在生产环境上部署深度学习

主讲人：吴书卫

关于 SKYMIND

Deeplearning4j 的商业支持机构

SKYMIND是一家提供企业级人工智能深度学习开源平台及企业支援的公司，肩负了提升深度学习开源平台核心竞争力的重要使命。

主要目的是帮助企业、政府及集团设计与部署深度学习架构

SKYMIND以「专注平台开发、创新、整合、人性化」为理念，通过技术与业务模式创新，构建完整的智能生态链，提升深度学习平台的核心竞争力，为企业、政府及集团提供可靠和稳定的全方位人工智能平台

skymind

摘要

关于深度学习

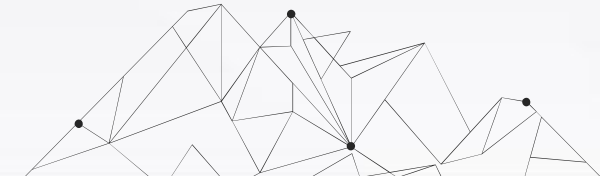
数据科学家在部署深度学习时遇到的难题

部署深度学习的解决方案

Deeplearning4j 深度学习框架

深度学习建模（模型训练）流程

运行模型

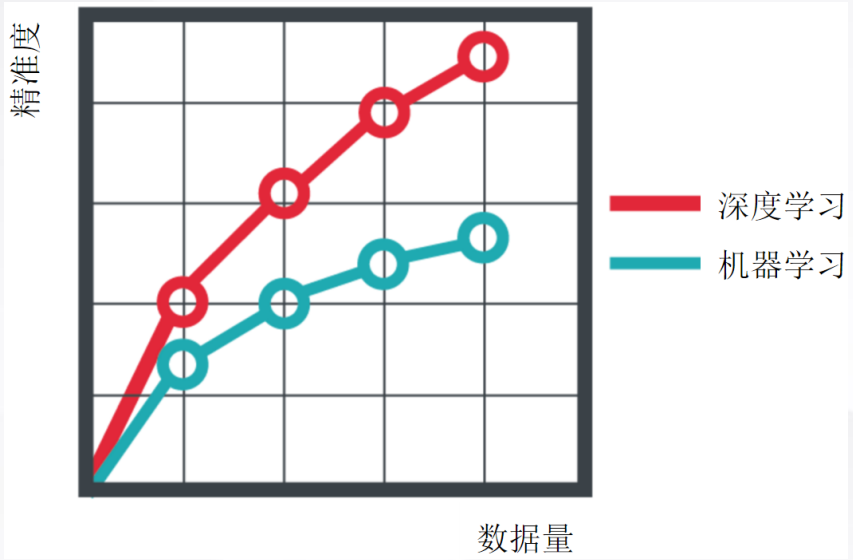


关于深度学习

深度学习的概念源于人工神经网络的研究：

- 机器学习研究中的一个新的领域，其动机在于**建立、模拟人脑**进行分析学习的神经网络
- 大大提高了计算的精度与准确率
- 能识别，分析并**学习**文字，图片，声音，视频以及时间序列数据。
- **能自动学习与发掘数据的特征**
- 未来计算机发展的必然趋势

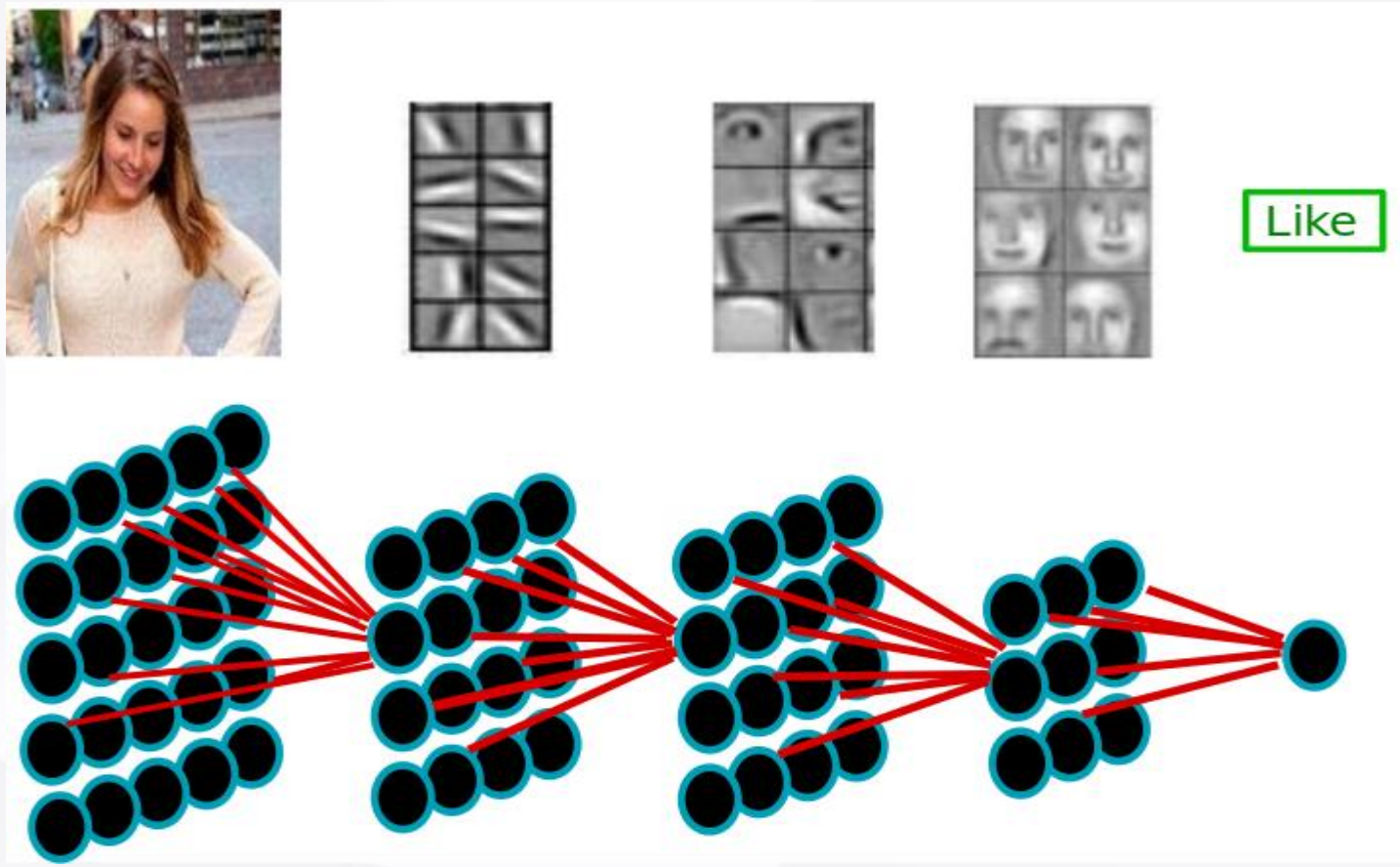
高精准度：自动学习与发掘数据的特征



深度学习的优势在于它能随着数据的增加，精准度也会随着提高

Data Sector	Use Case	Input	Transform	Neural Net
Text	Sentiment analysis	Word vector	Gaussian Rectified	RNTN or DBN (with moving window)
	Named-entity recognition	Word vector	Gaussian Rectified	RNTN or DBN (with moving window)
	Part-of-speech tagging	Word vector	Gaussian Rectified	RNTN or DBN (with moving window)
	Semantic-role labeling	Word vector	Gaussian Rectified	RNTN or DBN (with moving window)
Document	Topic modeling/ semantic hashing (unsupervised)	Word count probability	Can be Binary	Deep Autoencoder (wrapping a DBN or SDA)
	Document classification (supervised)	TF-IDF (or word count prob.)	Binary	Deep-belief network, Stacked Denoising Autoencoder
Image	Image recognition	Binary	Binary (visible and hidden)	Deep-belief network
		Continuous	Gaussian Rectified	Deep-belief network
	Multi-object recognition			Convolutional Net, RNTN (image vectorization forthcoming)
	Image search/ semantic hashing		Gaussian Rectified	Deep Autoencoder (wrapping a DBN)
Sound	Voice recognition		Gaussian Rectified	Recurrent Net
				Moving window for DBN or ConvNet
Time Series	Predictive analytics		Gaussian Rectified	Recurrent Net
				Moving window for DBN or ConvNet

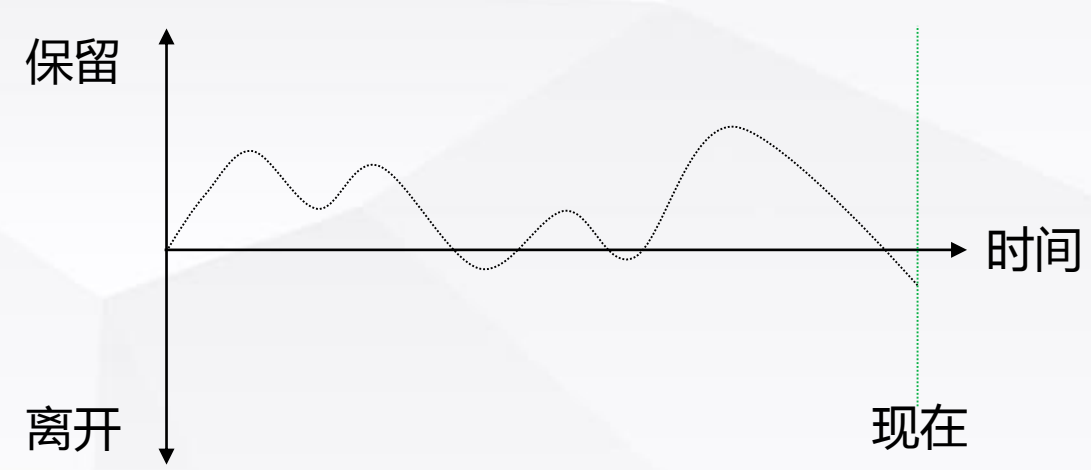
用列：TINDER 手机交友APP

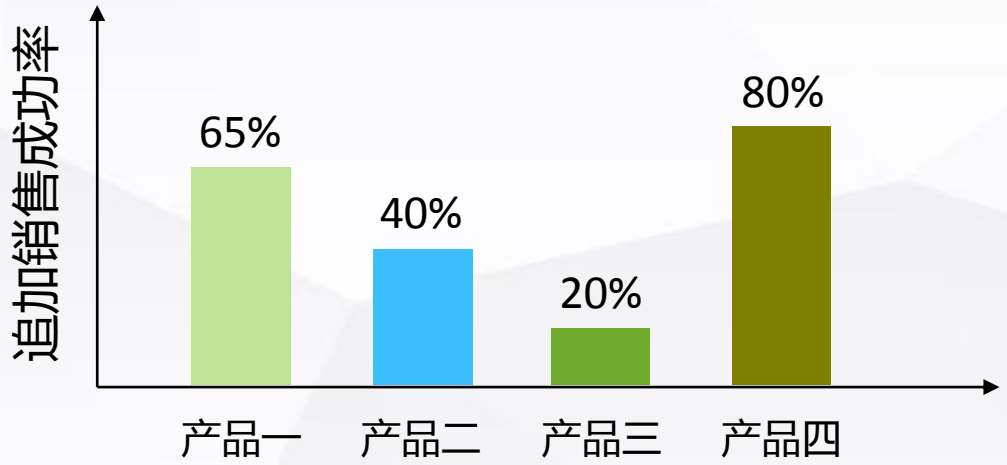


用列：用户分析

用户活动记录

Time	_source
June 10th 2015, 15:37:11.108	message: 2015-06-10 13:37:10 status installed cowsay:all 3.03+dfsg1-9 @version: 1 @timestamp: June 10th 2015, 15:37:11.108 type: syslog host: scw-08d553 path: /var/log/dpkg.log tags: _grokparsefailure syslog_severity_code: 5 syslog_facility_code: 1 syslog_facility: user-level syslog_severity: notice _source: {"message":"2015-06-10 13:37:10 status installed cowsay:all 3.03+dfsg1-9","@version":"1","@timestamp":"2015-06-10T13:37:11.108Z","type":"syslog","host":"scw-08d553","path":"/var/log/dpkg.log","tags":
June 10th 2015, 15:37:11.108	message: 2015-06-10 13:37:10 status installed cowsay:all 3.03+dfsg1-9 @version: 1 @timestamp: June 10th 2015, 15:37:11.108 type: syslog host: scw-08d553 path: /var/log/dpkg.log tags: _grokparsefailure syslog_severity_code: 5 syslog_facility_code: 1 syslog_facility: user-level syslog_severity: notice _source: {"message":"2015-06-10 13:37:10 status installed cowsay:all 3.03+dfsg1-9","@version":"1","@timestamp":"2015-06-10T13:37:11.108Z","type":"syslog","host":"scw-08d553","path":"/var/log/dpkg.log","tags":
June 10th 2015, 15:37:11.106	message: 2015-06-10 13:37:10 status half-configured cowsay:all 3.03+dfsg1-9 @version: 1 @timestamp: June 10th 2015, 15:37:11.106 type: syslog host: scw-08d553 path: /var/log/dpkg.log tags: _grokparsefailure syslog_severity_code: 5 syslog_facility_code: 1 syslog_facility: user-level syslog_severity: notice _source: {"message":"2015-06-10 13:37:10 status half-configured cowsay:all 3.03+dfsg1-9","@version":"1","@timestamp":"2015-06-10T13:37:11.106Z","type":"syslog","host":"scw-08d553","path":"/var/log/dpk
June 10th 2015, 15:37:11.106	message: 2015-06-10 13:37:10 status half-configured cowsay:all 3.03+dfsg1-9 @version: 1 @timestamp: June 10th 2015, 15:37:11.106 type: syslog host: scw-08d553 path: /var/log/dpkg.log tags: _grokparsefailure syslog_severity_code: 5 syslog_facility_code: 1 syslog_facility: user-level syslog_severity: notice _source: {"message":"2015-06-10 13:37:10 status half-configured cowsay:all 3.03+dfsg1-9","@version":"1","@timestamp":"2015-06-10T13:37:11.106Z","type":"syslog","host":"scw-08d553","path":"/var/log/dpk





数据科学家遇到的难题

数据传输

数据传输到另一个集群去处理会把影响整个深度学习模型训练流程的速度
数据传输到另一个集群去处理会吧整个深度学习模型训练流程复杂化

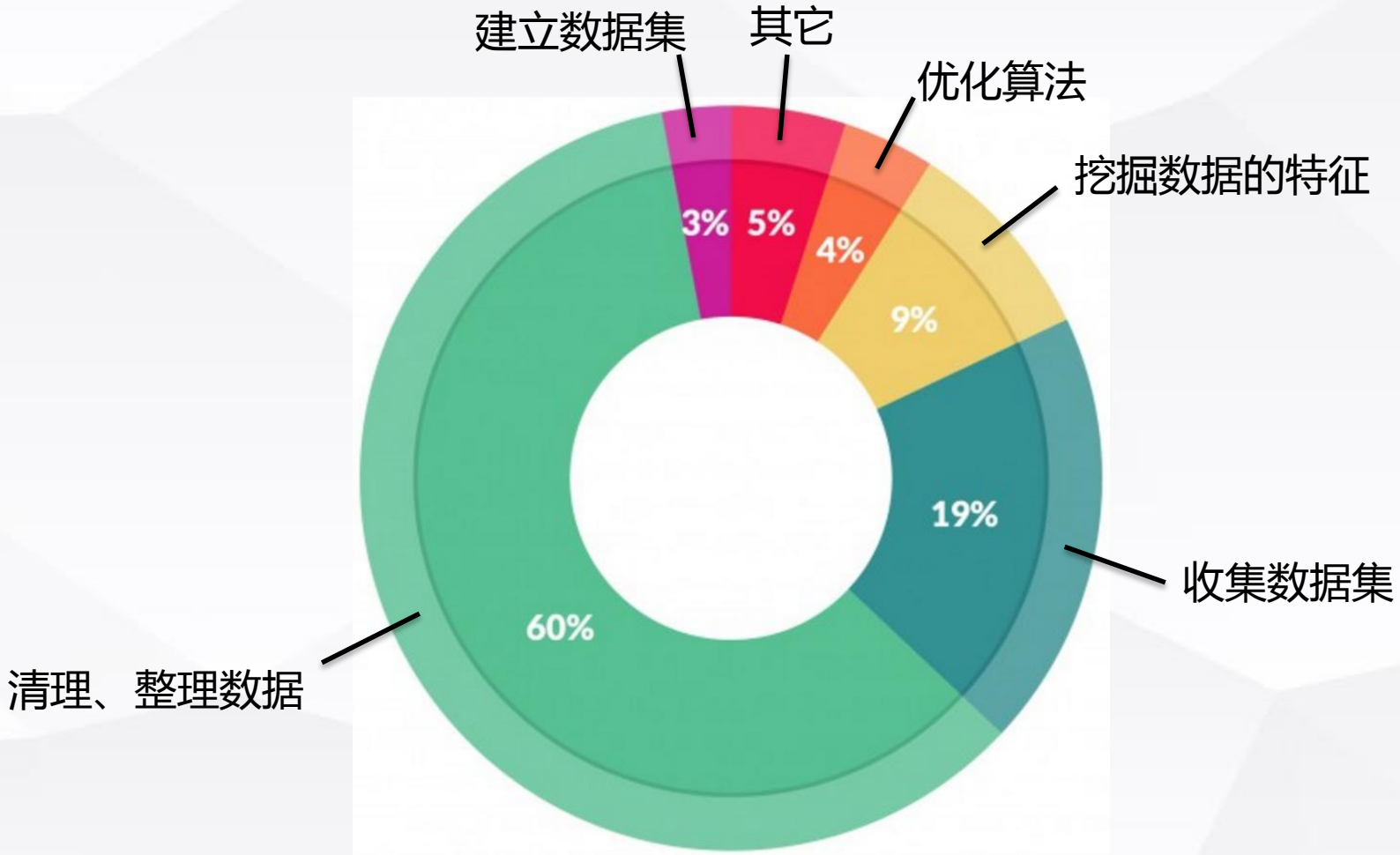
集成问题

数据摄取，抽取、转换、装载（ETL），矢量化，建模，评估与部署问题
大多数的机器学习工具是由基于过时（上一代）的架构而设计

传统架构

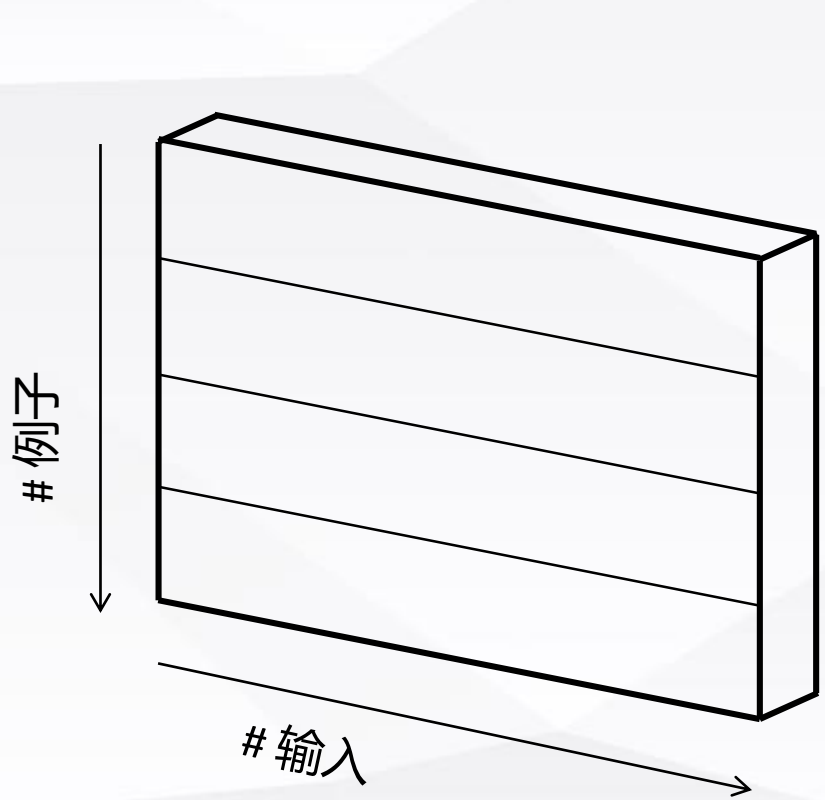
并行迭代算法架构是很少的

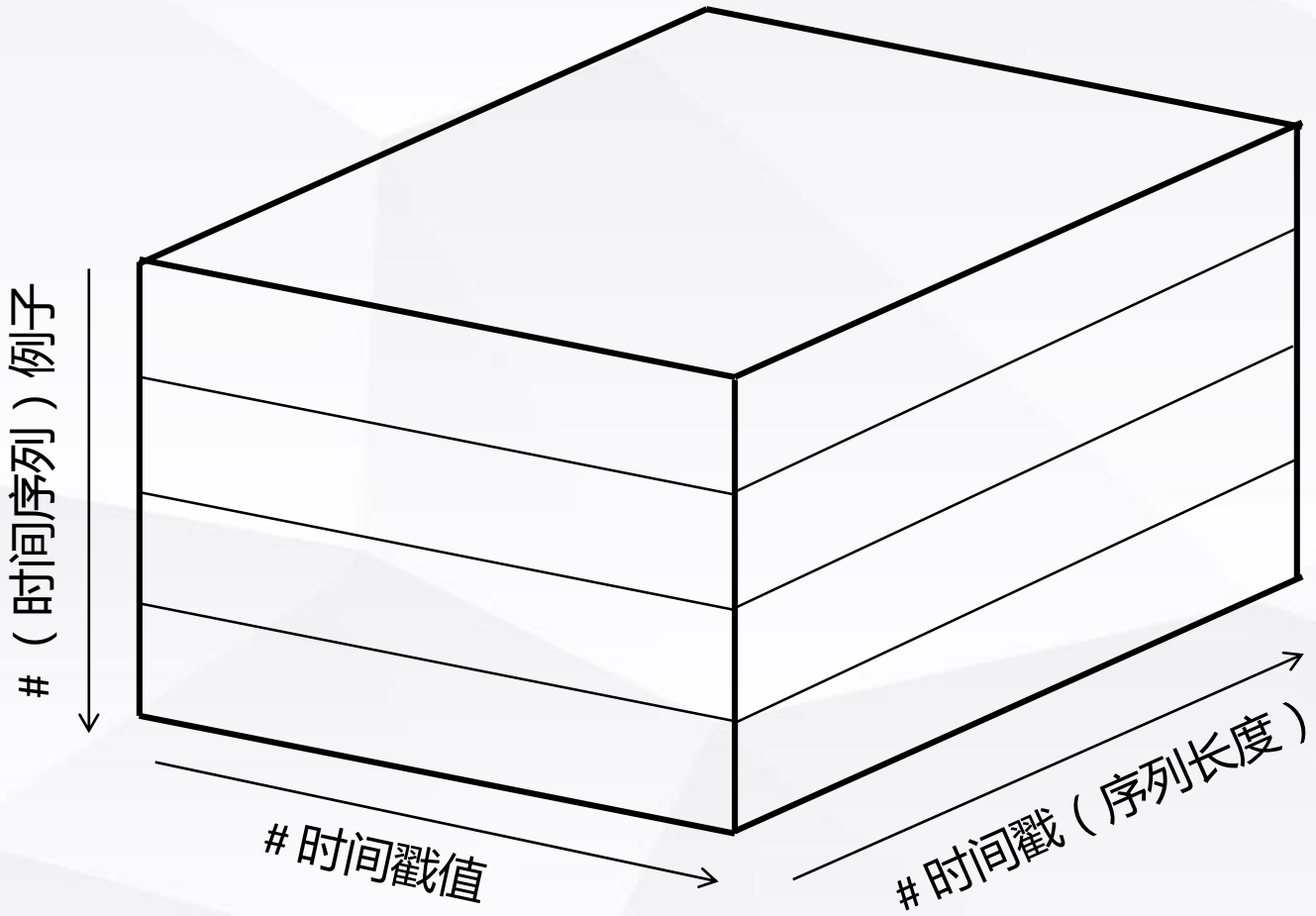
问题一：处理海量数据



参考：http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says

问题二：把海量数据向量化（Vectorization）





问题三：建模（训练模型）

建模、调模

GPU 集群

C 代码

数据输出、隔离、清理、转换、格式化、向量化

非常耗时间

大数据系统

Hadoop 集群：数据库

Java 代码

问题四：运用模型

建模、调模

GPU 集群

C 代码

部署深度学习模型在生产环境

重新编辑机器学习工具

大数据系统

Hadoop 集群：数据库

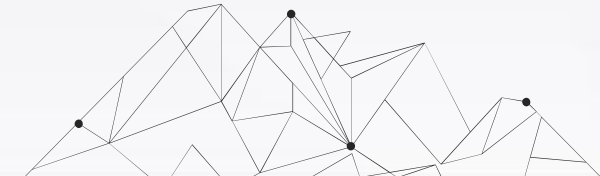
Java 代码

使用深度学习

不管拥有大数据或小数据，都可以方便的部署深度学习
　　避免为了升级到大数据系统（HADOOP）时把原本的机器学习工具都换掉

避免花费时间在数据矢量化与抽取、转换、装载（ETL）
　　专注于开发更好的深度学习模型

可以同时间实验、训练更多的深度学习模型
　　同时也要避免为了把深度学习部署到生产线时需要重新编辑机器学习工具



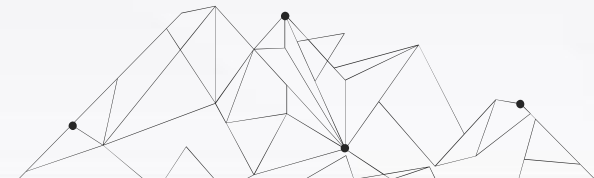
Deeplearning4j (DL4J) 系列工具

DataVec
- 深度学习专用的矢量处理器
- 数据标准化处理器
- 处理非结构化数据

Deeplearning4j
- 企业级商用的开源深度学习平台
- 专为Java和Scala编程的深度学习

ND4J
- 转为JVM开发的科学运算引擎
- JavaCPP:Java 到 Objective-C 的桥

Arbiter
- 深度学习模型检测、评估器
- 调整及优化机器学习模型



DataVec

主要解决数据输出、隔离、清理、转换、格式化、向量化等问题

- 机器学习的ETL（抽取、转换、装载）操作
- 主要目的是把原始数据（Raw Data）转化成可用的向量格式，让所有的深度学习工具都可以使用
  - 支持CSV、原始文本及、图像数据
- 拥有强大功能：数据特征处理、数据清理、数据规范化。这些功能都可以在Spark上
- 开源工具 ASF 2.0许可证：github.com/deeplearning4j/DataVec

ND4J

让数据科学家在同一个集群上充分的利用GPU，CPU和内存：训练和运行深度学习模型。

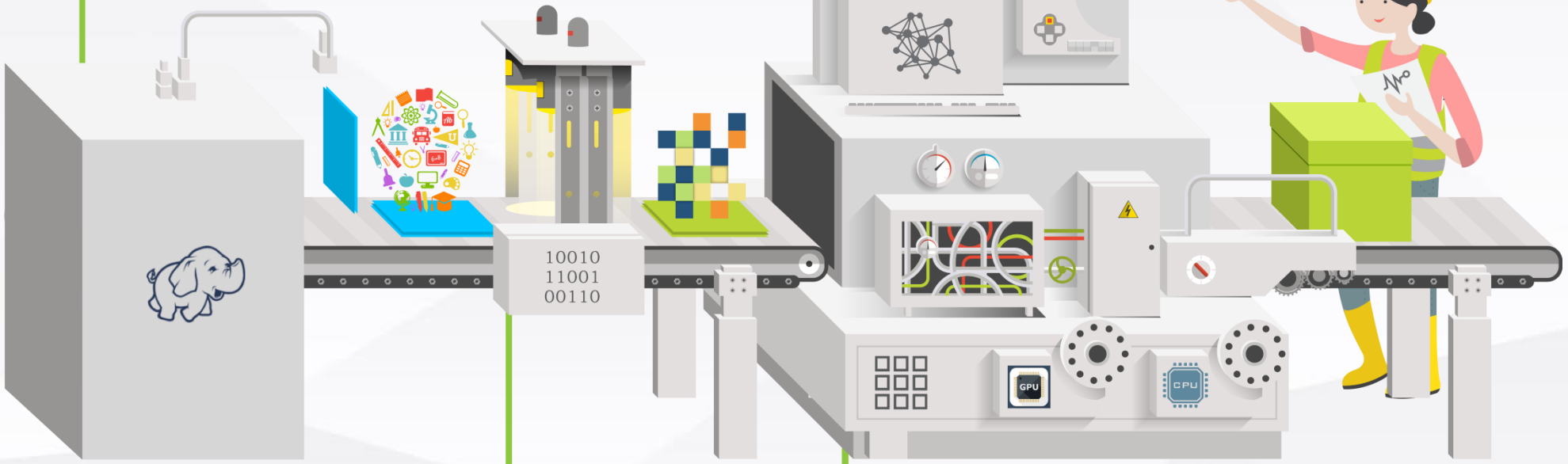
- JavaCPP: Java 到 Objective-C 的桥，可像其他 Java 对象一样来使用 Objective-C 对象。
- CPU 后瑞：OpenMP、OpenBlas 或 MKL、与SIMD的扩展
- GPU 后瑞：最新CUDA 及 CuDNN
- 开源工具 ASF 2.0许可证：github.com/deeplearning4j/nd4j

深度学习建模（模型训练）流程

HADOOP
HDFS
数据来源

DL4J
建模

ARBITER
模型评估

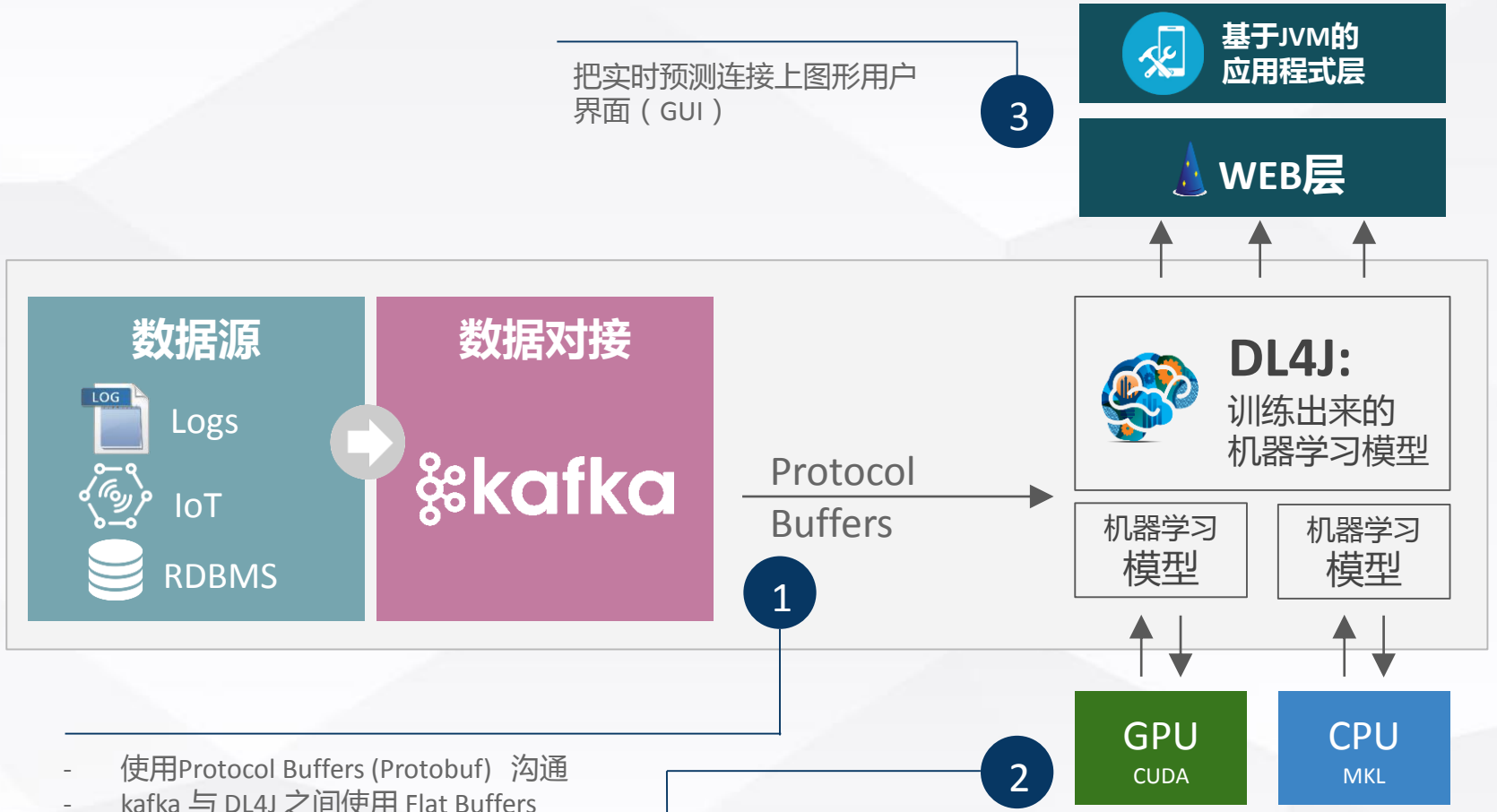


DATAVEC
数据提取、
转换和加载（ETL）

DATAVEC
向量化

ND4J
线性代数运行时：CPU、GPU

运行模型


把实时预测连接上图形用户界面（GUI）
3
基于JVM的应用程式层
WEB层
数据源
Logs
IoT
RDBMS
数据对接
kafka
Protocol Buffers
1
DL4J:
训练出来的机器学习模型
机器学习模型
机器学习模型
GPU
CUDA
CPU
MKL
2
- 使用Protocol Buffers (Protobuf) 沟通
- kafka 与 DL4J 之间使用 Flat Buffers
- 二进制格式
针对故障使用多种机器学习模型应对多种需求。每个模型将会在一台机上运算。无分布式运行（因此也没有平均化），因为之间的沟通是低效的。

解决方案

数据库 + 建模、调模 + 运行模型

Hadoop 集群
Java 代码

+

GPU 集群
C 代码

DataVec
- 深度学习专用的矢量处理器
- 数据标准化处理器
- 处理非结构化数据

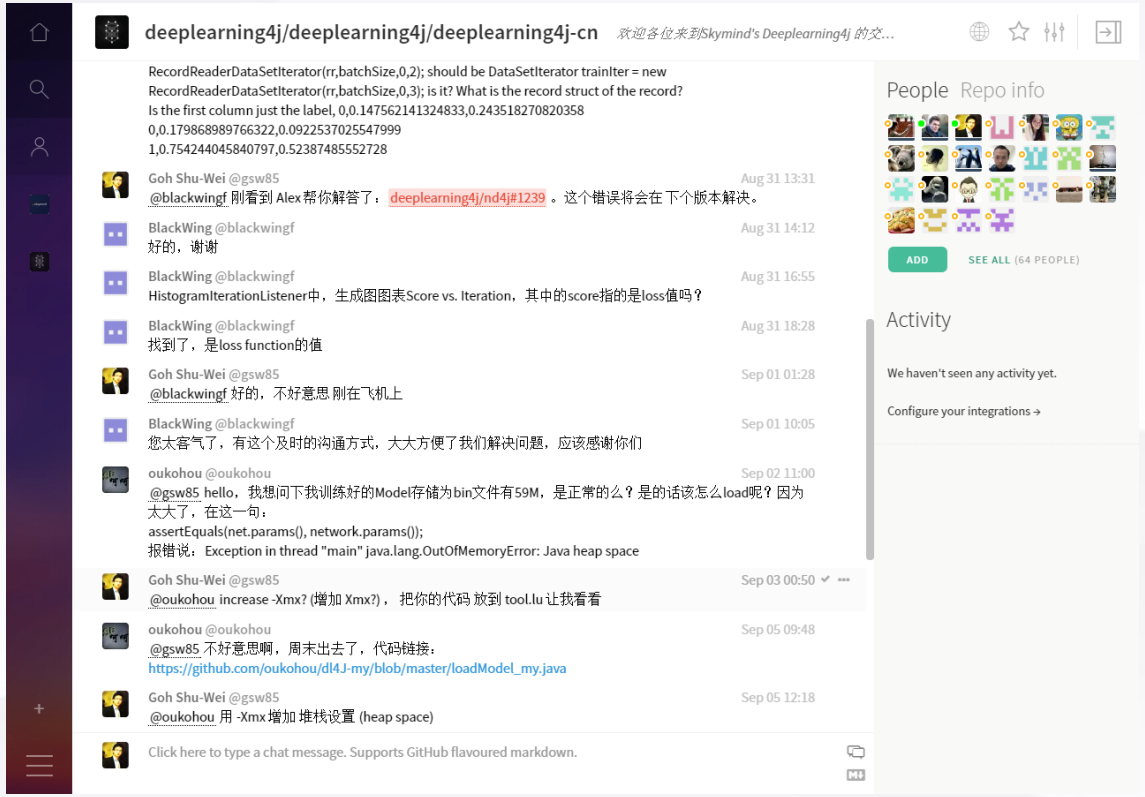
ND4J
- 转为JVM开发的科学运算引擎
- 可以在最低内存的配置下高效运行

Deeplearning4j
- 企业级商用的开源深度学习平台
- 专为Java和Scala编程的深度学习

Arbiter
- 深度学习模型检测、评估器
- 调整及优化机器学习模型

24/7 在线实时与SKYMIND的工程师和其它开发者分享、讨论心得

网址：gitter.im/deeplearning4j



微信群讨论区

T11

2016

THANKS

TalkingData