

大数据前沿知识 与课标内容解读

欧阳元新

北京航空航天大学 计算机学院

2018-02-08

- ▶ 大数据教育与信息技术教育的关系解读
- ▶ 大数据在课程标准中的设计与要求
- ▶ 教学建议与案例研讨

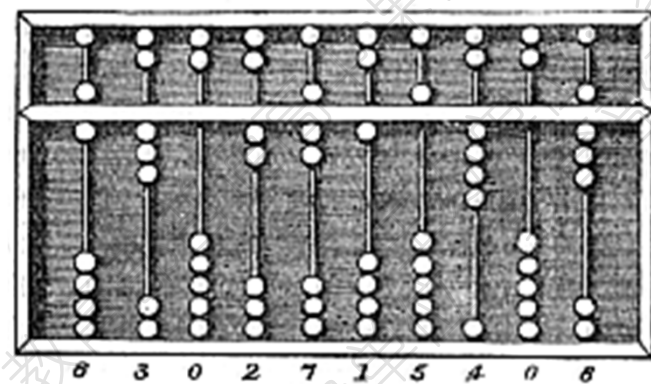
Computer = Compute + er

Compute what ?

计算的演化史



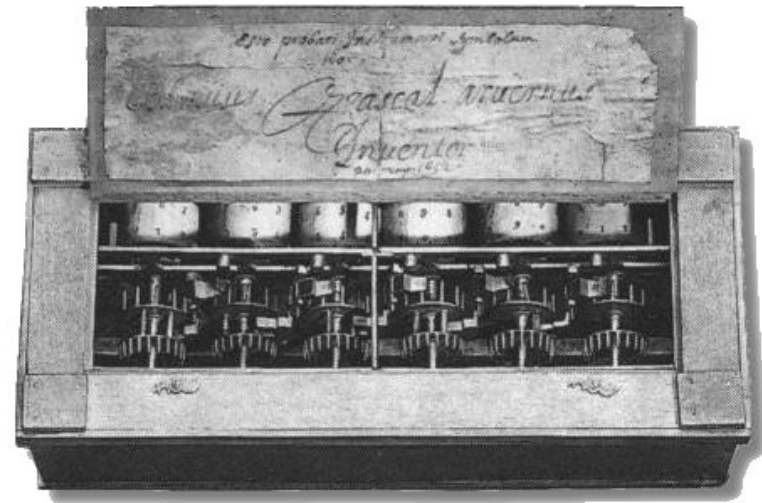
1	2	3	4	5
—	┌	└	正	正
/	//	///	////	/////



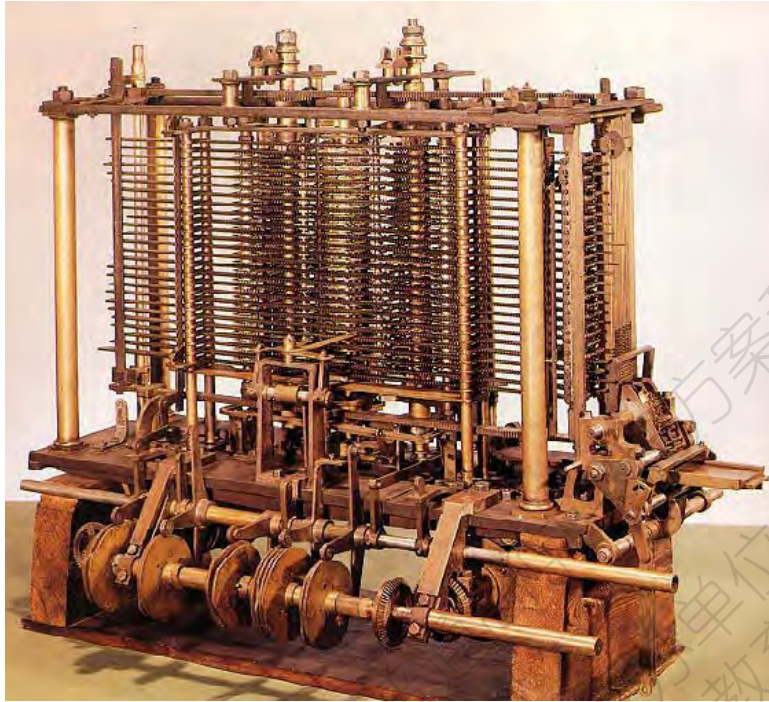
计算的演化史



- Blaise Pascal
- 1642 France
- 第一部机械计算设备
- 8 figures



计算的演化史

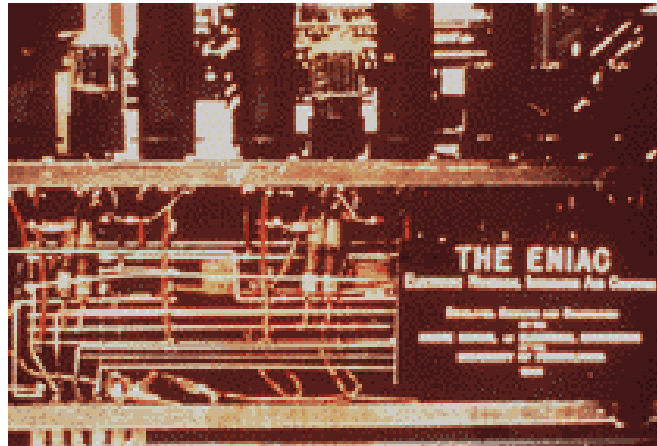


差分机 (1820s)



卡片制表系统 (1890s)

计算的演化史

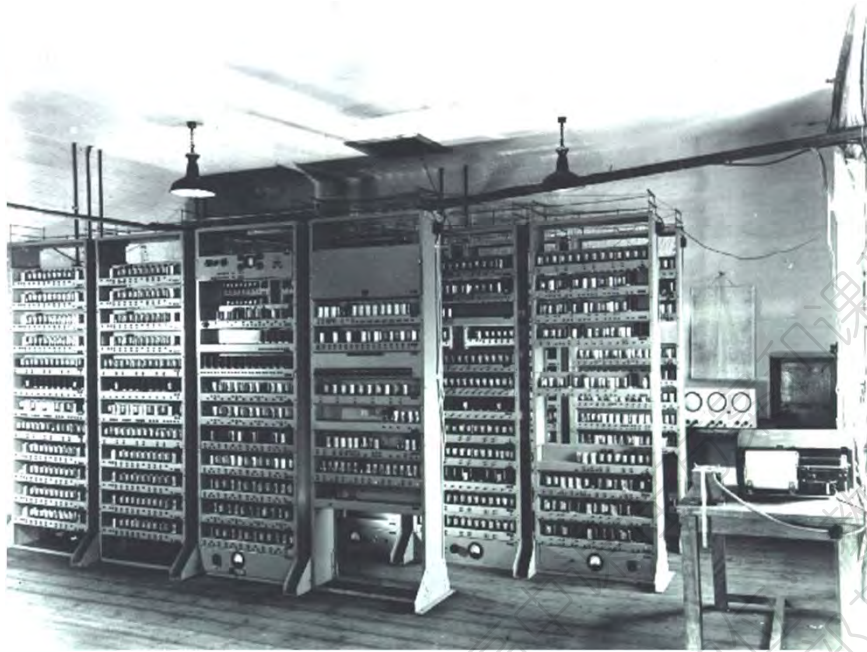


ENIAC (1946)

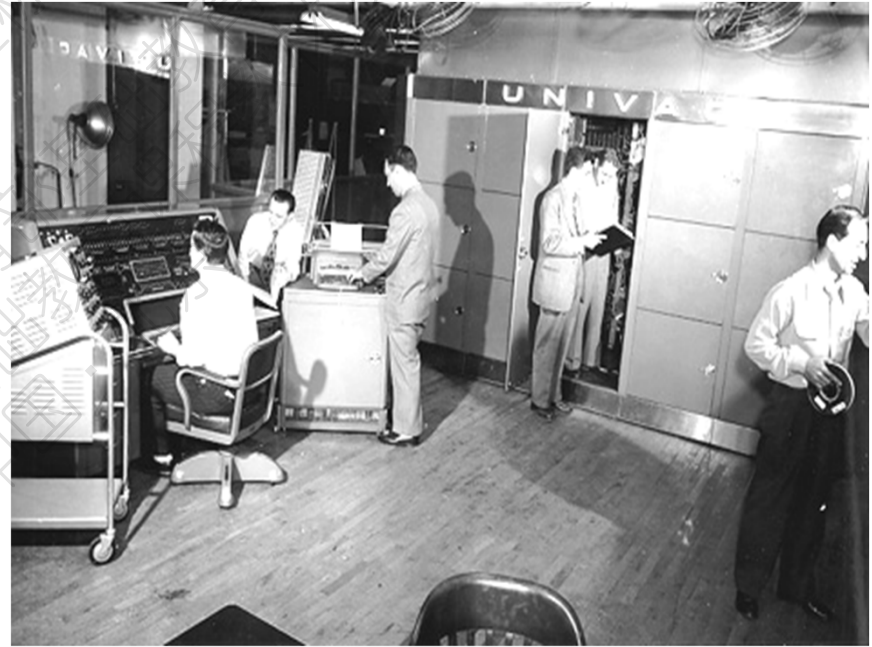


EDVAC (1947)

计算的演化史



EDSAC (1949)



UNIVAC (1951)

计算的演化史



IBM360 (1960s)



PDP-11 (1965)

计算的演化史

- **1964, Seymour Cray, Control Data Company delivers the CDC 6600**
- **supercomputer**
- **nanoseconds**



网络的演化史



Internet的早期历史

- 1950s: 计算机介入电话技术
- 1960s: 主机+终端系统
 - SABRE airline reservation system
- 1970s: 计算机相互通信
 - **ARPANET** packet switching network
 - TCP/IP internet protocols
 - Ethernet local area network
- 1980s 至今: 因特网大发展
 - Commercialization of Internet
 - E-mail, file transfer, web, P2P, . . .
 - Internet traffic surpasses voice traffic

三次工业革命

- 第一次工业革命18世纪60年代--19世纪中
 - 机械化，代表发明：蒸汽机、**铁路、轮船**
 - **物质**在全球范围的自由流动
- 第二次工业革命19世纪下半叶--20世纪初
 - 电气化，代表发明：电动机、**全球电网**
 - **能源**在全球范围的自由流动
- 第三次工业革命20世纪四五十年代到现在
 - 信息化，代表发明：计算机、**互联网**
 - **信息**在全球范围的自由流动

计算机科学的发展阶段

- CPU为核心
- 网络为核心



?

教育部普通高中课程方案和课程标准国家级示范校
主办单位：教育部基础教育司
承办单位：教育部基础教育课程教材发展中心
中国·北京

计算机科学的发展阶段

- CPU为核心
- 网络为核心



- 搜索为核心

教育部普通高中课程方案和课程标准国家级示范校
主办单位：教育部基础教育司
承办单位：教育部基础教育课程教材发展中心
中国·北京

计算机科学的发展阶段

YAHOO!



- 1994年 杨致远 费罗
 - 浏览器推出后，立刻迷上了
 - 他们制作了自己的主页，把喜欢的靓站网址收集起来，链接到自己的主页上。
 - 随着网址越来越多，编出一个专门用于整理特网上各个节点资料的程序。
 - 放到网站上后，访问量快速增长，以至于斯坦福大学网络都给挤爆了。杨致远敏锐地发现了其中蕴藏的巨大商机。

计算机科学的发展阶段

The logo for Yahoo!, featuring the word "YAHOO!" in a bold, red, sans-serif font with a registered trademark symbol.

- 网景（Netscape）资助
 - 1995年1月，网景浏览器一个最重要的按钮——网上搜索指向了雅虎
- 雅虎方向
 - 不只是提供**分类目录**的网站，而是一种**新媒体**，进入信息高速公路必经的门户
- 雅虎上市
 - 1996年3月7日，雅虎股票正式上市，市场价值达到8.5亿美元，是“红杉资本”投资时的**200倍（不到1年）**。

计算机科学的发展阶段



- 2000年
 - 互联网泡沫破灭
 - 雅虎开始把搜索引擎业务外包给谷歌，并宣称其为“互联网上最好用的搜索引擎”，因为当时搜索引擎没有大规模盈利模式，雅虎的门户网站显示型广告才是主流模式，所以雅虎并不在意搜索。
 - 在2000年按年向谷歌支付了720万美元的搜索技术服务费，这笔资金雪中送炭，帮助谷歌随后高速发展起来。

十五年周期定律

- IBM前CEO郭士纳提出
 - 计算模式每隔15年发生一次变革
- 1965年前后，以大型机市场化为标志
- 1980年前后，以个人计算机为标志
- 1995年前后，以互联网普及为标志
 - 每一次变革都引起企业间、产业间甚至国家间竞争格局的重大变化。互联网革命一定程度上是由美国“信息高速公路”战略所催熟

- 2010年前后？

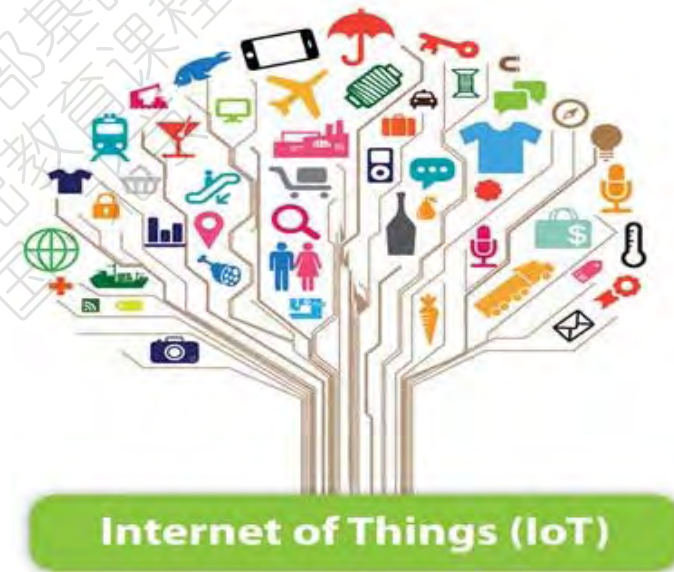
MADE IN CHINA 中国制造
2025

从互联网到物联网

- What is “thing”?
- How the things connected by internet?



Internet of Computers



Internet of Things (IoT)

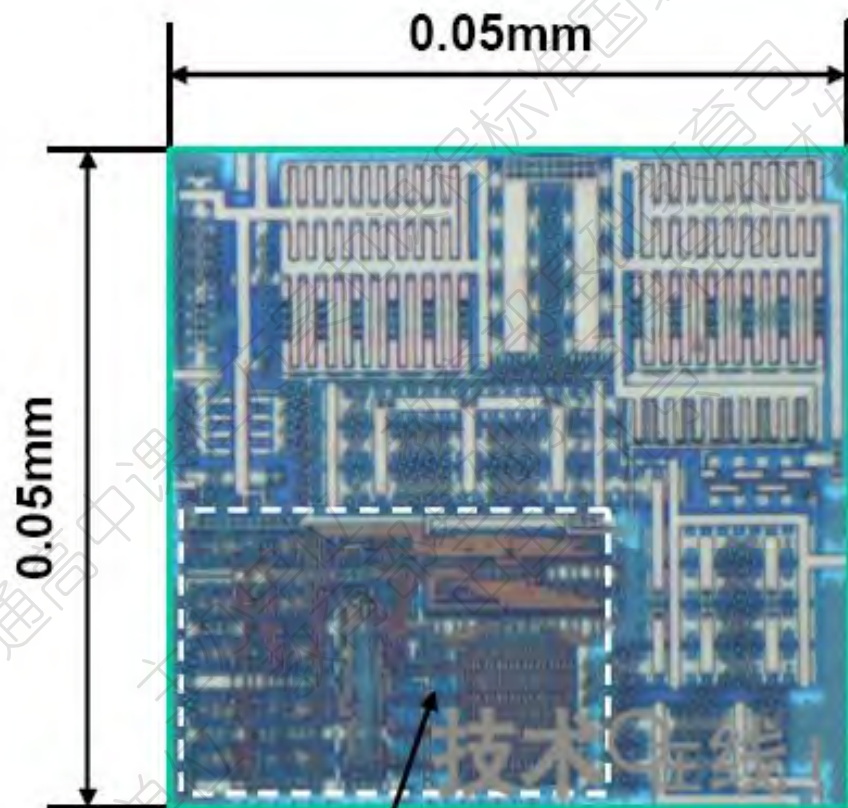
射频标签——RFID标签



RFID标签及封装



射频标签



128-bit Memory (21µm x 32µm)

从互联网到物联网

物联网的概念是什么时候提出的？

- A. 1995年 B. 1999年
C. 2005年 D. 2009年



Internet of Computers



Internet of Things (IoT)

1999年 MIT Auto-ID

- **EPC系统（Electronic Product Code）**
- 把所有物品通过**RFID**和条码等信息传感设备与互联网连接起来，实现智能化识别和管理功能的网络
- 通过**RFID**技术和互联网的结合应用，实现物品或商品的自动识别和信息的互联与共享

幕后推手：零售巨头沃尔玛

传统条码

(1) 国际编码格式



(2) 自用编码格式



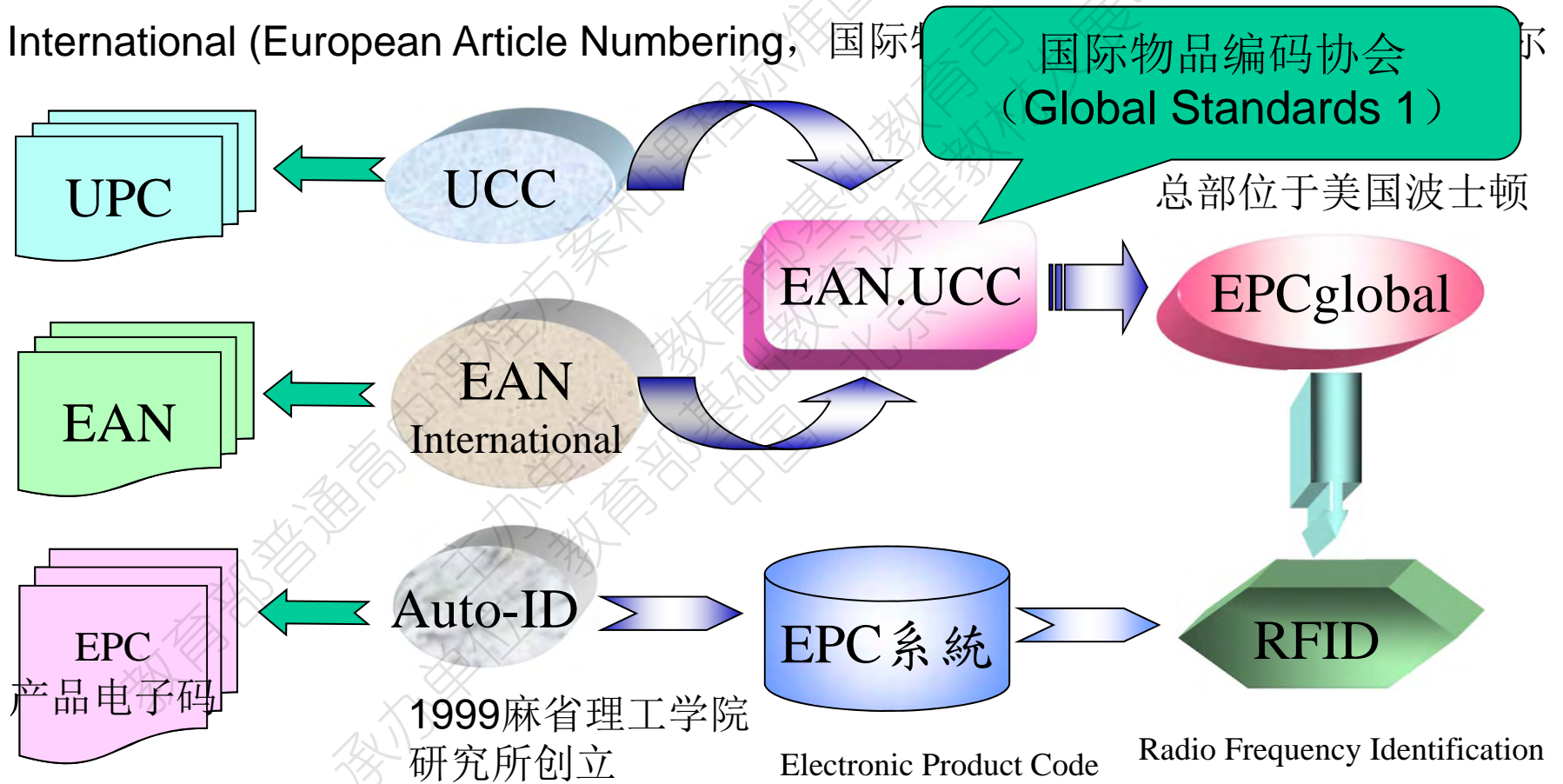
39码
128码
EAN码
PDF417条码
UPC码
二维条码
ISBN码与ISSN码

UPC:通用产品代码 EAN:国际物品编码

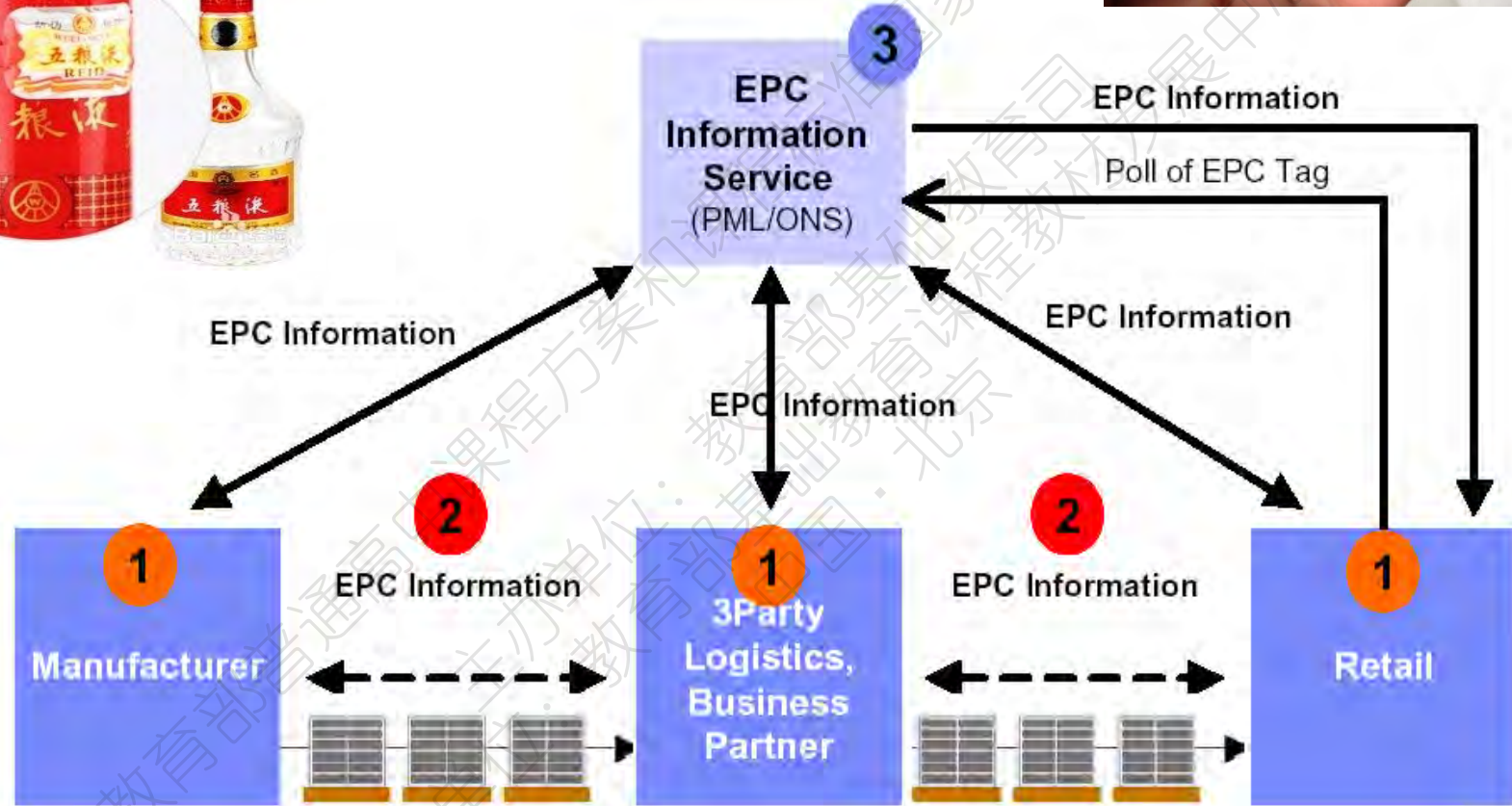
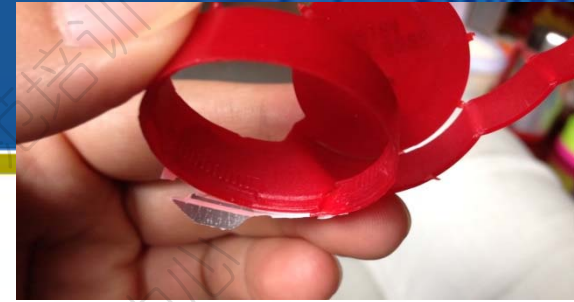
电子代码

UCC(Uniform Code Council, 统一编码协会): 总部位于美国, UPC码的发号机构

EAN International (European Article Numbering, 国际物品编码协会)



EPC系统

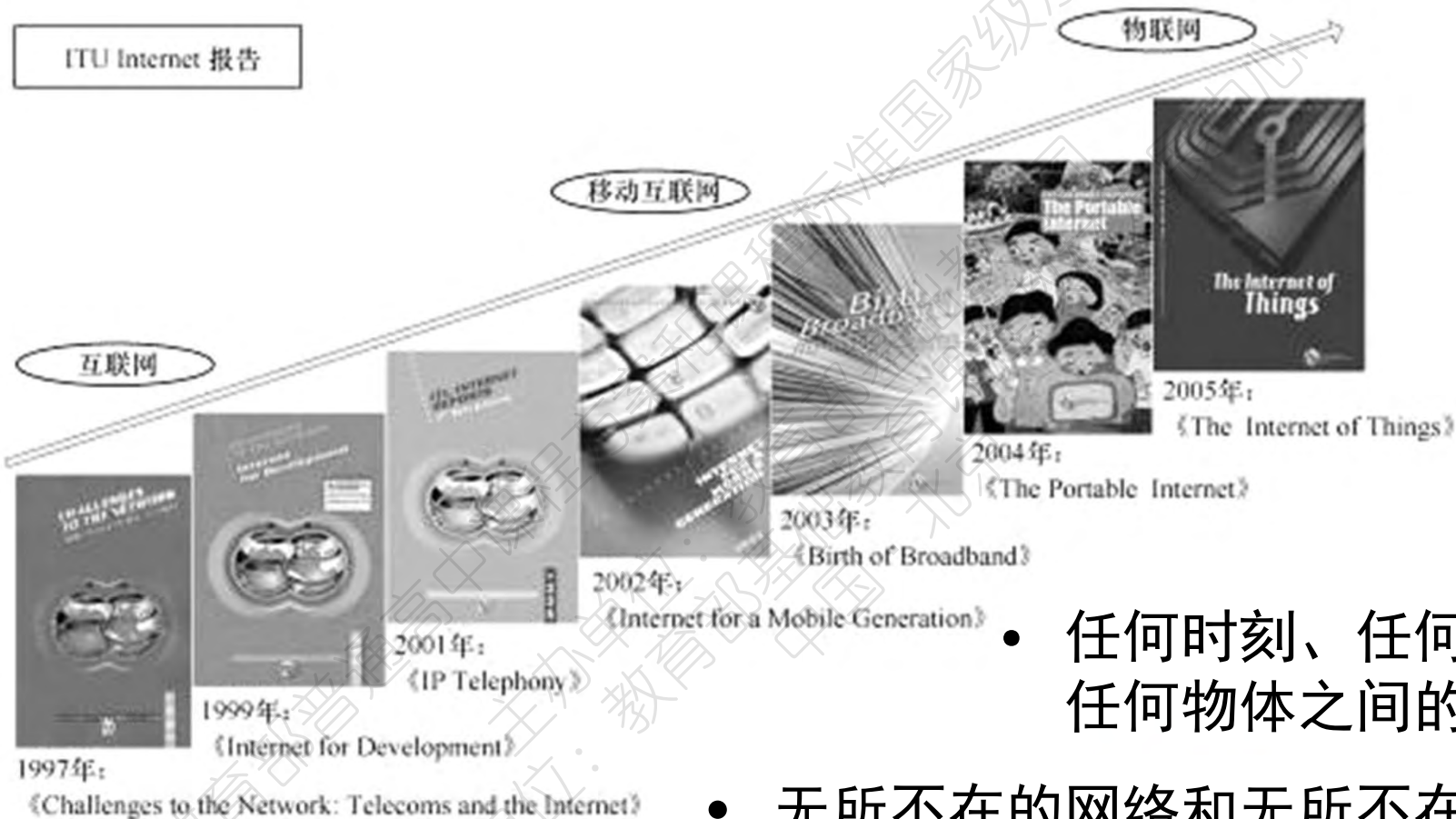


1 Internal EPC Network

2 B2B EPC Network

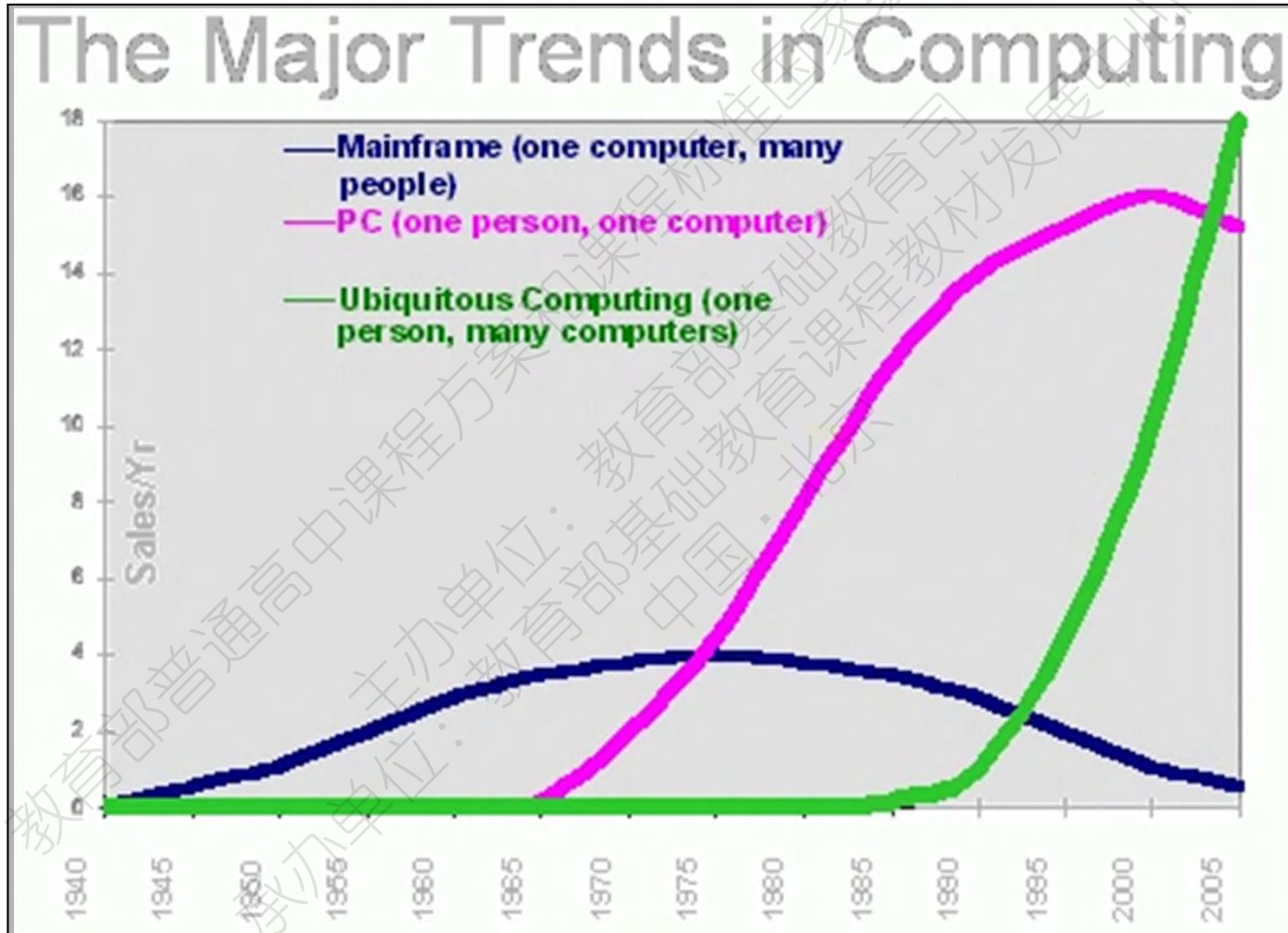
3 Industry EPC Network

2005年 - 重现江湖

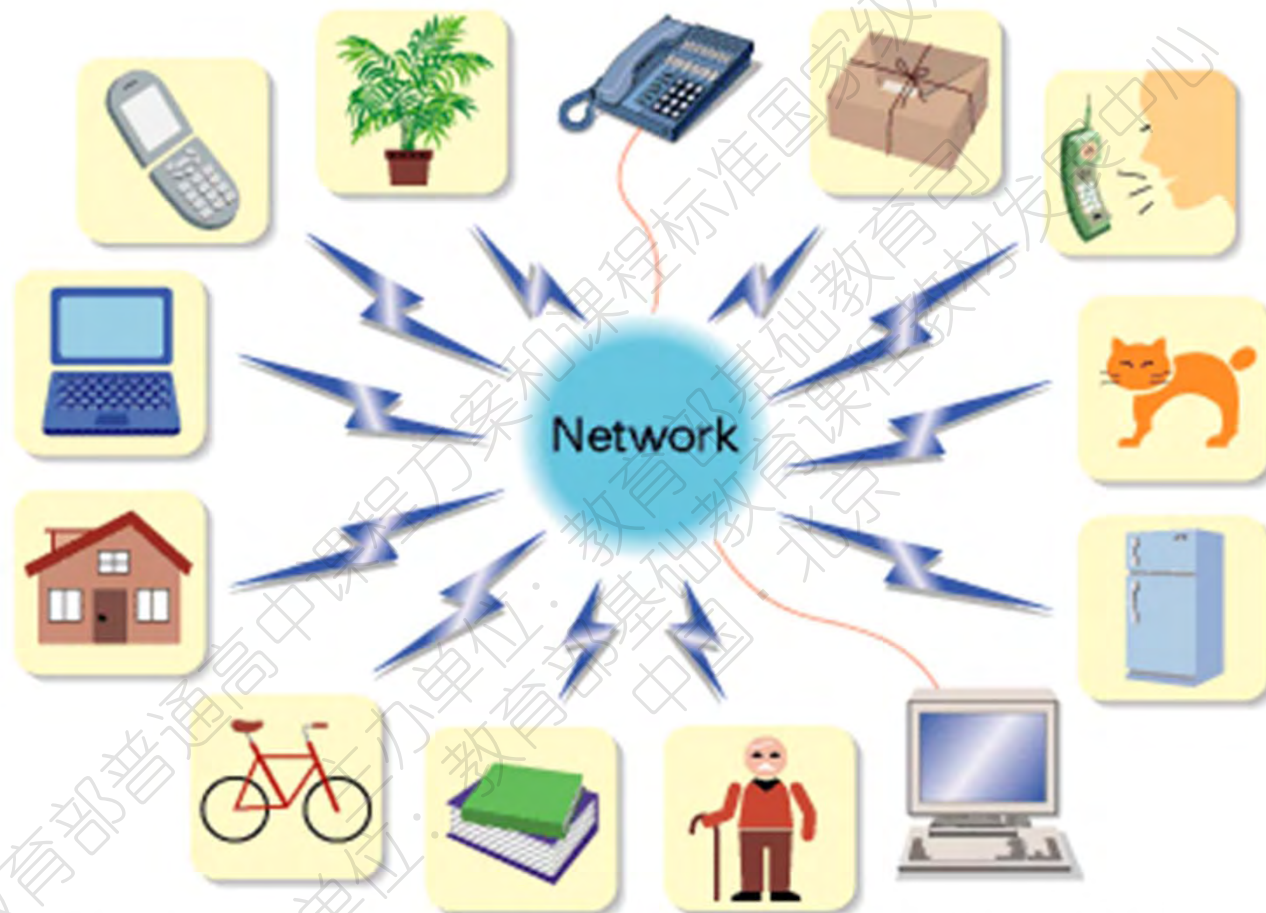


- 任何时刻、任何地点、任何物体之间的互联
- 无所不在的网络和无所不在计算
- 除RFID技术外，传感器技术、纳米技术、智能终端等技术也将得到更加广泛的应用

计算模式的发展趋势

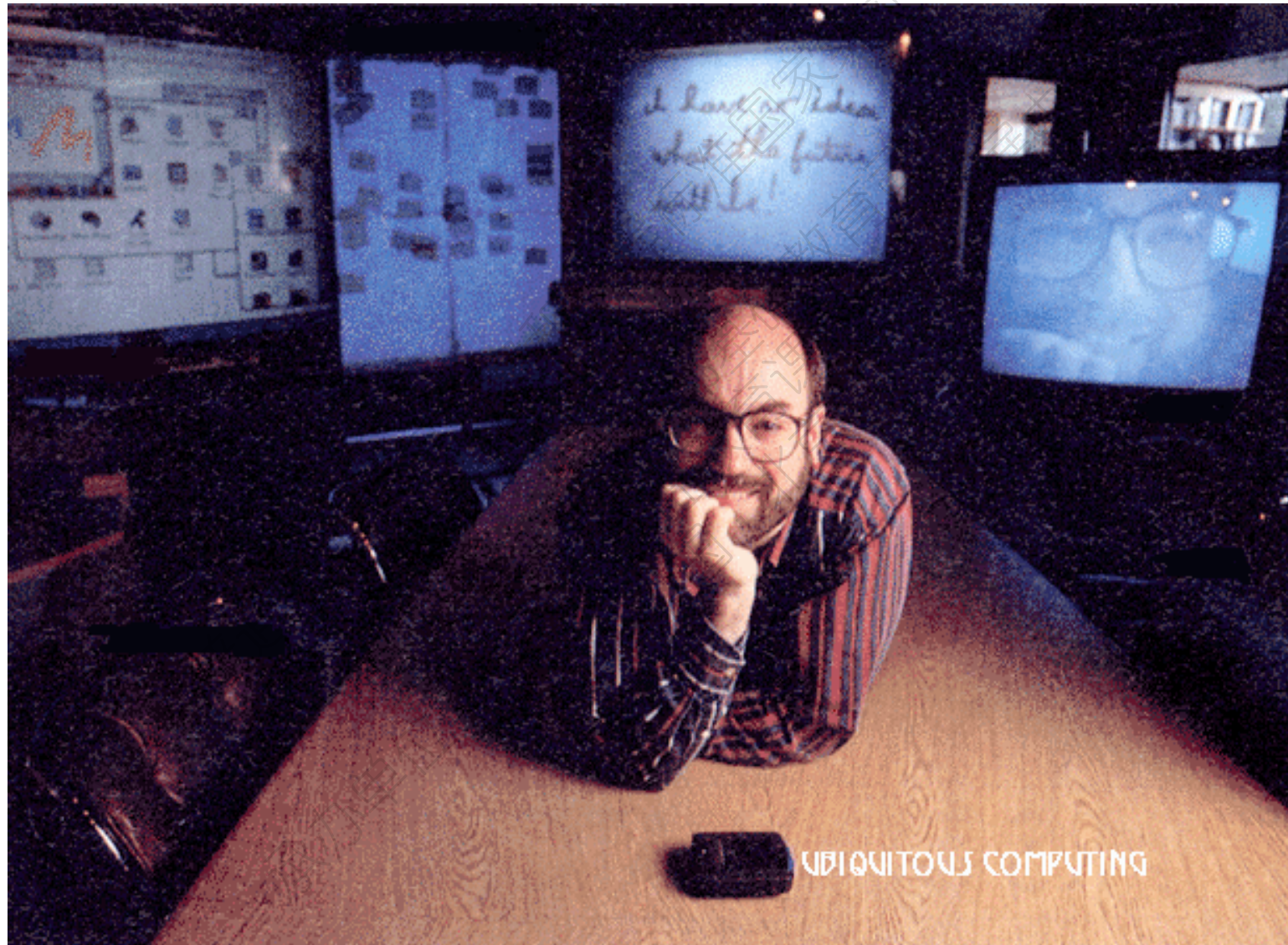


普适计算与泛在计算



Ubiquitous computing will enable diverse wireless applications, including monitoring of pets and houseplants, operation of appliances, keeping track of books and bicycles, and much more.

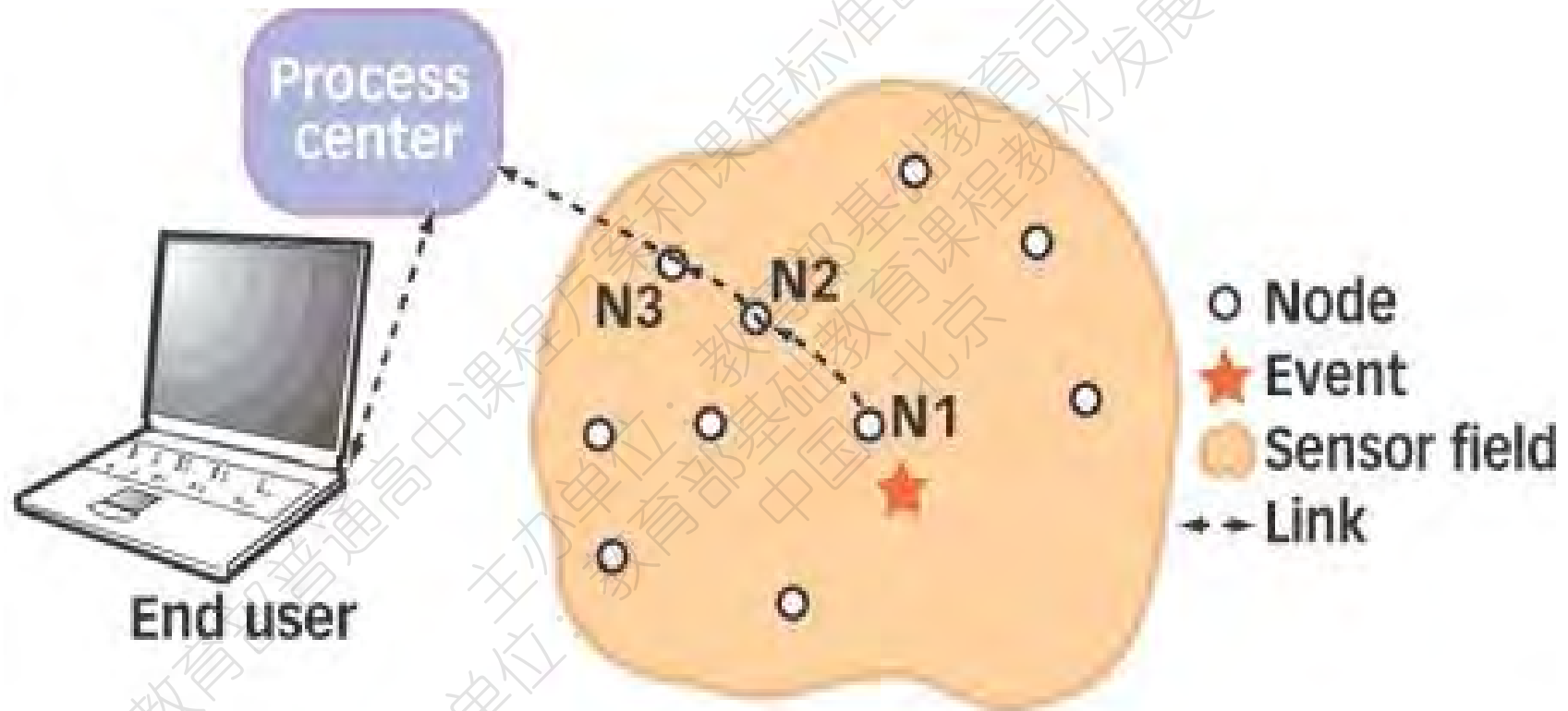
Mark Weiser



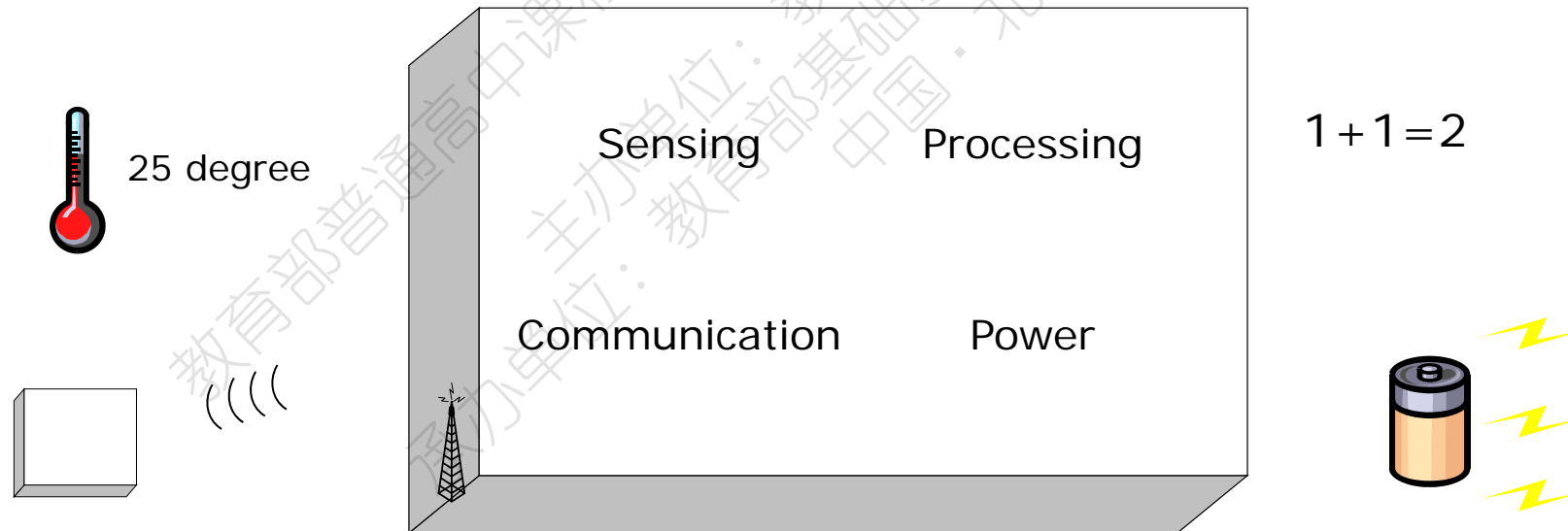
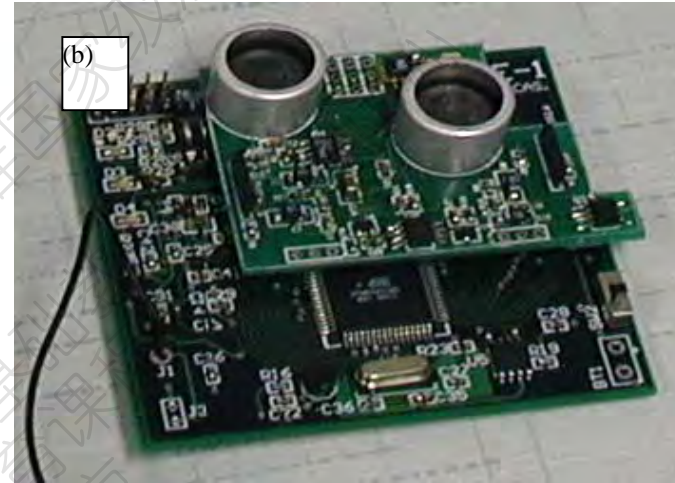
普适计算与泛在计算

- 强调把计算机嵌入到环境或日常工具中去，让计算机本身从人们的视线中消失，让人们注意的中心回归到要完成的任务本身
- 实现的前提条件
 - 尺寸大小不一、种类繁多的显示设备和廉价、低能耗计算设备
 - 存在将所有计算设备（如嵌入式计算设备、辅助设备）联接在一起的网络
 - 研制出用于实现普适计算应用系统的软件支撑系统

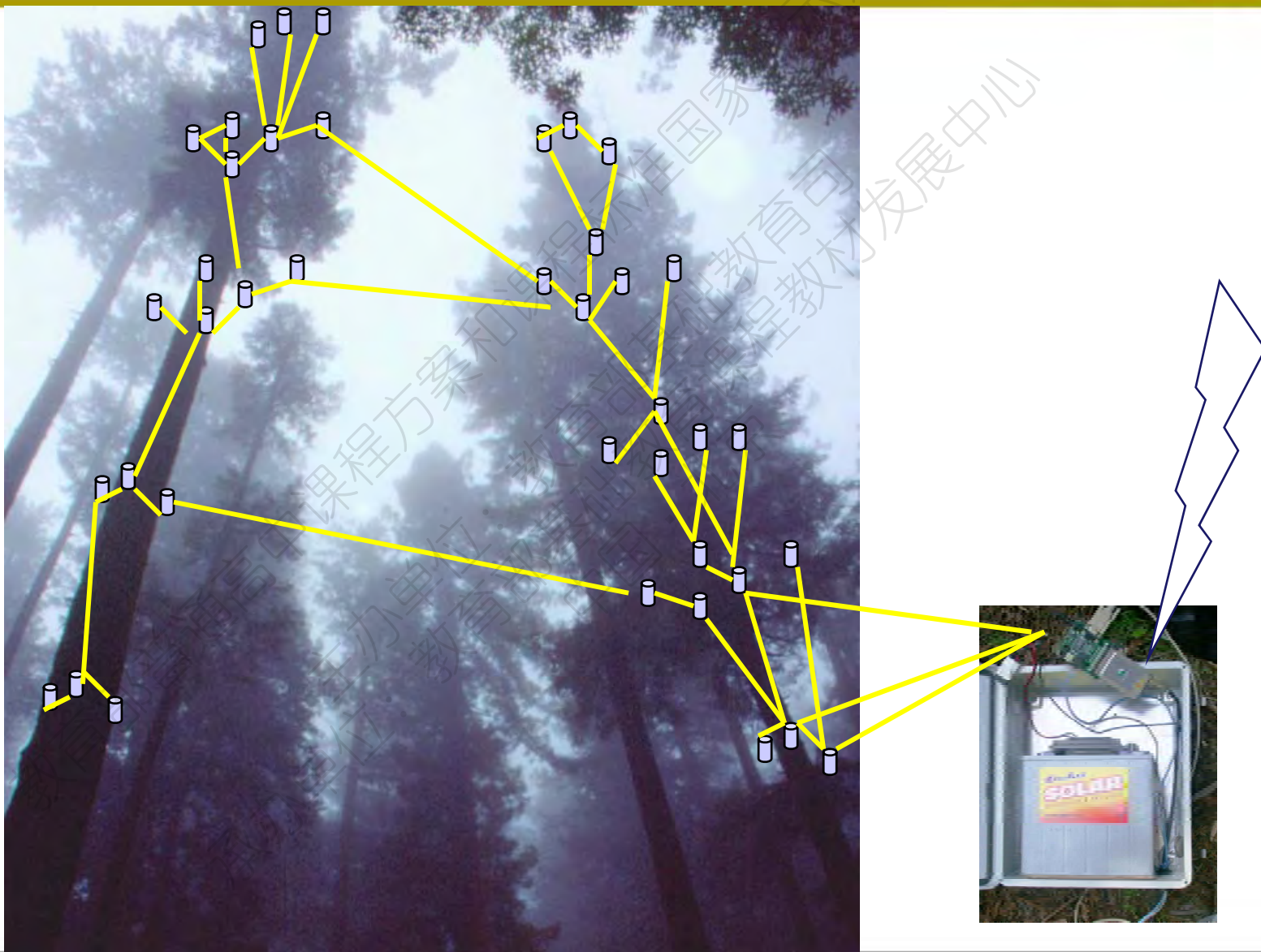
关于无线传感器网络



无线传感器网络节点

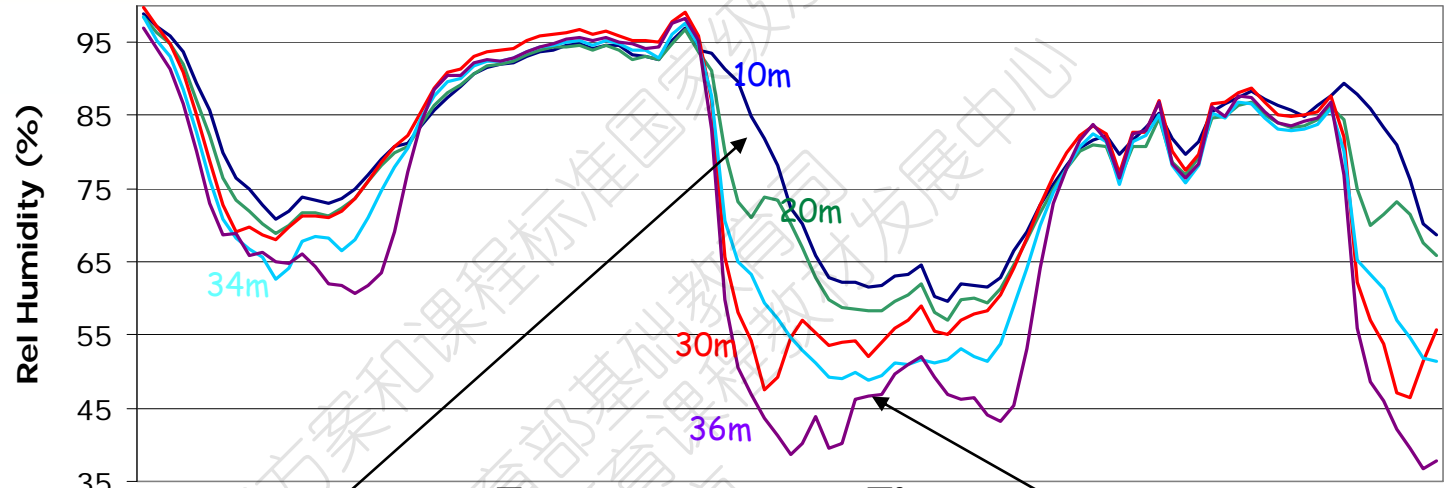


无线传感器（自组织多跳）网络



Humidity vs. Time

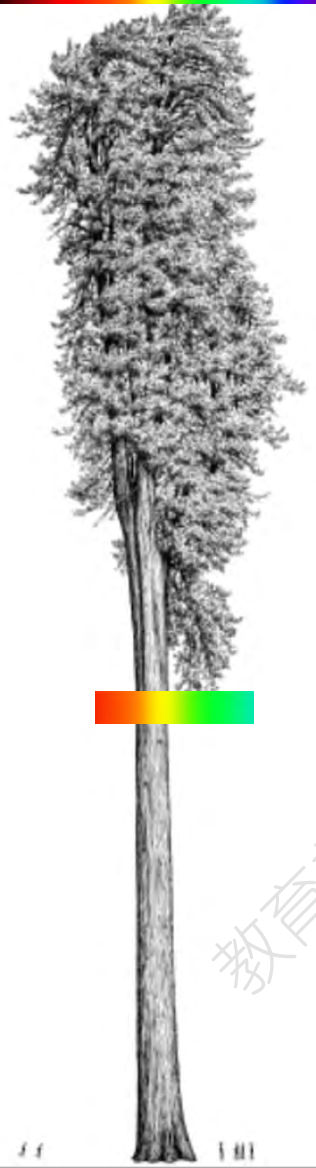
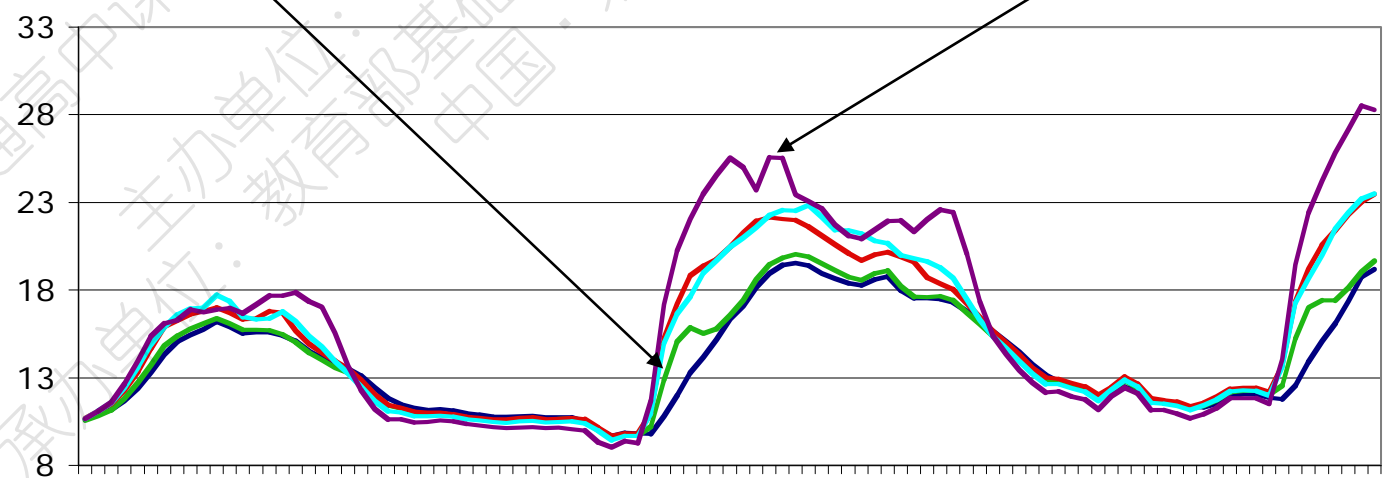
101 104 109 110 111



Temperature vs. Time

Bottom

Top

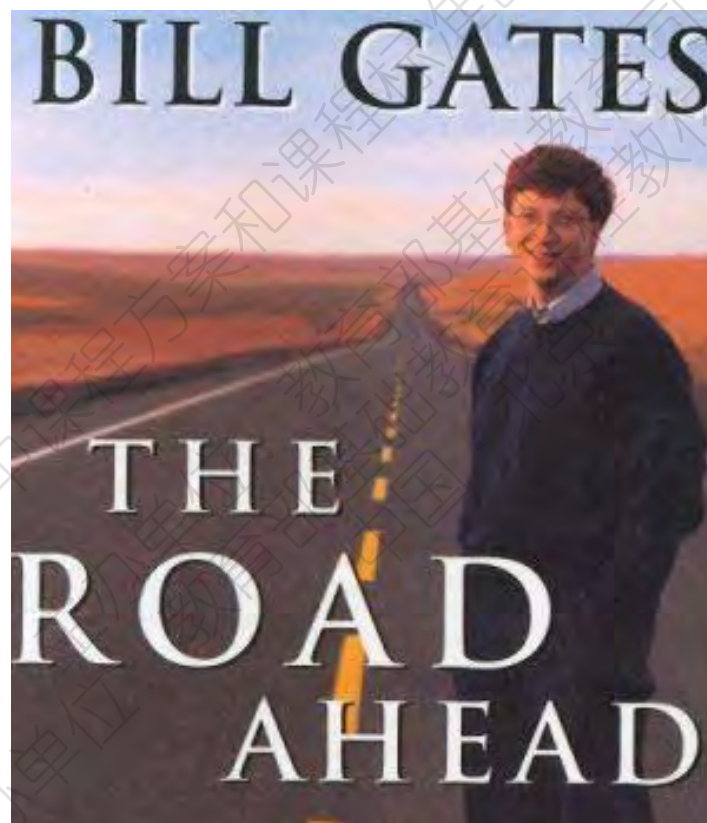


7/7/03 9:40 7/7/03 13:41 7/7/03 17:43 7/7/03 21:45 8/7/03 1:47 8/7/03 5:49 8/7/03 9:51 8/7/03 13:53 8/7/03 17:55 8/7/03 21:57 9/7/03 1:59 9/7/03 6:01 9/7/03 10:03

Date



倒叙……未来之路（1995）



强势复出（2009）

- 2008年 IBM 智慧的地球
- 2009年1月
 - IBM建议美国政府加大智慧型基础设施投资
 - 奥巴马将物联网写入美国复兴与再投资计划
- 2009年8月7日 温总理无锡重要讲话
 - 传感系统与3G TD技术
 - 中国的传感信息中心——“感知中国”中心
- 2009年9月15日 欧盟
 - 发布《物联网 战略研究路线图》

物联网与中国——战略意义

- 继计算机、互联网之后的第三次浪潮
- 全球尚没有成熟的标准体系
- 同一起跑线，是赶超世界的历史性机遇

2010物联网元年

物联网等新兴战略产业
首次写入国务院政府工作报告

《物联网“十二五”发展规划》



智能工业



智能物流



智能农业



智能电网



智能交通



智能环保



智能安防



智能家居



智能医疗

物联网白皮书—电信研究院2011



物联网感知层



- **感知层是实现物联网全面的感知的基础**
- 包括二维码标签和识读器、RFID标签和读写器、摄像头、GPS、传感器和M2M终端、传感器网络和传感器网关等
- 要解决的重点问题是感知和识别物体，采集和捕获信息
- 要突破的方向是具备更敏感、更全面的感知能力，解决低功耗、小型化和低成本的问题

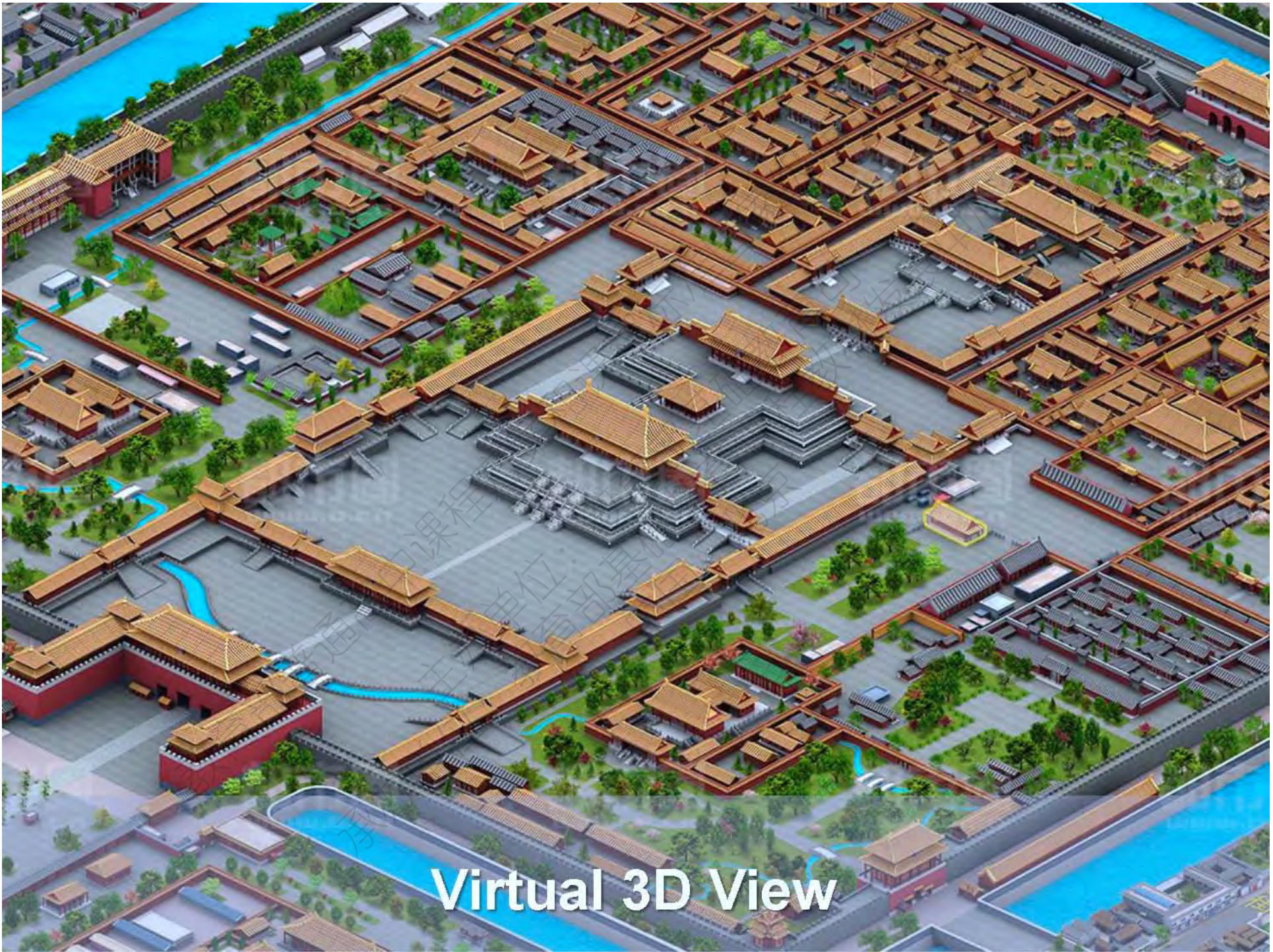


Satellite View

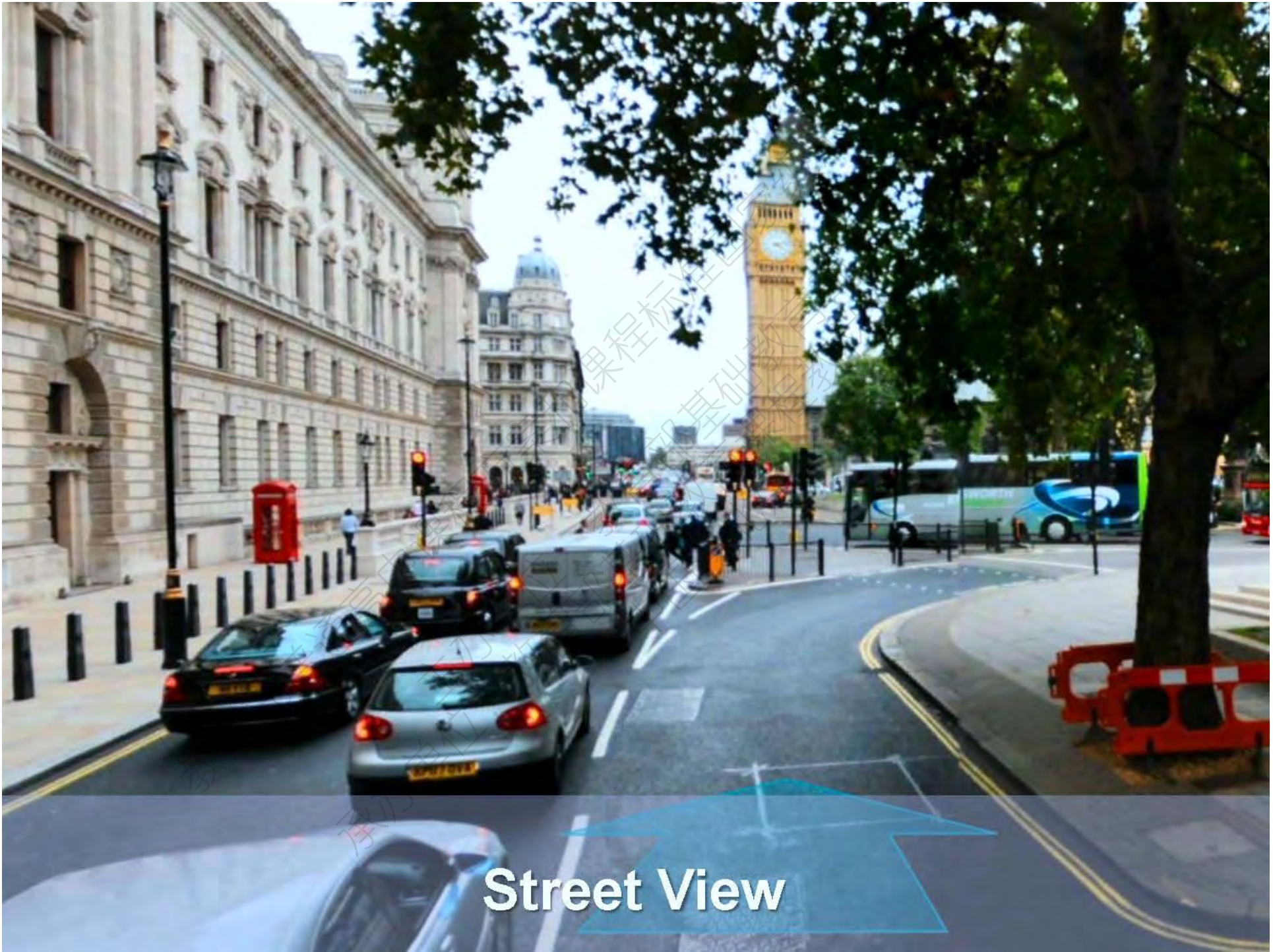


London

City View



Virtual 3D View



Street View

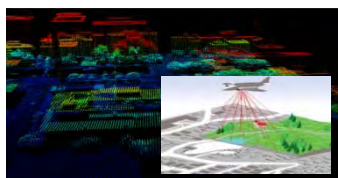


3D Textured TIN View

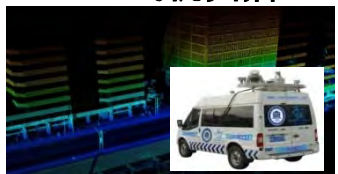
物联网连接了物理世界与信息世界

城市三维感知技术

“视觉”



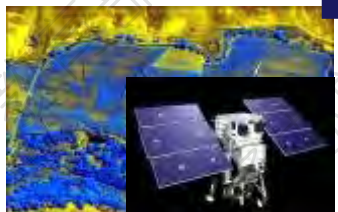
飞机扫描



车载激光扫描

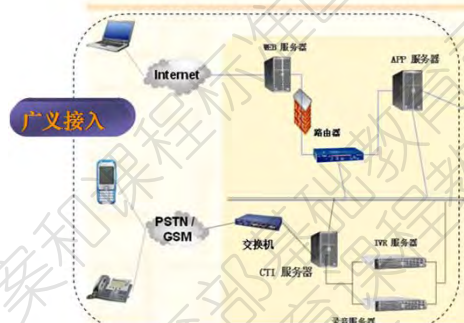


摄像头智能监控



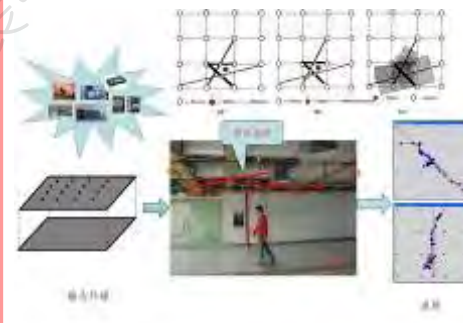
卫星遥感

“听觉”



传统电信网络

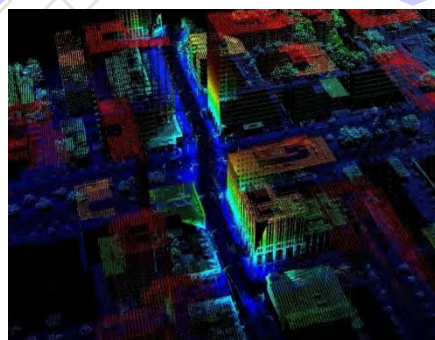
“触觉”



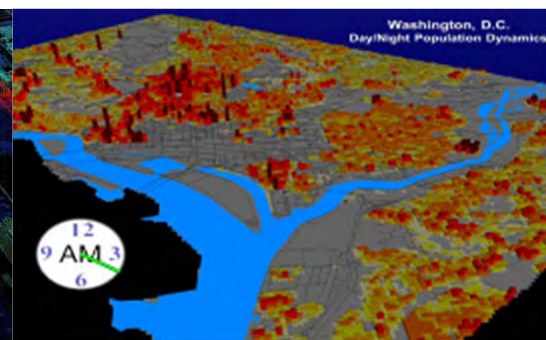
无线传感网络

融合

融合

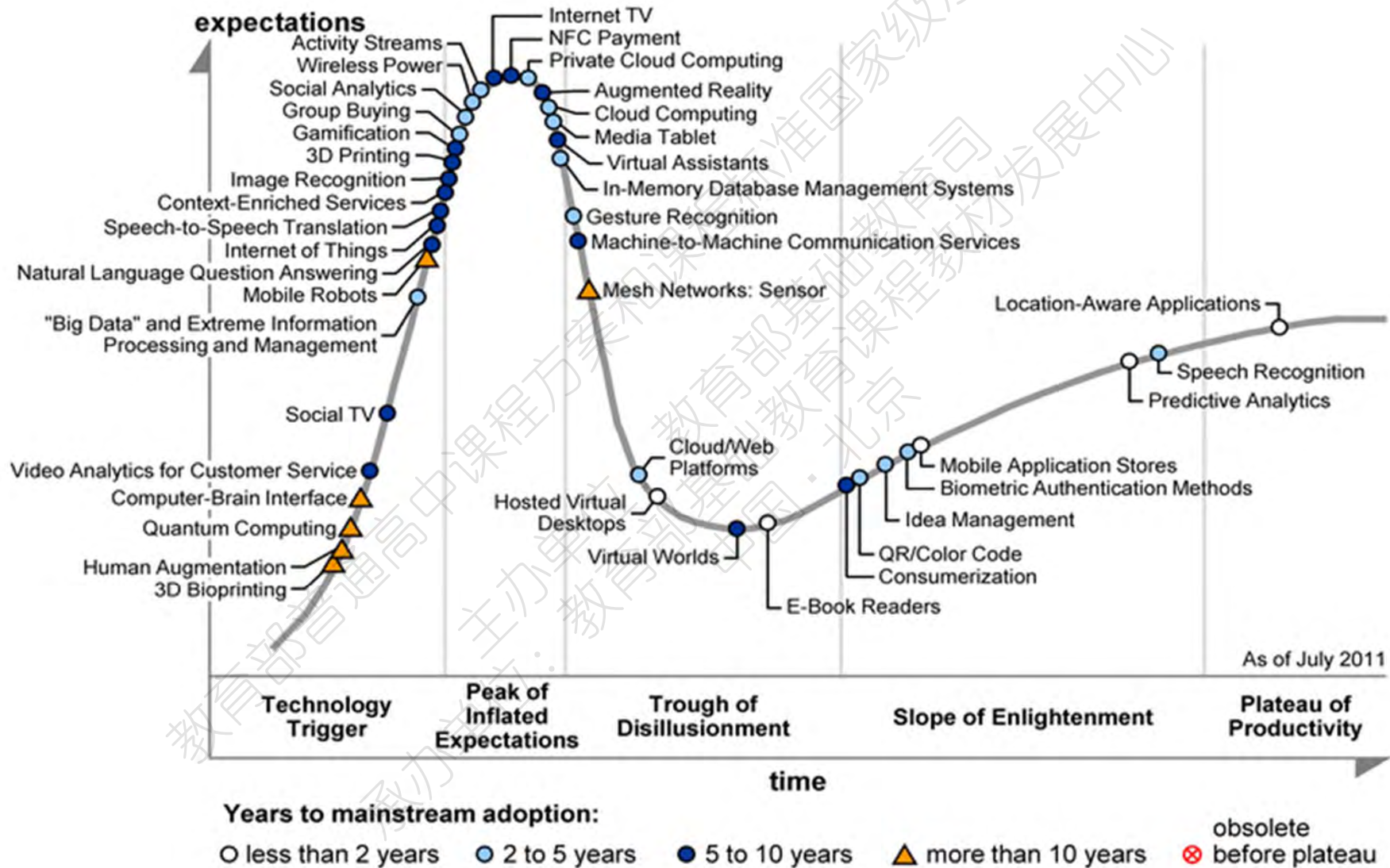


高精度三维建模



人口、车辆信息动态分布

Gartner 技术成熟度曲线 2011



全美航空公司1549航班迫降事件

2009年1月15日 纽约



全美航空公司1549航班迫降事件



程方案和课程标准国家级示范中心
教育部普通高
主办单位
教育部基础
中国
教育部基础
课程教材发展中心

计算机科学的发展阶段

- CPU为核心
- 网络为核心
- 搜索为核心



- 数据为核心

教育部普通高中课程方案和课程标准国家级示范校
主办单位：教育部基础教育司
承办单位：教育部基础教育课程教材发展中心
中国·北京

大数据时代的科技民生

“数据”概念的演变

数字

1, 2, 3, ...

数值

3.1415926

0.618

数值计算

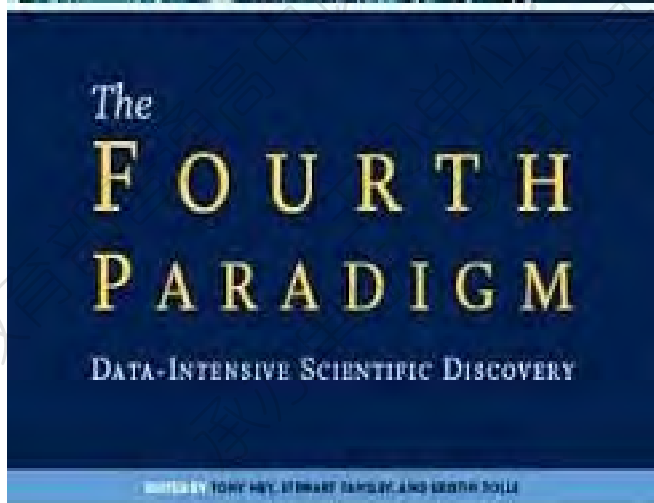
数据

数据处理

统计

大数据时代的科技民生

数据与科学



大数据时代的科技民生

数据与科学

- 图灵奖获得者 Jim Gray
- 数据密集型科学发现 (Data-Intensive Scientific Discovery)
- 科学技术发展的第四范式

大数据时代的科技民生

nature International weekly journal of science

Search [Advanced search](#)

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 527 > Issue 7576 > Outlook > Article

NATURE | OUTLOOK

Big data: The power of petabytes

Michael Eisenstein

Nature 527, S2–S4 (05 November 2015) | doi:10.1038/527S2a

Published online 04 November 2015

[PDF](#) [Citation](#) [Reprints](#) [Rights & permissions](#) [Article metrics](#)

Researchers are struggling to analyse the steadily swelling troves of '-omic' data in the quest for patient-centred health care.

Subject terms: [Genomics](#) · [Molecular biology](#) · [Computational biology and bioinformatics](#) ·

[Drug discovery](#)



It's gone. [Undo](#)

What was wrong with this ad?

- Irrelevant
- Repetitive
- Inappropriate



Editors' pick



Image credit: Peter Steffen/DPA/PA

大数据时代的科技民生

数据与经济

- 2015年诺贝尔经济学奖
- 普林斯顿大学安格斯·迪顿
(Angus Deaton)



数据与经济

- 官方颁奖词：
 - 减少贫困的经济政策，必须了解个人消费的选择
 - 个人选择的细节和汇总
 - 改变了微观经济学、宏观经济学和发展经济学

大数据时代的科技民生

数据与工业

- 工业1.0是蒸汽机带来的机械化
- 工业2.0是电力带来的电动化
- 工业3.0是计算机带来的自动化
- 工业4.0就是数据带来的智能化

数据与社会

- 《自然》2015年2月
- 查尔斯·赛费
- 不仅带来了商业和科学的革命，同时带来了社会的变化



大数据时代的科技民生

数据与社会

- 谷歌
- 80种语言的即时互译
- 每天10亿次的翻译服务
- 服务于联合国不同语种的翻译

大数据时代的科技民生

数据与日常生活

- 天气预报
- 导航
- 购物
- 订餐
-

大数据时代的科技民生

数据与社会科学

- 社交应用是数据的来源
- 数据应用又作用于社会

大数据时代的科技民生

数据与社会科学

- Quantitative Social Science
- 量化社会科学
- 哈佛大学Gary King教授
- 政治学、公共政策、法学、心理学

大数据时代的科技民生

数据与社会科学

- 64000篇国会议员的新闻稿
- 高达27%的议员发布的新闻稿内容只是单纯地想抨击对方,而不是想要解决问题

大数据时代的科技民生



© 视觉中国

大数据时代的科技民生

大数据的实质

- 所有数据
- 多源多态
- 新的处理方式
- 新的应用
- 推动了科技民生的变革

大数据时代的科技民生

大数据是什么？

- 工信部长苗圩：大数据是21世纪的石油和金矿，是一个国家提升综合竞争力的又一关键资源（人民日报2015年10月13日）

大数据时代的科技民生

大数据是什么？

- “数据” VS 石油和矿石
- “数据” VS 农具和机器
- “数据” VS 像高速路和机场
- 原材料、生产资料、基础设施

大数据时代的科技民生

- Wisdom

智慧

- Knowledge

知识

- Information

信息

- Data

数据

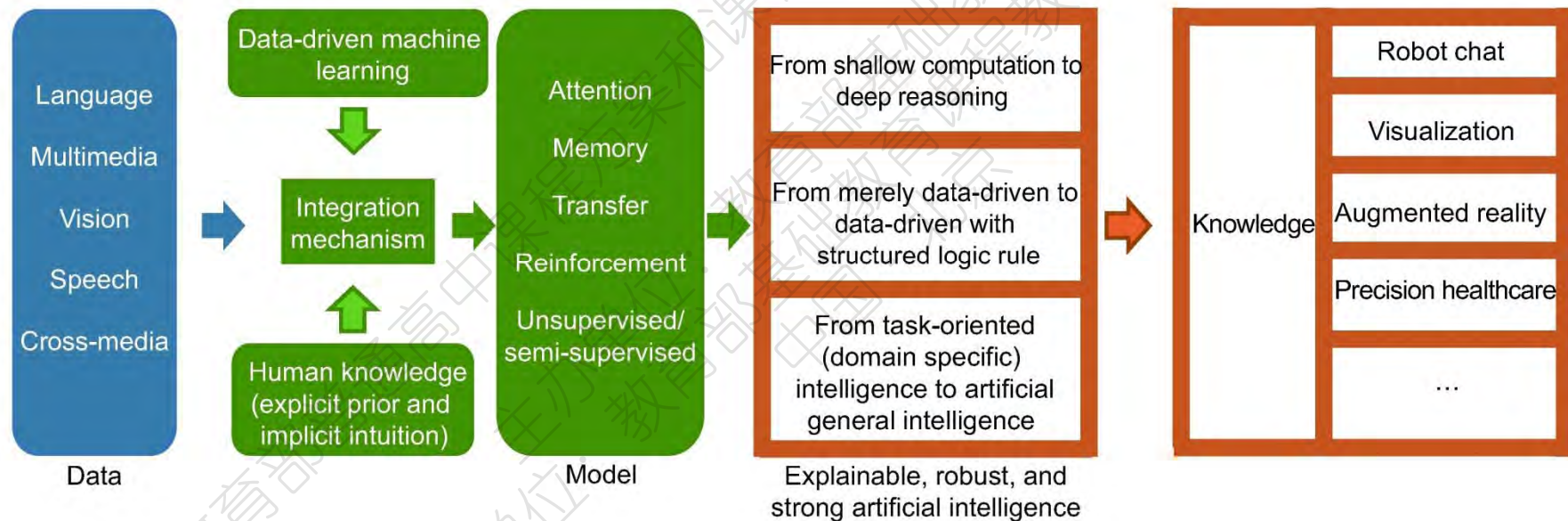
定性

+

定量

大数据时代的科技民生

- 下一代人工智能（AI 2.0）将改变计算本身，将大数据转变为知识，以支持人类社会更好决策



- 中国工程院院刊：人工智能迈向2.0时代（2017.2）

- ▶ 大数据教育与信息技术教育的关系解读
- ▶ 大数据在课程标准中的设计与要求
- ▶ 教学建议与案例研讨

教育部普通高中课程方案和课程标准国家级示范培训
主办单位：教育部基础教育课程教材发展中心
承办单位：教育部基础教育课程教材发展中心
中国·北京

大数据的特征

- 大数据的4V特征
 - **Volume** (巨大的数据量)
 - **Variety** (数据类型多)
 - **Velocity** (处理速度快)
 - **Value** (价值密度低)

大数据的特征

- 大数据的4V特征
 - **Volume** (巨大的数据量)
 - Variety (数据类型多)
 - Velocity (处理速度快)
 - Value (价值密度低)

大数据的特征

2

Data inflation

Unit	Size	What it means
Bit (b)	1 or 0	Short for “binary digit”, after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10} , bytes	From “thousand” in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From “large” in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From “giant” in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From “monster” in Greek. All the catalogued books in America’s Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America’s postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of <i>The Economist</i>
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Source: *The Economist*

大数据的特征

1956年，IBM正在移动一个只有5兆的硬盘。



大数据的特征

- 大数据的4V特征

- Volume (巨大的数据量)

- **Variety** (数据类型多)

- Velocity (处理速度快)

- Value (价值密度低)

大数据类别

- **数据类型**

- 结构化数据

- 关系数据等：数据的查询、统计、更新等操作效率低。

- 半结构化数据

- XML、**图数据**等（word、ppt）：转换为结构化存储或者按照非结构化存储。

- 非结构化数据

- 图片、视频等：不利于检索、查询和存储

大数据类别

- 行业数据

- 大规模的电子商务数据

- 社会数据（社会网络，互联网等），是一类重要的图数据

- 移动数据（通信记录、RFID、传感器网络）

- 医疗数据

- 天文学，大气科学，基因组学，生物地球化学，生物和其他复杂和/或跨学科的科研数据

大数据的特征

- 大数据的4V特征
 - Volume (巨大的数据量)
 - Variety (数据类型多)
 - **Velocity** (处理速度快)
 - Value (价值密度低)

处理速度快



$$\frac{7200000}{0.34} \approx 21,176,470 \text{ 条/秒}$$

"互联网有1万亿个网页，人类的大脑有1000亿个神经元。" Kelly 在他2010年出版的书《科技需要什么》中写道。

大数据的特征

- 大数据的4V特征

- Volume (巨大的数据量)

- Variety (数据类型多)

- Velocity (处理速度快)

- **Value** (价值密度低)

价值密度高



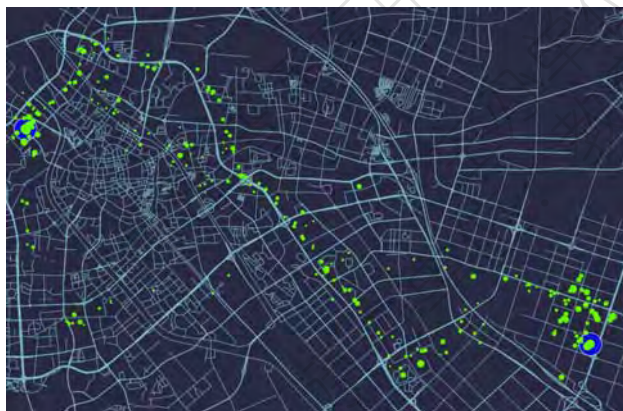
基于手机信令数据的居民行为挖掘

研究目标：基于居民历史手机信令数据，恢复其从居住地到工作地所经过的实际路径

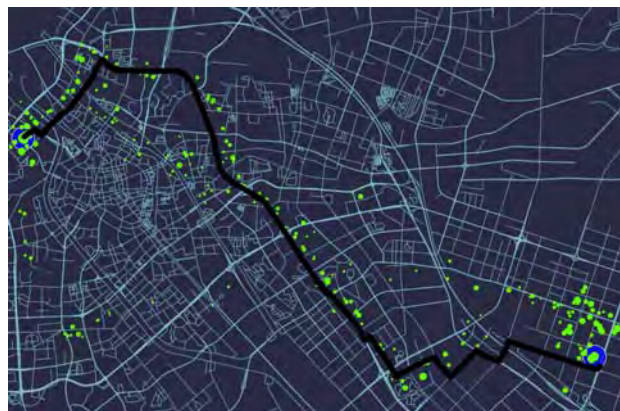


手机信令数据

- 无锡市移动手机信令数据
- 居民手机设备与临近基站联系，在基站上留下记录信息
- 设备匿名ID，基站ID，时间戳和标签



手机信令数据



用户通勤路径

基于手机信令数据的居民行为挖掘

为什么选择手机信令数据？

基于调查问卷（传统方法）

- 人力成本极高
- 受主观影响大
- 采样数量有限

1. How did you travel to work today? PLEASE TICK ALL MODES OF TRAVEL USED, NOT JUST THE MAIN ONE

Bicycle Go to Q2
Bus
Car - driver
Car - passenger Go to Q2
Foot
Motorbike
Train
Working from home Go to Q10
Other Go to Q2

IF OTHER, PLEASE SPECIFY _____ Go to Q2

ALL WHO TRAVELLED TO WORK TODAY

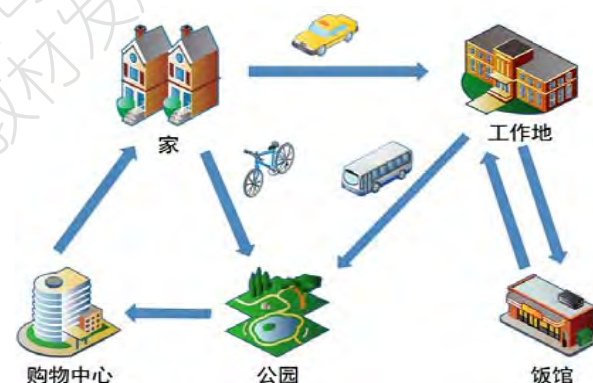
2. From the options selected in Q1, which was your main mode of transport (ie. most time spent)? PLEASE TICK ONE

Bicycle Go to Q3
Bus
Car - driver Go to Q3
Car - passenger
Foot
Motorbike Go to Q3
Train
Other as previously specified _____

ALL WHO USED CAR AS MAIN MODE OF TRANSPORT

3. Including yourself, how many people were travelling in the car?

ENTER THE NUMBER OF INDIVIDUALS IN THIS BOX _____



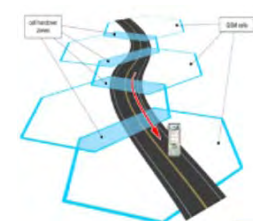
基于手机信令数据（本文方法）

- 快速
- 可靠
- 样本覆盖率高

每个人都有手机

时时刻刻用手机

大量手机信令数据



数据清洗

- 时间段：2013年10月24日至2014年3月24日
- 用户数量：600万（每小时平均4000万条原始记录）
- 问题：同一部手机设备所产生的两条相邻基站记录之间的时间间隔平均约为30分钟，但大量手机设备产生的数据极为稀少，时间间隔远大于30分钟
- 解决方案：清洗掉每天产生记录小于24条的设备，同时清洗掉可判定出其居住地和工作地相同的用户数据

基站聚类

问题

手机数据位置信息精确度低

- GPS设备精度：3米
- 基站位置精度：1000米

解决方案

对比：基于密度的聚类算法

- 聚类范围偏大，不利于路径匹配任务

Leader 聚类算法

- 充分利用了权重信息
- 聚类大小易于控制



通勤轨迹提取

重要位置检测

依据：停留时间长，产生的记录数量多

休闲时间：下午7点到早上6点

$$home = \{group_i | \max(R_i) \cap t \subset HomeTime\}$$

工作时间：下午1点到下午5点

$$work = \{group_i | \max(R_i) \cap t \subset WorkTime\}$$



通勤轨迹提取

问题：手机信令数据采样率低

- GPS数据：2秒
- 手机数据：1800秒
- 融合多天数据，提升采样率

提取居住地到工作地之间的聚类轨迹，

得到通勤聚类轨迹： $seq_i = \{g_1^i, g_2^i, \dots, g_n^i\}$



路径匹配


数据融合

问题：多天轨迹融合

- 极大提升了采样点的数量
- 临近采样点**拓扑关系**难以判断

数据融合算法

- **多天数据融合的关键**
- 采用**转移矩阵**进行数据融合
- 使用最短路径补齐不连续路段

 $seq_i = \{g_1^i, g_2^i, \dots, g_n^i\}$

 聚类间转移矩阵： $M_{G \times G}$

 路段间转移矩阵： $M_{R \times R}$

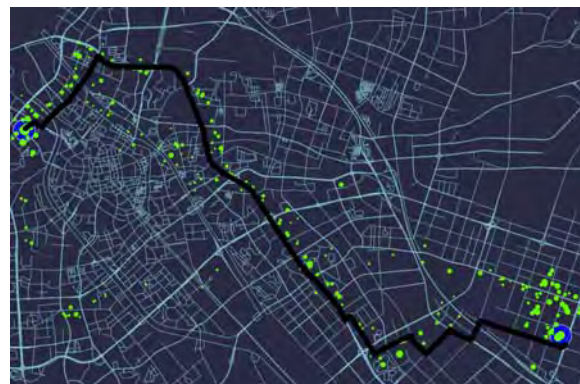
路径生成

路径生成算法

- **从加权有向图中生成最短路径**
- 对路段间转移矩阵进行参数更新

$$m_{i,j} = \frac{\maxValue - m_{i,j}}{m_{i,j}}$$

- 平衡路径总长度与路径总频数
- 应用Dijkstra算法生成最短路径



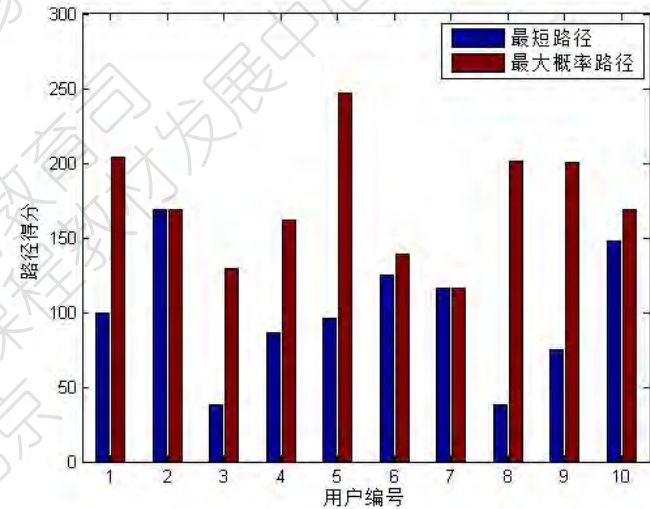
通勤路径可视化



实验验证与分析

评价方法

- 对照方法：最短路径
- 问题：缺少用户真实通勤路径信息
- 路径评分
- 路径可视化工具



路径评分评估

- 依据：最准确的路径周围附着最多的基站记录
- 评分：计算距离路径 $P = \{r_1, r_2, \dots, r_n\}$ 中每个路段300米内的基站上用户留下的记录的总条数

$$\text{数 } score = \sum_{i=1}^n \text{sizeof}(C_i)$$

大数据的本质是一次思维方式革命

教育部普通高中课程标准和课程标准国家级示范培训
主办单位：教育部基础教育司
承办单位：教育部基础课程教材发展中心
中国·北京

大数据的本质是一次思维方式革命

01 更多

不是随机样本
而是全体数据

大数据时代

收集与分析全体数据是可行和便宜的

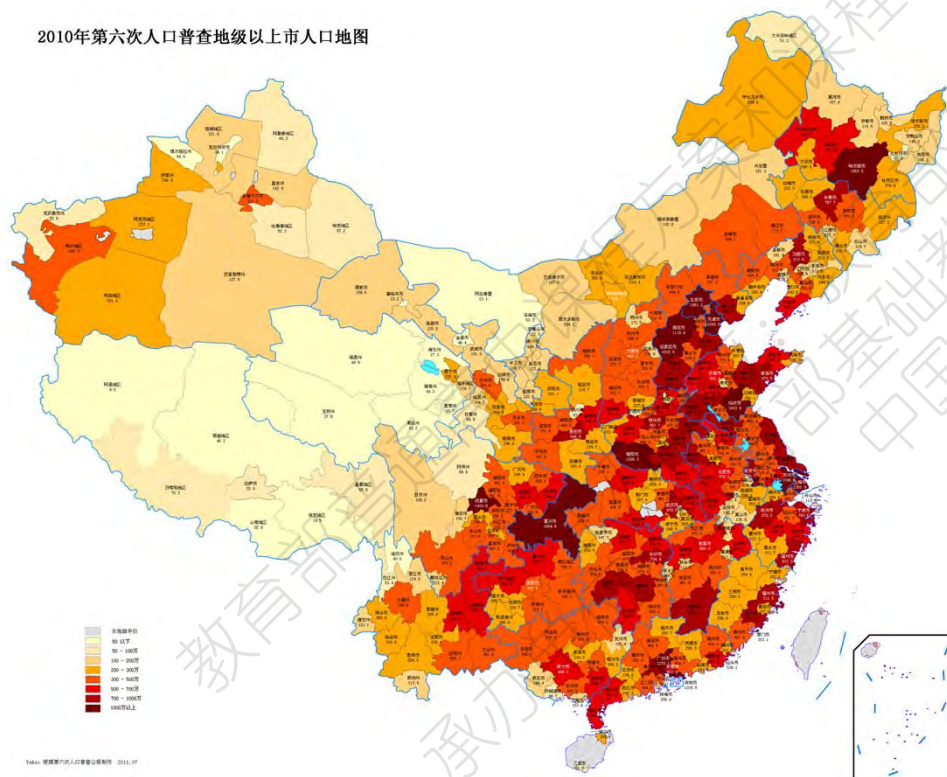
小数据时代

受制于技术只能收集与分析随机样本

人口普查

从1949年至今，中国分别在**1953年**、**1964年**、**1982年**、**1990年**、**2000年**与**2010年**进行过六次全国人口普查。

2010年第六次人口普查地级以上人口地图



1%人口抽样调查

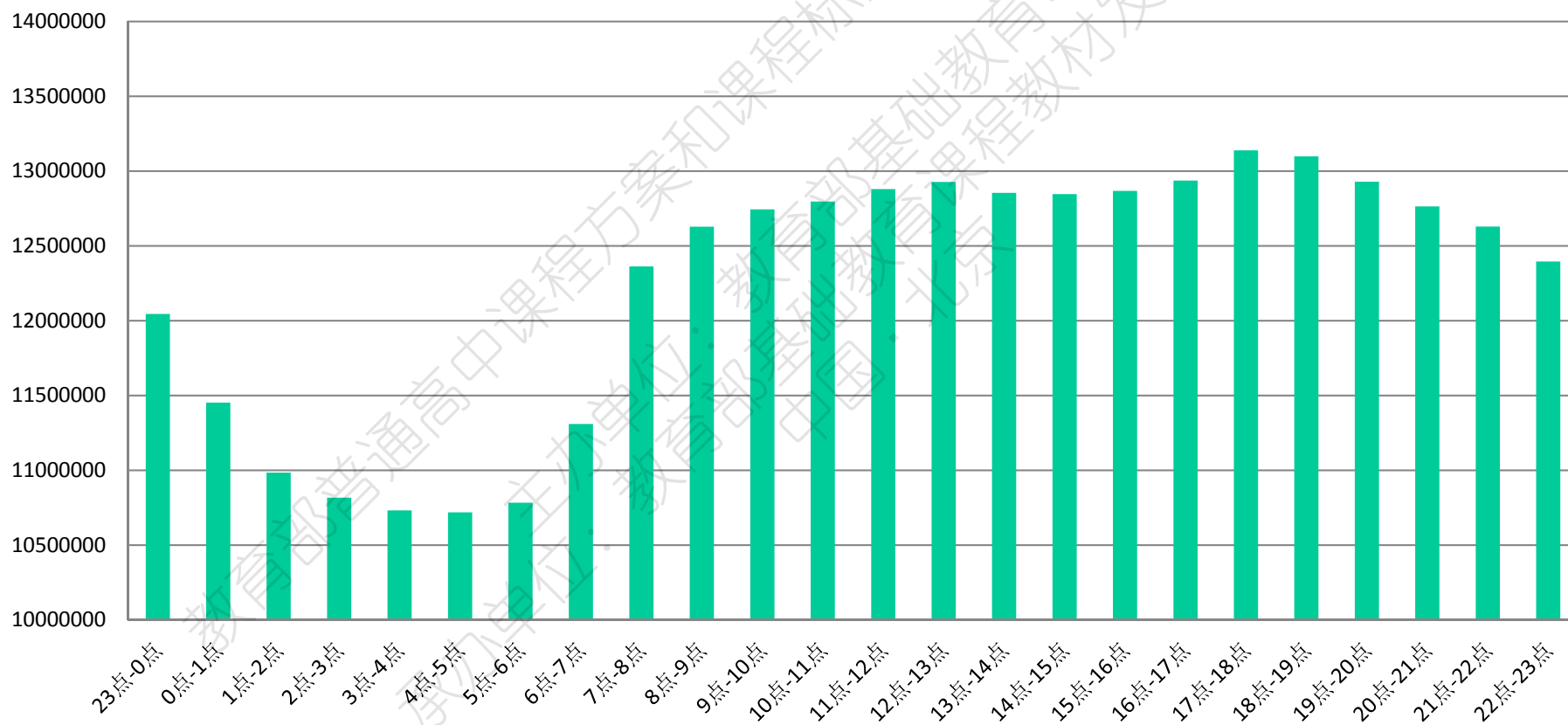
Result Accuracy



Time Spent

基于手机数据的人口普查

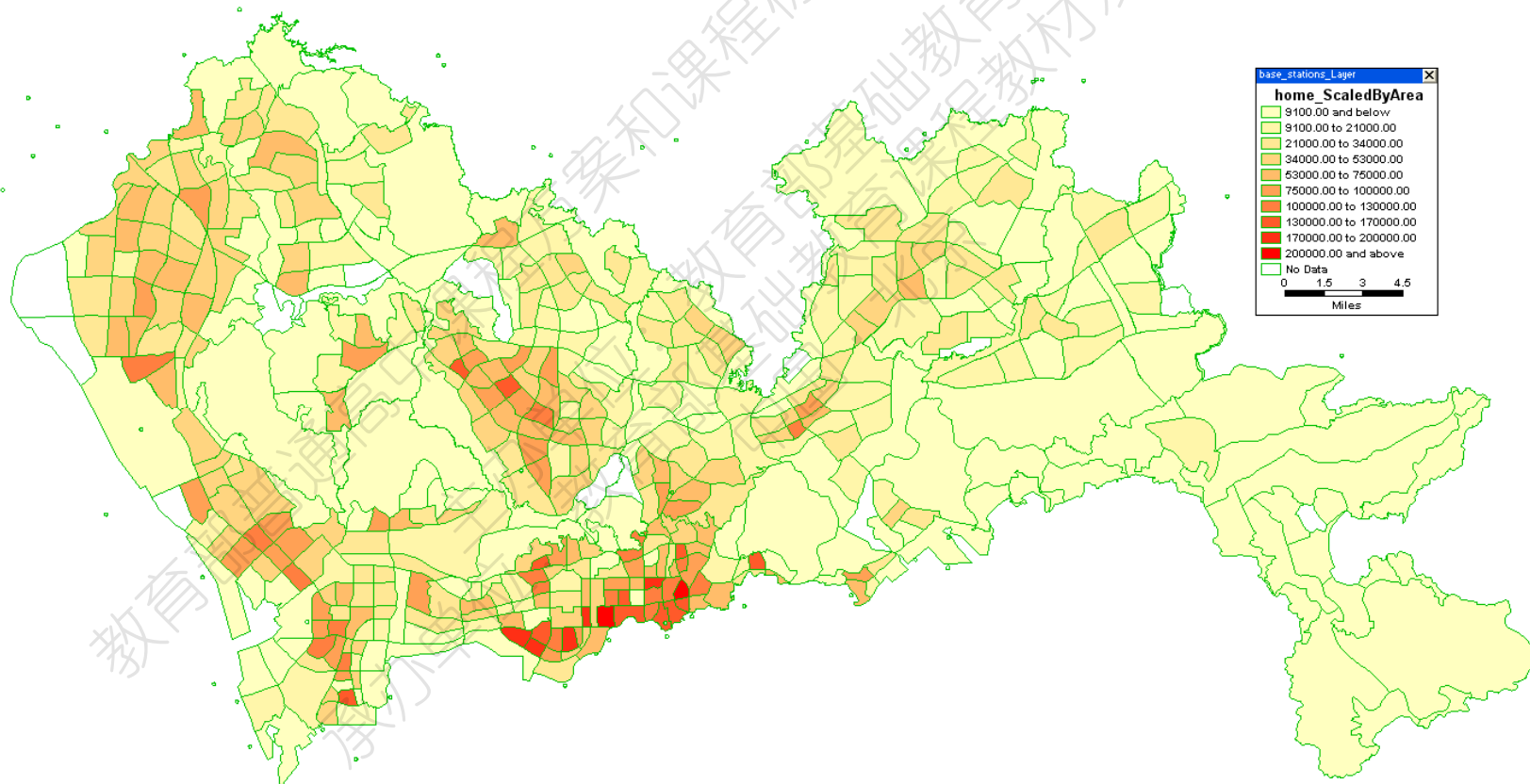
每个小时记录到的用户数的变化情况



基于手机数据的人口普查

根据数据推断居住地分布

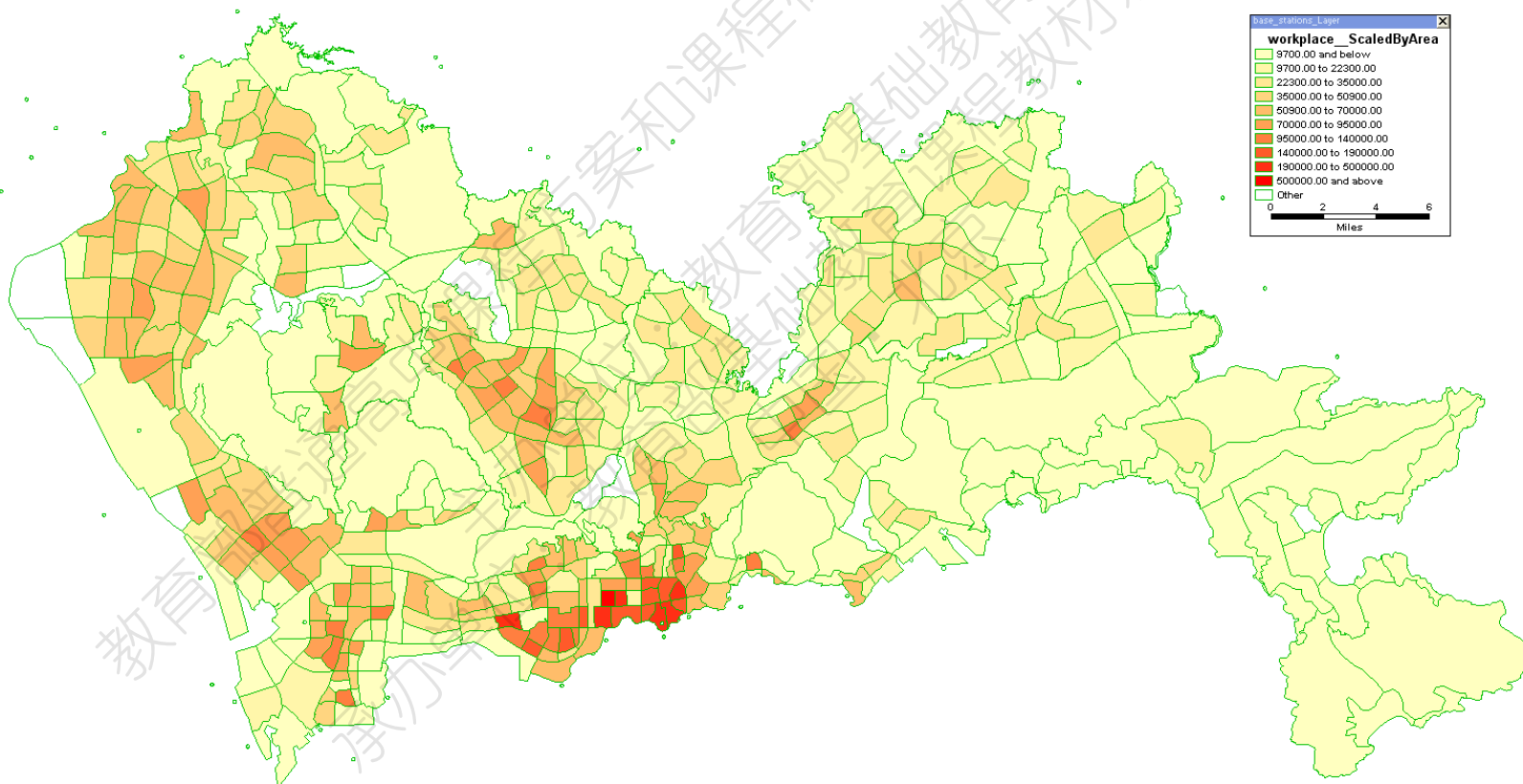
通过识别某用户**晚10点至早9点间**出现次数最多的地点来推测该用户的居住地



基于手机数据的人口普查

根据数据推断工作地分布

通过识别某用户**早9点至晚9点间**出现次数最多的地点来推测该用户的工作地



海地地震

- 2010年海地地震，海地人散落在全国各地，援助人员为弄清该去哪里援助手忙脚乱。传统上，他们只能通过飞往灾区上空来查找需要援助的人群。
- 一些研究人员采取了一种不同的做法：他们开始跟踪海地人所持手机内部的SIM卡，由此判断出手机持有人所处的位置和行动方向。正如一份联合国 (UN) 报告所述，此举帮助他们“准确地分析出了逾60万名海地人逃离太子港之后的目的地。”后来，当海地爆发霍乱疫情时，同一批研究人员再次通过追踪SIM卡把药品投放到正确的地点，阻止了疫情的蔓延。

如何估计全球华人分布情况？

微信全球活跃用户：9.8亿

其中海外用户：约2亿

海外用户中非华裔用户：约5千万

大数据的本质是一次思维方式革命

02 更杂

不是精确性
而是混杂性

大数据时代

追求大量数据，允许不精确的数据

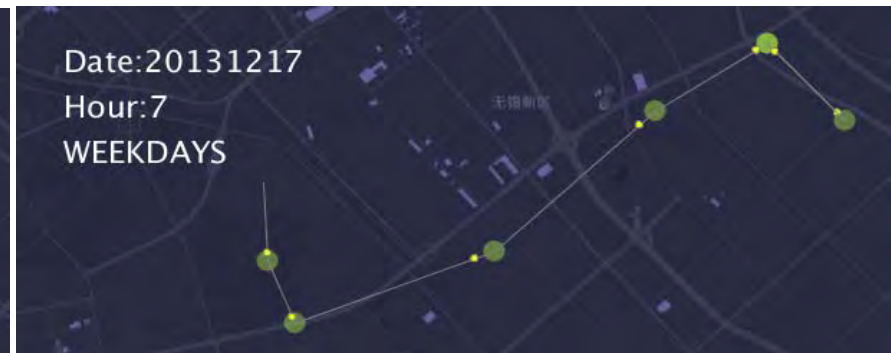
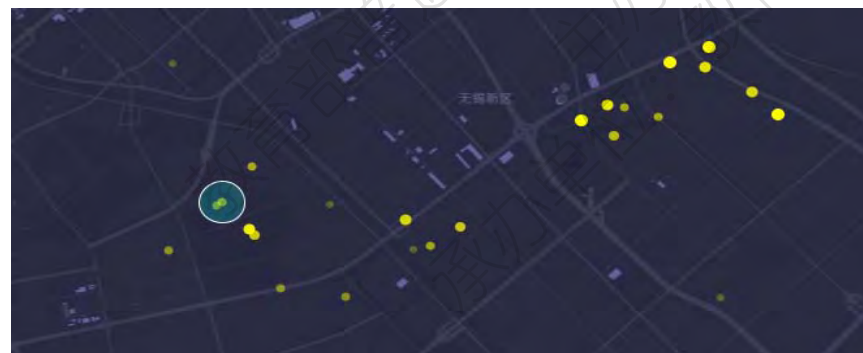
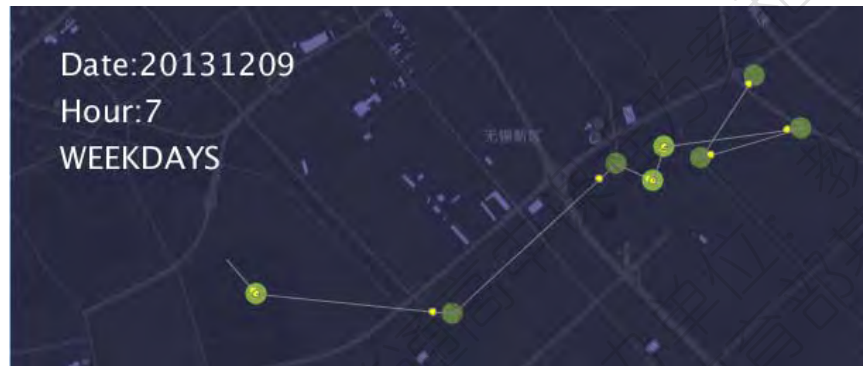
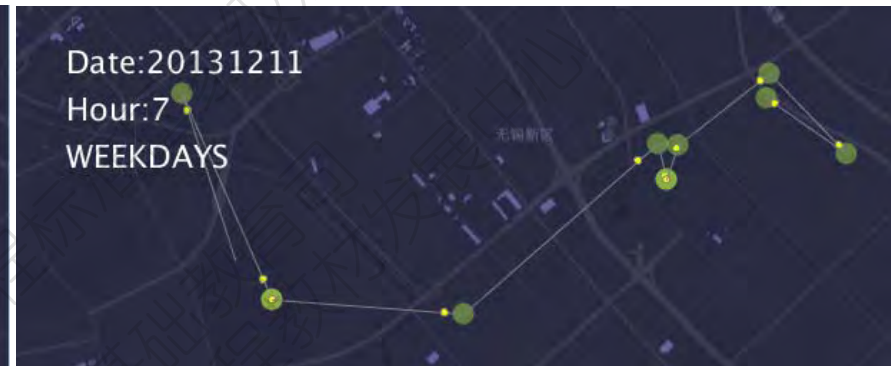
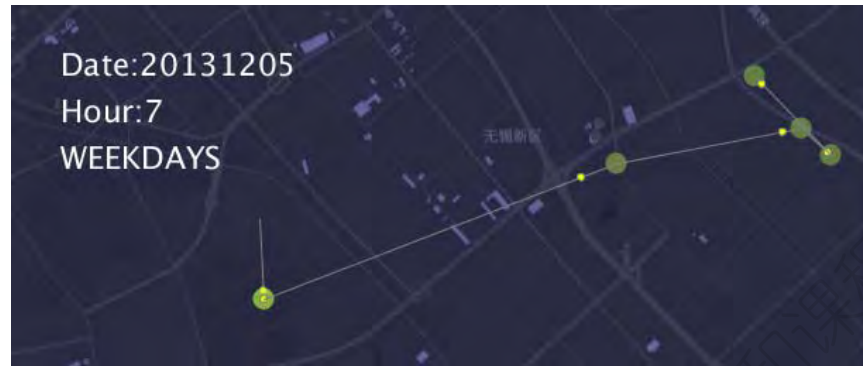
小数据时代

因信息量少，对数据精确性更苛刻

手表定律



个体行为的不确定性



传统数据库

SQL



增、删、查、改

原子性

	订单ID	产品	单价	数量	进价	折扣
▶	10001	17	14	12	7	0
	10002	42	10	10	5	0
	10003	72	35	5	20	0
	10004	14	19	9	11	0
	10005	51	42	40	27	0
	10006	41	8	10	5	0

大数据时代的准确与容错



大数据的本质是一次思维方式革命

03更好

不是因果关系
而是相关关系

大数据时代

相关关系大放异彩

小数据时代

相关关系是有用的

相关性



精准预测

- 马云对未来的预测，建立对在用户行为分析的基础上
- “2008年初，阿里巴巴平台上整个买家询盘数急剧下滑，欧美对中国采购在下滑。海关是卖出了货，出去以后再获得数据；而我们提前半年从询盘上推断出世界贸易发生了变化。”



- ▶ 大数据教育与信息技术教育的关系解读
- ▶ 大数据在课程标准中的设计与要求
- ▶ 教学建议与案例研讨

教育部普通高中课程方案和课程标准国家级示范培训
主办单位：教育部基础教育课程教材发展中心
承办单位：教育部基础教育课程教材发展中心
中国·北京

案例研讨：学生行为画像

中国移动 07:36 73%

返回 大数据研发“学生画像” ...

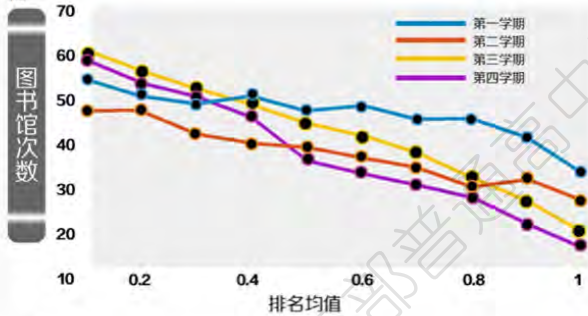


首页 > 专业百科

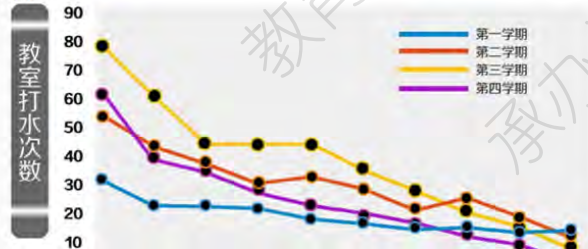
大数据研发“学生画像” 谁会成为“学霸”“学渣”可预测

华西都市报 2015年08月27日 09:27

图一



图二



中国移动 07:35 73%

返回 baijiahao.baidu.com

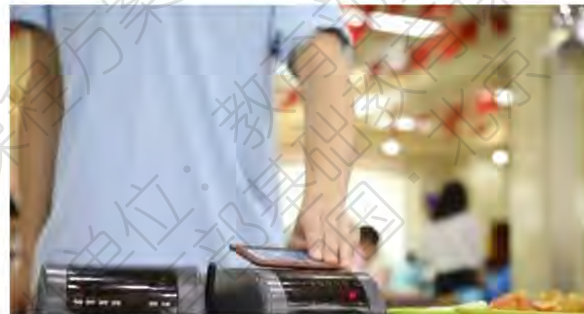
中科大大数据分析食堂消费 为真贫困学生提供补助



sweat笑笑木

百家号 17-07-13 17:18

关注



近日，知乎上的一个回答火了。一名贫困生每天在食堂吃饭不超过六块钱，一段时间后，突然收到了母校中科大的通知，让其领取生活补助360元。



返回 > 分享 收藏 打印

中国移动 07:36 73%

返回 腾讯大数据

QQ空间：2016考生行为数据大揭秘

2016-06-07 腾讯大数据

今日起，2016年全国高考正式拉开大幕，4月~6月也被称为学子“备考季”，是中高考、期末大考的集中时段。一大波考卷迎面来袭，今年考生的鸭梨是什么？新生代又会以怎样的方式进行备考？你了解新生代考生的行为习惯吗？QQ空间、QQ社交指数、腾讯大数据将联合为您解密！

QQ空间 考生社交行为白皮书 2016互联网 读懂新生代

腾讯大数据 QQ社交指数

案例研讨：个性化推荐技术



案例研讨：个性化推荐技术



案例研讨：个性化推荐技术

豆瓣：我买了俩馒头，他问我，你要不要来碗米饭？

淘宝：我吃完俩馒头，问我，你要不要来俩馒头？

百度：“老板，给我俩馒头”-“湖南株洲馒头机制造厂供应优质馒头机”

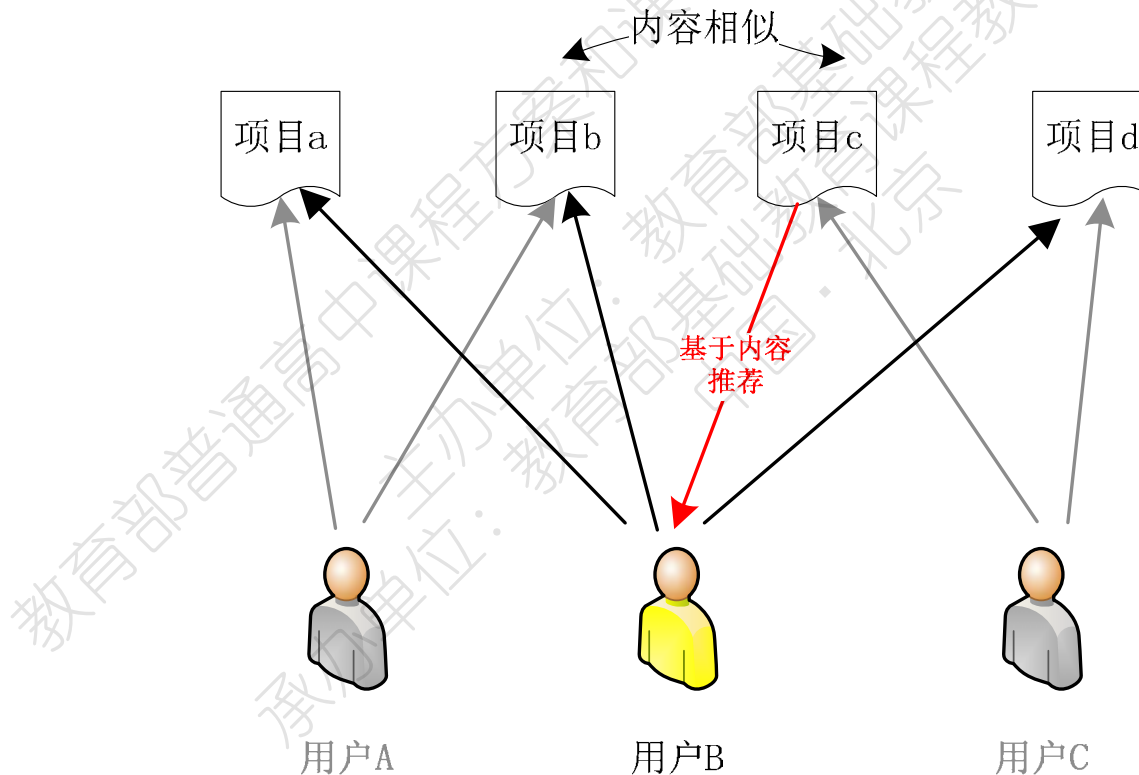
腾讯：正当我要买馒头时，在后面拍了拍我，“同学，来我这买，一模一样，还有豆沙馅”

360：让我摸一下，免费送馒头。

- 基于内容？
- 基于历史行为？
- 基于知识？

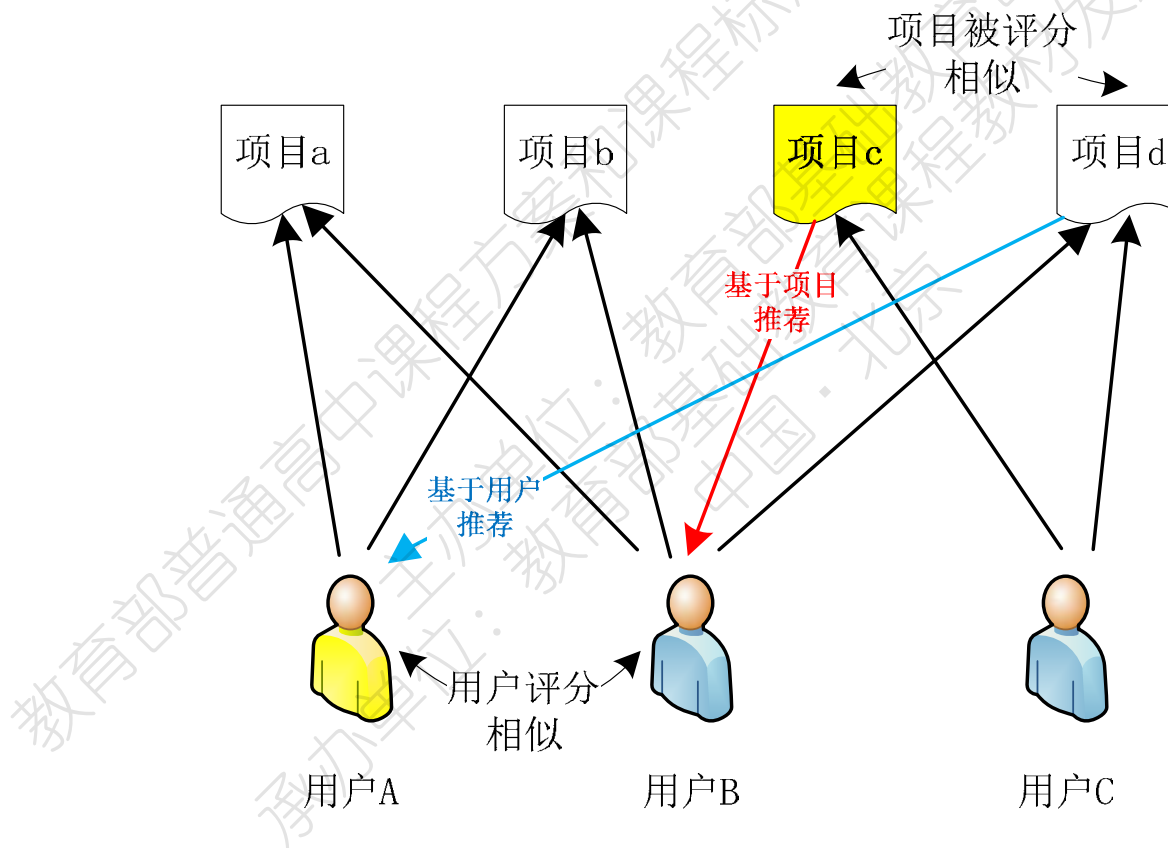
案例研讨：个性化推荐技术

- 基于内容的推荐模型：信息检索技术的延伸与发展
 - 使用从项目内容中抽取出的特征来代表各个项目；
 - 结合用户的行为历史和涉及到的项目特征进行建模。



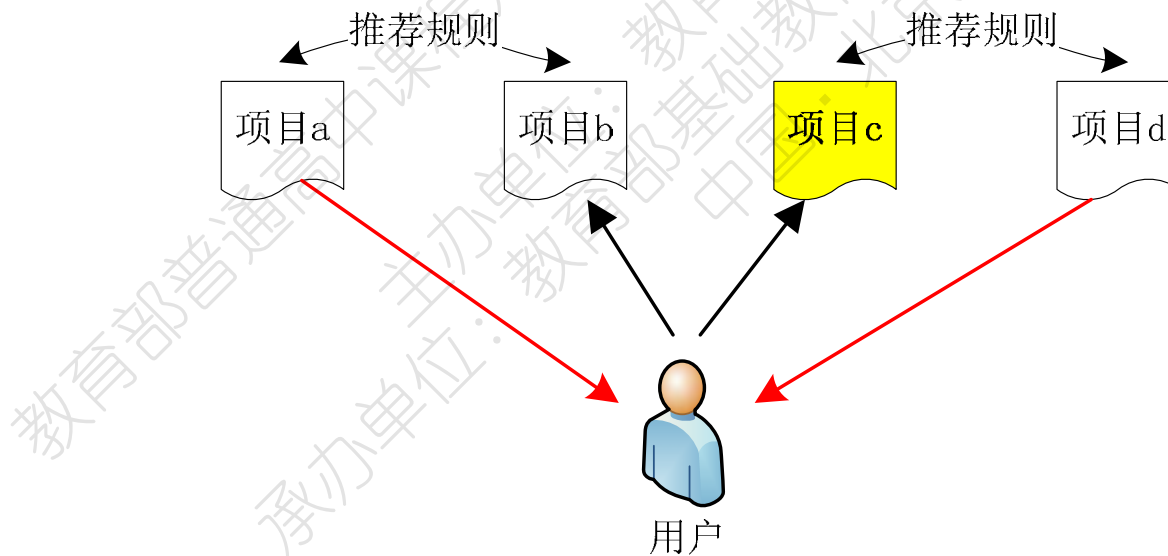
案例研讨：个性化推荐技术

- 基于协同过滤的推荐模型：相似的用户有相似的兴趣
 - User-based与Item-based

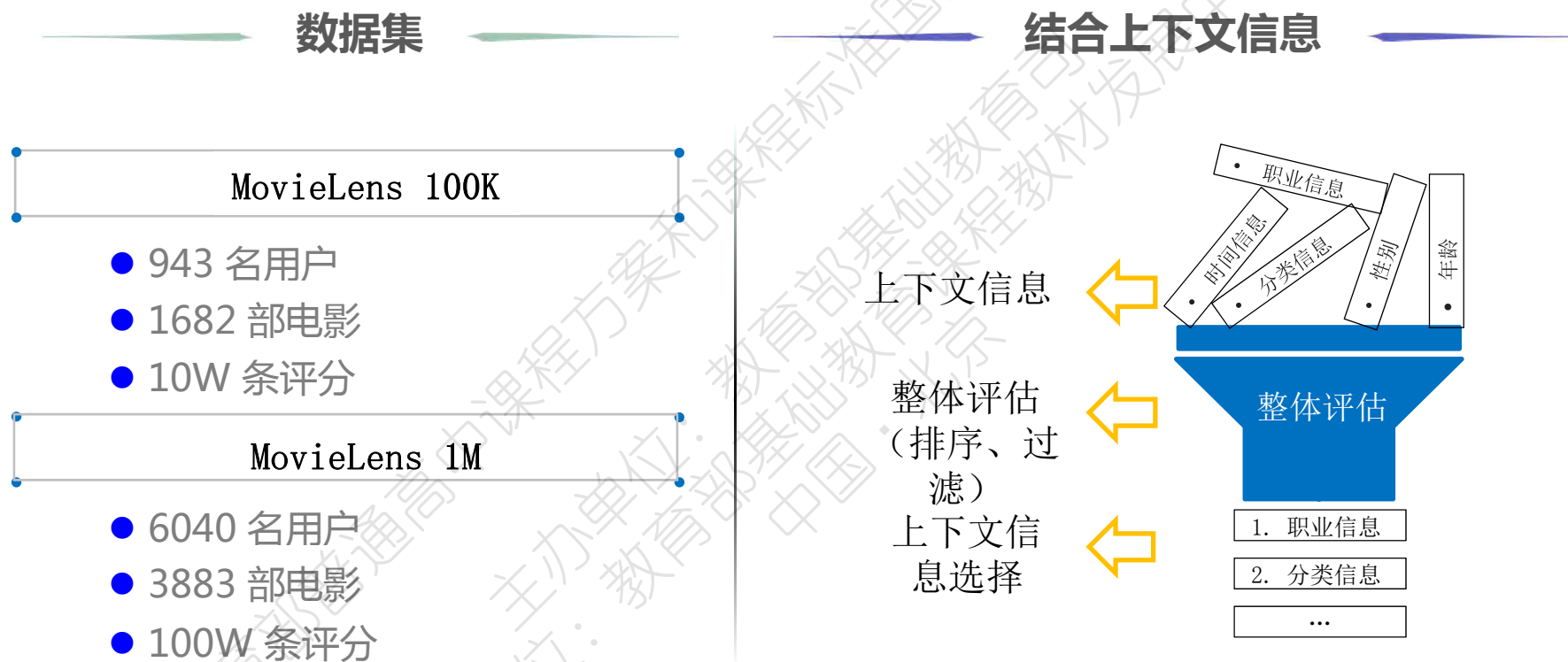


案例研讨：个性化推荐技术

- 基于知识的推荐模型：行为数据较少又有明确需求时
 - 依赖人工的先验经验，能较处理好冷启动问题
 - 利用知识库的延伸或扩展形成规则或相互关系
 - 用户需求与物品之间的相似度
 - 明确的推荐规则



混合推荐方法



教育信息技术演进历程

广播电视
自主学习
.....

D-Learning
(远程学习)
60s~80s

数字化
网络化
.....

E-Learning
(数字学习)
90s~00s

移动设备
无线网络
.....

M-Learning
(移动学习)
2005~

传感技术
情景感知
.....

U-Learning
(泛在学习)
2008~

大数据
人工智能
.....

S-Learning
(智慧学习)
2013~

让我们来做一个对比

- 2016年10月17日
- 神舟飞天瞬间，民航为其“让”出一条巨大通道
- 在神舟飞船向太空冲刺的20分钟里，如果你能够查看空管雷达，就会发现，从内蒙古西部到山东江苏交界的黄海之滨，我国民航客机规整地“让”出了一条巨大通道，而通道中央，就是神舟十一号拔地而起，造访天宫的轨迹。

让我们来做一个对比



Computer = Compute + er

计算工具
(辅助计算)



大数据智能
群体智能
跨媒体智能

.....

(辅助判断、决策)

+ Thank You

请各位批评指正！