北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

大数据时代的管理

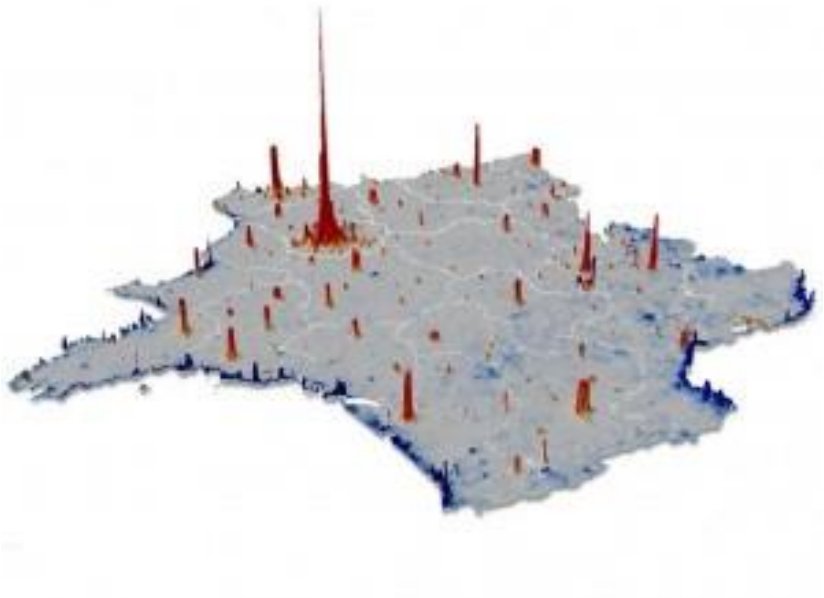
Management in Big Data Era

马宝君 博士 讲师

经济管理学院
电子商务中心
2014年11月3日

分享-手机：重塑人口普查

- 入户调查和邮寄问题的方式已经过时，比利时的研究者目前完成了通过获取手机记录完成人口调查的高质量数据。他们开发了一个数据模型，能利用手机信号塔同时识别出的上亿位手机用户的地理位置和通话记录，实时预测时当地人口密度或历时性“迁徙规律”。此外，还可以帮助政府掌握人群流动规律，以便应对埃博拉等疾病暴发和突发事件。



来源：http://www.looooker.com/archives/10782

课程停上提醒

第11周周一（11.10）晚

本课程因放假停上一次

前两次课程回顾：大数据时代的管理喻意

- **三个“融合”**
  - IT融合
  - 内外融合
  - 价值融合

- **三个“新”**
  - 新模式
  - 新业态
  - 新人群

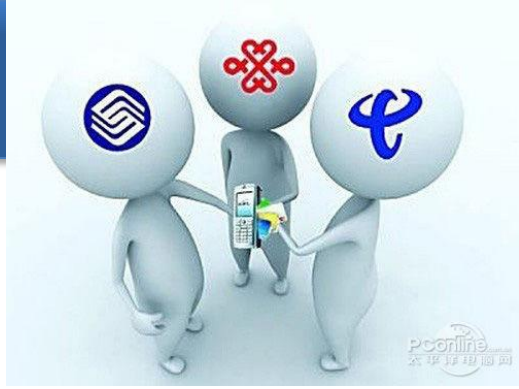
大数据分析

1. 企业核心能力

大数据中的企业：几个示例

- **Visa-提升异常交易识别**
  - 分析100%交易记录 (2005年2%)；200个属性 (2005年40个)
- **Citibank-提升客户贷款质量**
  - 分析市场、申请人状况、社交媒体记录、历史决策行为…
- **Walmart-提升网购服务质量（提高10%-15%交易量）**
  - 沃尔玛网上超市：http://www.walmart.com/
  - 分析社交媒体和产品热度，提供推荐…
- **某制造企业-提升生产率改善库存（+25%）**
  - 分析销售、生产、库存数据，识别需求及匹配的优化生产配置模式…
- **UP Medical Center-提升癌症诊断和治疗质量**
  - 分析基因顺序、肿瘤特征、分子差异、年龄因素…

例：中国移动通讯行业



- **现有模式**
  - 搭桥/修路/铺管道/建基站
  - 定价机制/容量/速度/维护/客服

- 覆盖率/规模大幅提升 =〉饱和/增速下降？
- 价值创造规模可观 =〉占比/单位效益下降？

- **其他压力**
  - 移动互联网、内容经营和信息消费的替代性和多元化服务

升级转型？

精益管理挑战

- **更好地了解顾客**
  - 客户特征和细分
  - 客户行为和粘性
  - 客户喜好和新需求
  - ……
- **更好地了解业务**
  - 业务活动轨迹
  - 产品体验与口碑
  - 业务关联与因果分析
  - ……
- **更好地了解对手和伙伴**
  - 行业动态与趋势
  - 对手优势特征
  - ……

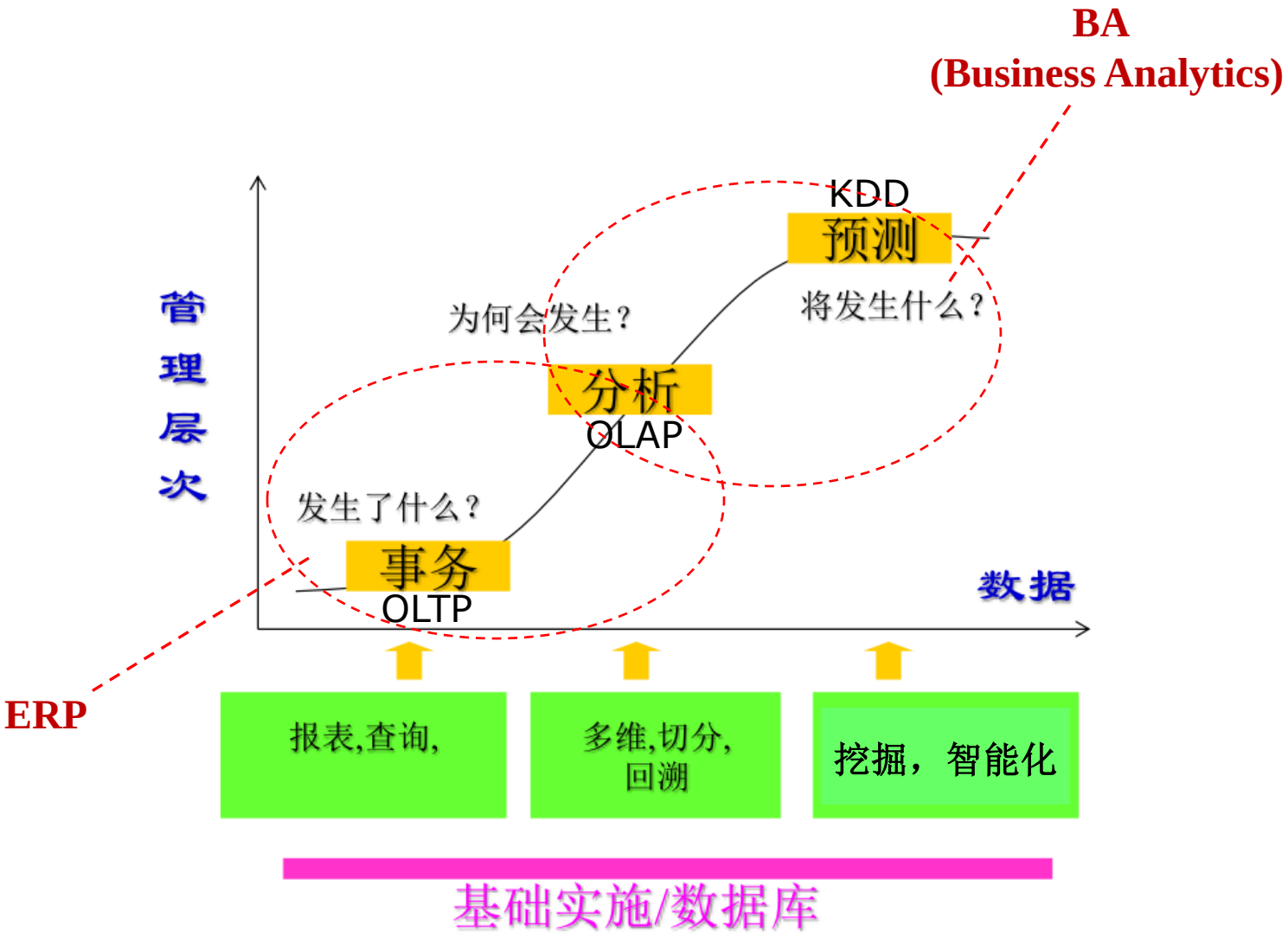
深度业务分析 (Business Analytics)

- **精准营销**
- **针对性KPI**
- **优化运营**

大数据分析

2. 深度业务分析(Business Analytics, BA)

企业数据分析的管理层次



“IT Doesn’t matter?”

- 《哈佛商业评论》(Harvard Business Review), 2003年5月期, Nicholas G. Carr
- IT是必需品，不是战略竞争优势？
  - 如：Microsoft Office、Email、Internet/Web Access？
  - CRM（Customer Relationship Management）、ERP（Enterprise Resource Planning）、BI（Business Intelligence）…？
- 论战结果：不能一概而论，新技术仍将发挥作用
  - 新的IT技术（只要没有完全成熟）仍然会持续给企业带来竞争优势的机会
  - *Unless nurtured and evolved, IT-enabled competitive applications, like many competitive advantages, don't endure*

IT Doesn’t Matter

by Nicholas G. Carr

As information technology’s power and ubiquity have grown, its strategic importance has diminished. The way you approach IT investment and management will need to change dramatically.

Letters to the Editor

Does IT Matter?
An HBR Debate

Order the article, “IT Doesn’t Matter”

E-mail us at hbr_letters@hbsp.harvard.edu

商务智能（Business Intelligence，BI）

- **商务智能**
  - 发现知识的过程
  - 从海量数据中
  - 支持管理决策

- **潜在**
- **新颖**
- **有用**

核心技术：深度业务分析（BA）

深度业务分析

（Business Analytics-BA）

Methods & Tools

深度业务分析
（Business Analytics-BA）

Methods & Tools

- 原方法
- 组合方法

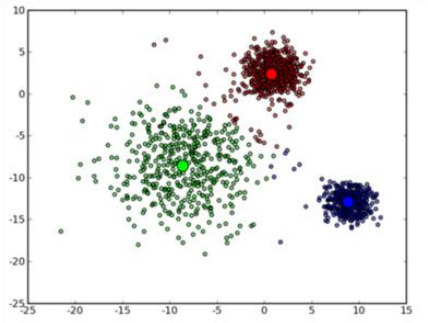
深度业务分析：原方法

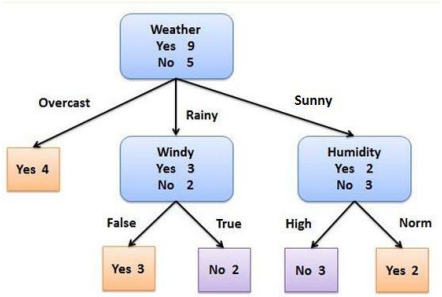
- 聚类（Clustering）
- 分类（Classification）
- 关联（Association）
- 模式（Pattern）
- ……

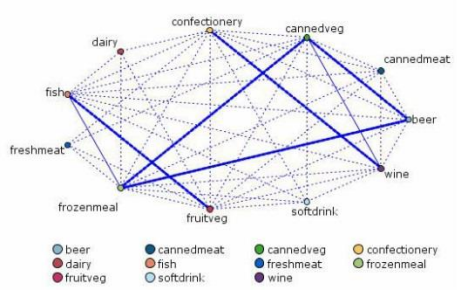
类别

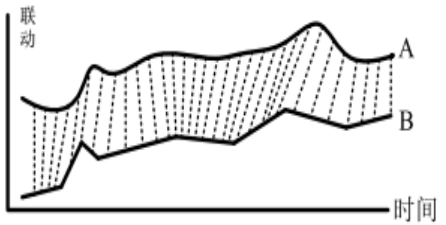
联系

轨迹





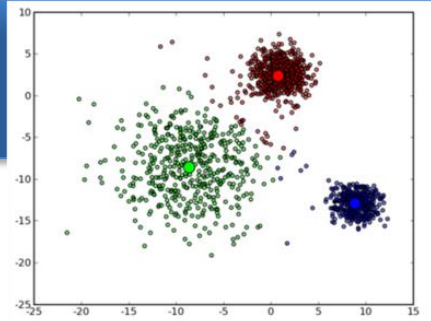




大数据深度业务分析方法

（1）聚类分析方法：Clustering

聚类分析方法（Clustering）



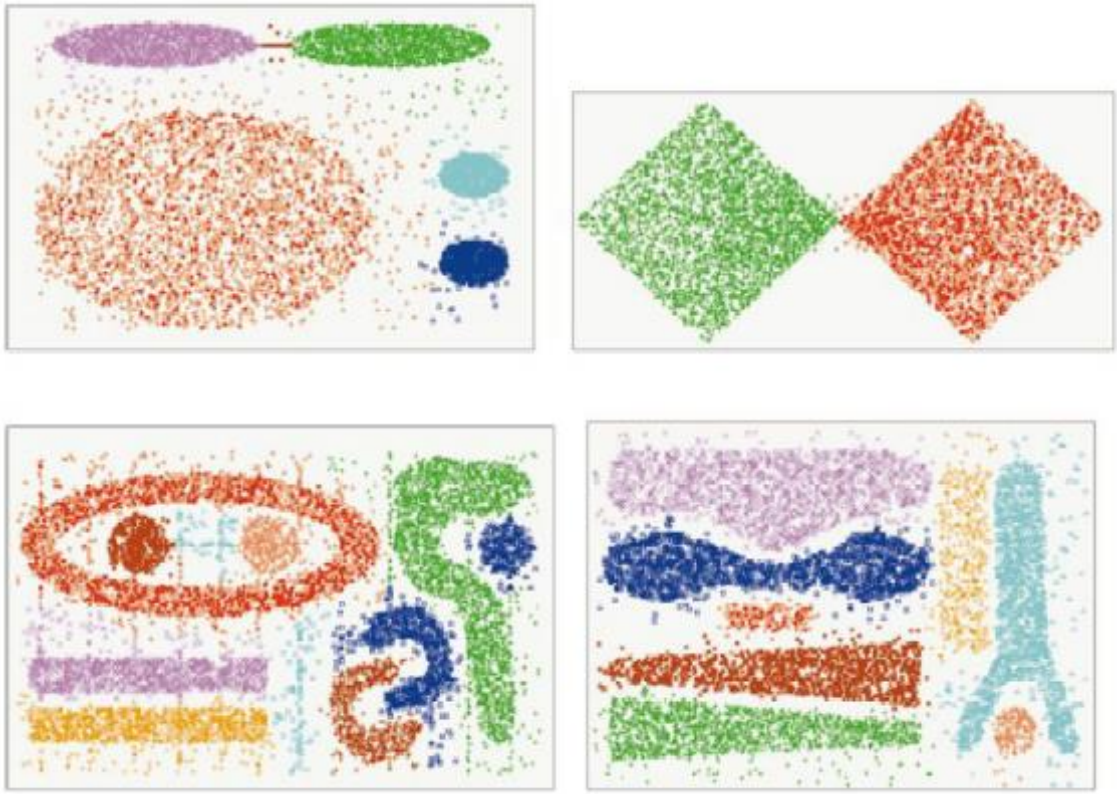
- “原来无类，聚之成类”（Cluster、Class）
- 例：某公司希望将客户按照一般特征（即数据属性）进行分组，基于这些分组信息，公司将针对不同的客户群进行有针对性的市场营销和广告活动。但该公司管理部门没有任何关于这些分组的预先定义的类别标准，他们知道客户的信息包括：
  - 收入、年龄、婚姻状况、子女数目、教育程度等（如下表所示）

收入	年龄	婚姻状况	子女数目	教育程度
$ 25000	35	已婚	3	高中
$ 15000	24	已婚	1	高中
$ 30000	21	未婚	0	高中
$ 23000	25	离异	0	高中
$ 21000	25	离异	3	大学
$ 72000	60	已婚	0	大学
$ 81000	32	已婚	0	研究生
$ 253000	43	已婚	3	研究生
$ 198000	51	离异	2	大学

第一组：高中有小孩；
第二组：高中无小孩；
第三组：大学有小孩；
第四组：收入较高、大学以上、没有小孩；
第五组：收入较高，大学以上、有小孩。

聚类标准：What is good clustering?

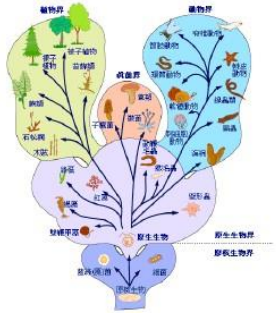
- **好的聚类方法**
  - 类内对象间较高的相似度
  - 类间对象间较低的相似度



聚类分析方法的一些应用

- **CRM中的客户分群**
  - Customer Segmentation

- **保险 Insurance**
  - Identifying groups of motor insurance policy holders with a high average claim cost 高索赔额的汽车保险投保人识别
- **城市规划 City-planning**
  - Identifying groups of houses according to their house type, value, and geographical location
- **WWW**
  - 根据 网络日志（Web log）数据发现相似的访问模式
- **生物：新物种的动植物分类（taxonomy）**



聚类方法准备：对象数据结构 & 相似度

- **数据矩阵 Data Matrix**
  - 行：对象，*n*个
  - 列：属性，*p*个

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- **相似矩阵/距离矩阵 Similarity/distance Matrix**
  - 行列都表示对象：*n*个对象
  - *d*(*i*, *j*)：对象*i*和对象*j*之间的距离/相似度

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

对象相似度度量（1）

- **基于内容的相似度（标准化后的对象向量）**
  - Distance （Minkowski Distance，明可夫斯基距离，明氏距离）

$$d(i,j)=\sqrt[q]{(|x_{i_1}-x_{j_1}|^q+|x_{i_2}-x_{j_2}|^q+...+|x_{i_p}-x_{j_p}|^q)}$$

where $i=(x_{i1}, x_{i2}, ..., x_{ip})$ and $j=(x_{j1}, x_{j2}, ..., x_{jp})$ are two p-dimensional data objects, and q is a positive integer

  - 如果q=1，d(i,j)就是曼哈顿距离（ Manhattan Distance ）

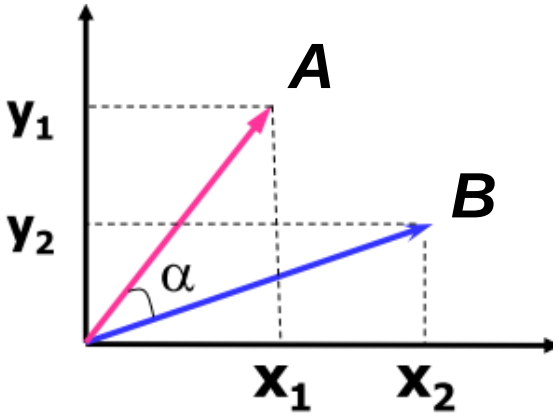
$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+...+|x_{i_p}-x_{j_p}|$$

  - 如果q=2，d(i,j)就是欧拉距离（Euclidean Distance ）

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

对象相似度度量（2）

- Cosine similarity
- A= (x_1,y_1) ,B= (x_2,y_2)



$$\cos(A,B)=\frac{x_1\times x_2+y_1\times y_2}{\sqrt{x_1^2+y_1^2}\times\sqrt{x_2^2+y_2^2}}$$

- $i=(x_{i1},x_{i2},\ldots,x_{ip}),\ j=(x_{j1},x_{j2},\ldots,x_{jp})$

$$\cos(i,j)=\frac{\sum_{k=1}^{p}x_{ik}\times x_{jk}}{\sqrt{\sum_{k=1}^{p}x_{ik}^2}\times\sqrt{\sum_{k=1}^{p}x_{jk}^2}}$$

聚类方法的种类

- **划分法 Partitioning approach**
  - 构建分区：K-means，k-medoids, CLARANS
- **层次法 Hierarchical approach**
  - 分层分解：Diana, Agnes, BIRCH, ROCK, CAMELEON
- **基于密度的方法 Density-based approach**
  - 基于连接性和密度函数: DBSCAN, OPTICS, DenClue
- **基于模型的方法 Model-based approach**
  - 根据假设为每个类构建一个模型： SOM, EM, COBWEB
- **基于频繁模式法 Frequent pattern-based approach**
  - 基于频繁模式的分析: pCluster
  - 多层次粒度结构: STING, WaveCluster, CLIQUE
- **……**

数据挖掘/商务智能分析方法 参考资料

- 韩家炜, Micheline Kamber, 裴健著, 范明, 孟小峰译. 数据挖掘:概念与技术(原书第3版),机械工业出版社, 2012.
- 威滕(新西兰)等著, 董琳等译. 数据挖掘：实用机器学习技术(第2版), 机械工业出版社, 2006.
- 刘红岩. 商务智能方法与应用, 清华大学出版社, 2013.
- 陈国青, 卫强, 张瑾. 商务智能原理与方法（第2版）, 电子工业出版社, 2014.

- http://baike.baidu.com/view/7893.htm?fr=aladdin
- http://baike.baidu.com/view/903740.htm