AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

aws INNOVATE
ONLINE CONFERENCE

分会场四：人工智能与机器学习

AWS 人工智能介绍与新进展

王世帅，AWS 解决方案架构师

目录

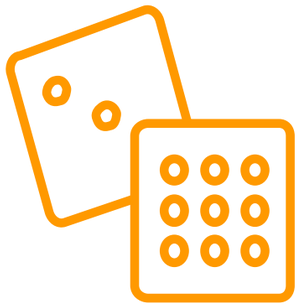
人工智能是数字化转型核心之一

客户体验

商业运营

决策制定

创新应用



竞争优势

40% IDC 预计，到2019年，40%的数字转型计划将由人工智能支持

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

亚马逊在人工智能领域的大量深度创新

1995

亚马逊自从成立以来一直在人工智能和机器学习领域进行大量投入，并且把我们的知识与能力与客户分享

2017

商品智能推荐

机器人与物流仓储

新产品

供应链管理

智慧呼叫中心

无人值守商店

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

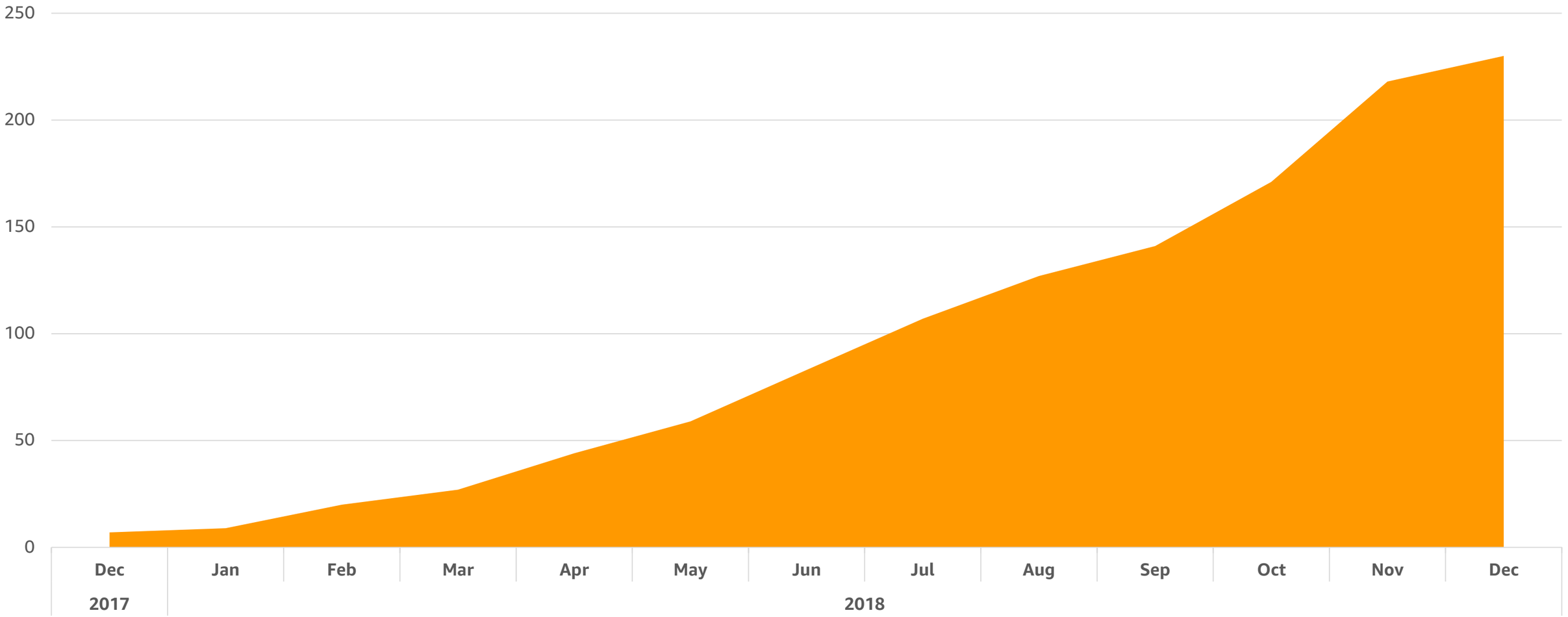
AWS 的使命

让每位开发人员掌握机器学习

aws INNOVATE
ONLINE CONFERENCE

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

一年内发布超过 200项机器学习新功能



AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

越来越多客户在 AWS 构建机器学习应用

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

AWS 上的机器学习

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

机器学习面对的三个重要挑战

弹性和费用

Optimized TensorFlow

MXNet Dynamic Training

Amazon Elastic Inference

AWS Inferentia

数据

Amazon SageMaker Ground Truth

Amazon SageMaker RL

易用性

AWS Marketplace for Machine Learning

Amazon SageMaker Neo

Amazon Textract

Amazon Forecast

Amazon Personalize

Amazon Polly

Amazon Lex

Amazon Rekognition

Amazon Transcribe

Amazon Comprehend

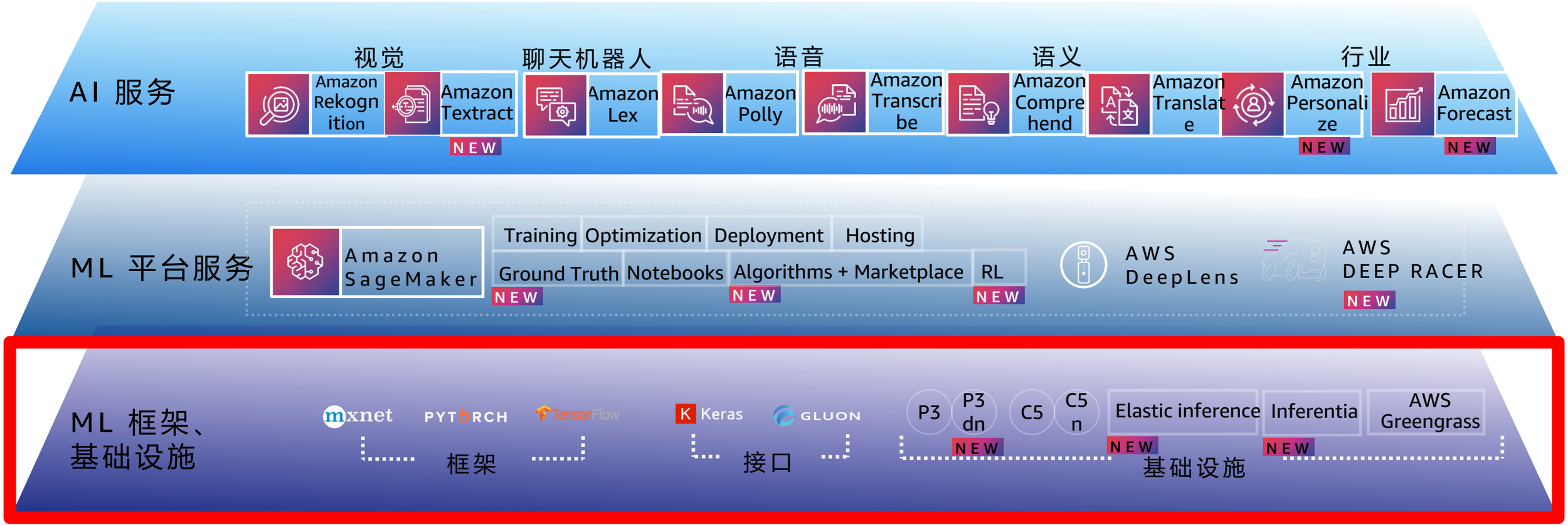
Amazon Translate

ML 框架和基础设施服务

aws INNOVATE
ONLINE CONFERENCE

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

AWS 上的机器学习



AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

AWS 优化的 TensorFlow NEW

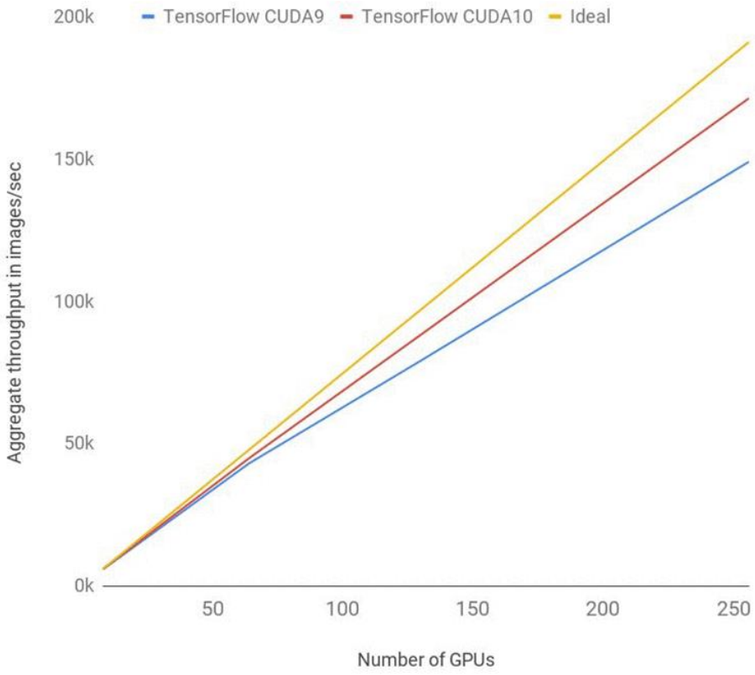
在256颗 GPU 上达到90% 的扩展性能

原始版本 TensorFlow	AWS 优化的 TensorFlow
65%	**90%**
在256 GPU 的扩展性能	在 256 GPU 的扩展性能
30m	**14m**
训练时间	训练时间

HOROVOD

TensorFlow 目前较快的纪录

Amazon SageMaker 和 AWS 深度学习 AMIs 都提供支持

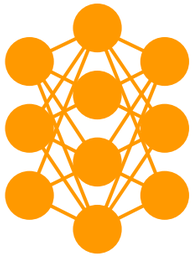


Amazon EC2 P3dn 实例 NEW

最大 P3 实例，针对分布式训练进行优化

减少机器学习训练时间

更好地利用 GPU



支持更大、更复杂的模型

主要特征

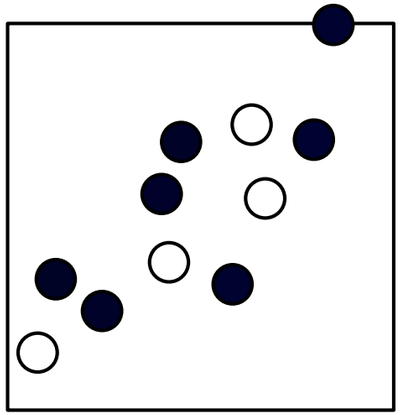
网络带宽为100Gbps
(P3的4倍)

8个 NVIDIA Tesla
V100 GPU

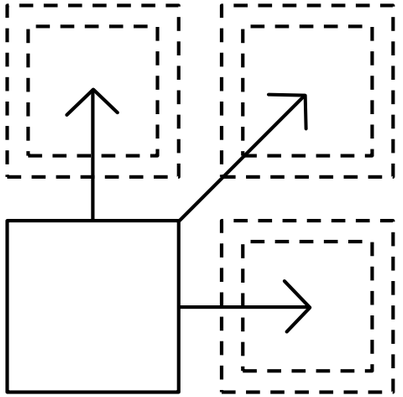
每个 GPU 的内存为 32GB
（共256GB,
是 P3 的2倍）

96个Intel Skylake
vCPU（比 P3 多50%）
，支持 AVX-512

预测在生产过程中的挑战



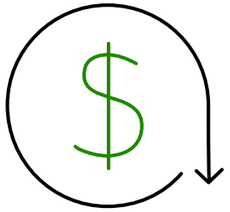
单一的配置不能满足所有的需求



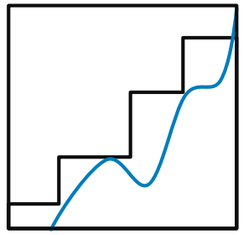
弹性是关键

Amazon Elastic Inference NEW

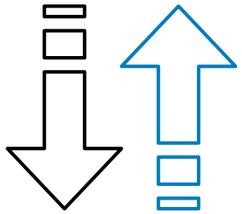
将深度学习推理成本降低多达75%



推理成本降低



匹配容量与需求



可提供1到32 TFLOPS
的处理速度

主要特征

与 Amazon EC2、
Amazon SageMaker 及
Amazon DL AMI 相集成

支持 TensorFlow、Apache
MXNet 和 ONNX 模型，即将
推出 PyTorch 框架

单精度与混合精度运算

将深度学习推理成本降低多达75%

c5.large 实例（cpu）	0.23 秒	
p2.xlarge (NVIDIA K80)	0.042 秒	$0.90 / 小時 (us-east-1)
c5.large + eia1.medium	0.046 秒	$0.22 / 小時(us-east-1)

成本降低75%

AWS Inferentia: 机器学习推理芯片 NEW

AWS 定制化高效率机器学习推理芯片

预计2019下半年推出

- 高吞吐量
- 低延迟
- 上百 TOPS

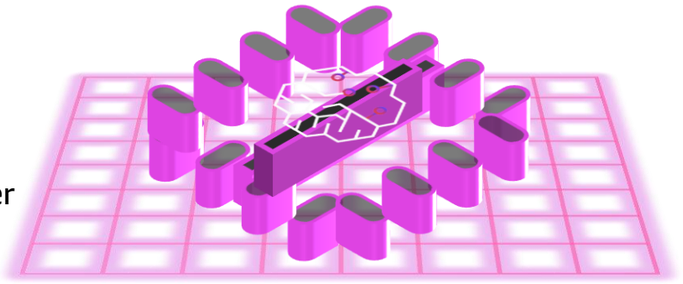
多种数据类型

INT8, FP16,
mixed precision

支持多种
ML 框架

TensorFlow, MXNet,
PyTorch, Caffe2, ONNX

- Amazon EC2 实例
- Amazon SageMaker
- Amazon Elastic Inference



ML 平台服务

aws INNOVATE
ONLINE CONFERENCE

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

AWS 上的机器学习

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Amazon SageMaker

一个**全托管的服务**，可以方便**数据科学家**和**研发人员**在智能应用的**生产环境**中快速和轻松得**构建**基于机器学习的模型。

预置 Jupyter notebook

数据探索、预处理

内置高性能机器学习算法

选择并优化你的机器学习算法

一键训练

创建并管理训练环境

优化

训练与超参数调优

一键部署

生产环境模型部署

全托管、自动扩展、健康检查、错误处理、安全检查

扩展与管理生产环境

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Amazon SageMaker 内置算法

Classification:

- Linear Learner
- XGBoost
- Factorization Machines

Regression:

- Linear Learner
- XGBoost
- Factorization Machines

Recommendations:

- Factorization Machines

Clustering:

- K-Means

Forecasting:

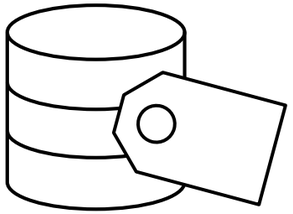
- DeepAR
- Linear Learner
- XGBoost

Dimensionality Reduction/Anomaly Detection:

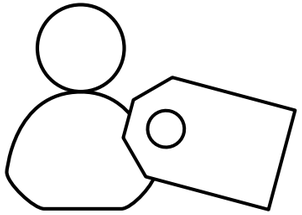
- PCA

Amazon SageMaker Ground Truth NEW

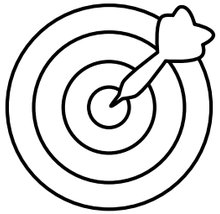
简单准确地标记机器学习训练数据



快速标记训练数据



轻松集成人工标识器



获得准确结果

主要特征

自动标记与机器学习

提供现成和自定义的图像边框、分割和文本工作流

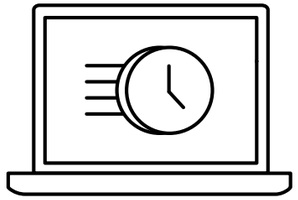
可使用私有和公有劳动力

标记管理

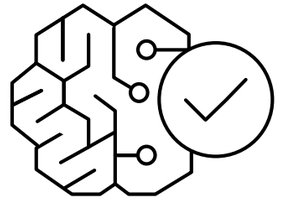
AWS Marketplace for Machine Learning NEW

在 AWS Marketplace 上获取机器学习算法和模型，通过简单的操作部署到 Amazon Sagemaker 上

浏览或搜索 AWS Marketplace



单击订阅



可在 Amazon SageMaker 中使用

主要特征

销售商

IP 保护

自动计费与计量

可选择付费、免费和开放源算法和模型

数据保护

采购者

150 多种可用模型和算法

已有的供货商

DimensionalMechanics

Open for Innovation KNIME

DEEPVISION

H_2O.ai

wipro

Knowledgent

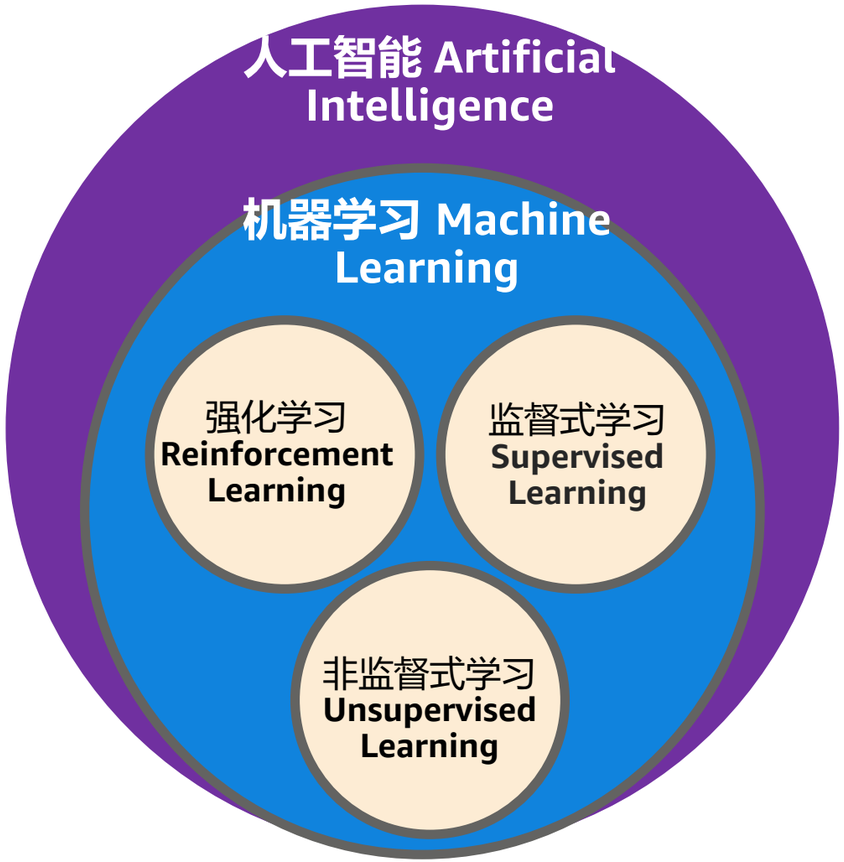
figure eight

TIBCO™

一些可用的算法与模型

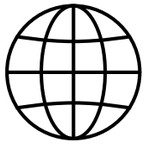
自然语言处理	语法与分析	文本 OCR	计算机视觉	命名实体识别	视频分类
语言识别	文本转语言	扬声器识别	文本分类	3D 图像	异常检测
文本生成	对象检测	回归	文本聚类	手写识别	等级

强化学习 (Reinforcement Learning, RL)

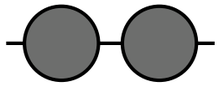


强化学习 (Reinforcement Learning, RL)

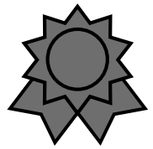
透过真实或模拟环境互动来学习



建立环境来模拟现实世界的问题



反复试验观察结果

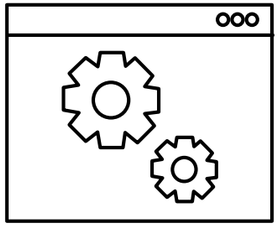


优化学习策略以获得最高的长期回报

训练机器模型进行复杂的决策

Amazon SageMaker RL NEW

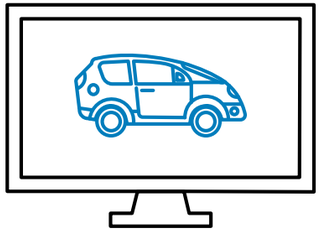
适用于每位开发人员和数据科学家的强化学习



完全托管

mxnet
TensorFlow
intel riselab UC Berkeley

广泛支持框架



广泛支持模拟环境

包括 SimuLink、MatLab

主要特征

支持 TensorFlow、Apache MXNet、Intel Coach 与 Ray RL	支持 2D & 3D 物理环境与 OpenAI Gym	支持 Amazon Sumerian 与 Amazon RoboMaker	示例 Notebook 和教程

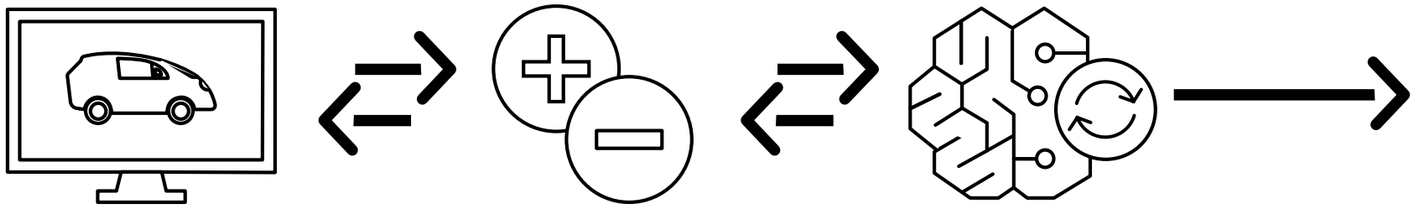
AWS DeepRacer NEW

由强化学习驱动的全自动1:18比例赛车

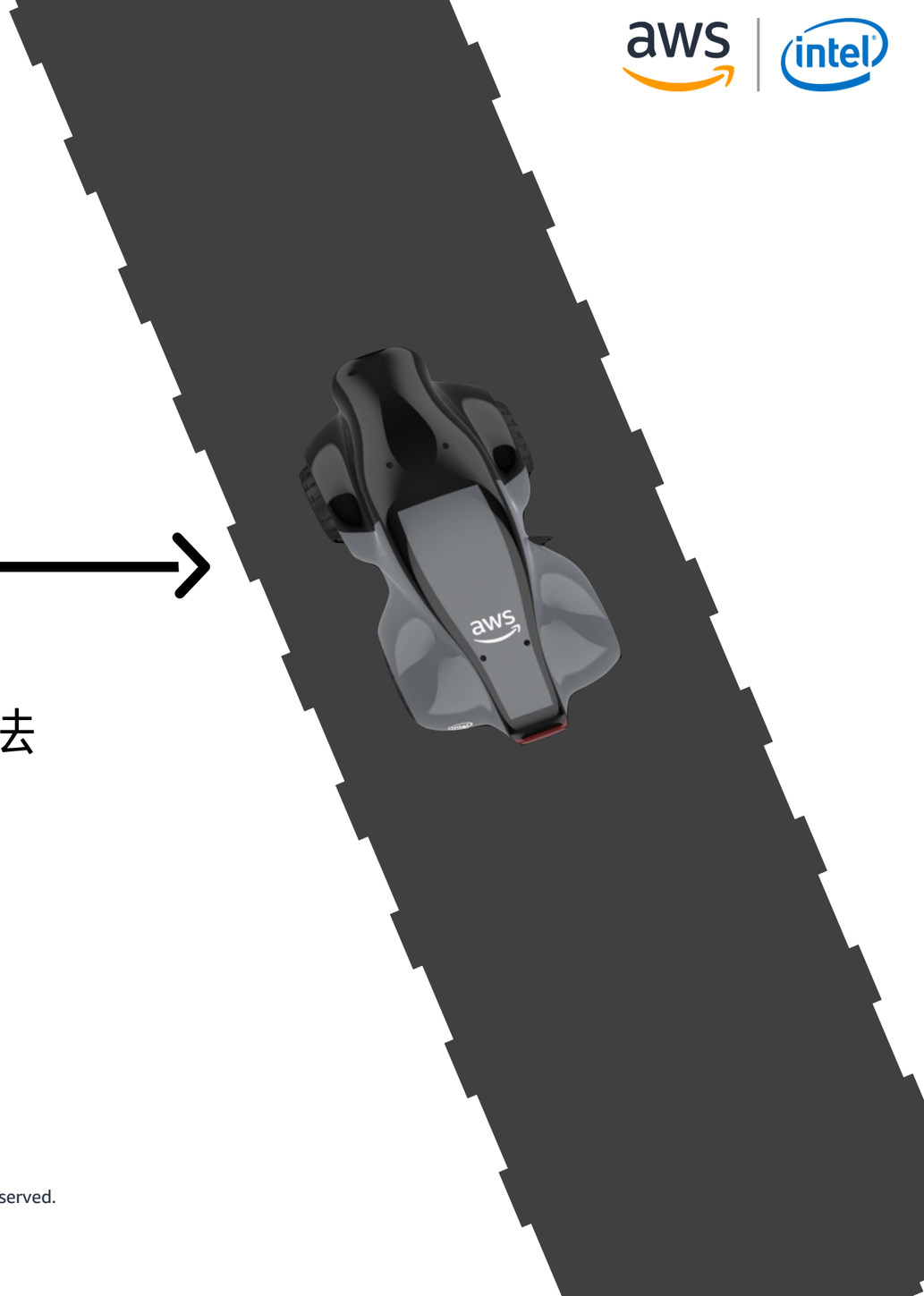
AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

AWS DeepRacer NEW

由强化学习驱动的全自动1:18比例赛车

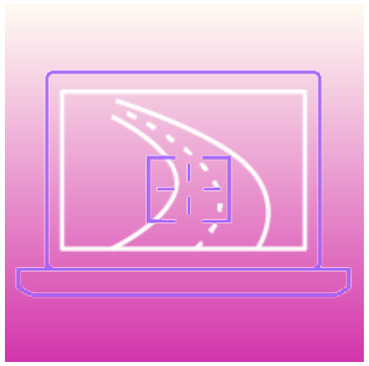


模拟环境　　记分函数　　强化学习算法



自动驾驶赛车杯 Deep Racer League NEW

现场赛



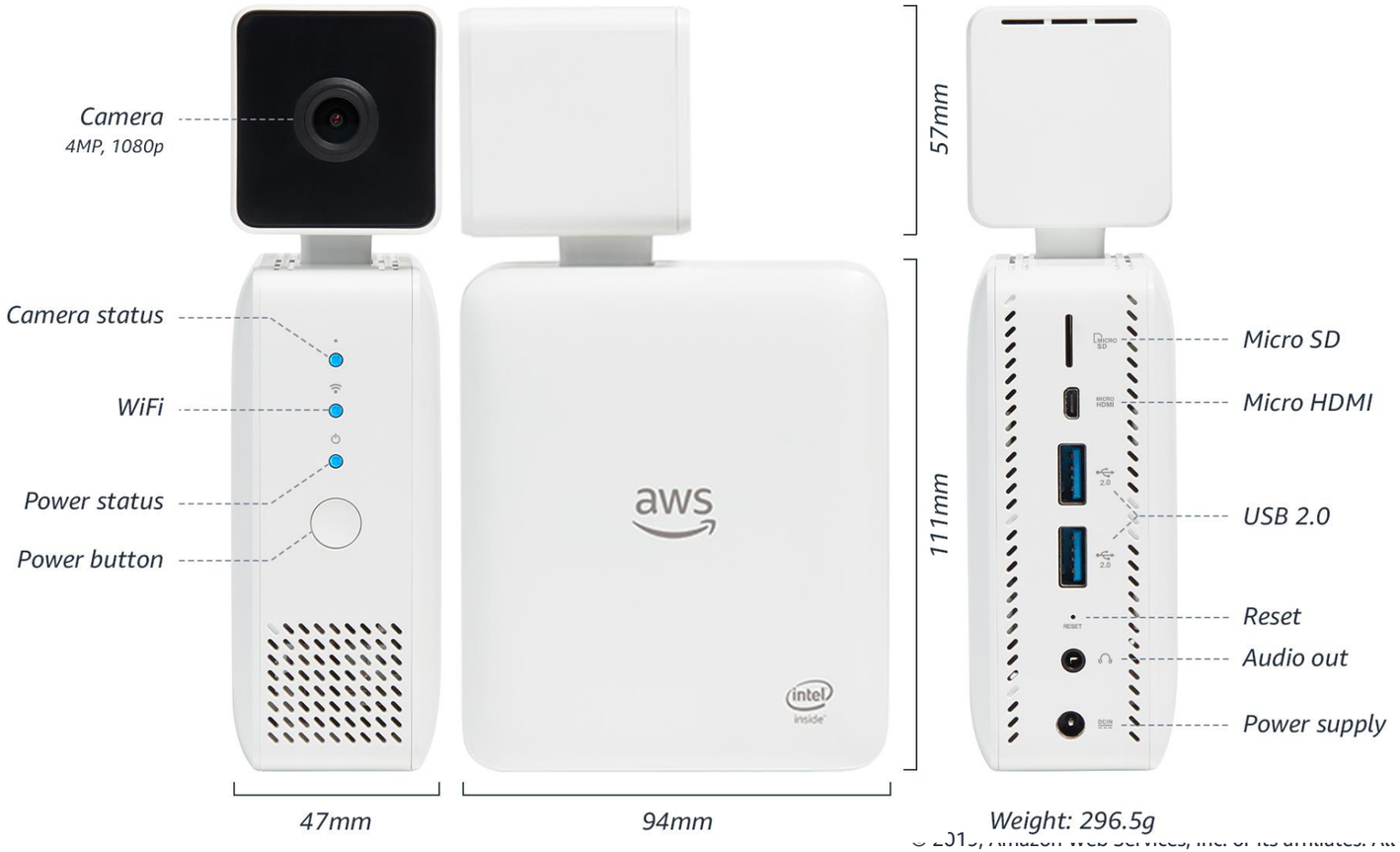
仿真环境

https://aws.amazon.com/deepracer/league/

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

AWS DeepLens

开发人员的深度学习摄像头



- Intel Atom® 处理器、 Ubuntu OS-16.04 LTS、 英特尔第九代显卡引擎、8GB RAM、 16GB 存储（可扩展）
- 完全可编程的视频摄像头
- 设备端优化的深度学习框架 Apache MXNet, Caffe, TensorFlow
- 教程，样例代码，示例，预制模型（对象检测、人脸检测、行为识别、画风转换、猫狗识别、热狗检测、 头部姿势检测 、社区项目等）
- 和 Amazon Sagemaker, Amazon Rekognition, Amazon Polly 集成

AI 服务

aws INNOVATE
ONLINE CONFERENCE

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

AWS 上的机器学习

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Amazon Rekognition 图像

基于深度学习的图像识别服务

对象和场景检测

面部分析

面孔比较

面部识别

名人识别

图像审核

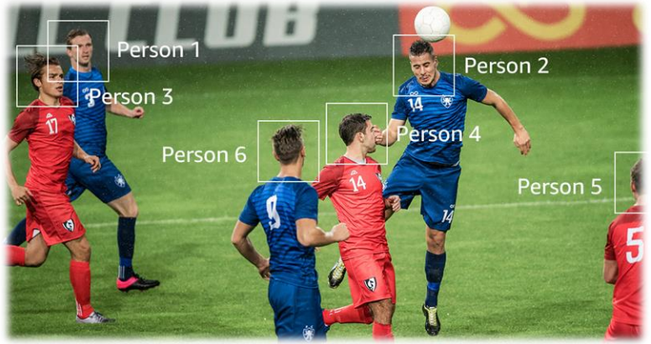
图像中的文本

从视觉内容中提取丰富的元信息

Amazon Rekognition 视频

基于深度学习视频分析

物体、活动检测



路径追踪

面部检测、识别

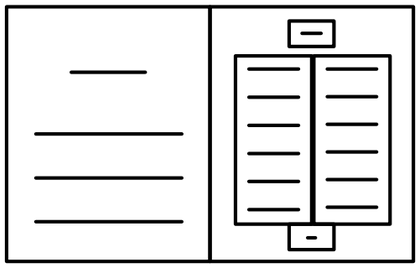
名人识别

不安全视频检测

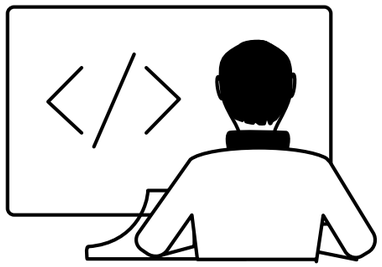
实时视频流处理

Amazon Textract

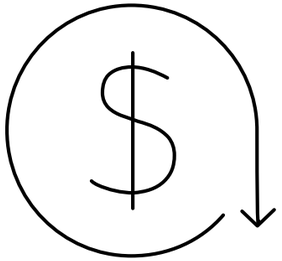
NEW



快速精准提取文本



消除人工干预



低成本

主要特征

OCR　　键值检测　　表格检测　　可调节的置信度阈值　　边界框坐标　　无需机器学习知识

Amazon Textract

NEW

对文档内容的有序整理

Polychronidou *et al. BMC Bioinformatics* 2018, **19**(Suppl 14):414 Page 76 of 176

Table 5 Comparison of clustering accuracy between TM-score and the various 3D descriptors (optimal number of clusters) for the highest accuracy is highlighted

TM-score	8	89.7%
FPFH	9	89.3%
3DSC	9	89.5%
RSD	7	92.0%
VFH	8	85.3%
Combined silhouette weights	7	**92.2%**
Combined equal weights	7	90.2%

most of the proteins have been correctly clustered, with few exceptions. Moreover, the clustering method discovered 7 clusters, instead of 6, splitting stereotyped subset #2 (green-blue color) into two clusters (indexes 4 and 7). The reason behind this separation is probably the pattern of somatic mutations in the immunoglobulin heavy-chain variable region gene (IGHV).

Clustering of all BcR IGs

The procedure followed for clustering the annotated dataset was repeated, this time using the whole BcR IG protein dataset, including both stereotyped (annotated) and non-stereotyped (non annotated) cases. For each type of descriptor, the optimal number of clusters was first determined, using the maximum average silhouette width method. Then, the proteins were clustered using the *k*-medoids method with the optimal number of clusters.

The performance of the various clusterings was evaluated using two types of measures. The first is the average silhouette width itself, which is a measure of the cluster compactness and separation. In general, clustering is based on the assumption that the underlying data form compact clusters of similar characteristics. Larger average silhouette width means that the result of a clustering algorithm consists of compact clusters which are well separated from each other, i.e. probably close to the actual data distribution. A small average silhouette width means e.g. that one of the clusters discovered by the clustering algorithm could be separated in two clusters, or that some of the discovered clusters could be merged together. The average silhouette width is an internal evaluation measure, in the sense that it uses only information contained in the dataset, without assuming any knowledge of ground truth class labels or clusterings.

The second type of evaluation measure is the Rand index, which is an external measure, in the sense that it makes use of ground truth knowledge. The evaluation using the Rand index is similar to the evaluation of the annotated dataset in the previous section, by comparing the produced clusterings to the ground truth clustering. However, only the annotated BcR IG were used for the computation of the Rand index. In other words, after computing a clustering of all proteins, both annotated and unannotated, we wanted to evaluate how well they have been clustered by examining the clustering distribution

OCR++

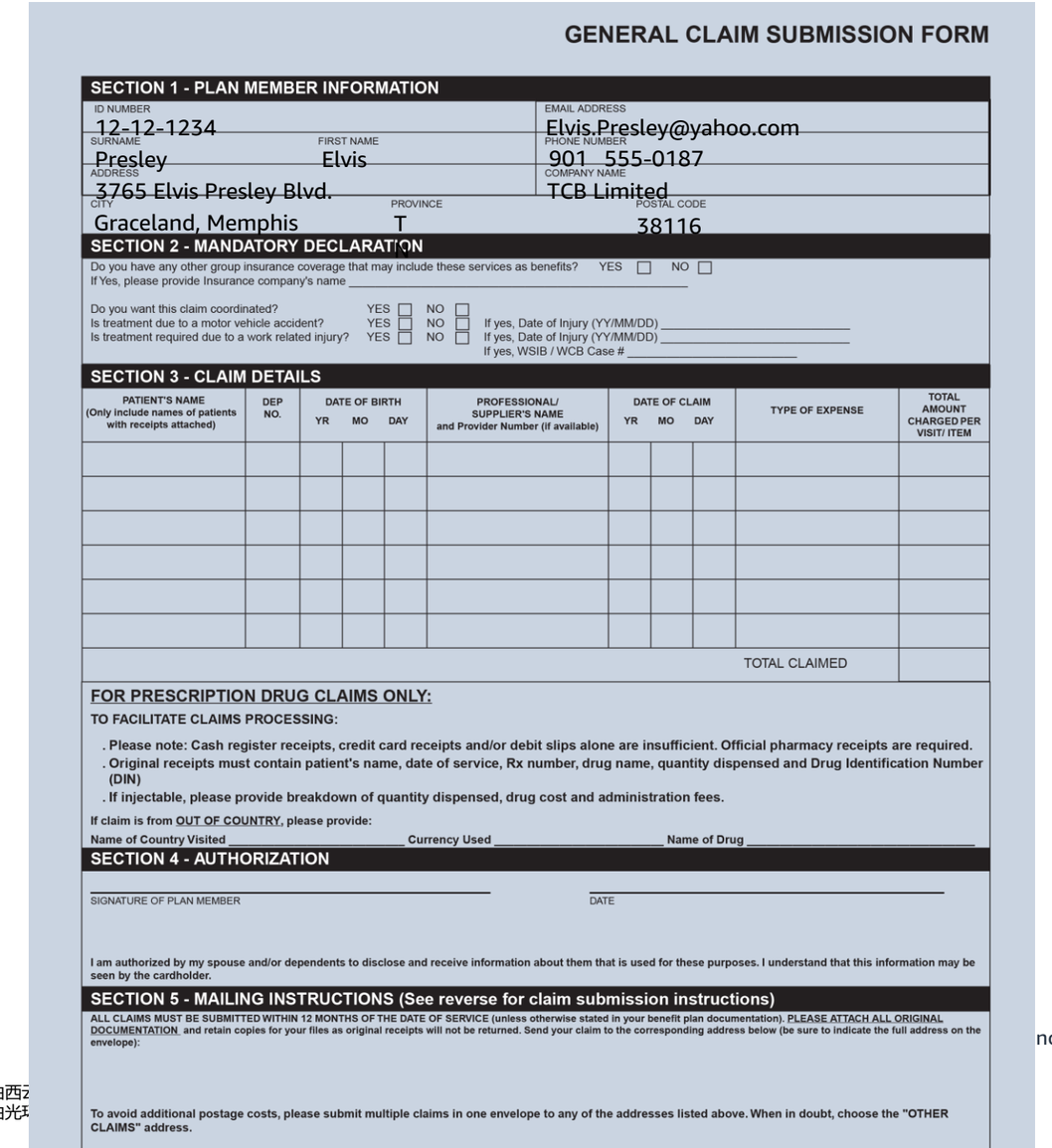
Amazon Textract

Method	Num. clusters	Rand index
TM-score	8	89.7%
FPFH	9	89.3%
3DSC	9	89.5%
RSD	7	92.0%
VFH	8	85.3%
Combined silhouette weights	7	92.2%
Combined equal weights	7	90.2%

Amazon Aurora

Amazon Textract NEW

自动文档处理，无需数据输入，也无需编写规则



GENERAL CLAIM SUBMISSION FORM

SECTION 1 - PLAN MEMBER INFORMATION

ID NUMBER: 12-12-1234
EMAIL ADDRESS: Elvis.Presley@yahoo.com
SURNAME: Presley
FIRST NAME: Elvis
PHONE NUMBER: 901 555-0187
ADDRESS: 3765 Elvis Presley Blvd.
COMPANY NAME: TCB Limited
CITY: Graceland, Memphis
PROVINCE: T
POSTAL CODE: 38116

SECTION 2 - MANDATORY DECLARATION

Do you have any other group insurance coverage that may include these services as benefits? YES ☐ NO ☐
If Yes, please provide Insurance company's name _____

Do you want this claim coordinated? YES ☐ NO ☐
Is treatment due to a motor vehicle accident? YES ☐ NO ☐ If yes, Date of Injury (YY/MM/DD) _____
Is treatment required due to a work related injury? YES ☐ NO ☐ If yes, Date of Injury (YY/MM/DD) _____
If yes, WSIB / WCB Case # _____

SECTION 3 - CLAIM DETAILS

PATIENT'S NAME (Only include names of patients with receipts attached)	DEP NO.	DATE OF BIRTH YR	MO	DAY	PROFESSIONAL/ SUPPLIER'S NAME and Provider Number (if available)	DATE OF CLAIM YR	MO	DAY	TYPE OF EXPENSE	TOTAL AMOUNT CHARGED PER VISIT/ ITEM
									TOTAL CLAIMED	

FOR PRESCRIPTION DRUG CLAIMS ONLY:

TO FACILITATE CLAIMS PROCESSING:

. Please note: Cash register receipts, credit card receipts and/or debit slips alone are insufficient. Official pharmacy receipts are required.
. Original receipts must contain patient's name, date of service, Rx number, drug name, quantity dispensed and Drug Identification Number (DIN)
. If injectable, please provide breakdown of quantity dispensed, drug cost and administration fees.

If claim is from OUT OF COUNTRY, please provide:
Name of Country Visited _____ Currency Used _____ Name of Drug _____

SECTION 4 - AUTHORIZATION

SIGNATURE OF PLAN MEMBER _____ DATE _____

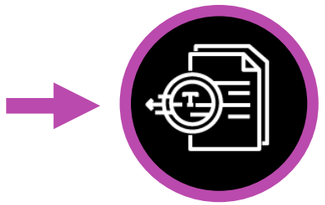
I am authorized by my spouse and/or dependents to disclose and receive information about them that is used for these purposes. I understand that this information may be seen by the cardholder.

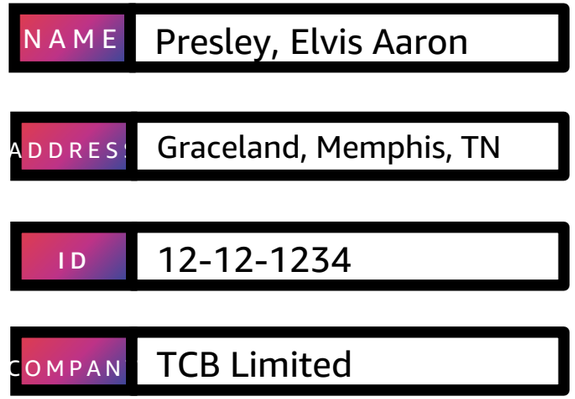
SECTION 5 - MAILING INSTRUCTIONS (See reverse for claim submission instructions)

ALL CLAIMS MUST BE SUBMITTED WITHIN 12 MONTHS OF THE DATE OF SERVICE (unless otherwise stated in your benefit plan documentation). PLEASE ATTACH ALL ORIGINAL DOCUMENTATION and retain copies for your files as original receipts will not be returned. Send your claim to the corresponding address below (be sure to indicate the full address on the envelope):

To avoid additional postage costs, please submit multiple claims in one envelope to any of the addresses listed above. When in doubt, choose the "OTHER CLAIMS" address.

OCR++





AWS 中国（宁夏）区域由西云
AWS 中国（北京）区域由光环

Amazon Polly

使用深度学习将文本转换为逼真的语音

逼真的语音

全托管

57种声音

28种语言
支持中文

低延迟，实时

语音质量与发音

1. 自动、精确的文本处理
2. 清楚、易懂
3. 为文本增加语义理解能力
4. 定制化读音

```
<speak>
<amazon:effect name="whispered">
        If you make any noise,
</amazon:effect>
        she said,
<amazon:effect name="whispered">
        they will hear us.
</amazon:effect>
</speak>
```

Amazon Transcribe

全托管和不断训练的自动语音识别 (ASR) 服务，可接收音频并自动生成准确的文字记录。

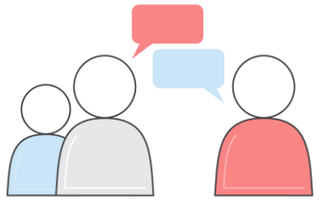
支持普通语音和电话语言记录

Amazon S3 集成

时间戳

标点符号 格式化

话者识别



实时音频流

词汇定制

示例用例：

- 呼叫中心
- VOD 字幕生成
- 广播字幕生成
- 录制会议

Amazon Lex

使用文本和语音构建自然的会话交互界面

语音、文本
“聊天机器人”

和 Alexa
相同的技术

移动，Web 和
设备语音交互

和 Slack & Messenger
文本交互
(更多支持中...)

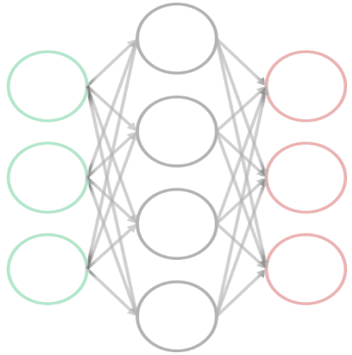
企业接口

Salesforce
Microsoft Dynamics
Marketo
Zendesk
Quickbooks
Hubspot

Amazon Translate

全托管翻译服务，使用深度学习提供高质量、快速且价格合理的语言翻译服务。

实时翻译



深度学习引擎

21 种语言之间互译，
417种组合

语言检测，
自定义术语

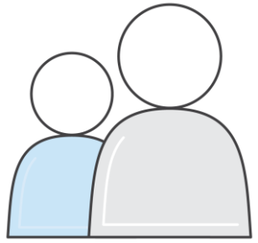
示例用例：

- 多语言应用程序(支持中文)
- 文本分析工作
- 与其他 AWS 服务协作 (如 Amazon Polly 或 Amazon Transcribe)

Amazon Comprehend

全托管的自然语言处理 (NLP) 服务，使用深度学习智能发掘任何文本的内容。

情感分析

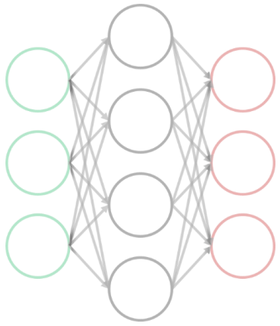


命名实体

关键词

多语言支持

主题建模



自定义实体
自定义分类器

示例用例：

- 社交媒体分析
- 个性化内容推荐
- 电子邮件分析
- 语义搜索
- 知识管理 / 发现

Amazon Comprehend Medical NEW

医学命名实体与关系提取 (NERe API)

受保护的健康信息标识 (PHId API)

实体

- 药物治疗
- 医疗条件
- 试验、处理和程序
- 解剖学
- 受保护的健康信息标识(PHI)

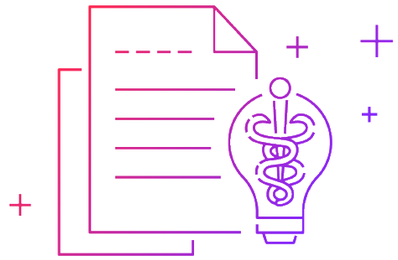
关系提取

- 药物和剂量
- 测试和结果

实体特征

- 否定
- 诊断、体征或症状

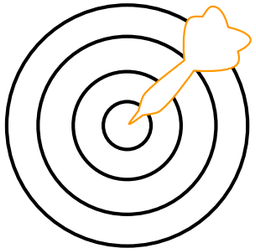
将一个复杂的过程提取为一个简单的 API 调用



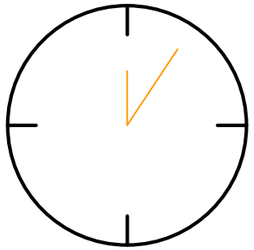
Amazon Personalize NEW

基于在 Amazon.com 中使用的相同技术，

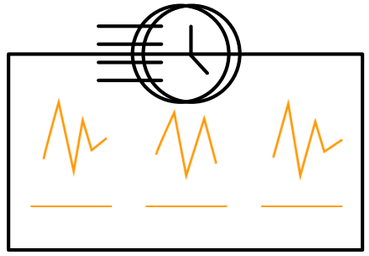
实时提供个性化信息与推荐服务，无需机器学习经验



提供优质的推荐服务



实时



使用简单

几乎适用于任何产品或内容

主要特征

对意图变化作出反应

自动机器学习

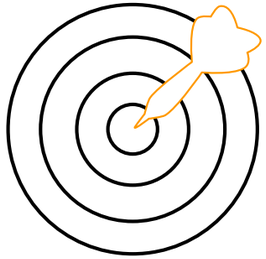
支持深度学习算法

从 Amazon SageMaker 引入现有算法

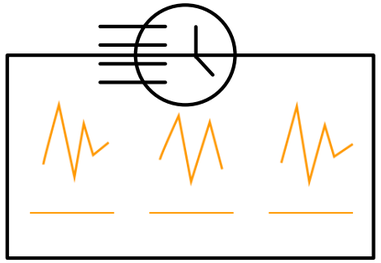
Amazon Forecast NEW

基于在 Amazon.com 上使用的相同技术，

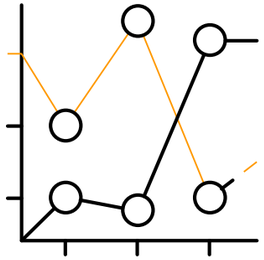
提供精确的时间序列预测服务，无需机器学习经验



整合时间和相关变数的分析，
预测更精准



使用方便，可通过控制台和
API 设置域



适用于任何历史时间序列

主要特征

同时考虑多个时间序列

自动机器学习

通过控制台评估模型的准确性

直观显示控制台中的预测值并将结果导入到业务应用程序中

计划预测与模型再训练

从 Amazon SageMaker 引入现有算法

机器学习

视觉
Amazon Rekognition
Amazon Textract NEW
聊天机器人
Amazon Lex
语音
Amazon Polly
Amazon Transcribe
语义
Amazon Comprehend
Amazon Translate
Amazon Personalize NEW
行业
Amazon Forecast NEW

Amazon SageMaker
Training Optimization Deployment Hosting
Ground Truth NEW Notebooks Algorithms + Marketplace NEW RL NEW
AWS DeepLens
AWS DEEP RACER NEW

mxnet PYTORCH TensorFlow Keras GLUON
框架 接口
P3 P3dn NEW C5 C5n NEW Elastic inference Inferentia NEW AWS Greengrass
基础设施

分析

Amazon Athena
Amazon EMR Hadoop, Spark, Presto, Pig, Hive…19 total
Amazon Redshift + Redshift Spectrum
AWS Glue
Amazon Elasticsearch Service
Amazon Kinesis
Amazon QuickSight

存储

Amazon EBS
Amazon S3 Standard-IA
Amazon S3 Standard
NEW Amazon S3 Intelligent-Tiering
Amazon S3 One Zone-IA
Amazon Glacier
NEW Amazon S3 Glacier Deep Archive

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

感谢参加 AWS INNOVATE 2019 在线技术大会

我们希望您在这里找到感兴趣的内容！

也请帮助我们完成**投票打分**和**反馈问卷**。

欲获取关于 AWS 的更多信息和技术内容，可以通过以下方式找到我们：


微信公众号：AWSChina


新浪微博：https://www.weibo.com/amazonaws/


领英：https://www.linkedin.com/company/aws-china/


知乎：https://www.zhihu.com/org/aws-54/activities/


视频中心：http://aws.amazon.bokecc.com/


更多线上活动：https://aws.amazon.com/cn/about-aws/events/webinar/

aws INNOVATE
ONLINE CONFERENCE

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营