数据科学导论

Introduction to Data Science

第一章 数据科学基础

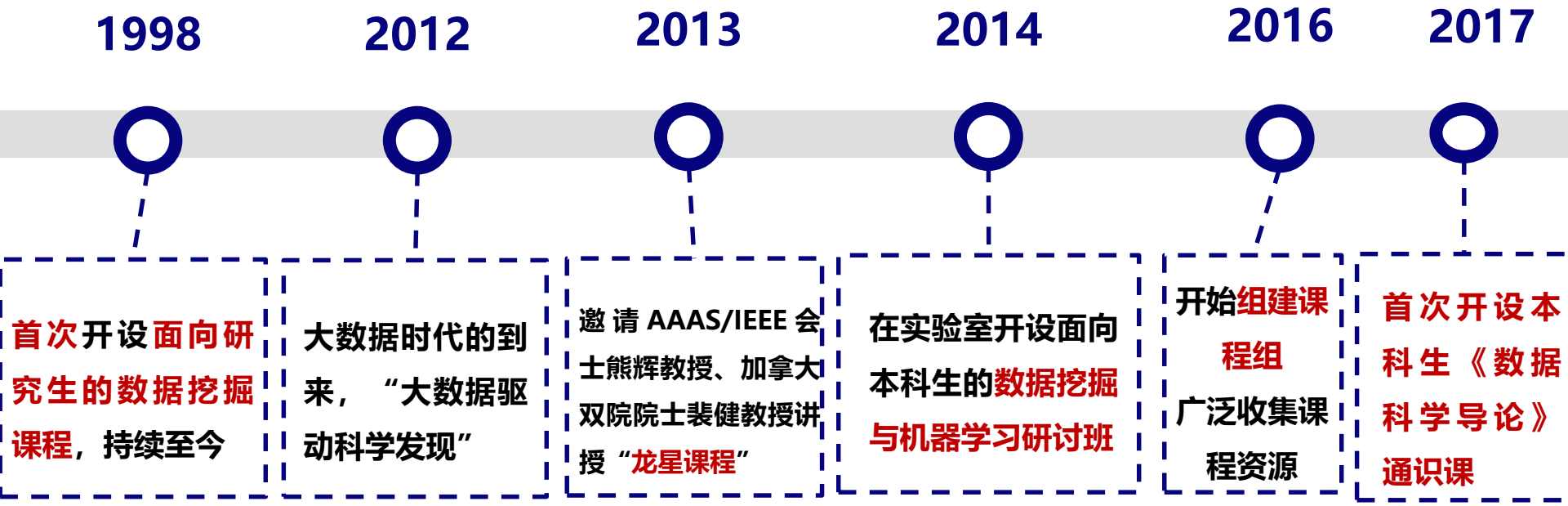
陈恩红，黄振亚，刘淇
Email: cheneh@ustc.edu.cn, huangzhy@ustc.edu.cn

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/DS2021.html

助教：刘嘉聿
ds_intro2021@163.com

11/10/2021

建设历程


1998
首次开设面向研究生的数据挖掘课程，持续至今
2012
大数据时代的到来，“大数据驱动科学发现”
2013
邀请 AAAS/IEEE 会士熊辉教授、加拿大双院院士裴健教授讲授“龙星课程”
2014
在实验室开设面向本科生的数据挖掘与机器学习研讨班
2016
开始组建课程组
广泛收集课程资源
2017
首次开设本科生《数据科学导论》通识课

课程目标

- 全面了解数据科学的基础知识
  - 包括数据分析的常用技术、发展前沿和应用案例
  - 了解数据的“能”与“不能”
- 树立数据科学的基本思路
- 初步掌握使用数据分析手段解决实际应用问题的能力

用科学的方法研究和应用数据

选修数据科学与导论课程的同学将来可能从事不同领域的科学研究或者技术开发，

希望这门课程带给你们的是终身受用的数据思维和创新能力。

数据科学基础

- 数据
  - 从计算机科学的角度，所有能够输入到计算机并被计算机程序处理的符号的总称
- “人-机-物”三元融合，世界已经成为数据化的世界

Google 谷歌

Baidu 百度

当文字成为数据

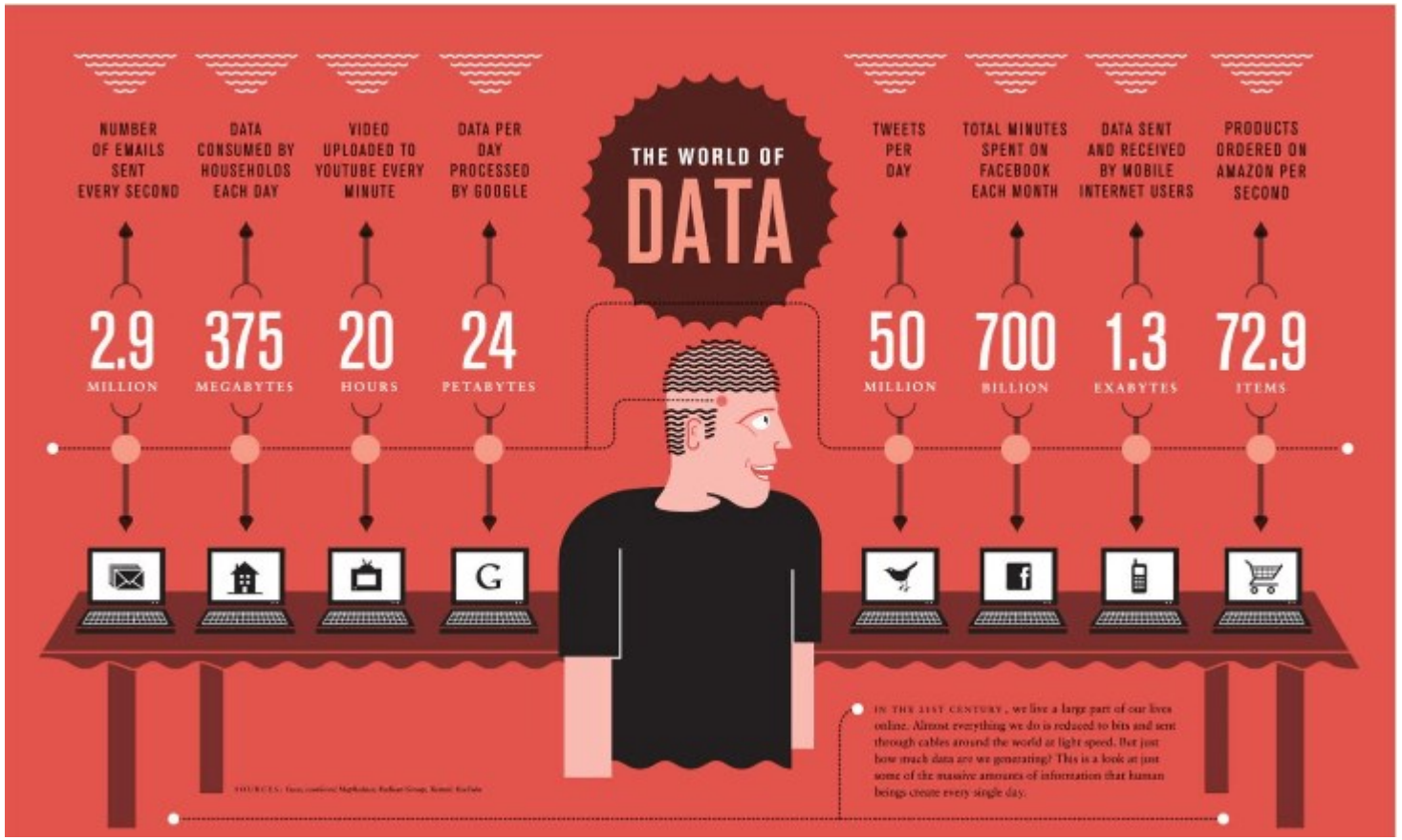
当方位成为数据

当沟通成为数据

一切事物的数据化

数据科学基础

- 我们生活在数据中，所有人都在制造和分享数据



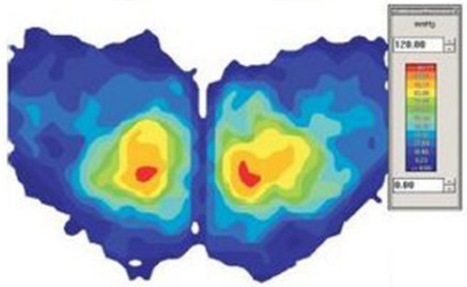
数据科学基础

- 我们生活在数据中，所有人都在制造和分享数据

amazon alexa

数据科学基础

- 案例：从最不可能的地方获得数据 (互联网+物联网)
  - 当一个人坐着的时候，他的身形、姿势和重量分布都可以量化和数据化。



测量人对椅子施加的压力，
用0~256的数值量化

- 把人体屁股特征转化成了数据，产生独属于每个乘坐者的精确数据资料。
- 汽车可以准确的识别乘坐者的身份：**汽车防盗系统**

数据科学基础

背景：铺天盖地的“大数据”字眼

大数据“完全占领”了互联网和IT领域之后，开始进入各行各业，形成了政府大数据、教育大数据、医疗大数据、交通大数据、金融大数据、保险大数据、公安大数据、法院大数据、旅游大数据、……

数据科学基础

- **李德毅院士：大数据本身，既不是科学也不是技术，它反映的是网络时代的一种客观存在**

你们说的大数据到底是啥？**大数据的输入和输出是？**

我不认为数据等同于价值，**哪些数据才有价值？**

大数据到底是**噱头+忽悠**，还是**真金白银**啊？

我没看清楚大数据的价值，但很清楚**大数据的大成本**，真能赚回来吗？

未来真的**不会大数据就不能赢了吗？**

我用SQL Server用的好好的，一定要**现在就转大数据吗？**

所谓的大数据牛的公司，**到底牛在哪？**

大数据就是数据，没什么可神秘的。它是一种原材料，数据库、数据挖掘、云计算、高性能计算、机器学习等都可以看作是对这种原材料进行存储烹饪加工等的手段和技术，目的就是做出各种美食（例如让AlphaGo打败李世石）

背景：铺天盖地的“大数据”字眼

从2008年9月，《Nature》杂志首次出版一期大数据专刊，科学家们提出“大数据真正重要的是新用途和新见解，而非数据本身”

数据科学基础

大数据有多大?

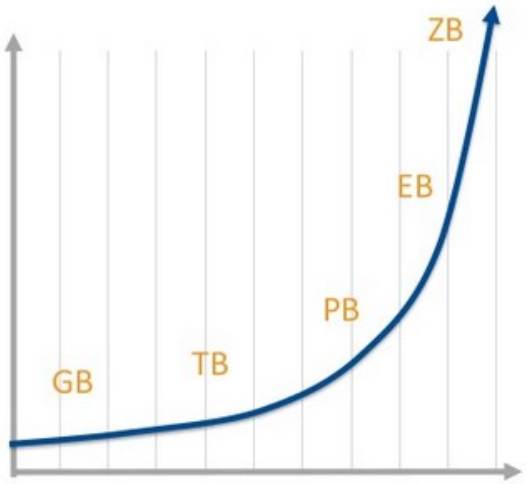
PB是大数据层次的临界点

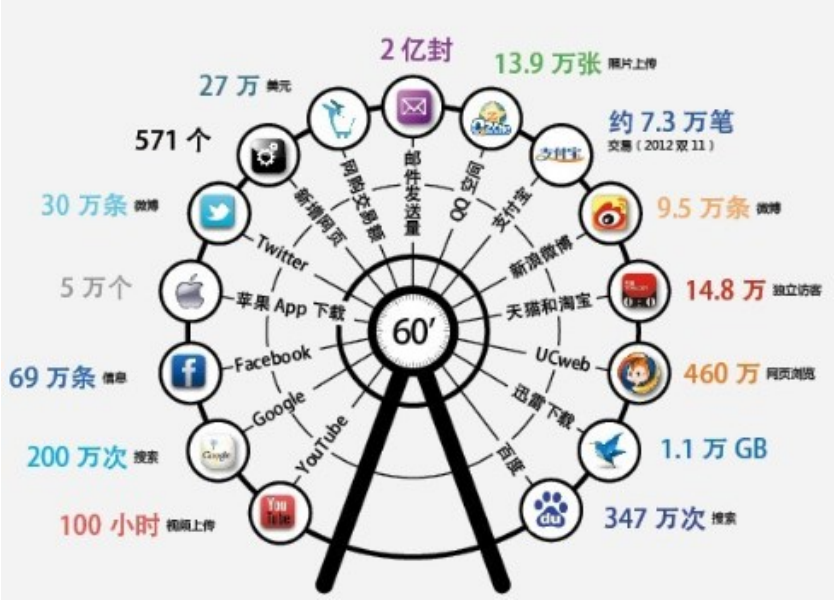
- 数据量已到ZB等级

KB->MB->GB->TB->PB->EB->ZB->YB->NB->DB

PB以上级别的数据，最有效的传输方式是空运，而不是网络

- 而大数据不仅仅只是量大!

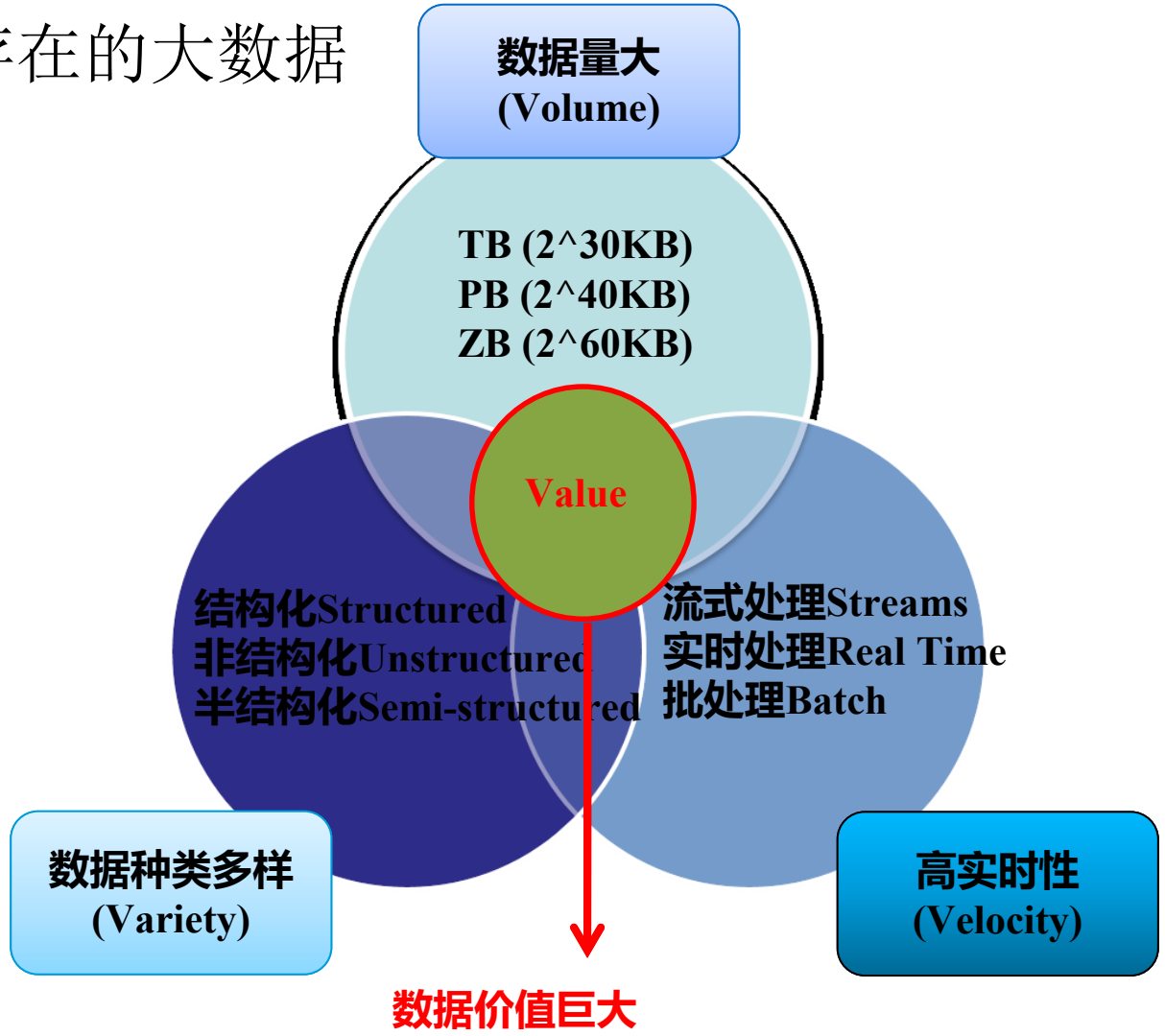




60秒，我们能产生多少数据?

数据科学基础

- 客观存在的大数据



数据科学基础

- 客观存在的大数据---Volume(数据量巨大)

仅是历史数据

阿里所保有的、经过清洗的历史数据已超过100PB。

——阿里数据仓库负责人七公（汪海）

百度现在的数据规模已经到了EB级，每天处理的数据量到了上百PB。

——百度大数据部总监薛正华

全球数据总量在2010年达到1.2ZB，预计2020年达到44ZB，每两年增长一倍。

——IDG数字宇宙报告2014

$$1\ ZB = 2^{10}\ EB = 2^{20}\ PB = 2^{30}\ TB = 2^{40}\ GB$$

- 1 ZB = 地球上沙粒的总量，1 EB = 4000个美国国会图书馆的藏书

数据科学基础

- 客观存在的大数据--- Variety(数据类型多)

数据形式的多样:

- 结构化数据，半结构化数据，非结构化数据
- 关系数据库数据、xml/JASON文档、音视频数据

数据来源的多样性:

- 不同的IT应用系统
- 各种设备（手机、手环）
- 互联网、物联网
- 其它

时空数据

图像数据

文本数据

事务数据

视频数据

音频数据

数据科学基础

- 客观存在的大数据--- Velocity(高实时性)

1秒定律：对于大数据应用而言，必须要在1秒钟内形成答案，否则这些结果可能就是过时的、没有意义的

在百度输入关键字：
“汽车维修”、“挖掘机 学习”

某在线电影网站

某IT业界资讯网站

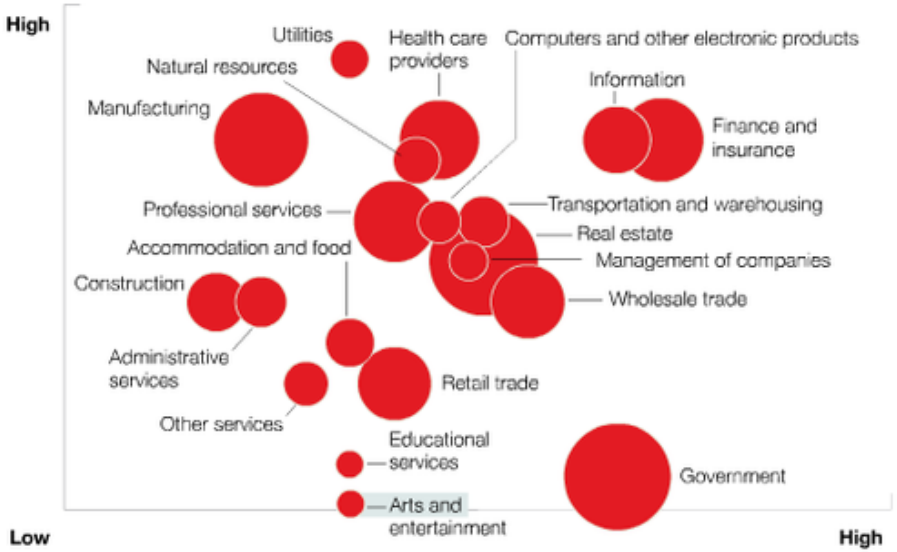
例如用户在合肥某台PC上，打开百度输入关键字片刻之后，再打开其它网站，就会看到相关的广告，并且所推荐的是地理位置信息相关的（**合肥、安徽**）

数据科学基础

- 客观存在的大数据--- Value(价值巨大但价值密度低)

挖掘大数据中的价值类似**沙里淘金**，需要从海量数据中挖掘稀疏但珍贵的信息

所有产业都可以应用大数据产生价值



各产业GDP占比
(以美国经济为例)

图：麦肯锡对各个行业从大数据中获得价值难易程度的分析 (2011年)

国际战略布局与思考

美国的大数据规划 – 上升为国家意志

- ✓ 2012年3月29日，美国联邦政府整合6个部门宣布2亿美元的“Big Data Research and Development Initiative”
- ✓ 目标：国家安全、新兴产业、科学发现与新型学科

美国成立国家级研究机构–突破核心科学与技术挑战

- ✓ 2012年，美国能源部斥资2500万美元，在Lawrence Berkeley National Laboratory组织下，汇集美国6大国家实验室和7所著名大学，建立可扩展网络数据管理、分析与可视化(SDAV)研究所。

欧盟的大数据规划 - 数据基础设施为先导

- ✓ **地平线(Horizon 2020) - The Framework Program for Research and Innovation**
- ✓ **GRDI 2020 - Global Research Data Infrastructures**
- ✓ **FP7 Call 8 Intelligent Information Management - Big Data**

GRDI2020
A Vision for Global Research Data Infrastructures

HORIZON 2020
THE FRAMEWORK PROGRAMME FOR RESEARCH AND INNOVATION

数据科学基础

- **大数据已成为国家基础性战略资源，日益对全球经济运行机制、社会生活方式和国家治理能力产生重要影响**

- **党中央、国务院高度重视大数据发展及其创新应用**
  - 2014.11《关于促进电子政务协调发展的指导意见》
  - 2015.08《促进大数据发展行动纲要》
  - 2015.10 十八届五中全会明确提出实施**国家大数据战略**

《“十三五”国家科技创新规划》发布，部署启动大数据等15个重大项目

2016-08-09 17:31

发改委：将组织建设13类国家级大数据实验室

2016年08月31日 10:29:46 来源： 新华网

国务院关于印发
新一代人工智能发展规划的通知

国发〔2017〕35号

- 十三五规划建议中的“大数据”

中共中央关于制定国民经济和社会发展第十三个五年规划的建议

(2015年10月29日中国共产党第十八届中央委员会第五次全体会议通过)

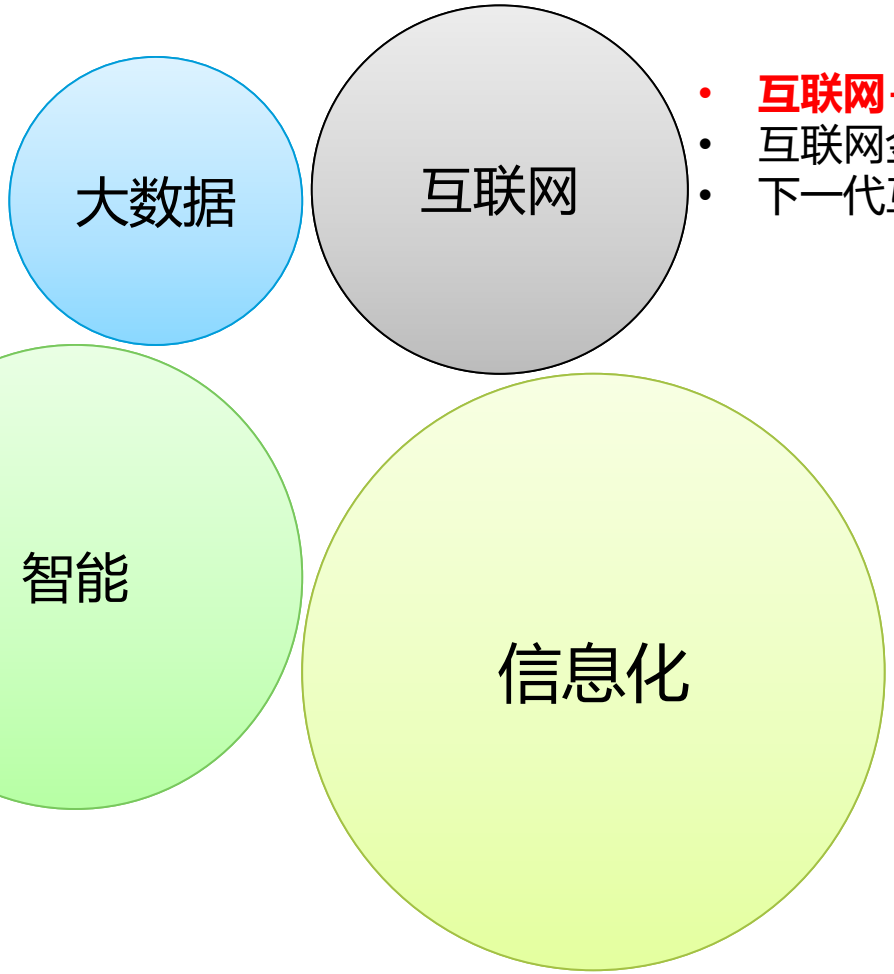
到二〇二〇年全面建成小康社会，是我们党确定的“两个一百年”奋斗目标的第一个百年奋斗目标，“十三五”时期是全面建成小康社会决胜阶段

数据科学基础

- 十三五规划建议中的“大数据”

关键词

- **国家大数据战略**
- 大数据技术



- **互联网+**
- 互联网金融
- 下一代互联网

- 智能消费
- **智能制造**
- 智能电网

- 社会信息化
- 流通信息化
- 农业信息化
- 教育信息化
- 信息化战争
- 社会治理信息化
- 信息基础设施
- 信息技术

数据科学基础

- 十三五规划建议中的“大数据”

21世纪的“数据”相当于
20世纪的“石油”

数据科学基础

- 十四五规划建议中的“大数据”

中共中央关于制定国民经济和社会发展第十四个五年规划的建议

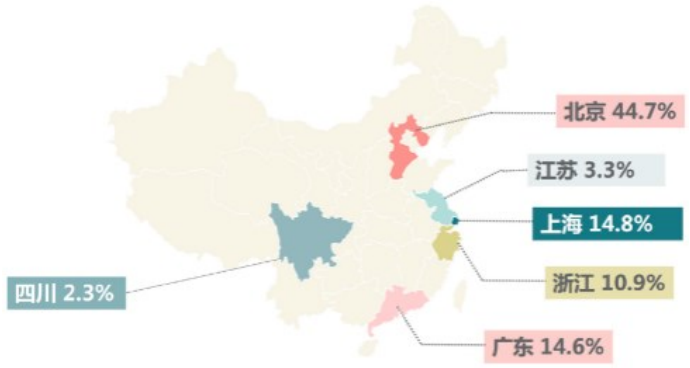
(2020年10月29日中国共产党第十九届中央委员会第五次全体会议通过)

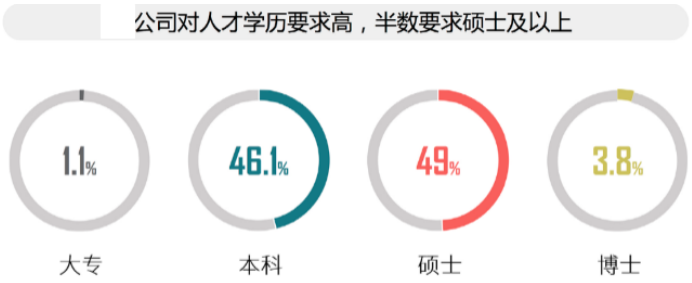
系统布局新型基础设施，加快第五代移动通信、工业互联网、**大数据中心**等建设。

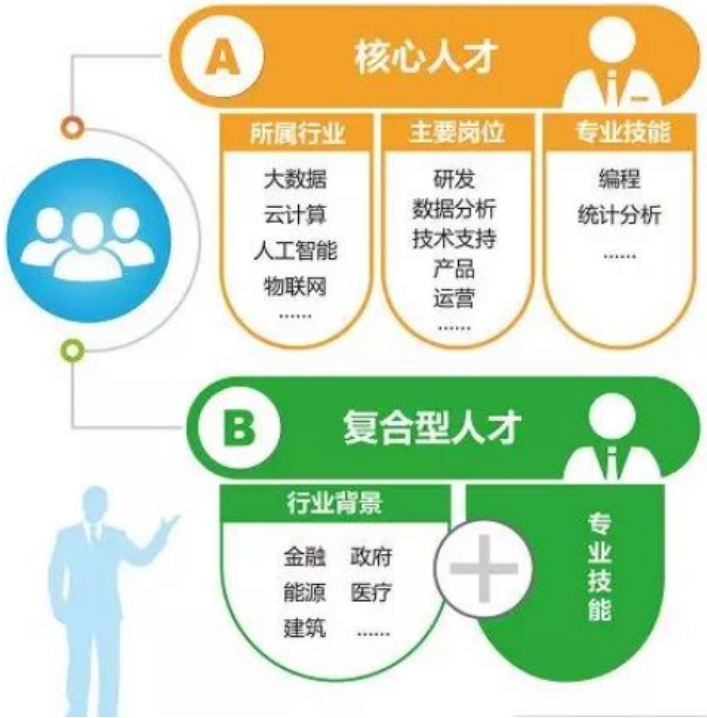
推进土地、劳动力、资本、技术、**数据**等要素市场化改革。

数据科学基础

- 大数据人才缺口
  - 市场对大数据人才的需求日益增加，供求关系不成正比，2025年人才缺口可达到230万
  - 同时，产业发展对大数据人才提出更高要求







数据科学基础

- **目前，全国已有众多省份成立了大数据分析相关的省级重点实验室**

北京市	大数据管理与分析方法研究重点实验室
上海市	上海市数据科学重点实验室
广东省	广东省大数据分析与处理重点实验室
江苏省	江苏省大数据分析技术重点实验室
浙江省	浙江省大数据智能计算重点实验室
湖南省	大数据研究与应用湖南省重点实验室
…	…
安徽省	**大数据分析与应用安徽省重点实验室**

http://bigdata.ustc.edu.cn

数据科学基础

- 大数据人才缺口
  - 传统的工科人才培养模式难以满足新兴产业的发展要求，亟需探索大数据工科人才培养的新途径和新模式
  - 大数据工科人才培养面临的挑战

校内外都存在针对大数据技术的广泛需求，但专业划分偏窄、偏细，使得现有大数据相关的技术体系和积累较为零散，给跨学科交叉创新与技术应用带了诸多限制

新工科时代涌现出大量的新兴经济行业（如共享经济、数字经济），使得传统的专业设置略显陈旧，培养的人才知识体系不能有效应对新兴行业的大数据人才需求

很多工科学生的理论知识扎实，但实验实训、人文基础不够，创新性受限，向专业数据科学研发人才的转化率较低

数据科学基础

- 大数据新工科人才需要具备以下素质

理论基础扎实，能理解运用数据科学中的理论模型

实践能力强，具有处理大数据的能力

跨界能力强，能够解决特定行业的大数据应用问题

数据科学基础

- 国外的著名高校大部分设立了大数据科学相关专业和机构
  - 斯坦福大学、麻省理工学院、加州大学伯克利分校等
- 2015年，复旦大学成立了大数据学院和大数据研究院
- 2016年，北京大学、对外经济贸易大学及中南大学分别成功申请了数据科学与大数据技术专业
- 2017年3月，教育部公布了第二批32所高校新增数据科学与大数据技术专业
- 2018年3月，教育部公布了第三批248所学校新增数据科学与大数据技术专业
- **2018年，中国科学技术大学成立大数据学院 http://sds.ustc.edu.cn/**

数据科学基础

- 数联寻英《大数据人才报告》
  - 当前全国的大数据人才仅46万,3-5年内大数据人才的缺口将高达150万
- 中国商业联合会数据分析专业委员会
  - 未来中国基础性数据分析人才缺口将达到1400万
- 在BAT企业招聘的职位里，60%以上都在招大数据人才

改变这个世界的四种力量

暴力

知识

世界著名未来学家托夫勒
《第三次浪潮》作者

大数据

金钱

数据科学基础

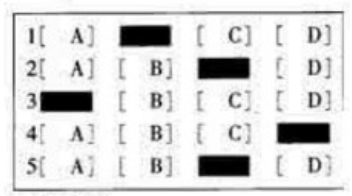
- 数据蕴含着巨大的价值
  - 健康医疗方面
    - 病人数据资料推动个性化药物治疗

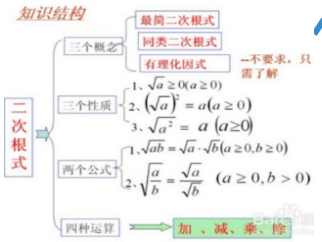
第三次药物革命的代表将是靶向的、个性化的药物。这些药物可以针对每个人的基因进行定向治疗，使治疗能够更加精准、有效且副作用更少。现在看似同样的疾病、同样的治疗，对不同的患者可能会产生完全不同的治疗结果，这就是因为每个个体都是有差异的，年龄、性别、体重、饮食结构等都不同，更不用说基因遗传的不同了。

数据科学基础

- 数据蕴含着巨大的价值
  - 教育方面：“因材施教”

学生的**学习行为数据**

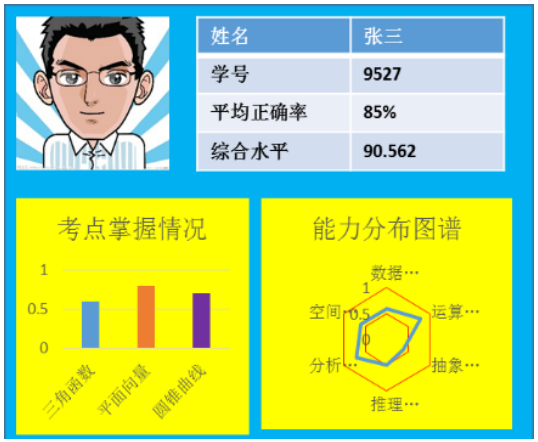




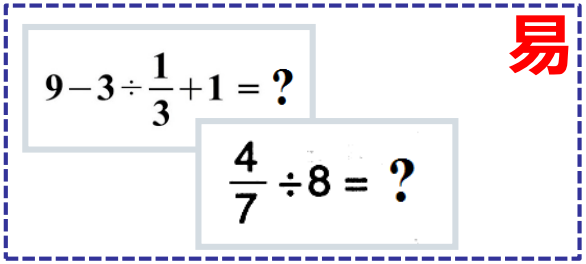
试题-知识点

大数据分析

学生认知水平画像



试题难度等特征的预测

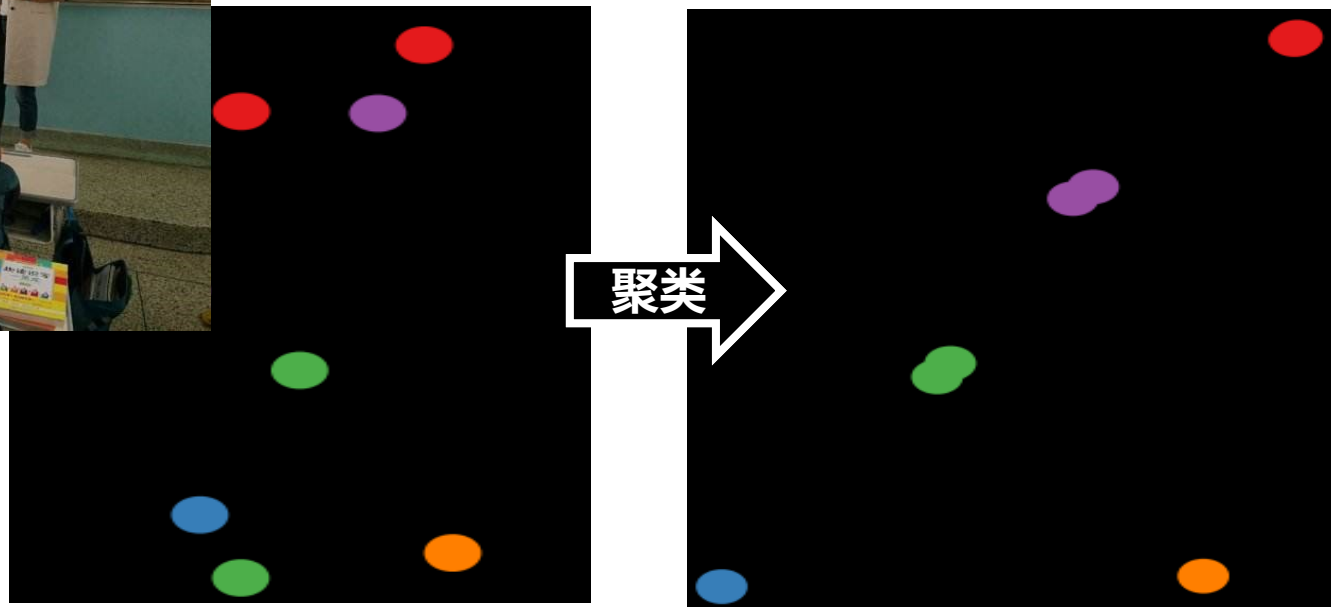


个性化学习推荐

数据科学基础

- 数据蕴含着巨大的价值
  - 教育方面：“优化教师教学”

根据考试数据对班级进行简单聚类，根据聚类结果，发现70%的类里，两个班集是同位授课教师



数据科学基础

- 数据蕴含着巨大的价值
  - 教育方面：“优化教师教学”

浙江省教育考试院
ZHEJIANG EDUCATION EXAMINATIONS AUTHORITY

组织机构 信息公开 政策法规 政策解读 2018年11月27日 星期二 11:11:39 请输入关键字

高考 | 学考选考 | 研究生考试 | 成人高考 | 自学考试 | 社会考试 | 教师资格考试 | 海外考试

关于英语科目考试成绩的说明

[发布时间:2018-11-27 阅读量:1570]

浙江省高考英语科目一年安排2次考试，考生可报考2次，选用其中较高1次的成绩。在2018年11月刚结束的英语科目考试中，根据答卷试评情况，发现部分试题与去年同期相比难度较大。为保证不同次考试之间的试题难度大体相当，浙江省招委组织专家研究论证，在制订评分细则时，决定面向所有考生，对难度较大的第二部分（阅读理解）、第三部分（语言运用）的部分试题进行难度系数调整，实施加权赋分。其他试题未作调整。

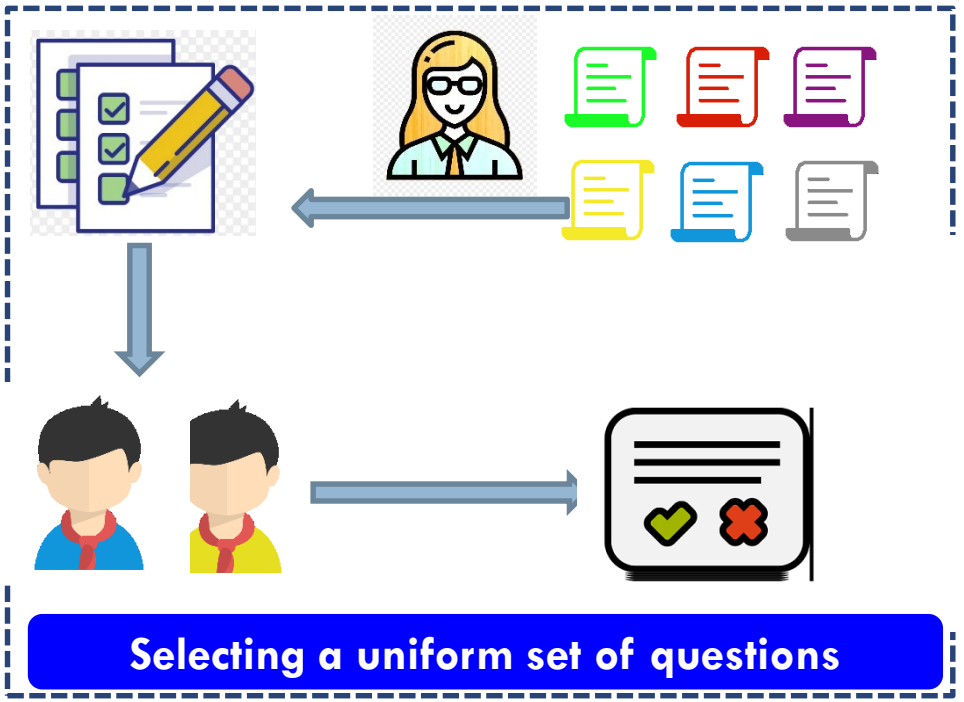
浙江省教育考试院
2018年11月27日

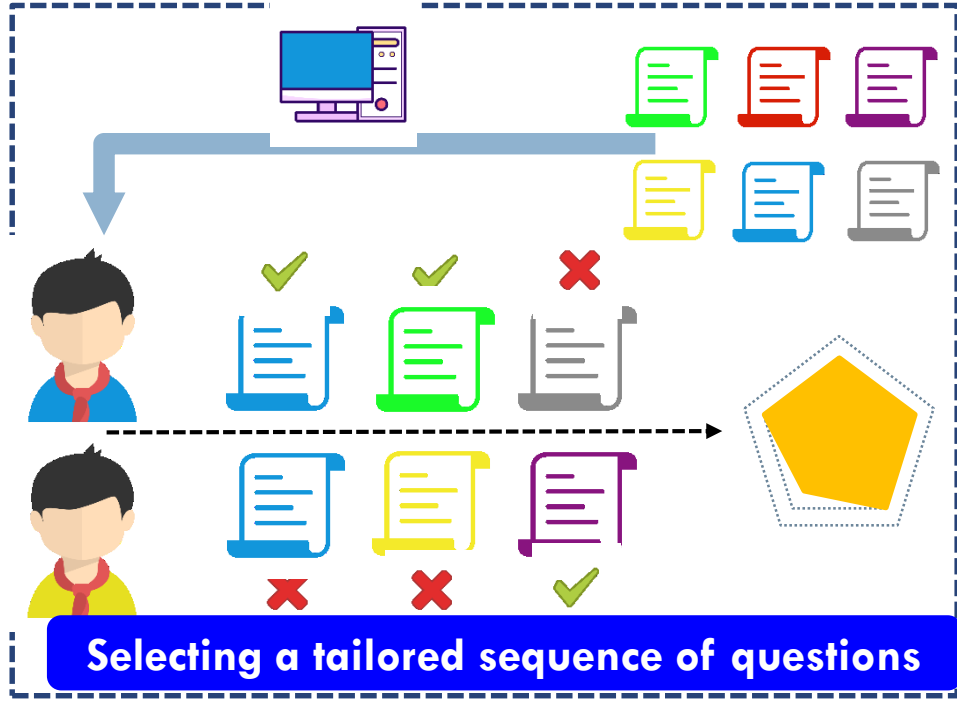
数据科学基础

- 数据蕴含着巨大的价值
  - 教育方面： “优化考试过程”
  - 数据驱动的自适应测试

GRE

Paper & Pencil Testing ← **Testing** → **Adaptive Testing**





数据科学基础

- 数据蕴含着巨大的价值
  - 社会科学方面
    - 社交媒体比问卷调查提供了更有代表性的结果
    - 智能引导社会成员的行为

15万名奥巴马支持者在Facebook安装了“奥巴马2012”应用，而通过这个程序，总统竞选团队可以间接得到这些支持者数百万的Facebook好友信息。

有一种说法称，特朗普的团队聘用数据分析公司，做了精准的广告投放，影响了那些徘徊不定的选民，拿下了决定性的关键州选举人票

数据科学基础

- 数据蕴含着巨大的价值
  - 社会科学：自动写稿、评论

机器人写稿时代来了！今日头条、腾讯、南周齐发力，媒体人将迎下岗潮？

2017-07-19 10:58 腾讯 / 机器人 / 媒体

数据科学基础

- 数据蕴含着巨大的价值
  - 影视娱乐：纸牌屋效应

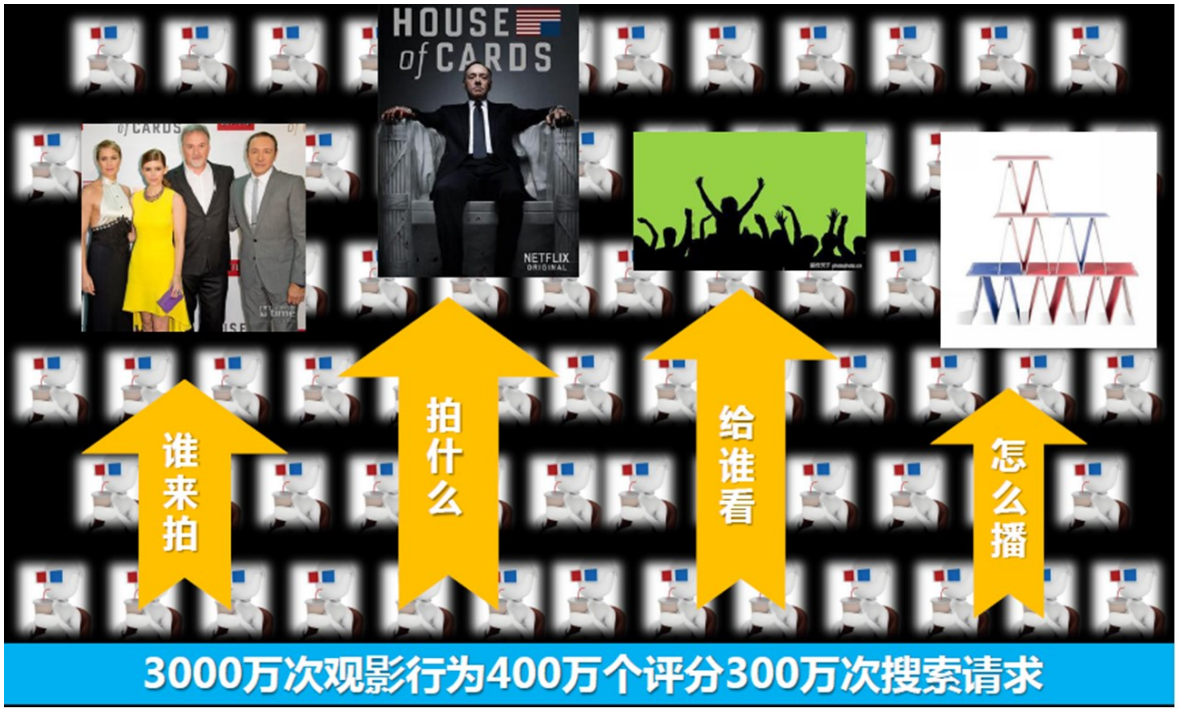
美国宫斗剧《纸牌屋》

- 出品公司：Netflix
- 导演：大卫·芬奇等
- 主演：凯文·史派西等
- 自2013年以来已播出四季
- 获得多项艾美奖提名
- 第一季总播放量：1906万
  平均每集播放量:146万
  平台：搜狐独家引进

美国把《纸牌屋》的成功归功于大数据的成就。虽然,任何一个项目的成功,绝对不是某单一因素决定的,但是《纸牌屋》代表了一个非常可喜的象征意义

数据科学基础

- 数据蕴含着巨大的价值
  - 影视娱乐：纸牌屋效应



数据科学基础

- 数据蕴含着巨大的价值
  - 电子商务方面：计算广告

媒体端

IMEI号和当前场景信息

①媒体端告诉后台，该机主的IMEI号和场景信息

②后台取出该IMEI号对应的用户画像标签

后台

用户画像标签：男、青年、已婚、无子、白领、喜欢汽车、无车、IT、喜欢理财、…

DMP

点击率预测

点击率预测模型

③后台取出当前激活的各个物料及其标签

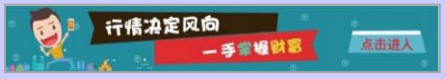
④进行点击率预测，并展现概率最高的物料

0.68%

0.52%

0.82%

0.60%



物料ID1，APP，理财，…

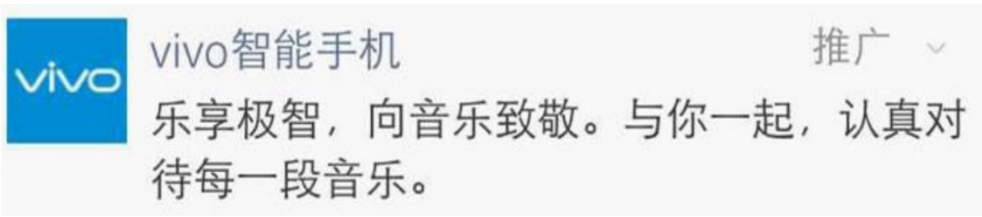
物料ID2，商品，食品，蟹，…

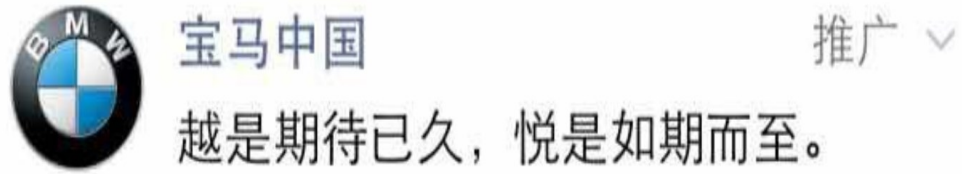
物料ID3，汽车，本田

物料ID4，工具，学习，词典，

数据科学基础

- 数据蕴含着巨大的价值
  - 电子商务方面：计算广告





数据科学基础

- 数据蕴含着巨大的价值
  - 电子商务方面：计算广告

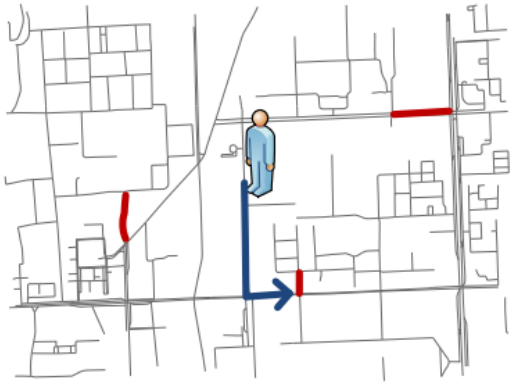
数据科学基础

- 数据蕴含着巨大的价值
  - 电子商务方面：精准搜索、个性化消费推荐

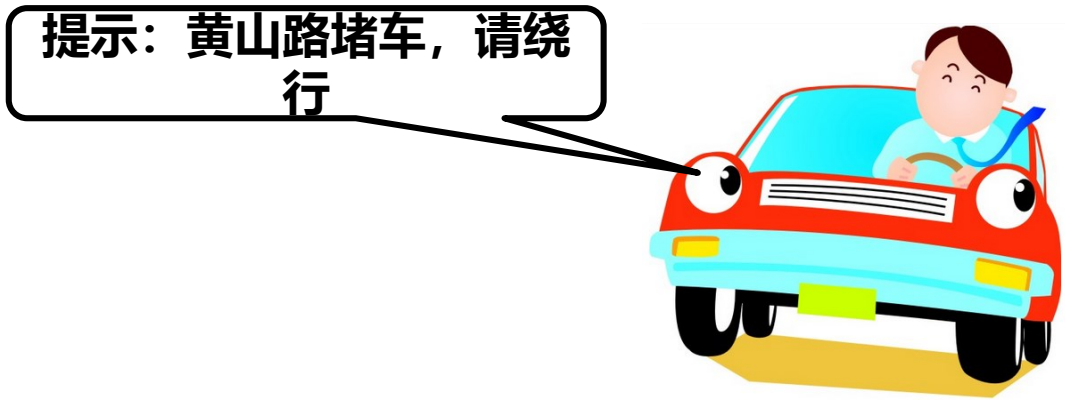
数据科学基础

- 数据蕴含着巨大的价值
  - 城市交通方面

A) Taxi recommender



B) Passenger recommender



- 数据蕴含着巨大的价值
  - 司法管理方面：智慧司法

案件：涉案当事人王某被法院判定赔偿人民币20万元，王某宣称无力支付。

从前：法院查询王某名下财产，包括银行存款、房产、现金等，发现其的确没有执行能力，但怀疑其有故意转移财产的嫌疑。王某社会关系复杂，这让执行法官无从下手，可能需要求助公安机关。

现在：创建涉案当事人画像系统后，打破了各部门之间数据壁垒，使得数据可以互通互联。执行法官通过系统获知王某的社交关系信息、房产交易信息等，最终查明王某通过给好处费的方法蓄意将变卖房产后的现金存放在友人处。王某故意隐匿财产，法院对其强制执行。

- 数据蕴含着巨大的价值
  - 公司管理方面：大数据智能化人力资源管理

大数据智能化招聘管理工具

外部招聘是互联网公司完善人才梯队、提升业务竞争力的重要手段。在传统招聘中，主要依赖招聘人员的知识储备与经验判断。因此，熊辉老师团队开发了智能化招聘的一系列管理工具，基于内外部大数据精准HR更有效、高效地获取高科技人才。

面试官评估与人才特征

该工具学习提取优秀面试官与优秀员工特征，帮助HR在招聘过程中更合理的安排不同阶段面试人员，并提供候选人考评依据

招聘市场趋势挖掘

该工具挖掘外部公司与时长任意时刻所处的招聘状态，与因此产生的招聘需求与招聘主题匹配。从而预测未来的招聘趋势

智能生产招聘广告

该工具把业务部门提出的模糊的、片面的招聘需求转化为正面的、清晰的人才筛选标准，并智能生成相应招聘广告

招聘人才圈子发现

构建人才转移网络，挖掘网络中存在的人才转移圈子，为公司招聘与员工求职提供依据，同时可以用来预测人才流动

预测外部人才跳槽倾向

该工具通过挖掘外部人才社交网络信息，智能预测人才一段时间内的跳槽倾向，从而为公司有针对性地竞争优秀人才提供依据

数据科学基础

- 数据蕴含着巨大的价值
  - 熊辉教授把人力资源管理细分为三大层次
    - 对人的管理（录、离、升、降、调）
    - 对组织的管理（组织领导力、组织结构稳定性、组织激励机制）
    - 对文化的管理（企业愿景、合理的价值评估和价值分配）

ACM杰出科学家
美国罗格斯大学商学院教授
海外杰青、长江学者
中科大大师讲席教授
百度公司商业智能实验室主任

大数据 / 人工智能分析技术 ＋ 人力资源管理领域知识 ＝ 大数据驱动智能化人才管理

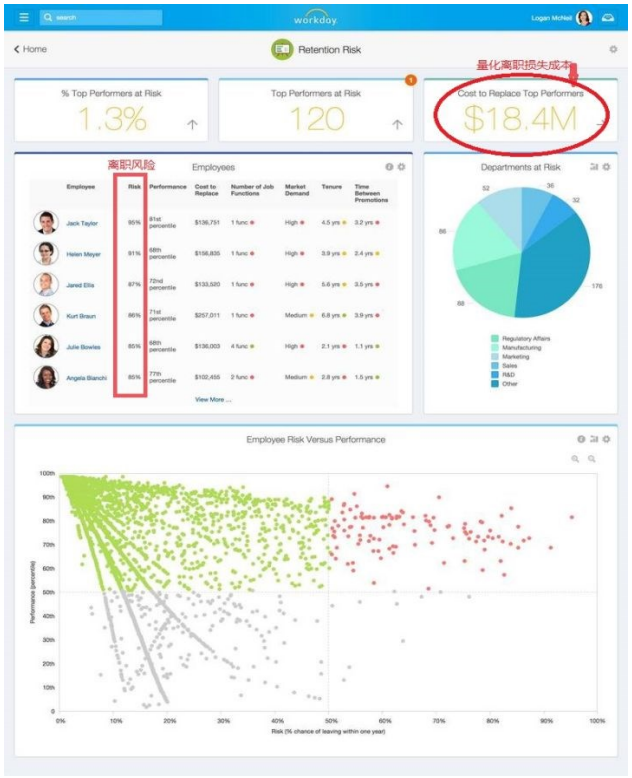
用数据说话 → 用数据决策 → 用数据管理 → 用数据创新

数据科学基础

- 数据蕴含着巨大的价值
  - 公司管理方面：大数据智能化人力资源管理

分析出了离职指数最高的前30名员工，3个月内其中29人向人力部门提出离职申请

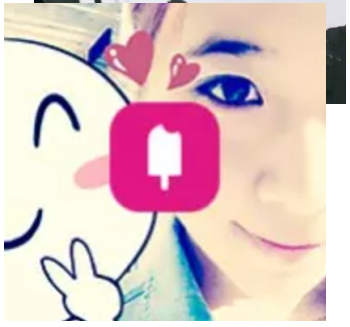
DVD4.0机器学习之离职分析预测系统截图



Workday公司离职风险评估系统截图

数据科学基础

- 数据
  - 智



偶然心儿跳动的时候我看到了你
舞台上不必张惶
眼睛充满生命的火焰
幻化成水滴飘在空中
——少女诗人小冰

长按二维码得首诗

来也

您最贴心的私人助理「来也」

我们能帮您：打车、订机票、订火车票、订酒店、叫外卖、下午茶、买生鲜、夜宵、订座、发快递、挂号、送药、保洁、跑腿、维修等等，还可以帮您设置各种提醒。我们相信「来也」能让您的生活变得更简单更美好。

数据科学基础

- 数据蕴含着巨大的价值
  - 智能助手、艺术创作方面
    - 流行音乐的旋律与编曲生成

【KDD18最佳论文揭晓】中科大等斩获最佳学生论文，刘兵获创新奖，清华大学唐杰任副主席

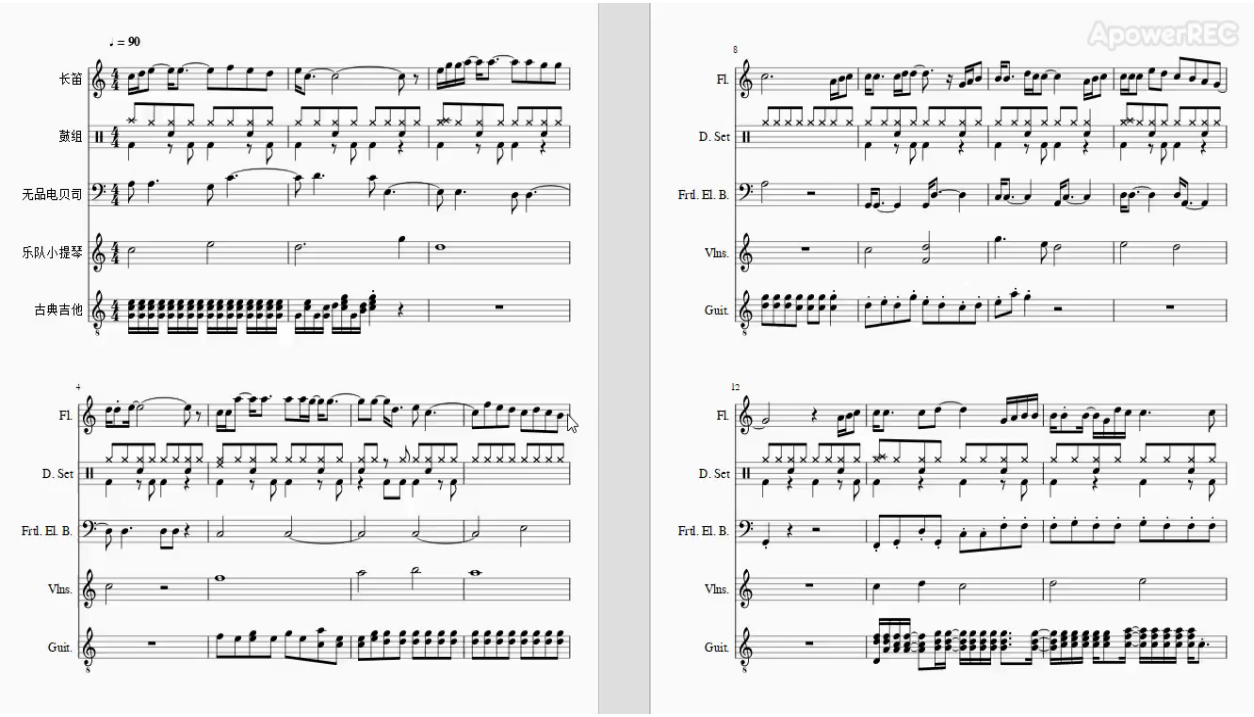
新智元 2018-08-21 06:24:19 阅读(0) 评论(0)

首页 新闻博览

我校获数据挖掘领域顶级国际会议KDD 2018最佳学生论文奖

数据科学基础

- 数据蕴含着巨大的价值
  - 智能助手、艺术创作方面
    - 流行音乐的旋律与编曲生成



More demos:

https://musicshow.chinacloudsites.cn/

数据科学基础

- 数据蕴含着巨大的价值
  - 智能助手、艺术创作方面
    - 机器作诗

用户写作意图

江、船、秋风

春风、杨柳

松、竹、山、牧童

诗歌自动生成系统

对应诗词

江北江南万顷秋，
船头人去水悠悠。
一帆一棹秋风急，
又有离人万里愁。

杨柳千条拂地垂，
一川春水浸桃花。
游人不识湖中路，
游遍人间野水涯。

乔松古木两三间，
松竹阴中一径斜。
白鸟不知山路远，
牧童踏过野人家。

- 机器作诗 PK 古代诗人

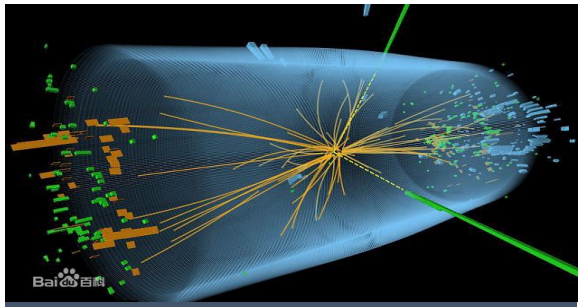
哪首诗是人写的?

秋夕湖上	秋夕湖上
一夜秋凉雨湿衣， 西窗独坐对夕晖。 湖波荡漾千山色， 山鸟徘徊万籁微。	荻花风里桂花浮， 恨竹生云翠欲流。 谁拂半湖新镜面， 飞来烟雨暮天愁。
机器	**宋代诗人葛绍体**

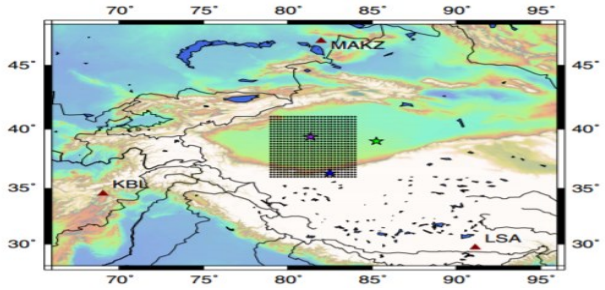
数据科学基础

- 数据蕴含着巨大的价值
  - 科学技术研究方面
    - 大数据推动科学新技术发现

天文大数据搜索新星



物理大数据预测分子属性



大数据地震速报、余震预测

生物大数据改良基因

科学
大数据

专利数据挖掘保护知识产权

数据科学基础

- **科技大数据来自于物理世界**
  - 科学实验数据或传感数据
  - 技术描述型数据—专利、论文
- **集多种特点于一身**
  - 采集的高代价性
  - 复杂性
    - 超高维度
    - 高度计算复杂性
    - 高度的不确定性
  - 学科知识壁垒
  - 信息与通信技术高度集成性

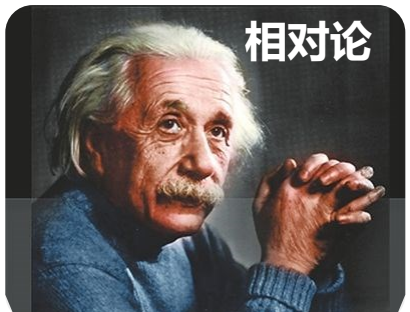
关系型数据库的鼻祖Jim Gray（右）

单一学科 → 数据驱动

多学科交叉

数据科学基础

- **2007年，Jim Gray总结出了四个科学范式**

几千年前	几百年前	几十年前	今天
经验科学	**理论科学**	**计算科学**	**数据密集型科学**
• **第一范式** • 以**归纳法**为主，带有盲目性的观测和实验 • **科学实验**	• **第二范式** • 以**演绎法**为主，关注理论总结和理性概括 • **数学模型**	• **第三范式** • 重视**数据模型构建**、**定量分析方法**，利用计算机来分析和解决 • **科学计算**	• **第四范式** • 先有了**大量的已知数据**，然后通过计算得出之前未知的理论 • **机器学习**
 自由落体实验	 相对论	 冯诺依曼计算机	 天文大数据

数据科学基础

- 把握大数据带来的机遇
  - 零售业
    - Winners: Amazon, Ebay
    - Losers: 传统书店、电子产品零售店
  - 旅游业
    - Winners: Expedia, Ctrip
    - Losers: 旅行中介商
  - 金融服务业
    - Winners: E*trade, TD Ameritrade
    - Losers: 股票中介商公司

数据科学基础

视频数据

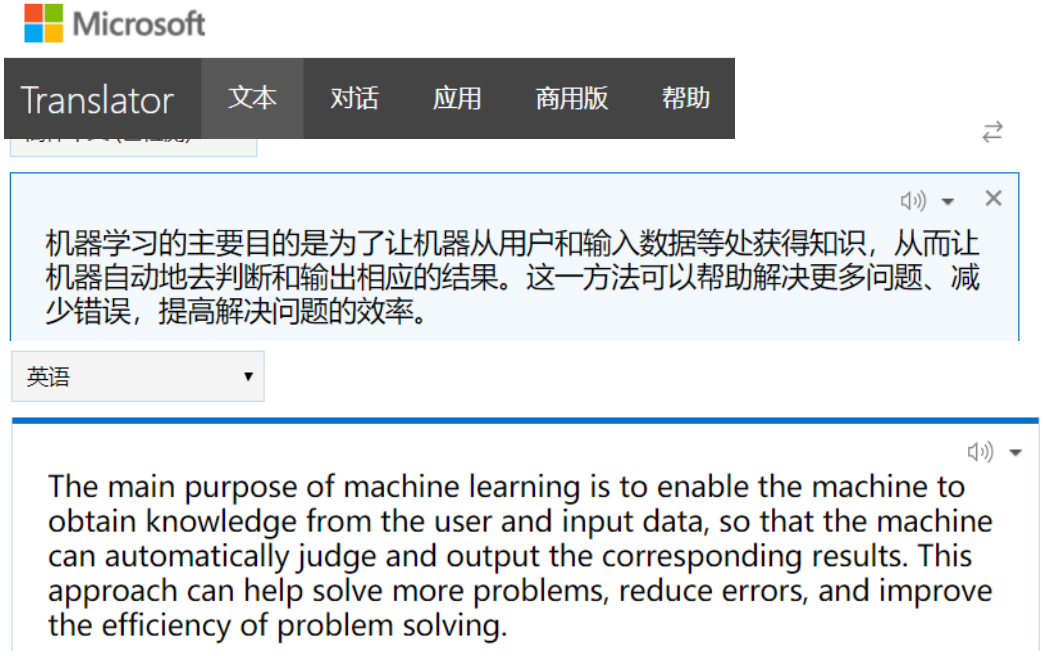
- 把握大数据带来的机遇
- 影像租赁业
  - Winners: 视频流媒体公司(Netflix, Amazon, Hulu)
  - Losers: DVD租赁公司
- 软件应用业
  - Winners: 软件数据服务公司(Saleforce.com)
  - Losers: 软件产品公司
- 新闻报纸业
  - Winners: Google, Twitter, Facebook, Bloomberg
  - Losers: 传统报纸业, Washington Post, WSJ
- 出租车行业
  - Winners: Uber, DiDi

数据科学基础

- 大数据带来的技术创新-当前进展
  - 语音识别
    - 微软英语语音识别实现词错率5.9%的突破，第一次超越人类。近来，科大讯飞等的语音识别词错率仅有3%左右

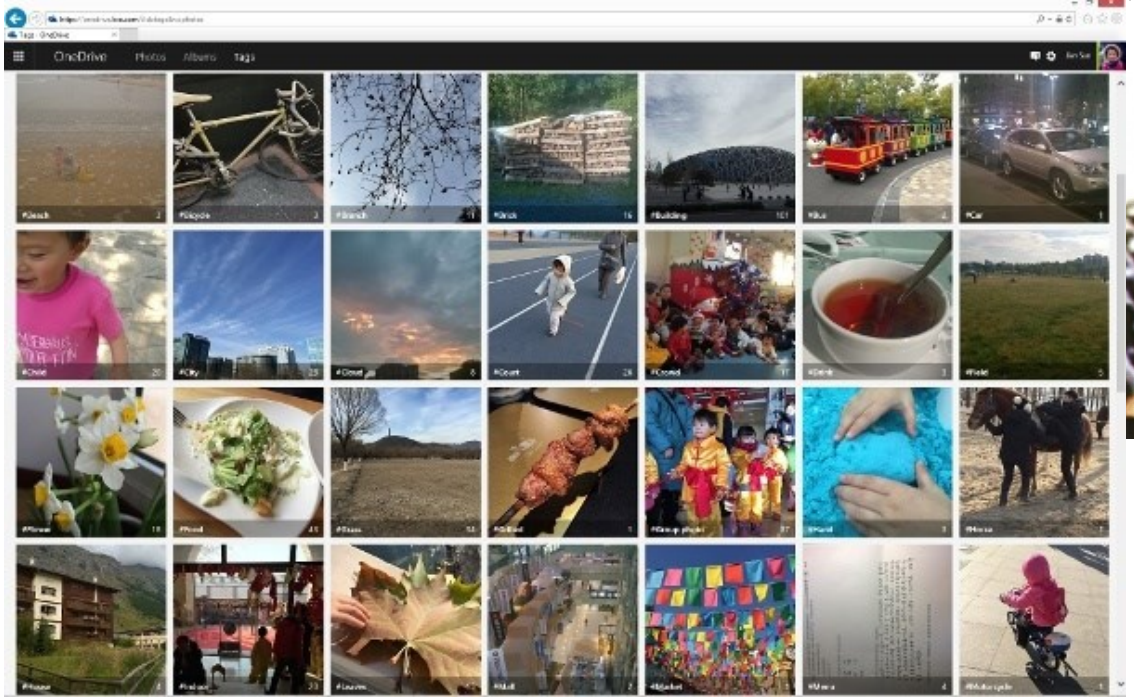
数据科学基础

- 大数据带来的技术创新-当前进展
- 机器翻译
  - 2018年3月，微软亚洲研究院与雷德蒙研究院宣布，其共同研发的机器翻译系统在通用新闻报道测试集newstest2017的中-英测试集上，**达到了可与人工翻译媲美的水平**



数据科学基础

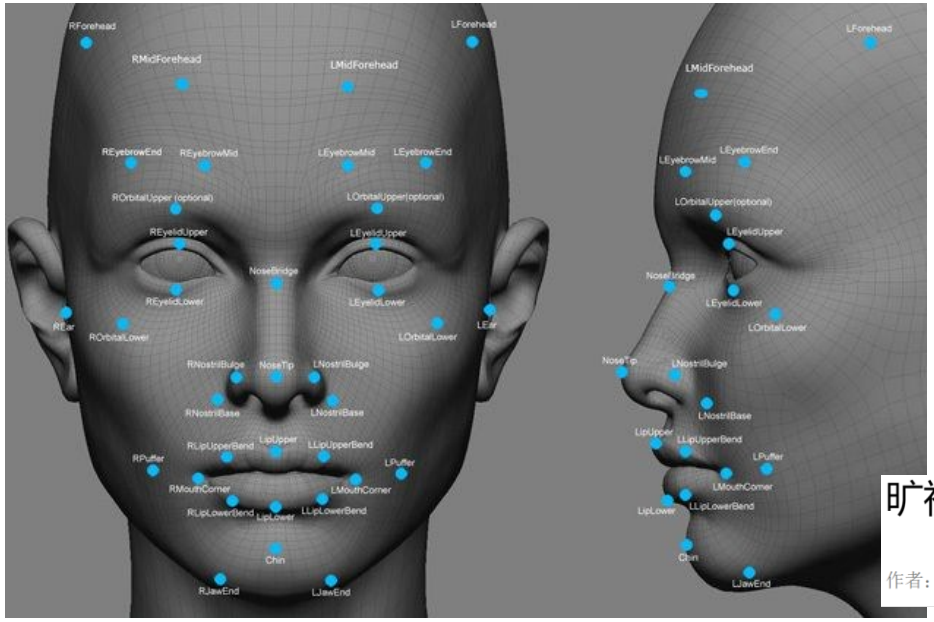
- 大数据带来的技术创新-当前进展
- 图像识别
  - ImageNet图像数据库上，人工智能已达到2.99%的错误率（公安部三所），低于人类5.1%的错误率



李飞飞
斯坦福大学、谷歌Ai前任首席科学家

数据科学基础

- 大数据带来的技术创新-当前进展
- 人脸识别
  - 2014年，Facebook人脸识别系统DeepFace达到97.53%准确率，达到人类水平
  - 不少科研人员从研究机构离职，或自己创办公司，或加入创业公司



SEETA

中科视拓

Face++ 旷视

旷视科技融资4.6亿美元布局城市大脑 创AI融资记录

作者：　　来源：中国证券报·中证网　　2017-11-01 10:55

数据科学基础

- 大数据带来的技术创新-当前进展
- 自然语言处理

通用语言理解评估（GLUE）基准

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B
1	Alibaba DAMO NLP	StructBERT		90.3	75.3	97.1	93.9/91.9	93.0/92.5
2	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8
3	ERNIE Team - Baidu	ERNIE		90.1	72.8	97.5	93.2/91.0	92.9/92.5
4	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5
5	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5
6	Huawei Noah's Ark Lab	NEZHA-Large		88.7	67.4	97.2	93.2/91.0	92.2/91.6
7	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1
8	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0
9	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9
10	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7
11	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6

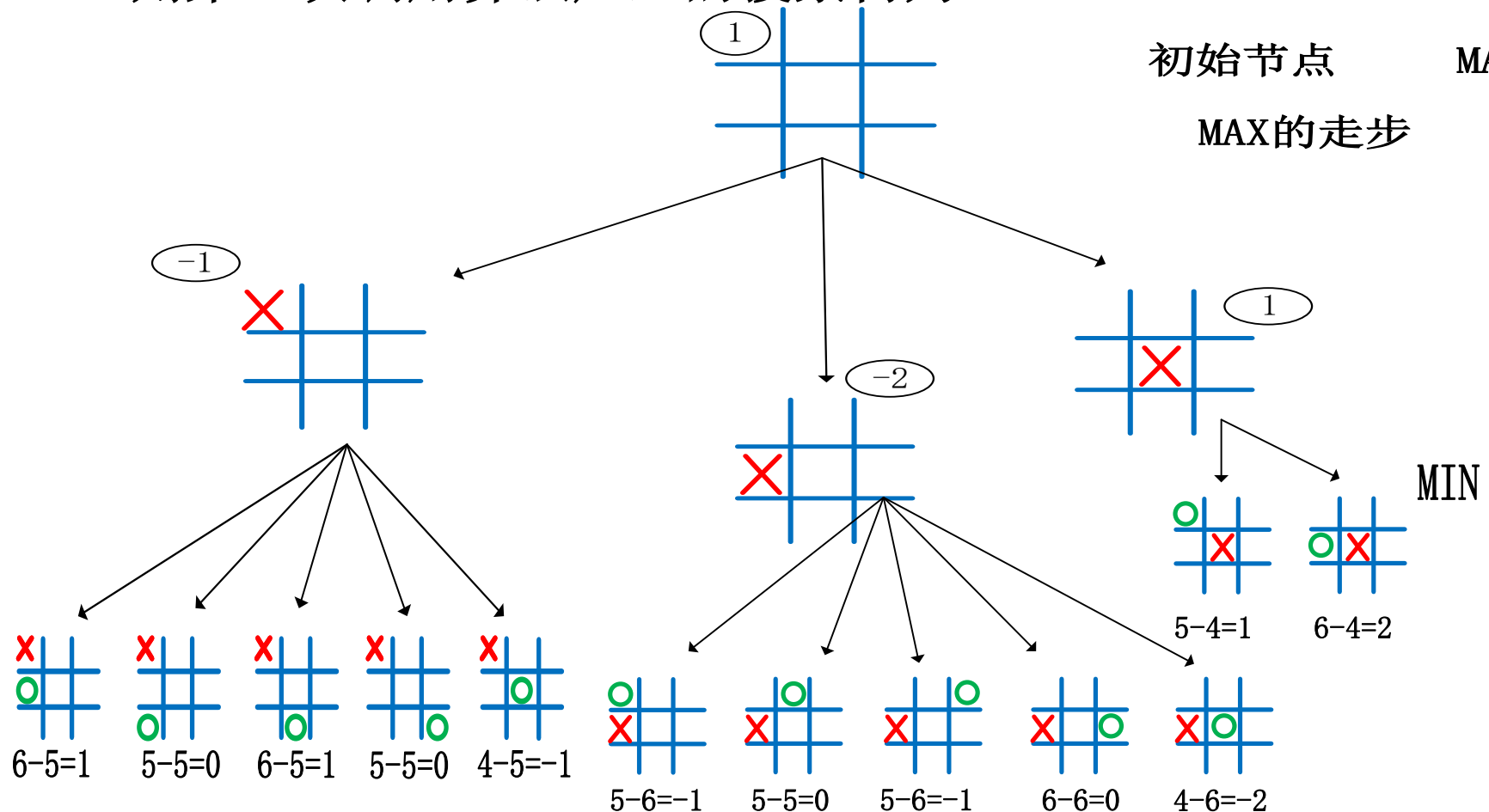
数据科学基础

- 大数据带来的技术创新-当前进展

人机对弈

- 2016年，AlphaGo以4：1的战绩击败李世石，机器第一次在围棋领域战胜人类顶尖高手
- 2017年5月，AlphaGo的升级版Master在围棋快棋上击败柯杰，聂卫平等高手，取得60胜0负的战绩
- 2017年10月，AlphaGo Zero从0学起，在不到3 天的时间内以100:0完虐AlphaGo

对抗搜索：设考虑走两步的搜索过程，利用棋盘对称性的条件，则第一次调用算法产生的搜索树为：



数据科学基础

- 大数据与人工智能
  - ABC当前AI的技术体系