中国国际大数据大会

China International Big Data Summit

资料汇编

（仅供内部交流，请勿外传）

百度大数据引擎

2014.8.20

Baidu 百度

目录

3

全球最大的中文搜索引擎，最大的中文网站

百度每天响应来自 138个国家和地区的 70亿次搜索请求，平均每个中国网民每天使用10次百度。

百度的大数据特征

容量
Volume

- 数据总量EB级
- 每日新增800TB
- 网页量>5000亿
- 单集群数万台服务器

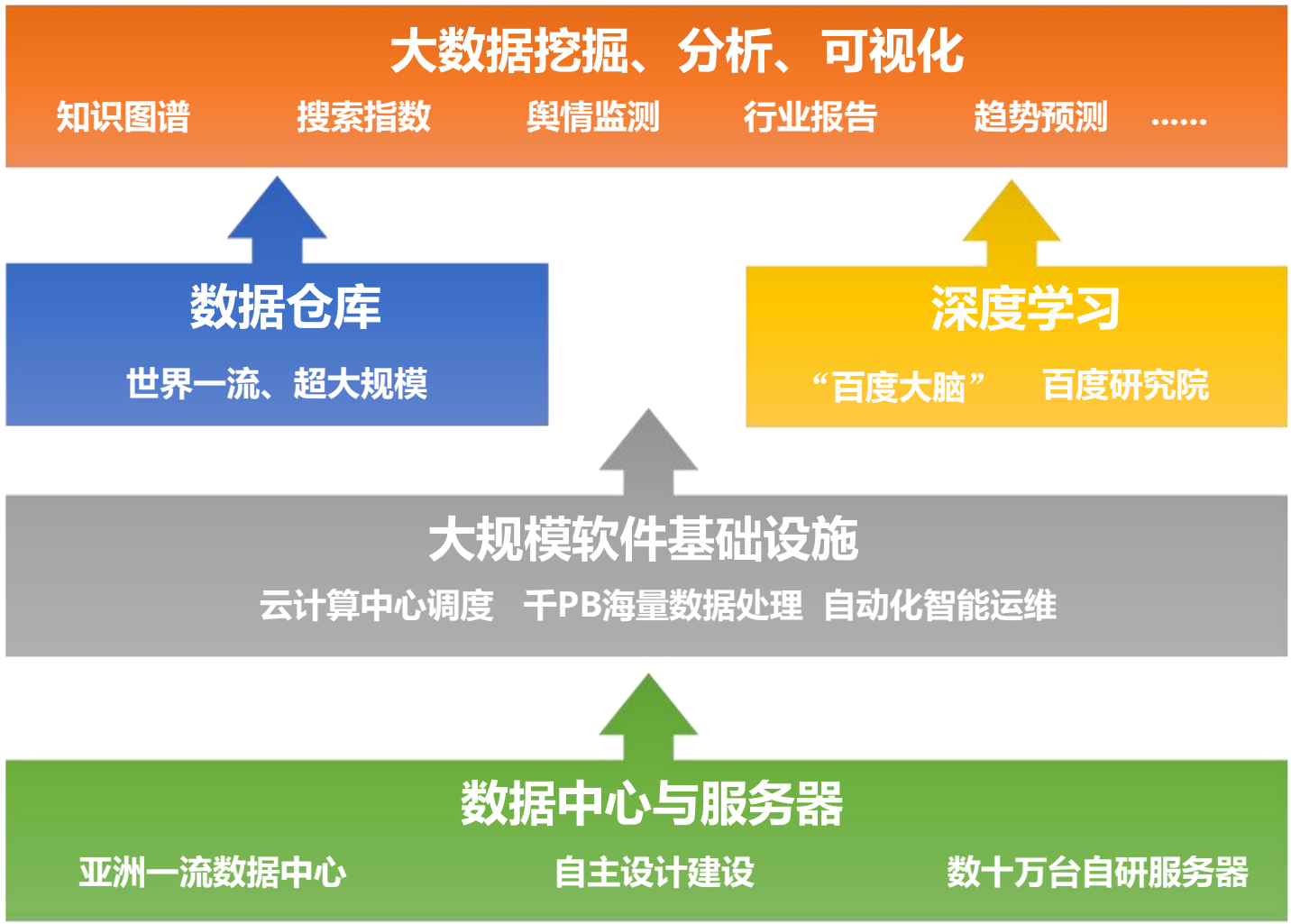
时效性
Velocity

- 毫秒-秒级响应时间
- 极速搜索最快0.04秒

多样性
Variety

- 内容：网页、广告、日志、UGC
- 类型：文本、图片、视频
- 形式：结构化、半结构化、非结构化

突破关键技术，构建完整、自主创新的大数据架构



基础设施自主设计

ARM服务器

- 全球首个**ARM**架构服务器规模化应用
- 存储密度提升**70%**

GPU服务器

- 单**GPU**计算能力可比**百片CPU**
- **GPU**实现**深度神经网络并行训练**
- 训练时间从数月缩短到**一周**

自研万兆交换机

- **业内最大规模部署**自研万兆交换机
- 接入成本下降**83%**

整机柜服务器

- **国内首次**规模部署
- 高效部署，提高交付效率**10倍**

百度IDC

- 三大自建数据中心
- 软硬件一体化设计
- 全年约一半时间完全**免费冷却**
- 国内大型数据中心**PUE**第一的最佳成绩：**年均PUE 1.32，最佳PUE 1.16**

自研大规模软件基础设施

集群操作系统

支持在线离线业务混布

完善的资源隔离方案

最大化资源利用

最小运维成本

服务实现动态伸缩

屏蔽底层硬件

故障处理全自动化

服务零宕机时间

志愿计算

高效利用空闲计算资源

额外提供10000台服务器计算能力

接入10W+台服务器

提供80万+CPU核

2W T 内存的超强计算能力

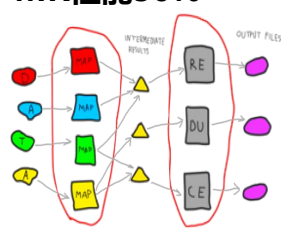
统一资源池管理

大规模分布式计算

高吞吐离线计算平台

单集群规模数万台

自主研发技术提升MR性能50%



大规模机器学习平台

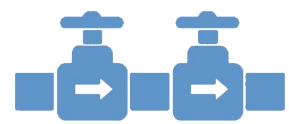
样本/特征达千亿

支持30+机器学习算法

实时流式计算平台

延迟毫秒-秒级

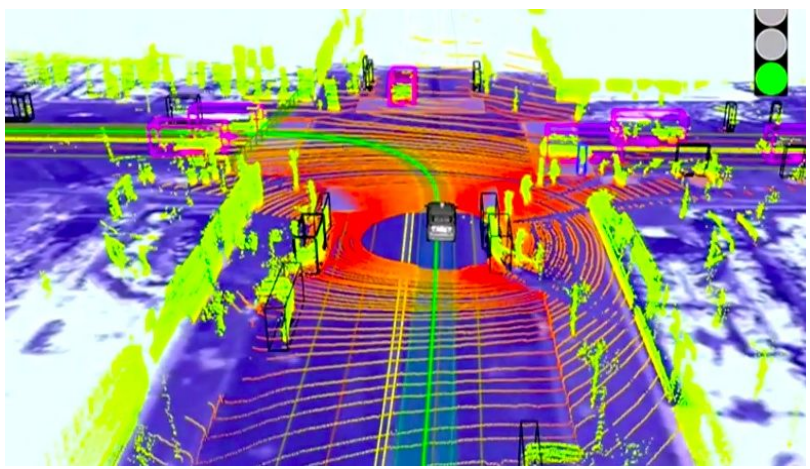
吞吐10GB/s

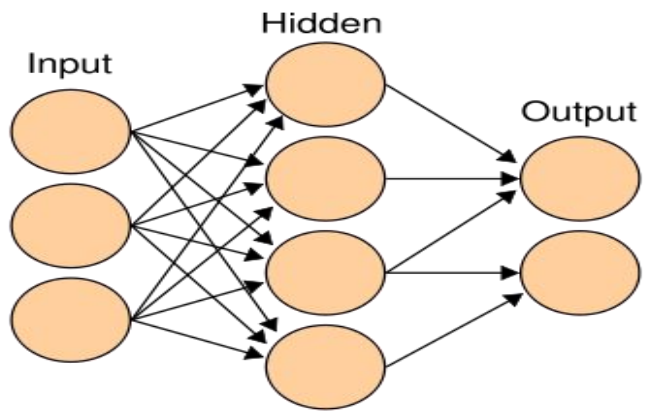


人工智能及深度学习技术

- 2014年成立百度研究院
  - 硅谷人工智能实验室
  - 北京深度学习实验室（原深度学习研究院）
  - 北京大数据实验室
- 深度学习、大规模机器学习、统计建模
- 计算机视觉、自然语言处理
- 智能交互、无人车
- 带动AI和大数据领域发展

人工智能世界级专家
百度首席科学家
吴恩达





目录

百度智能搜索

- 理解自然语言文本，推送精准答案

知识图谱，智能推荐

<海贼王> 视频

<海贼王> 百科

<海贼王> 漫画

<海贼王> 知识库 Uri

[海贼王] 搜索结果页

从不同的来源 **提取** 知识

归并, **消歧**, 和 **保存** 到知识库

组织知识库内容, **服务**用户

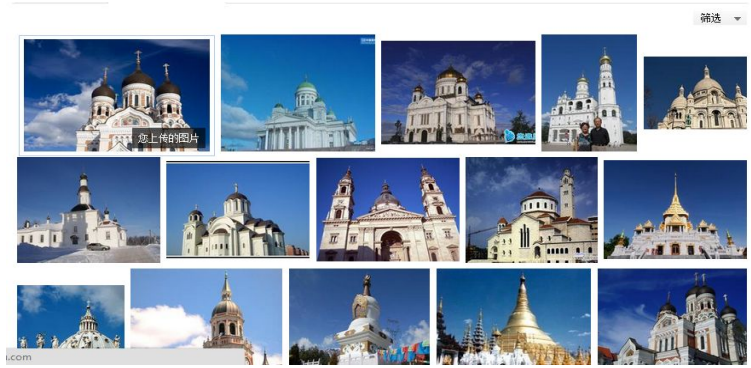
大数据驱动智能交互

语音

图像

基于图片的搜索

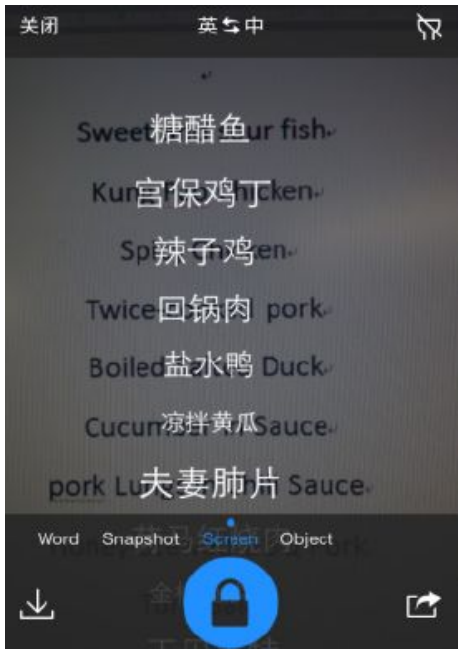
百度的结果



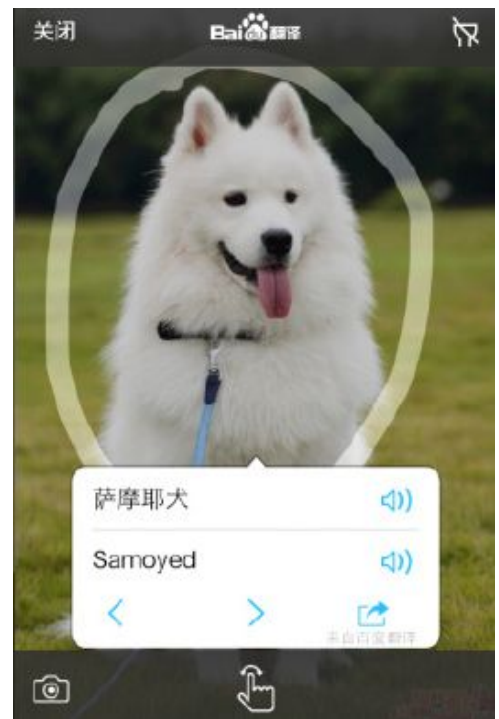
其他搜索引擎结果

- 更加自然的人机交互方式（感知，展现，…）
- 对非结构化数字媒体内容的语义解析

百度翻译



- 最懂中文的机器翻译
  - 翻译质量持平或领先业界水平
  - 时效性更新快
  - 多语种翻译（支持30个语言方向且不断增加中）
  - 语音同声翻译
  - 语音会话模式



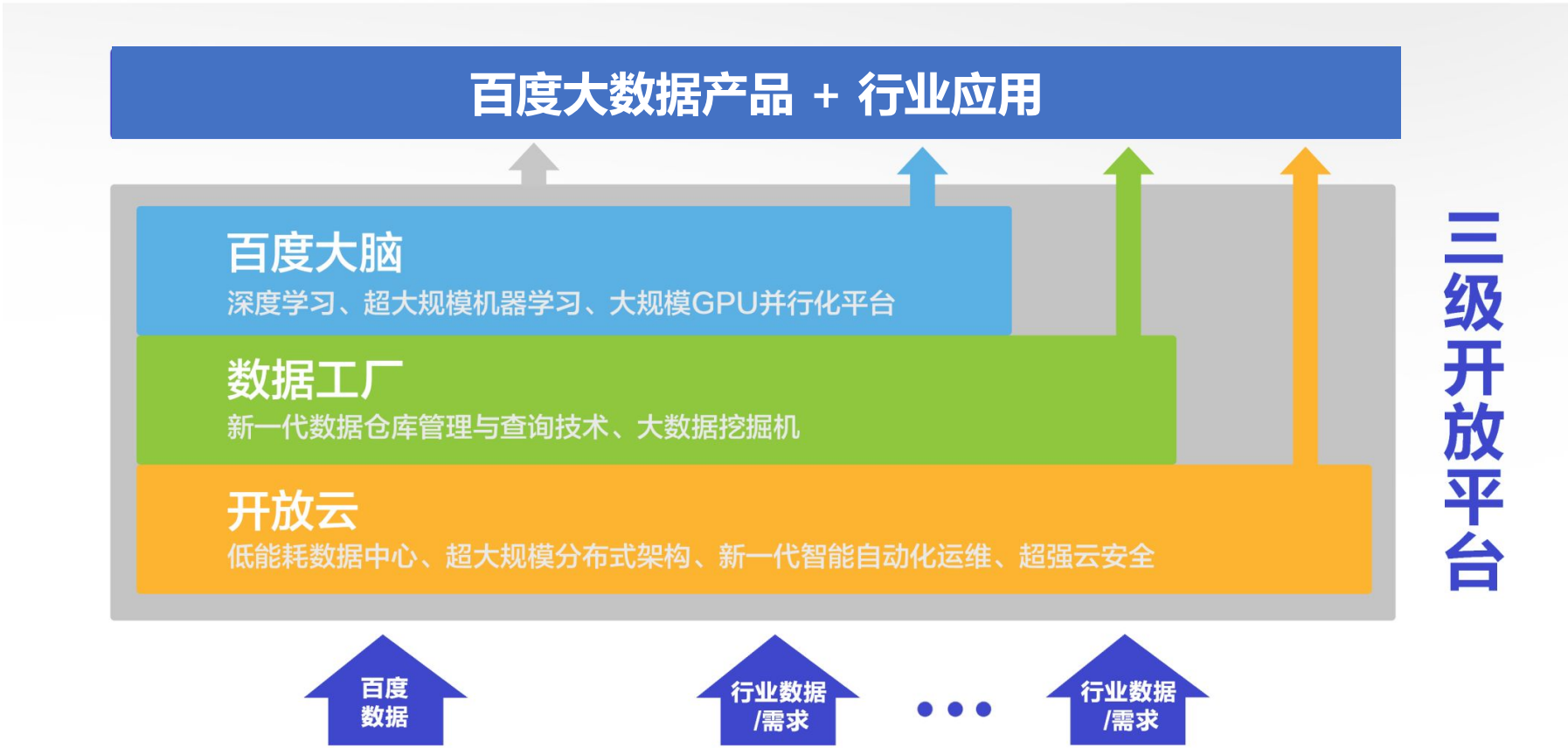
- 实物翻译
  - 将百度领先的机器翻译和图像识别技术完美结合。用户仅需对准实物拍摄，对需要翻译的物品画圈，即可进行识别、判断
  - 支持两万余个对象种类的识别和翻译，包括常见的日常用品以及人物照片等，可以从容应对用户的日常翻译需求

百度预测：大数据知天下

景点预测	城市预测	高考预测	世界杯预测	疾病预测	经济指数预测
• 未来2日的拥挤及舒适度预测	• 未来2周内人数规模预测	• 2014年全国高考语文题目，百度高考作文预测命中了全国18卷中12卷作文方向	• 淘汰赛阶段16场预测15中，并成功预测冠军球队，在各种预测产品中一枝独秀	• 4种疾病：流感、肝炎、肺结核、性病 • 未来6天内发病指数预测	• 中小企业景气指数 • 宏观经济指数预测

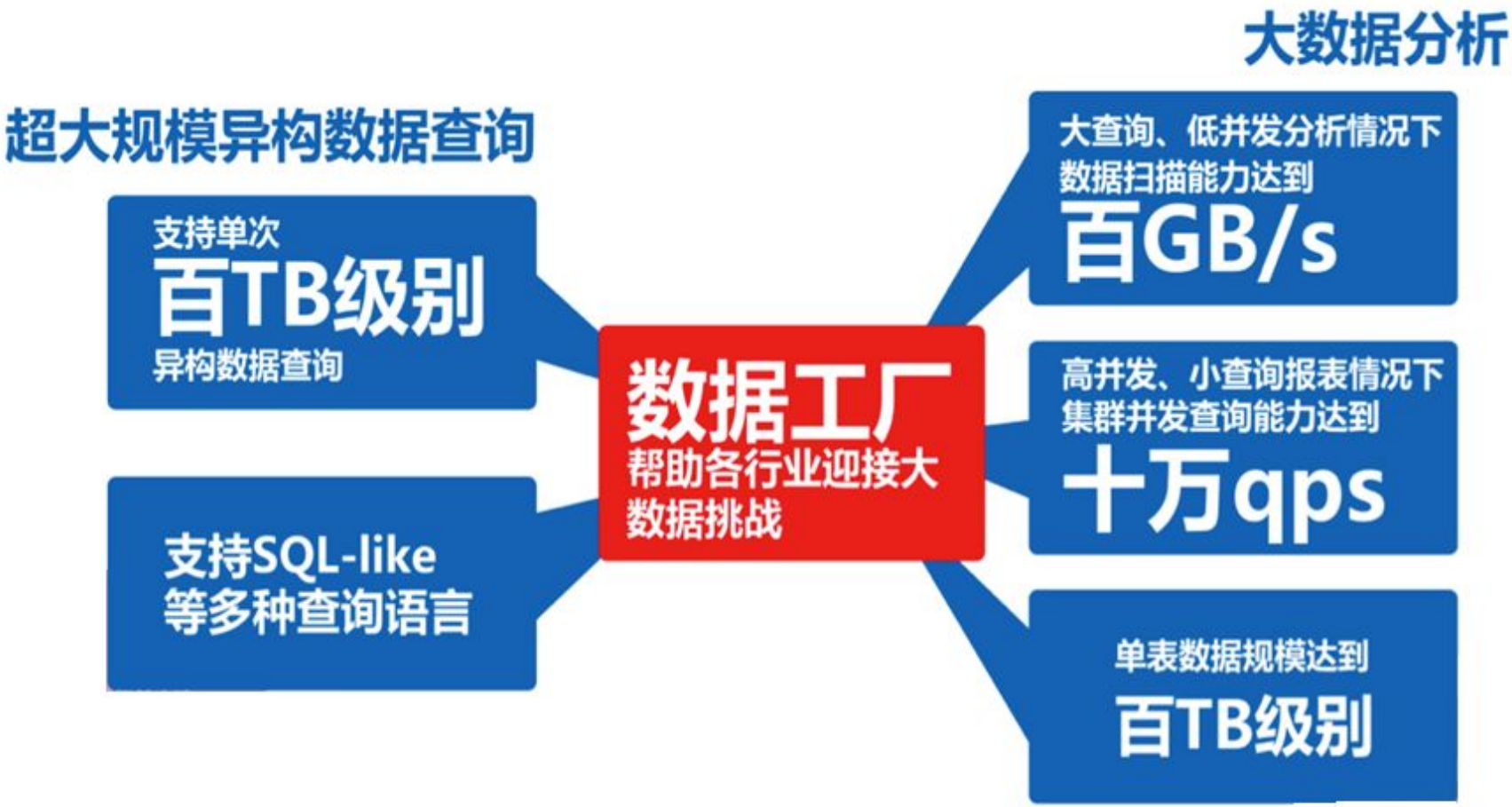
目录

开放的百度大数据引擎

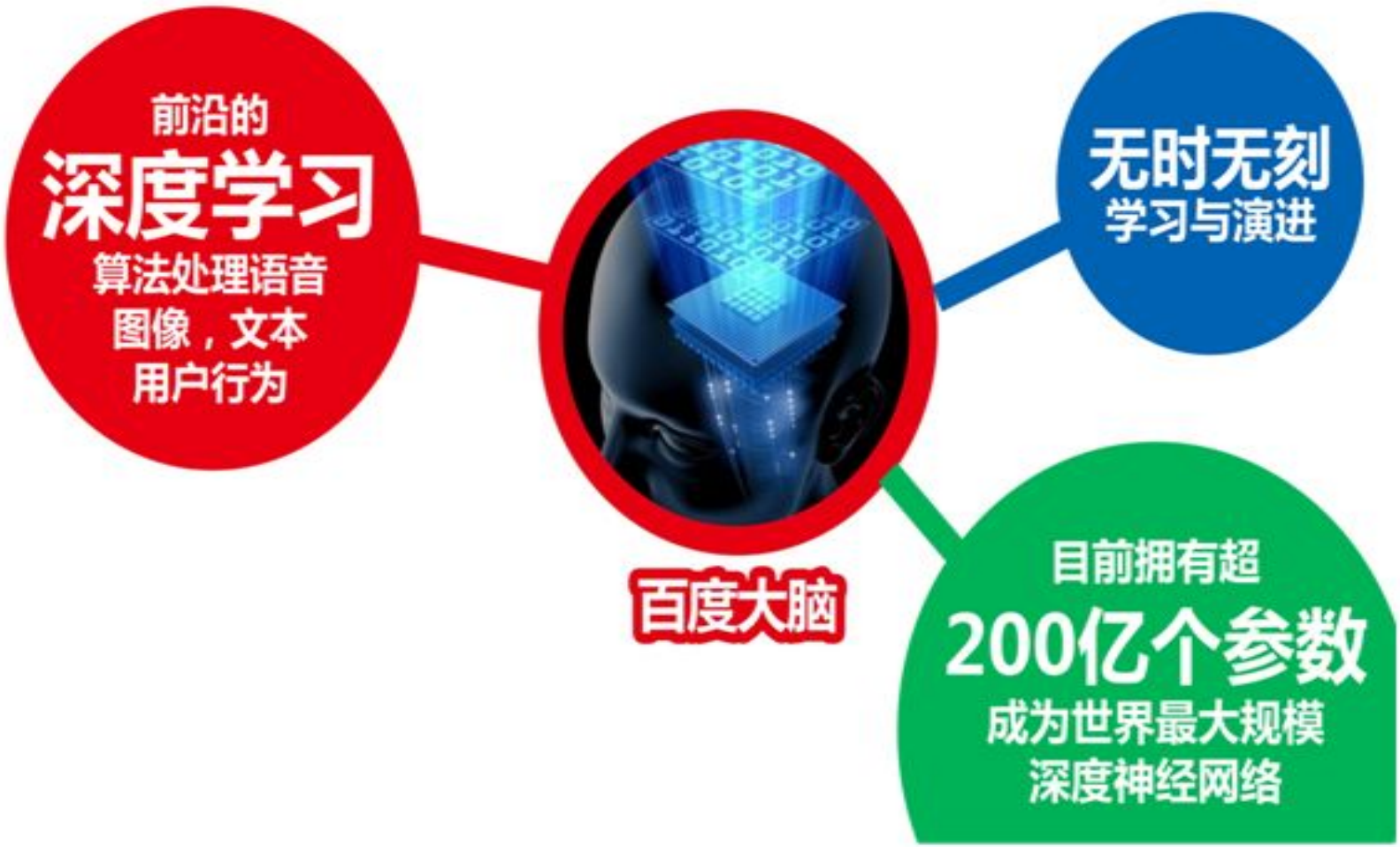

百度大数据产品 + 行业应用
百度大脑
深度学习、超大规模机器学习、大规模GPU并行化平台
数据工厂
新一代数据仓库管理与查询技术、大数据挖掘机
开放云
低能耗数据中心、超大规模分布式架构、新一代智能自动化运维、超强云安全
三级开放平台
百度数据
行业数据/需求
行业数据/需求

开放云，解决数据存储与计算规模瓶颈

数据工厂，提供强大的数据分析能力



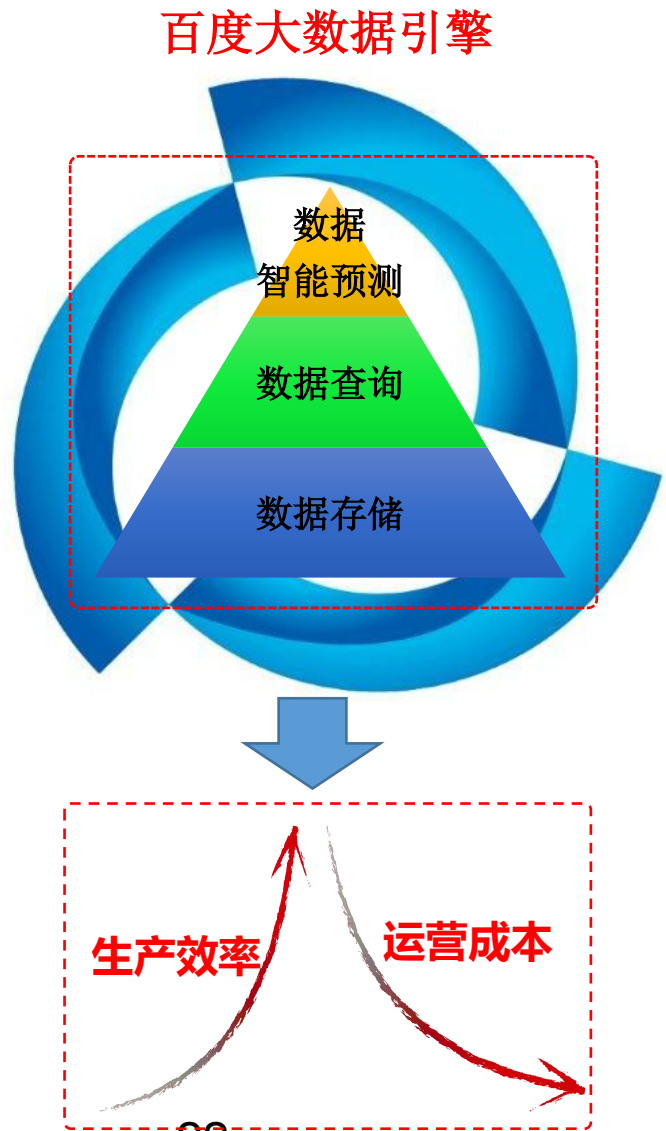
百度大脑，实现数据智能


前沿的
深度学习
算法处理语音
图像，文本
用户行为
百度大脑
无时无刻
学习与演进
目前拥有超
200亿个参数
成为世界最大规模
深度神经网络

目录

工业领域——智能监控与运维

- 百度基础设施
  - 200+万块硬盘
  - 硬件故障率：硬盘>内存>电源
- 预测故障，提前拷贝数据，更换硬盘
  - 百度经过对近9亿条实例进行采集处理
  - 选取15万个训练样本
  - 从历史的硬盘故障病例中，选取了240个特征
  - 人工智能模型预测硬盘故障
- 对比无预测的故障恢复的好处
  - 提前一天预测出硬盘故障
  - 提前拷贝数据，更换硬盘
  - 准确率 >85%
  - 极大节省带宽和计算资源
- **正在IT、发电机组、发动机组、汽车制造和基站等的智能监控与运维方面开展合作**



医疗领域——百度医疗大脑

- 借助百度大数据引擎平台，及其云计算、大数据、人工智能等核心技术，构建“百度医疗大脑”
- **与“祥云”医疗集团、政府机构等开展合作**

个人健康跟踪与评估

最佳临床治疗方案选择

百度医疗大脑

传染病监测预警

重大慢性疾病预防控制

疾病分析模型 | 治疗路径分析 | 药用药效分析 | 健康指标特征 | 疾病人群画像 | 传播感染预测 | …

临床电子病历 | 可穿戴设备数据 | 健康体检数据 | 临床专家经验 | 疾病科研成果 | 公共健康数据 | …

金融领域

- 基础架构
  - 分布式存储
    - 结构化、非结构化数据存储
  - 分布式数据仓库
    - 超大规模异构数据查询
    - SQL查询语言
    - 多维分析
    - 高可用性、高并发、低延迟
- 数据智能分析
  - 人群分析
    - 品牌用户的目标人群及人群特征
  - 品牌分析
    - 用户对产品的评价、关注度及同业的对比
  - 媒体分析
    - 了解用户常访问的媒体，有利于广告投放
- 数据可视化
  - 交互式体验
  - 有利于用户理解和分析数据
- **与银行开展合作**

24

大数据引擎逐渐成为推动行业升级的重要力量

工业领域

- 工业增质提效转型动力
- 有效分配资源、提高产能、产业链升级
- 硬盘预测，发动机组、汽车制造和基站故障等

医疗领域

- 医疗数据激增，大数据应用开始布局
- 整合临床、健康、公共卫生数据，改良科研、疾病控制、临床支持和重大预警

金融领域

- 信息化程度高，率先实践
- 消费者洞察，个人定制，品牌分析，改良现有产品形态

百度的大数据实践奠定了行业数据应用的基础，行业意识到数据挖掘的价值，大数据引擎逐渐成为行业升级助推器


谢谢！