

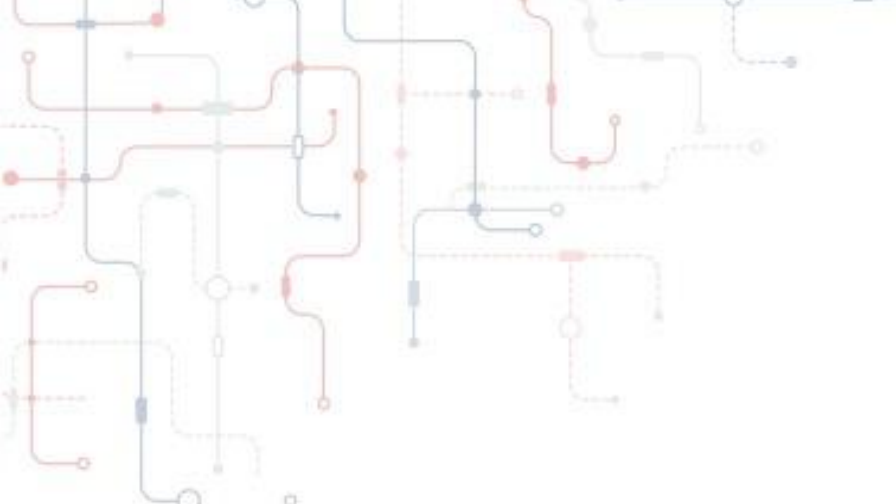
机器学习，化繁为简

——**Apache MXNet**框架以及**Amazon Sagemaker**
端到端机器学习平台介绍

AWS解决方案架构师 吴鹏程

pengchew@amazon.com





目录

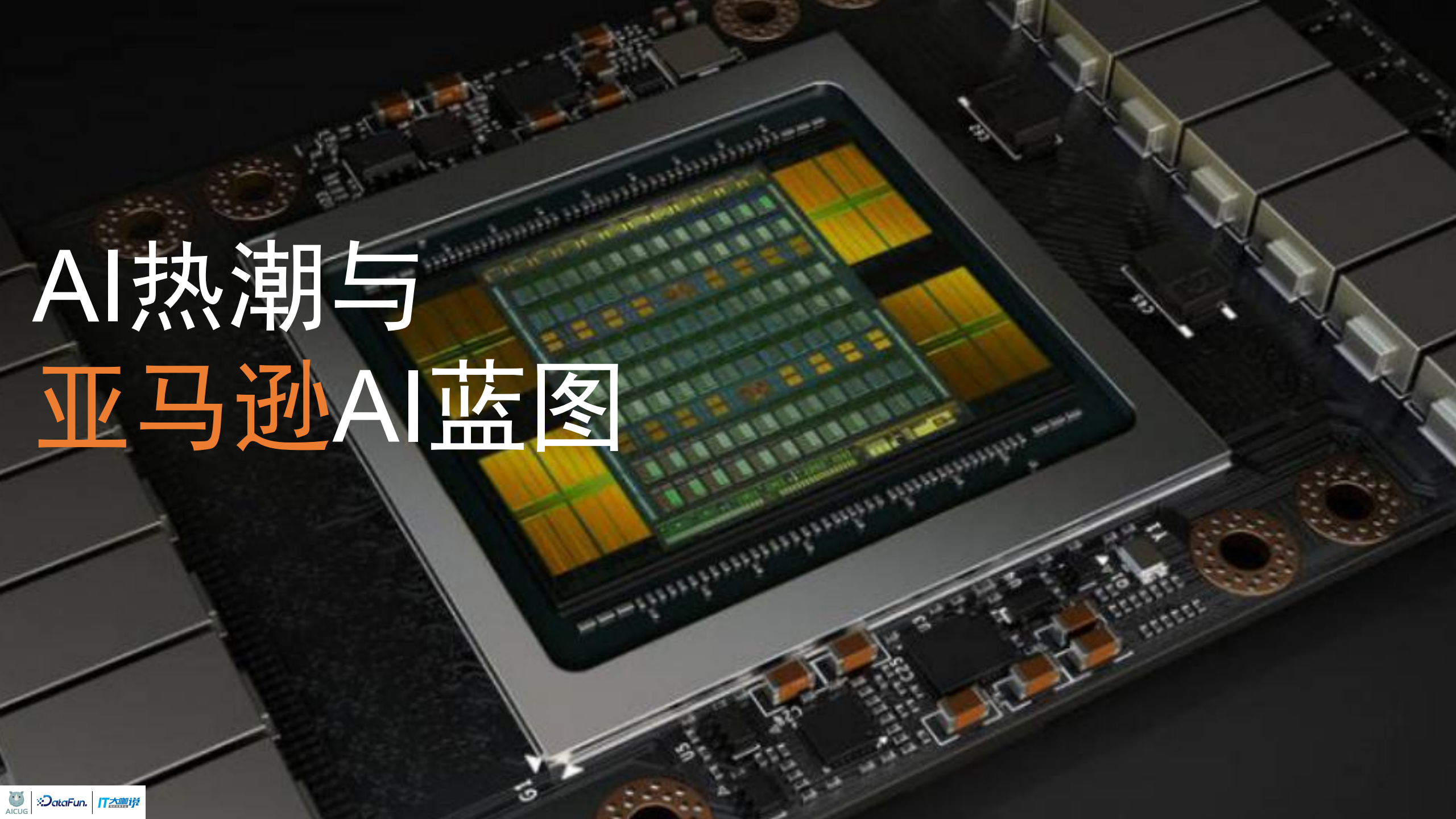
| AI热潮的与亚马逊AI蓝图

| 深度学习之MXNet以及Gluon

| Amazon Sagemaker机器学习平台简介

| 演示

| 总结



AI热潮与 亚马逊AI蓝图

算法 | 框架



数据爆发
式增长



计算能力
大大提升

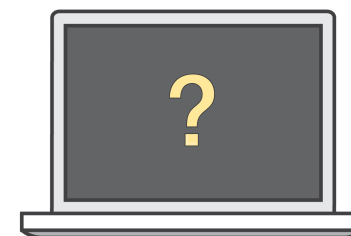
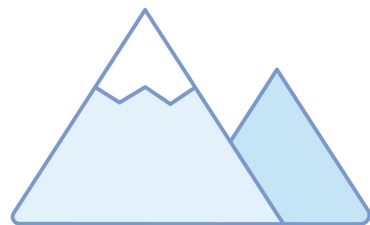
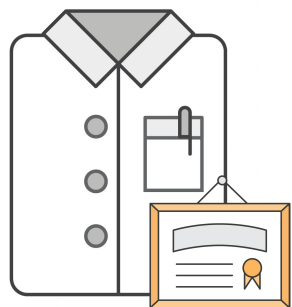


推理设备
的廉价化

*AI*的文艺复兴时代

AI如此多娇，但是桎梏因素很多

掌握人工智能专业知识的人才不多

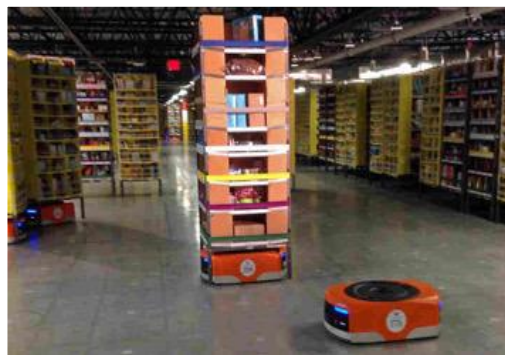


构建和扩展人工
智能的技术产品
有难度

在生产经营中部
署人工智能应用
费时且昂贵

缺乏成本低，易使
用，可扩展的人工智
能产品和服务

亚马逊的知名AI产品



人工智能与AWS生态系统

» 平台/垂直行业应用

» 混合部署、迁移和应用服务

» 安全服务

» 基础云服务



服务支持

AWS专注于把AI的知识和功能提供给所有用户

易用性

视觉
Rekognition Image
Rekognition Video



语音
Polly
Transcribe



语言
Lex Translate
Comprehend



Amazon
Machine
Learning



APACHE
Spark™ MLlib

Amazon
SageMaker



Mechanical
Turk

amazon mechanical turk

AWS
DeepLens



PYTORCH

AWS Deep Learning AMLs

GPU
(P3)

CPU

FPGA

容器化

无服务器化

边缘设备

更多服务
敬请期待

定制化



很多企业在尝试机器学习



缺少
ML 专家



无法释放商业潜力

AMAZON ML LAB
提供急需的机器学习经验



充分利用亚马逊机器学习专家在几十年间积累的经验、形成的技术，如 **AMAZON ECHO**, **AMAZON ALEXA**, **PRIME AIR** 和 **AMAZON GO**



头脑风暴

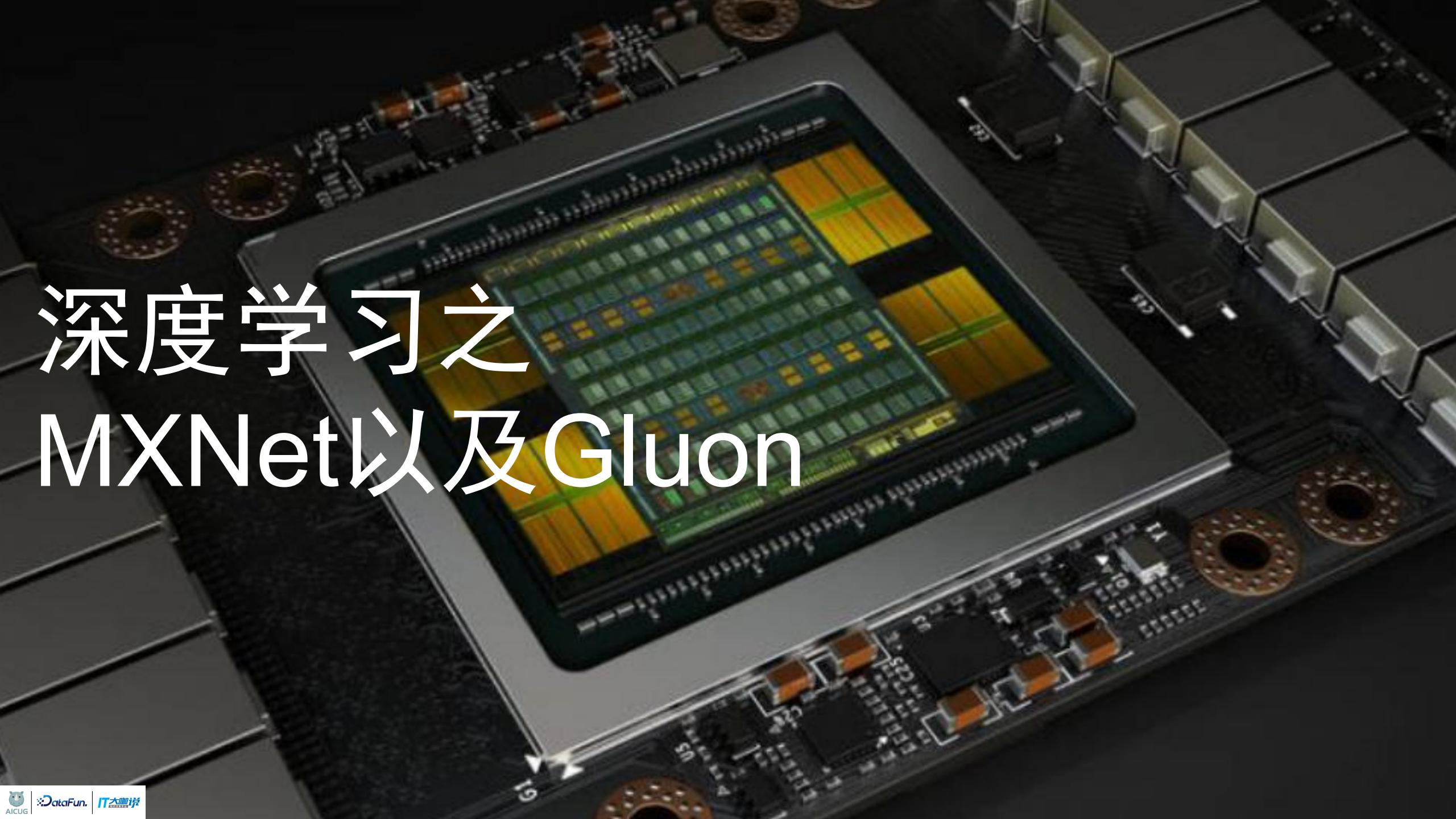


建模

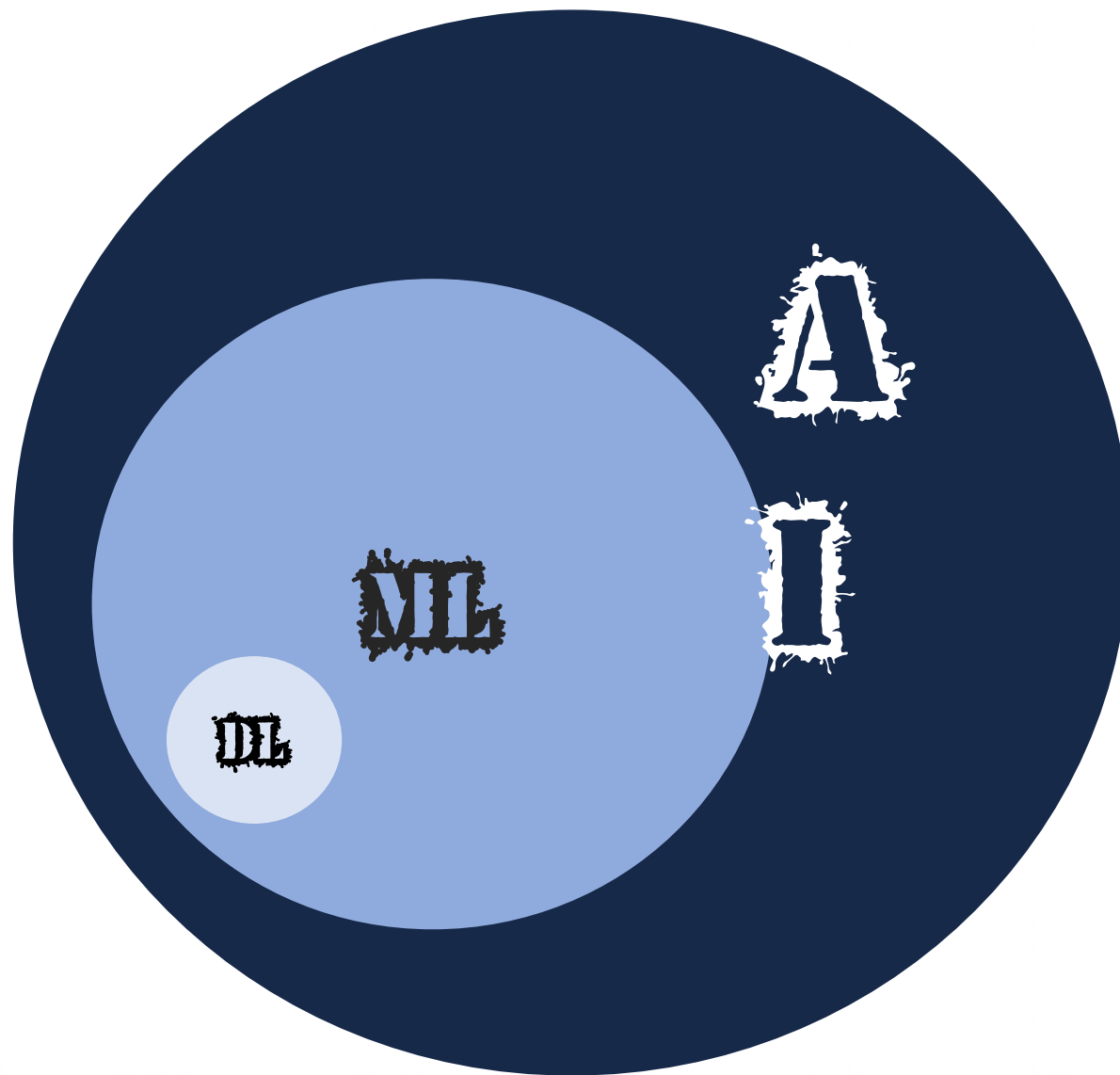


传授

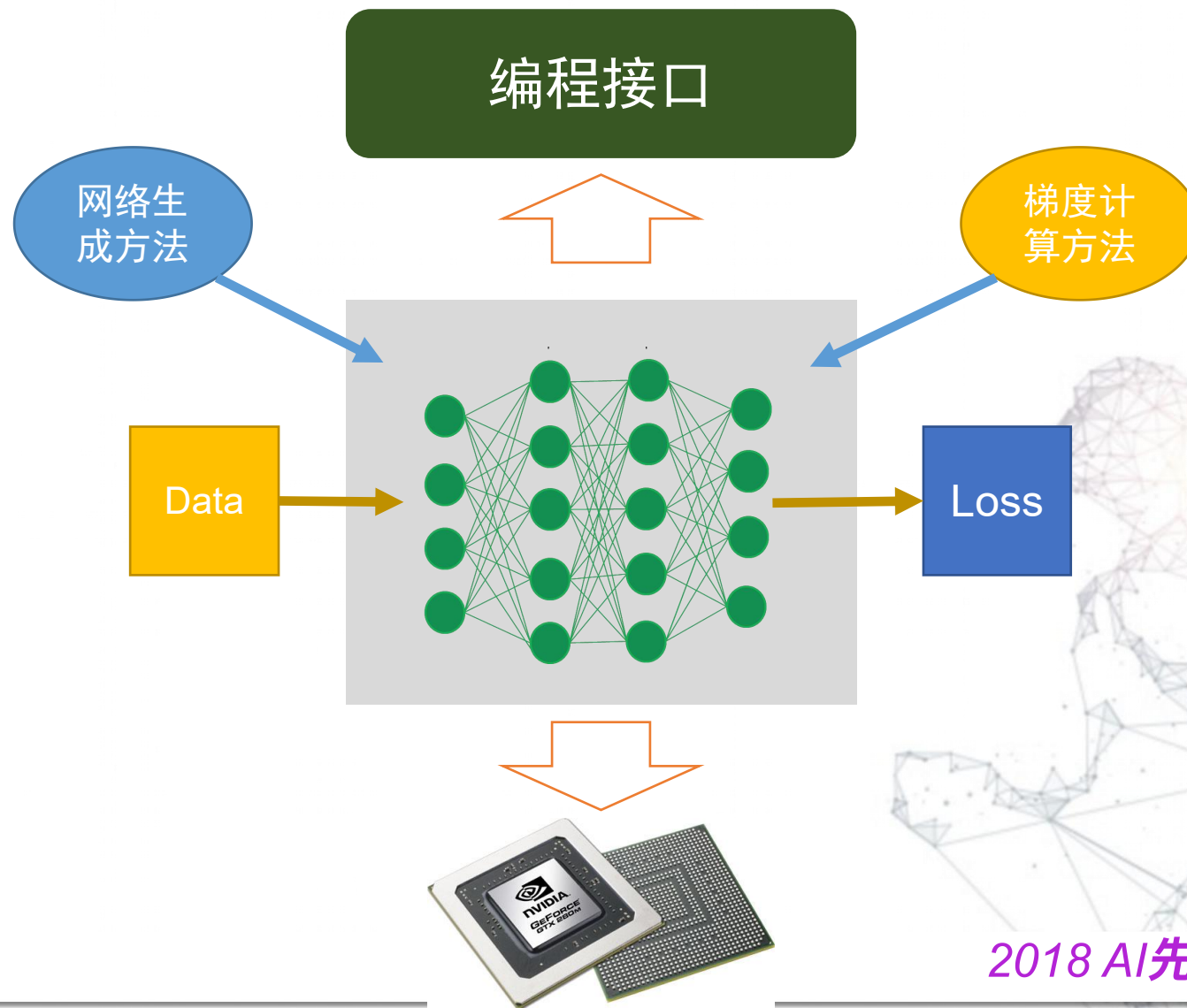
2018 AI先行者大会



深度学习之 MXNet以及Gluon



- 张量表达
- 梯度计算
- 网络定义
- 硬件支持
- 编程接口

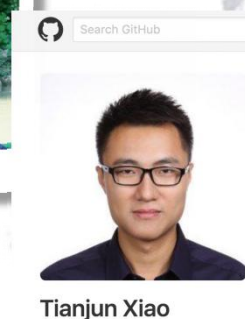
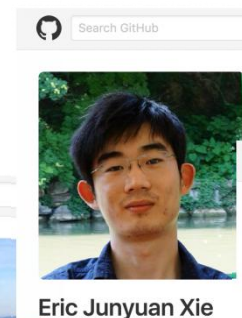
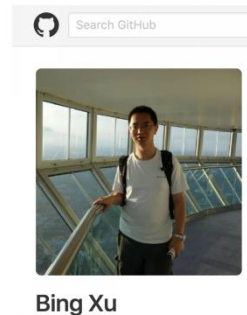
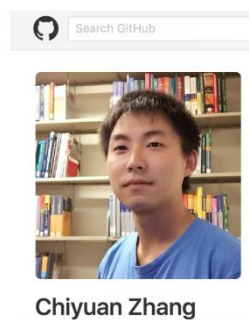
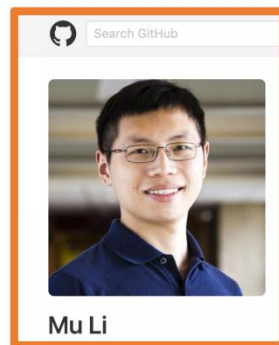


MXNet 背景介绍

MXNet诞生于卡耐基-梅隆大学，也在这里发展壮大，它已经成为我所见过的最具可扩展能力的深度学习框架，同时也是计算机科学领域辉煌成果的一种典型体现。多个不同学科在这里交汇并共同碰撞出火花：我们可以想象将线性代数引入大规模分布式计算，从而构建起一种全新的深度学习实现方案。我们对于Amazon投资支持MXNet的决定感到兴趣，并期待着看到MXNet能够走上新的高峰。

——Andrew Moore
卡耐基-梅隆大学计算机科学系主任

<https://github.com/dmlc/mxnet>



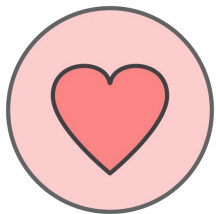
2018 AI先行者大会

支持MXNet的组织

MXNet 是亚马逊选择的深度学习库。它拥有类似于 Theano 和 TensorFlow 的数据流图，为多 GPU 配置提供了良好的配置，有着类似于 Lasagne 和 Blocks 更高级别的模型构建块，并且可以在你可以想象的任何硬件上运行（包括手机）。支持Python、Clojure、R、Julia、Perl、C++、Scala、Matlab，和 Javascript 的接口。

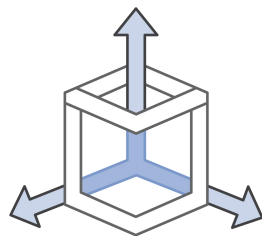


APACHE MXNET的优点



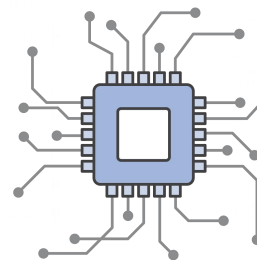
编程友好

简单,多语言支持



可移植性

CPU, GPU,
ARM, IoT 设备



性能

跨GPU性能几乎线性增长



开放

成为Apache孵化项目



AWS 优化

在AWS上有专门优化

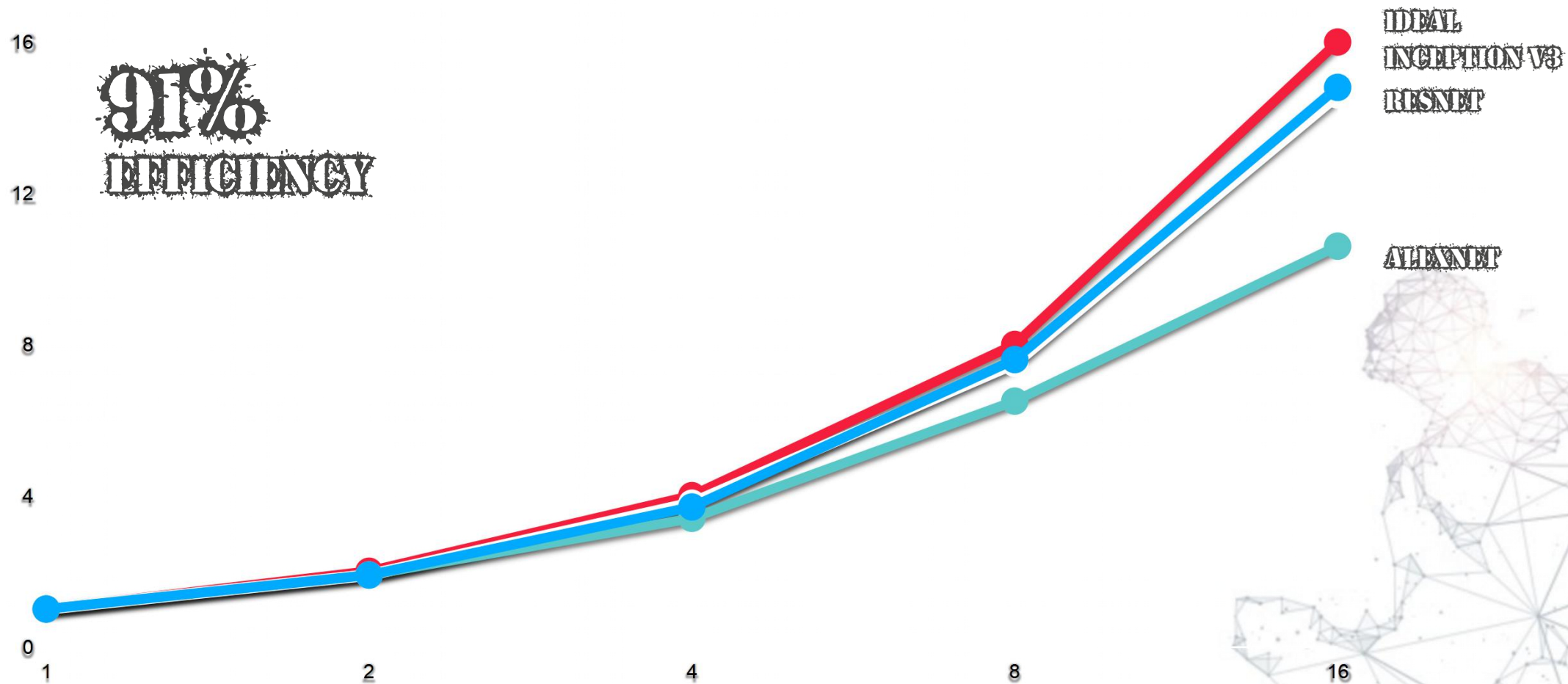
2018 AI先行者大会

深度学习框架的比较

	APACHE KNNET	TENSORFLOW	COGNITIVE TOOLKIT
行业所有者	N/A □ APACHE 开源项目	GOOGLE	MICROSOFT
编程方法	命令式和声明式	声明式	声明式
语言支持	R, PYTHON, SCALA, JULIA, C++, JAVASCRIPT, GO, MATLAB AND MORE.	PYTHON, C++, EXPERIMENTAL GO AND JAVA	PYTHON, C++, BRAINSRIPT.
代码长度 † ATHANNET (PYTHON)	44 SLOC	107 SLOC (使用 TERSLIM)	214 SLOC
内存使用 (TERSLIM)	2.6GB	7.2GB	N/A

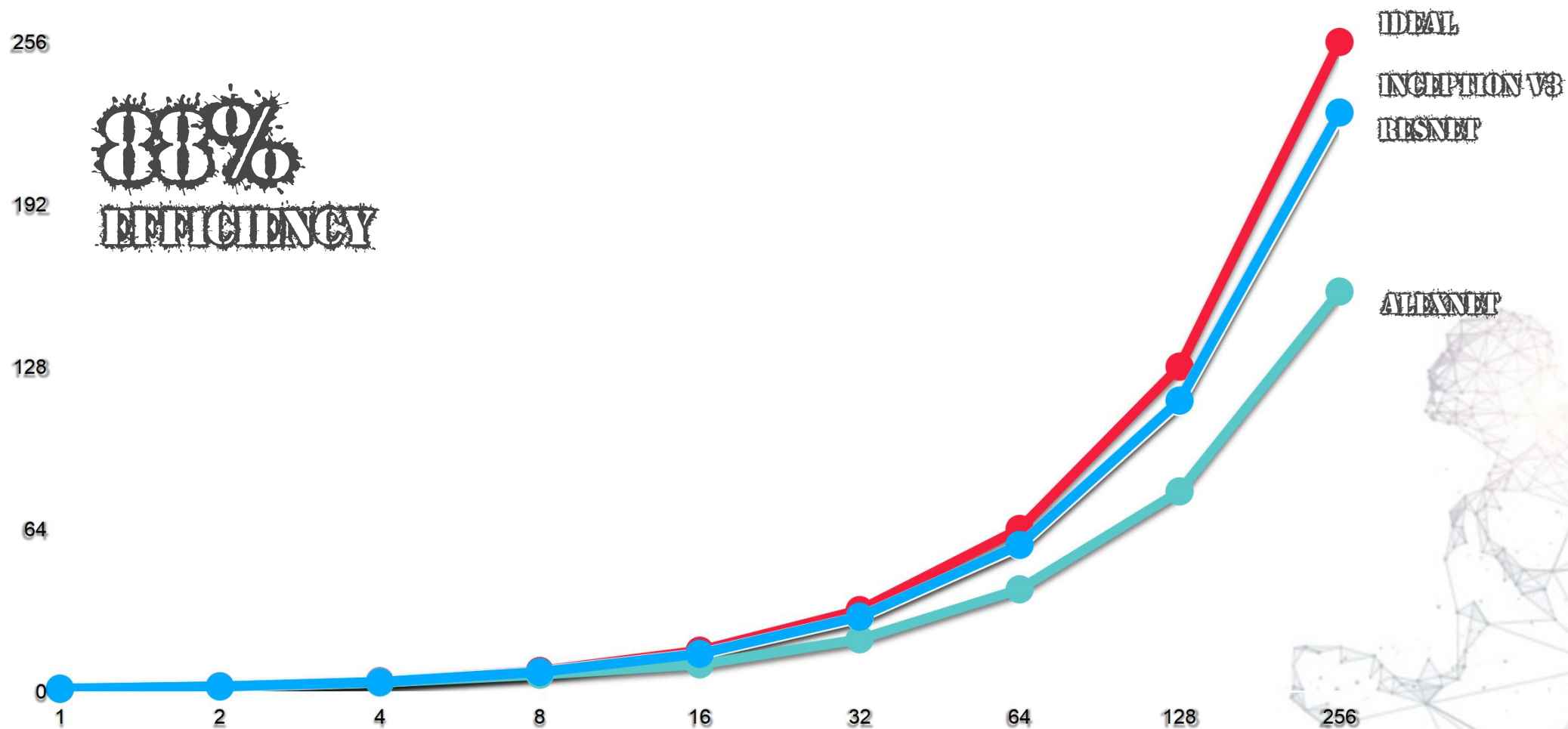
*sloc – source lines of code

MXNET跨GPU实现分布式训练



2018 AI先行者大会

MXNET跨实例实现分布式训练



Apache MXNet: Amazon内部的使用

- 应用研究
- 核心研究
- Alexa
- 需求推理
- 风险分析
- 搜索
- 推荐系统
- Ai服务 | Rek, Lex, Polly
- 问答系统
- 供应链优化
- 广告
- 机器翻译
- 视频内容分析
- 机器人控制
- 大量计算机视觉处理..
- 大量自然语言处理..

基于MXnet构建卷积神经网络

```
• data = mx.symbol.Variable('data')
• # 第一个卷积层
• conv1 = mx.sym.Convolution(data=data, kernel=(5,5), num_filter=20)
• tanh1 = mx.sym.Activation(data=conv1, act_type="relu")
• pool1 = mx.sym.Pooling(data=tanh1, pool_type="max", kernel=(2,2), stride=(2,2))

• #第二个卷积层
• conv2 = mx.sym.Convolution(data=pool1, kernel=(5,5), num_filter=50)
• tanh2 = mx.sym.Activation(data=conv2, act_type="relu")
• pool2 = mx.sym.Pooling(data=tanh2, pool_type="max", kernel=(2,2), stride=(2,2))

• # 第一个全连接层
• flatten = mx.sym.Flatten(data=pool2)
• fc1 = mx.symbol.FullyConnected(data=flatten, num_hidden=500)
• tanh3 = mx.sym.Activation(data=fc1, act_type="relu")

• # 第二个全连接层
• fc2 = mx.sym.FullyConnected(data=tanh3, num_hidden=10)
• # softmax损失层
• lenet = mx.sym.SoftmaxOutput(data=fc2, name='softmax')
```


Gluon 简介



GLUON

简单，易于理解的代码

Gluon是AWS和微软联合发布的新的开源深度学习接口，可帮助开发人员更轻松、更快速地构建机器学习模型，而不牺牲任何性能。Gluon 采用一系列预先构建并经过优化的神经网络组件，为定义机器学习模型提供清晰、简洁的 **API**。开发人员可以像使用任何其他数据结构那样定义和操作机器学习模型，也可以基于它快速构建原型和利用动态神经网络图实现全新模型架构，而又不减缓训练进度。Gluon 现已在 [Apache MXNet](#) 中可用，日后将“进驻”Microsoft Cognitive Toolkit 及其他架构。

可以使用**简单，清晰，简洁**的代码定义神经网络

即插即用神经网络构建块，包括预定义层**PREDEFINED LAYERS**，优化器**OPTIMIZERS**和初始化器**INITIALIZERS**

弹性、命令式编程

消除神经网络模型定义的桎梏，并将**模型与训练算法结合**在一起
直观，**易于调试**，熟悉的代码

动态图

在训练过程中，算法可以在定义和训练期间**动态调整网络**，以解决数据馈送大小可变的高级场景
动态图是在自然语言处理（**NLP**）的重要创新领域。**GLUON同时支持**灵活的动态图和高效的静态图，既保留动态图的易用性，也具有静态图的高性能，

高性能

以前，对编程友好的 **API** 或动态图都会减缓训练速度，**GLUON不会以牺牲训练速度为代价**
当需要从原型设计转向生产时，可以轻松缓存神经网络以**实现高性能和减少内存占用**

基于Gluon构建卷积神经网络

- # code from <https://github.com/zackchase/mxnet-the-straight-dope>

```
num_fc = 512
net = gluon.nn.Sequential()
with net.name_scope():
    net.add(gluon.nn.Conv2D(channels=20, kernel_size=5, activation='relu'))
    net.add(gluon.nn.MaxPool2D(pool_size=2, strides=2))
    net.add(gluon.nn.Conv2D(channels=50, kernel_size=5, activation='relu'))
    net.add(gluon.nn.MaxPool2D(pool_size=2, strides=2))
    # The Flatten layer collapses all axis, except the first one, into one axis.
    net.add(gluon.nn.Flatten())
    net.add(gluon.nn.Dense(num_fc, activation="relu"))
    net.add(gluon.nn.Dense(num_outputs))
```

- # Handy class available here
- https://github.com/sunilmallya/dl-twitch-series/blob/master/cnn_mnist_gluon_simplified.ipynb

李沐大神的MXNet公开课



视频中心

首页 产品 市场活动 成功案例 教学 在线研讨会

在线研讨会 Webinars



MXNet公开课 | 第十六课：词向量 (word2vec)



MXNet公开课 | 第十五课：门控循环单元 (GRU)



MXNet公开课 | 第十四课：实现、训练和应用循环神经网络



MXNet公开课 | 第十三课：正向传播、反向传播和通过时间反向传播



MXNet公开课 | 第十二课：循环神经网络



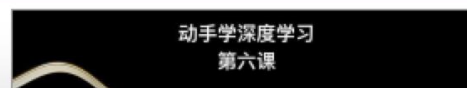
MXNet公开课 | 第十一课：样式迁移



MXNet公开课 | 第十课：语义分割 (全连接卷积网络)



MXNet公开课 | 第九课：物体检测 · 再续

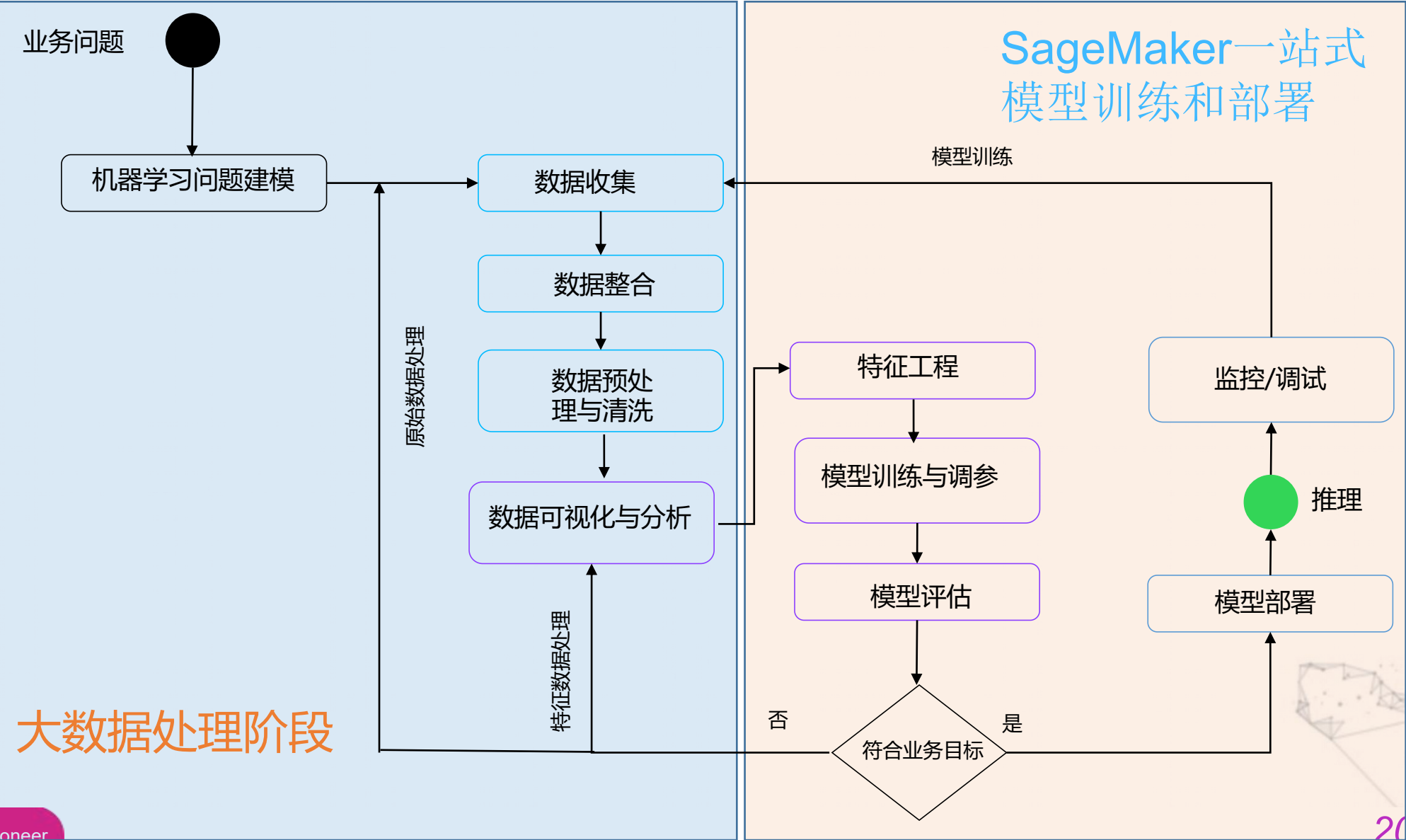


更多视频参见：<http://aws.amazon.com/bokecc.com/news/56/3.html>



Amazon SageMaker 机器学习平台简介

机器学习典型工作流程



训练

- 部署与管理notebook环境
- 部署与管理训练训练集群
- 构建数据接口
- 对大数据集进行扩展性算法设计
- 分布式机器学习训练
- 模型数据安全设计

部署

- 部署与管理模型推理集群
- 管理与部署推理API
- 监控并调试模型推理
- 模型版本化管理与性能监测
- 新版本模型自动部署与A/B测试

SageMaker——端到端机器学习模型构建，训练与部署平台

部署

全托管的线上推理
集群

一键部署机器学习
模型

预配置Notebook
计算实例



**Amazon
SageMaker**

构建模型

深度优化的机器学习
算法

训练

一键启动自定义的
机器学习、深度学习
算法

PYTORCH

TensorFlow

mxnet

Caffe2

Chainer

K

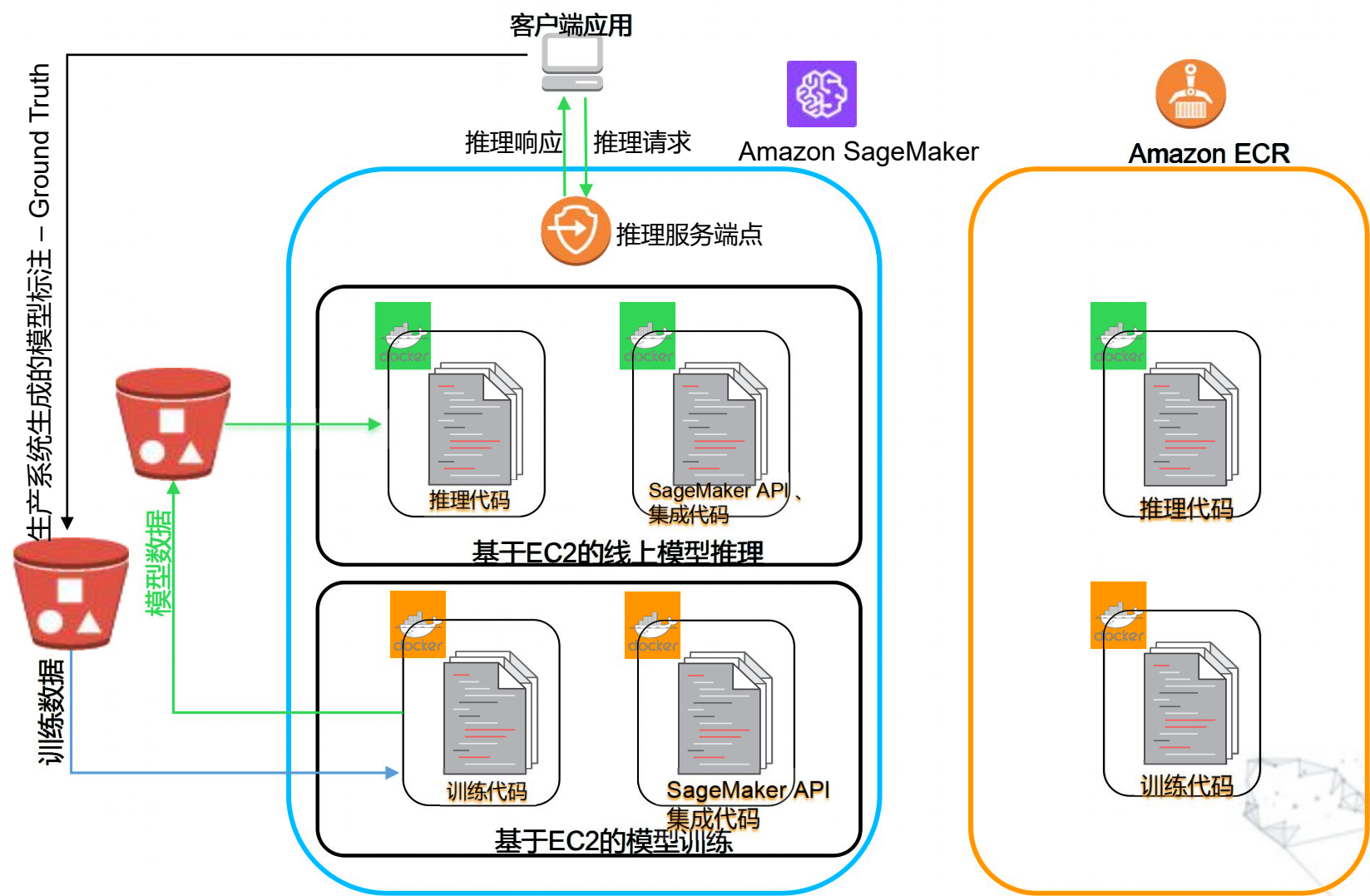
GLUON

torch

The Microsoft
Cognitive Toolkit

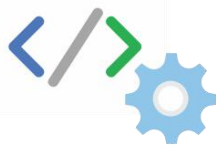
2018 AI先行者大会

Sagemaker的工作原理



1

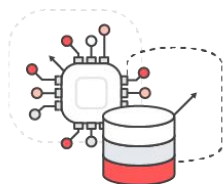
零配置进行数据分析与探索



NoteBook计算环境



编写Notebook

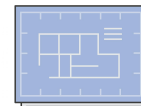


访问数据湖



直接访问AWS数据库服务

EBS



只需要添加数据

- 推荐与个性化算法
- 欺诈分析
- 推理
- 图片分类
- 客户趋势分析
- 市场与商品推广
- 日志处理与分析
- 语音文字转化
- 更多...

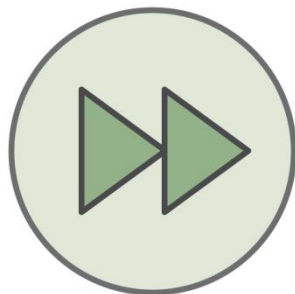
2 Amazon SageMaker: 10 倍优化的算法



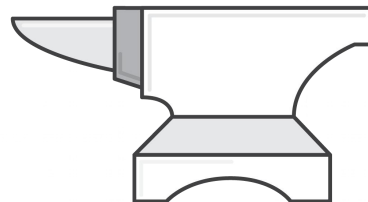
算法



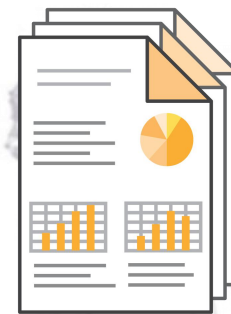
流式数据注入



更快的训练过程

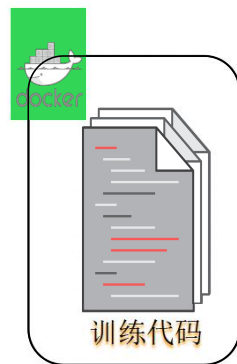


基于大数据量的可靠训练模型与算法



10多种经典机器学习算法

算法和代码的灵活性



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

亚马逊提供的算法



利用您已有的算法脚本



IM Estimators
Apache Spark



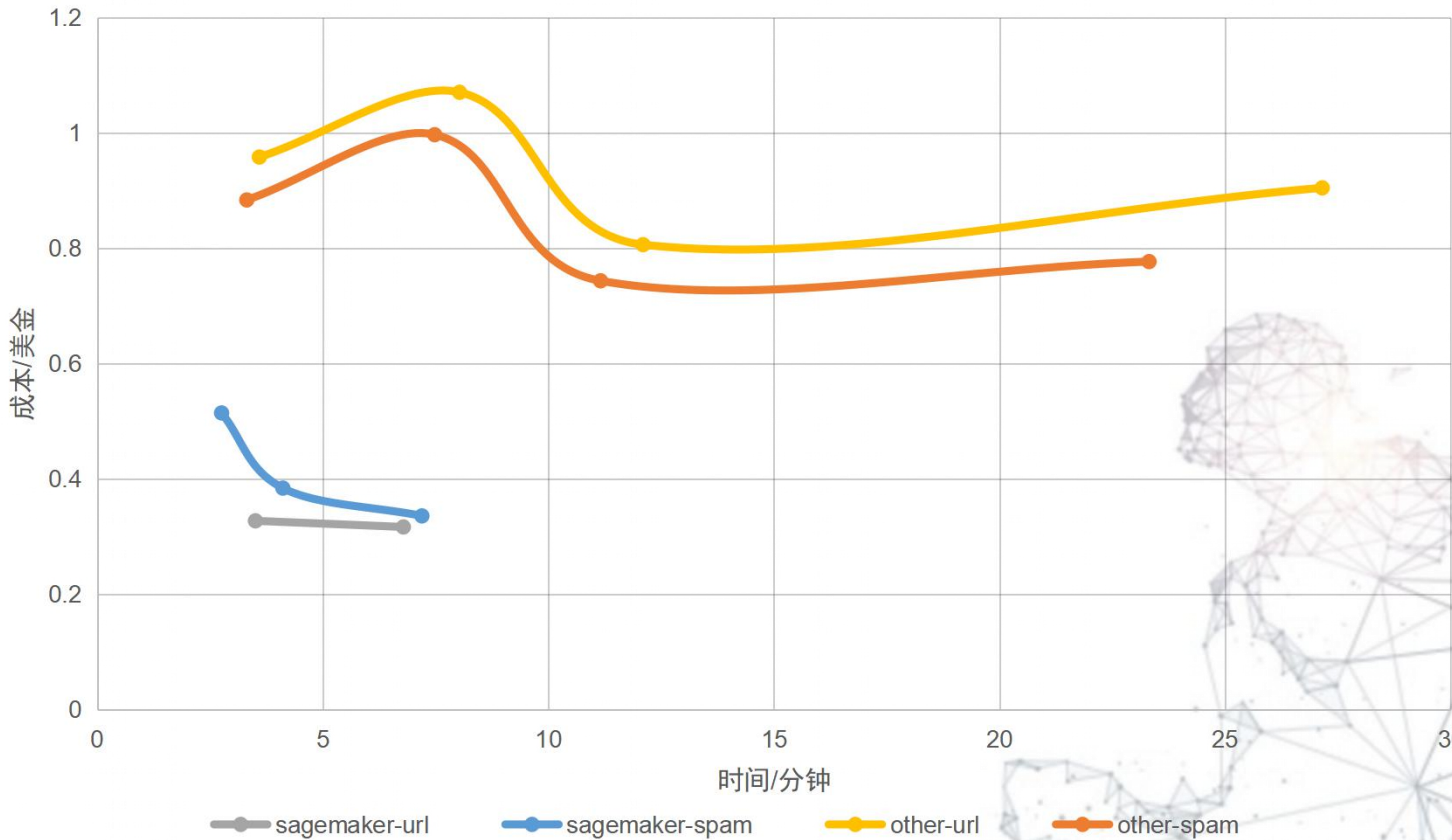
利用Docker封装您定制化的算法

Linear Learner — 线性模型

Regression (mean squared error)	
SageMaker	其他
1.02	1.00
1.09	1.02
0.332	0.108
0.086	0.120
83.3	0.15

Classification (F1 Score)	
SageMaker	其他
0.980	0.981
0.870	0.980
0.997	0.997
0.978	0.960
0.914	0.859
0.470	0.752
0.903	0.908
0.508	0.508

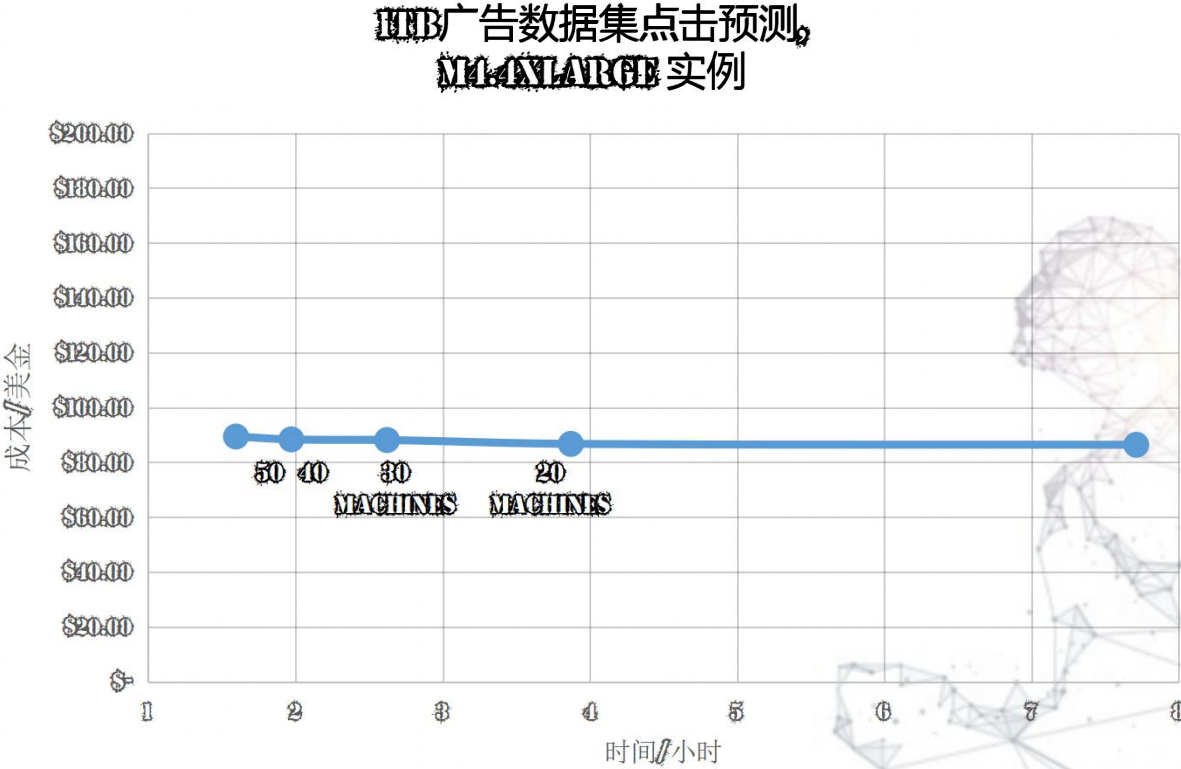
30GB 数据集的Web URL的线性分类



Factorization Machines — 因子分解机算法

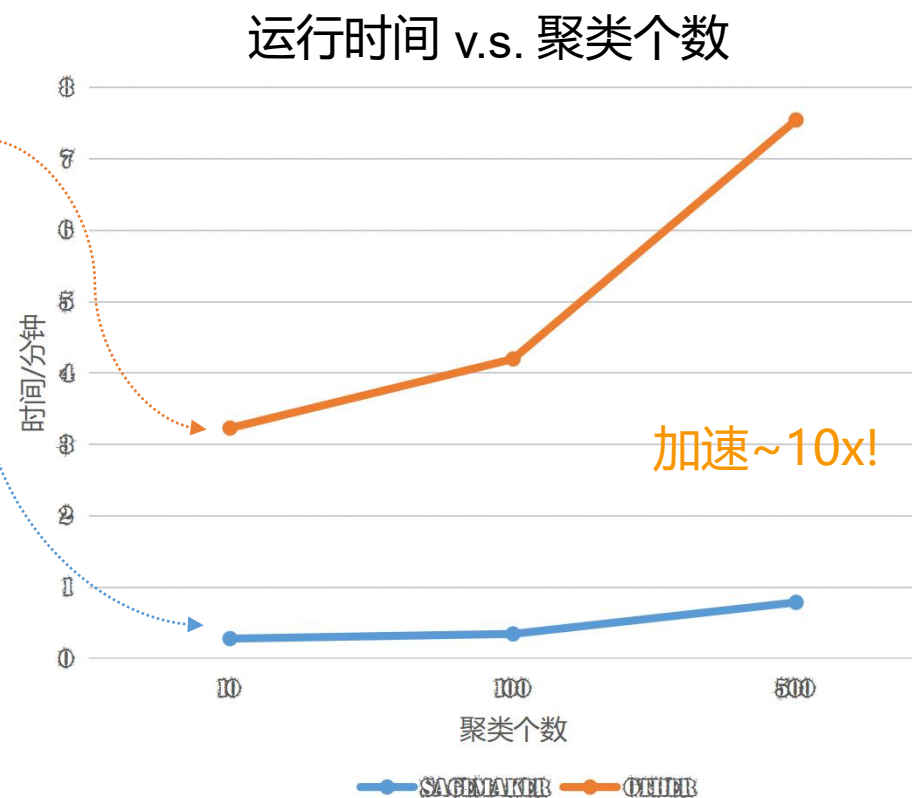
$$\tilde{y} = w_0 + \langle w_1, x \rangle + \sum_{i,j>i} x_i x_j \cdot \langle v_i, v_j \rangle$$

	Log_loss	F1 Score	Seconds
SageMaker	0.494	0.277	820
Other (10 lter)	0.516	0.190	650
Other (20 lter)	0.507	0.254	1300
Other (50 lter)	0.481	0.313	3250



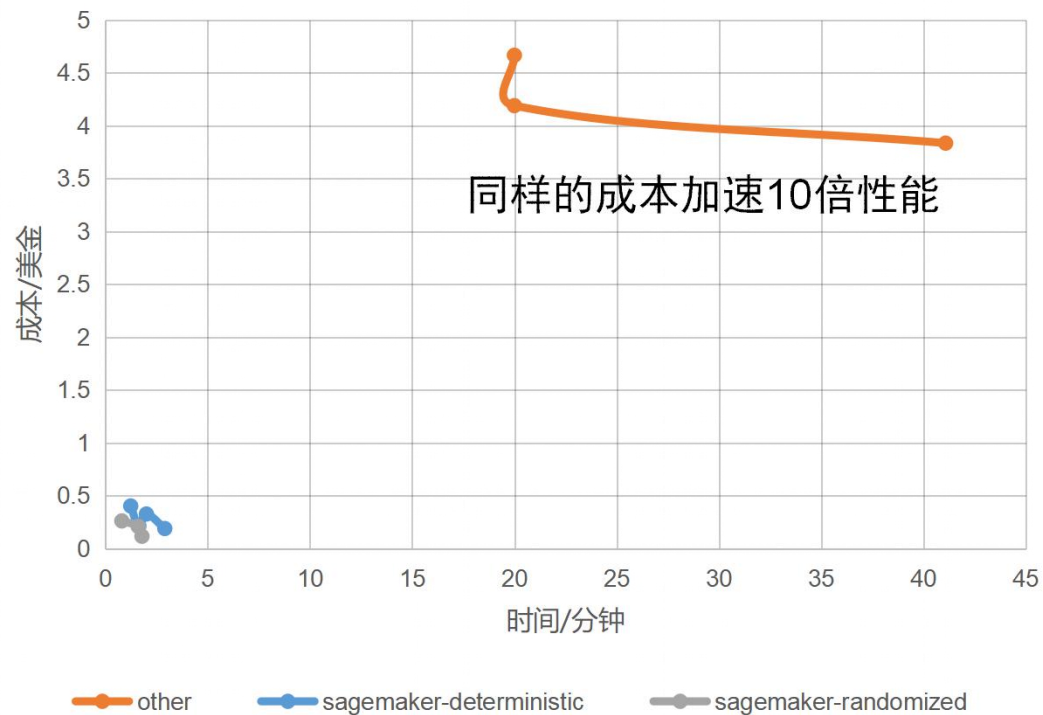
K-Means 聚类

	k	SageMaker	其他
Text 1.2GB	10	1.18E3	1.18E3
	100	1.00E3	9.77E2
	500	9.18E2	9.03E2
Images 9GB	10	3.29E2	3.28E2
	100	2.72E2	2.71E2
	500	2.17E2	Failed
Videos 27GB	10	2.19E2	2.18E2
	100	2.03E2	2.02E2
	500	1.86E2	1.85E2
Advertising 127GB	10	1.72E7	Failed
	100	1.30E7	Failed
	500	1.03E7	Failed
Synthetic 1100GB	10	3.81E7	Failed
	100	3.51E7	Failed
	500	2.81E7	Failed

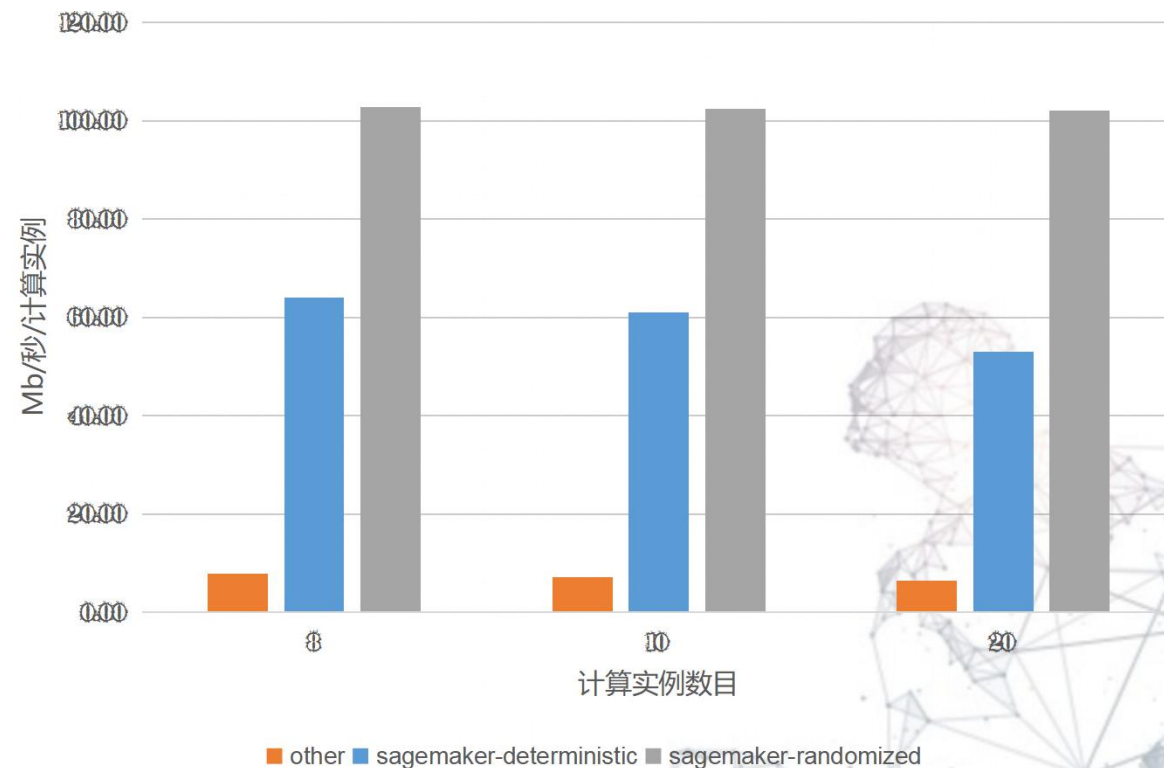


PCA — 主成分分析

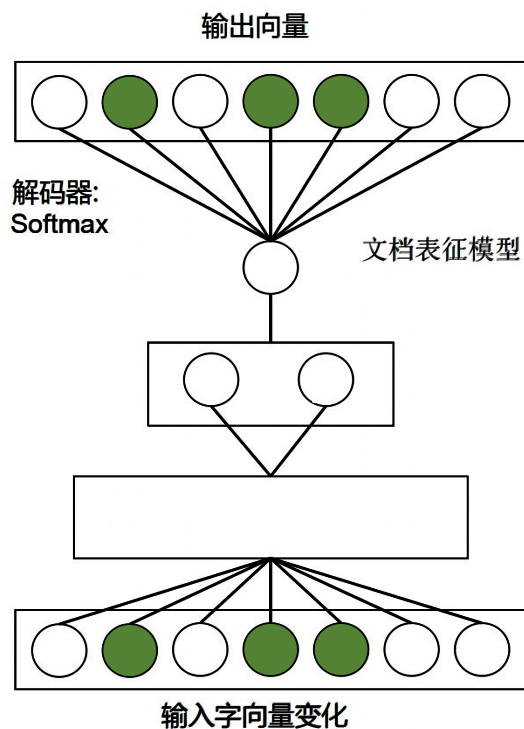
成本 v.s. 时间



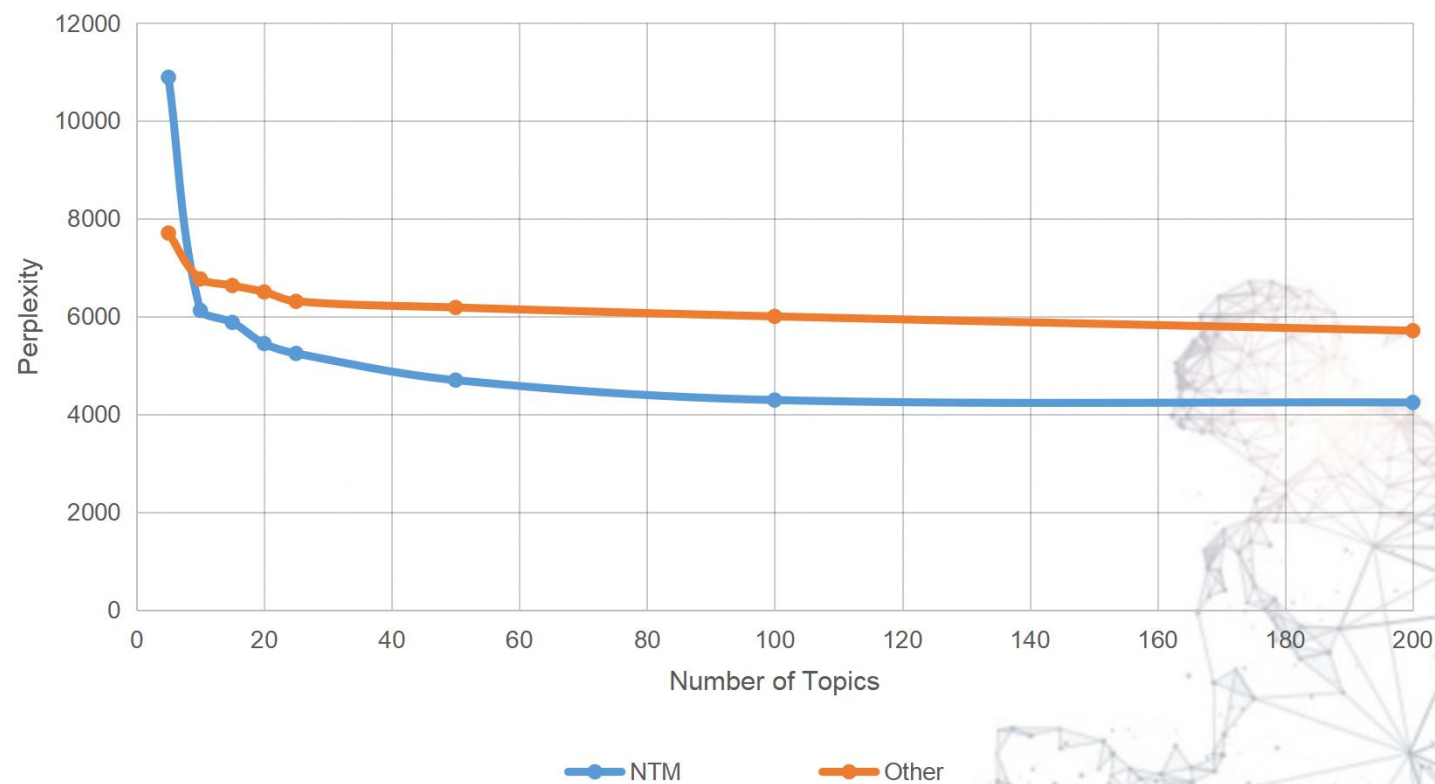
吞吐量与扩展性



Neural Topic Modeling — 基于神经网络的文档建模



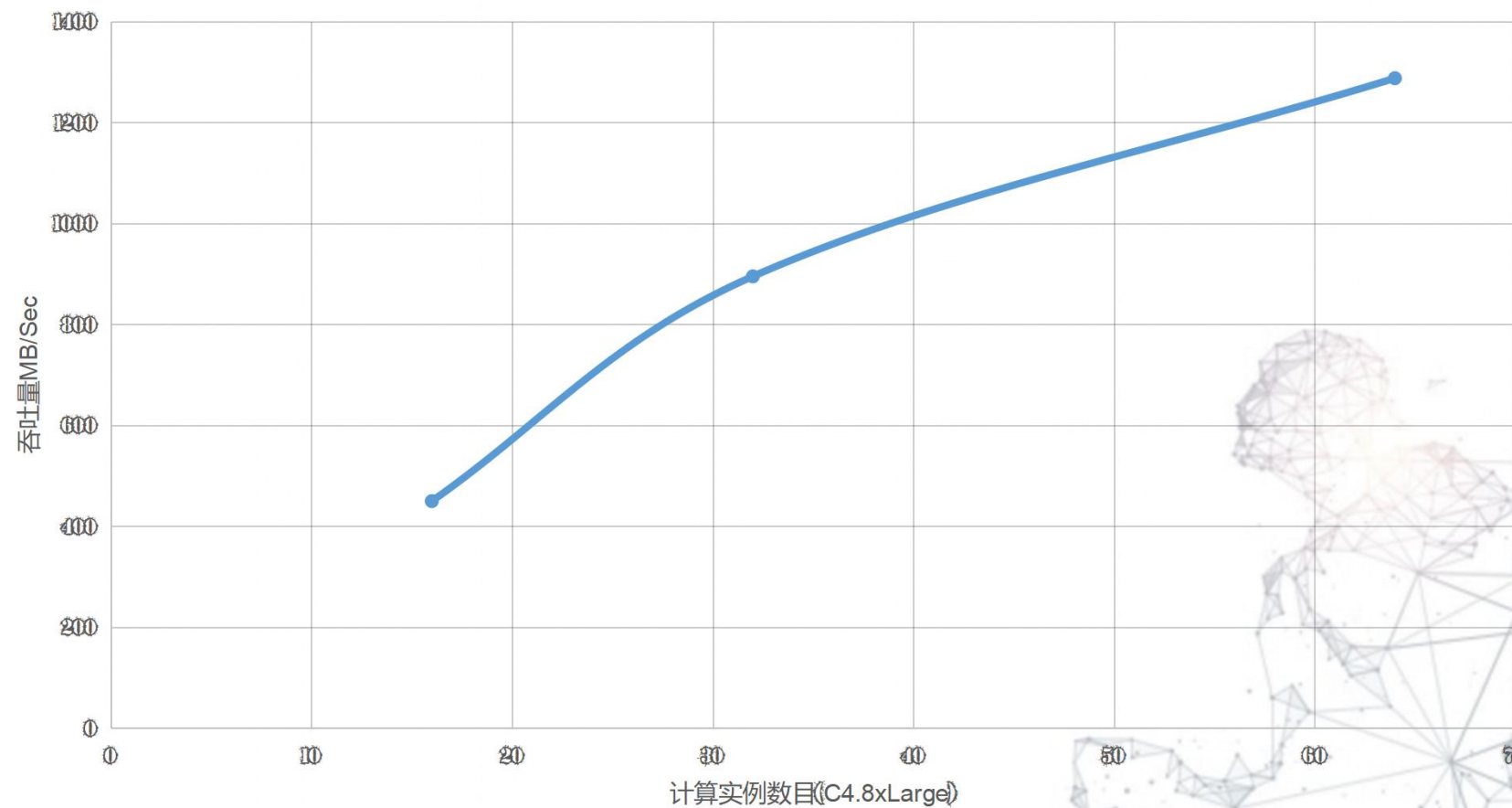
困惑度 (PERPLEXITY) VS. 主题节数 (NUMBER OF TOPICS)
(~200K 文档, ~100K 词库)



Boosted Decision Trees

dmlc
XGBoost

吞吐量 vs. 计算实例



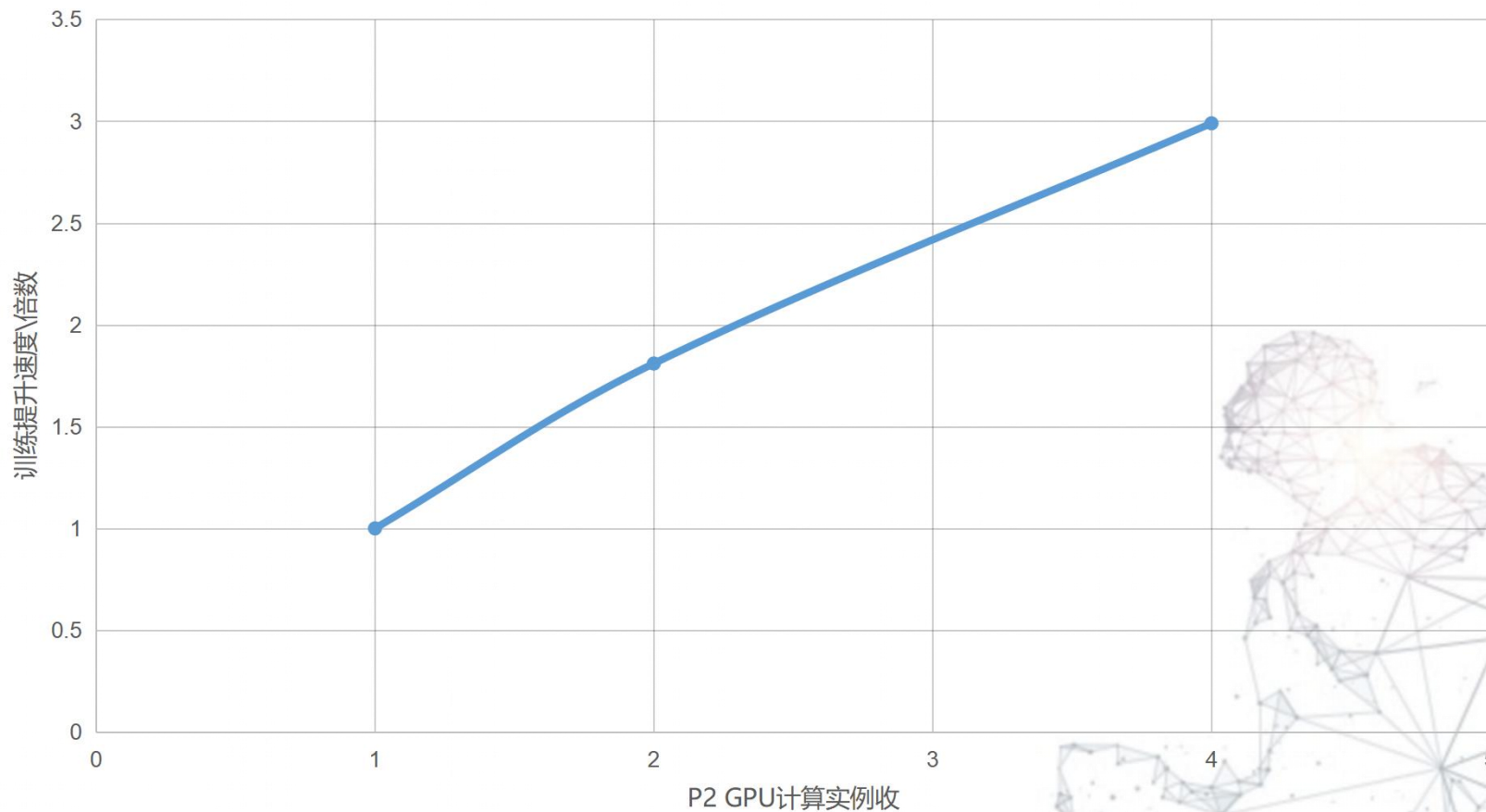
2018 AI先行者大会

Image Classification — 基于神经网络的图片分类

基于MXNet的ResNet模型

用户可以利用已有的训练模型实现迭代训练(transfer learning)

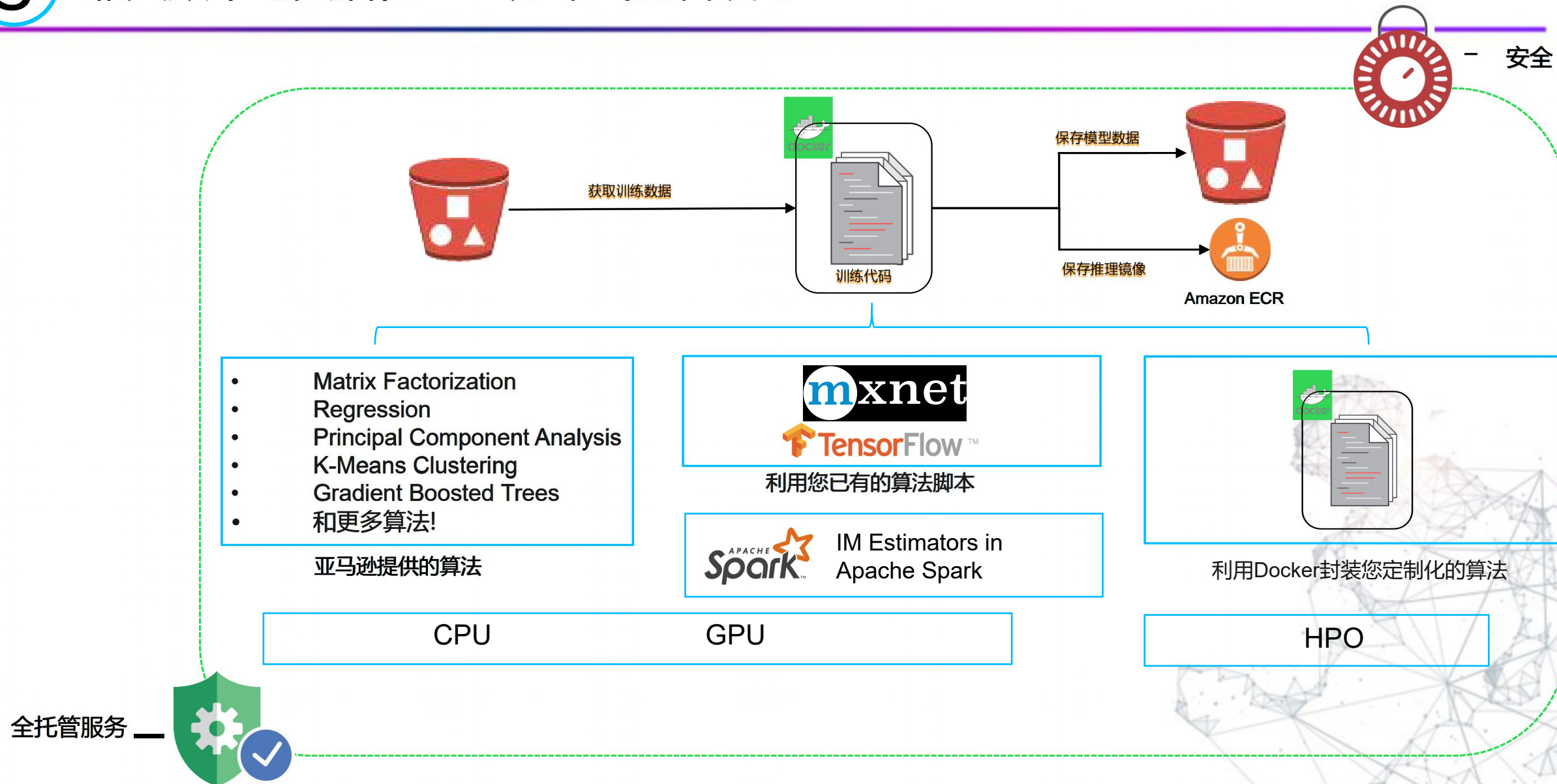
水平扩展提升训练速度



2018 AI先行者大会

3

兼具效率与灵活性——分布式托管训练



3

兼具效率与灵活性——智能超参数调整 (Hyperparameter tuning)

超参数调整是受监督的机器学习回归问题

- 给定一组输入要素（超参数），超参数调整可针对您选择的指标优化模型
- 您可以选择您使用的算法定义的任何指标

为了解决回归问题，超参数调整可以猜测哪些超参数组合可能获得最佳结果，并运行训练作业来测试这些猜测

- 在测试了第一组超参数值之后，超参数调优使用Amazon SageMaker贝叶斯优化实现

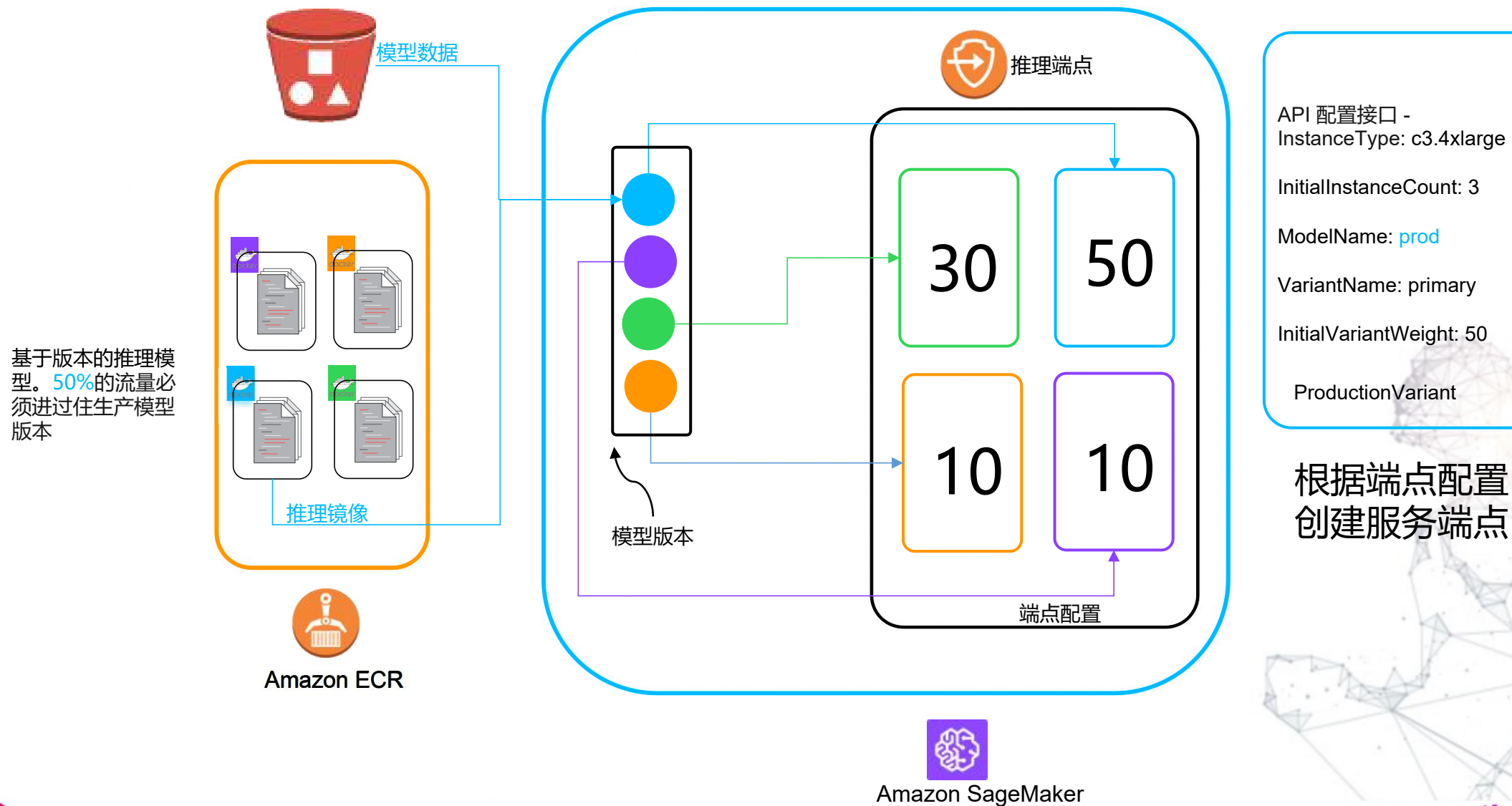
在为下一个训练工作选择最佳超参数时，超参数调优会考虑到目前为止它对此问题所知的一切

- 有时它会选择一个可能会对迄今为止发现的最佳结果进行渐进式改进的点。这允许超参数调整以利用最熟知的结果
- 有时它会选择一组远离其尝试过的超参数。这允许它探索空间以试图找到不太了解的新区域

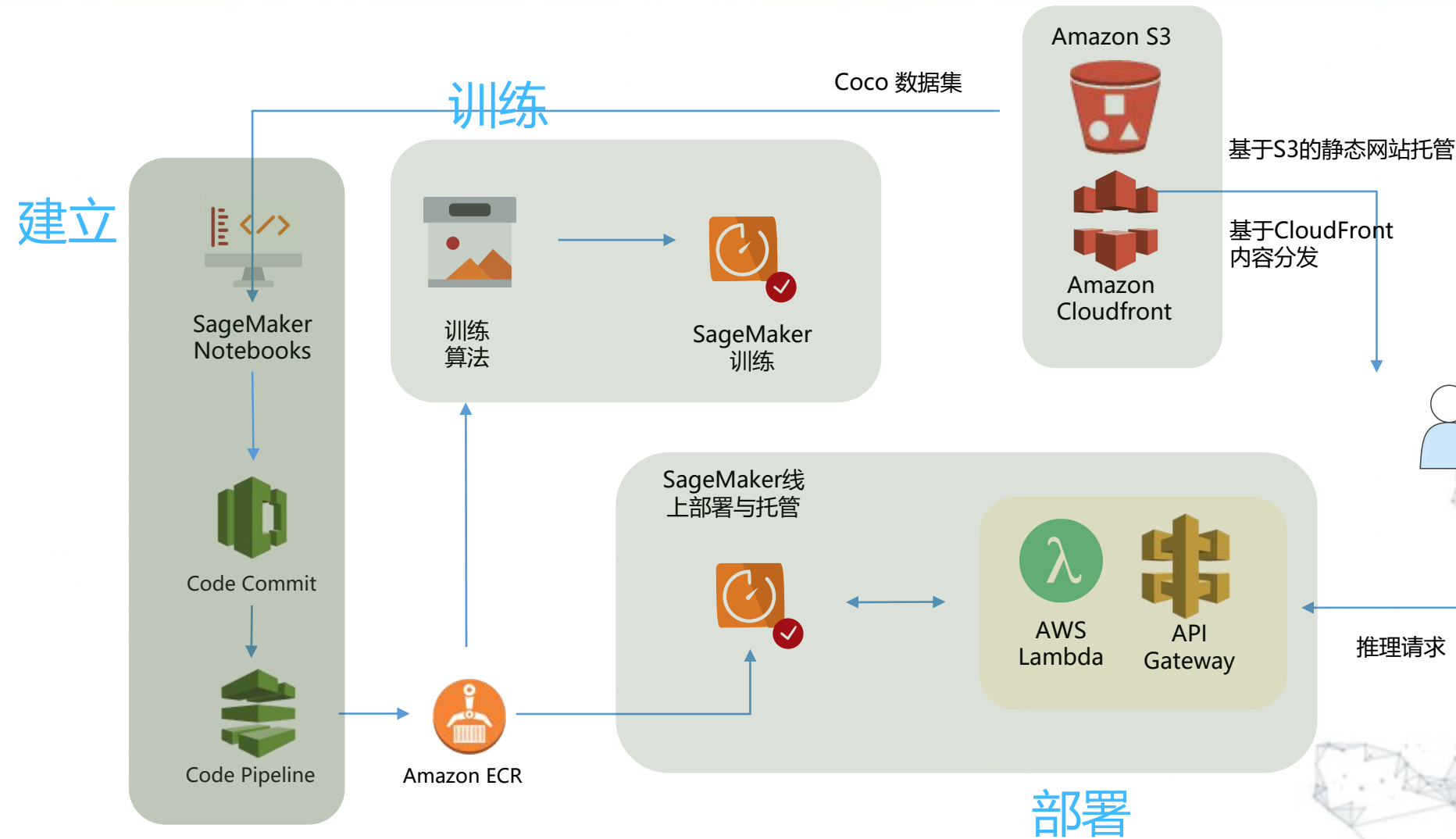
learning_rate	Continuous ▼		-	
lr_scheduler_fa	Static ▼	0.1		
lr_scheduler_st	Static ▼			
optimizer	Categorical ▼	sgd ✕ adam ✕ rmsprop ✕ nag ✕		
momentum	Continuous ▼		-	
weight_decay	Continuous ▼		-	
beta_1	Continuous ▼		-	
beta_2	Continuous ▼		-	

4

轻松部署训练模型



基于 SageMaker 端到端参考架构

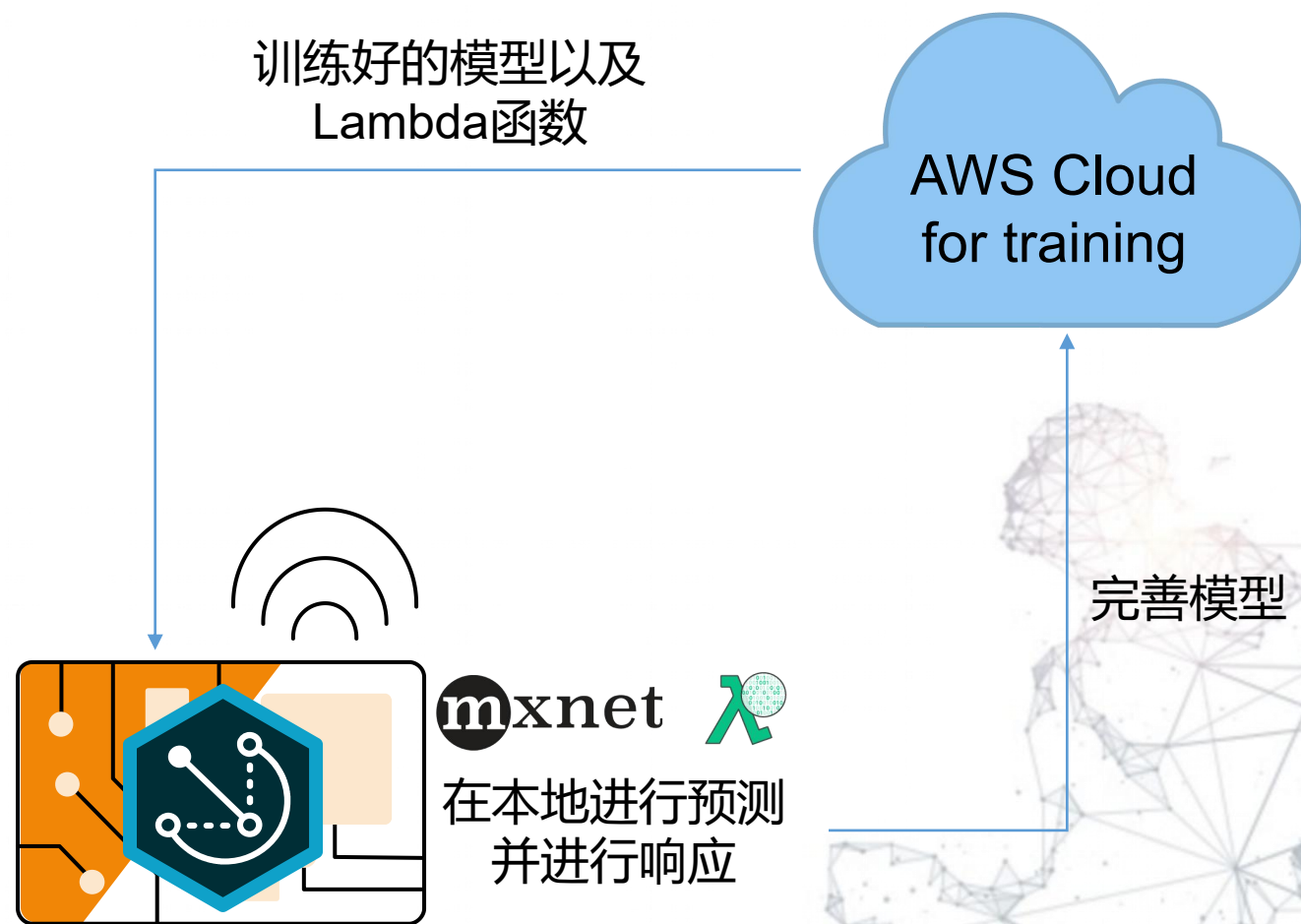


为终端赋予智能——AWS Greengrass 上的 ML Inference

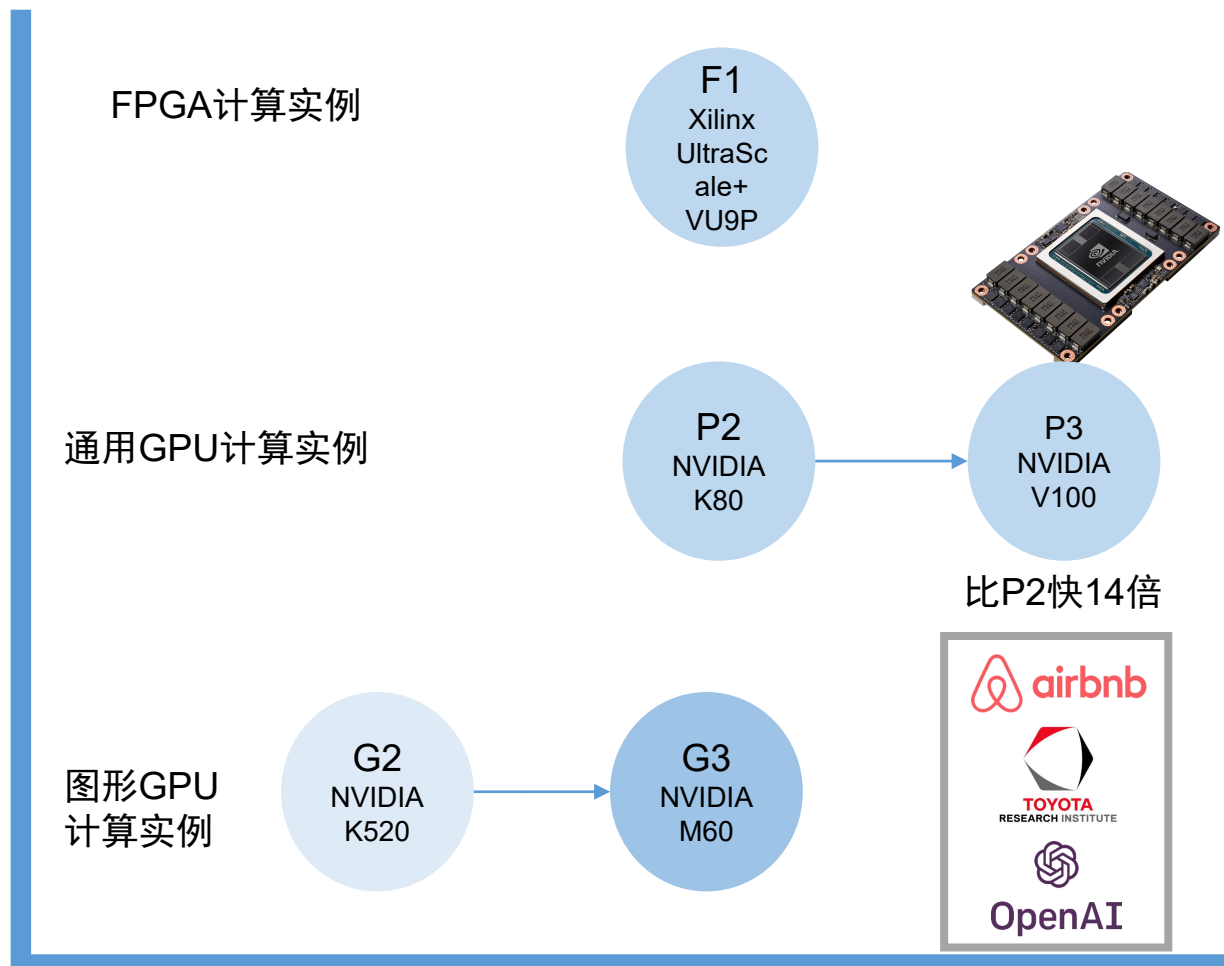
- 训练在云端
- 海量的计算能力
- 海量的数据集

预测在边缘设备

- 低延迟
- 节省带宽
- 合规性/隐私



最灵活的选择——基于**EC2**虚拟机的框架和接口



AWS Deep Learning AMI

预配置的环境便于快速构建深度学习应用程序

TensorFlow

mxnet

GLUON

theano

PYTORCH

K

Chainer

Caffe

The Microsoft
Cognitive Toolkit

Caffe2

2018 AI先行者大会

- Expedia 拥有来自30万家酒店的超过1,000万张图片
- 利用图片加速商业机会的转换
- 使用 Keras 框架运行在 EC2 GPU 实例上, 使用了10万个图像对预先训练的卷积神经网络进行微调
- 酒店的描述可以自动的提供最佳的图片

Some images are really good



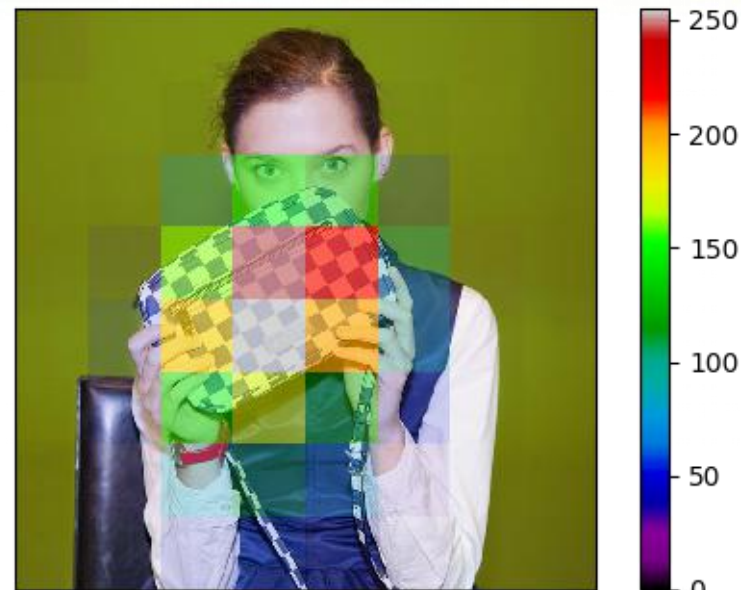
Others not so much



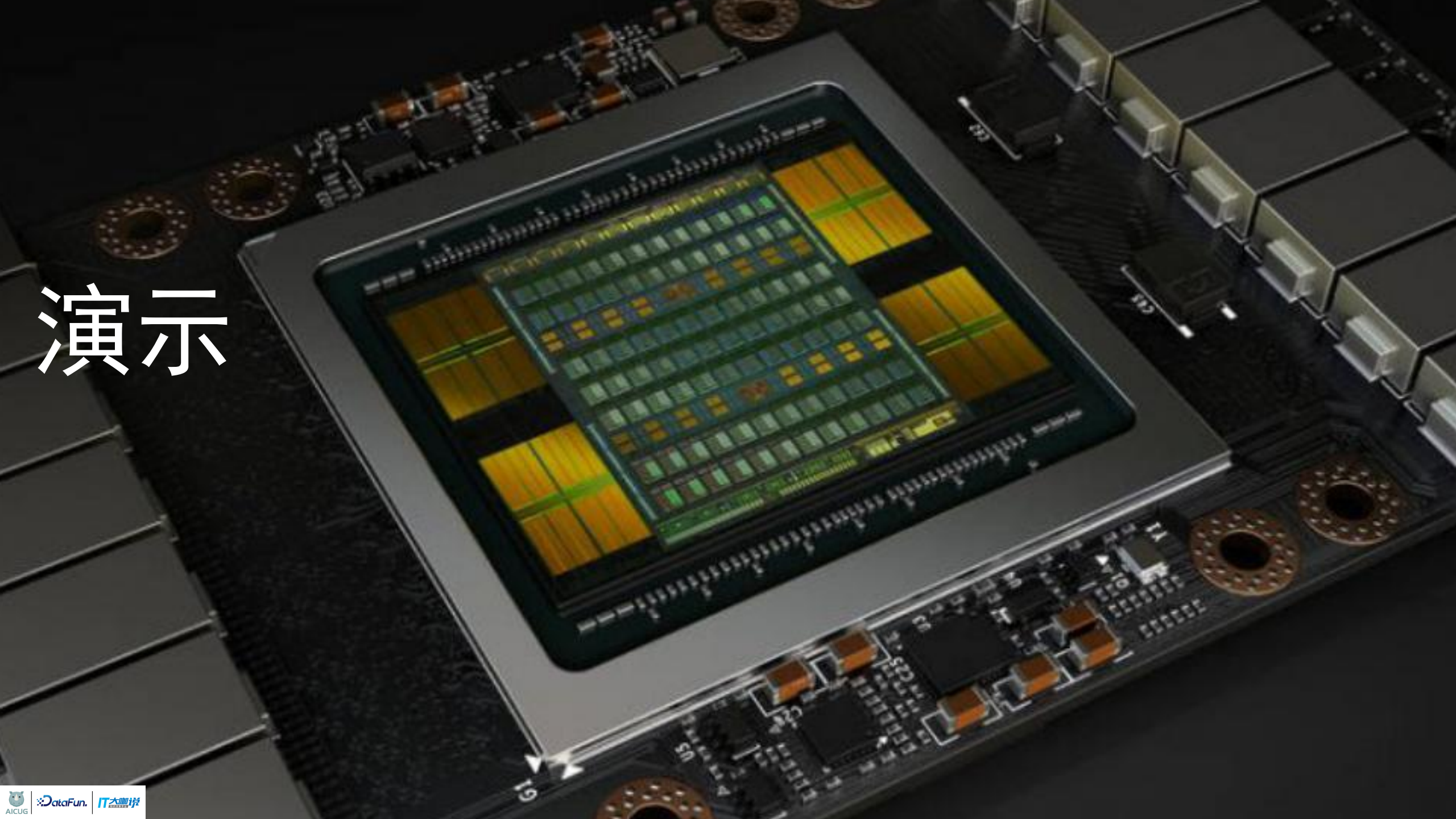
<https://news.developer.nvidia.com/expedia-ranking-hotel-images-with-deep-learning/>

- 源自Instagram 的17,000 张图片
- 7个品牌
- 基于ImageNet 的深度学习预先训练的模型
- TensorFlow 框架下微调，运行在 EC2 GPU实例上
- 此外还实现了颜色抽取

	Chanel	Coach	Gucci	Marc Jacobs	Kate Spade	No Handbag	Prada	Vuitton
Chanel	0.83	0.00	0.01	0.02	0.00	0.00	0.00	0.01
Coach	0.01	0.85	0.00	0.05	0.05	0.01	0.04	0.03
Gucci	0.01	0.00	0.85	0.02	0.00	0.01	0.01	0.02
Marc Jacobs	0.00	0.03	0.01	0.78	0.00	0.01	0.03	0.00
Kate Spade	0.00	0.01	0.01	0.01	0.87	0.00	0.00	0.00
No Handbag	0.09	0.06	0.08	0.09	0.04	0.97	0.04	0.09
Prada	0.03	0.03	0.02	0.03	0.01	0.00	0.85	0.01
Vuitton	0.01	0.00	0.00	0.02	0.00	0.01	0.01	0.81



演示



AI

复杂度和学习成本
是主要问题



简化深度学习
代码编写的工作量



Sagemaker将深度学习的训练和推理工作复杂化为简单的底层框架和计算资源定制



程度更高的EC2机器学习平台



请关注AWS微信公众号:

[awschina](#)

访问网站:

www.amazonaws.cn

THANKS

