大数据、智能化从概念到日常应用

何伏刚 博士

中国人民公安大学 副教授

http://hefugang.coding.me/

2017年3月5日/4月20日/5月6日

Hefg, Ph.D. Associate Professor in PPSUC

Teaching & Research

Work With Me

Fugang He (何伏刚)

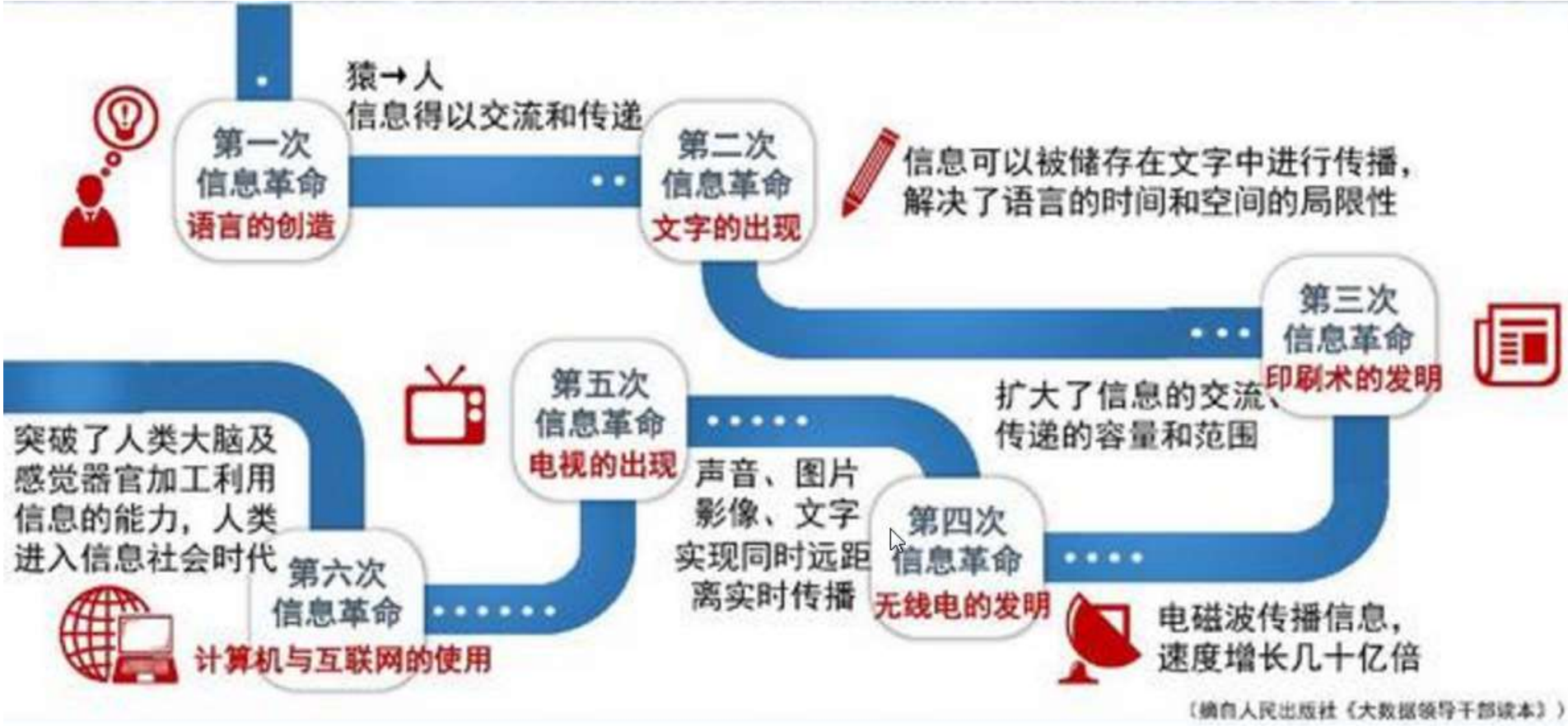
(hefugang@gmail.com), male, born in 1980, Ph.D, associate professor.

Fugang He received his PhD at Beijing Normal University. He is at present an Associate Professor at the Department of Police Command and Tactics in the College of Police Training, People's Public Security University of China. Dr. He's research focuses on the use of information systems for police training, police command and criminal law enforcement. He has papers in Modern Education Technology, Journal of People's Public Security University of China and Distance Education in China. He served as researcher in RCDE. He is also the

目录

目录

信息技术的发展


猿→人
信息得以交流和传递
第一次
信息革命
语言的创造
第二次
信息革命
文字的出现
信息可以被储存在文字中进行传播，
解决了语言的时间和空间的局限性
第三次
信息革命
印刷术的发明
扩大了信息的交流、
传递的容量和范围
第四次
信息革命
无线电的发明
电磁波传播信息，
速度增长几十亿倍
第五次
信息革命
电视的出现
声音、图片
影像、文字
实现同时远距
离实时传播
第六次
信息革命
计算机与互联网的使用
突破了人类大脑及
感觉器官加工利用
信息的能力，人类
进入信息社会时代
（摘自人民出版社《大数据领导干部读本》）

互联网的发展

互联网的发展过程，本质是让互动变得更加高效。

未来生活的一天

目录

1.1 认识大数据时代

- 大数据时代的生活令人神往，你对客观世界的认识更进了一步，所做的决策也不再仅仅依赖主观判断。甚至于你的一个习惯动作、你的一次消费行为、你的一份就诊记录，都正在被巨大的数字网络串联起来。移动互联网风潮汹涌。大数据正悄悄包围着我们。甚至连世界经济格局也在酝酿着巨大变革！

最早提出“大数据”时代到来的是全球知名咨询公司麦肯锡称：“**数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。**”

2012年3月份**美国奥巴马政府**发布了**“大数据研究和发展倡议”，**投资2亿以上美元，正式启动“大数据发展计划”。计划在科学研究、环境、生物医学等领域利用大数据技术进行突破。奥巴马政府的这一计划被视为美国政府继信息高速公路(Information Highway)计划之后在信息科学领域的又一重大举措。

2012年5月，联合国发表名为**《大数据促发展：挑战与机遇》**的政务白皮书中，指出大数据对于联合国和各国政府来说是一个历史性的机遇，还探讨了如何利用包括社交网络在内大数据资源造福人类。联合国的大数据白皮书还建议联合国成员国建设**“脉搏实验室”**“Pulse Labs”网络开发大数据的潜在价值

《大数据时代：生活、工作与思维的大变革》一书的作者维克托·迈尔·舍恩伯格，如是说，“**如果你是一个个人，如果你拒绝的话，可能会失去生命，如果是一个国家的话，拒绝大数据时代的话，可能失去这个国家的未来，失去一代人的未来。**” 这一句话恐怕不能算作耸人听闻，因为每当人们站在现在这个节点的时候，总会去眺望未来，但是未来往往在你不经意当中已经悄悄地来到你的身边。

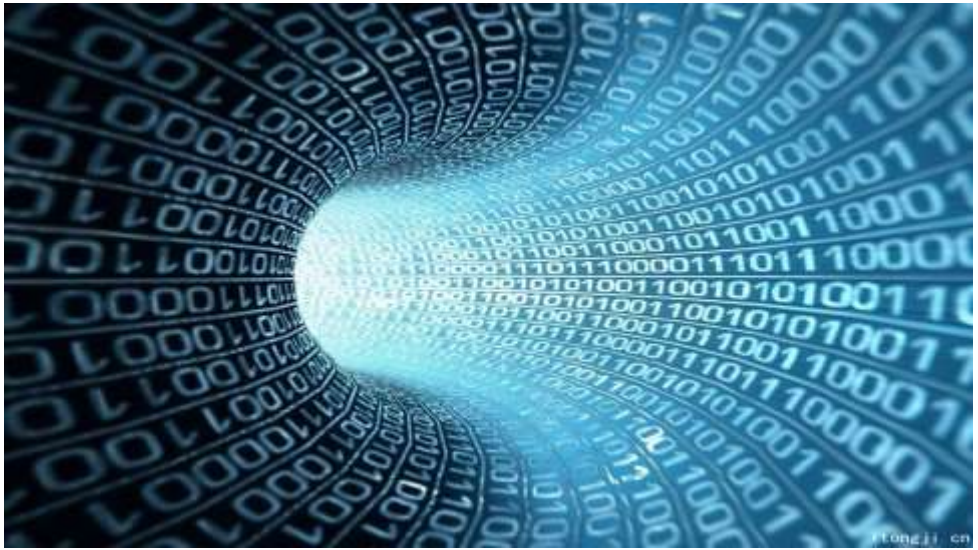
拒绝大数据时代，可能会失去生命！

大数据时代到来的必然性：

- 硬件成本的降低
- 网络带宽的提升
- 云计算的兴起
- 网络技术的发展
- 智能终端的普及
- 电子商务、社交网络、电子地图等的全面应用
- 物联网

随着一系列标志性事件的发生和建立，人们越发感觉到大数据时代的力量。因此2013年被许多国外媒体和专家称为**“大数据元年”**。

当今『大社会』，
三分技术，七分数据，
得数据者得天下。



1.2 什么是大数据

- 大数据很抽象，表示数据规模的庞大。
- 大数据泛指巨量的数据集，因可从中挖掘出有价值的信息而受到重视。《华尔街日报》将**大数据时代、智能化生产、无线网络革命**称为引领未来繁荣发展的重大技术变革。

目前对大数据尚未有一个公认的定义，不同的定义基本上是从特征出发，试图给出大数据的定义。

大数据的定义：

- 维基百科给出的定义：大数据是指利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间的数据集。
- 大数据 (big data)：是所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理、并整理成为帮助企业经营决策有更积极目的的资讯。大数据的4V特点：Volume（大量）、Velocity（高速）、Variety（多样）、Value（价值密度低）。

规模性（Volume）

高速性（ Velocity ）

多样性（ Variety ）

价值性（Value）（IDC）

真实性（Veracity）（IBM）

“4V”定义

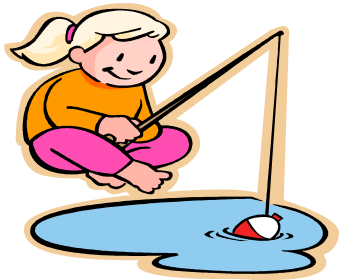
大数据的数据有多大？

我国网民数量居世界之首，每天产生的数据量也位于世界前列。

淘宝网站	◆单日数据产生量超过5万GB ◆存储量4000万GB
百度公司	◆目前数据总量10亿GB ◆存储网页1万亿页 ◆每天大约要处理60亿次搜索请求
一个8Mbps的摄像头	◆一小时能产生3.6GB的数据 ◆一个城市每月产生的数据达上千万GB
医院	◆一个病人的CT影像数据量达几十GB ◆全国每年需保存的数据达上百亿GB

从数据库(database，DB)到大数据(big data，BD)

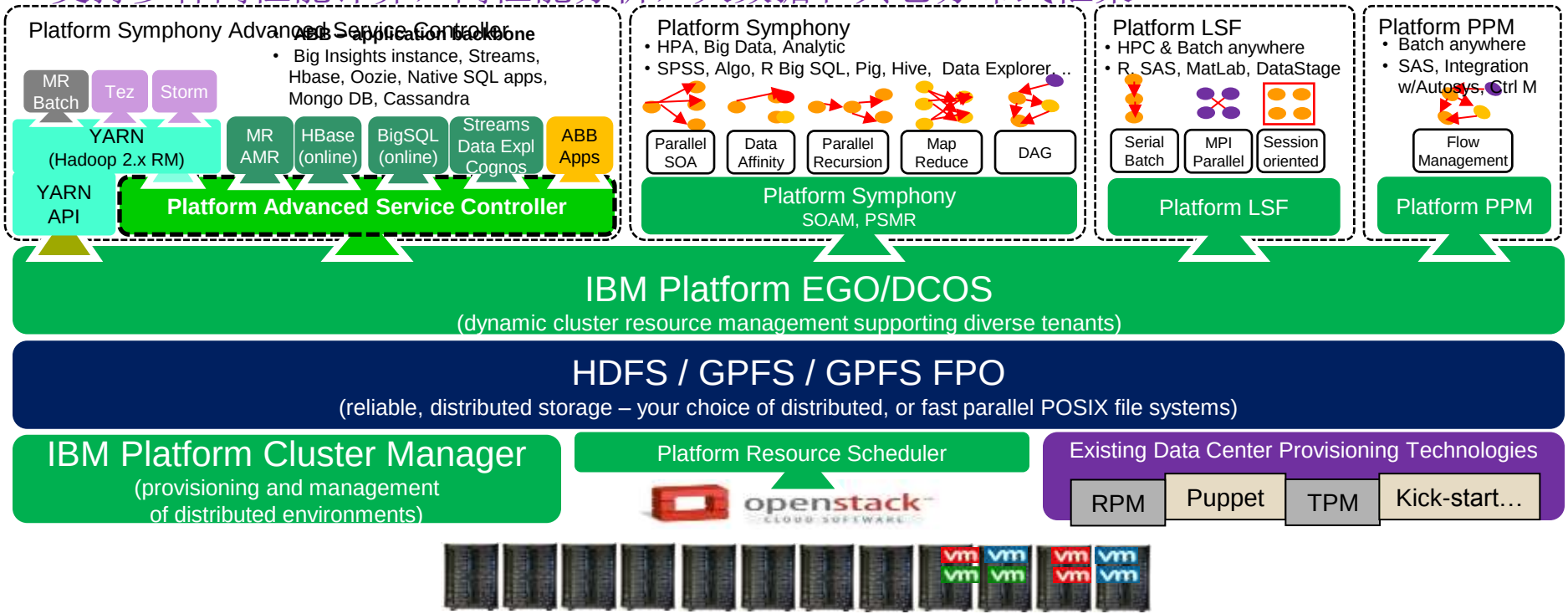
“池塘捕鱼”VS“大海捕鱼”　　“鱼”是待处理的数据



数据规模	小（以MB为处理单位）	大（以GB、TB、PB为处理单位）
数据类型	单一（结构化为主）	繁多（结构化、半结构化、非结构化）
模式和数据的关系	先有模式后有数据 (先有池塘后有鱼)	先有数据后有模式 模式随数据增多不断演变
处理对象	数据（池塘中的鱼）	(“鱼”，通过某些“鱼”判断其他种类的“鱼”是否存在)
处理工具	One size fits all	No size fits all

IBM 大数据平台架构

支持多种高性能计算，高性能分析，大数据和其它分布式框架



实际生产环境验证的多租户，共享资源框架。支持包括**Hadoop**在内的分布式负载。

《纽约时报》2012年2月的一篇专栏中所称，“大数据”时代已经降临，在商业、经济及其他领域中，**决策将日益基于数据和分析而作出，而并非基于经验和直觉。**

哈佛大学社会学教授加里·金说：“这是一场革命，庞大的数据资源使得各个领域开始了**量化进程**，无论学术界、商界还是政府，所有领域都将开始这种进程。”

亚马逊前任首席科学家Andreas Weigend说：“**数据是新的石油**。”

1.3大数据的应用领域

大数据就如同蜜蜂，其主要价值是传播花粉，自己生产的蜂蜜价值并不大。

2013年世界范围内狭义的大数据产业产值只有186亿美元，但广义的大数据应用几乎覆盖所有产业。据麦肯锡公司预测，开放数据仅在教育、保健等7个行业便可释放3.2万亿～5.4万亿美元的经济价值。

1.3大数据的应用领域

教育学　公安　公共服务

天文学　电子政务　传媒业

生物医学　商业智能　国家安全

气候学　企业管理　市场营销

金融学　生活娱乐　总统选举

大数据的应用领域——政治领域

不要总喊“狼来了”，“狼”已经来了！

大数据帮助奥巴连任

分析川普获胜

对数以千万计的选民邮件进行了大数据挖掘，精确预测出了更可能拥护奥巴马的选民类型，并进行了有针对性的宣传，从而帮助奥巴马成为了美国历史上唯一一位在竞选经费处于劣势下实现连任的总统。

大数据的应用领域——政治领域

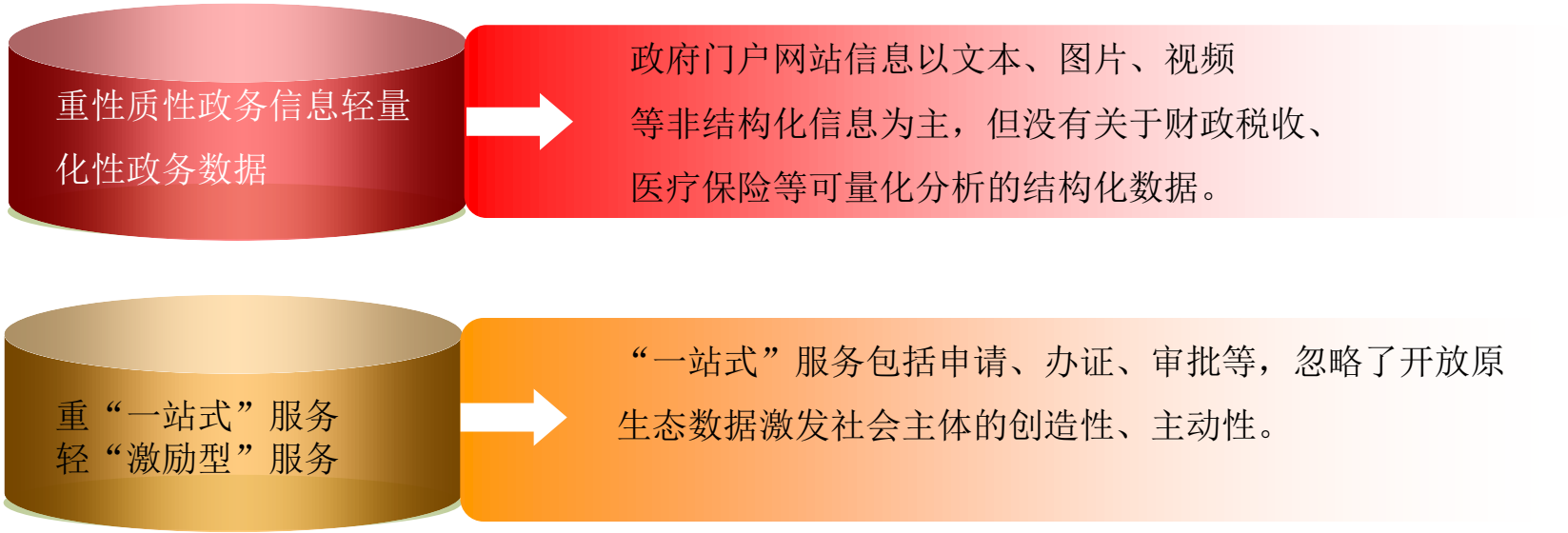
奥巴马的例子告诉我们，只要**数据量够大，够及时，挖掘够深刻**，我们完全可以洞悉每个选民的投票几率。迅速普及的互联网与移动互联网，悄然为记录人的行为数据提供了最为便利、持久的载体。最重要的是，在这些强大的数据收集终端面前，人们没有掩饰的意图，从而创造着过去无法收集与分析的海量数据，这**让所有社会科学领域能够从宏观群体走向微观个体，让跟踪每一个人的数据成为了可能，从而让研究人性成为了可能。**

大数据的应用领域——政治领域

大数据在我国电子政务领域的应用

大数据的发展，将极大地改变政府的管理模式，有利于节约政府投资、加强市场监管能力、提高政府决策能力、提升公共服务能力，实现区域化管理。

我国电子政务的发展现状:



大数据的应用领域——政治领域

目前，我国有60多个城市，将构建“智慧城市”的目标列入“十三五规划”中。智慧城市即利用大数据的整合和分析来治理社会。两会期间，有代表提议将发展大数据上升为国家战略。

大数据背景下，我国电子政务创新模式的转变：

1. 关注焦点——从信息向数据转变
2. 增值基础——从公开向发布、开放转变
3. 行为方式——从独立向协同转变

大数据的应用领域——政治领域

关注焦点——从信息向数据转变

数据是生成信息和知识的原生素材

如将我国城镇居民医疗数据与保险数据对比分析可以优化保险企业报销比例，发现虚假报销行为；与制药厂数据比对分析可以调节药品的生产量与销售渠道。但医疗政策信息并不能发挥这样的作用。

数据是舆情监督的有利依据

2008年，政府公布4万亿元的经济刺激计划，但社会对资金的具体去向及其准确数额却无从得知；汶川地震灾后重建近2亿资金的用途也因没有准确的数据公开而被暗箱操作，违规使用……当今电子政务中，信息公开实质是性质性信息的公开，而数据才是舆情监督的真正证据。

大数据的应用领域——政治领域

从信息向数据的转变是政府从后台走向阳光的转变。

大数据的应用领域——政治领域

增值基础——从公开向发布、开放转变

大数据时代，数据增值的关键在于数据的整合与分析，整合的前提就是数据的开放。

数据公开、发布是一条一条的；

数据开放是一片一片的。

政府态度从被动转为主动；

数据从点对点转为面对面。

数据公开是意识上的、被动的；

数据发布是行动上的、主动的。

大数据的应用领域——政治领域

行为方式——从独立向协同转变

内部协同：各地区政府、各层级政府和各部门之间

某市电子政务数据交换平台实现了工商、国税、质检、公安、社保等20多个部门涉税数据的共享，国税局与地税局通过数据比对，发现了25000条数据差异，落实纳税企业5000多户，补缴税款2700多万元。

外部协同：政府与社会之间自上而下、自下而上的互动

大数据的应用领域——金融领域

你开心他就买，你焦虑他就抛！

华尔街“德温特资本市场”公司首席执行官保罗·霍廷每天的工作之一，就是利用电脑程序分析全球3.4亿微博账户的留言，进而判断民众情绪，再以“1”到“50” 进行打分。根据打分结果，霍廷再决定如何处理手中数以百万美元计的股票。霍廷的判断原则很简单：如果所有人似乎都高兴，那就买入；如果大家的焦虑情绪上升，那就抛售。

这一招收效显著——当年第一季度，霍廷的公司获得了7%的收益率。

大数据的应用领域——金融领域

大数据助力推进高频金融交易和小额信贷

高频交易：实时性要求高、数据规模大。目前沪深两市每天4个小时交易时间会产生3亿条以上逐笔成交数据，通过对历史和实时数据的挖掘创新，以创造和改进数量化交易模型，并将之应用于基于计算机模型的实时证券交易过程中。

小额信贷：阿里巴巴和建行在2007年推出一个专注于小企业的贷款计划——**e贷通**，阿里巴巴利用拥有的用户信息及交易数据，通过大数据技术自动判定是否给予企业贷款；而建行坐拥巨额资金，希望贷款给无信用记录但发展势头良好的小企业。到2012年底，阿里在累计服务小微企业超过20万家，放贷300多亿元，坏账率仅为0.3%左右，低于商业银行水平。

大数据的应用领域——金融领域

大数据协助金融企业精准营销

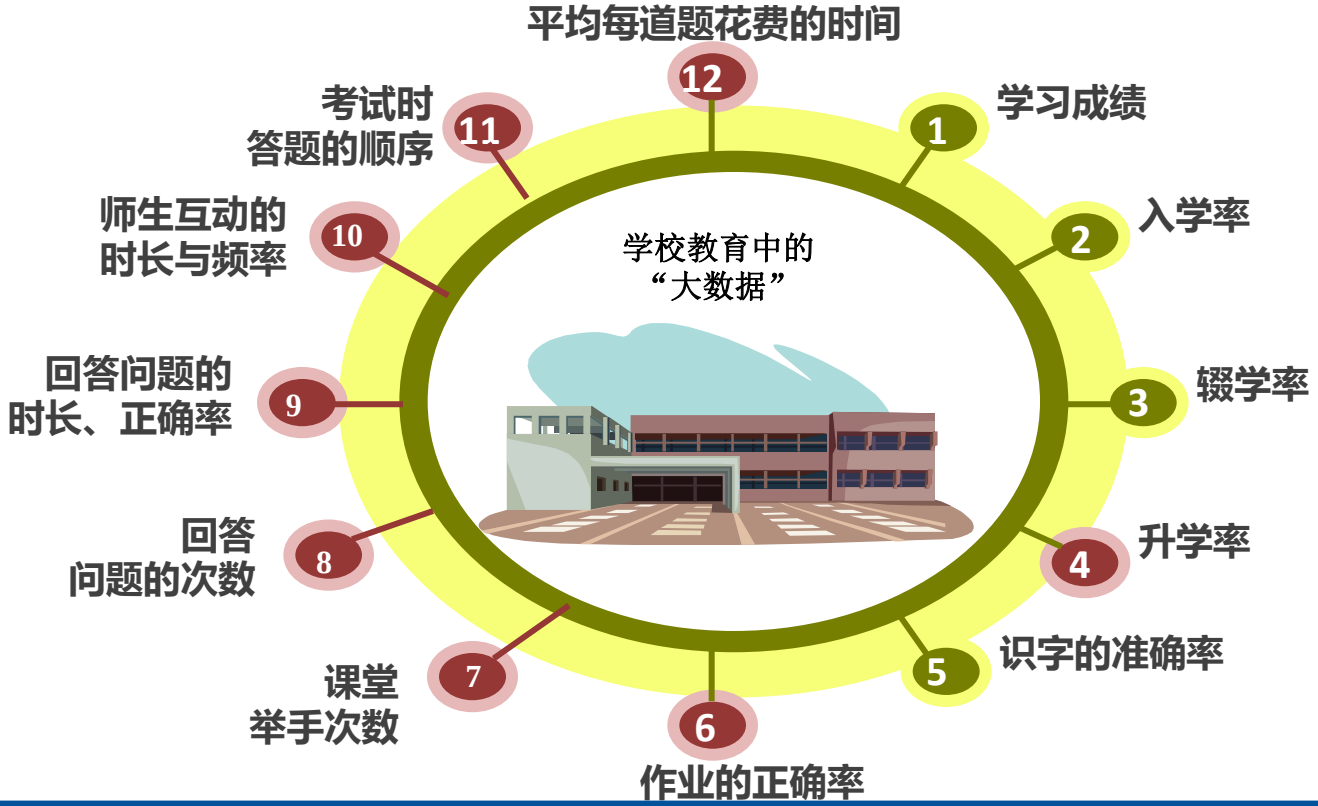
招商银行通过数据分析识别出招行信用卡高价值客户经常出现在星巴克、DQ、麦当劳等场所后，通过“多倍积分累计”“积分店面兑换”等活动吸引优质客户；通过构建客户流失预警模型，对流失率等级前20%的客户发售高收益理财产品予以挽留，使得金卡和金葵花卡客户流失率分别降低了15个和7个百分点；通过对客户交易记录进行分析，有效识别出潜在的小微企业客户，并利用远程银行和云转介平台实施交叉销售，取得了良好成效。

招商银行
CHINA MERCHANTS BANK

金葵花 理财
SUNFLOWER
专 注 您 所 关 注

大数据的应用领域——教育领域

大数据分析已经被应用到教育中，成为教学改革的重要力量。——北师大的案例



大数据的应用领域——教育领域

在加拿大，教育科技公司**“渴望学习”**（Desire 2 Learn）已经面向高等教育领域的学生，推出了基于过去的学习成绩数据预测并改善未来学习成绩的大数据服务项目。通过监控学生**阅读电子化的课程材料、提交电子版的作业、通过在线与同学交流、完成考试与测验，**就能让计算程序持续、系统地分析每个学生的教育数据。老师得到的不再是过去那种只展示学生分数与作业的结果，而是像阅读材料的时间长短等这样更为详细的重要信息。这样老师就能及时诊断问题的所在，提出改进的建议，并预测学生的期末考试成绩。

通过大数据你可以知道：

- 一个学生成绩不好是由于他因为周围环境而分心了吗？
- 期末考试不及格是否说明学生未掌握学习内容，还是因为他请了很多病假的缘故？

大数据的应用领域——生活方面

大数据首次播报春运迁徙实况

40天，36亿人次。这是2014年春运的总时间和总出行人数。在这场堪称人类历史上最大规模的短期迁徙中，**人群从哪儿去了哪儿？哪些线路最热门？**在以往，这些问题可能难以精确回答。但随着技术进步，通过应用“大数据”这一技术利器，人们已经接近“在迷宫中感受全局”地看见春运的全景。

大数据的应用领域——生活方面

国内有2亿手机用户使用百度地图，用户每次位置变化，百度都能得到数据。把手机网民的定位信息汇总成大数据进行分析，就能勾勒出人们的迁徙轨迹。

此次百度图景化地展示春运情况，是基于LBS(基于地理位置的服务)技术的一次创新。它的数据每8小时更新一次，囊括了全国铁路、公路和航空在内的线路。

新闻视频：2014年1月25日，《“据”说春运》

当你有了锤子，好像什么问题都看上去像钉子！

今天，大数据似乎成了“灵丹妙药”，“包治百病”，无所不能。但千万别把“大数据”用做解决世界上所有问题的全能办法，无论是管理城市到消除贫困，制止恐怖袭击、疾病流行到拯救地球环境等，以为有了“大数据”，就没有解决不了的问题，这也是一种误解。人类的思想、个人的文化和行为模式、不同国家及社会的存在发展都非常复杂、曲折和独特，显然不能全部由计算机来“数字自己说话”。**无论到何时，其实都还是人在思考和“说话”。**

当前，大数据的应用只是冰山一角，绝大部分隐藏在表面之下。

未来，大数据所带来的精彩值得期待！

目录

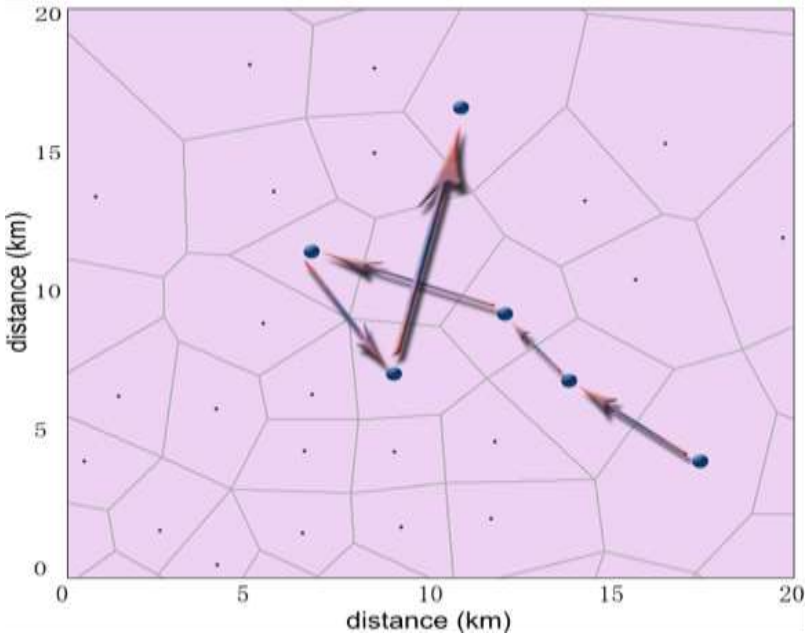
2.1大数据面临的机遇

案例：大数据在应急管理中的应用

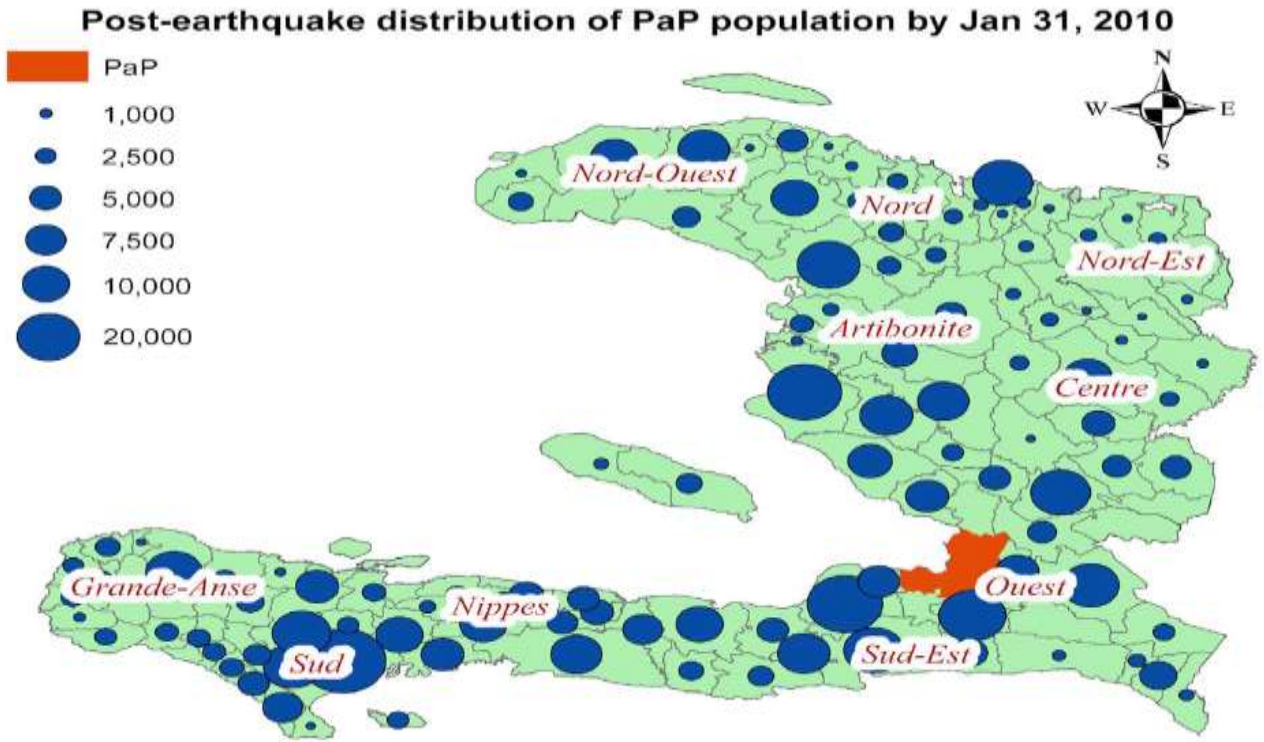
全球每年约有2.58亿人受到如台风、地震、洪水、干旱等自然灾害的影响，大部分自然灾害最直接的后果是导致人口移动和迁徙，社会管理系统趋于瘫痪，大量救援物资无法及时、准确分发给灾民。

手机数据辅助定位灾民移动和分布

- MOU & NDA with Digicel Haiti（海地）
- 约 10亿条通话记录!
- 数据库文件 250GB
- 单个文本文件25GB



震后灾民流动情况分析

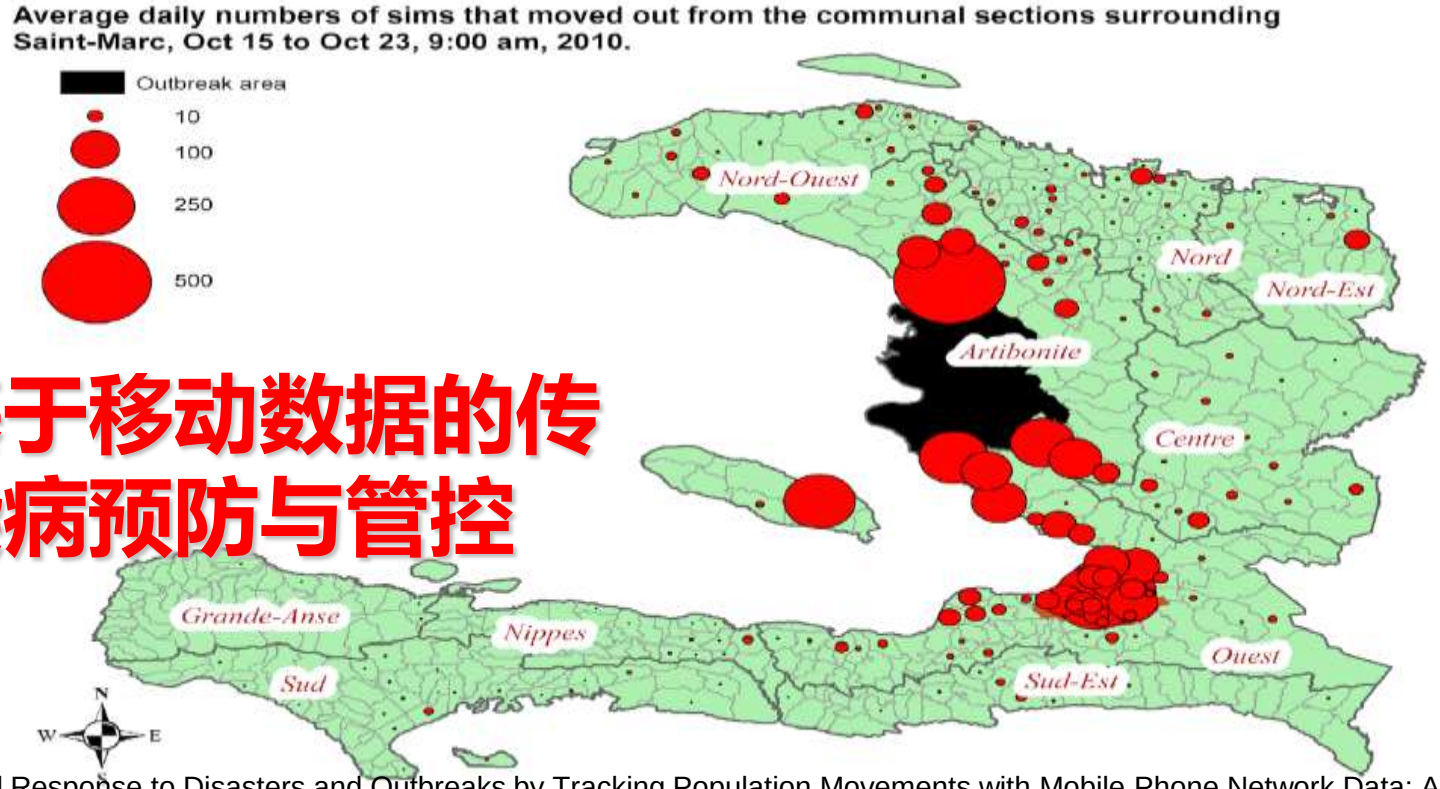


Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. **PLoS Medicine**, 2011; 8 (8): e1001083 DOI: 10.1371/journal.pmed.1001083

Haiti Earthquake - Population Movements out of Port-au-Prince - 31 January 2010

Karolinska Institutet

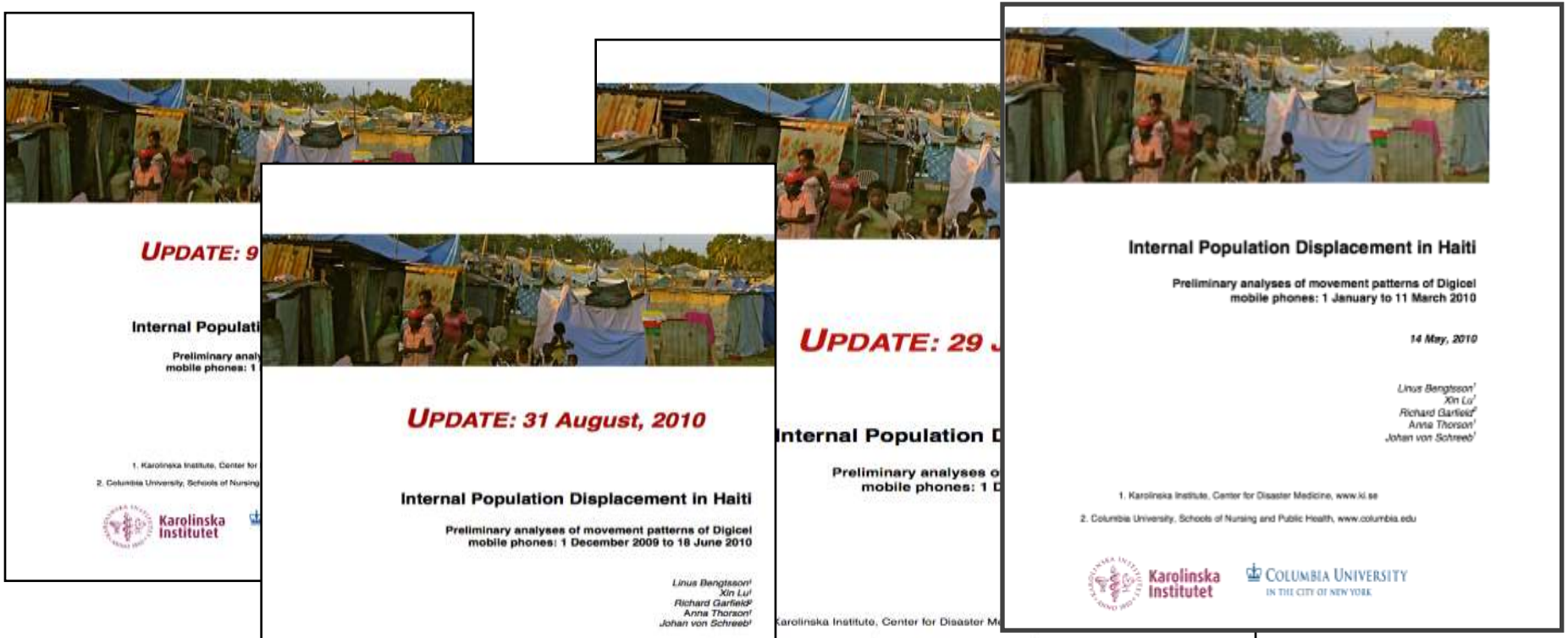
霍乱感染地区人口流动及风险评估



基于移动数据的传染病预防与管控

Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. **PLoS Medicine**, 2011; 8 (8): e1001083 DOI: 10.1371/journal.pmed.1001083

数据分析报告提交联合国救援组织



国际媒体关注

GLOBAL UPDATE

Haiti: Cellphone Tracking Helps Groups Set Up More Effective Aid Distribution, Study Says

By DONALD G. McNEIL Jr.
Published: September 5, 2011

Populations on the run during disasters can be tra
signals, which could help guide life-saving aid to t
new study has concluded.

CONNECT WITH FAST COMPANY:

FAST COMPANY

TECHNOLOGY | CO.DESIGN | CO.EXIST | LEADERSHIP | MAGAZINE | NEWSLETTE

How Tech Is Helping the Haiti Recovery

BY: JOCELYN C. ZUCKERMAN February 10, 2011

cording to the Red Cross. Phoneseanable
urtesy of the Bill and Melinda Gates

BBC Mobile | News | Sport | Weather | Travel | TV

NEWS TECHNOLOGY

Home | UK | Africa | Asia | Europe | Latin America | Mid-East | US & Canada | Business | Health | Sci/Environm

2 September 2011 Last updated at 16:22 GMT

236 Share

Mobile phones help to target disaster aid, says study

Mobile phones of people fleeing natural disasters can be used to target emergency aid according to a new study.

The report reveals how scientists mapped population movements following the Haiti earthquake based on location data from two million handsets.

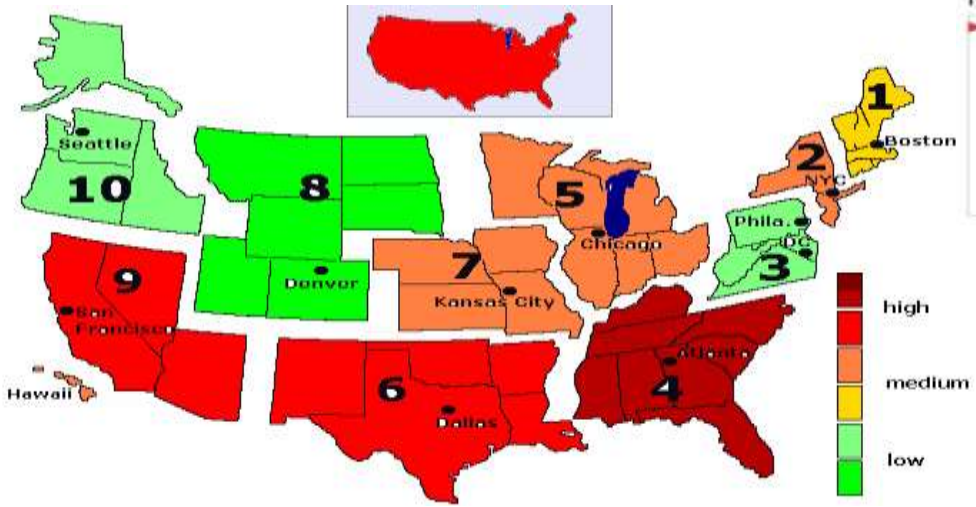
That information allowed aid organisations to channel relief supplies to those areas most in need.

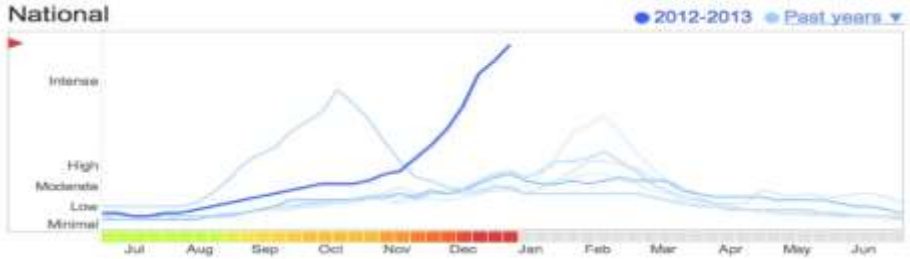
Researchers are now setting up a non-profit organisation to provide location analysis for

Data from the Digicel network in Haiti was used to estimate population movement

大数据应用案例 - 谷歌流感趋势

- 谷歌设计人员认为：人们输入的搜索关键词代表了他们的即时需要，反映出用户情况。为便于建立关联，设计人员编入“一揽子”流感关键词，包括温度计、流感症状、肌肉疼痛、胸闷等。只要用户输入这些关键词，系统就会展开跟踪分析，创建地区流感图表和流感地图。为验证“谷歌流感趋势”预警系统的正确性，谷歌多次把测试结果与美国疾病控制和预防中心的报告做比对，证实两者结论存在很大相关性。





大数据应用案例 - 百度迁徙地图

- 百度迁徙地图，是百度在2014年春运期间推出的一项技术项目。百度迁徙利用大数据，对其拥有的LBS（基于地理位置的服务）大数据进行计算分析，采用的可视化呈现方式，动态、即时、直观地展现中国春节前后人口大迁徙的轨迹与特征。

大数据应用案例 - 百度迁徙地图（续）

- **LBS（基于位置服务）**

基于位置的服务，它是通过电信移动运营商的无线电通讯网络（如 GSM 网、CDMA 网）或外部定位方式 (如 GPS）获取移动终端用户的位置信息（地理坐标，或大地坐标），在地理信息系统（外语缩写：GIS、外语全称：Geographic Information System）平台的支持下，为用户提供相应服务的一种增值业务。

它包括两层含义：首先是确定移动设备或用户所在的地理位置；其次是提供与位置相关的各类信息服务。意指与定位相关的各类服务系统，简称"定位服务"，另外一种叫法为MPS-Mobile Position Services, 也称为"移动定位服务"系统。[2] 如找到手机用户的当前地理位置，然后在上海市6340平方公里范围内寻找手机用户当前位置处1公里范围内的宾馆、影院、图书馆、加油站等的名称和地址。所以说LBS就是要借助互联网或无线网络，在固定用户或移动用户之间，完成定位和服务两大功能。

大数据应用案例 - 高德实时路况

- 实时路况可以通过安装在道路上的监测设备，或将定位设备安装在车上，实时动态地对其所经过的路段的通行情况的调查系统来获取。
- 数据来源于用户，服务于用户。

这些案例告诉我们，基于大数据我们能够迅速、直观、理性地对复杂系统或问题进行分析、预测、评估、决策和实施管理，即“**基于数据（或证据）的决策**”而不是靠似是而非的“理论”、基于过多假设的“模型”的决策或“事后诸葛亮”。

美国前白宫经济委员会主任、哈佛大学教授Lawrence Summers所认为的，“我们正处在历史的转折点，200年后的人书写我们这段历史时，他们会发现我们所处的时代人类思考方式发生了重大的变化，那就是我们比以往任何时代在更多事情上变得更加理性，我们更多以数据为依据分析思考问题。”

机遇——大数据技术促进国家和社会发展

大数据技术的运用前景是十分光明的。当前，我国正处在全面建成小康社会征程中，工业化、信息化、城镇化、农业现代化任务很重，建设下一代信息基础设施，发展现代信息技术产业体系，健全信息安全保障体系，推进信息网络技术广泛运用，是实现四化同步发展的保证。大数据分析对我们深刻领会世情和国情，把握规律，实现科学发展，做出科学决策具有重要意义，我们必须重新认识数据的重要价值。

机遇——大数据时代呼唤创新型人才

盖特纳咨询公司预测大数据将为全球带440万个IT新岗位和上千万个非IT岗位。麦肯锡公司预测美国到2018年需要深度数据分析人才44万——49万，缺口14万——19万人；需要既熟悉本单位需求又了解大数据技术与应用的管理者150万，这方面的人才缺口更大。

能理解与应用大数据的创新人才更是稀缺资源。

2.2大数据面临的挑战

挑战——大数据技术的运用仍有困难

目前，大数据技术的运用仍存在一些困难与挑战，体现在大数据挖掘的四个环节中。

数据收集

要对来自网络包括物联网和机构信息系统的数据附上时空标志，**去伪存真**，尽可能收集异源甚至是异构的数据，还可与历史数据对照，**多角度**验证数据的全面性和可信性。

数据存储

要达到**低成本、低能耗、高可靠性目标**，要用到冗余配置、分布化和云计算技术，存储时对数据进行分类，通过过滤和去重，减少存储量，并加入便于检索的标签。

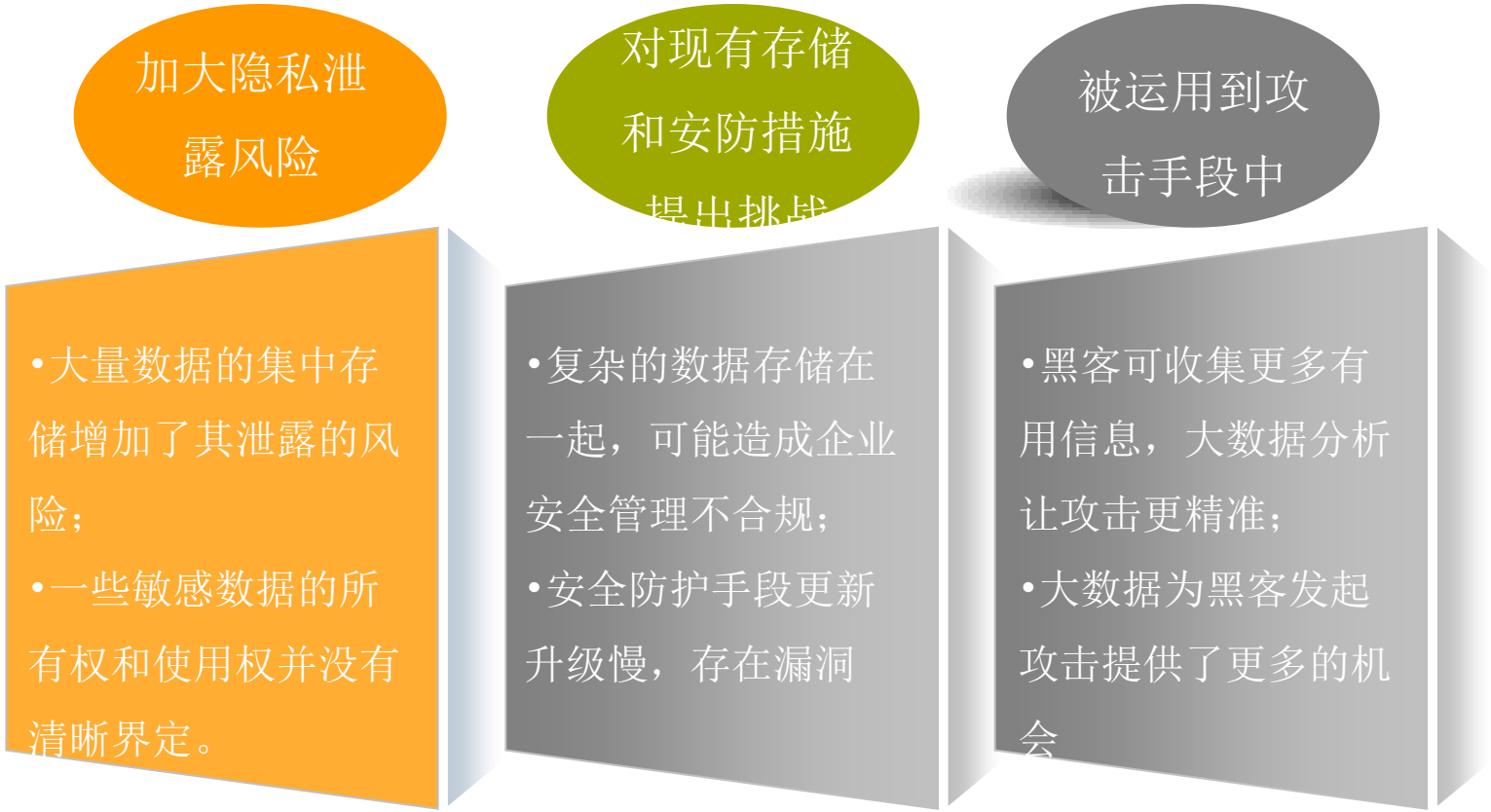
数据处理

大数据的复杂性使得难以用传统的方法描述与度量，需要将高维图像等多媒体数据降维后度量与处理，利用**上下文关联**进行语义分析，从大量动态及可能模棱两可的数据中综合信息，并**导出可理解的内容**。

结果的可视化呈现

使结果更直观以便于洞察。目前，尽管计算机智能化有了很大进步，但还只能针对小规模、有结构或类结构的数据进行分析，谈不上深层次的数据挖掘，现有的数据挖掘算法在不同行业中难以通用。

挑战——大数据给信息安全带来新挑战



“棱镜门”引爆大数据时代争议

事情的起因是美国中情局前职员斯诺登向媒体爆料，过去6年间，美国的情报部门通过一个代号为“棱镜”的项目，从多家知名互联网公司获取**电子邮件、在线聊天内容、照片、文档、视频等网络私人数据，跟踪用户一举一动**。他说，自己只需要坐在办公桌前，动动指头，敲敲键盘，就能了解很多人的私密信息。

斯诺登的爆料引起一片哗然，根据他提供的资料，被卷入“棱镜门”事件的公司包括微软、雅虎、谷歌、苹果、Facebook等9大IT业巨头。在“棱镜门”事件开始发酵之后，这些公司先是赶紧出面否认与美国政府的监视项目进行过合作，并相继发表声明，呼吁政府采取更透明态度，以证明他们的“清白”。

一方面我们通过对大量用户数据的分析，公司、企业、政府都可以更好的了解用户行为、消费习惯的等等，从而可以提供更好的服务。但是另外一方面，这又不可避免的对用户的隐私构成威胁、挑战。很多人已经意识到，在数据的应用方面，**相关法律法规的制定变得越来越重要。**作为用户，需要明确界定自己在数据的使用方面具有什么权力和义务；作为企业和政府，需要逐渐的定位清楚，在多大程度上可以并且用什么样的方式来使用用户的数据。

在现有的互联网结构下，我们所有的网络行为对于服务提供商来说都是**透明**的。人们既想借助互联网平台与别人交流，又想自己不被窥探，这是完全不可能的。**网络隐私安全**将是未来一个巨大的问题。

各种大数据背后的复杂系统管理

数据类型一：社会网络抽样数据，管理：网络大数据与社会管理
数据类型二：手机定位与移动通话数据，管理：交通、救援、公安
数据类型三：卫星图像数据，管理：侦察、态势分析、大气科学等
数据类型四：人口数据，管理：人口管理、城镇化建设管理等
数据类型五：在线社会网络数据，管理：舆情、开源情报分析等
数据类型六：教育、市场与医疗数据，管理：医疗、教育管理等
数据类型七：犯罪数据、警局报告数据，管理：团伙犯罪网络分析
数据类型八：物联网数据，管理：信号监测与食品安全管理等
数据类型九：组织结构、经济发展数据，管理：组织演化规律研究
数据类型十：人力资源数据，管理：人力资源结构、更替规律分析

挑战——大数据呼唤智能化时代的到来

实际上，大数据、物联网、云计算、网络科学、社会网络、数据挖掘、证析、仿真、计算实验等都有很密切的关系：

1、大数据挖掘、开源情报分析
2、大数据与证析（基于证据的决策）
3、大数据分析与仿真验证
4、大数据分析与云计算
5、大数据与网络科学研究
6、大数据与社会网络分析
7、大数据与计算实验

结论：大数据时代追求的“不是随机样本，而是全体数据”。全体数据正好刻画了复杂系统的整体。

总的发展趋势是数据越来越多，问题越来越非结构化，关系越来越复杂和网络化。

目录

3.1 人工智能的概念

- 什么是人工智能
  - 顾名思义，人工智能就是人造智能，其英文表示是“Artificial Intelligence”，简称AI。当然，这只是人工智能的字面解释或广义解释。目前的“人工智能”一词是**指用计算机模拟或实现的智能**，同时，人工智能又是一个学科名称。
  - 人工智能研究的是如何使机器（计算机）具有智能的科学和技术，特别是自然智能如何在计算机上实现或再现的科学和技术。因此，从学科角度讲，当前的人工智能是计算机科学的一个分支。

基于应用领域的领域划分

- 1. 难题求解
- 2. 自动定理证明
- 3.自动程序设计
- 4. 自动翻译
- 5.智能控制
- 6.智能管理
- 7.智能决策
- 8.智能通信
- 9. 智能仿真
- 10.智能CAD
- 11.智能CAI

基于应用系统的领域划分

- 1.专家系统
- 2.知识库系统
- 3.智能数据库系统
- 4.智能机器人系统

人工智能的基本技术

- 1 推理技术
- 2 搜索技术
- 3 知识表示与知识库技术
- 4 归纳技术
- 5 联想技术


李开复未来十年人工智能趋势的崛起
国家会议中心
00:19

3.2 人工智能的发展概况

- 人工智能学科的产生
- 现在公认，人工智能学科正式诞生于1956年。需要指出的是，人工智能学科虽然正式诞生于1956年的这次学术研讨会，但实际上它是逻辑学、心理学、计算机科学、脑科学、神经生理学、信息科学等学科发展的必然趋势和必然结果。单就计算机来看，其功能从数值计算到数据处理，再下去必然是知识处理。实际上就其当时的水平而言，也可以说计算机已具有某种智能的成分了。
- 天才的英国计算机科学家图灵（A.M.Turing）就于1950年发表了题为“计算机与智能”的论文，提出了著名的“图灵测试”，为人工智能提出了更为明确的设计目标和测试准则。

人工智能的发展

1946年第一台电子管计算机诞生

1969年，科学家们用集成电路计算机完成登月任务。当时的一个程序大小只有6M。相当于现在的一张高清照片。

今天，你口袋里的智能手机的计算处理能力相当于1969登月时候地球上所有计算机处理能力的总和。

符号主义途径发展概况

- 1956年之后的10多年间，人工智能的研究取得了许多引人瞩目的成就。从符号主义的研究途径来看，主要有：
- (1)1956年，美国的纽厄尔、肖和赛蒙合作编制了一个名为逻辑理论机(Logic Theory Machine，简称LT)的计算机程序系统。
- (2)1956年，塞缪尔研制成功了具有自学习、自组织、自适应能力的跳棋程序。
- (3)1959年，籍勒洛特发表了证明平面几何问题的程序，塞尔夫里奇推出了一个模式识别程序；1965年罗伯特(Roberts)编制出了可以分辨积木构造的程序。
- (4)1960年，纽厄尔、肖和赛蒙等人通过心理学试验总结出了人们求解问题的思维规律，编制了通用问题求解程序(General Problem Solving简称GPS)。

- (5)1960年，麦卡锡研制成功了面向人工智能程序设计的表处理语言LISP。该语言以其独特的符号处理功能，很快在人工智能界风靡起来。它武装了一代人工智能学者，至今仍然是人工智能研究的一个有力工具。
- (6)1965年，鲁宾逊(Robinson)提出了消解原理，为定理的机器证明做出了突破性的贡献。
- 以上是以推理为中心，是人工智能的早期，称为人工智能的推理期。
- 1965年，美国斯坦福大学的费根鲍姆教授研制了基于知识DENDRAL专家系统，标志人工智能新时期的开始。随后出现许多专家系统，用于诊病、找矿等工作。1977费根鲍姆教授提出“知识工程”的概念，使人工智能进入以知识为中心的知识期。

连接主义途径发展概况

- 从连接主义的研究途径看，早在20世纪40年代，就有一些学者开始了神经元及其数学模型的研究。例如，1943年心理学家McCulloch和数学家Pitts提出了形式神经元的数学模型—现在称之为MP模型，1944年Hebb提出了改变神经元连接强度的Hebb规则。MP模型和Hebb规则至今仍在各种神经网络中起重要作用。

- 神经网络学科的发展和应用也迎来了脑神经科学、认知科学、心理学、微电子学、控制论和机器人学、信息技术和数理科学等学科的相互促进、相互发展的空前活跃时期，特别是在计算机科学研究领域，将从根本上改变人们传统的数值、模拟、串行、并行、分布等计算与处理概念的内涵和外延，出现了分布式并行新概念、数值模拟混合的新途径，探索和开创光学计算机、生物计算机、第n代计算机的新构想，为21世纪计算机科学与技术的飞速发展奠定了思想和理论基础。

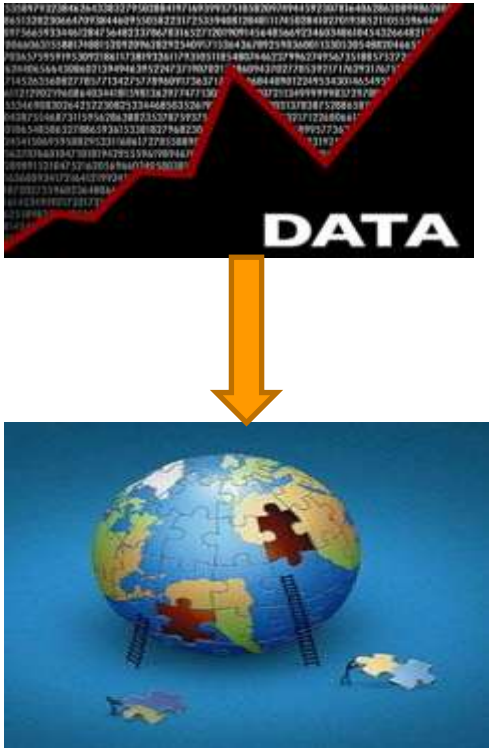
当前发展趋势

- 首先指出，由于人工智能技术的飞速发展和作者视野的限制，所以，很难在这样一个小节的篇幅里，对人工智能的当前发展趋势作出全面、准确的评估。但一般认为，当前人工智能的发展，呈现出如下特点：
  - (1)传统的符号处理与神经计算各取所长，联合作战。
  - (2)一批新思想、新理论、新技术不断涌现。
  - (3)以Agent（称为“主体”或“智能主体”、“智能体”等）技术和分布式人工智能(DAI）正异军突起，蓬勃发展。
  - (4)应用研究愈加深入而广泛。当今的人工智能研究与实际应用的结合越来越紧密，受应用的驱动越来越明显。事实上，现在的人工智能技术已同整个计算机科学技术紧密地结合在一起了，其应用也与传统的计算机应用越来越相互融合了。


谷歌云首席科学家李飞飞：人工智能 2017
GIF
00:13

从大数据、虚拟现实、物联网到智能化应用

- 阿尔法狗的发展
  - 智能机器人：siri，微软小冰
  - 医疗行业、汽车行业
- 虚拟现实：VR，AR，MR，IR
  - AR眼镜，AR小程序等
- 物联网、智能物流供应链：
  - amazon，JD，智能拣货机器人
  - 智能家居：扎克伯格
- 人工智能能否替代人？？三体、机器人



人机大战

深蓝 国际象棋

“算”——超级计算机系统

阿尔法狗 围棋

“想”——深度学习的机器学习算法

围棋挑战——被称作人工智能的“阿波罗计划”

如何深度学习？

阿尔法狗学习围棋跟机器识别小狗的原理一样

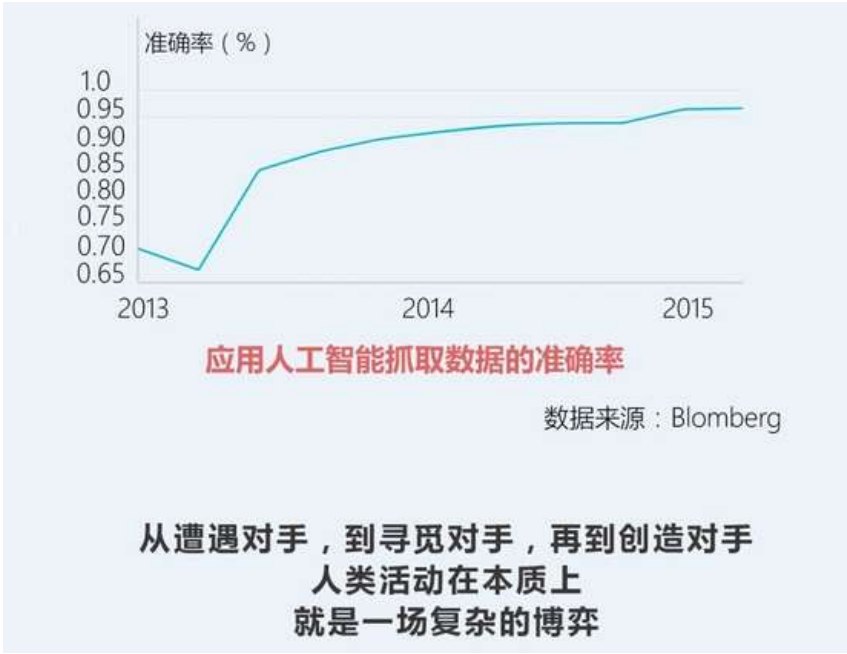
如何深度学习？


AlphaGo与主要对手的实力对比
注：评估时间为2015年10月
入门
业余
职业
段位
1 3 5 7 9 1 3 5 7 9
李世乭
AlphaGo

不断学习


大师秘笈
全真模拟
订正本
给AlphaGo输入
3000万步
人类围棋大师的走法
让AlphaGo自我对弈
3000万局
积累胜负经验
在自我对弈的训练中
形成全局观
对局面作出评估
构成策略网络
给出落子选择
构成评价网络
修正落子选择
尤其是在中盘和官子部分，AlphaGo展现出强大的落子选择能力。随着训练增加，AlphaGo还在进步。

人工智能的无限可能——政治、军事、贸易



虚拟现实技术

- 沉浸感、交互感、想象性
- 买VR眼镜送岛国动作片

虚拟现实技术——VR、AR、MR、CR

VR：你看到的一切都是假象

（Virtual Reality）=虚拟现实

纯虚拟场景，即看到的场景和人物全是假像，把意识代入一个虚拟的世界。

VR设备代表：Oculus Rift

VR类影片代表：《黑客帝国》、《盗梦空间》等。

AR：你能分清哪个是真的，哪个是假的

（Augmented Reality）=增强现实=真实世界+数字化信息

即看到的场景和人物一部分是真一部分是假，是把虚拟的信息带入到现实世界中。

虚拟物体与真实物体能被区分。虚拟物体的相对位置，会随设备的移动而移动。

虚拟现实技术——VR、AR、MR、CR

MR：你已经分不清哪个是真 哪个是假

（Mix Reality）=混合现实=VR + AR合并现实和虚拟世界而产生的新的可视化环境。虚拟物体与真实物体很难被区分。虚拟物体的相对位置，不会随设备的移动而移动。

MR设备代表：Magic leap

CR（Cinematic Reality）=影像现实

虚拟场景跟电影特效一样逼真，达到“欺骗”大脑的目的，有别于通过屏幕投射显示技术。

关键：设备、系统平台、后台、硬件

智能物流及供应链——无人仓、无人机

- 代表企业：亚马逊、京东...

Google Atlas机器人

从互联网到物联网再到智联网

- 扎克伯格的“贾维斯”

人工智能与未来工作生活

烹饪200种美食的机器人

智能驾驶

两天盖好一栋房子

比博尔特跑得还快的智能警察

农业自动化

我国人工智能研究发展简况

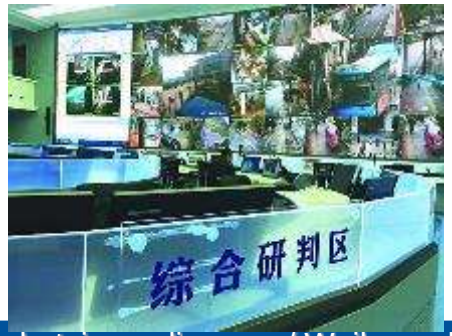
- 由于众所周知的原因，我国人工智能的研究起步较晚。20世纪70年代末，我国才有一批学者认真地开始了人工智能的研究。1977年，涂序彦（现任中国人工智能学会理事长）和郭荣江在《自动化》第1期上发表了国内首篇关于AI的论文—《智能控制及其应用》，拉开了我国人工智能研究的序幕。从此，我国在人工智能方面的研究便蓬勃兴起。
- 中国人工智能离世界多远？

目录

大数据侦查破案案例

公安大数据的特点

- 公安业务数据之“量大”
  - 地市级公安部门数据累积量已达到或接近PB级
    - 苏州市公安局日监控数据达770TB
    - 北京市公安局日监控数据达2000TB
    - 公安部案件文本数据量约24TB
    - 全国犯罪人员DNA指纹库样本达2000多万



公安大数据的特点

- 公安业务数据之“多样”
  - 公安大数据包含了几乎所有的数据结构类型
    - 结构化数据
      - 案件信息库、人员信息库、车辆信息库
    - 半结构化数据
      - 网页数据，html文件等
    - 非结构化数据
      - 语音数据、视频监控数据、图片数据、生物特征数据

公安大数据的特点

- 公安业务数据之“实时”
  - 公安大数据具有很强的实时性






公安大数据的特点

- 公安业务数据之“价值”
  - 公安大数据的价值较低

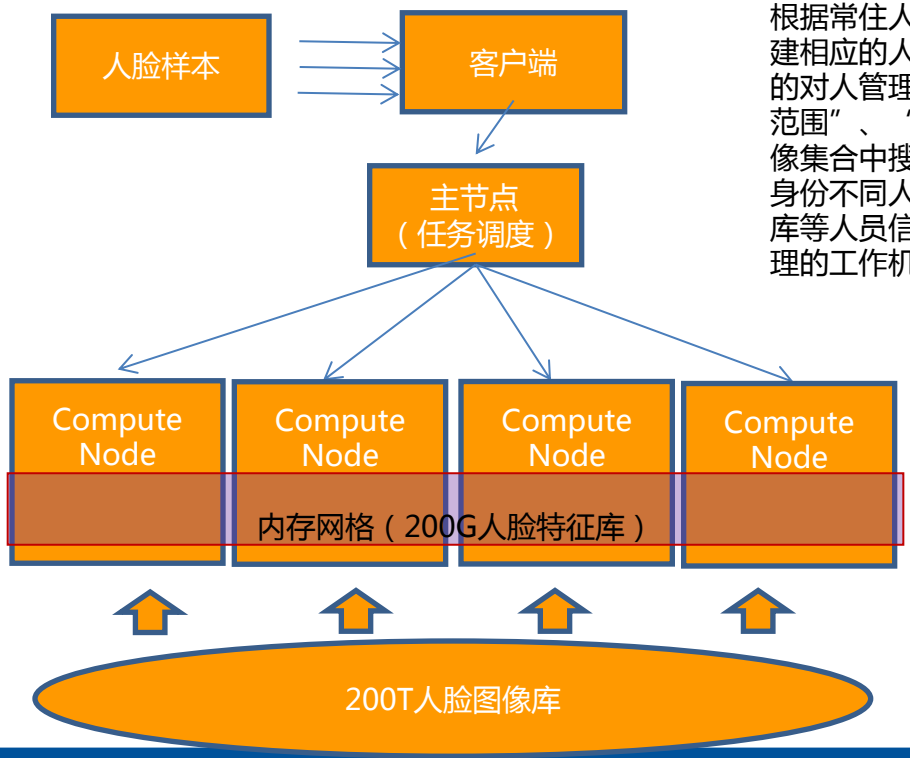
公安大数据实践应用一：寻找犯罪活动规律

◆大数据应用案例 - IBM大数据分析预防犯罪

- 在美国南卡罗来那州的查尔斯顿， 警方利用 IBM 的数据分析工具， 帮助当地的 400 多名警察更加准确地进行犯罪模式的分析。 警方利用分析预测工具进行警力调配， 发现犯罪热点地区提前预防犯罪发生， 从而减少了当地的发案率。
- 进入大数据时代，警察的职责也将发生转变：从案发后追捕罪犯，到分析犯罪数据，识别犯罪模式，部署警力，预防犯罪。在案发前，终结犯罪。
- 大数据帮助分析情报并找到可疑的规律，在犯罪分子出现之前布置警力，预防犯罪。在大海捞针时，大数据帮助把“针”放大几百倍。
- 北京的天罗地网系统。

公安大数据实践应用二：案件线索比对查询

◆指纹人脸比对系统



目前公安机关已建立人口信息系统、出入境系统、警综系统、在逃人员系统、监管系统、交管系统等，有大量的人员基本信息及照片，因此可根据常住人口、流动人口、重点人员、在逃人员的基本信息及照片，组建相应的人像基础数据库，建立人像智能识别应用平台，做到真正有效的对人管理。实现人像数据的检索挖掘。系统可根据“性别”、“年龄范围”、“分库”等条件对人像库中的记录进行筛选，在符合条件的人像集合中搜索与用户提供的查询照片高度相似人员资料列表。可进行同身份不同人及同人不同身份的可以人员人脸数据的挖掘，发现既有人口库等人员信息库中双重户口、虚假身份等问题，并建立起查证、打击处理的工作机制。

优势（相对传统数据库方案）：

- 海量分布式存储
- 线性扩展
- 并行对比
- 快速查询

→ SOA调度，并行对比

→ MapReduce特征转化、加载

→ Hbase存储

◆ 交通卡口大数据分析应用案例

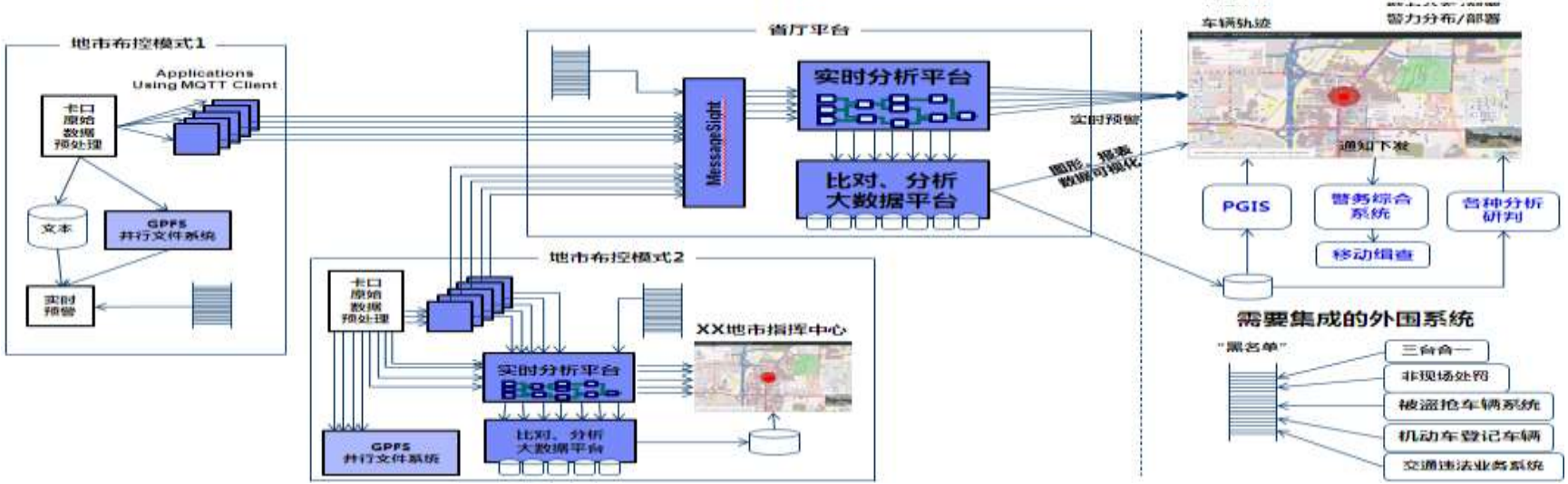
- 嫌疑车辆预警的时效性 <8 秒
  - 高清图片处理，解析文本传输、比对和存储、管理（归档） -- 海量

    21个地市平均1.2PB 的高清图片（2年）； 5000个高清卡口产生1亿条（1K）/天；几十万条黑名单
  - *套牌车、盗抢车辆、报废车、违法车等*
  - 高清图片的实时访问 （从地市到省厅）
- 各种高级分析、研判和关联分析
  - *同行车辆分析、昼伏夜出车辆、关联业务等*
- 多个边界系统的有效整合及信息交换
- 跨部门、多警种的高效联动

◆ 智能监控系统中卡口数据分析

- 目前现状
  - ✓ 全省21个地市，共1000多个卡口，全省**每天产生1300万条数据**
  - ✓ 地市卡口数据分为图片(压缩后标清150K，**高清350K**)和解析后的**文本1K/每条**；
  - ✓ 黑名单大约从几万~几十万条，可以按**平均33万条**计算；
  - ✓ 在线数据90天，其余数据迁移到备份系统

- 业务需求
  - ✓ 全省1000多个摄像头，将增加到**5000个**；
  - ✓ 将**全部转变为高清图片**采集，将来需要实现视频流的采集；
  - ✓ 在线数据存放由目前90天，增加到**存放2年**；
- 面临挑战
  - ✓ 无法应对将来数据增长的挑战；实时预警要求；研判和关联分析



◆ 智能监控系统中卡口数据分析

- **同行车辆分析：**
  通过分析嫌疑车辆前后几分钟的车辆（时间自定义），找出在不同地市不同卡口多次同行的车辆。对确认的同行车辆进行关联追踪，寻找出其他同行车辆，并以线条方式在地图上展示同行车辆的轨迹，查看同行车辆轨迹信息

- **昼伏夜出分析：**
  分析判断部分车辆经常白天某个时间点进城后不出城或是晚上某个时间点进城或出城，针对该些具有规律性的车辆进行筛选，筛选出该车辆进行特别关注，并对该车辆进行特别标注

- **与车辆登记库、交通违法库关联分析：**
  针对部分车辆只提供车辆品牌等特征并非知道该车辆号牌号码、车牌颜色等特征，需与车辆登记库进行关联，查询出该车辆的登记信息，并根据车辆登记信息关联交通违法库，查询出该车辆是否为涉案等违法车辆


人民网 安徽频道
ah.people.cn
3DMGAME

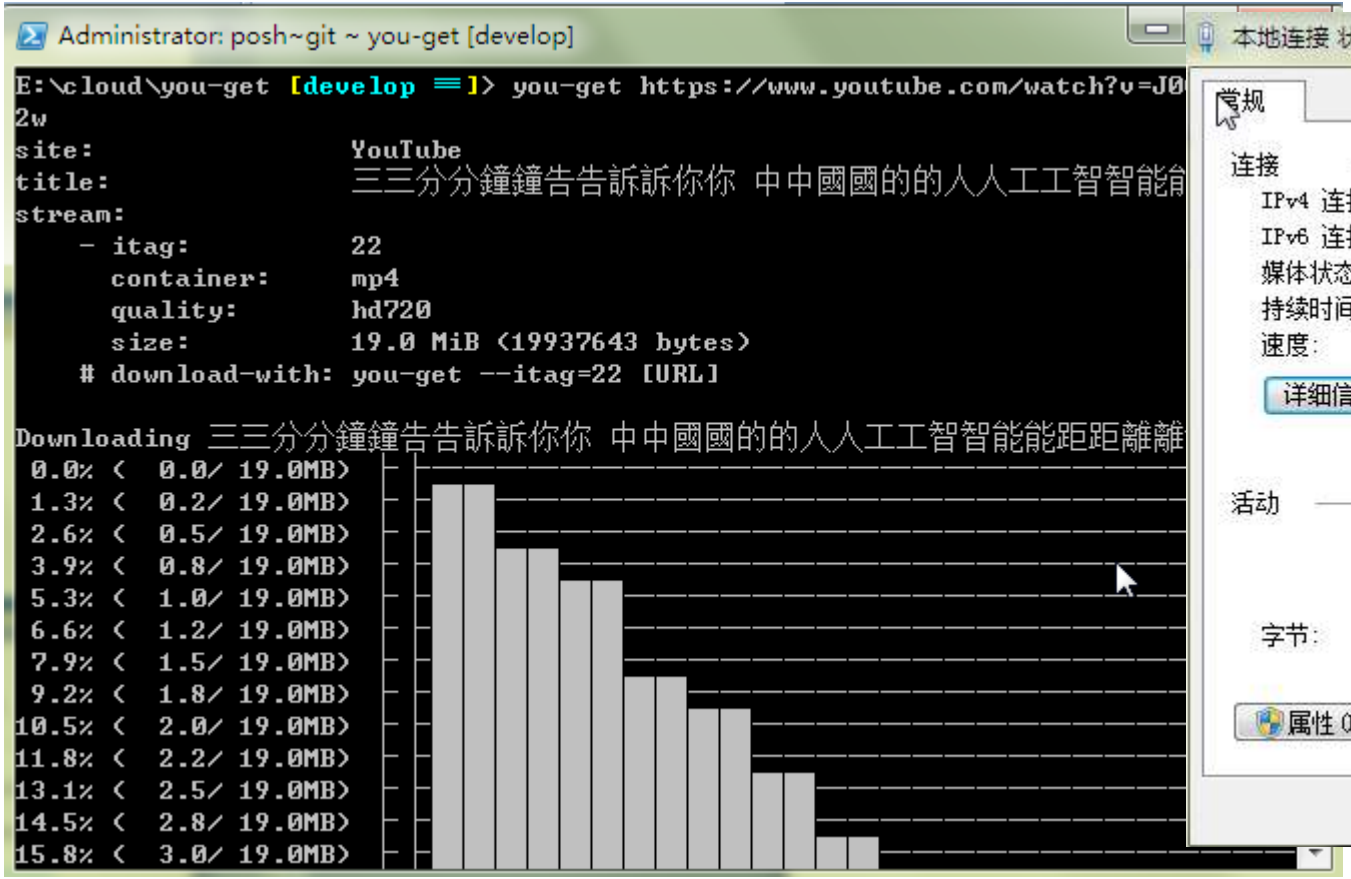
未来机器人

目录

其它日常应用

- 公有云、私有云、混合云
  - 私有云采用与公共云同样一种基于互联网的架构，但它们是专门供本企业使用的，可以通过防火墙与公共互联网隔离开来，从而加强安全级别、提高性能水平。
  - 而混合云集公共云和私有者这两者之所长;在这种云环境中，企业使用私有云来处理最重要的计算任务，使用公共云来处理偶尔出现的需求高峰或不太敏感的任务
- 监控
- 存储

其它日常应用


Administrator: posh~git ~ you-get [develop]
E:\cloud\you-get [develop ≡]> you-get https://www.youtube.com/watch?v=J0
2w
site: YouTube
title: 三三分分鐘鐘告告訴訴你你 中中國國的的人人工工智智能能
stream:
- itag: 22
container: mp4
quality: hd720
size: 19.0 MiB (19937643 bytes)
download-with: you-get --itag=22 [URL]
Downloading 三三分分鐘鐘告告訴訴你你 中中國國的的人人工工智智能能距距離離
0.0% (0.0/ 19.0MB)
1.3% (0.2/ 19.0MB)
2.6% (0.5/ 19.0MB)
3.9% (0.8/ 19.0MB)
5.3% (1.0/ 19.0MB)
6.6% (1.2/ 19.0MB)
7.9% (1.5/ 19.0MB)
9.2% (1.8/ 19.0MB)
10.5% (2.0/ 19.0MB)
11.8% (2.2/ 19.0MB)
13.1% (2.5/ 19.0MB)
14.5% (2.8/ 19.0MB)
15.8% (3.0/ 19.0MB)


本地连接 状态
常规
连接
IPv4 连接: Internet
IPv6 连接: Internet
媒体状态: 已启用
持续时间: 00:48:43
速度: 1.0 Gbps
详细信息(E)...
活动
已发送
已接收
字节: 17,669,860 296,685,353
属性(P)
禁用(D)
诊断(G)
关闭(C)

谢谢

THANK YOU FOR YOUR ATTENTION

13811517244

hefugang@ppsuc.edu.cn

http://hefugang.github.io/