

大数据编程

1-1

大数据概述

中央财经大学 商学院
姚凯
2016

课程简介



课程简介

- ❖ 大数据的基本概念
- ❖ 大数据的核心技术
- ❖ 基于大数据的商业应用
- ❖ 大数据发展过程中的问题

课程目的

- ❖ 了解大数据的基本概念
- ❖ 深入理解大数据在现实生活中的应用
- ❖ 提高实际动手能力，掌握常用大数据编程技术
- ❖ 锻炼解决问题和独立思考的能力

课程内容

第一周	大数据概述		第十周	并行计算和云计算
第二周	大数据应用技术		第十一周	并行编程
第三周	数据获取方法		第十二周	线程同步
第四周	编写数据爬虫		第十三周	分布式计算
第五周	互联网广告		第十四周	海量数据处理实验
第六周	精准广告投放算法		第十五周	Hadoop
第七周	个性化推荐应用		第十六周	Hadoop平台下的大数据分析工具
第八周	个性化推荐应用		第十七周	成果展示
第九周	小组		第十八周	期末考试

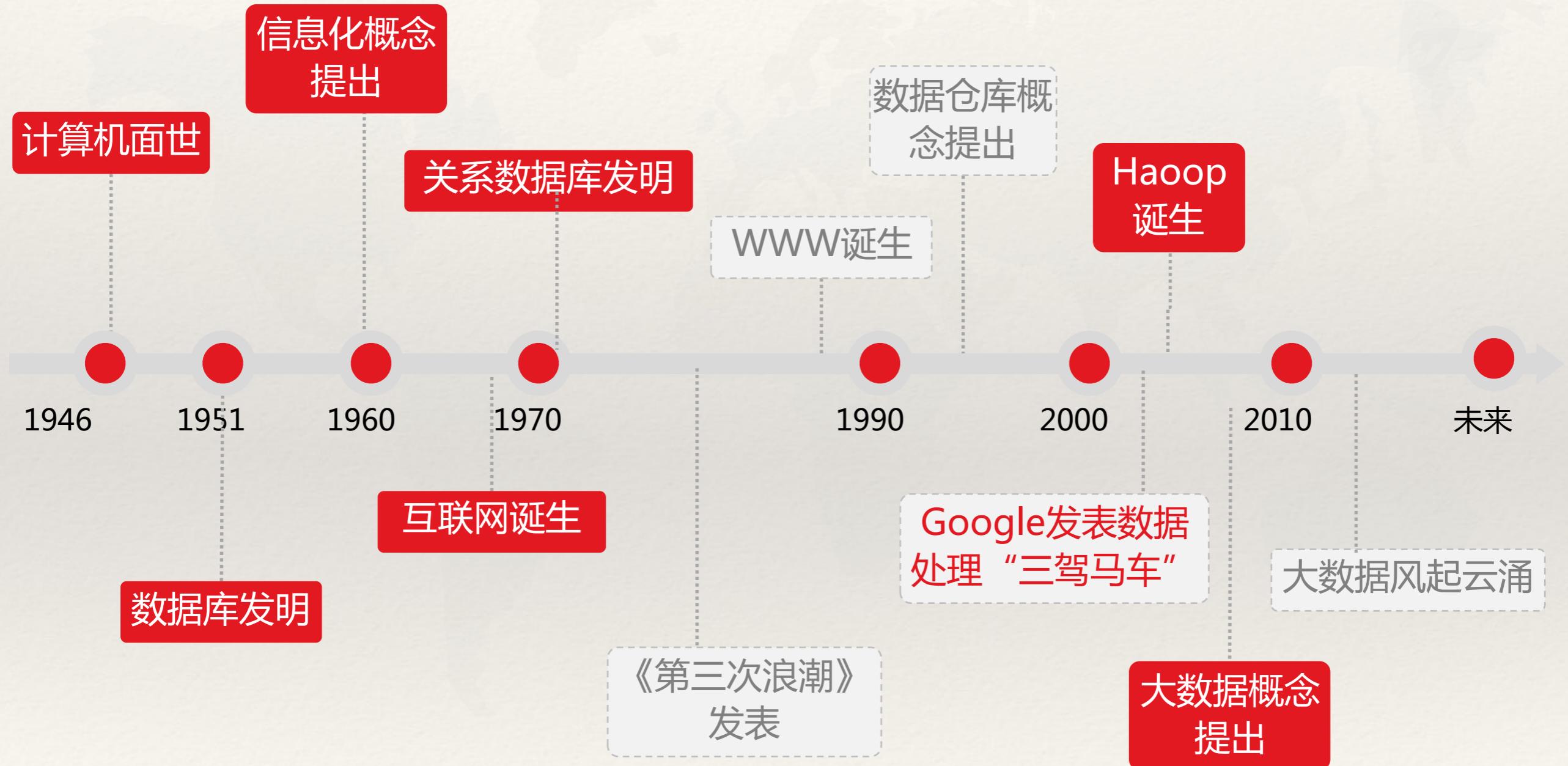
个人简介

- ❖ 姚凯
- ❖ 计算机本科、硕士
- ❖ 市场营销博士
- ❖ yaokai@cufe.edu.cn
- ❖ 课程资料：<https://github.com/jasonyaopku/BigDataProgramming>

主要内容

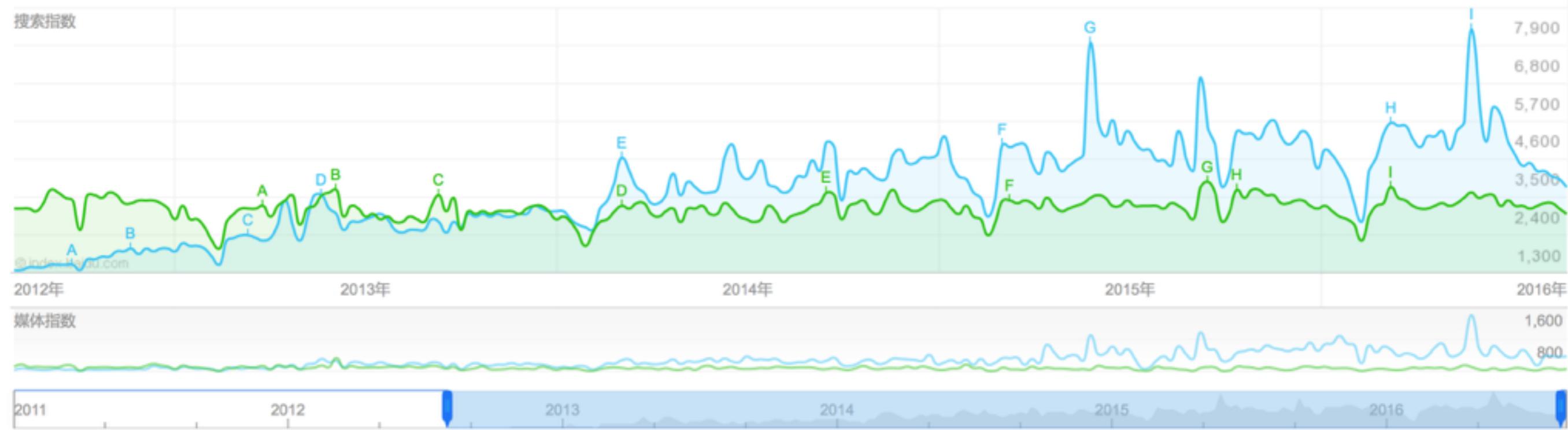
- ❖ 什么是大数据?
- ❖ 大数据特点
- ❖ 大数据应用
- ❖ 大数据实例
- ❖ 大数据面临的挑战

大数据时代大事记

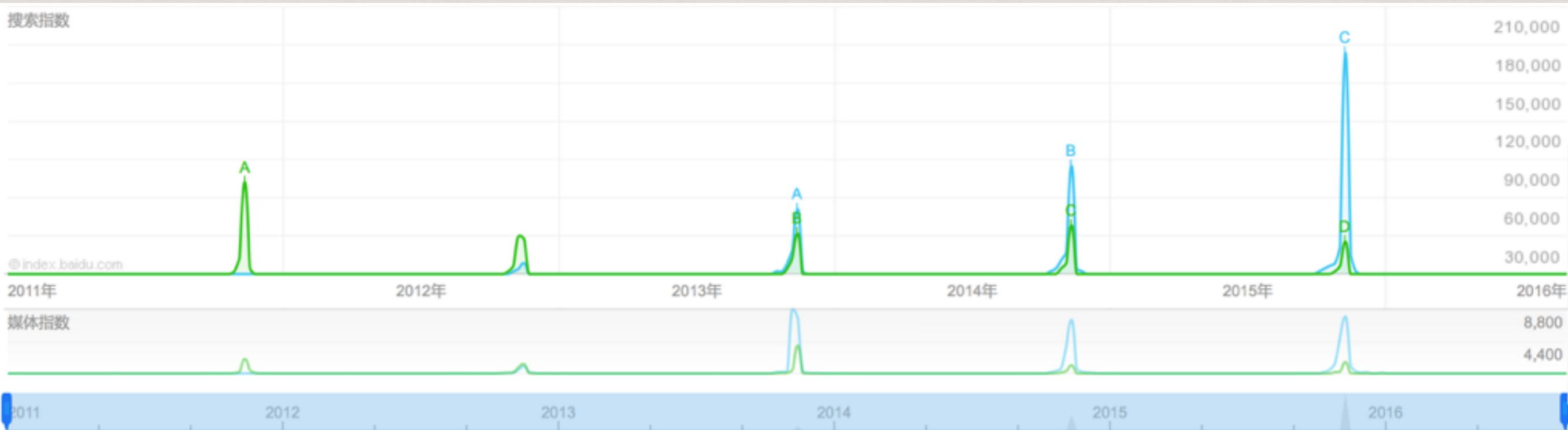
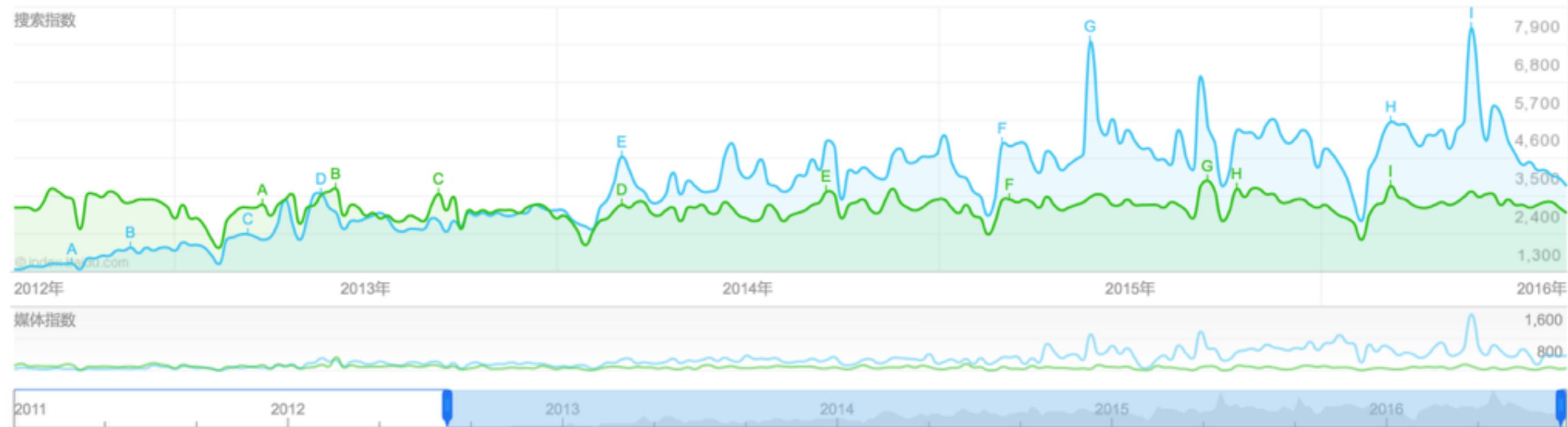


大数据发展

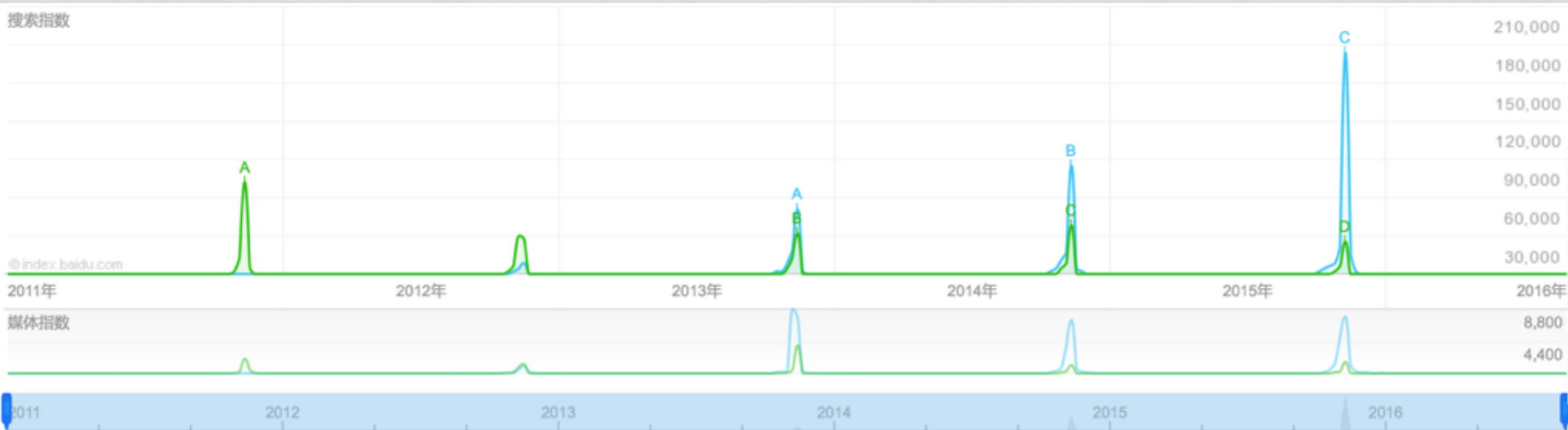
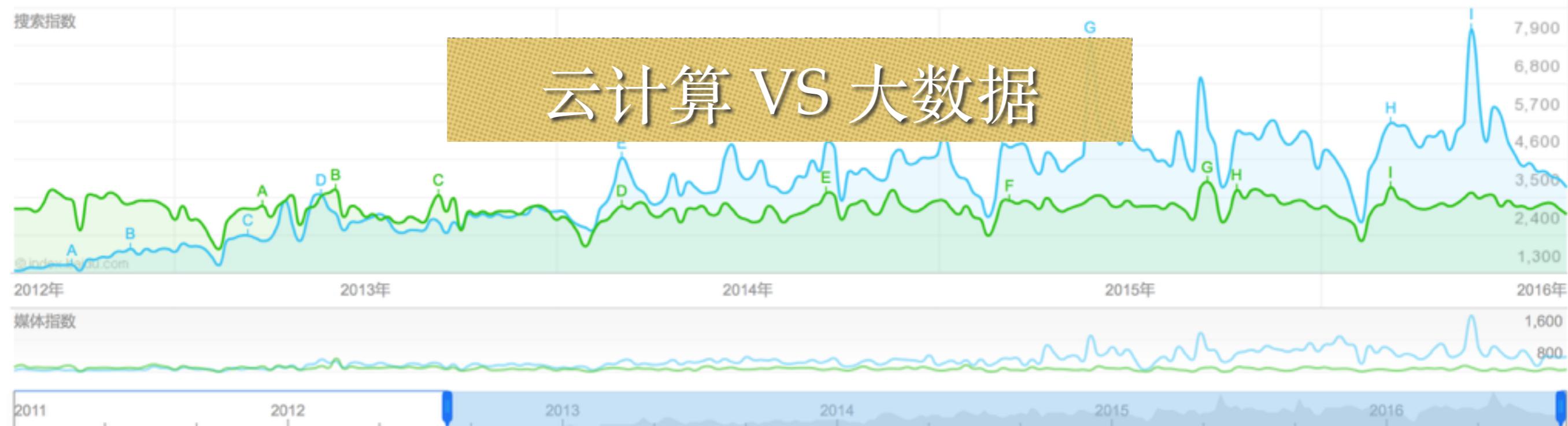
大数据发展



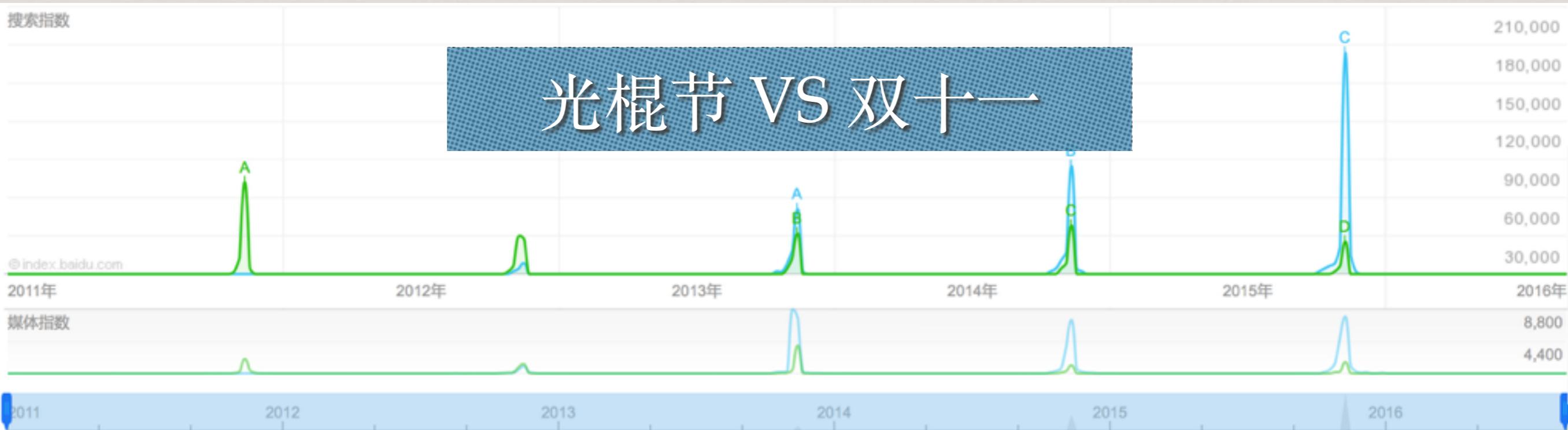
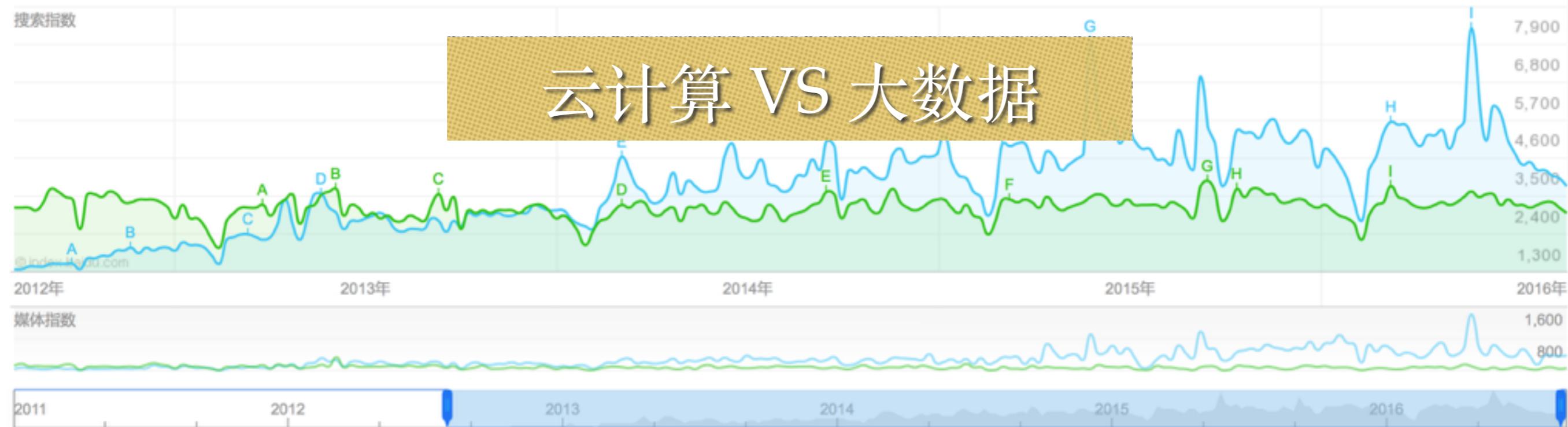
大数据发展



大数据发展



大数据发展



大数据是人类发展的必由之路

大数据

信息化的自然延伸

存储与计算资源极大
丰富的自然结果

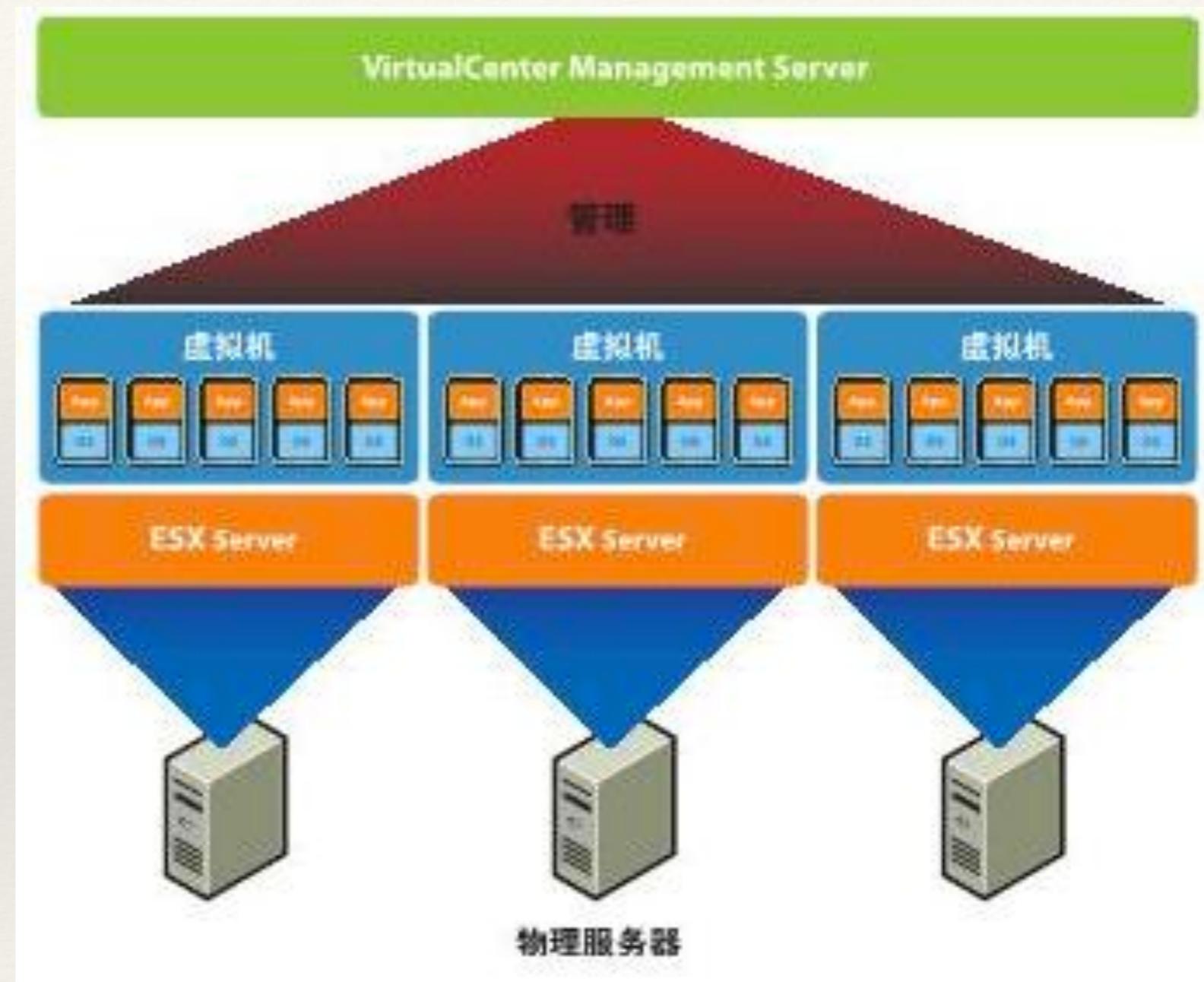
人类大一统追求的自
然体现

数据记录方式

	账本	交易系统	在线交易系统	记录一切信息
便捷		★	★★	★★★
规范		★	★★	★★★
容量	★	★★	★★	★★★
保存	★	★★	★★	★★★
交换	★	★	★★★	★★★
查询		★	★★	★★★
分析		★	★★	★★★
挖掘			★	★★★

大数据相关概念：虚拟化

虚拟化是一种资源管理技术，是将计算机的各种实体资源，如服务器、网络、内存及存储等，予以抽象、转换后呈现出来，打破实体结构间的不可切割的障碍，使用户可以比原本的组态更好的方式来应用这些资源



大数据相关概念：虚拟化

虚拟硬件

虚拟机

虚拟内存

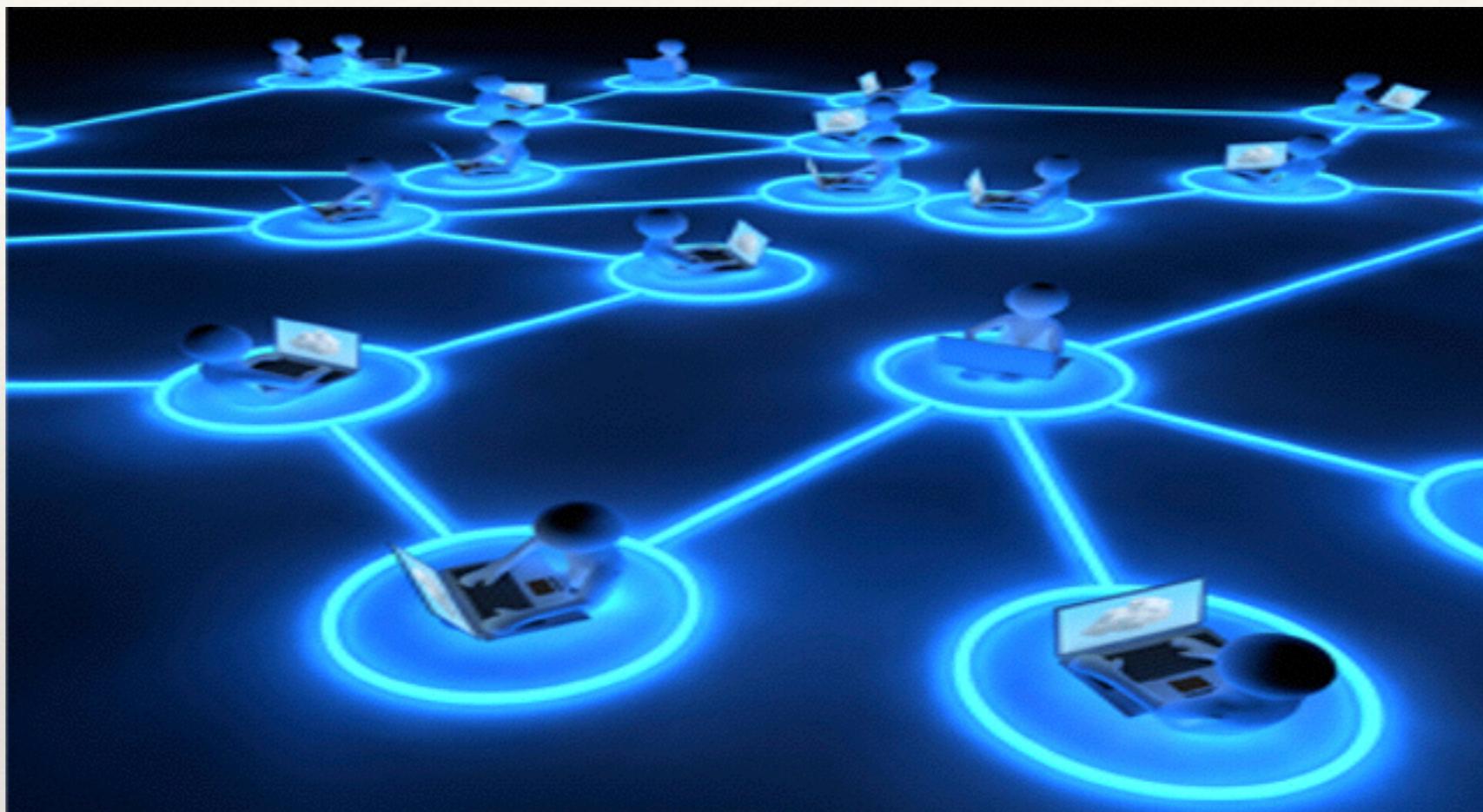
虚拟桌面

虚拟网络

虚拟存储

大数据相关概念：分布式计算

分布式计算是将一个计算任务分成许多小的部分，然后把这些部分分配给许多计算机进行处理，最后把这些计算结果综合起来得到最终的结果的计算过程



大数据相关概念：分布式计算

分布式
存储

分布式
计算模型

分布式
系统

分布式
算法

大数据相关概念：云计算

云计算是一种按使用量付费的服务模式，这种模式提供可用的、便捷的、按需的网络访问，进入可配置的计算资源共享池（资源包括网络，服务器，存储，应用软件，服务），这些资源能够被快速提供，只需投入很少的管理工作，或与服务供应商进行很少的交互。



云计算类型

SaaS (Software-as-a-Service): 软件即服务，厂商将应用软件统一部署在自己的服务器上，客户可以根据自己实际需求，通过互联网向厂商定购所需的应用软件服

PaaS (Platform-as-a-Service): 平台即服务，把服务器平台作为一种服务提供的商业模式，把客户开发的或收购的应用程序部署到供应商的云计算基础设施上去

IaaS (Infrastructure-as-a-Service): 基础设施即服务，消费者通过 Internet 可以从完善的计算机基础设施获得服务

云计算：分类

SaaS

Blog

邮件

存储

其他

PaaS

Web

分析

通信

其他

IaaS

虚拟机

存储

运维

安全

云计算类型

SaaS (Software-as-a-Service): 软件即服务，厂商将应用软件统一部署在自己的服务器上，客户可以根据自己实际需求，通过互联网向厂商定购所需的应用软件服

PaaS (Platform-as-a-Service): 平台即服务，把服务器平台作为一种服务提供的商业模式，把客户开发的或收购的应用程序部署到供应商的云计算基础设施上去

IaaS (Infrastructure-as-a-Service): 基础设施即服务，消费者通过 Internet 可以从完善的计算机基础设施获得服务

什么是大数据

什么是大数据

大数据指的是所涉及的数据量规模巨大到无法通过传统方式，在合理时间内达到截取、管理、处理、并整理成为人类所能解读的信息。
《维基百科》

什么是大数据

大数据指的是所涉及的数据量规模巨大到无法通过传统方式，在合理时间内达到截取、管理、处理、并整理成为人类所能解读的信息。
《维基百科》

大数据一般会涉及 2 种或 2 种以上数据形式。它要收集超过 100TB 的数据，并且是高速、实时数据流；或者是从小数据开始，但数据每年会增长 60% 以上。

IDC

什么是大数据

大数据指的是所涉及的数据量规模巨大到无法通过传统方式，在合理时间内达到截取、管理、处理、并整理成为人类所能解读的信息。
《维基百科》

大数据一般会涉及 2 种或 2 种以上数据形式。它要收集超过 100TB 的数据，并且是高速、实时数据流；或者是从小数据开始，但数据每年会增长 60% 以上。

IDC

大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

Gartner

主要内容

- ❖ 什么是大数据?
- ❖ 大数据特点
- ❖ 大数据应用
- ❖ 大数据面临的挑战

身边的大数据

Google

通过用户行为预测他们喜欢什么广告

Apple Siri

与用户交互，理解语音并与其他应用联动

百度地图

多种多样的导航模式，以及本地服务推荐

**Eugene
Goostman**

让许多人认为它是13岁小男孩，第一个通过图灵测试的程序

人脸识别

“高斯”系统比人更准确的识别人脸

淘宝推荐引擎

精准预测用户购物和阅读偏好

大数据的特点

大数据的特点

Volume 体量

大数据的特点

Volume 体量

Velocity 速度

大数据的特点

Volume 体量

Velocity 速度

Variety 多样

Volume 体量

1 Byte = 8 bit

1 EB = 1,024 PB = 1,048,576 TB

1 KB = 1,024 Bytes = 8192 bit

1 ZB = 1,024 EB = 1,048,576 PB

1 MB = 1,024 KB = 1,048,576 Bytes

1 YB = 1,024 ZB = 1,048,576 EB

1 GB = 1,024 MB = 1,048,576 KB

1 BB = 1,024 YB = 1,048,576 ZB

1 TB = 1,024 GB = 1,048,576 MB

1 NB = 1,024 BB = 1,048,576 YB

1 PB = 1,024 TB = 1,048,576 GB

1 DB = 1,024 NB = 1,048,576 BB

Volume 体量



Volume 体量

- ❖ 互联网络的广泛应用, 使用网络的人、企业、机构增多, 数据获取、分享变得相对容易
- ❖ 各种传感器数据获取能力的大幅提高, 使得人们获取的数据越来越接近原始事物本身
- ❖ 早期网络上的数据以文本和一维的音频为主、维度低、单位数据量小。近年来, 图像、视频等数据占据主要带宽。
- ❖ 数据量大还体现在人们处理数据的方法和理念发生了根本的改变。

Velocity 速度

- ❖ 数据的产生、发布越来越容易,产生数据的途径增多
- ❖ UGC (User Generated Content)的流行加快了数据产生速度
- ❖ 数据的价值是随着时间的推移 而迅速降低的,如果数据尚未得到有效的处理,就失去了价值,大量的数据就没有意义。
- ❖ 对不断激增的海量数据数据的实 时处理要求,是大数据与传统海
量数据处理技术的关键差别之一。

Variety 多样

- ❖ 以往的数据尽管数量庞大,但通常是事先定义好的结构化数据。便于人类和计算机存储、处理、查询的方向抽象的结果。
- ❖ 随着互联网和软硬件的飞速发展, 非结构化数据大量涌现, 非结构化数据没有统一的结构属性
- ❖ 数据分析需求和技术的发展, 使得数据多样性提高
- ❖ 海量数据、高纬度能够帮助大数据达到小数据无法达到的应用效果

大数据的特点

Volume 体量

Velocity 速度

Variety 多样

大数据的特点

Value (价值)

Veracity (真实性)

可能的误区

1. 我们用几千维的信息来描述消费者特征，模型匹配度特别高
2. 我们有十几亿互联网用户数据，来跟我们合作吧
3. 我们用非常复杂的模型，利用计算机集群为消费者提供更加优质的服务

小结

- ❖ 大数据概述
 - ❖ 虚拟化
 - ❖ 分布式计算
 - ❖ 云计算
- ❖ 大数据特点
 - ❖ 3V+2V

大数据编程

1-2

大数据概述

中央财经大学 商学院
姚凯
2016

回顾

- ❖ 虚拟化
- ❖ 分布式
- ❖ 云计算及主要内容
- ❖ 3V
- ❖ 2V

主要内容

- ❖ 什么是大数据?
- ❖ 大数据特点
- ❖ 大数据应用
- ❖ 大数据实例
- ❖ 大数据面临的挑战

大数据的应用

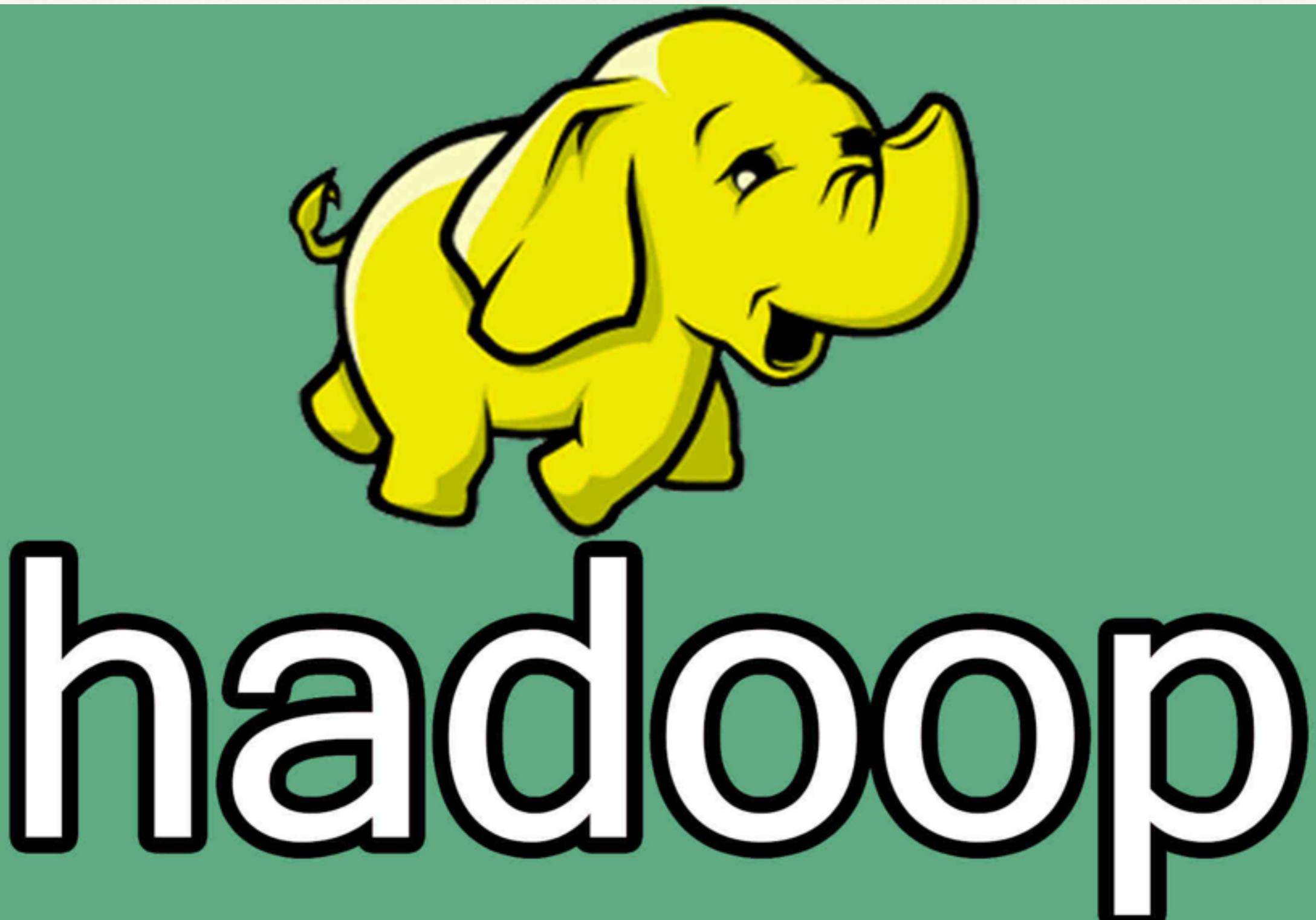
大数据的应用

- ❖ 分布式框架
- ❖ 个性化商品推荐
- ❖ 联合分析
- ❖ 精准广告投放

大数据的应用

- ❖ 分布式框架
- ❖ 个性化商品推荐
- ❖ 联合分析
- ❖ 精准广告投放
- ❖ 客户终身价值
- ❖ 个人征信
- ❖ 搜索引擎
- ❖ 时空大数据

Hadoop分布式框架



Hadoop分布式框架



个性化商品推荐

Amazon Try Prime All - iwatch

Hello, Jasonya .. Your Account - Try Prime student

Departments - Browsing History - jasonya..'s Amazon.com Today's Deals Gift Cards & Registry Sell Help

Cell Phones & Accessories Carrier Phones Unlocked Phones Prime Exclusive Phones Accessories Cases Wearable Technology Best Sellers Deals Trade-In All Electronics

Prime student FREE Two-Day Shipping for college students Learn more +

Back to search results for "iwatch"

Apple Watch Sport, Space Grey Aluminum Case/Black Band, 42mm by Apple ★★★★☆ 718 customer reviews | 381 answered questions

Available from these sellers.

- System on a chip: Apple S1
- Storage: 8 GB
- Initial operating system: watchOS 1.0
- Display: 42 mm: 24.34 30.42 mm; 38.96 mm (1.534 in) diagonally, 312380 pixels, 326 PPI
- Battery: 246mAh

96 new from \$329.99 107 used from \$224.78 19 refurbished from \$249.99

Report incorrect product information.

Important Information: Amazon global delivery to Hangzhou, Huzhou, Jiaxing and Shaoxing, China may experience delays due to the G20 Summit in Hangzhou from 8/20-9/7. [help page](#)



Roll over image to zoom in

Customers Who Bought This Item Also Bought

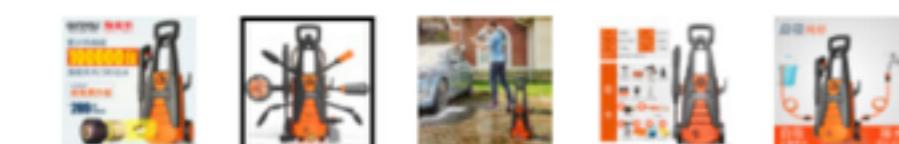
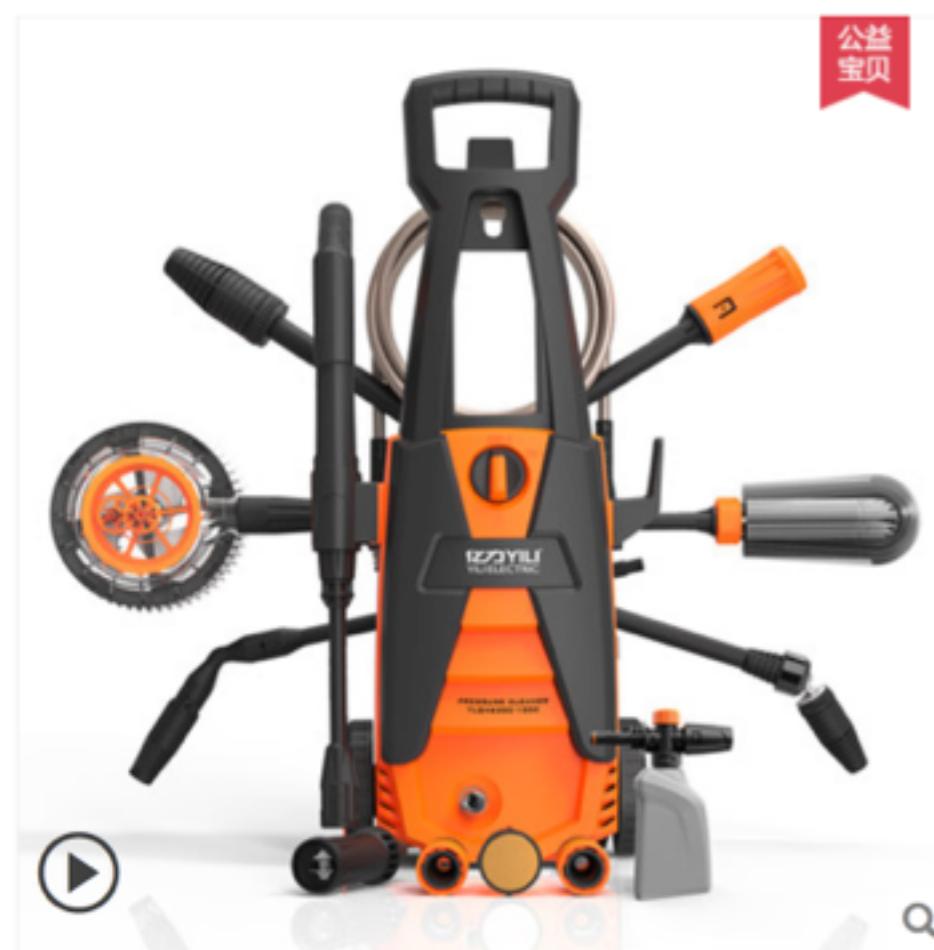
Page 1 of 10

 ArmorSuit MilitaryShield For Apple Watch 42mm Screen Protector [Full Coverage] [2 Pack] Anti... \$21.99 ✓Prime	 Apple Watch Case, SUPCASE [Unicorn Beetle Pro] Rugged Protective Case with Strap Bands... ★★★★☆ 1,301	 Apple Watch Case, Spigen [Rugged Armor] Resilient [Black] - [Include 2 Screen Protectors] Ultimate... ★★★★☆ 1,476	 Apple 1.49-Inch Sport Smart Watch - Rose Gold Aluminum Case with Lavender Band ★★★★☆ 330	 Infland 4007176 Apple Watch 42 mm Band, Infland Stainless Steel Metal Replacement Strap... ★★★★☆ 322	 CATALYST CASE for APPLE WATCH 42mm - STEALTH BLACK ★★★★☆ 151	 Apple Watch Band, Penom Fully Magnetic Closure Clasp Mesh Loop Milanese... ★★★★☆ 3,154	 #1 Apple Watch Stand, Aeris Bamboo Wood Charging Stand Bracket Docking... ★★★★☆ 2,794	 Original Apple Watch 42mm (fits 5.5" - 8.2" wrists) - Silver Aluminum Case... ★★★★☆ 128	 Apple Watch Stand, Aeris Rose Wood Charging Stand Bracket Docking... ★★★★☆ 2,794
Display a menu 843	\$21.99 ✓Prime	\$301.00 ✓Prime	\$19.99 ✓Prime	\$58.00 ✓Prime	\$12.99 ✓Prime	\$299.00 ✓Prime	\$10.99 ✓Prime		

个性化商品推荐



个性化商品推荐



分享 ★ 收藏商品 (74157人气)

举报

服务承诺 正品保证 极速退款 赠运费险
七天无理由退换

支付方式 ▾

亿力高压洗车机家用220v全自动车载洗车器清洗机刷车水泵便携水枪

旗舰系列 免费升级 精英升豪华 咨询客服 抢优惠

天猫 购物券 全天猫实物商品通用

去刮券 >

价格 ￥416.00-877.00

促销价 ￥309.00-859.00

运费 浙江嘉兴 至 北京 快递 0.00

付款后，预计9月4日(周日)送达

月销量 3780

累计评价 37997

送天猫积分 154起

颜色分类



数量

1 ▲ ▼ 件 库存521件

立即购买

加入购物车

看了又看



¥78.00



亿力电
购券
前100名顾客
免费升级
价值499元清洗机

418+
469.00



亿力电
购券
小巧便携
自吸两用

379+
428.00



联合分析

大小

颜色

容量

价格



联合分析



精准广告投放



王宝强要求法院强制马蓉交出儿女：别耽误上学



重庆晨报 08-29 22:11

大数据之“大”的三个具体含义



搜狐网



王健林说这辈子在中国我没对手了，马云说我看未必

搜狐财经 08-30 16:43

实时热点

- 高晓松节目遭停播
- 郑州现山寨凯旋门
- 上海离婚买房疯狂
- 梁家辉被开罚单
- 新车加油打不着火
- 杨洋示范撩妹吻戏
- 云南首育小熊猫
- 房贷利息抵扣个税
- 王健林小目标走红
- 中国残奥团出征
- 男子法院偷回罚金
- 中国驻吉使馆爆炸
- 台风狮子山袭东北 新
- 纪晓岚和珅重聚 新
- 婚期临近酒店关门 新
- 警方退还被盗手机 新

换一换

客户终身价值

CUSTOMER LIFETIME VALUE

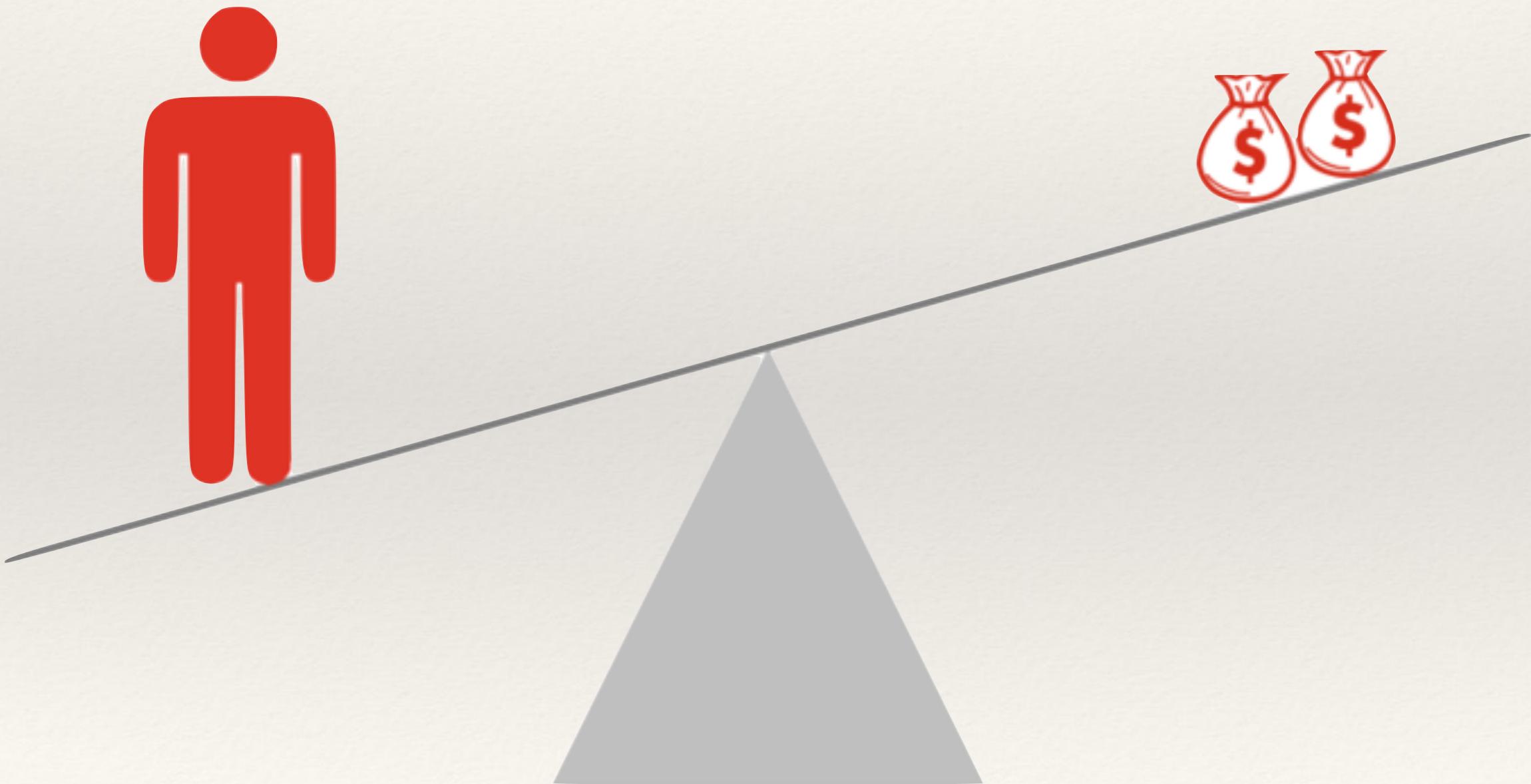


客户终身价值

Customer Acquisition Cost
(CAC)



Customer Lifetime Value
(CLV)



个人征信



个人征信



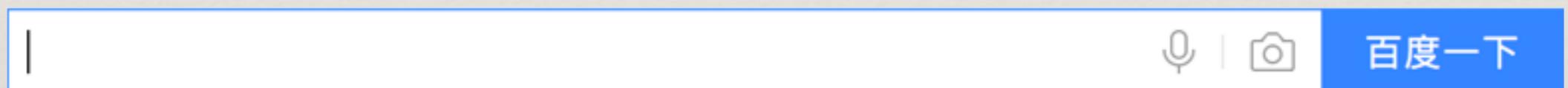
搜索引擎

Google

Google Search

I'm Feeling Lucky

搜索引擎



时空大数据



时空大数据



主要内容

- ❖ 什么是大数据?
- ❖ 大数据特点
- ❖ 大数据应用
- ❖ 大数据实例
- ❖ 大数据面临的挑战



利用大小数据竞选

大数据

facebook

**Facebook helps you connect and share with
the people in your life.**



小数据

每周：

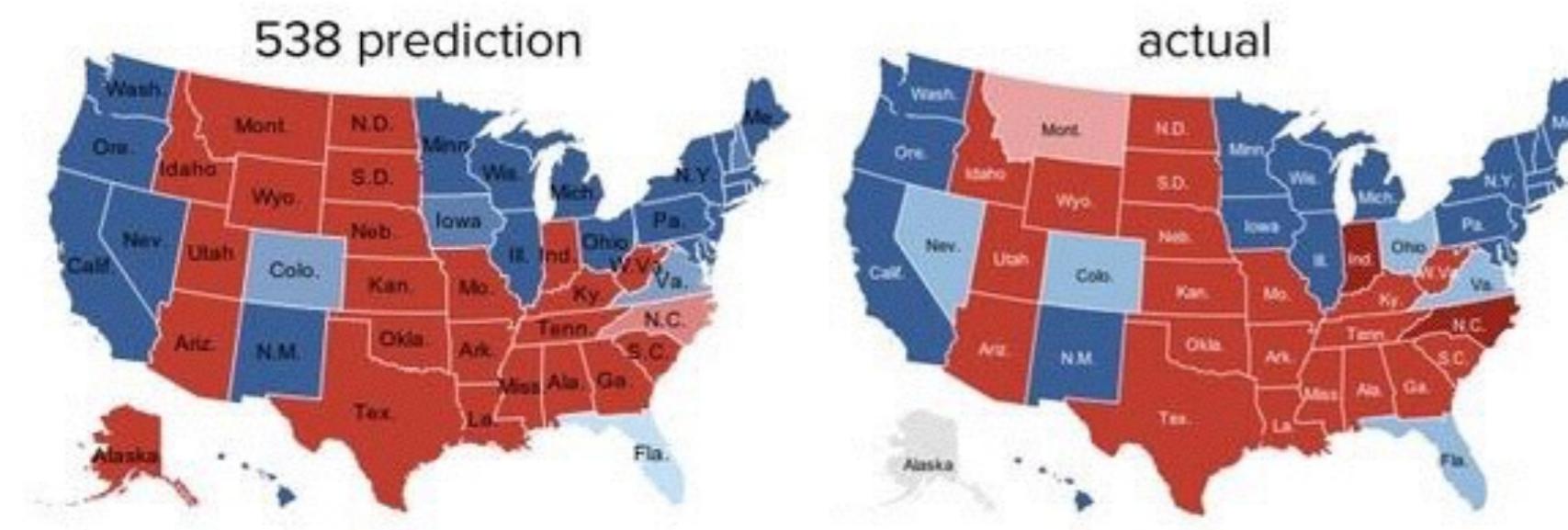
5000~10000份短问卷

1000次深度访谈

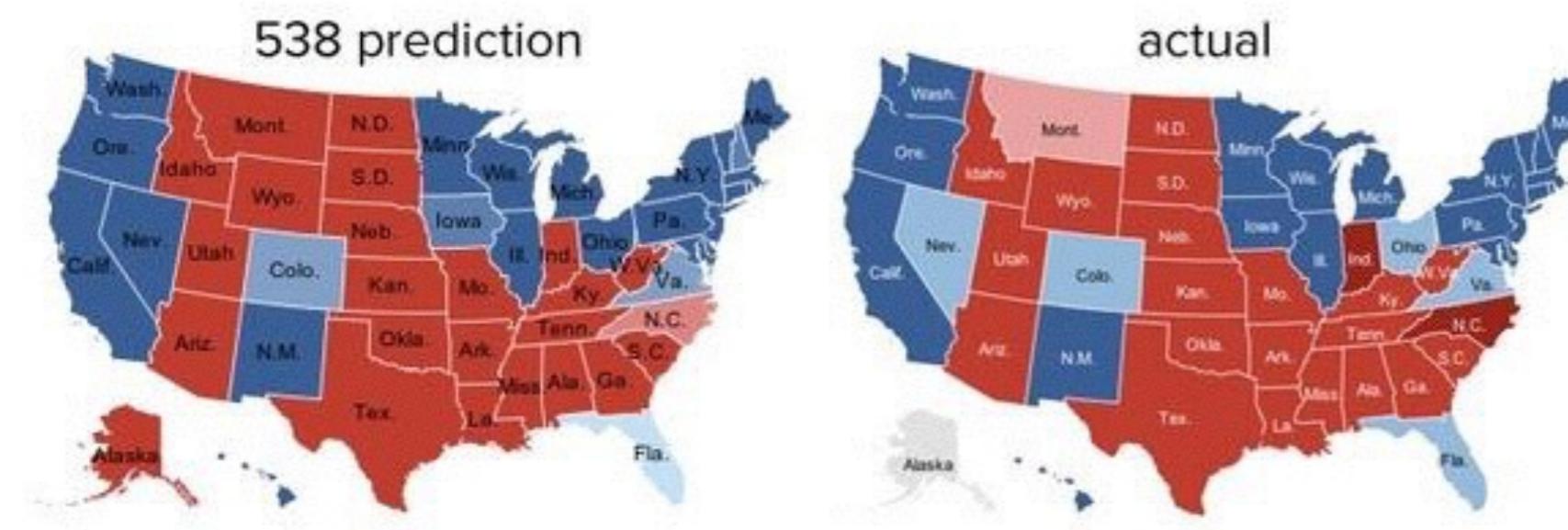
App实时传输数据

不断迭代更新预测模型

预测和真实结果



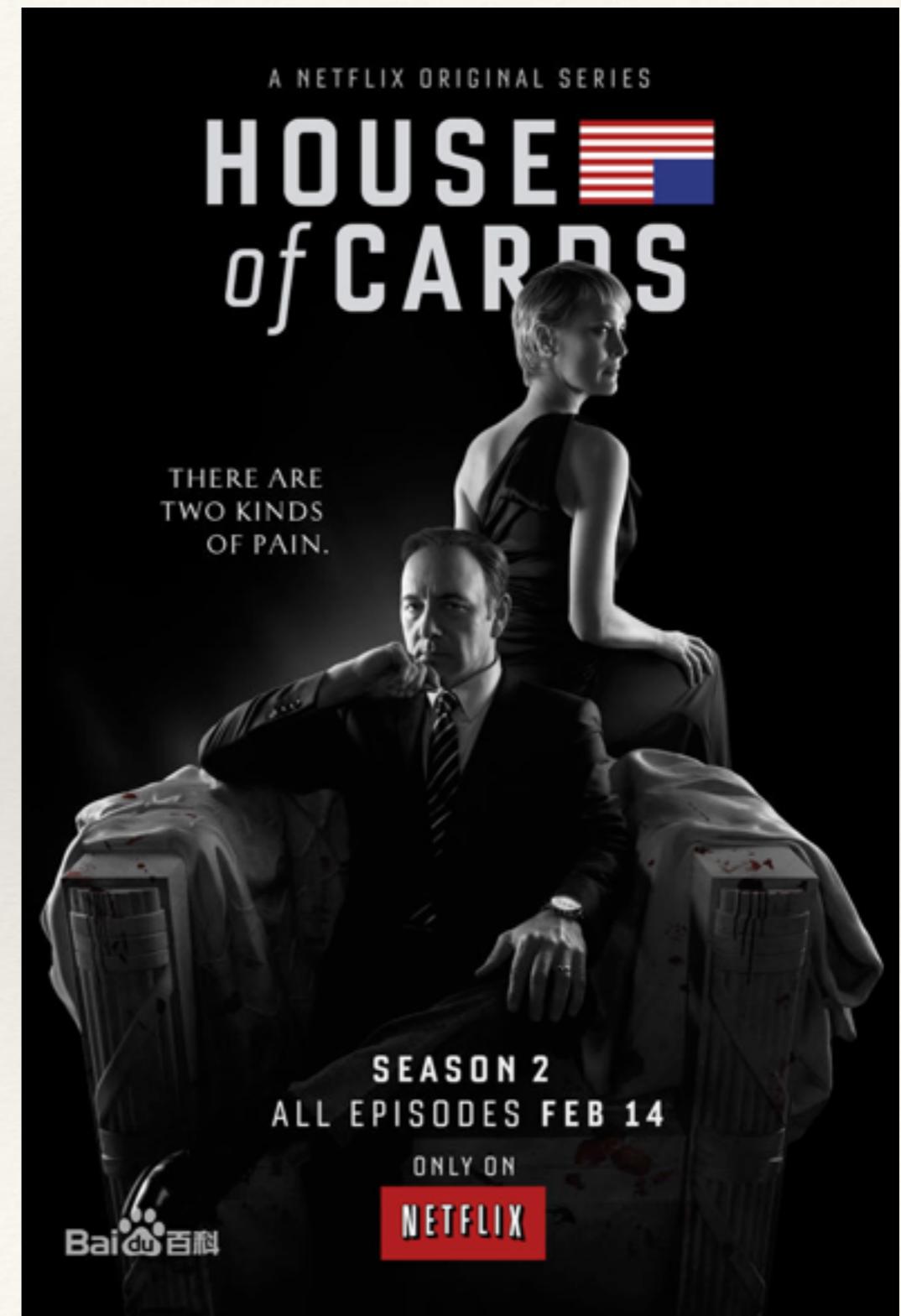
预测和真实结果



俄亥俄州

57.68% | **57.16%**
Model | Actual

Netflix — 纸牌屋



Netflix —— 纸牌屋

纸牌屋探秘：Netflix的大数据炼金术

作者：王萌

星期三, 二月 13, 2013

 动态 大数据, 热点

无评论



大卫芬奇执导的《纸牌屋》是首个将第一季全集同时上线的电视剧，Netflix勇气何来？

Netflix在《纸牌屋》中的“大数据制作”，可能预示着影视制作行业即将迎来一个重要拐点。

影视投资充满风险，收视率、票房与投资回报率的可预测性向来很差，但Netflix最新投资的电视剧——新版“House of Cards”（纸牌屋），让人们见识了大数据分析对Netflix这样的新媒体公司的价值。大数据分析正深入到电影的创作环节，这对整个影视创作行业，从剧本选择，导演演员的选择，拍摄和后期制作，乃至营销，都会产生深刻的影响。

本周热门

本月热门

评论排行

最新评论

标签云

- 薪酬支付-区块链2.0智能合约的商业应用
- 智能电商时代已经到来，但你可能还没有准备好
- DARPA暗网搜索引擎将引发互联网二次革命
- 中国网络安全企业50强(2016年上半年)
- 不作恶？别逗了，Google正在收集你的手机电话记录
- Facebook开源三款图像识别人工智能软件
- 云计算为企业软件市场带来的变化
- 2016年存储市场10大趋势
- 物联网正在给批发业带来新的商业模式
- 2016年物联网市场5大趋势

CTOCIO专栏



Andrew Chen



王萌



宋妍



关志刚



刘朝阳



Steve Mushero



Rik Ferguson



Jean-Paul Smets



张霖

Netflix —— 纸牌屋

- ❖ 2011年3月，拿下MRC公司《纸牌屋》剧集的首播权，Netflix CEO表示对涉足原创内容制作不感兴趣
- ❖ 首席联络官+首席营销官调整公关策略
- ❖ 两个方向，一是放大数据分析的作用，二是通过这种联系将 Netflix包装成技术型的HBO。
- ❖ Netflix从产业底端逆袭，改变变现流程

主要内容

- ❖ 什么是大数据?
- ❖ 大数据特点
- ❖ 大数据应用
- ❖ 真实案例
- ❖ 大数据面临的挑战

大数据面临的挑战

- ❖ 大数据不能完全替代传统数据
 - ❖ 数据价值
 - ❖ 信息爆炸 
 - ❖ 大数据技术实用性

大数据面临的挑战

- ❖ 数据安全及保护
 - ❖ 个人隐私受到威胁
 - ❖ 需要保护的数据量给存储带来压力

大数据面临的挑战

- ❖ 数据分析与精准预测
 - ❖ 历史数据难以刻画未来需求
 - ❖ 训练数据并不能代表真实数据