Amazon SageMaker

端到端的托管机器学习平台

Damon Deng，AWS资深架构师

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

解决计算机科学中十分困难的问题

学习

语言

洞察

解决

解释

ML @ AWS: Our mission

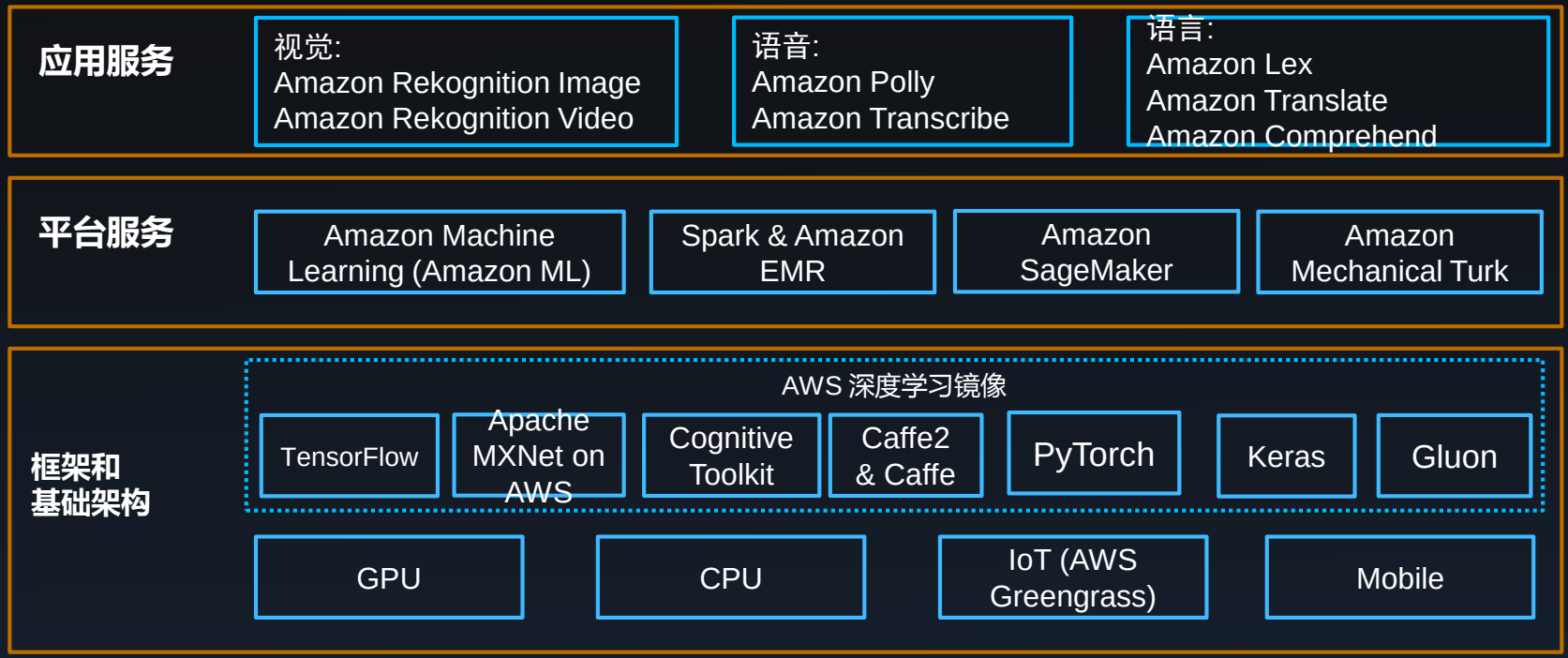
为开发者和数据科学家打造机器学习平台

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

数以万计的用户在AWS上运行机器学习

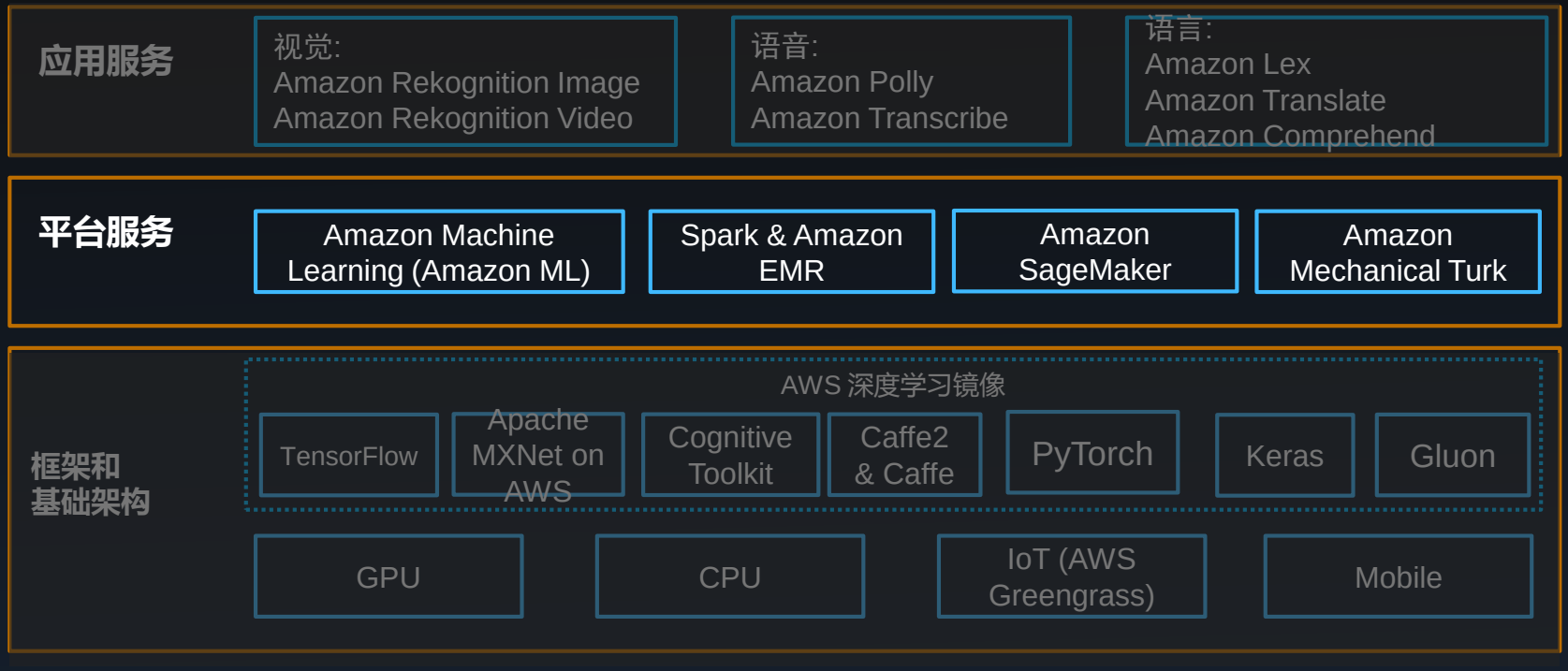
AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

AWS上的机器学习技术堆栈



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

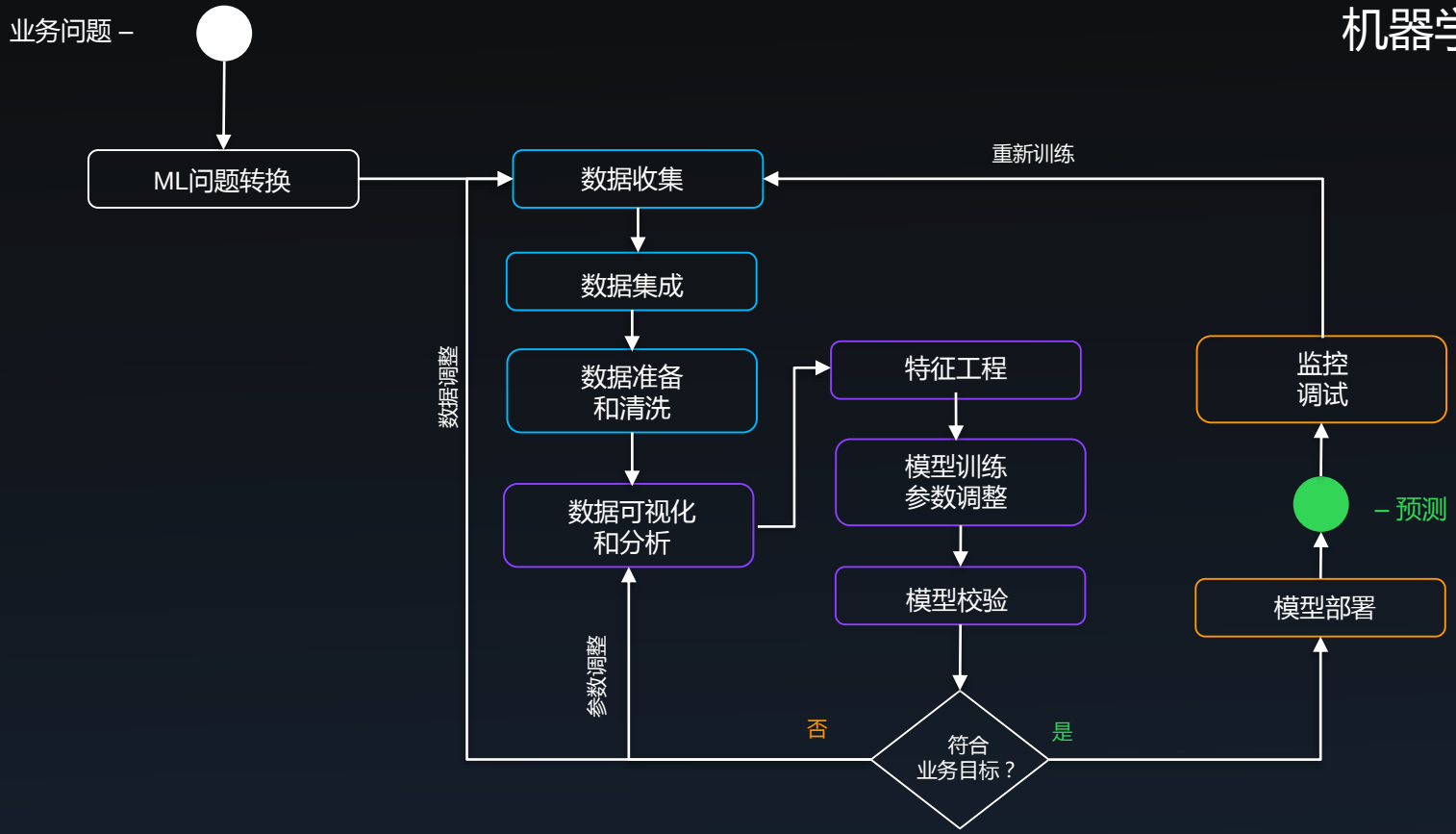
AWS上的机器学习技术堆栈



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

机器学习流程

机器学习流程



NEW!

Amazon SageMaker

A **fully managed service** that enables **data scientists** and **developers** to quickly and easily **build** machine-learning based models **into production** smart applications.

一个**全托管服务**，可以帮助**数据科学家**和**开发者**快速而轻松地**构建**基于机器学习的模型的**生产环境**智能应用

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

Amazon SageMaker

构建

预配置的
notebook
实例

高度优化的
机器学习算法

训练

机器学习，深度学习，定制算法
一键训练

超参数优化
更容易的训练过程

部署

全托管的
可扩展平台

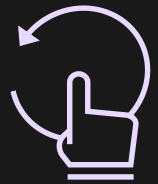
部署过程
不需要工程支撑

AWS 中国（宁夏）
AWS 中国（北京）区域由光环新网运营

aws

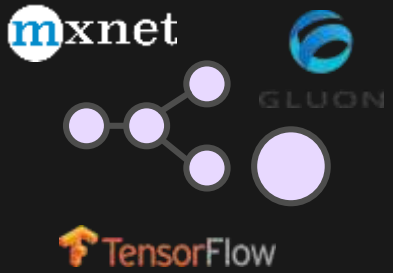
Amazon SageMaker

从创新想法到实际模型实现的十分快速、简单的方法



端到端机器学习平台

零配置



多样的模型训练

按秒付费

/S 中国（宁夏）区域由西云数据运营
/S 中国（北京）区域由光环新网运营

Amazon SageMaker

构建 | **训练** | **部署**

Amazon高性能，可扩展的算法

分布式TensorFlow, Apache MXNet, Chainer, PyTorch

自带算法

超参数调优

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 的组件

构建 | **训练** | **部署**

Amazon高性能，可扩展的算法

分布式TensorFlow, Apache MXNet, Chainer, PyTorch

自带算法

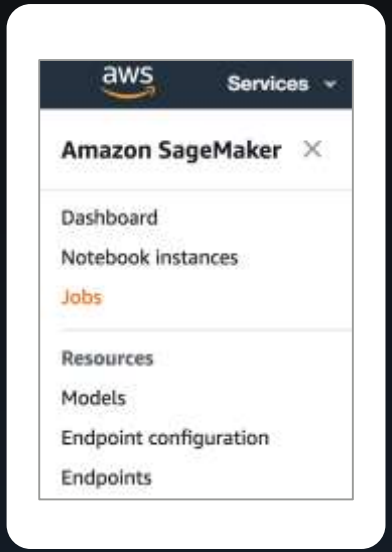
超参数调优

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

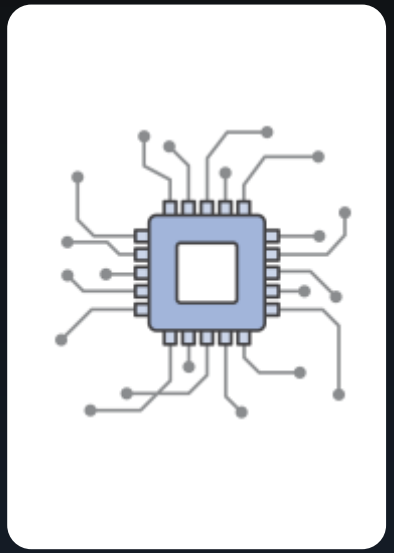
构建

使用Amazon SageMaker托管的笔记本实例…

… 或者通过Amazon EMR和Amazon SageMaker Spark SDK使用Apache Spark...



... 或Amazon SageMaker控制台的点击操作...



... 或者您的设备(Amazon Elastic Compute Cloud (Amazon EC2), laptop等.)

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

构建 | **训练** | **部署**

Amazon高性能，可扩展的算法

分布式TensorFlow, Apache MXNet, Chainer, PyTorch

自带算法

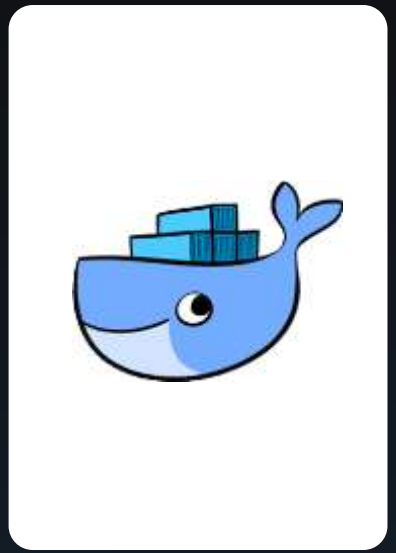
超参数调优

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

训练

一键式训练

流式数据集+分布式计算



Docker / Amazon Elastic Container Service (Amazon ECS)

训练好的模型既可以部署在本地也可以部署在Amazon SageMaker, AWS Greengrass, AWS DeepLens

Amazon SageMaker 的组件

构建

训练

部署

Amazon高性能，可扩展的算法

分布式TensorFlow, Apache MXNet, Chainer, PyTorch

自带算法

超参数调优

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

部署

一键式部署

低延迟，高吞吐，
高可靠

自动A/B测试

自带模型

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的组件

构建 | 训练 | 部署

Amazon高性能，可扩展的算法

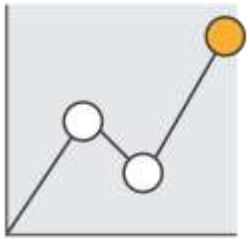
分布式TensorFlow, Apache MXNet, Chainer, PyTorch

自带算法

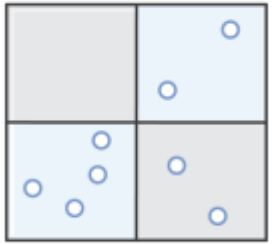
超参数调优

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

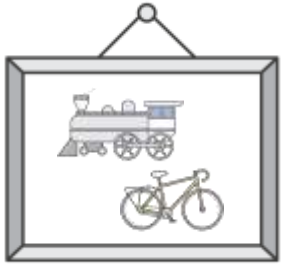
内置算法



用于监督学习的XGBoost，FM，线性和预测算法



Kmeans，PCA和Word2Vec用于聚类和预处理



卷积神经网络的图像分类

用于话题建模的LDA和NTM，用于翻译的seq2seq

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的组件

构建

训练

部署

Amazon高性能，可扩展的算法

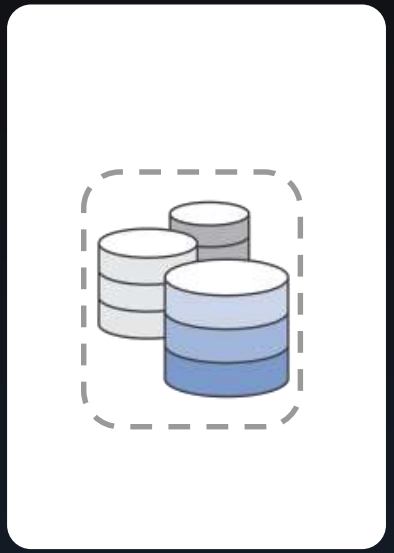
分布式TensorFlow, Apache MXNet, Chainer, PyTorch

自带算法

超参数调优

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

主流深度学习框架容器



采样数据…

…在单独的Notebook实例中探索和细化模型

使用相同的代码在实例集群上对完整数据集进行训练…

… 部署在生产环境

Amazon SageMaker的组件

构建 | 训练 | 部署

Amazon高性能，可扩展的算法

分布式TensorFlow, Apache MXNet, Chainer, PyTorch

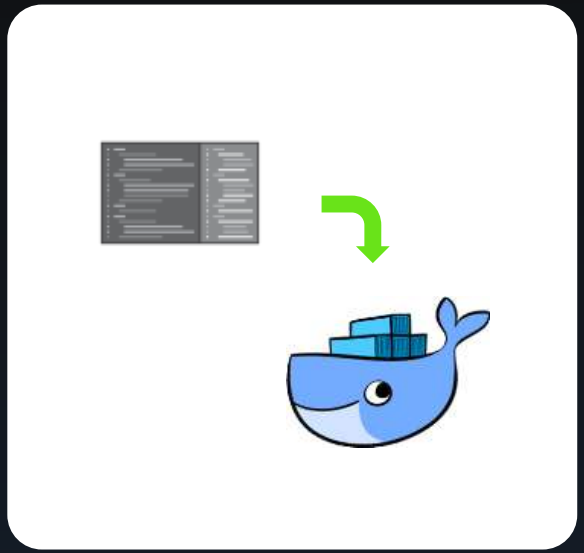
自带算法

超参数调优

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

自带算法

选择您首选的算法...



... 把算法代码加入到Docker容器中...

...发布到Amazon ECS

Amazon SageMaker的组件

构建 | **训练** | **部署**

Amazon高性能，可扩展的算法

分布式TensorFlow, Apache MXNet, Chainer, PyTorch

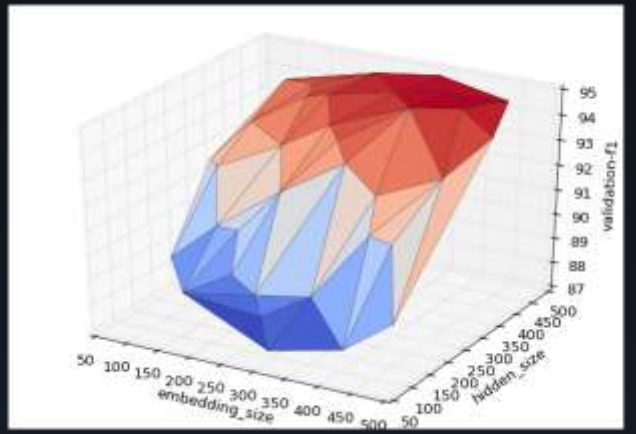
自带算法

超参数调优

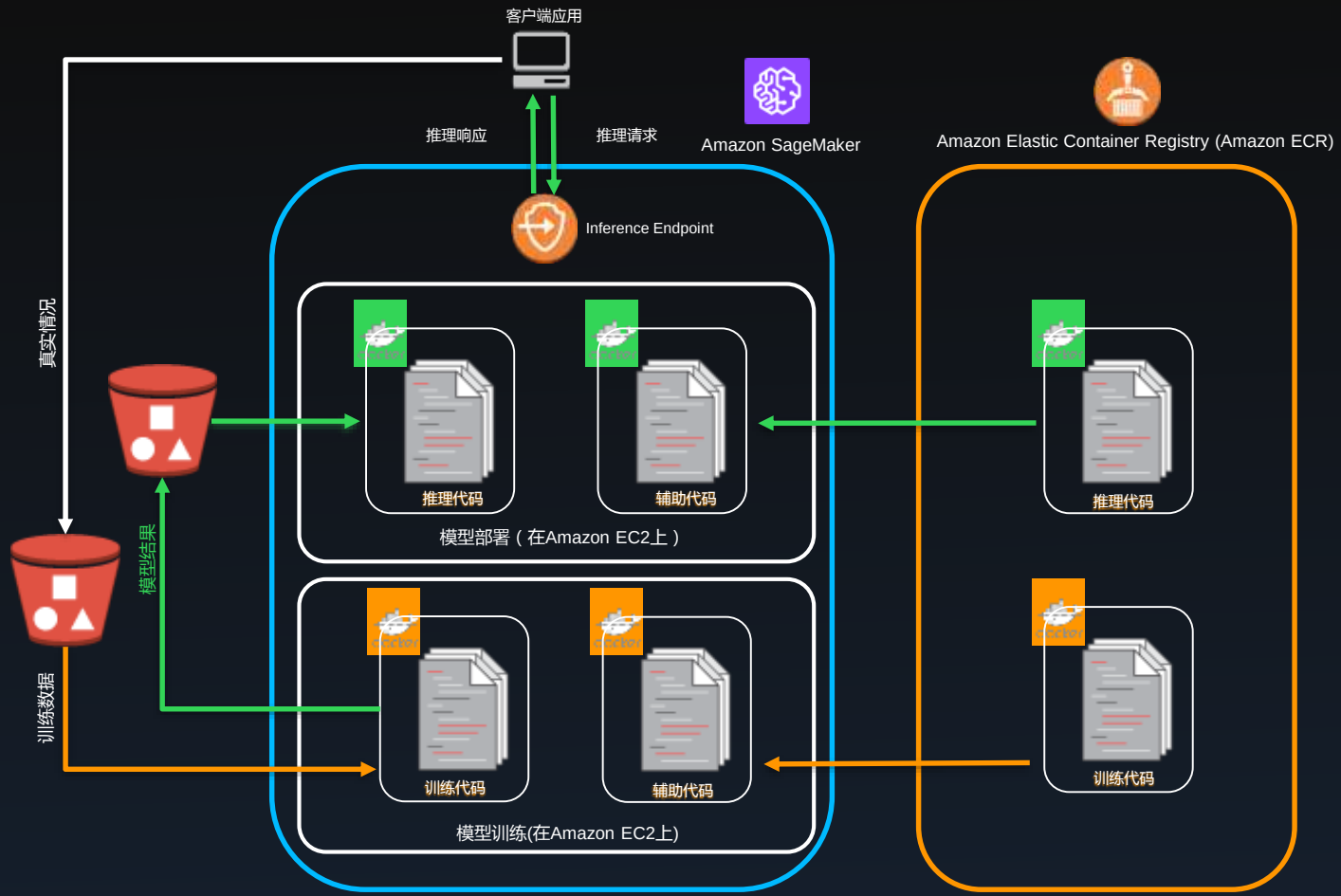
AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

超参数挑优（模型自动调优）

用不同的超参数进行大量的训练作业...



... 搜索超参数空间提高模型精度



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

① Notebook 实例

② 算法

③ 模型训练服务

④ 模型部署服务

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

1

零配置 数据探索分析平台

Notebook实例

“Just add data”

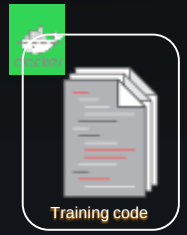
- 推荐、个性化
- 欺诈检测
- 预测
- 图像分类
- 流失预测
- 市场活动、邮件的定位
- 日志处理，异常检测
- 语音文字转换
- 其它更多…

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

② Amazon SageMaker: 10倍速 算法支持

I
算法



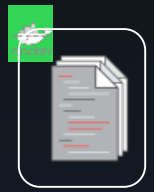
- Matrix Factorization
- Regression
- Principal Component Analysis
- K-Means Clustering
- Gradient Boosted Trees
- 更多算法!

AWS提供的算法

mxnet TensorFlow™

自带脚本 (SM构建好的容器)

Apache Spark 集成



自带算法（自己通过容器构建）

3 全托管的、灵活的 分布式训练

模型训练服务

安全保障

获取训练数据

训练代码

保存模型

保存推理镜像

Amazon ECR

- Matrix Factorization
- Regression
- Principal Component Analysis
- K-Means Clustering
- Gradient Boosted Trees
- 更多算法!

AWS提供的算法

mxnet TensorFlow

自带脚本 (SM构建容器)

Apache Spark 集成

自带算法（自己通过容器构建）

CPU GPU

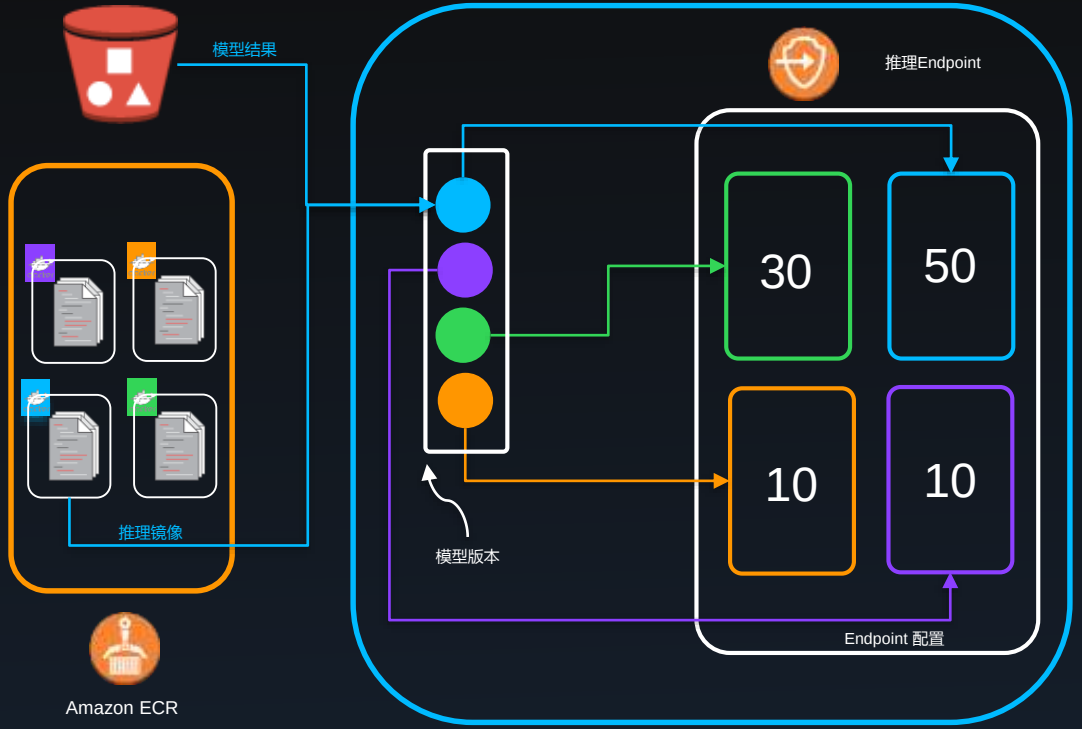
HPO

全托管

4 轻松将模型部署到Amazon SageMaker

模型部署服务

在推理容器中保存多个版本的镜像. Prod 是主要版本，支持50% 用户流量



实例类型: c3.4xlarge

初始实例数量: 3

模型名称: prod

版本名称: primary

初始版本权重: 50

生产版本

一键部署

mxnet

TensorFlow™

AWS提供的算法

AWS 中国（宁夏）区域由西云数据运营

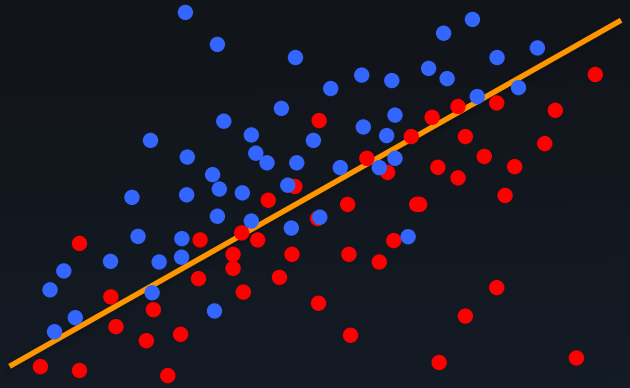
AWS 中国（北京）区域由光环新网运营

aws

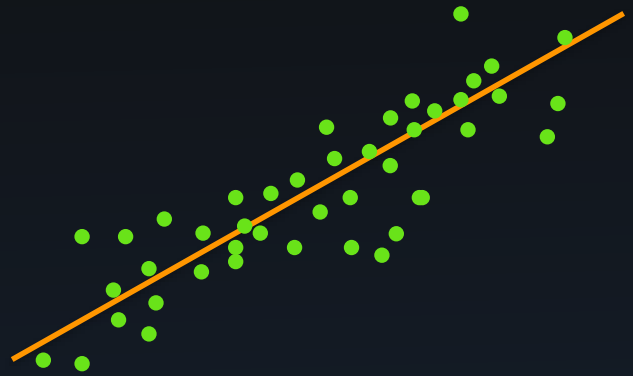
内置的机器学习算法

线性算法

二元分类
预测0/1结果



回归
预测一个数值



$$\tilde{y} = \begin{cases} 1 \text{ if } \langle w, x \rangle > t \\ 0 \text{ else} \end{cases}$$

$$\tilde{y} = \langle w, x \rangle + t$$

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

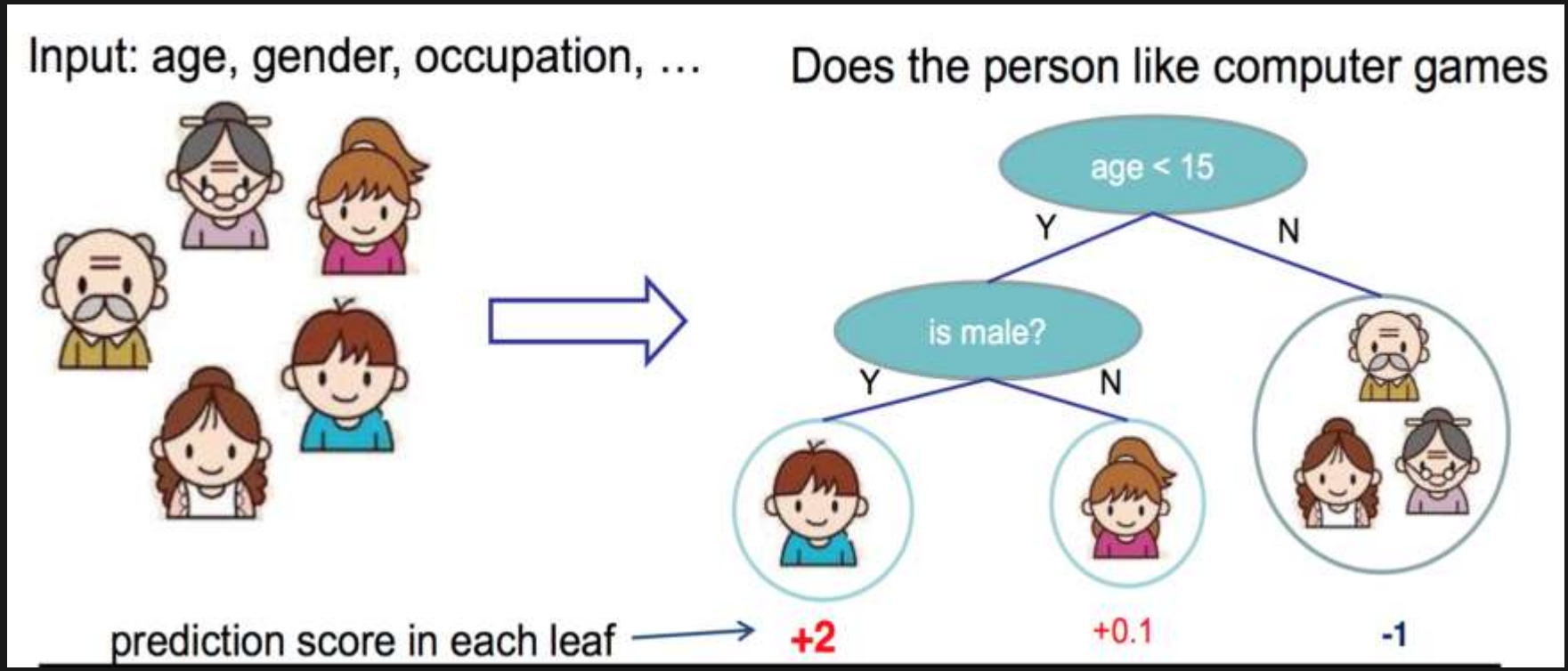
线性学习器的使用场景

- 分类
  - 根据过去客户的反馈，选择是否发邮件给这个特定的客户？ 是/否
  - 根据过去客户的分类，判断该客户属于哪个细分类别？ “空巢老人”，“郊区居民”或“城市白领”
- 回归
  - 根据过去邮件的投资回报率（ROI），邮件此客户的投资回报率是多少？

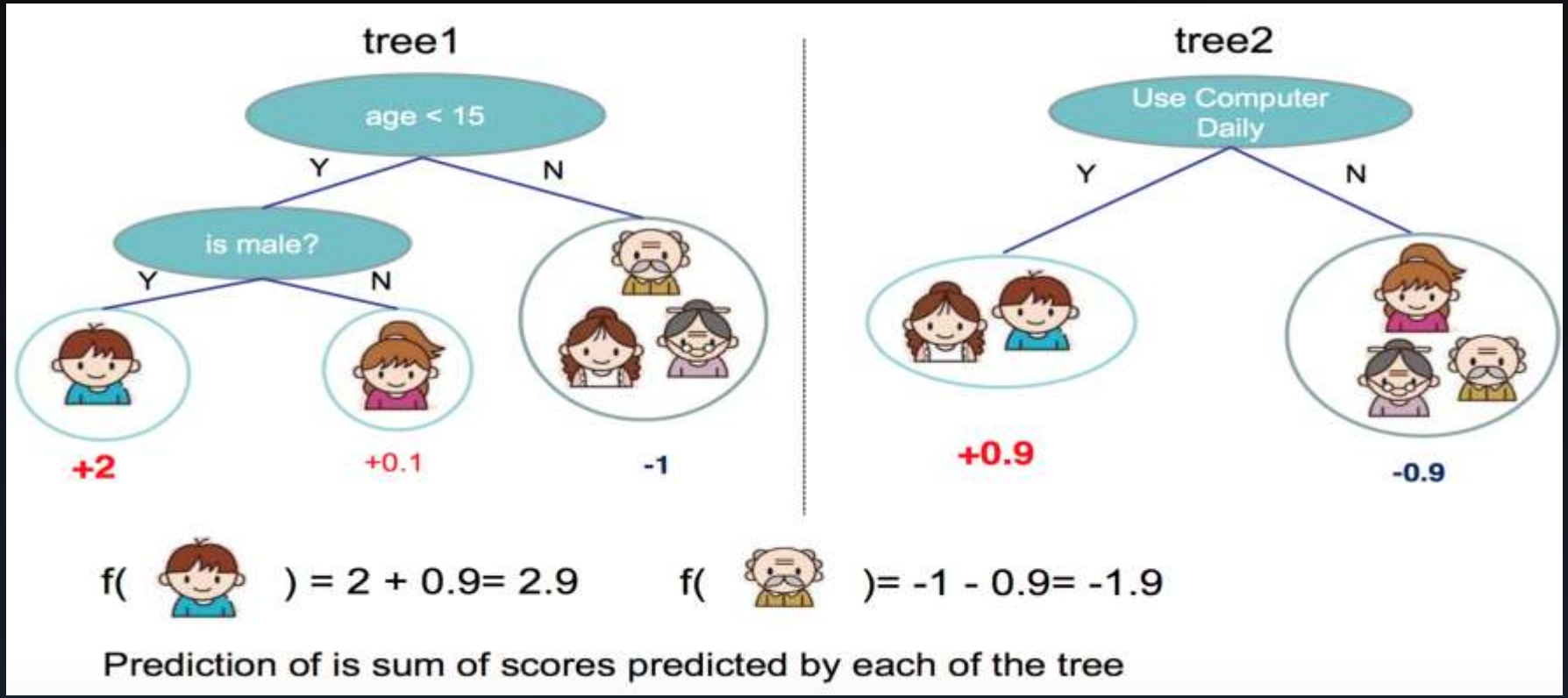
XGBoost

- Extreme Gradient Boosting
  - 基于Gradient Boosting决策树算法（GBDT）
  - 通过组合一组相对简单，能力较弱的模型，把它们的预测结果相加来预测目标变量

XGBoost



XGBoost



XGBoost 的使用场景

- 分类
- 回归
- 排行

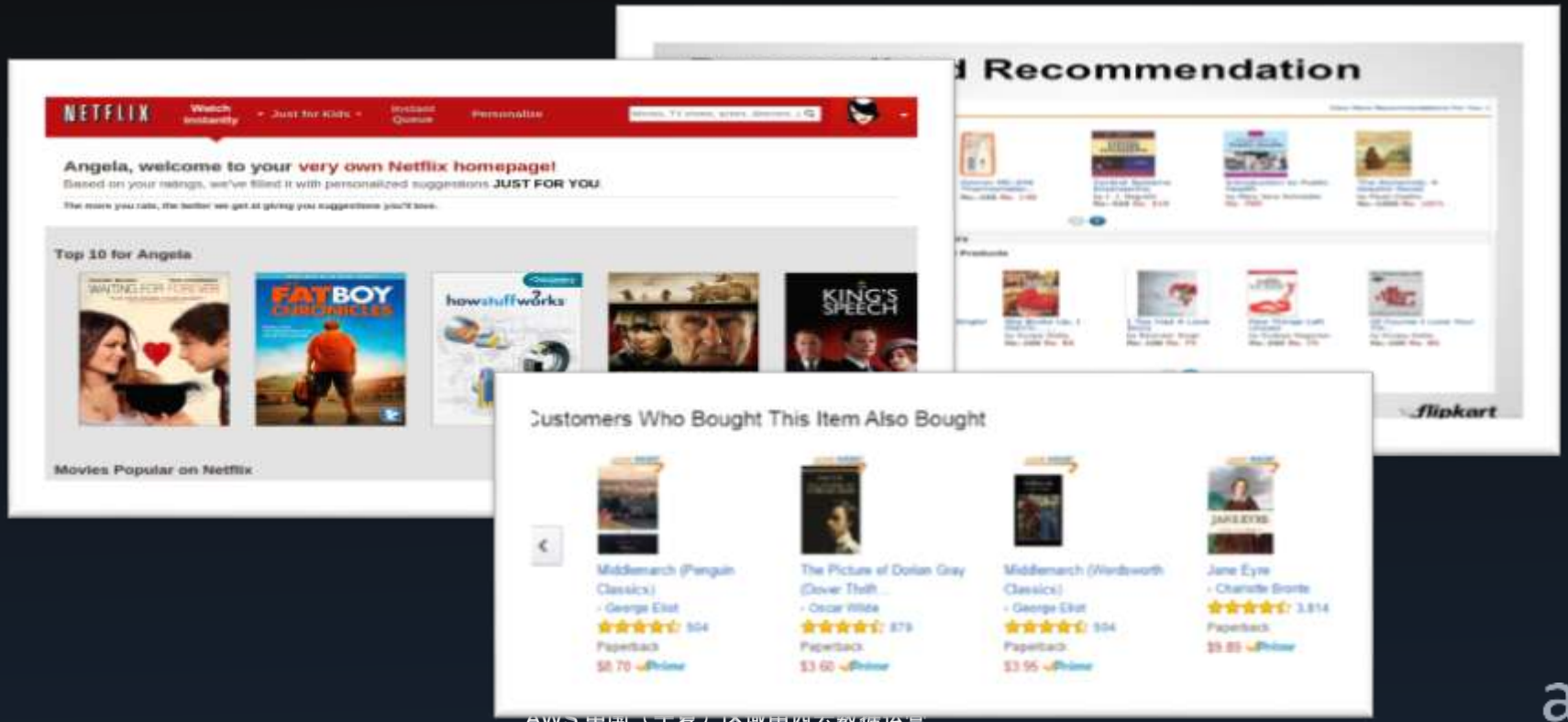
因子分解机（Factorization Machines）

- 线性回归的泛化
  - 每个特征的单独权重 vs k维向量代表特征之间的关系

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
User 1	0	3	0	3	0	0
User 2	4	0	0	2	0	0
User 3	0	0	3	0	0	5
User 4	0	0	0	0	3	0
User 5	4	0	0	4	0	0

$$\tilde{y} = w_0 + \langle w_1, x \rangle + \sum_{i,j>i} x_i x_j \cdot \langle v_i, v_j \rangle$$

因子分解机使用场景



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

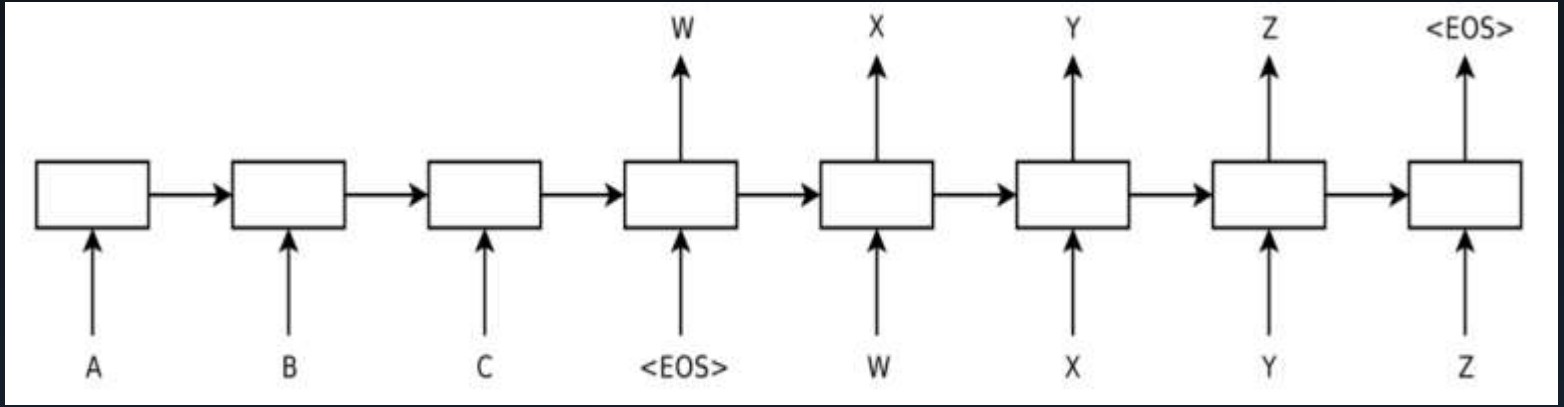
图像分类

- 将图像分类为多个类别中的一类
- ResNet
  - 非常深的网络（默认为152层）
- 两种使用模式
  - 全量学习 (从随机参数开始训练，需要大量数据，结果准确)
  - 迁移学习 (利用公开成熟的模型，替换最后的一层或几层全联通层，不需要很多数据也能训练)

图像分类

Sequence to Sequence（seq2seq)

- 输入一个序列并获得另一个序列作为输出。
- 编码器和解码器



seq2seq 的使用场景

- 机器翻译
  - 以一种语言输入一个句子，并预测该句子在另一种语言中的含义
- 文字摘要
  - 输入较长的单词串，并通过作为摘要的较短的单词串输出
- 语音转文字
  - 输入一段音频，通过转化输出相应的文字

DeepAR

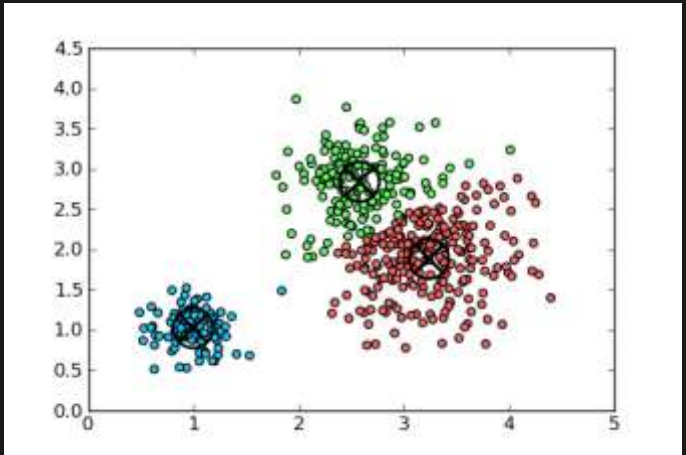
- 时间序列预测
- 亚马逊内部使用的算法
- 训练一组相关的时间序列，以获得更多的见解和更高的预测能力
- 最小化特征引擎
- 预测
  - 值（销量为 x）
  - 概率（出售金额在 x 和 y 之间的概率 z）

DeepAR 的使用场景

- 预测
  - 产品需求量
    - 供应链优化
  - 服务器负载
  - 网页请求

K-Means 聚类

- 将数据分成 K 个离散的群集
- 基于特征的相似性
- 群集内的成员尽可能相似，群集间的成员尽可能不同



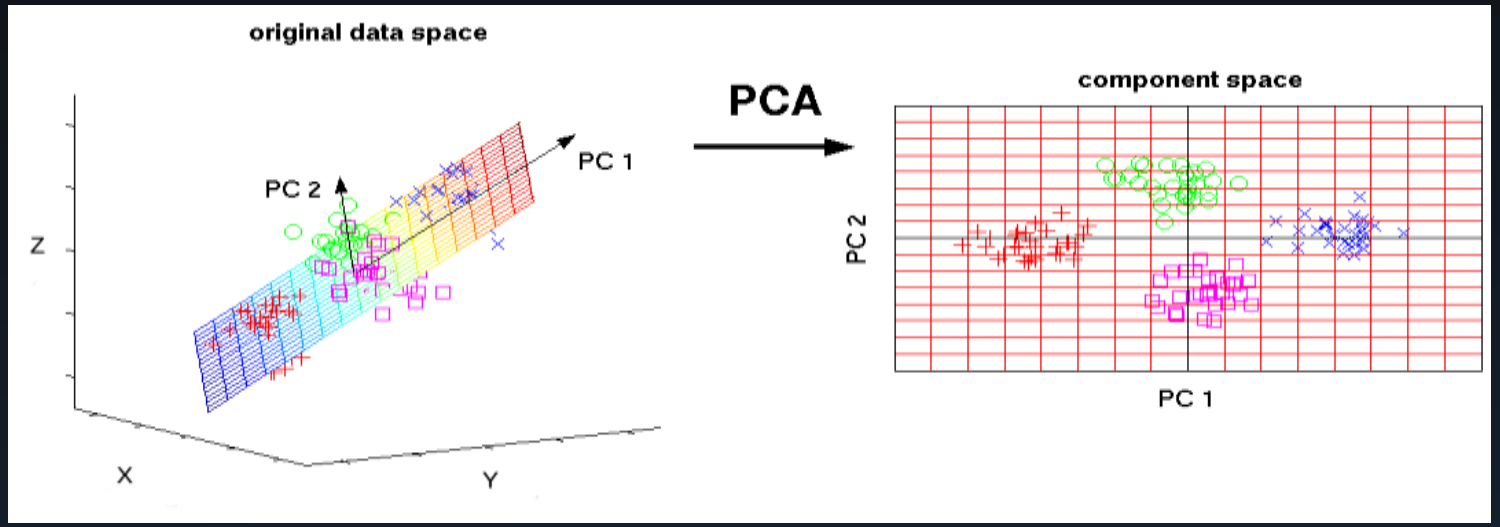
K-Means聚类的使用场景

- 搜索
- 客户划分
  - 根据购物记录进行划分
  - 根据网站、应用、平台上的用户行为进行划分
  - 根据喜好和行为特征建构用户描述信息

- 库存分类
  - 根据销售情况分组
  - 根据生产情况分组

主成分分析 (PCA)

- 数据降维（降低特征的数量）
- 将特征映射到具体的成分



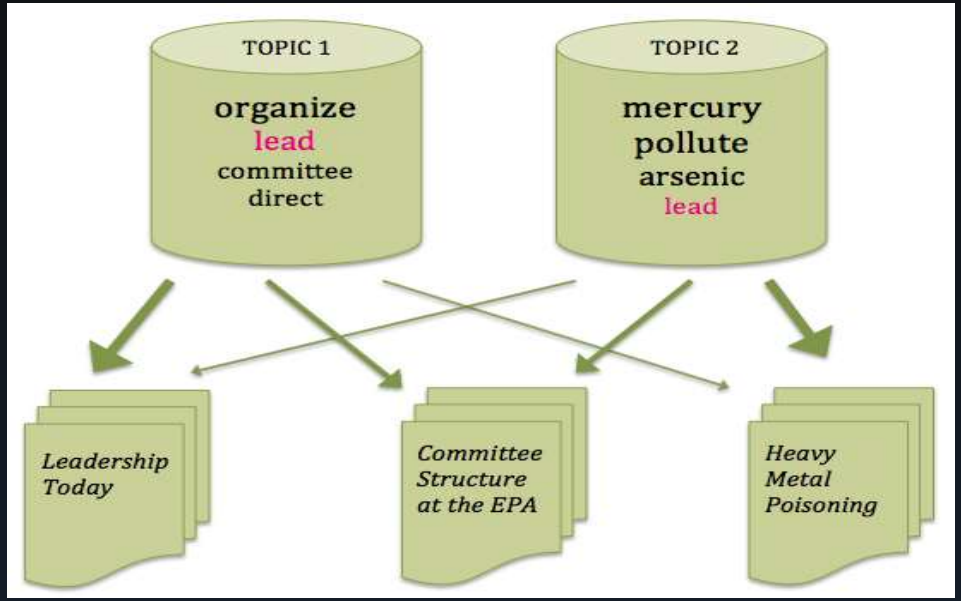
主成分分析 (PCA) 使用场景

- 数据压缩
- 图像处理
- 探究性数据分析
- 高维度数据的模式识别
- 金融，生物信息，心理学，数据挖掘

隐含狄利克雷分布 (LDA)

- 在文本语料库中，发现文档中的主题
  - 每次输入都是一份文档
  - 特征是每个单词是否存在（或出现个数）
  - 对文档的分类为该文档的主题
- 主题通过对每个文档中出现的单词的概率分布进行机器学习
- 每个文档最终被描述为一些主题的集合

隐含狄利克雷分布 (LDA)



https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Latent Dirichlet Allocation (LDA) 使用场景

- 根据相关性和相似性对文档进行分类和组织
- 文档摘要
- 根据含义对大规模文档进行情感分类, 如文本，图像，歌词

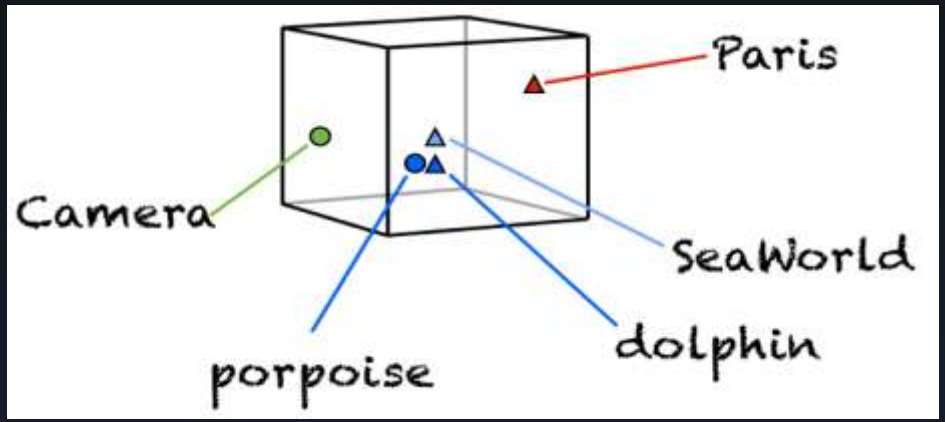
Neural Topic Modelling (NTM)

- 在文本语料库中，发现文档中的主题
- LDA vs NTM
  - 两种不同的算法会在同一数据集上产生不同的结果
  - NTM 通常具有较低的混淆度
  - LDA 在少数主题上训练非常快，但不像 NTM 那样扩展到更多主题

结论：如果使用场景中需要判断很多主题和更好的“合适度(fit)”，则使用 NTM，否则使用 LDA。

BlazingText

- 生成 Word2Vec
- 生成文档中各单词的矢量表示
- 获取其中的意义，单词和上下文之间的语义关系



BlazingText 的使用场景

- 用于自然语言处理 (NLP)
  - 情绪分析
    - 更好的了解客户
    - 确定产品趋势
  - 机器翻译
    - 为网站提供多语言支持
  - 命名实体识别
    - 从文字中获取组织与主要参与者信息

客户案例

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

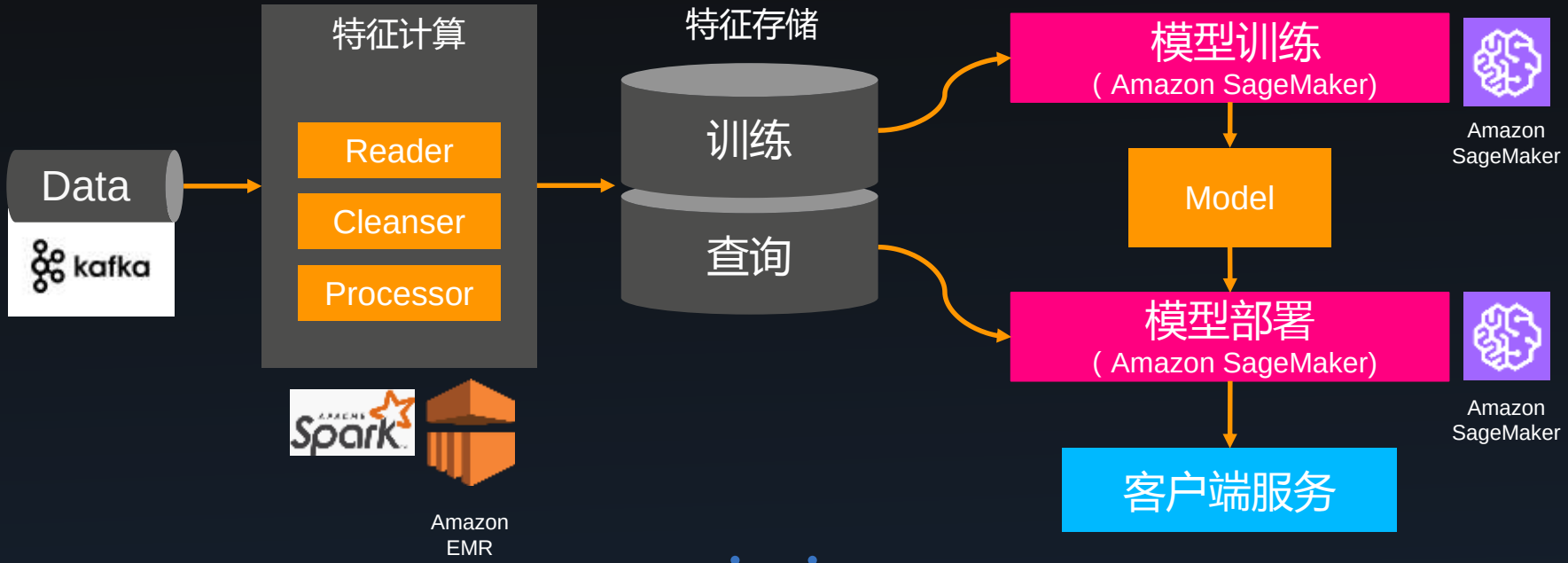
Amazon SageMaker: 初始用户

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

Intuit使用Amazon SageMaker获得的好处

From	To
需要临时设置 和管理notebook环境	使用Amazon SageMaker notebook轻松完成数据探索工作
有限的模型部署选择	通过虚拟化手段 达成极强的灵活性
团队之间需要争抢计算资源	自动扩展的模型部署环境

使用Amazon SageMaker在AWS上构建几乎实时的欺诈检测

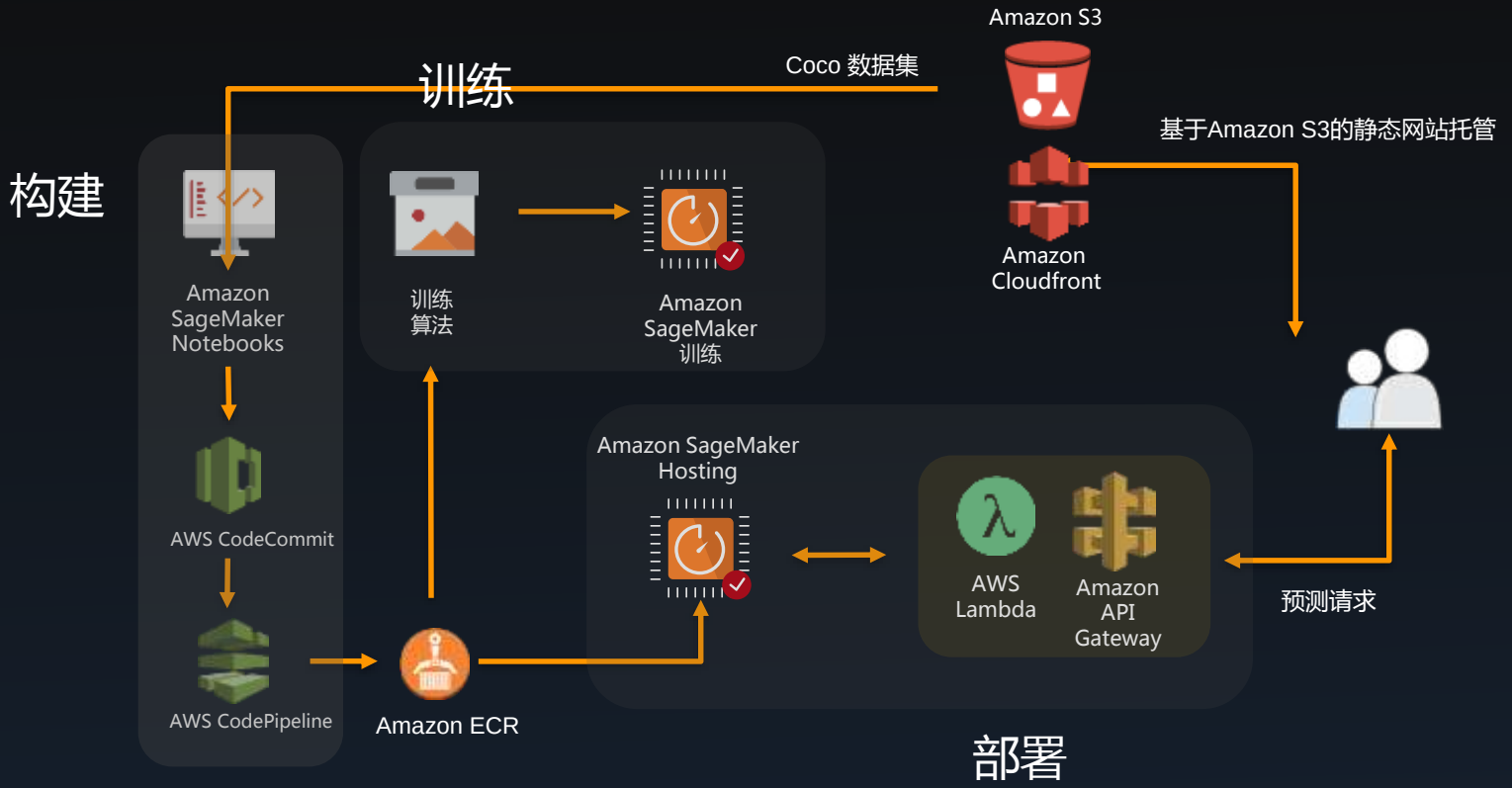


intuit

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

Amazon SageMaker 样例端到端架构: 风格转移



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

端到端的托管机器学习平台

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


Elastic
Map
Reduce
Lead scoring
Storage
Bulk/Batch
AMAZON WEB SERVICES
BUILD ON
REAL TIME
DATA
Amazon EC2
Model Repository
Developers
Task Management
Feature
Staging
Revenue

本PPT来自2018携程技术峰会
更多技术干货，请关注“携程技术中心”微信公众号

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