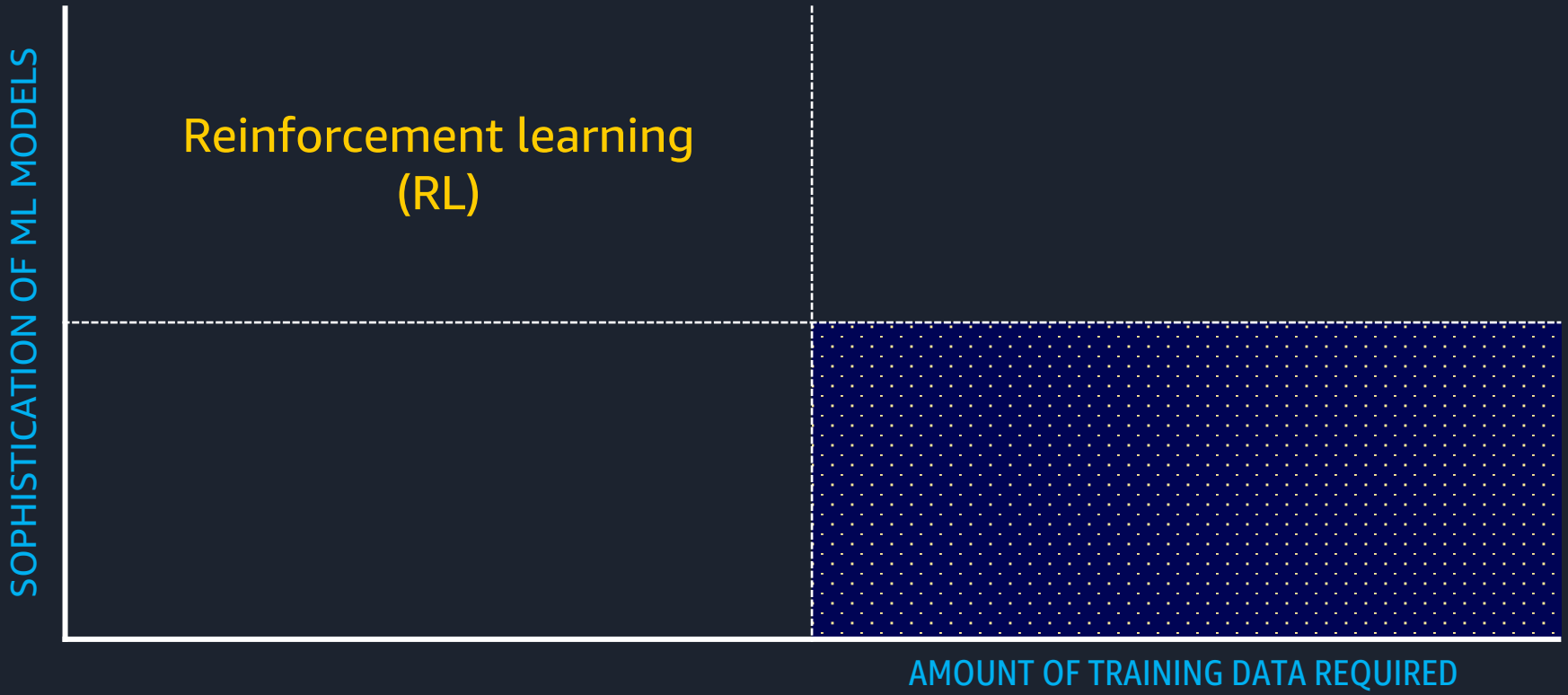
Types of Machine Learning



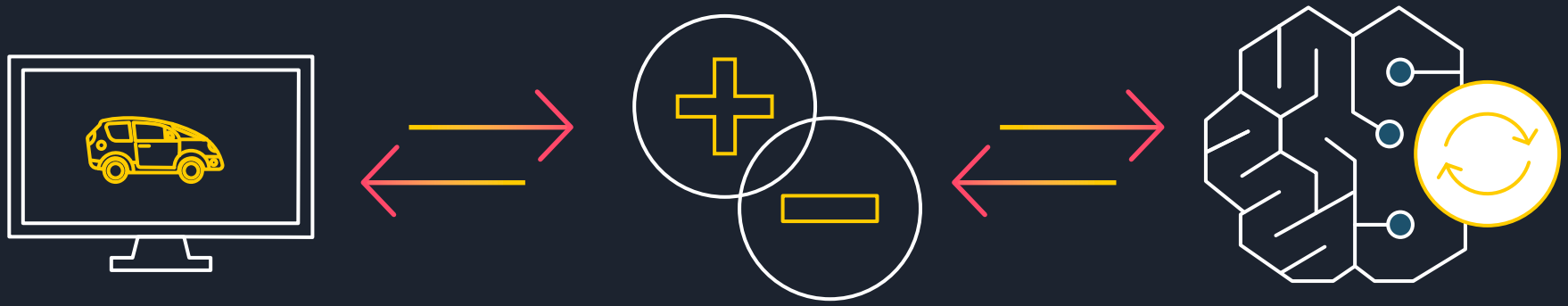
Types of Machine Learning



Types of Machine Learning



How does RL work?



Simulation environment

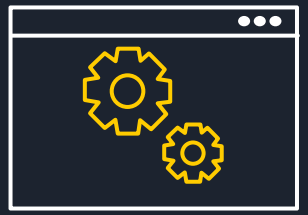
Scoring function

RL algorithm

USE CASES

Supply chain simulation, manufacturing process, robot manipulation, autonomous car, drone navigation…

Amazon SageMaker RL
Reinforcement learning for every developer and data scientist



Fully managed

mxnet
TensorFlow
intel riselab UC Berkeley

Broad support for frameworks

Broad support for simulation environments including SimuLink and MatLab

KEY FEATURES

TensorFlow, Apache MXNet, Intel Coach, and Ray RL support	2D & 3D physics environments and OpenAI Gym support	Supports Amazon Sumerian and Amazon RoboMaker	Example notebooks and tutorials

aws machine learning

Machine Learning Primer

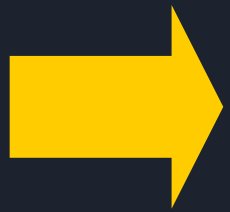
Machine Learning

gender	age	smoker	eye color	lung cancer
male	19	yes	green	no
female	44	yes	gray	yes
male	49	yes	blue	yes
male	12	no	brown	no
female	37	no	brown	no
female	60	no	brown	yes
male	44	no	blue	no
female	27	yes	brown	no
female	51	yes	green	yes
female	81	yes	gray	no

Machine Learning

gender	age	smoker	eye color
male	19	yes	green
female	44	yes	gray
male	49	yes	blue
male	12	no	brown
female	37	no	brown
female	60	no	brown
male	44	no	blue
female	27	yes	brown
female	51	yes	green
female	81	yes	gray

lung cancer
no
yes
yes
no
no
yes
no
no
yes
no



Train
ML Model

male	77	yes	gray
male	19	yes	green
female	44	no	gray

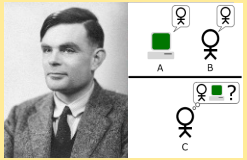
yes
no
no

Make predictions

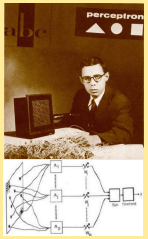
AI, Machine Learning, and Deep Learning

Artificial Intelligence

Any techniques that allows computer to mimic human intelligence



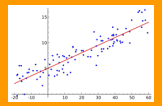
Turing Test

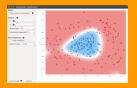


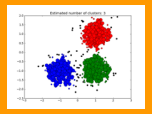
Perceptron

Machine Learning

A technique that allows computer to perform tasks without being explicitly programmed

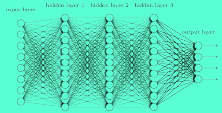






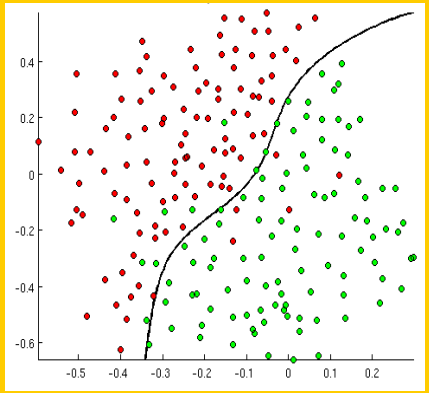
Deep Learning

A subfield of machine learning that uses neural network to learn

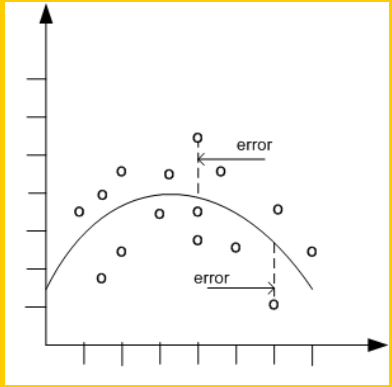


Some common classes of machine learning problems

Classification

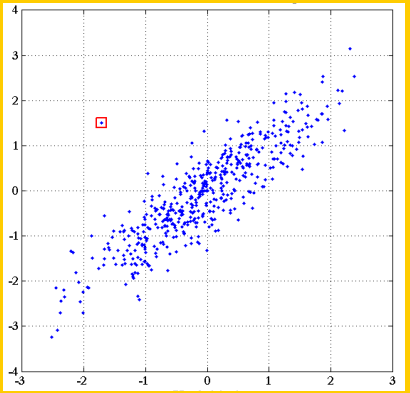


Regression



Recommenders

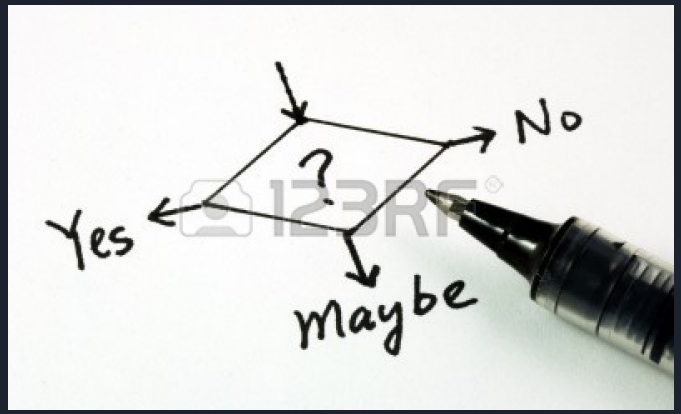
Anomaly Detection



 |

aws machine learning

Requirements for Problem solving with ML



Valuable business problem involving decision

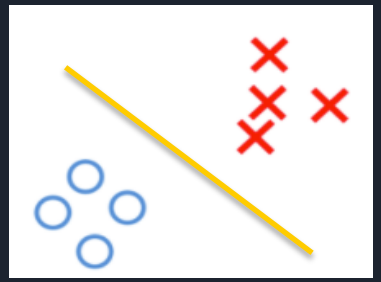
- Existing process
- Metrics

Available data

- Related to the decision
- Historical
- Outcomes

3 Types of Machine Learning

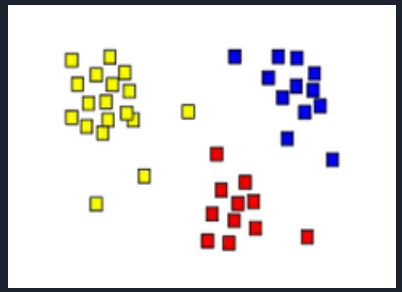
Supervised Machine Learning



Task driven

- Classification, Regression
- Training Data: (X,Y) (Features, Labels)
- Predict: Y, minimizing loss

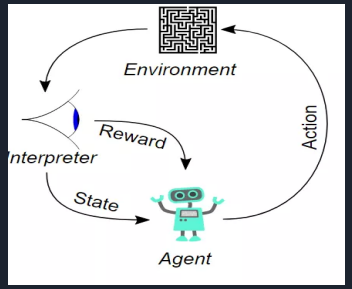
Unsupervised Machine Learning



Data driven

- Clustering
- Training Data: X (features only)
- Find similar points in high dimensional X-space

Reinforcement Learning



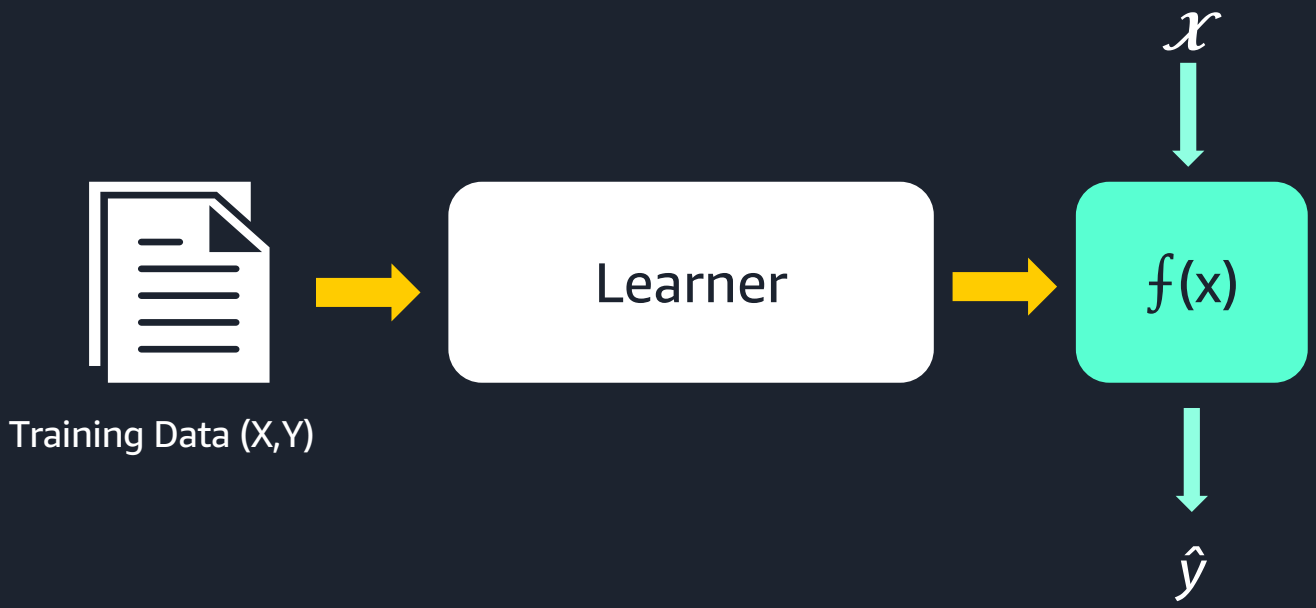
Decision making

- Robotics, games
- Training data: (State, Action, Reward)
- Maximize long term rewards

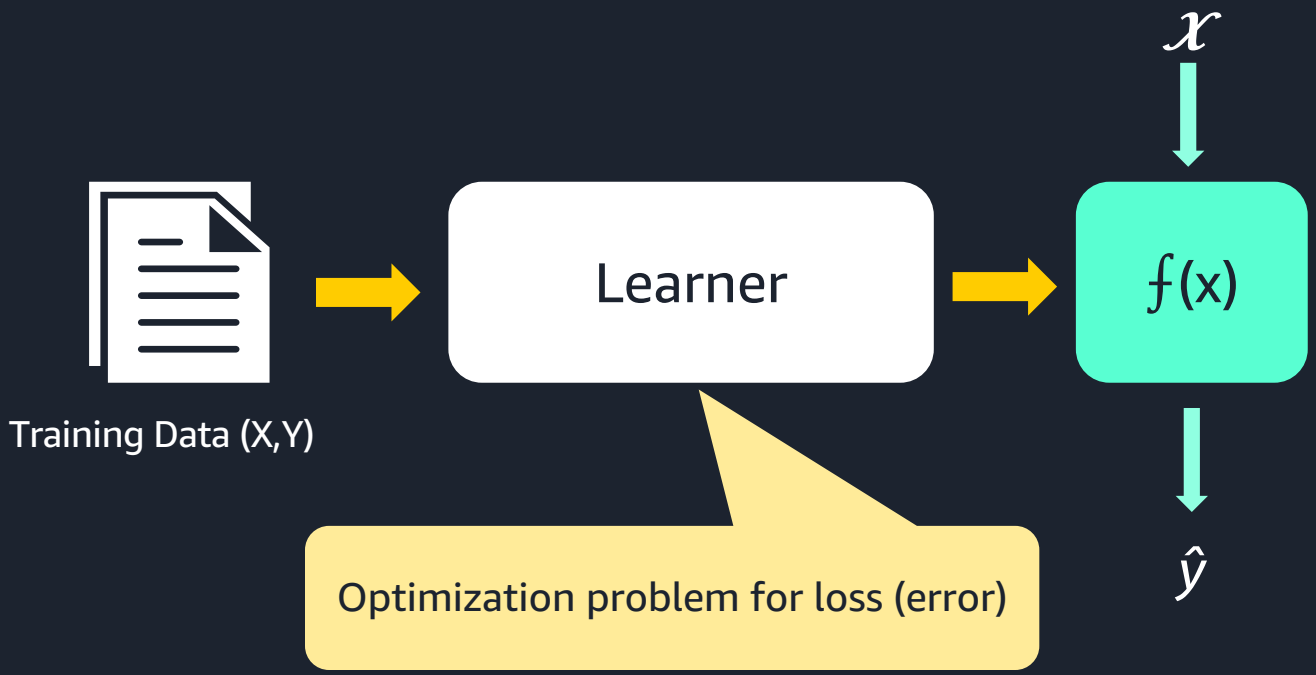
Example scenarios

SCENARIO	ML TECHNIQUE	ML TYPE
1. Identify company logos in pictures	Classification	Supervised
2. Estimate the sales price of a house in a market	Regression	Supervised
3. Forecast the number of visitors at a museum	Regression	Supervised
4. Recommend a book to a reader	Collaborative Filtering	Supervised
5. Segment customers into different groups	Clustering	Unsupervised
6. Transform correlated features into uncorrelated features	Dimensionality Reduction	Unsupervised
7. Robots learn to pick up objects		Reinforcement
8. Stock market bots learn how to make trade		Reinforcement

The Learning Framework



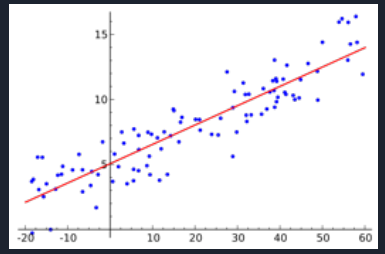
Loss Minimization



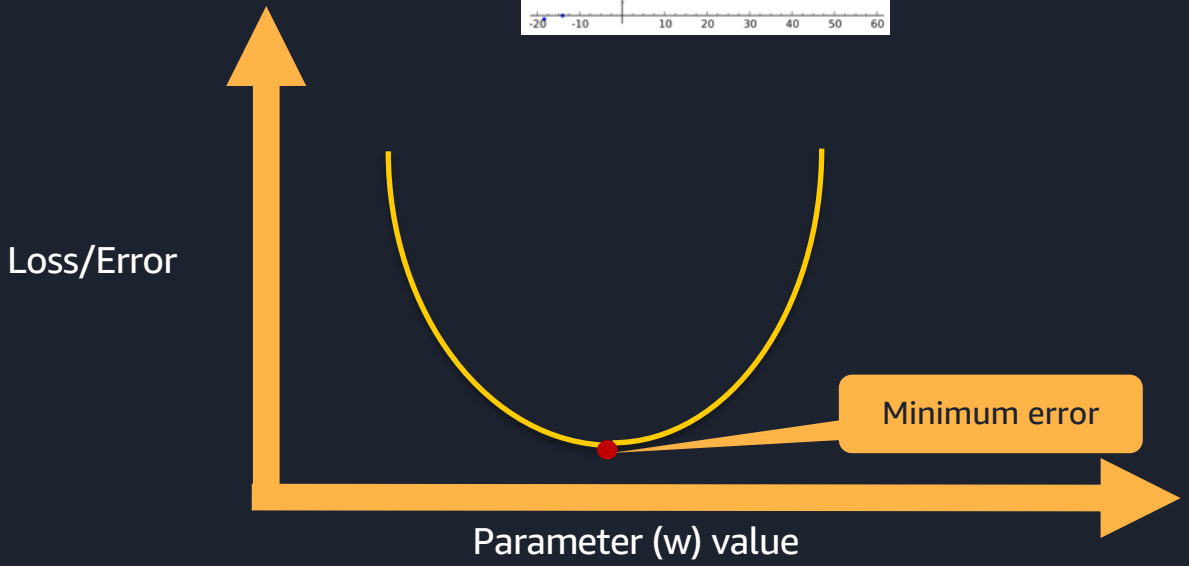
aws machine learning

Loss curve

Linear Regression

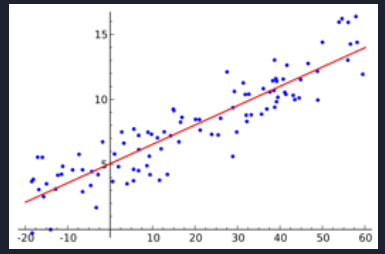


$f(x) = W*x + b$



Optimization algorithm

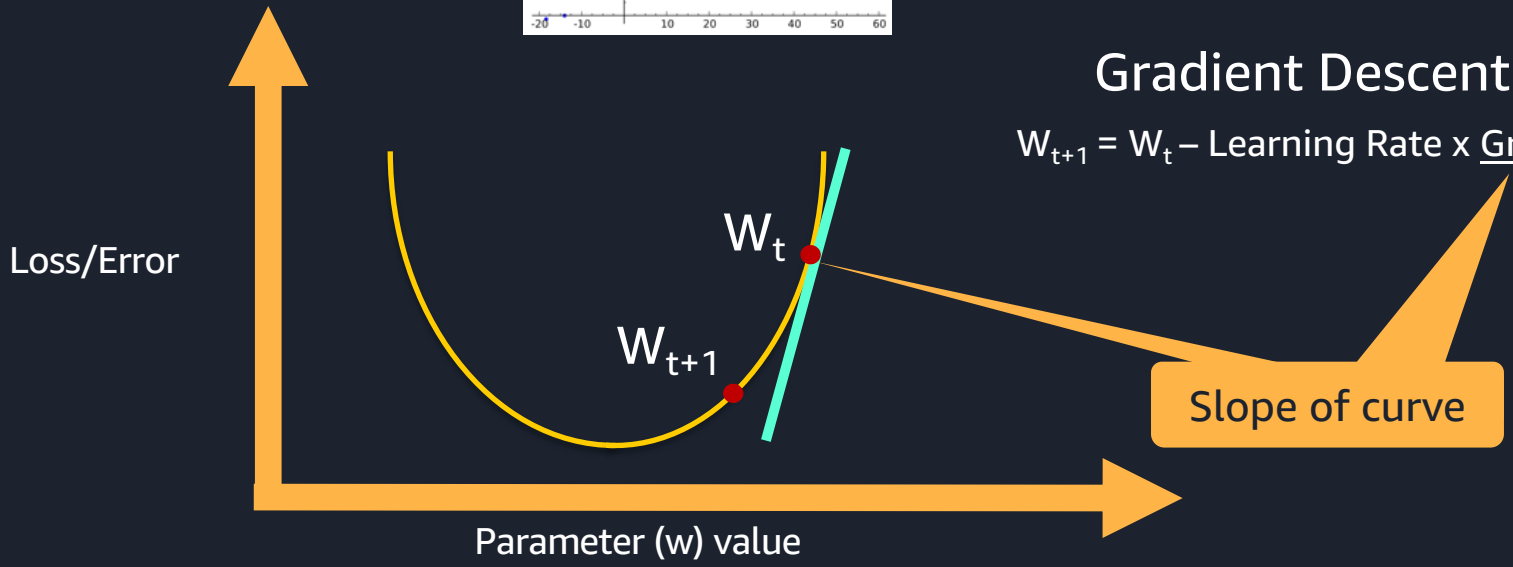
Linear Regression



$f(x) = W*x + b$

Gradient Descent

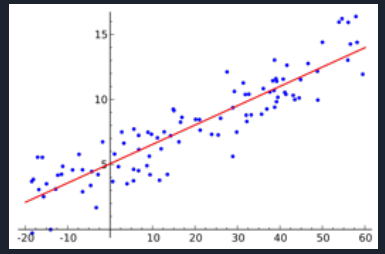
$W_{t+1} = W_t$ – Learning Rate x Gradient



aws machine learning

Optimization algorithm

Linear Regression



$f(x) = W*x + b$

Gradient Descent

$W_{t+1} = W_t$ – Learning Rate x Gradient

Loss/Error

W_t

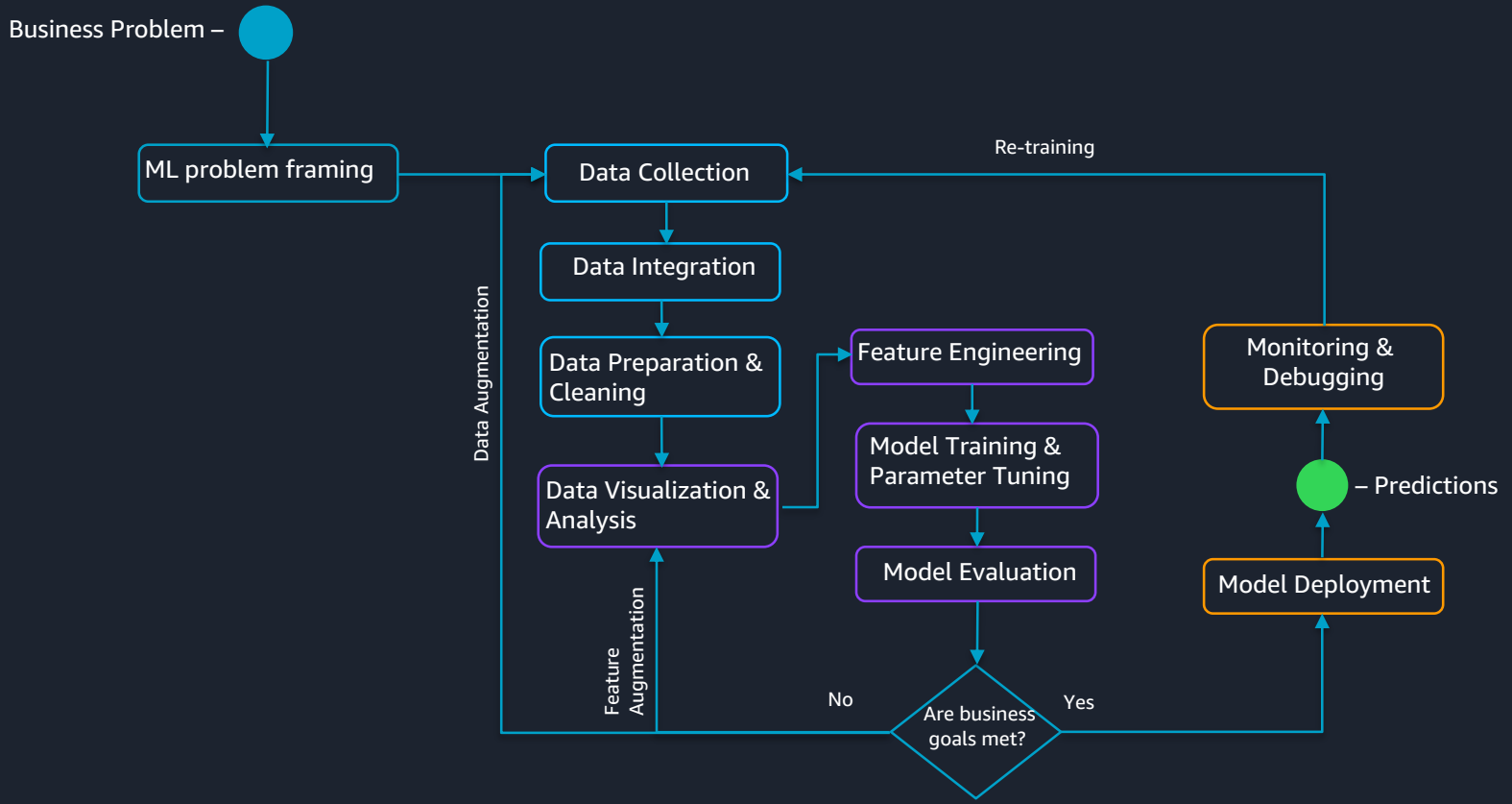
W_{t+1}

Change weight

Parameter (w) value

aws machine learning

Machine learning process



aws machine learning

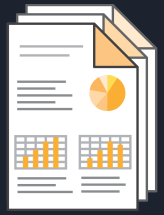
aws machine learning

Amazon SageMaker

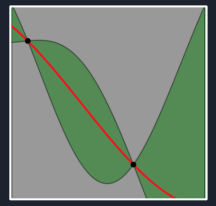
Amazon SageMaker Components

Hosted Notebook

Training Service



Built-in Algorithms



Hyper-parameter Optimization

Hosting

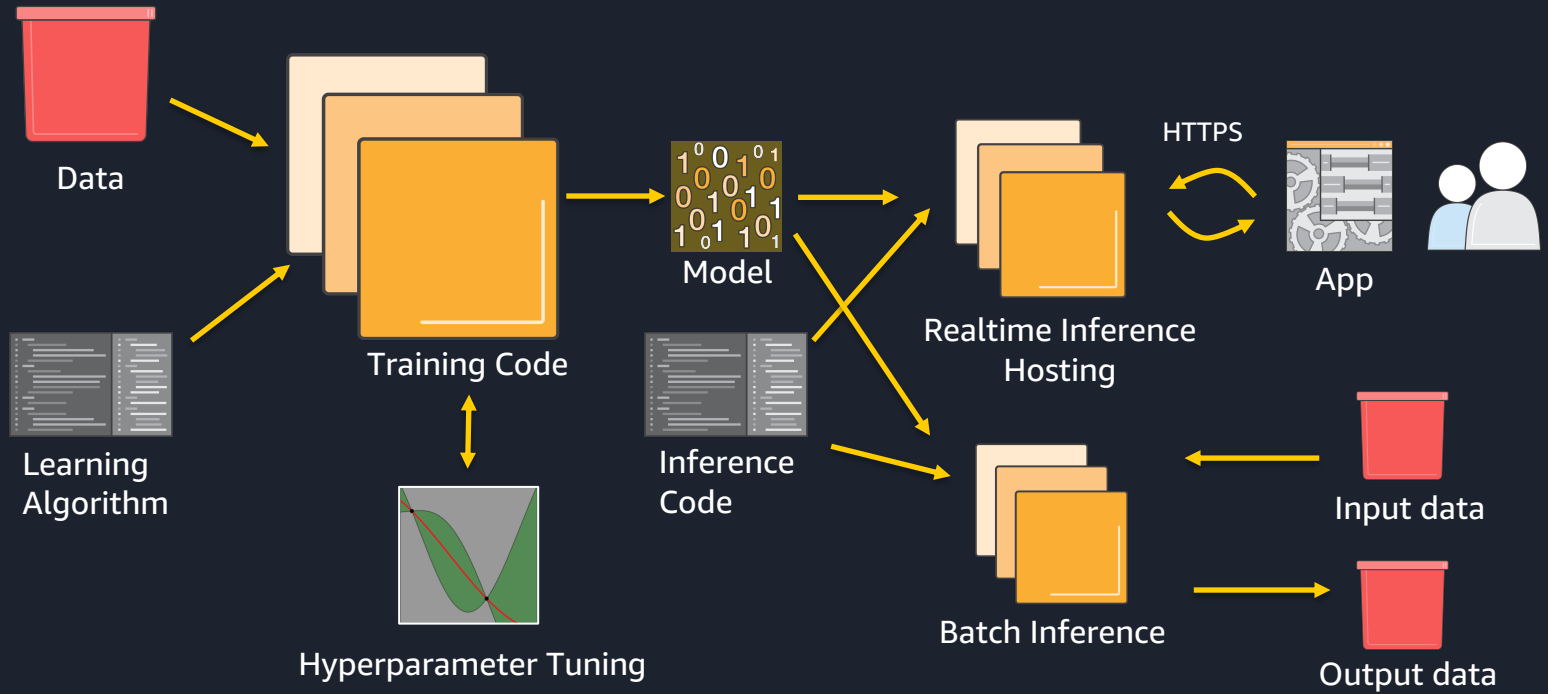
Console

Python SDK

Spark SDK

Step Functions

Amazon SageMaker Training and Inference Flow



Each algorithm solves a type of prediction problem

Classification
- Linear Learner
- XGBoost
- KNN

Working with Text
- BlazingText
  - Supervised
  - Unsupervised

Sequence Translation
- Seq2Seq

Computer Vision
- Image Classification
- Object Detection
- Semantic Segmentation

Recommendation
- Factorization Machines

Anomaly Detection
- Random Cut Forests
- IP Insights

Regression
- Linear Learner
- XGBoost
- KNN

Topic Modeling
- LDA
- NTM

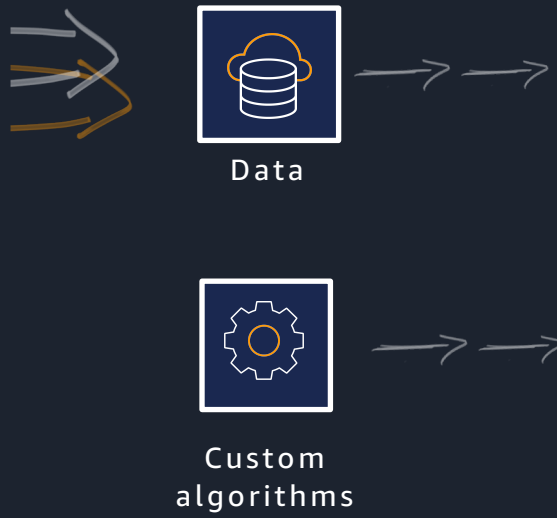
Forecasting
- DeepAR

Clustering
- KMeans

Feature Reduction
- PCA
- Object2Vec

Amazon SageMaker – Bring-Your-Own-Algorithm

Build your own DL algorithms, SageMaker handles the rest



Amazon SageMaker Training Service Supported Containers

GLUON

Keras

PYTORCH

scikit learn

mxnet

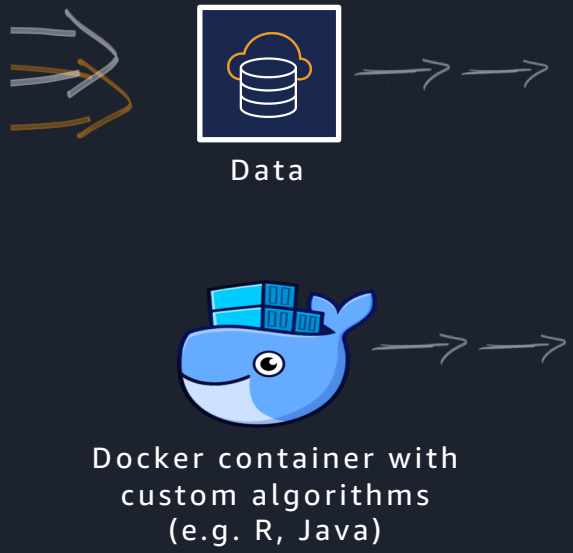
TensorFlow

Chainer

aws machine learning

Amazon SageMaker – Bring-Your-Own-Container

Bring your custom code and container, train at scale in SageMaker



aws machine learning

Walking through an example notebook

Helping you get started on ML on AWS

Machine Learning University

Uses the same materials used to train Amazon developers

Foundational knowledge with real-world application

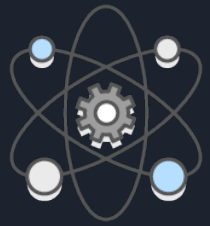
Structured courses and specialist certification

 |

AWS Machine Learning Specialty Certification!

aws machine learning

Amazon ML Solutions Lab



Amazon ML Solutions Lab provides ML expertise

Leverage Amazon experts with decades of ML experience with technologies like Amazon Echo, Amazon Alexa, Prime Air and Amazon Go

Brainstorming

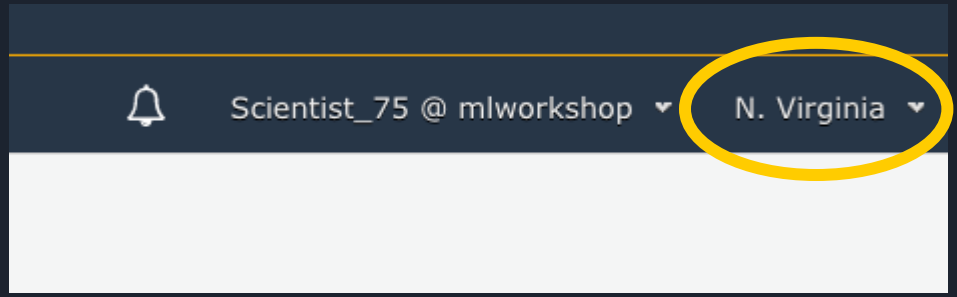
Modeling

Teaching

Setting up Notebook Instances for the workshop

https://amzn.to/2JKkT09
(or https://mlworkshop.signin.aws.amazon.com/console)

Username: Scientist_{#}
Password: TBD



Set region to
N.Virginia
(aka "us-east-1")

 |

aws machine learning

Sample projects (https://bit.ly/2lVxqV9)

- Classical ML:
  - Banking fraud, predictive maintenance
- NLP (natural language processing):
  - Topic modeling, books on tape
- CV (computer vision):
  - Object detection for ships
- Other:
  - Forecasting and recommendation
- Your own:
  - Bring your own dataset in your own AWS account

Agenda

Day 1	Day 2	Day 3
AI/ML on AWS Intro lab	Feature engineering Model evaluation	Moving to production Build
Team up Define problem	Build	Present
Write-up	Working model	Solution architecture

Problem statement

- Step 1: What is the problem?
- Step 2: Why does the problem need to be solved?
- Step 3: How would I solve the problem?

Problem statement

- Articulate your problem
- Start simple
- Identify your data sources
- Design your data for the model
- Determine where data comes from
- Determine easily obtained inputs
- Think about potential bias

