

Como a AWS pode te ajudar?

AI/ML

Gabriel Martini
Arquiteto de Soluções
martinig@amazon.com



Agenda

- Introdução a **AI/ML**
- AI APIs **AWS**
- Lab **APIs**
- Modelos de **ML** na **AWS**
- Lab **SageMaker**



Amazon Web Services (AWS)

Infraestrutura Global





The broadest and deepest cloud platform for today's builders

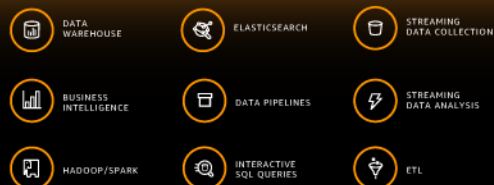
Technical & Business Support



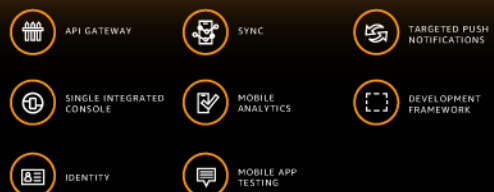
Marketplace



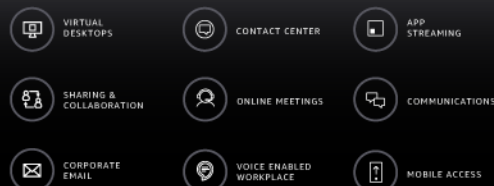
Analytics



Mobile Services



Enterprise Apps



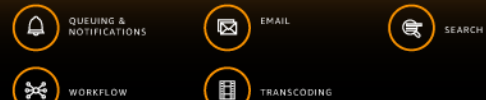
Core Services



DevOps



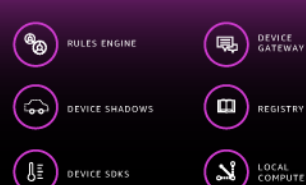
App Services



Infrastructure



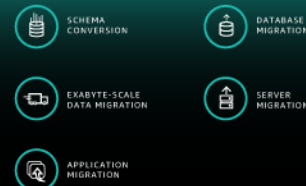
IoT



Blockchain



Migration



Management Tools



Machine Learning



Security & Compliance

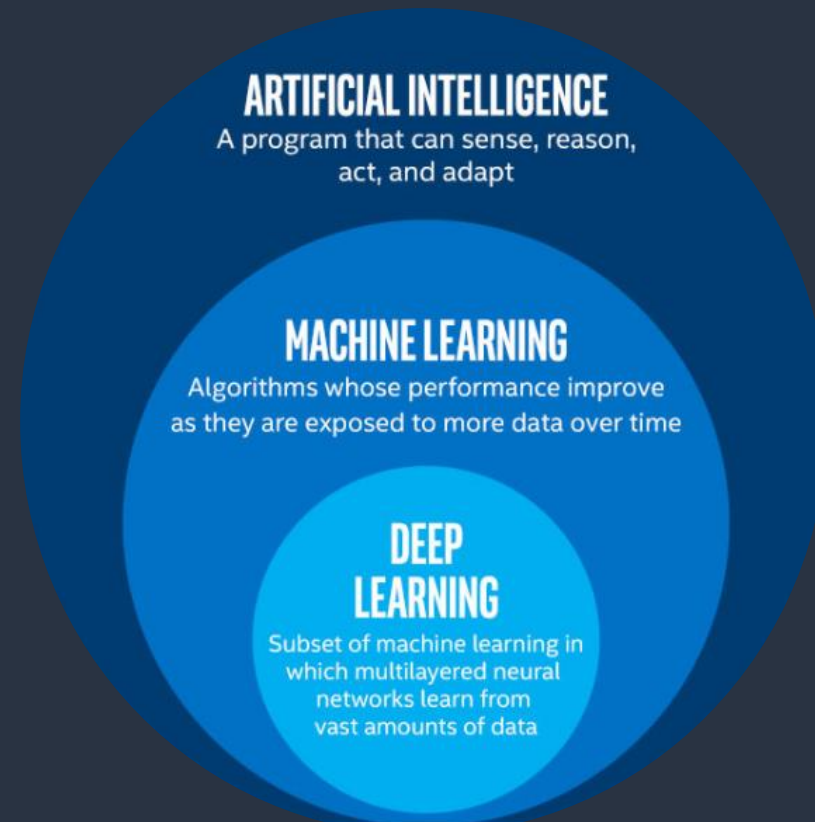


The background of the entire image is a soft-focus bokeh of numerous warm-toned lights, likely from a string of outdoor lights. In the foreground, on the right side, a single light bulb is in sharp focus, hanging from a black cord. The bulb is illuminated, showing a warm yellow glow. A dark, semi-transparent horizontal band runs across the middle of the image, serving as a background for the text.

AI/ML

AI/ML

“A IA é um campo da ciência da computação dedicado a resolver problemas cognitivos geralmente associados à inteligência humana, como aprendizado, solução de problemas e reconhecimento de padrões. ”



AI/ML na AWS

Serviços para AI/ML

Conjunto de recursos mais amplo e profundo

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	INFRASTRUCTURE								
 TensorFlow  mxnet  PYTORCH	 GLUON  K Keras	 EC2 P3 & P3DN	 EC2 G4 EC2 C5	 FPGA	 DL CONTAINERS & AMIs	 ELASTIC CONTAINER SERVICE	 ELASTIC KUBERNETES SERVICE	 GREENGRASS	 ELASTIC INFERENCE	 INFERENCE



AWS Compute

AI/ML na AWS

Serviços para AI/ML

Conjunto de recursos mais amplo e profundo

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	INFRASTRUCTURE								
 TensorFlow  mxnet  PYTORCH	 GLUON  Keras									
		EC2 P3 & P3DN	EC2 G4 EC2 C5	FPGAS	DL CONTAINERS & AMIs	ELASTIC CONTAINER SERVICE	ELASTIC KUBERNETES SERVICE	GREENGRASS	ELASTIC INFERENCE	INFERENCE

Amazon EC2

General
Purpose

M5



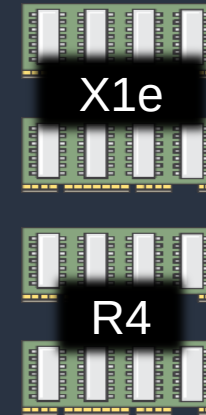
Compute
Optimized



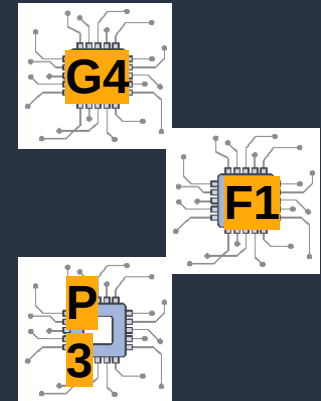
Storage and IO
Optimized



Memory
Optimized



Accelerated
Computing



Computação acelerada – P3

Até 8 GPUs **NVIDIA Tesla V100** GPUs, cada uma emparelhando **5.120 núcleos CUDA** e **640 núcleos Tensor**

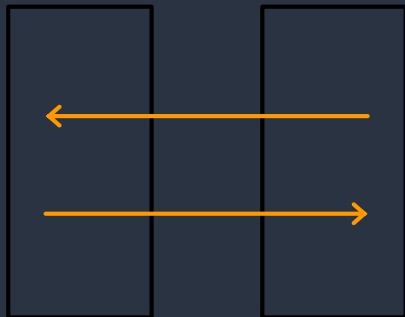
Suporte a NVLink para comunicação peer-to-peer entre GPUs

Oferece até 100 Gbps de largura de banda de rede agregada

Suporte para **NVIDIA Collective Communications Library** (NCCL) com **Elastic Fabric Adapter** (EFA)

Instance	GPUs	vCPU	Mem (GiB)	GPU Mem (GiB)	GPU P2P	Storage (GB)	Dedicated EBS Bandwidth	Networking Performance
p3.2xlarge	1	8	61	16	-	EBS-Only	1.5 Gbps	Up to 10 Gigabit
p3.8xlarge	4	32	244	64	NVLink	EBS-Only	7 Gbps	10 Gigabit
p3.16xlarge	8	64	488	128	NVLink	EBS-Only	14 Gbps	25 Gigabit
p3dn.24xlarge	8	96	768	256	NVLink	2 x 900 NVMe SSD	14 Gbps	100 Gigabit

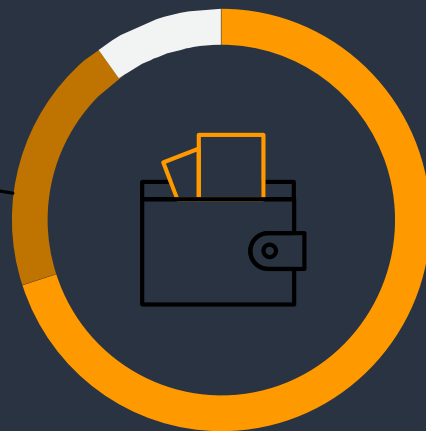
Redução de custos



Infraestrutura Spot

É o mesmo que sob demanda e RIs: mesmo hardware, mesmos recursos, mesma escolha de instâncias, mesma área geográfica (regiões e AZs)

70–90% off



Preço

Economize pagando menos do que o valor sob demanda



Flexível

Utilize diversas instâncias para chegar em suas necessidades

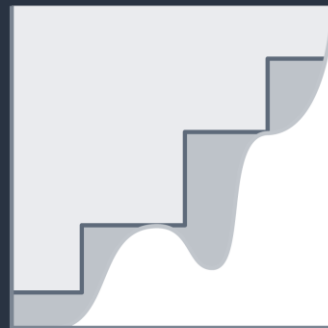
Redução de custos

Elastic Inference

Accelerator Type	F32 Throughput in TFLOPS	F16 Throughput in TFLOPS	Memory in GB
m1.eia2.medium	1	8	2
ml.eia2.large	2	16	4
ml.eia2.xlarge	4	32	8
ml.eia1.medium	1	8	1
ml.eia1.large	2	16	2
ml.eia1.xlarge	4	32	4



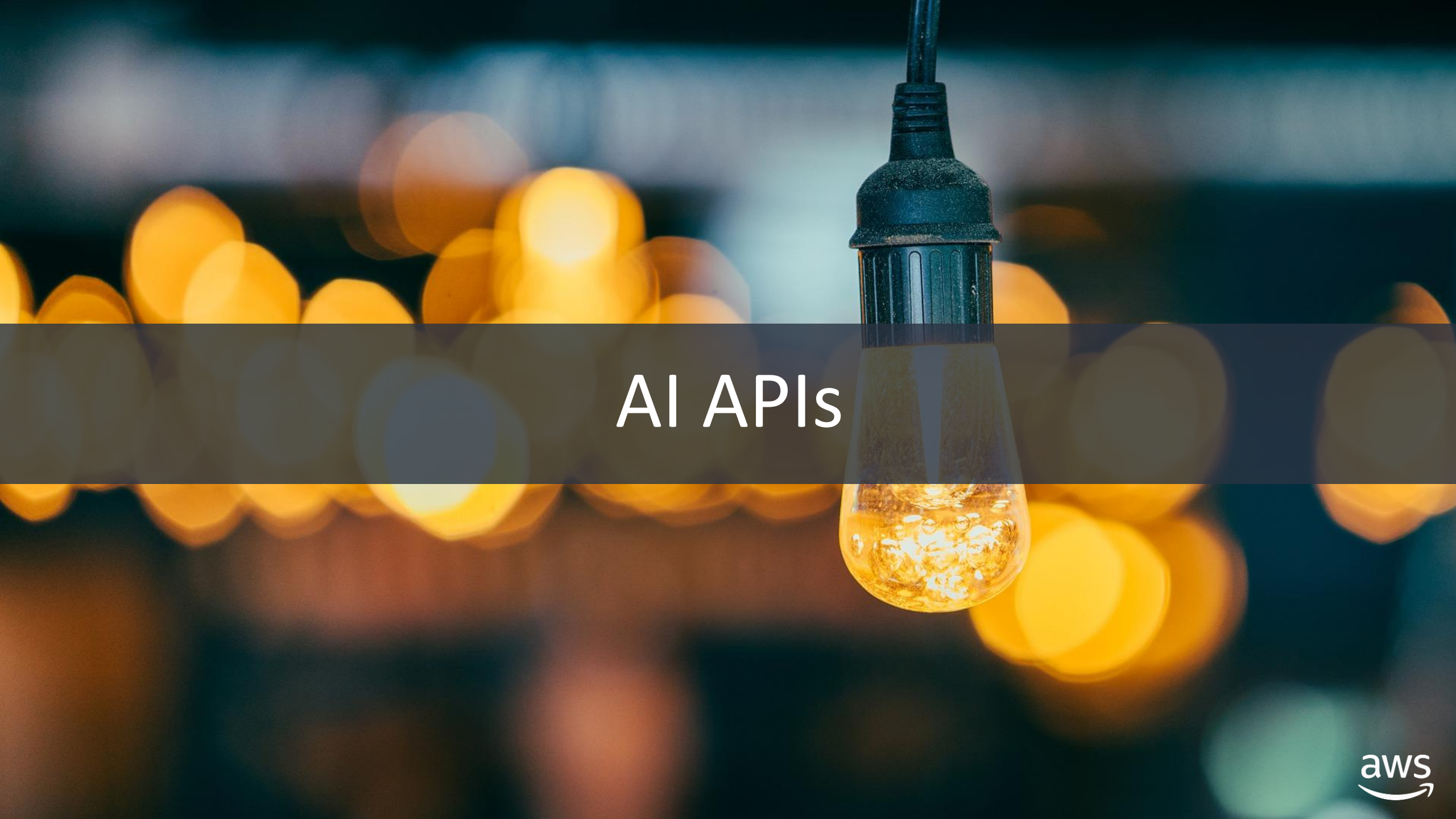
Reduza os valores
de inferência



Capacidade sob demanda



Disponibilidade entre
1 até 32 TFLOPS por
acelerador



AI APIs

AI/ML na AWS

Serviços para AI/ML

Conjunto de recursos mais amplo e profundo

AI Services

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

ML Services

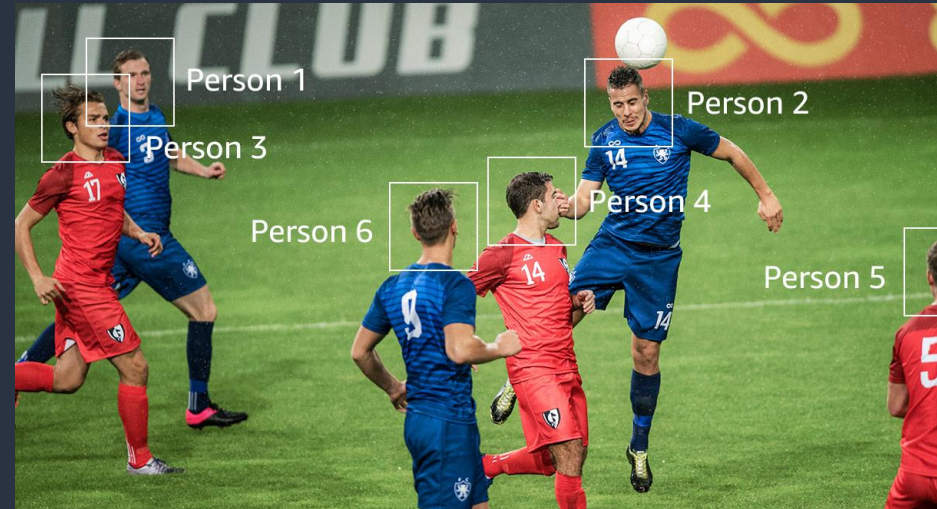
 Amazon SageMaker	Ground Truth	Notebooks	Algorithms + Marketplace	Reinforcement Learning	Training	Optimization	Deployment	Hosting
--	--------------	-----------	--------------------------	------------------------	----------	--------------	------------	---------

ML Frameworks + Infrastructure

FRAMEWORKS	INTERFACES	INFRASTRUCTURE								
 TensorFlow  mxnet  PYTORCH	 GLUON  Keras	 EC2 P3 & P3DN	 EC2 G4 & C5	 FPGAs	 DL CONTAINERS & AMIs	 ELASTIC CONTAINER SERVICE	 ELASTIC KUBERNETES SERVICE	 GREENGRASS	 ELASTIC INFERENCE	 INFERENCE

Visão Computacional

Amazon Rekognition



O serviço pode identificar os objetos, pessoas, texto, cenas e atividades, bem como detectar qualquer conteúdo inadequado. O Amazon Rekognition também fornece análise facial altamente precisa e reconhecimento facial nas imagens e vídeos que você fornece. Você pode detectar, analisar e comparar rostos para uma ampla variedade de casos de uso de verificação de usuários, contagem de pessoas e segurança pública.

Amazon Rekognition

- Detecção de objetos

rekognition.detect_labels

```
{ "Image":  
  { "S3Object":  
    { "Bucket": "console-sample-  
images",  
      "Name": "a.jpg"  
    }  
  }  
}
```



```
{ "Labels": [  
  { "Name": "Automobile",  
    "Confidence":  
98.87621307373047,  
    "Instances": [],  
    "Parents": [ { "Name": "Vehicle"  
}, { "Name": "Transportation" } ] },  
  { "Name": "Vehicle",  
    "Confidence": 98.87621307373047,  
    "Instances": [],  
    "Parents": [ { "Name":  
"Transportation" } ] }, .....  
]
```

Reconhecimento óptico de caracteres

Amazon
Textract



Sample document

Employment Application

Applicant Information

Full Name: Jane Doe

Phone Number: 555-0100

Home Address: 123 Any Street, Any Town, USA

Mailing Address: same as home address

Previous Employment History

Start Date	End Date	Employer Name	Position Held	Reason for leaving
1/15/2009	6/30/2013	Any Company	Head Baker	Family relocated
8/15/2013	present	Example Corp.	Baker	N/A, current employer

Raw text | Forms | Tables

Search

Lines

Employment Application Applicant Information Full Name: Jane Doe

Phone Number: 555-0100 Home Address: 123 Any Street, Any Town, USA

Mailing Address: same as home address Previous Employment History

Start Date End Date Employer Name Position Held

Reason for leaving 1/15/2009 6/30/2013 Any Company Head Baker

Family relocated 8/15/2013 present Example Corp Baker

N/A, current employer

É um serviço que extrai automaticamente texto e dados de documentos digitalizados. O Amazon Textract vai além do simples reconhecimento óptico de caracteres (OCR) para também identificar o conteúdo dos campos nos formulários e informações armazenadas nas tabelas.

Amazon Textract

- Extração de texto

`textract.detect_document_text`

```
{ "Document":  
  { "S3Object":  
    { "Bucket": "console-sample-  
images",  
      "Name": "x.jpg"  
    }  
  }  
}
```



```
'Blocks': [ {  
  'BlockType': 'LINE',  
  'Confidence': ...,  
  'Text': 'string', ...  
}]
```

Processamento de linguagem natural

Amazon Comprehend



Input text
[Supported languages](#)

Amazon.com, Inc. is located in Seattle, WA and was founded July 5th, 1994 by Jeff Bezos, allowing customers to buy everything from books to blenders. Seattle is north of Portland and south of Vancouver, BC. Other notable Seattle - based companies are Starbucks and Boeing.

272 of 5000 characters used.

[Clear text](#) [Analyze](#)

Insights [Info](#)

[Entities](#) | [Key phrases](#) | [Language](#) | [Sentiment](#) | [Syntax](#)

Analyzed text

Amazon.com, Inc. is located in Seattle, WA and was founded July 5th, 1994 by Jeff Bezos, allowing customers to buy everything from books to blenders. Seattle is north of Portland and south of Vancouver, BC. Other notable Seattle - based companies are Starbucks and Boeing.

Utiliza o aprendizado de máquina para ajudá-lo a descobrir insights e relacionamentos em seus dados não estruturados. O serviço identifica o idioma do texto; extrai frases-chave, lugares, pessoas, marcas ou eventos; entende o quão positivo ou negativo é o texto; analisa o texto usando tokenização e partes do discurso; e organiza automaticamente uma coleção de arquivos de texto por tópico.

Amazon Comprehend

- Análise de texto

`comprehend.detect_syntax`

```
{ "Image":  
  { "S3Object":  
    { "Bucket": "console-sample-  
images",  
      "Name": "y.jpg"  
    }  
  }  
}
```



```
'SyntaxTokens': [ {  
  'TokenId': 123,  
  'Text': 'string',  
  'BeginOffset': 123,  
  'EndOffset': 123,  
  'PartOfSpeech': { 'Tag': 'ADJ',  
    'Score': ... } }  
]
```


Demonstração

AWS Console:
<https://dashboard.eventengine.run/>

Laboratórios:
<https://bit.ly/2KVvJ3M>



ML Services

AI/ML na AWS

Serviços para AI/ML

Conjunto de recursos mais amplo e profundo

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

ML Frameworks + Infrastructure

FRAMEWORKS	INTERFACES	INFRASTRUCTURE							
 TensorFlow	 mxnet	 GLUON							
PYTORCH	 Keras		EC2 P3 & P3DN	EC2 G4 EC2 C5	FPGAS	DL CONTAINERS & AMIs	ELASTIC CONTAINER SERVICE	ELASTIC KUBERNETES SERVICE	GREENGRASS
									 ELASTIC INFERENCE
									 INFERENCE

SageMaker pipeline



Rotulação



Desenvolvimento



Testes



Treinamento e
otimização



Entrega



Gerenciamento e
escala

Rotulação dos dados

SageMaker GroundTruth



Mechanical
Turk workers



Private labeling
workforce



Third-party
vendors



Automatic
annotations

Notebooks para desenvolvimento

SageMaker Notebook



Notebook instances					Actions ▾	Create notebook instance
Q Search notebook instances					< 1 >	⚙
	Name ▾	Instance	Creation time ▾	Status ▾	Actions	
ⓘ	ml-demos	ml.c5.xlarge	May 14, 2019 16:02 UTC	✔ InService	Open Jupyter Open JupyterLab	

jupyter Image-classification-fulltraining (unsaved changes) 

File Edit View Insert Cell Kernel Widgets Help Not Trusted conda_mxnnet_p36



End-to-End Multiclass Image Classification Example

- [1. Introduction](#)
- [2. Prerequisites and Preprocessing](#)
 - [A. Permissions and environment variables](#)
- [3. Training the ResNet model](#)
- [4. Deploy The Model](#)
 - [A. Create model](#)
 - [B. Batch transform](#)
 - [C. Realtime inference](#)
 - [a. Create endpoint configuration](#)
 - [b. Create endpoint](#)
 - [c. Perform inference](#)
 - [d. Clean up](#)

Introduction

Welcome to our end-to-end example of distributed image classification algorithm. In this demo, we will use the Amazon sagemaker image classification algorithm to train on the [caltech-256 dataset](#).

To get started, we need to set up the environment with a few prerequisite steps, for permissions, configurations, and so on.

Prerequisites and Preprocessing

Treinamento do modelo

1



Algoritmos
Built-in

2



Script
Mode

3



Docker
Container

4



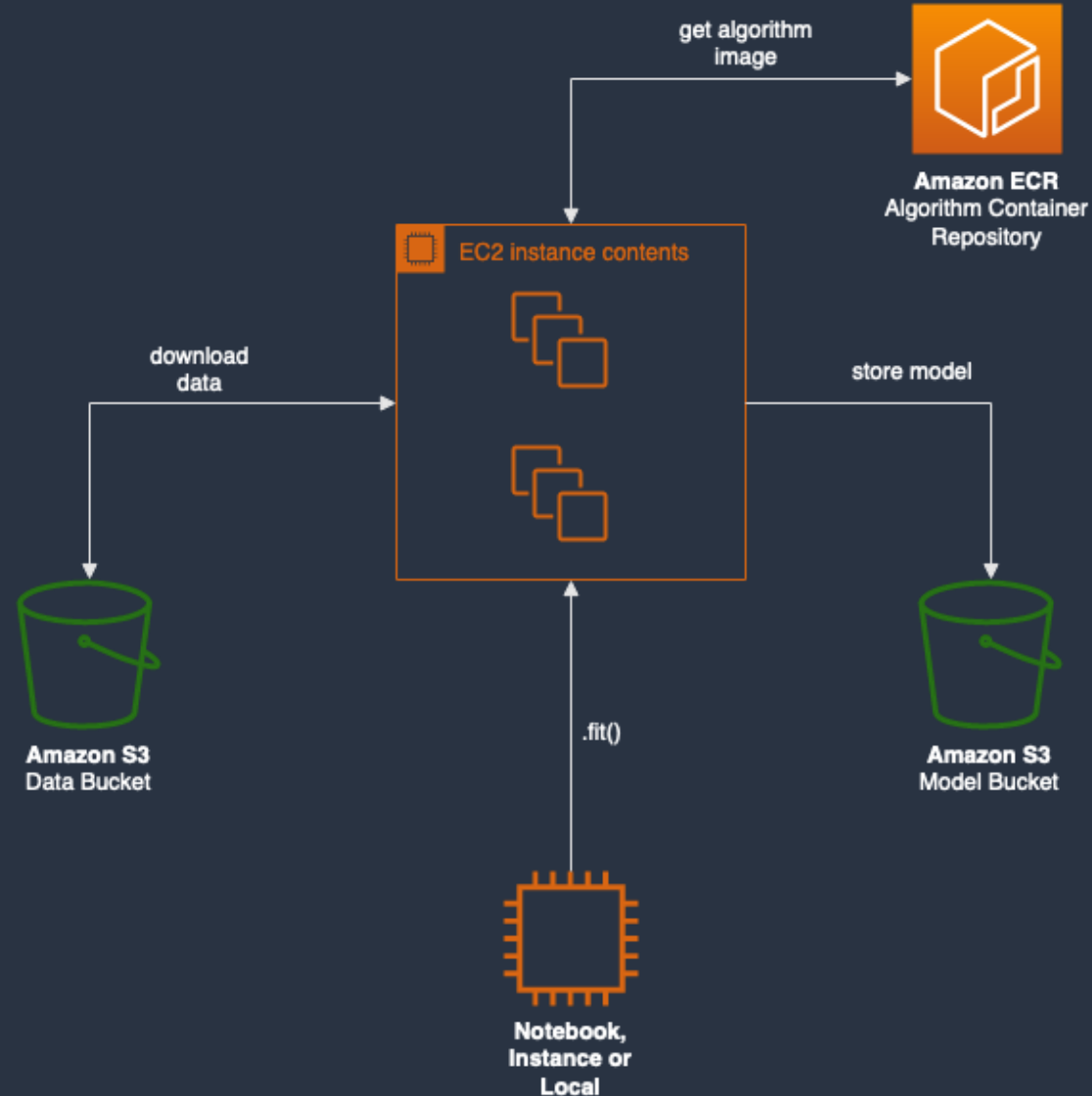
ML
Marketplace

5



Notebook
instance

Treinamento do modelo



Otimize a performance do modelo

SageMaker Neo



Get accuracy
and performance



Automatic optimization



Broad framework
support



Broad hardware support

Implantação

SageMaker Endpoint

```
endpoint_config_response = sage.create_endpoint_config(  
    EndpointConfigName = endpoint_config_name,  
    ProductionVariants=[  
        'InstanceType':'ml.m4.xlarge',  
        'InitialInstanceCount':1,  
        'ModelName':model_name,  
        'VariantName':'AllTraffic']])
```



Caso de cliente

Combate ao tráfico de pessoas

O Marinus Analytics usa o Amazon Rekognition para encontrar vítimas de tráfico de pessoas e depois ajudar a polícia com os traficantes.



Caso de cliente

Tornar o deep learning mais econômico

A Autodesk usa o TensorFlow no Amazon SageMaker para criar modelos de aprendizado profundo para vários casos de uso. Esses modelos permitem à Autodesk e seus clientes explorar milhares de alternativas de projeto em potencial, pesquisar semanticamente projetos, otimizar os processos de construção de engenharia e otimizar os fluxos de trabalho de renderização.

Com o Amazon Elastic Inference, use a computação de GPU de maneira econômica para seus modelos de aprendizado profundo. Eles esperam economizar 75% em custos em comparação com a execução em GPUs tradicionais.



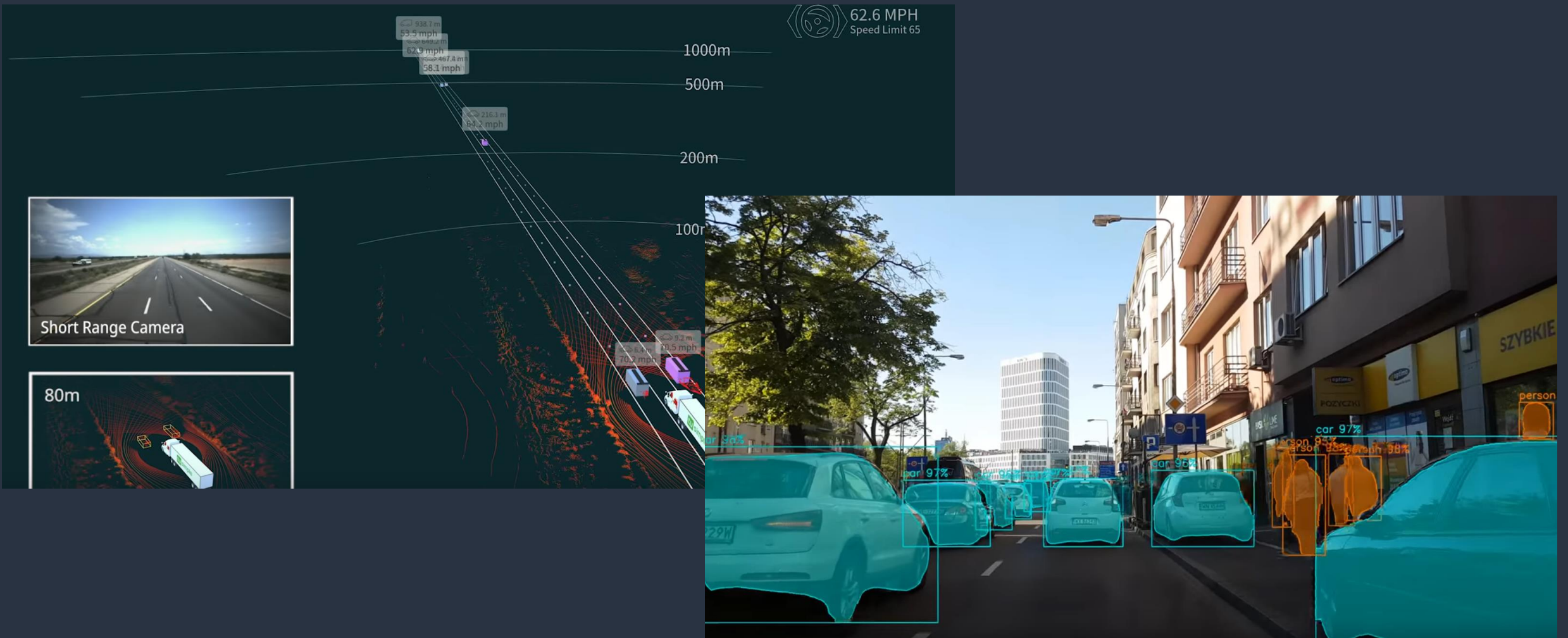
Caso de cliente

Prevenção de fraudes ataques em tempo real

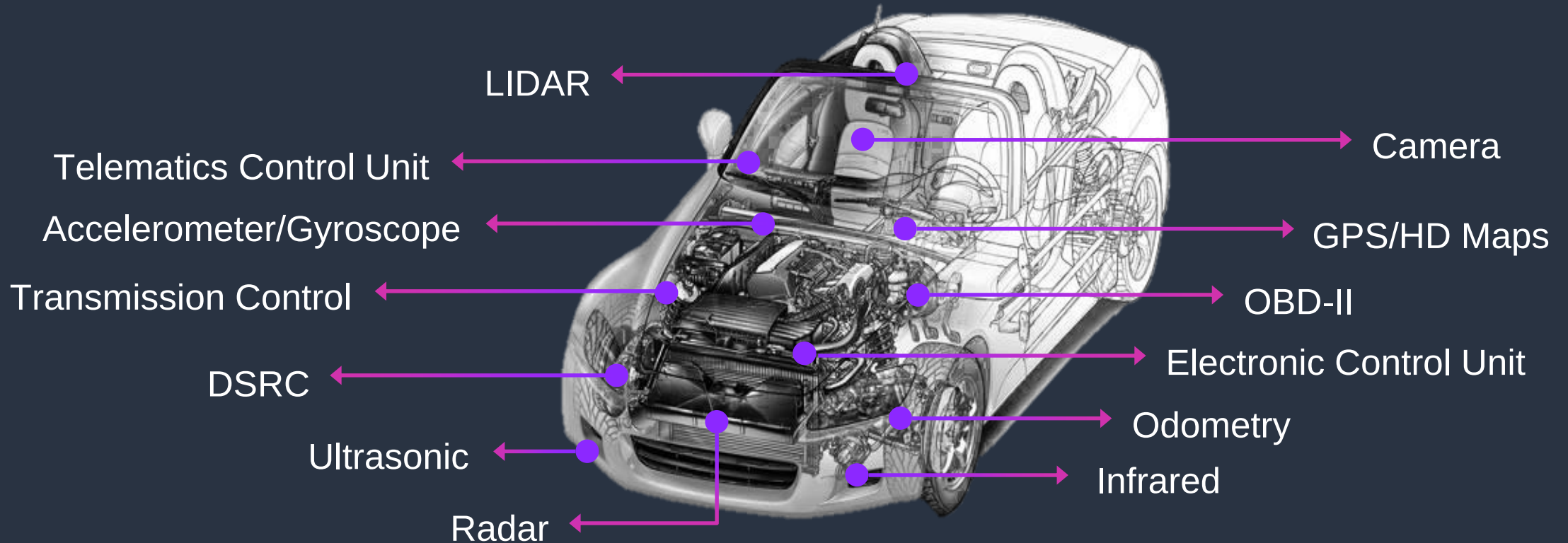
Usando o Amazon SageMaker, o NuData Security evita fraudes no cartão de crédito, analisando dados anônimos do usuário para detectar atividades anômalas antes que ocorra uma transação fraudulenta. Com o SageMaker, o NuData reduziu o tempo de desenvolvimento de aprendizado de máquina em 60%, simplificou sua arquitetura de aprendizado de máquina em 95% e trabalhou com um grande banco para bloquear passivamente quase 100% do tráfego de tentativas fraudulentas dentro da tolerância de atrito do consumidor do banco.



Carros Autônomos



Carros Autônomos



Carros Autônomos



Secure data
consumption



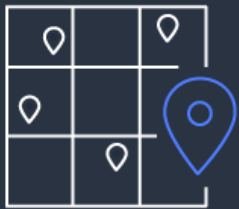
Driver
behavior



Anomaly
detection



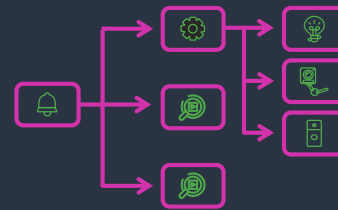
Diagnostics
alerts



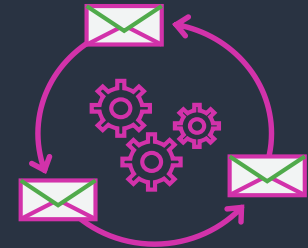
Location-
based services



Software
Updates OTA



Fleet configuration
management



Provisioning & lifecycle
management

DeepRacer

AWS DeepRacer

Um carro de corrida totalmente autônomo de escala 1/18 projetado para ajudá-lo a aprender sobre o aprendizado por reforço através da direção autônoma



- Crie modelos de aprendizado de máquina no Amazon SageMaker
Treine, teste e interaja na pista usando o simulador de corrida AWS DeepRacer 3D
- Competir na primeira liga mundial de corrida autônoma do mundo, disputar prêmios e uma chance de ganhar a cobiçada AWS DeepRacer Cup

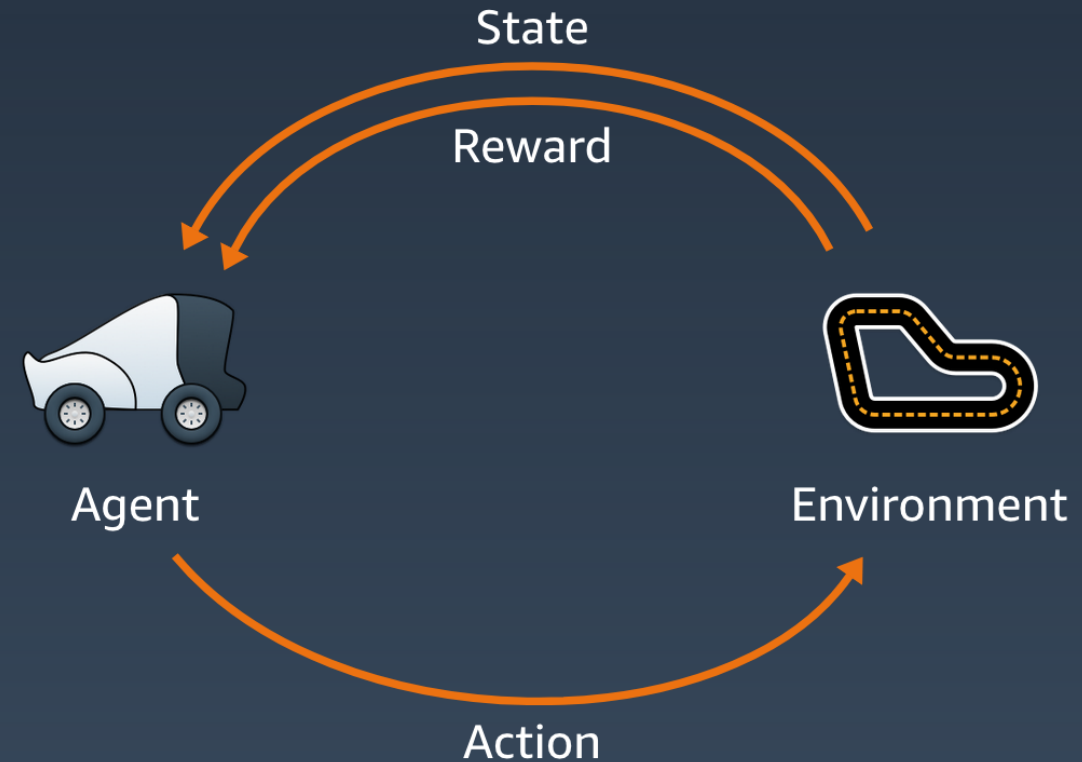
DeepRacer



DeepRacer

O aprendizado por reforço (RL) é um tipo de aprendizado de máquina, no qual um agente explora um ambiente para aprender a executar tarefas desejadas, executando ações com bons resultados e evitando ações com maus resultados.

Um modelo de aprendizado por reforço aprenderá com sua experiência e, com o tempo, será capaz de identificar quais ações levam às melhores recompensas.



DonkeyCar

Crie você mesmo:

<https://aws.amazon.com/blogs/machine-learning/build-an-autonomous-vehicle-on-aws-and-race-it-at-the-reinvent-robocar-rally/>



<https://www.donkeycar.com/>

Demonstração

AWS Console:

<https://dashboard.eventengine.run/>

Laboratórios:

<https://bit.ly/2KVvJ3M>

<https://amzn.to/2La7mj5>

Obrigado!

Qualquer dúvida estou á disposição!

Gabriel Martini

Arquiteto de Soluções

martinig@amazon.com

NOS AJUDE A MELHORAR,
FAÇA UMA RÁPIDA
AVALIAÇÃO SOBRE O
EVENTO

