aws machine learning

Getting started with Machine Learning on AWS

Julien Simon
Global Evangelist, AI & Machine Learning, AWS
@julsimon

AI vs. Machine Learning vs. Deep Learning

Artificial Intelligence: design software applications which exhibit human-like behavior, e.g. speech, natural language processing, reasoning or intuition

Machine Learning: using statistical algorithms, teach machines to learn from featurized data without being explicitly programmed

Deep Learning: using neural networks algorithms, teach machines to learn from complex data where features cannot be easily expressed

Types of Machine Learning

Supervised learning

- Run an algorithm on a labeled data set.
- The model learns how to correctly predict the right answer.
- Regression and classification are examples of supervised learning.

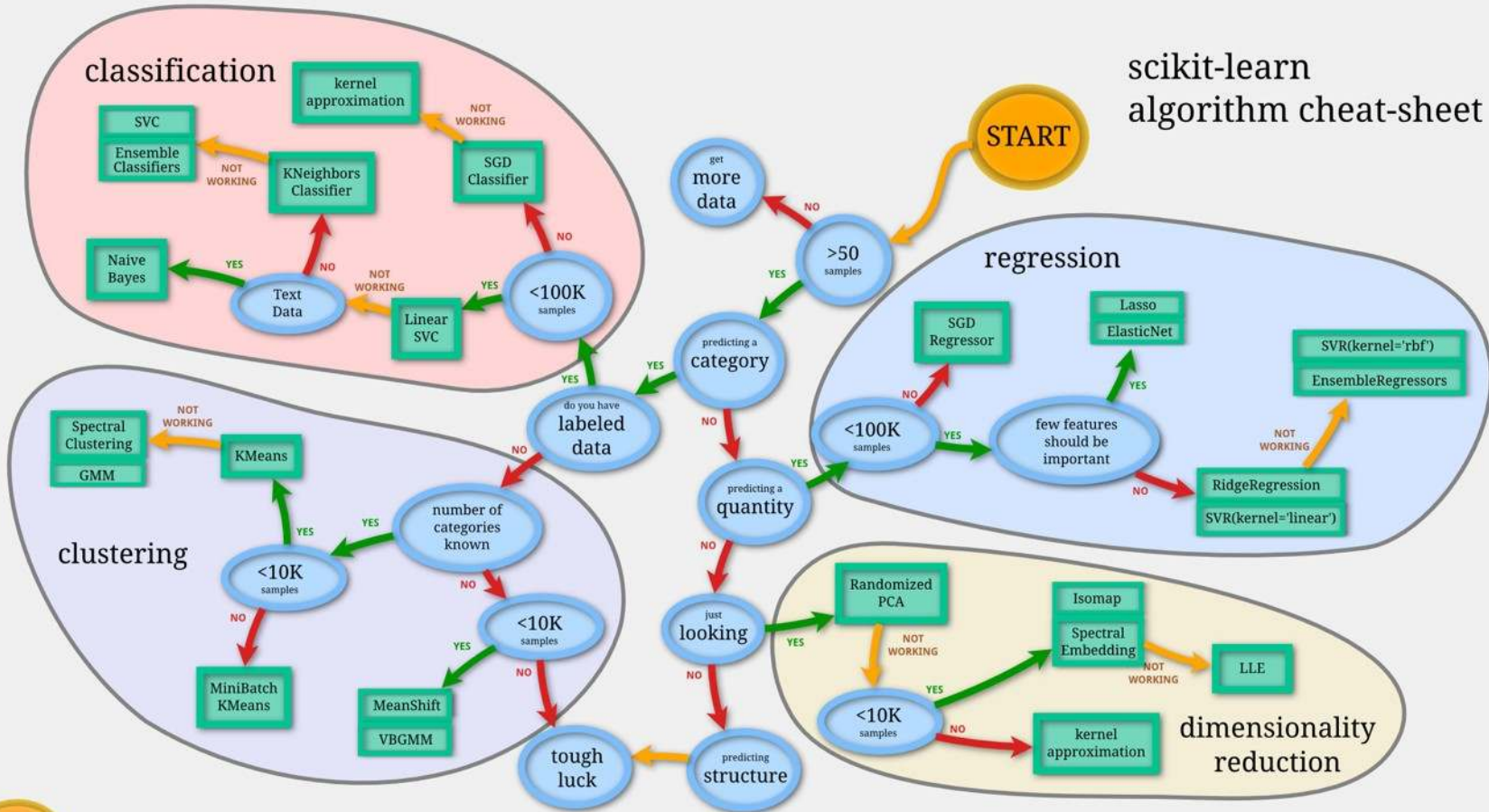## Unsupervised learning

- Run an algorithm on an unlabeled data set.
- The model learns patterns and organizes samples accordingly.
- Clustering and topic modeling are examples of unsupervised learning.
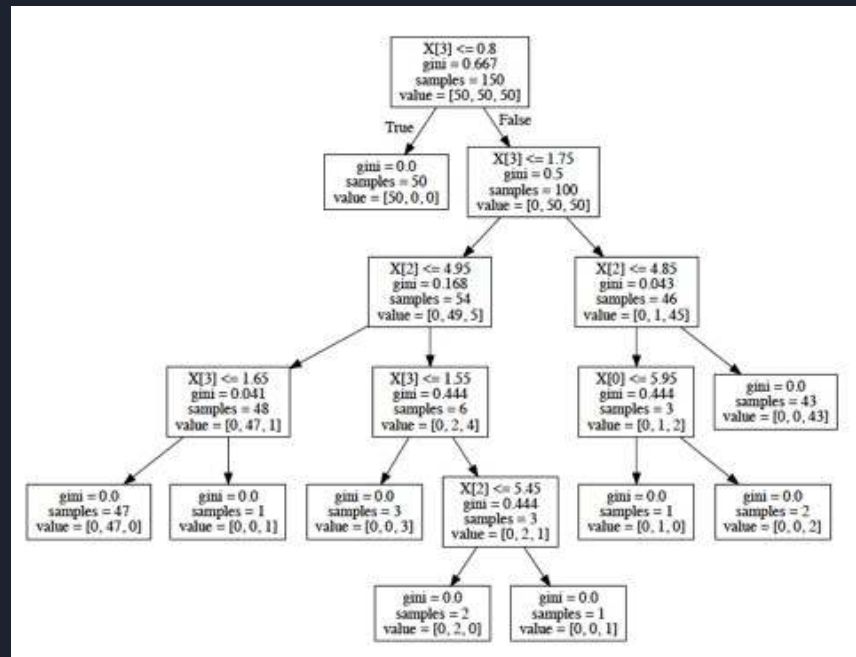
Types of Machine Learning
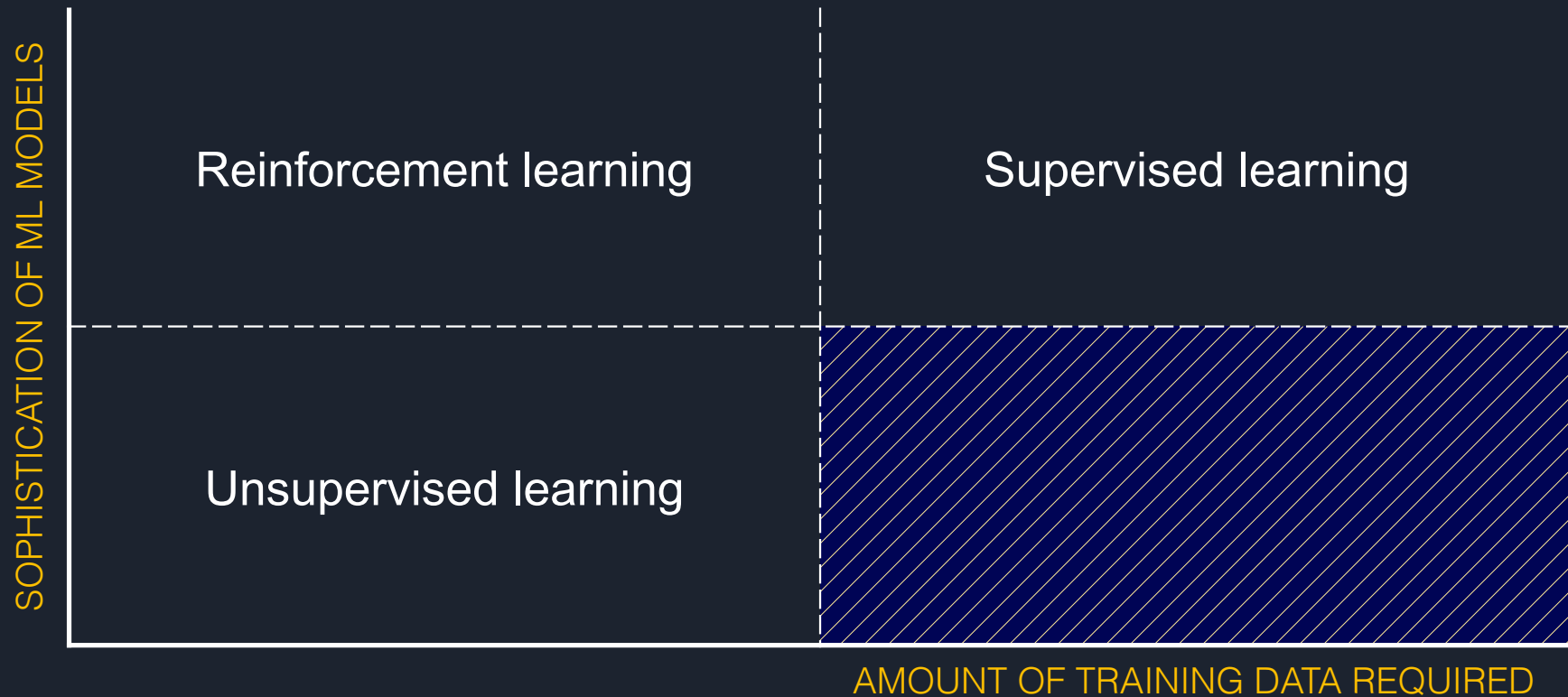
scikit-learn
algorithm cheat-sheet

Algorithm example: decision trees

- Supervised learning algorithm
- Goal: build a decision tree for regression or classification
- Data set : features + target attribute (value or class identifier)
- Intuition: find the "best" feature thresholds to go left or right
- "Easy" to interpret
- Advanced variants with multiple trees:
  Random Forests, XGBoost, etc.

The Machine Learning cycle

1 - Set expectations

- What is the business question you're trying to answer?
  - One sentence on the whiteboard
  - Must be quantifiable
- Do you have (enough) data that could help?
- Involve everyone and come to a common understanding
  - Business, IT, Data Engineering, Data Science, Ops, etc.

« We want to see what this technology can do for us »

« We have tons of relational data, surely we can do something with it »

« I read this cool article about FooBar ML, we ought to try it »

2 - Define clear metrics

- What is the business metric showing success?
- What's the baseline (human and IT)?
- What would be a significant and reasonable improvement?
- What would be reasonable further improvements?

« The confusion matrix for our support ticket classifier has significantly improved ». Huh?

« P90 time-to-resolution is now under 24 hours ». Err….

« Misclassified emails have gone down 5.3% using the latest model ». So?

« The latest survey shows that 'very happy' customers are up 9.2% ». Woohoo!

aws machine learning

3 - Assess needs (not wants) and skills

- Building a data set describing the problem?
- Cleaning, preparing and curating it?
- Writing and tweaking ML algorithms?
- Managing ML infrastructure?

Fully managed ←—————— ? —————→ 100% DIY

4 - Pick the best tool for the job

- Cost, time to market, accuracy: pick two
- The least expensive and fastest option won't probably be the most accurate.
  - Maybe enough to get started,
    and learn more about the problem.
- Improving accuracy will take increasingly more time and money.
  - Diminishing returns! Know when to stop.
- Keep an eye on actionable state of the art advances, ignore the rest
  - Transfer learning
  - AutoML

Cost

Time   Accuracy

aws machine learning
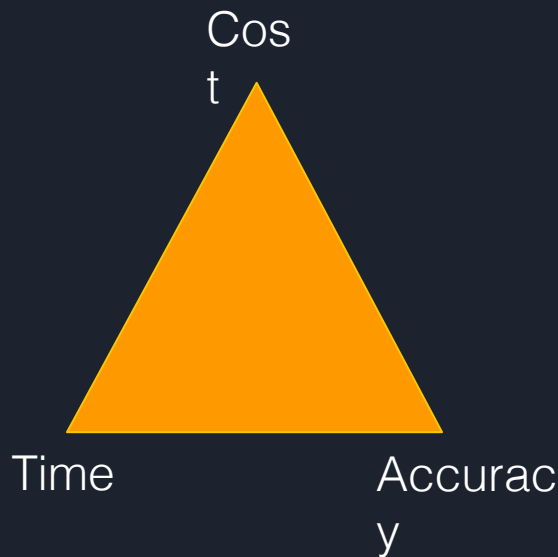
5 - Use proven best practices

- No, things are not different this time.

- AI / ML is software engineering
  - Dev, test, QA, documentation, Agile, versioning, etc.
  - Involve all teams

- Sandbox tests are nice, but truth is in production
  - Get there fast, as often as needed
  - CI / CD and automation are required
  - Devops for ML

Universal Pictures

6 - Iterate, iterate, iterate
aka Boyd's Law (1960)

- Start small
- Try the simple things first
- Go to production quickly
- Observe prediction errors
- Act: fix data set? Add more data? Tweak the algo? Try another algo?
- Repeat until accuracy gains become irrelevant
- Move to the next project

Machine Learning innovation at Amazon
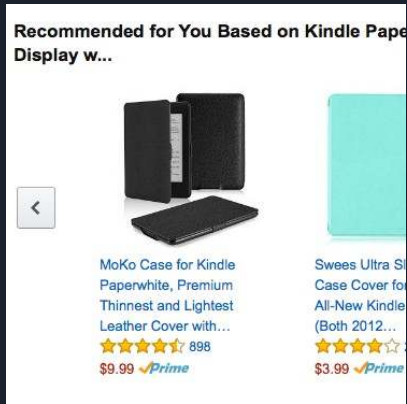
Our mission at AWS

Put machine learning in the
hands of every developer

MACHINE LEARNING IS HAPPENING
IN COMPANIES OF EVERY SIZE AND INDUSTRY

Tens of thousands customers have chosen AWS for their ML workloads | More than twice as many customers using ML than any other cloud provider

aws machine learning

https://aws.amazon.com/machine-learning/customers/

The AWS ML Stack

Broadest and most complete set of Machine Learning capabilities

AI SERVICES

VISION	SPEECH		TEXT			SEARCH	CHATBOTS	PERSONALIZATION	FORECASTING	FRAUD	DEVELOPMENT	CONTACT CENTERS
Amazon Rekognition	Amazon Polly	Amazon Transcribe	Amazon Comprehend	Amazon Translate	Amazon Textract	Amazon Kendra	Amazon Lex	Amazon Personalize	Amazon Forecast	Amazon Fraud Detector	Amazon CodeGuru	Contact Lens
		+Medical	+Medical									For Amazon Connect

ML SERVICES

Amazon SageMaker

Ground Truth	ML Marketplace	SageMaker Studio IDE								Neo	A2I
		Built-in algorithms	Notebooks	Experiments	Processing & Model Evaluation	Model training & tuning	Debugger	Autopilot	Model hosting	Model Monitor	

ML FRAMEWORKS & INFRASTRUCTURE

TensorFlow mxnet PYTORCH

GLUON K Keras learn MXBOARD DeepGraphLibrary

Deep LearningAMIs & Containers	GPUs &CPUs	ElasticInference	Inferentia	FPGA

aws machine learning

AI Services

The AWS ML Stack

Broadest and most complete set of Machine Learning capabilities

AI SERVICES

VISION	SPEECH		TEXT			SEARCH	CHATBOTS	PERSONALIZATION	FORECASTING	FRAUD	DEVELOPMENT	CONTACT CENTERS
Amazon Rekognition	Amazon Polly	Amazon Transcribe	Amazon Comprehend	Amazon Translate	Amazon Textract	Amazon Kendra	Amazon Lex	Amazon Personalize	Amazon Forecast	Amazon Fraud Detector	Amazon CodeGuru	Contact Lens
		+Medical	+Medical									For Amazon Connect

ML SERVICES

Amazon SageMaker

Ground Truth — ML Marketplace

SageMaker Studio IDE

| Built-in algorithms | Notebooks | Experiments | Processing & Model Evaluation | Model training & tuning | Debugger | Autopilot | Model hosting | Model Monitor |

Neo — A2I

ML FRAMEWORKS & INFRASTRUCTURE

TensorFlow — mxnet — PYTORCH

GLUON — K Keras — DeepGraphLibrary

| Deep LearningAMIs & Containers | GPUs &CPUs | ElasticInference | Inferentia | FPGA |

aws machine learning

The AWS ML Stack

Broadest and most complete set of Machine Learning capabilities

AI SERVICES

	VISION	SPEECH		TEXT			SEARCH	CHATBOTS	PERSONALIZATION	FORECASTING	FRAUD	DEVELOPMENT	CONTACT CENTERS
	Amazon Rekognition	Amazon Polly	Amazon Transcribe	Amazon Comprehend	Amazon Translate	Amazon Textract	Amazon Kendra	Amazon Lex	Amazon Personalize	Amazon Forecast	Amazon Fraud Detector	Amazon CodeGuru	Contact Lens
			+Medical	+Medical									For Amazon Connect

ML SERVICES

Amazon SageMaker

Ground Truth	ML Marketplace	SageMaker Studio IDE								Neo	A2I
		Built-in algorithms	Notebooks	Experiments	Processing & Model Evaluation	Model training & tuning	Debugger	Autopilot	Model hosting	Model Monitor	

ML FRAMEWORKS & INFRASTRUCTURE

TensorFlow   mxnet
GLUON   K Keras

PYTORCH   DeepGraphLibrary

Deep LearningAMIs & Containers | GPUs &CPUs | ElasticInference | Inferentia | FPGA

aws machine learning

The machine learning workflow is iterative and complex

Prepare

Build

Train & Tune

Deploy & Manage

Collect and prepare training data

Choose or build an ML algorithm

Set up and manage environments for training

Train, debug, and tune models

Manage training runs

Deploy model in production

Monitor models

Validate predictions

Scale and manage the production environment

aws machine learning

Amazon SageMaker helps you build, train, and deploy models

Prepare **Build** **Train & Tune** **Deploy & Manage**

Web-based IDE for machine learning

Automatically build and train models

Fully managed data processing jobs and data labeling workflows

101011010
010101010
000011110

Collect and prepare training data

One-click collaborative notebooks and built-in, high performance algorithms and models

Choose or build an ML algorithm

One-click training

Set up and manage environments for training

Debugging and optimization

Train, debug, and tune models

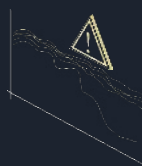Visually track and compare experiments

Manage training runs

One-click deployment and autoscaling

Deploy model in production

Automatically spot data drift

Monitor models

Add human review of predictions

Validate predictions

Fully managed with auto-scaling for 75% less

Scale and manage the production environment

Same service and APIs, from experimentation to production

Model options

AWS Marketplace
for Machine
Learning

Training code

Amazon SageMaker
AutoPilot

Factorization Machines
Linear Learner
Principal Component
Analysis
K-Means Clustering
XGBoost

mxnet Chainer
TensorFlow PyTorch
scikit learn dmlc XGBoost

Built-in Algorithms (17)
No ML coding required

Built-in Frameworks
Bring your own code
Use open source containers

Bring Your Own
Full control, run your container
R, C++, etc.

Fully managed training, spot instances included

Frameworks and Infrastructure

The AWS ML Stack

Broadest and most complete set of Machine Learning capabilities

AI SERVICES

VISION	SPEECH		TEXT			SEARCH	CHATBOTS	PERSONALIZATION	FORECASTING	FRAUD	DEVELOPMENT	CONTACT CENTERS
Amazon Rekognition	Amazon Polly	Amazon Transcribe	Amazon Comprehend	Amazon Translate	Amazon Textract	Amazon Kendra	Amazon Lex	Amazon Personalize	Amazon Forecast	Amazon Fraud Detector	Amazon CodeGuru	Contact Lens
		+Medical	+Medical									For Amazon Connect

ML SERVICES

Amazon SageMaker	Ground Truth	ML Marketplace	SageMaker Studio IDE								Neo	A2I
			Built-in algorithms	Notebooks	Experiments	Processing & Model Evaluation	Model training & tuning	Debugger	Autopilot	Model hosting	Model Monitor	

ML FRAMEWORKS & INFRASTRUCTURE

TensorFlow · mxnet · PYTORCH · GLUON · K Keras · learn · HOROVOD · DeepGraphLibrary

Deep LearningAMIs & Containers	GPUs &CPUs	ElasticInference	Inferentia	FPGA

aws machine learning

AWS: The platform of choice for TensorFlow

https://aws.amazon.com/tensorflow

89% of all deep learning workloads in the cloud run on AWS

85% of all TensorFlow workloads in the cloud run on AWS

Source: Nucleus Research, T147, October 2019

aws machine learning

Amazon Elastic Inference

Lower
inference
costs
up to 75%

Match
capacity
to demand

Available between
1 to 32 TFLOPS
per accelerator

Integrated with
Amazon EC2 and
Amazon SageMaker

Support for TensorFlow,
Apache MXNet (Incubating)
—PyTorch coming soon

Single and mixed-
precision operations

aws machine learning

Amazon EC2 Inferentia

- Fast, low-latency inferencing at a very low cost
  - 64 TeraOPS on 16-bit floating point (FP16 and BF16) and mixed-precision data.
  - 128 TeraOPS on 8-bit integer (INT8) data.
- Neuron SDK: https://github.com/aws/aws-neuron-sdk
  - Available in Deep Learning AMIs and Deep Learning Containers
  - TensorFlow and Apache MXNet, PyTorch coming soon

https://ml.aws

https://aws.amazon.com/sagemaker
https://github.com/aws/sagemaker-python-sdk
https://github.com/awslabs/amazon-sagemaker-examples

https://youtube.com/juliensimonfr
https://medium.com/@julsimon

Published August 2020

Discount link for the paper edition on Amazon (US only)
https://www.amazon.com/gp/mpc/AOHJSZC7A0AV5

Discount code for the e-book edition on Packt
20SAGEMAKER
https://www.packtpub.com/product/learn-amazon-sagemaker/9781800208919

Valid until November 11th

Learn
Amazon SageMaker

A guide to building, training, and deploying machine learning models
for developers and data scientists

Packt>
www.packt.com

Julien Simon
Foreword by Francesco Pochetti,
Senior Data Scientist at Mash & AWS ML Hero

aws machine
learning