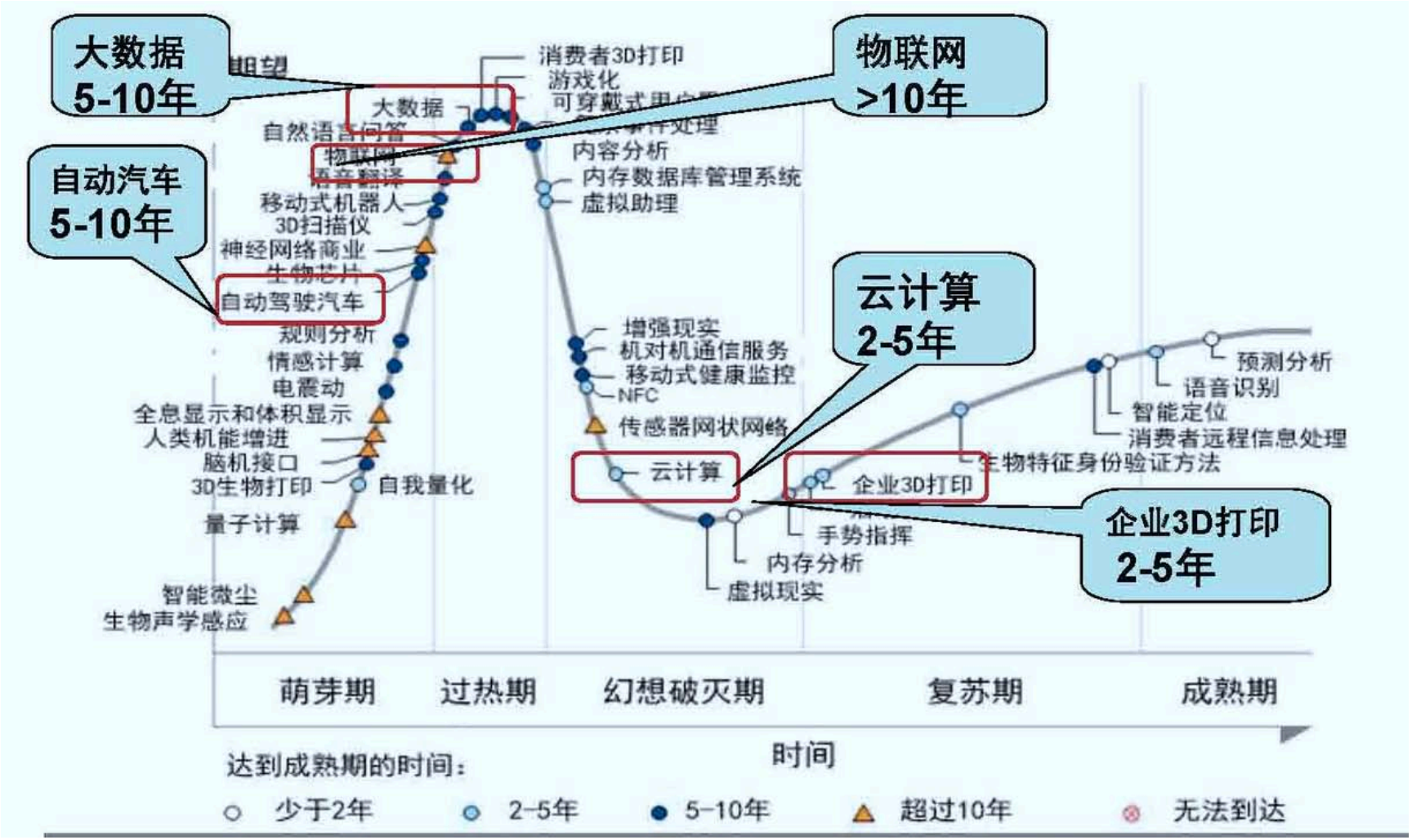
大数据思维
及电商大数据应用

邱宝军

7/4/2015

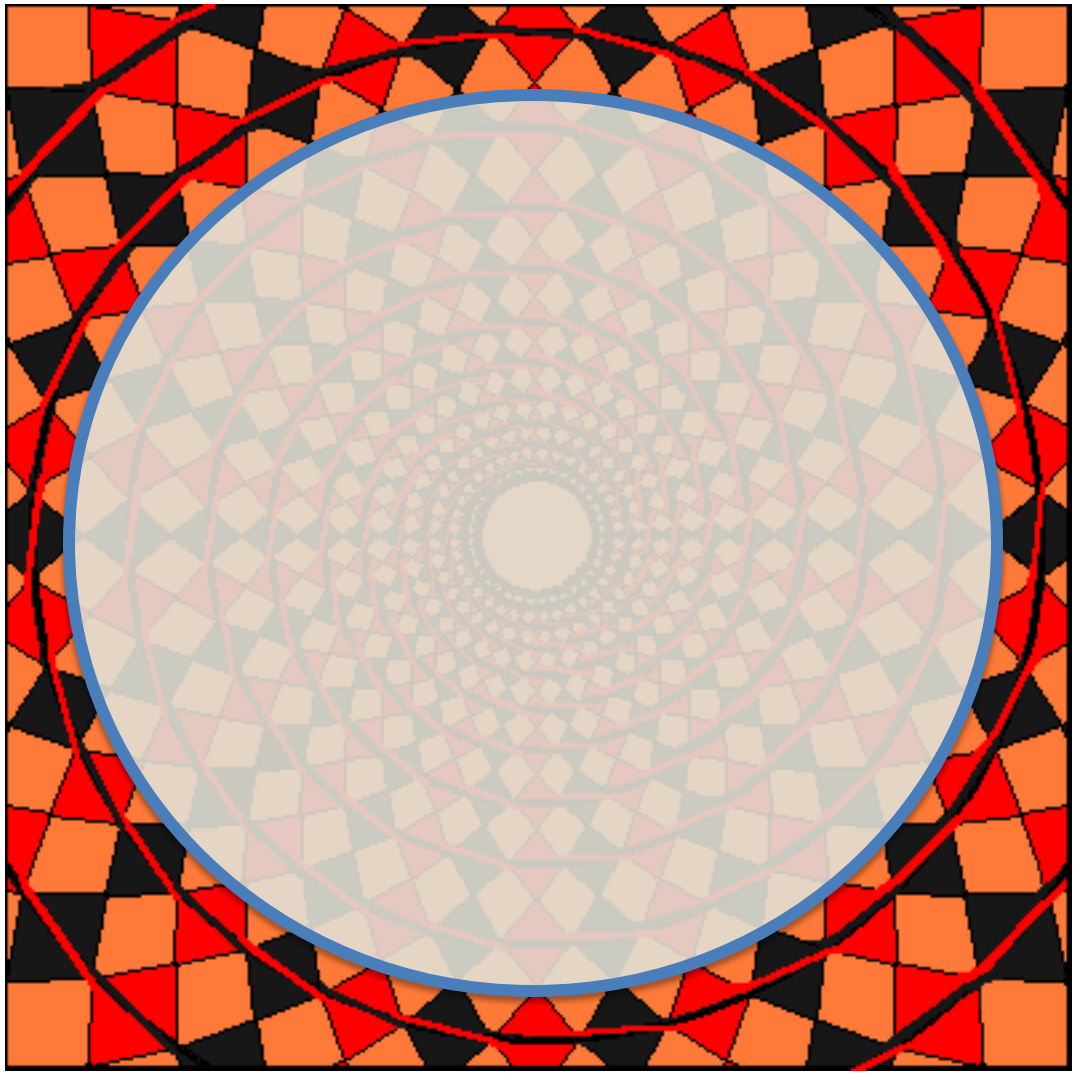
“大数据”生命周期


大数据
5-10年
物联网
>10年
自动汽车
5-10年
云计算
2-5年
企业3D打印
2-5年
消费者3D打印
游戏化
大数据
自然语言问答
物联网
语音翻译
移动式机器人
3D扫描仪
神经网络商业
生物芯片
自动驾驶汽车
规则分析
情感计算
电震动
全息显示和体积显示
人类机能增进
脑机接口
3D生物打印
自我量化
量子计算
智能微尘
生物声学感应
内容分析
内存数据库管理系统
虚拟助理
增强现实
机对机通信服务
移动式健康监控
NFC
传感器网状网络
云计算
内存分析
虚拟现实
手势指挥
企业3D打印
生物特征身份验证方法
消费者远程信息处理
智能定位
语音识别
预测分析
萌芽期
过热期
幻想破灭期
复苏期
成熟期
时间
达到成熟期的时间：
少于2年
2-5年
5-10年
超过10年
无法到达

自我介绍

- 邱宝军
  - 北京大学：数学学士、智能科学硕士
  - 宾夕法尼亚州立大学：统计/计算机硕士，计算机博士（人工智能与大数据）
  - 美国微软，研发工程师：Bing核心排序算法
  - 美国eBay，主管科学家：电商搜索引擎与大数据
  - 杭州雪肌科技(传感器与大数据)：共同创始人，现股东&顾问
  - 中国农产品贸易中心（全球）集团：董事
  - 国美在线：大数据VP

看到漩涡了吗？



以色列保释官

- 8位以色列保释官
  - 一天到晚，除去三餐，都在处理保释申请
  - 每份保释申请：6分钟
  - 35%的批准率
  - 65%→0%?

思维偏差

- 视觉偏差
- 立场偏差
- 生理状况导致偏差
- 相似度启发偏差
- ...

{ Talking } Data
用数据说话

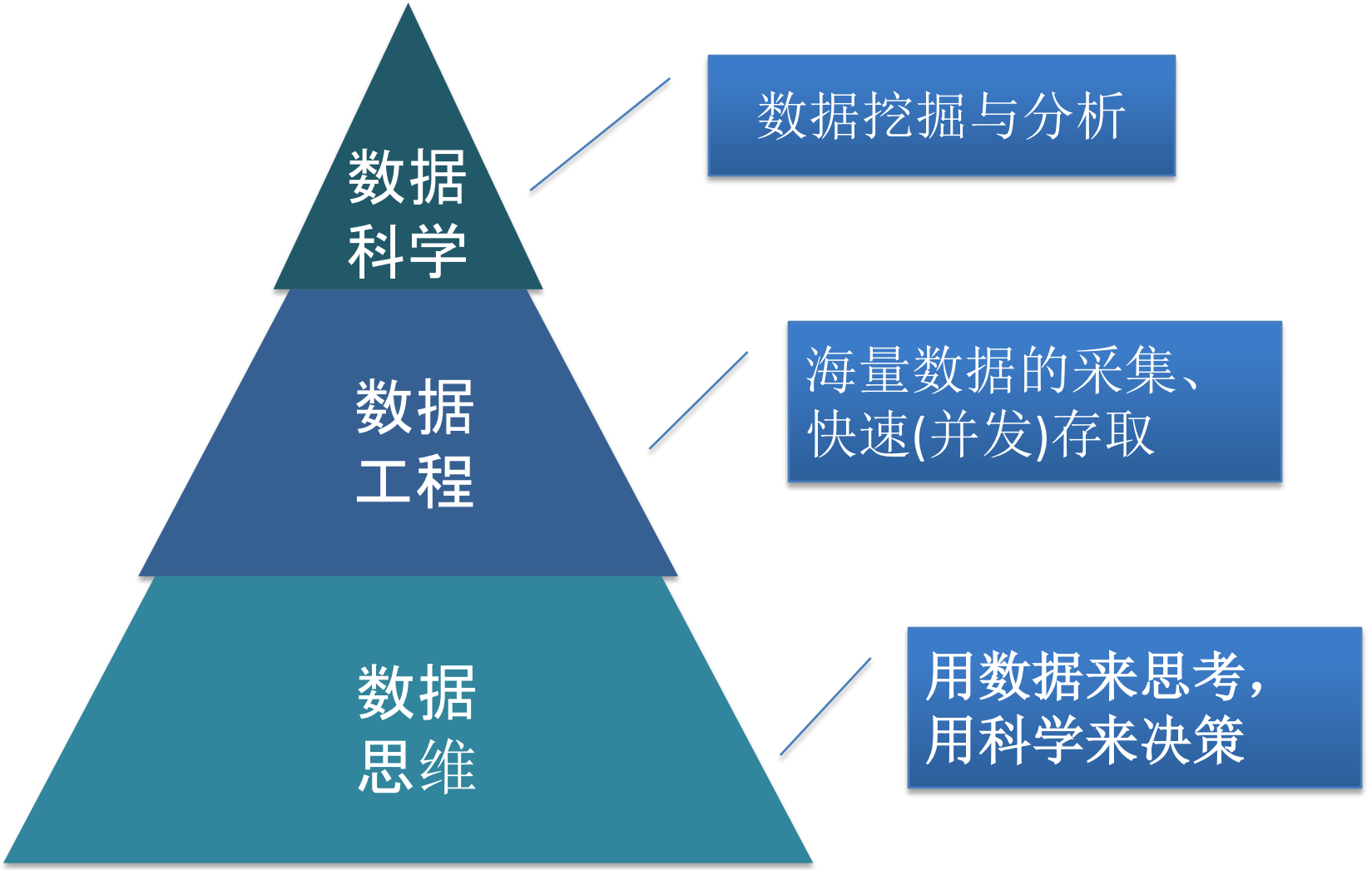
从文化的高度认识
数据思维

- 美国文化
  - 参议院+众议院
  - 人口普查----计算机的发明
  - 南北战争：谢尔曼将军“向大海进军”

- 中国传统文化
  - 胡适/汪洋：中国的“差不多先生”
  - 诸葛亮 vs. 司马懿

- 智慧 vs. 数据（+逻辑）思维
  - 可解释、可传承、可积累

大数据



数据思维（1）

—用数据来思考，用科学来决策

- 定量思维：测
  - 信息获取：用户的所有信息、每一次操作（甚至是鼠标的轨迹）
  - 信息到数据：销售数据、交易额、点击率、转化率、顾客满意度、用户增长率
  - 海底捞：客户满意度
  - Nala.com.cn

数据思维（2）

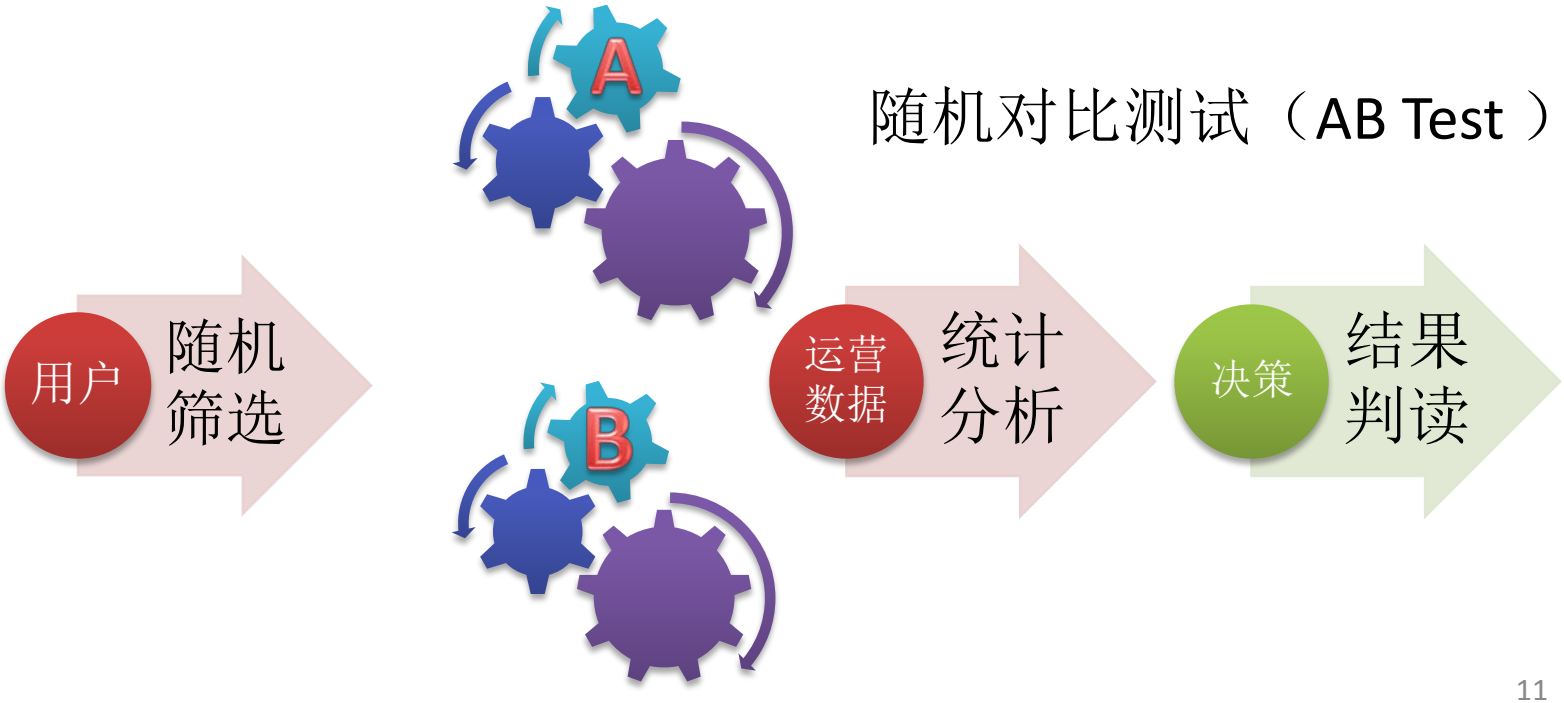
—用数据来思考，用科学来决策

- 相关思维：连
  - 用户属性/行为/事物之间
  - 移动数据+厦门旅游数据
    - 广告→中央台 vs. 广东台
  - 美国佛罗里达飓风 vs. 蛋挞销售
  - 不同地区和时间，感冒患者在Google上进行相关搜索的量 vs. 感冒的流行度

数据思维（3）

—用数据来思考，用科学来决策

- 实验思维：试
  - 假设→试验→结论



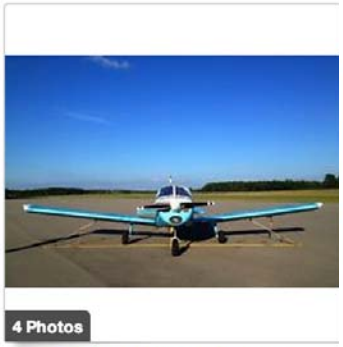
随机对比测试（AB Test ）

小图片 vs. 大图片

Categories

Toys & Hobbies (157,504)
- Models & Kits (111,462)
- Diecast & Toy Vehicles (35,364)
- Radio Control & Control Line (8,523)
- More

Collectibles (68,421)
- Transportation (29,183)
- Postcards (21,189)
- Militaria (8,436)
- Photographic Images (4,225)



1967 Piper Cherokee PA28-140

$22,500.00 0 bids

Cessna 172L 1971

$39,500.00 Buy It Now

1961 Cessna 210

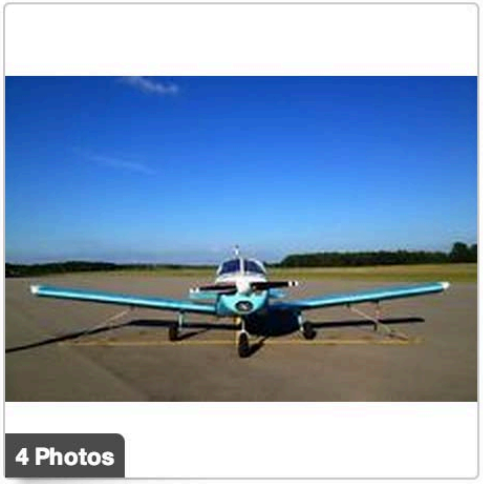
$33,433.33 4 bids
$48,000.00 Buy It Now

小图片

Categories

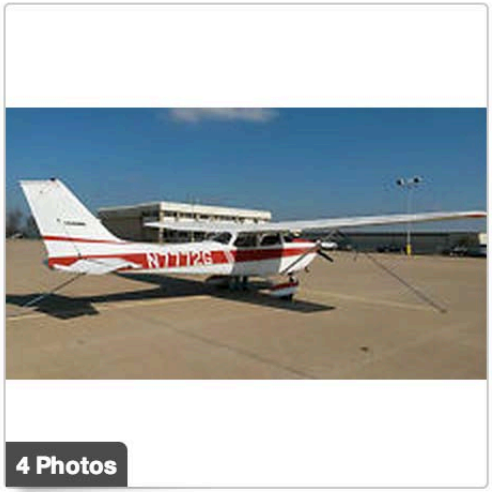
Toys & Hobbies (157,504)
- Models & Kits (111,462)
- Diecast & Toy Vehicles (35,364)
- Radio Control & Control Line (8,523)
- More

Collectibles (68,421)
- Transportation (29,183)
- Postcards (21,189)
- Militaria (8,436)
- Photographic Images (4,225)



1967 Piper Cherokee PA28-140

$22,500.00 0 bids



Cessna 172L 1971

$39,500.00 Buy It Now

大图片

随机对比测试（AB Test ）

- 客观的限制
- 伦理的限制

- 情感的限制
  - Amazon：定价方法测试
  - 美国大陆航空公司(Continental Airlines)：飞机晚点向旅客致歉方法测试

数据思维
测-连-试

描述　　预测　　攻略

易导致误解的大数据流行观点

- 大数据采用全体数据，而不是随机样本
- 大数据关注混杂性，而不是精确性
- 大数据关心相关性，而不是因果性

易导致误解的大数据流行观点

大数据采用全体数据，而不是随机样本

- 问题须分类：
  - 有些问题抽样就可完美解决，如检测某地土质
  - 有些问题数据越全越好，如推荐系统
- 数据的“全”是相对的
- 样本的关键在于具有“代表性”，而不是越大越好

- 1936：Roosevelt vs. Alf Landon
  - Literary Digest
    - 从车辆注册信息和电话号码簿中随机筛选了上千万人，寄出问卷
    - 230 万人电话回应：Landon > Roosevelt
- 波士顿：颠簸的街道App
  - 司机的手机App利用加速度感应器感应路面的坑洼，并自动通知市政厅检修路面
  - 波士顿政府：大数据为这座城市提供了实时信息，帮助我们解决问题并做出长期投资计划

易导致误解的
大数据流行观点

大数据关注混杂性，而不是精确性

- 数据一旦很大，噪声会相互抵消
- 随着训练数据的增加，简单模型的表现越来越接近（甚至超越）复杂模型

- 同样量级：精确数据 > 混杂数据

易导致误解的
大数据流行观点

大数据关心相关性，而不是因果性

- 跺地板 → 加州人躺下
- Google预测流感大爆发：2009✓，2012✗

If you torture the data hard enough, they will confess to anything

2014，美国国防高级研究计划局（DARPA）启动“大机理”（Big Mechanism）项目，目的是发展可以发现隐藏在大数据中的因果关系模型

正确利用大数据（1）
-业务驱动，应用导向

- 基于业务，提出大数据问题
- 基于业务，获取和清洗大数据
- 基于业务，解释和修正大数据模型

选商家

- 有两个淘宝在线店铺，数据显示：

  他们都做成了1000宗生意

  A店铺的客户好评率20%

  B店铺的客户好评率80%

- 某“客户”要采购两店铺都经营的宝贝，选哪个店铺？

	电冰箱	手电筒
A	100/900	100/100
B	0/100	800/900

美国竞选民意调查

- 1936：Roosevelt vs. Alf Landon
  - Literary Digest
    - 从车辆注册信息和电话号码簿中随机筛选了上千万人，寄出问卷
    - 230 万人电话回应：Landon > Roosevelt
- 2012：Obama vs. Romney
  - 第一轮总统辩论中，Romney 以巨大优势胜出
  - 大量民调→Obama 的领先优势化为乌有
  - YouGov→Obama仍领先
    - 辩论前后都参与民调的占70%，他们的立场几乎不变
- Bradley Effect
  - Tom Bradley：洛杉矶的黑人市长
  - 1982 年竞选加州州长，民调一路领先，最后落选

获取“正确”的决策数据

	目标	数据上“隐含的假设”
选店铺	Pr(好评| 某商品，某店铺)	总好评率高→ 特定商品好评率高
美国民调	Pr(某候选人得票 | 非偏的选民样本)	选民邀请随机， 选民接收民调率相似， 选民都在说实话， ……

随机对比测试（AB Test）

- 优化目标
  - 点击率，点击前耗时，成交率，成交商品数，成交额，成交满意率
  - 新用户 vs. 老用户
  - 关键指标的选取与演化
    - 如：初期用户增长率，新用户留存率，后期转化率
- 测试对象随机筛选
  - IP？用户ID？浏览器？
- 测试结果的误差分析与p值计算
  - 差别的显著程度：样本数，测试时长
- 周期性/季节性
  - 同一天的不同时段，工作日 vs. 周末，年底 vs. 非年底
- 多模块之间的干扰
  - 测试的延迟效应
  - 单变量测试 vs. 多变量同步测试

正确利用大数据（2）

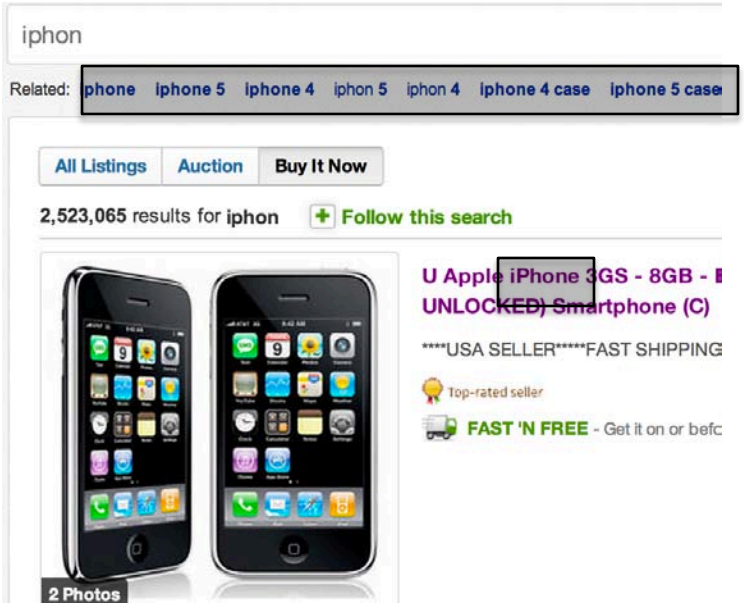
- 培养数据意识，使用数据思维：测-连-试
  - 例子：林彪，美国加油站数量
- 切勿绝对化：不分析因果，不需要采样，不需要精确数据
- 不要轻视也不要迷信大数据
- 不要盲目追求数据的规模
- 要注重“小数据”问题
- 要注意“大数据”的成本

- 大数据方法往往是归纳，注意归纳法的缺陷（黑天鹅现象），适时与演绎法结合
  - 德国哲学家波普尔：证伪主义（科学理论无法被归纳法证实，只能被实验发现的反例证伪），科学始于问题
  - 伽利略、牛顿：科学始于观察（观察→归纳→科学理论）
    - “科学始于数据”？
  - 2014，美国国防高级研究计划局（DARPA）启动“大机理”（Big Mechanism）项目，目的是发展可以发现隐藏在大数据中的因果关系模型

电商大数据挖掘

- 用户浏览器打入网址，回车
  - DNS → DC → 服务器
  - 用户个人信息/历史资料获取
  - 个性化的首页
- 用户搜索框输入
  - 查询自动补全，拼写纠错
  - 查询理解/扩展（用户意图预测），商品获取，排序（用户意图预测，个性化，不良交易预防）；相关搜索；商品推荐；广告显示；相关导购信息;
- 用户点击商品
  - 商品信息获取，商品推荐

- 拼写纠正/同义词挖掘
相关搜索



- 商品推荐

Apple MacBook Air MD760LL/A 13.3-Inch Laptop (NEWEST VERSION)

by Apple

356 customer reviews | 80 answered questions

List Price: $1,099.00
Price: $1,048.99 Prime
You Save: $50.01 (5%)

In Stock.
Ships from and sold by Amazon.com. Gift-wrap available.

Want it Saturday, March 8? Order within **1 hr 14 mins** and choose **Two-Day Shipping** at checkout. Details

Capacity: **13.3-inch**

- Amazon：35%页面销售源自于推荐引擎
- eBay：70%流量为搜索流量
  - 搜索与数据挖掘核心部门
    - 每年通过数据挖掘改善产品，增加营收几亿美元
  - 我领导的项目：每年提高8000万美金营收
- eBay：假货与不良交易检测

- Farecast：机票价格预测
- Decide：定价模型（价格预测）

- Target：预产期预测

淘宝数据分析

- 商品搜索排名不高
  - 关键词不对，不良记录，类目属性有错
- 商品有展现无点击：流量不精准
  - 关键词不精准，价格过高，款式不对，主图不吸引人，宝贝销量不够形成公信力
- 商品有点击无转化
  - 商品销量不够形成公信力，商品有中差评未解释，商品详情页不动人，其他主图不吸引人
- 店铺流量、点击或转化降低
  - 定位出哪些商品的相应指标在降低
- 分析比较各商品指标高低，设计实验寻找改善措施

其他大数据案例（选）

- 奥巴马选举
  - 基于民调分析各区域选情，找到支持率不高人群，选择适当媒体投放宣传片
- 家具店精准营销
  - 在百度买“家具”相关关键字，获取精准客流；通过数据分析找爆款
- 遥感数据→投资情报
  - 禾讯：基于遥感的农业分析
  - RS Metrics：Walmart停车场/储油罐的使用率→投资情报
- 德国队赢得2014世界杯
  - 大数据的胜利，传感器和摄像头获得运动员数据，帮助教练调整布局和战术，帮助运动员自我了解和提高
- IBM：大数据指导药物合成
  - 研究人员利用化合物-蛋白互作组的海量数据在超级计算机上模拟药物相互作用
- 社会治安与火灾预警大数据

数据是资产

搜索引擎

- 谷歌 vs 微软必应

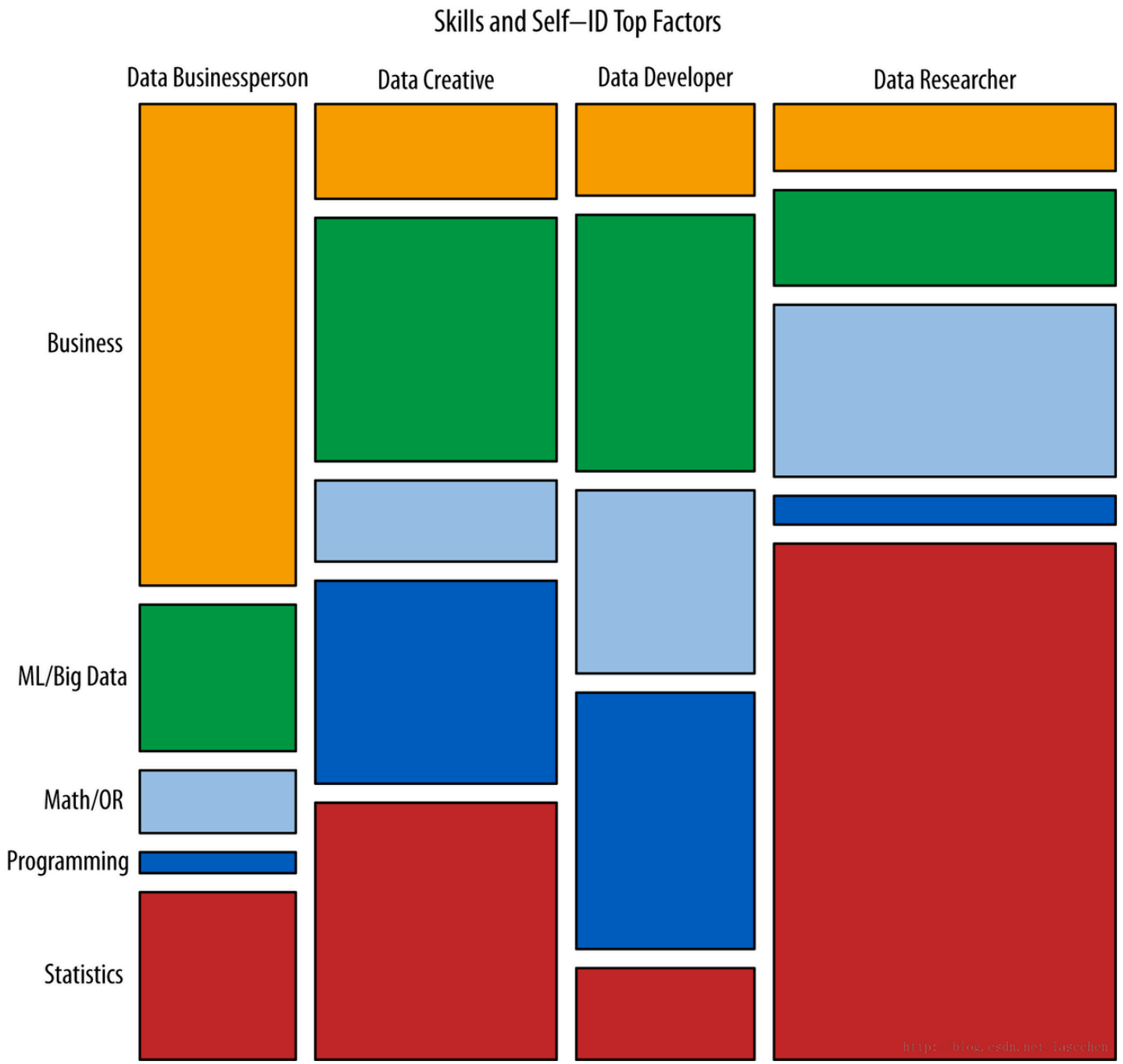
阿里巴巴：

- 电商消费数据
  - 内部使用：搜索、推荐、广告、决策
  - 金融：企业或个人征信 （银行）
  - 经济：经济走向、区域经济 （政府）
  - 产业：行情分析 （企业）
  - 广告：用户特征与喜好分析（广告主）
  - 保险：目标客户
- 布局大数据经济：投资相关领域和流量入口、获取更多更广泛数据；数据连接、分析与挖掘→产生价值

- Uber/滴滴打车

- 移动
  - 厦门旅游广告：中央台 vs 广东台

大数据科学家



小结

- 数据思维是一种优秀思维习惯
- 随机对比测试
- 小数据同样重要，要懂统计
- 业务与应用驱动数据应用
- 数据是资产
- 数据科学家

- Q/A

qiubaojun@139.com