



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



一、百度技术概览

二、百度大数据实践

三、百度大数据引擎

四、大数据引擎助力产业升级

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。

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

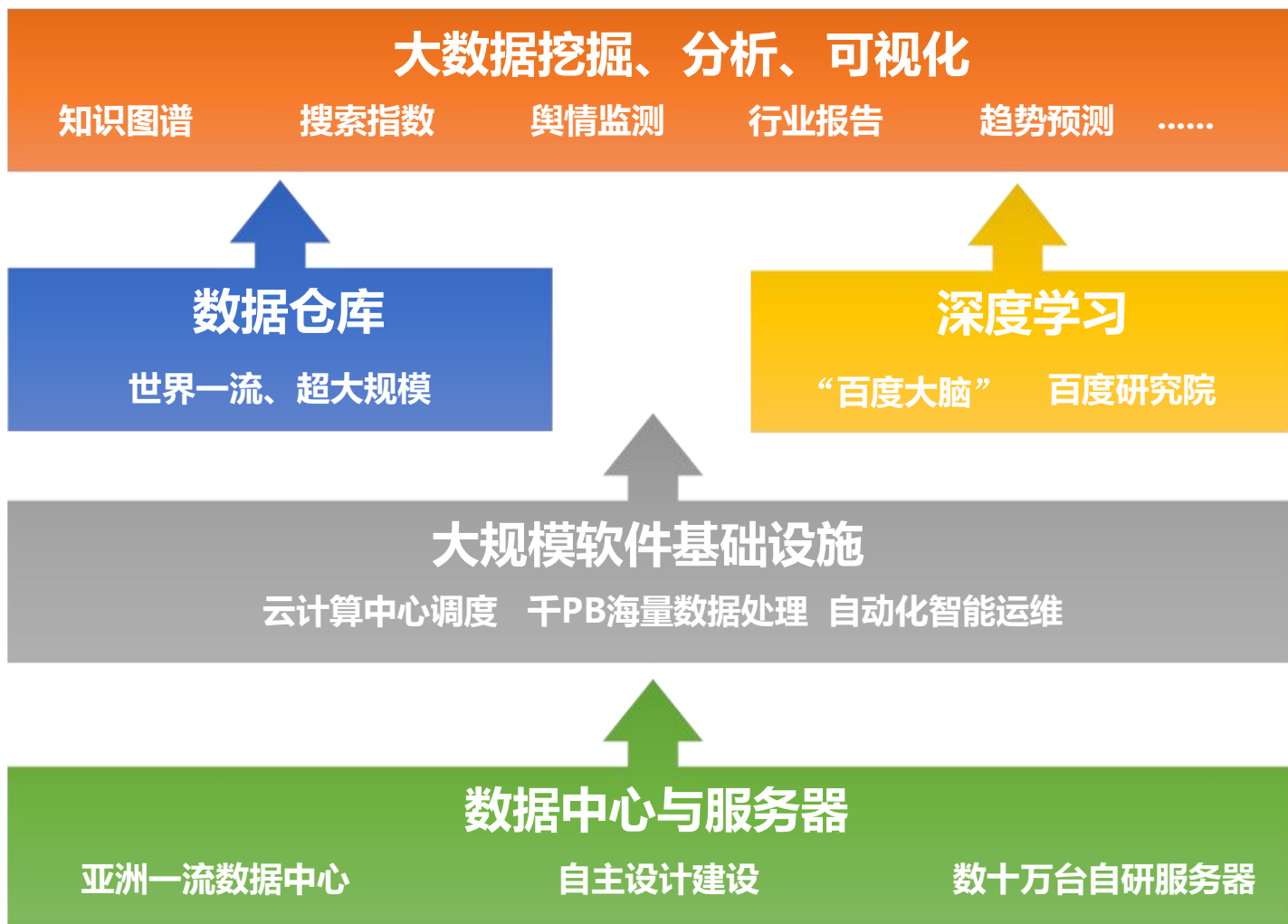
Velocity

- 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



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**大规模部署
- 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**

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

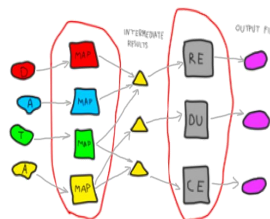
统一资源池管理

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

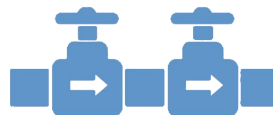
样本/特征达千亿

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

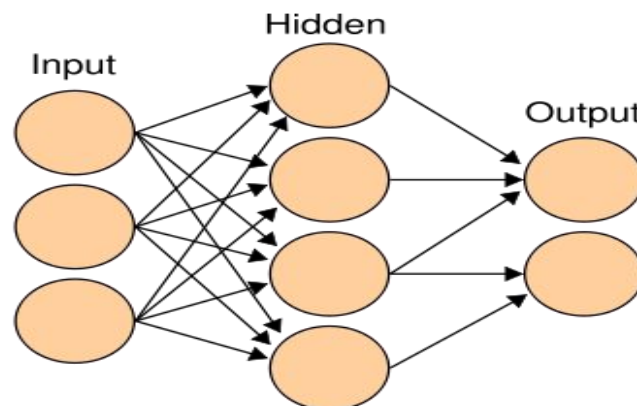
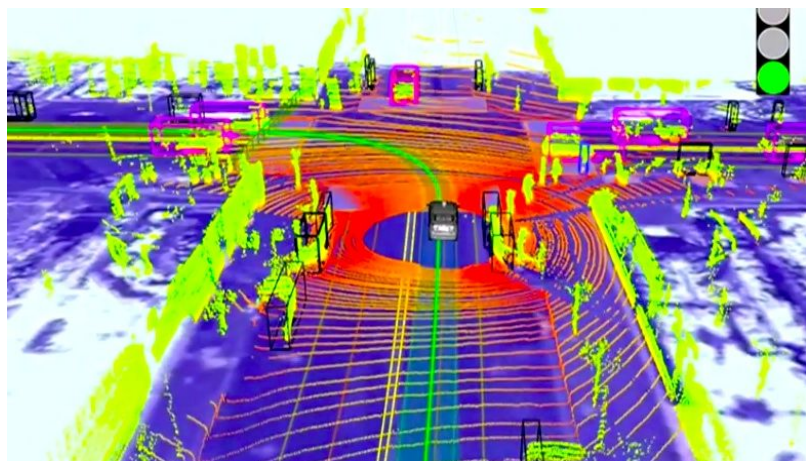


人工智能及深度学习技术

- 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



一、百度技术概览

二、百度大数据实践

三、百度大数据引擎

四、大数据引擎助力产业升级

• 理解自然语言文本，推送精准答案

Baidu 百度 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

182以上的篮球明星

百度一下

182以上的篮球明星 (1511个) :

性别: [全部](#) [男](#) [女](#)



勒布朗詹...	乔丹	林书豪	姚明	安东尼戴...	雷阿伦
202cm	198cm	191cm	229cm	208cm	196cm



易建联	詹姆斯哈登	西热力江	莫里斯	史蒂夫纳什	卡尔马龙
211cm	196cm	191cm	211cm	191cm	210cm

1 2 3 4 5 ... 64 下一页

报错

[身高182以上的韩国男明星。 - 心灵驿站 - 丝路花语 - 新疆最大...](#)

9条回复 - 发帖时间: 2006年8月13日

丝路花语 论坛《天山脚下》心灵驿站 身高182以上的韩国男明星。12下一页 返回列表 发新帖查看: 4275|回复: 10 身高182以上的韩国男明星。...

bbs.xj163.cn/thread-5... 2006-08-21 - 百度快照 - 评价

Baidu 百度 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

谢霆锋的爸爸的儿子的前妻的年龄

百度一下



谢霆锋的爸爸的儿子的前妻的年龄:

34岁

张柏芝 (Cecilia Cheung), 中国香港著名女演员、歌手。出生于香港油麻地。1998年被周星驰发掘, 成为《喜剧之王》的女主角, 由此一炮而红。因其形象秀丽秀美, 有着四分之一... [详情>>](#)

来自百度百科 | 报错

推理过程:



百度知识图谱真心强大 谢霆锋的前妻的儿子的爸爸的年龄



2013年8月22日 - 百度知识图谱真心强大 谢霆锋的前妻的儿子的爸爸的年龄 @康斯坦丁:2013年百度世界大会开的如火如荼,知识图谱正悄然上线,下一代搜索引擎雏形曝光...

www.qiyue.com/yule/ht... 2013-08-22 - 百度快照

JUMP杂志的官方中文网站

网站导航：多语言选择 | 订阅指南 | **最新连载**

[日漫首页](#) |
 [**航海王**](#) |
 [火影忍者](#) |
 [境·界](#) |
 [银魂](#) |
 [阿拉蕾](#) |
 [家庭教师](#)
[圣斗士星矢](#) |
[龙珠](#) |
[网球王子](#) |
[海贼王](#) |
[游戏王](#)

航海王
第716话更新啦

人气：151988337 分享到：

航海王简介

剧情梗概：获得12位皇族御守

赏格：两亿

章节列表

JUMP杂志最新连载

task_name: comic_debug_disam(1328)			
Entity Name		Entity URI	
周淑贞 4601		comic.basic.0011158.1373525136.0	
Properties			
Property Name	Property Value		Property Misc
rating	98		{ "value_index": "969"
ch	comic		{ "path": "Vrsp/V"
gatherrurl	http://v.baidu.com/v?word=%E8%A3%D4%F4%CD%F5		{ "value_index": "969"
trunk	周淑贞		{ "value_index": "969"
sites__split	[{"site_logo": " http://list.video.baidu.com/logo/sohu.gif ", "obj_url": " http://open.video.baidu.com:8080/api/video?id=4601&id=comic&site=sohu&com=f&stoken=a34bc81981ad8e3af62248a59f691657 ", "site_name": "搜狐", "site_snm": ""}]		{ "value_index": "03"
area	"日本"		{ "value_index": "969"
intro	"ONE PIECE"在故事中的“一个大大宝”之是故事裡男主角“蒙奇·D·路飞”为了当上“海贼王”而踏上“伟大航道”后所经历的一系列“冒险”“旅程”“冒险”中，路飞和他的伙伴们经历了具有挑战、成长力量世界第一的宝藏“ONE PIECE”，许多人为了解“ONE PIECE”，争相比高，许多国家开始被立宪而立，而形成了“大航海时代”。于是路飞为了要展现与同伴他而航海的回忆，让发 竟会到那的途中而在，在这途的路上也获得许多珍贵的宝物。于是路飞和伙伴们一起踏上了“伟大航道”。		{ "value_index": "969"

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

语音

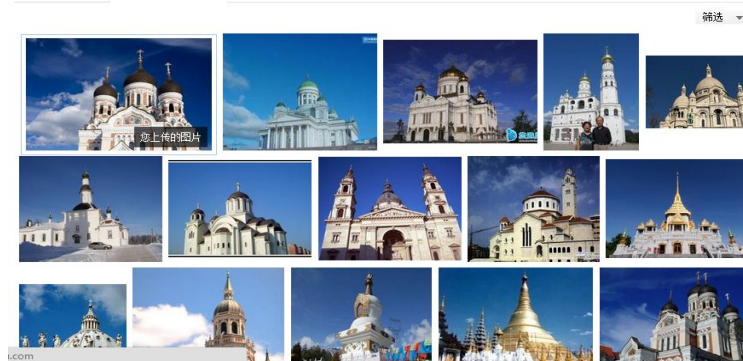


图像

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

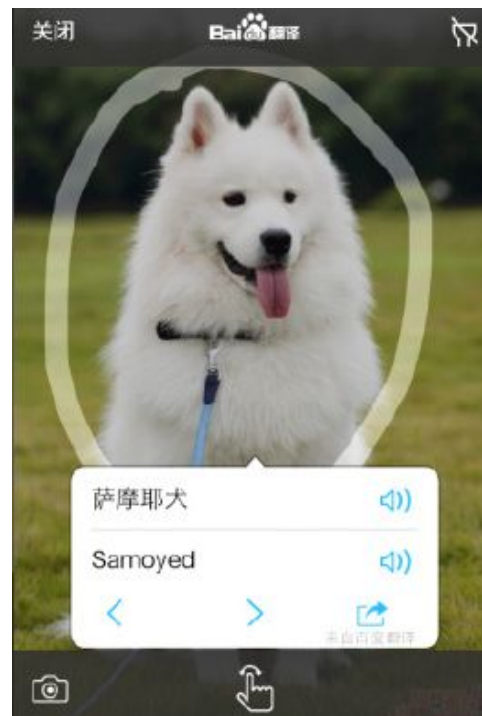


• 最懂中文的机器翻译

- 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



景点预测

- **未来2日**的拥挤及舒适度预测

城市预测

- **未来2周**内人数规模预测

高考预测

- **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**

世界杯预测

- 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**

疾病预测

- **4种疾病**：流感、肝炎、肺结核、性病
- **未来6天**内发病指数预测

经济指数预测

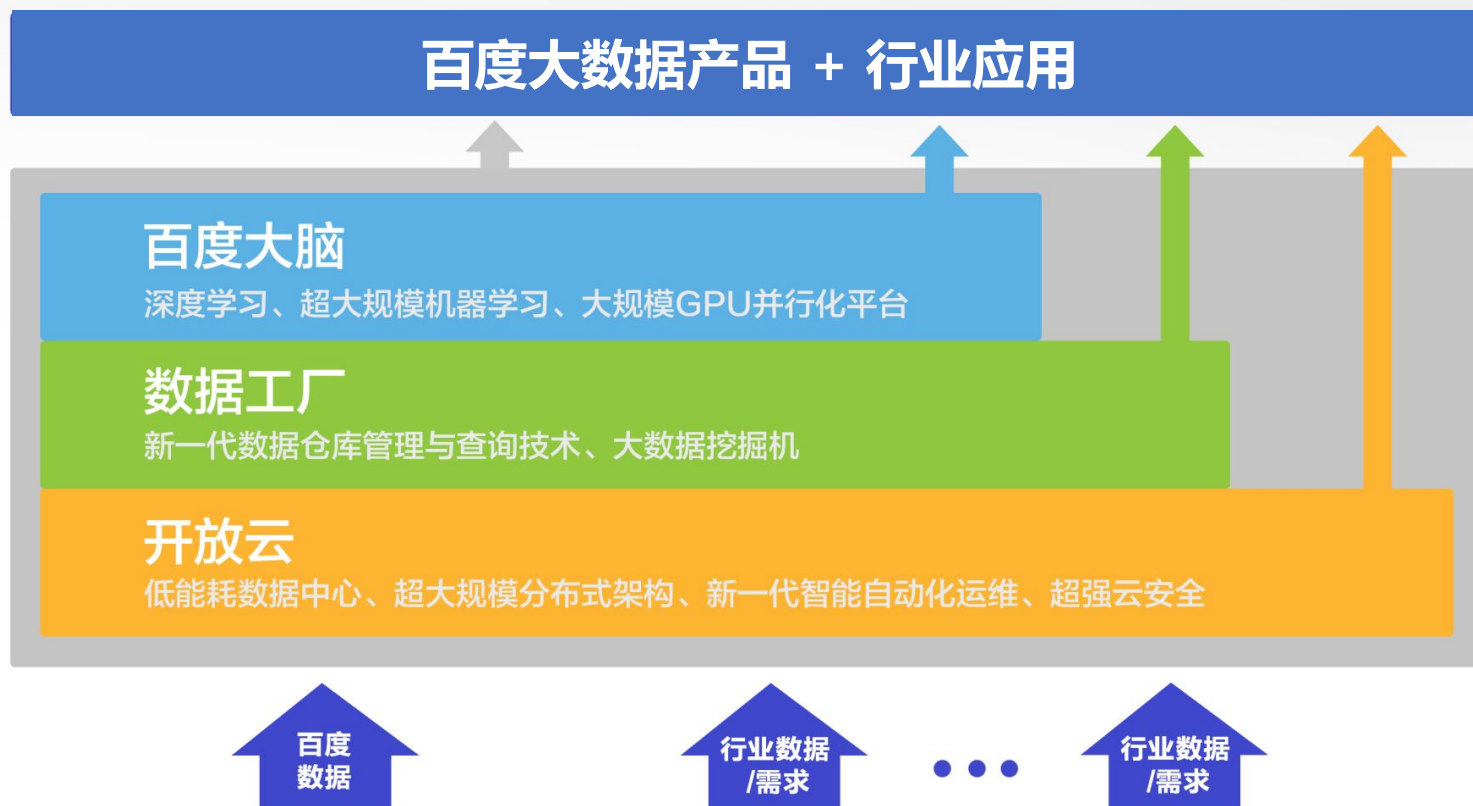
- **中小企业**景气指数
- **宏观经济**指数预测

一、百度技术概览

二、百度大数据实践

三、百度大数据引擎

四、大数据引擎助力产业升级



三级
开放
平台

超大规模存储

拥有并管理数据达
EB级别

单集群
百亿文件数

数十万
台服务器规模

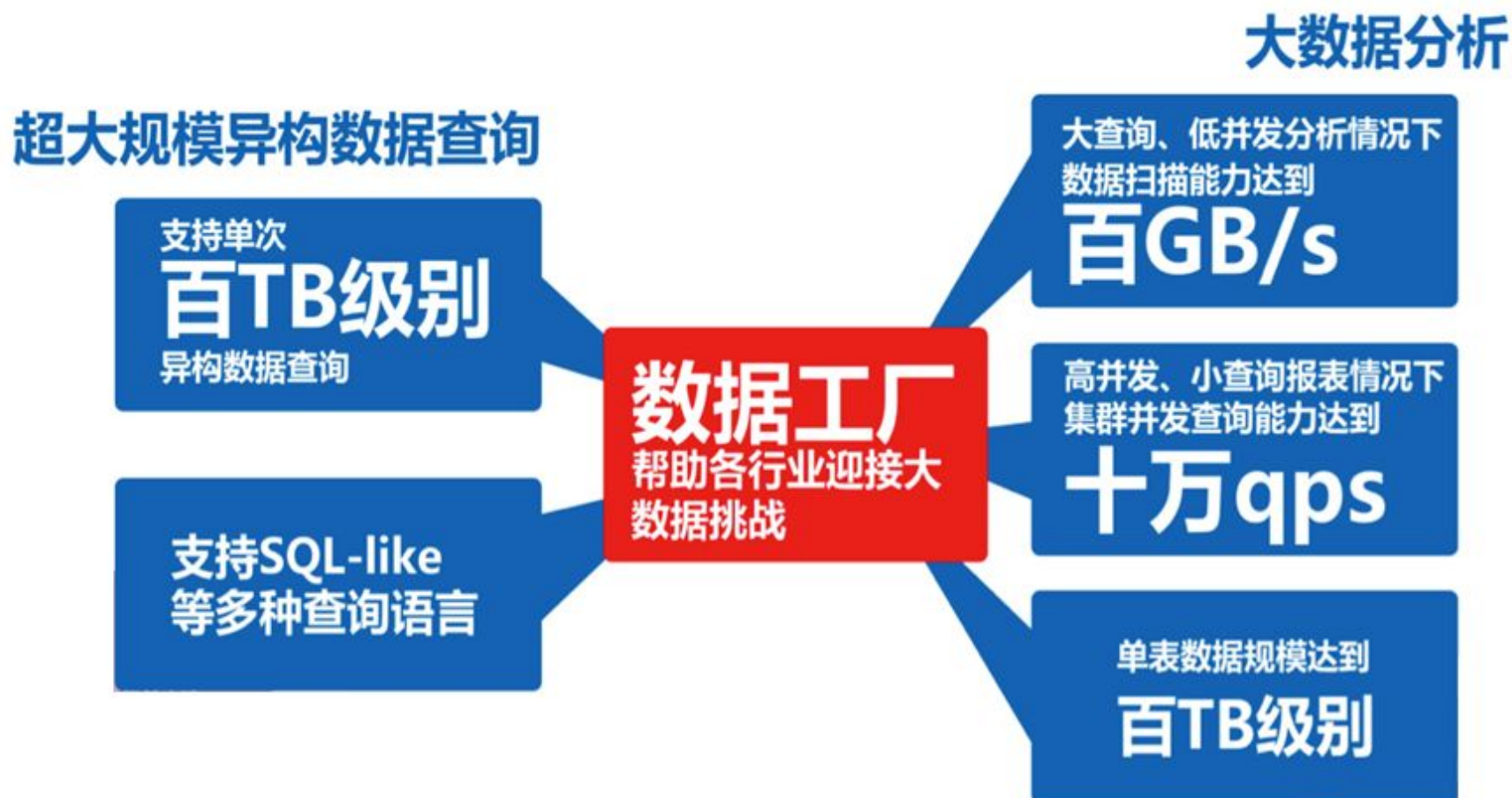
开放云

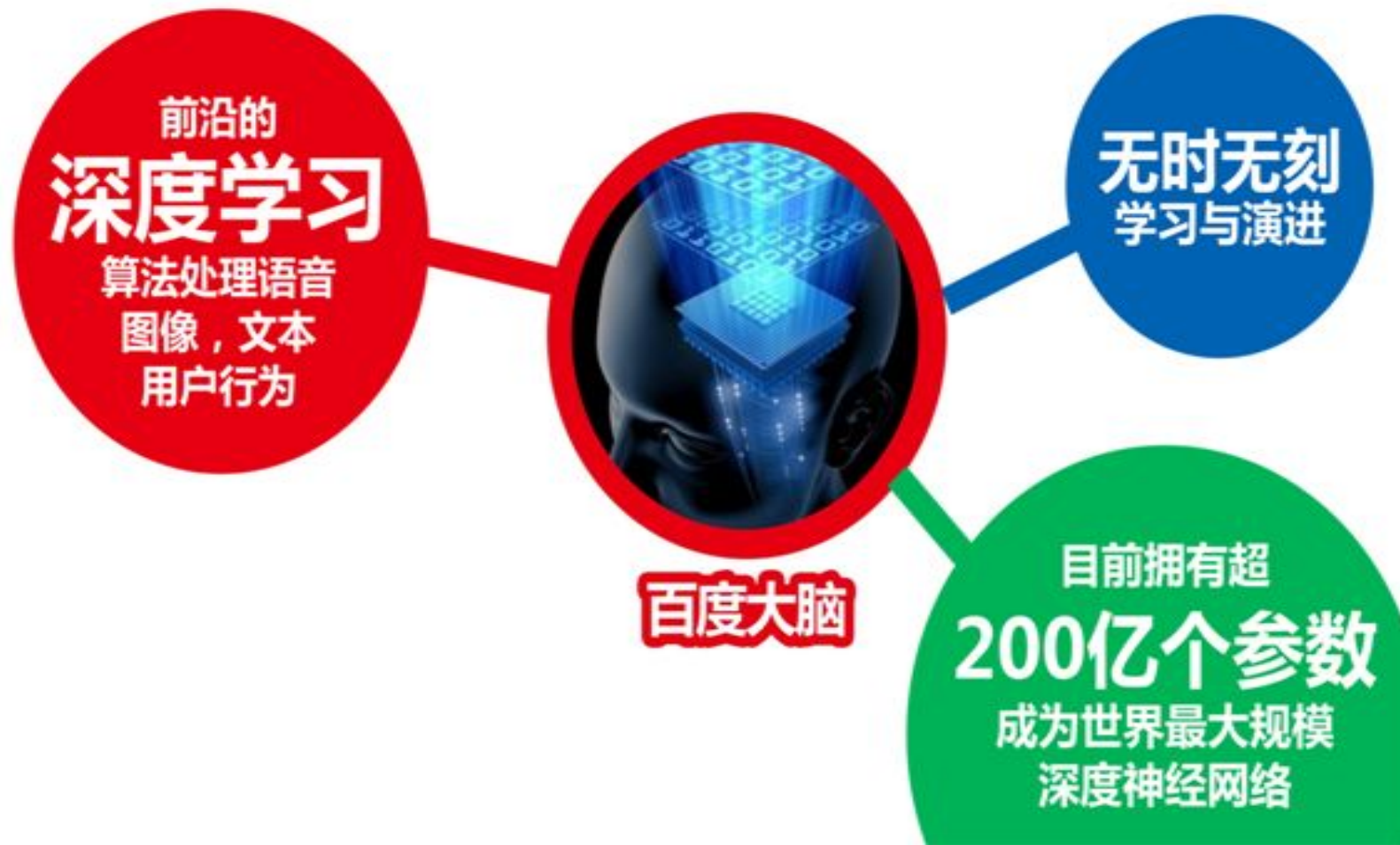
分布式计算

数据处理能力
100GB/s
毫秒级响应

单集群离线计算规模超
数十万台，达到
上百PB/天处理能力

单集群CPU利用率近
90%





一、百度技术概览

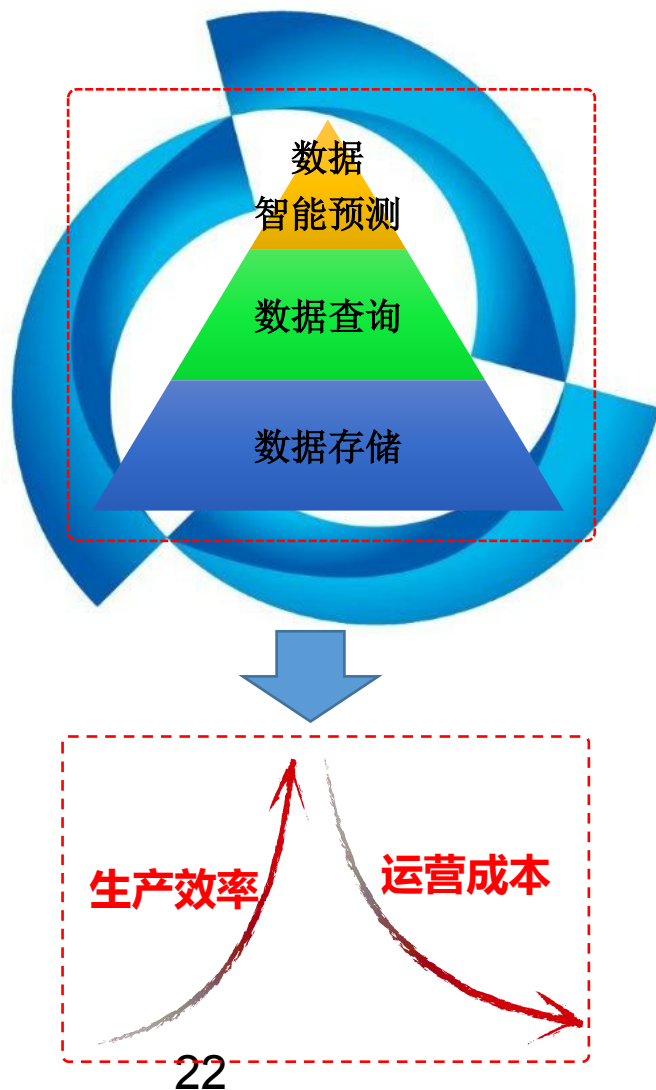
二、百度大数据实践

三、百度大数据引擎

四、大数据引擎助力产业升级

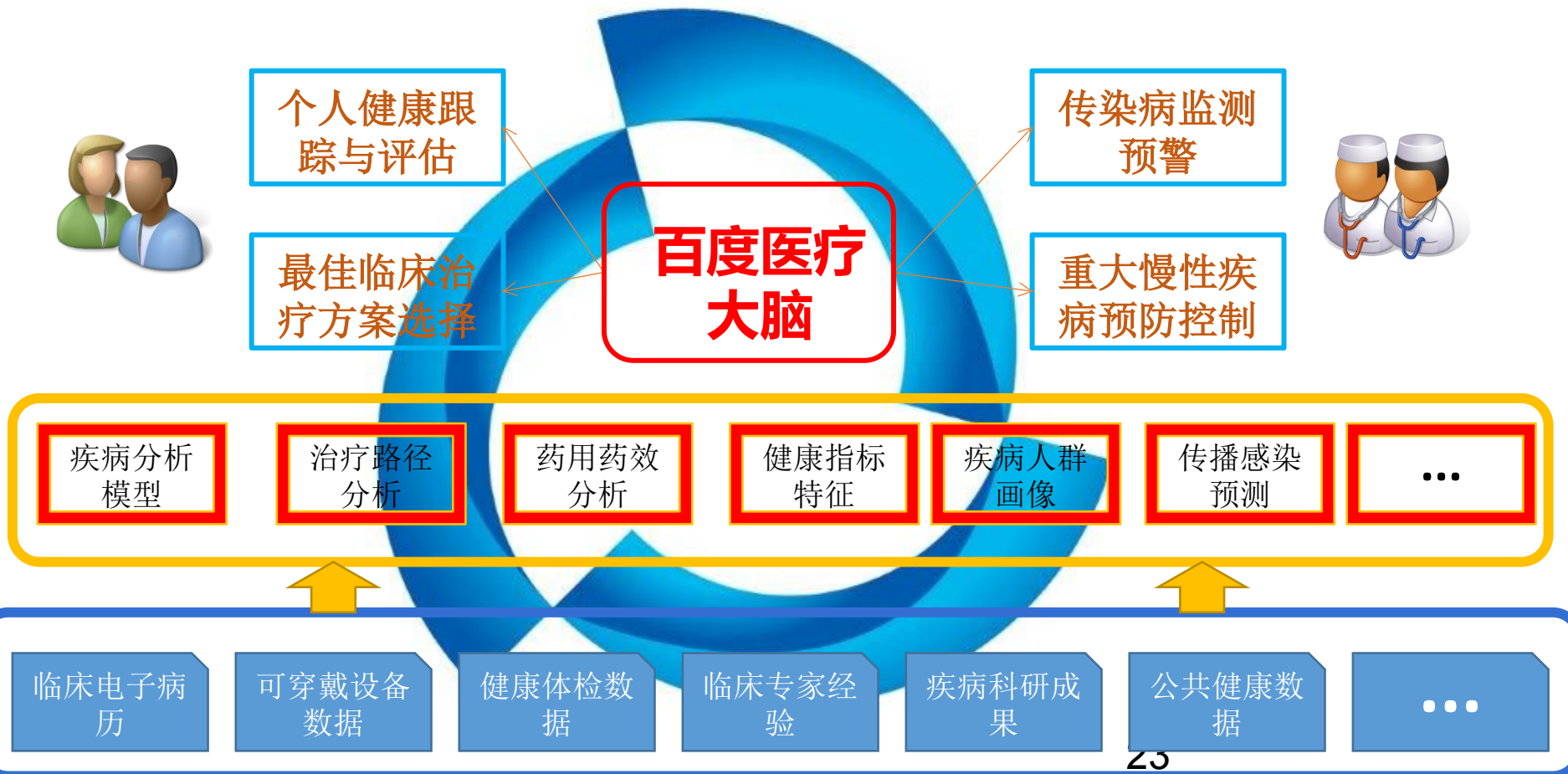
- 百度基础设施
 - 200+万块硬盘
 - 硬件故障率：硬盘>内存>电源
- 预测故障，提前拷贝数据，更换硬盘
 - 百度经过对近9亿条实例进行采集处理
 - 选取15万个训练样本
 - 从历史的硬盘故障病例中，选取了240个特征
 - 人工智能模型预测硬盘故障
- 对比无预测的故障恢复的好处
 - 提前一天预测出硬盘故障
 - 提前拷贝数据，更换硬盘
 - 准确率 > 85%
 - 极大节省带宽和计算资源
- 正在IT、发电机组、发动机组、汽车制造和基站等的智能监控与运维方面开展合作

百度大数据引擎



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



- 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



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！