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17日

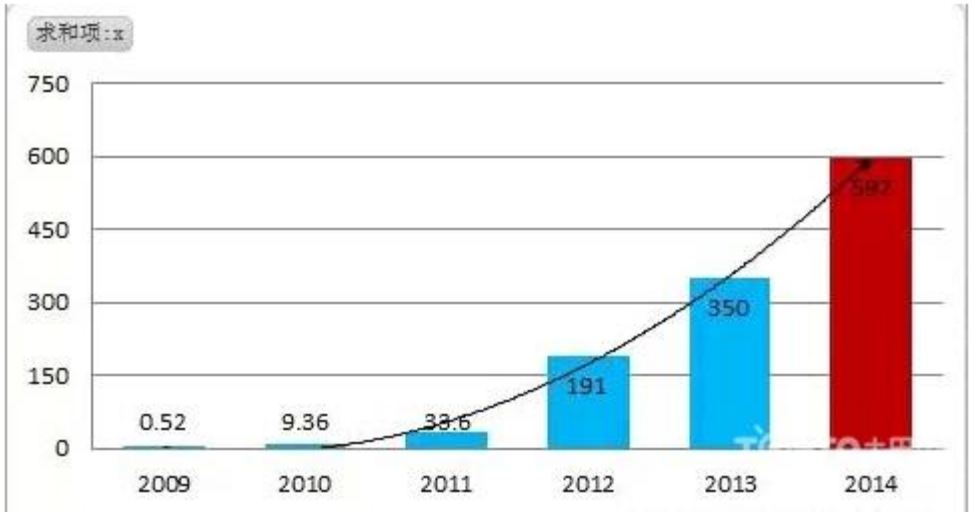
12月15日上课时间调整？

2014 年 12 月				December		
星期一	星期二	星期三	星期四	星期五	星期六	星期日
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15 18:30-20:20 此次课程不上	16	17	18	19 改到周五 18:30-20:20	20	21
22	23	24	25	26	27	28
29	30	31				

课前分享1：拼数据的双十一

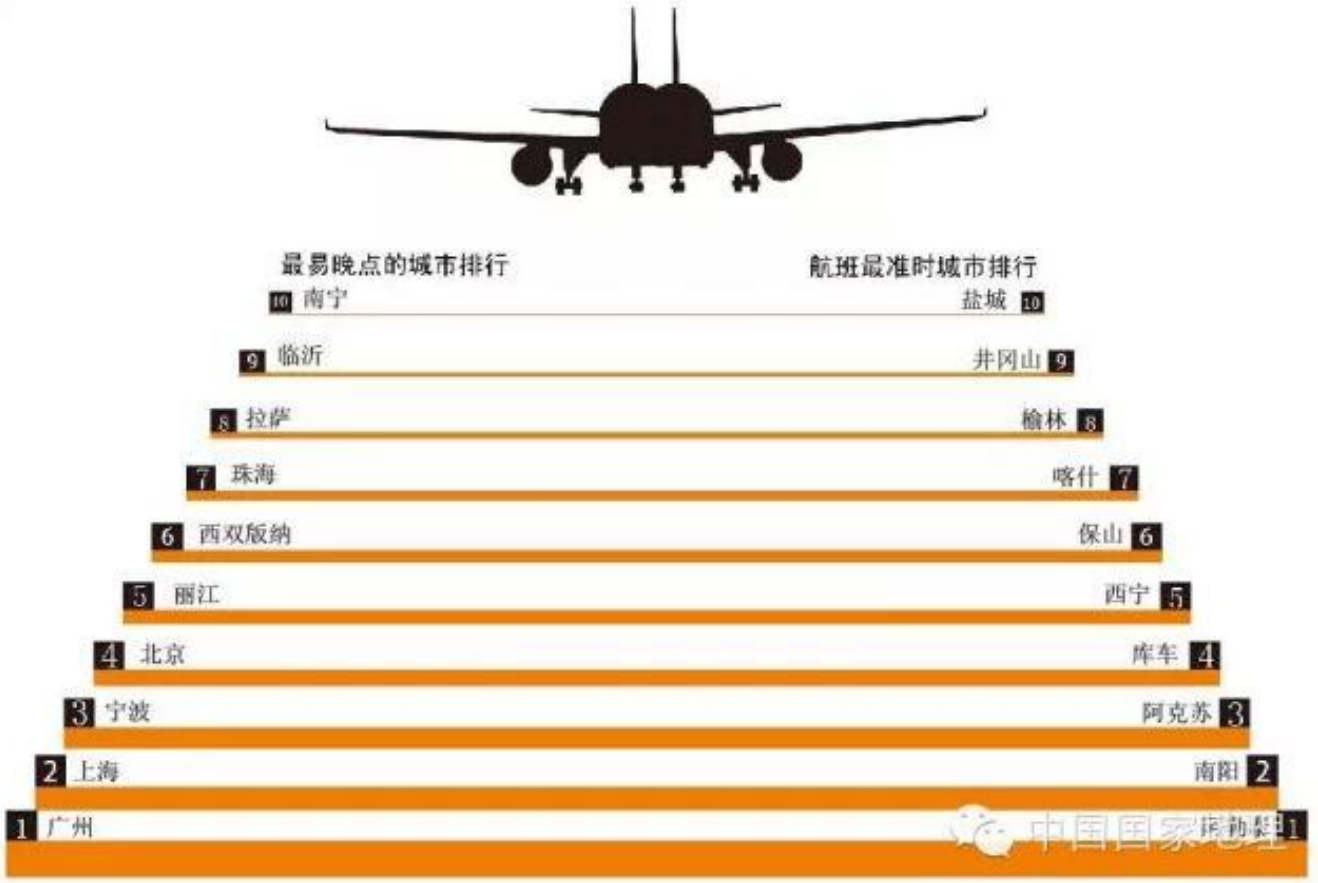
天猫双11成交额播报

2014.11.11		湖畔会	2013.11.11	
3分钟	10亿		55秒	1亿
5分钟	20亿		6分7秒	10亿
14分钟	50亿		13分22秒	20亿
38分钟	100亿		38分	50亿
2时04分	155亿		2时08分	90亿
12时58分	350亿		13时39分	200亿
21时13分	500亿		17时31分	300亿
24时	571亿		24时	350亿



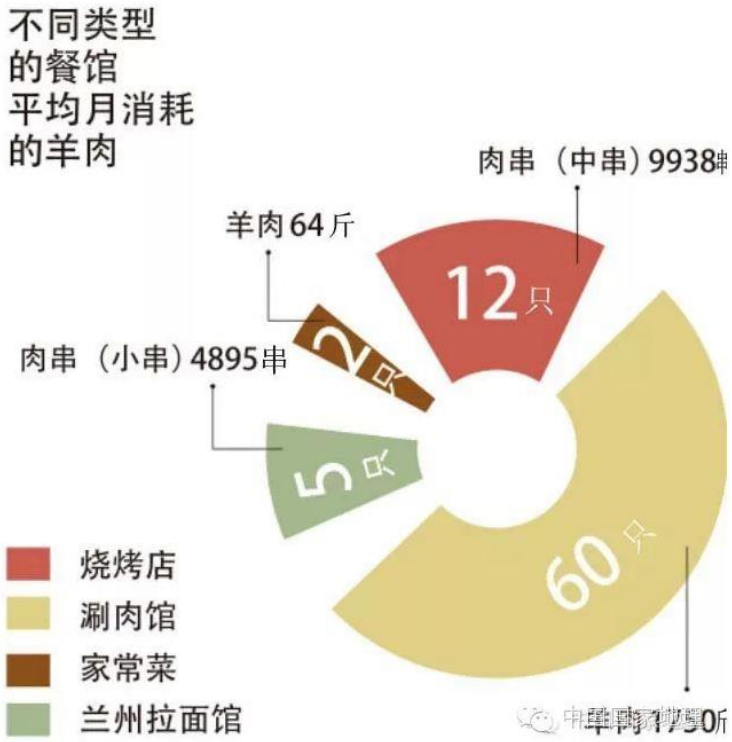
2014年双十一消费排行榜出炉

课前分享2：大数据让“马云们”知道了太多的秘密



2012年2月底，淘宝网根据“淘宝旅行”推出的航班延误险的销售与赔偿状况，以春运的“1月8日—2月16日”为统计周期，整理出了航班延误的城市排行榜。数据显示：广州是这一时间段飞机延误最常发生的城市，而飞向新疆阿勒泰的航班最准时靠谱。

北京餐馆每月消耗多少只羊？



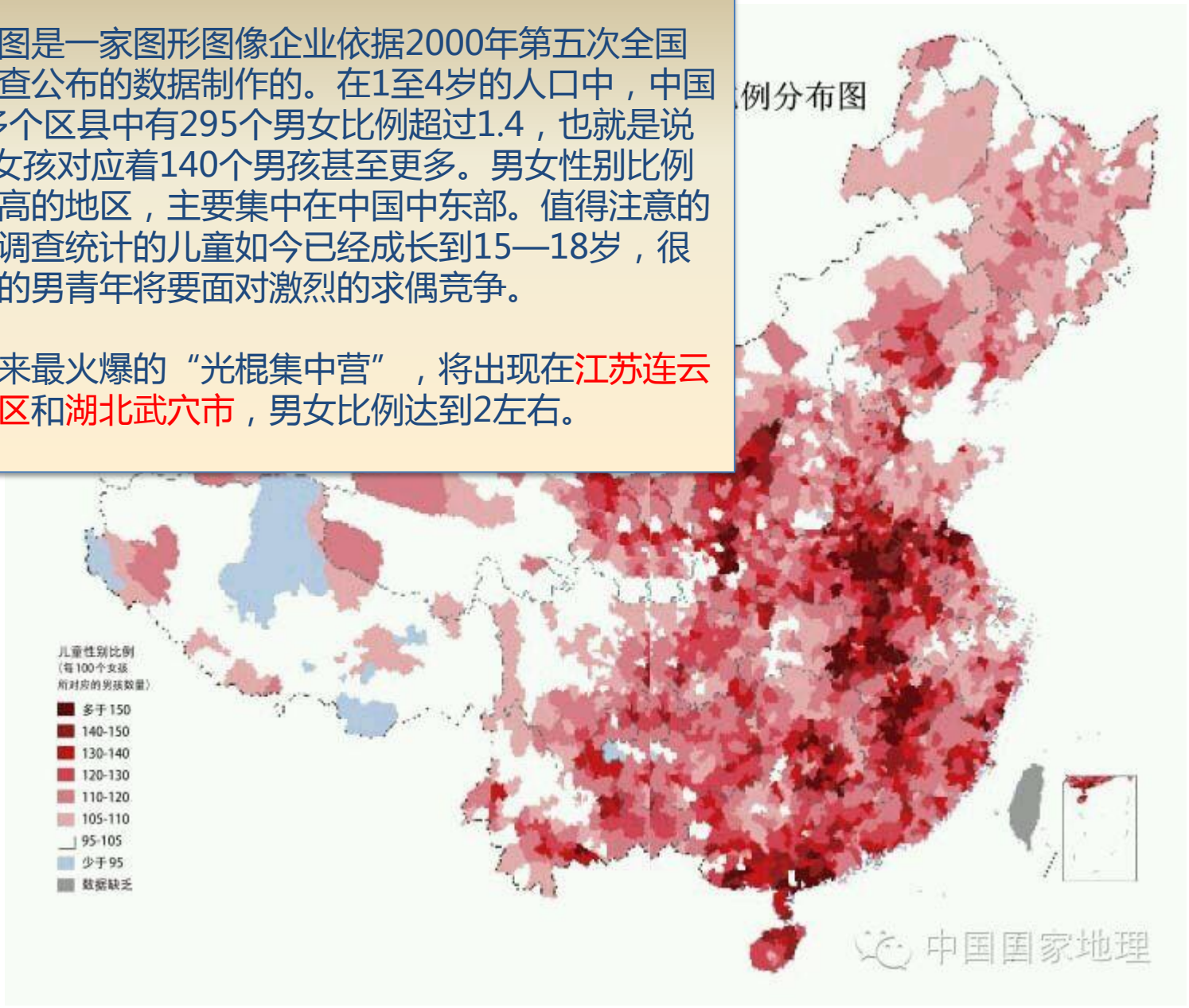
这是一组根据北京不同餐馆中无线点餐系统汇集的数据，其中一些内容颇为有趣：

- 在北京不同的商圈，以学生为主体的学院路—五道口地区，快餐店生意最好；
- 在国贸CBD商务区，川菜颇受欢迎；
- 而在丰台等地的传统居民区，火锅店最为热闹
- 在北京的烧烤店中，除了肉串之外，男性顾客最爱点的食物是烤腰子和板筋，女性顾客最青睐的则是烤虾和骨肉相连；
- 而在咖啡店中，卡布奇诺咖啡销售状况最好。

中国未来最火爆的“光棍集中营”在哪里？

- 这张地图是一家图形图像企业依据2000年第五次全国人口普查公布的数据制作的。在1至4岁的人口中，中国2800多个区县中有295个男女比例超过1.4，也就是说100个女孩对应着140个男孩甚至更多。男女性别比例严重偏高的地区，主要集中在中国中东部。值得注意的是，被调查统计的儿童如今已经成长到15—18岁，很多地区的男青年将要面对激烈的求偶竞争。

- 中国未来最火爆的“光棍集中营”，将出现在江苏连云港赣榆区和湖北武穴市，男女比例达到2左右。

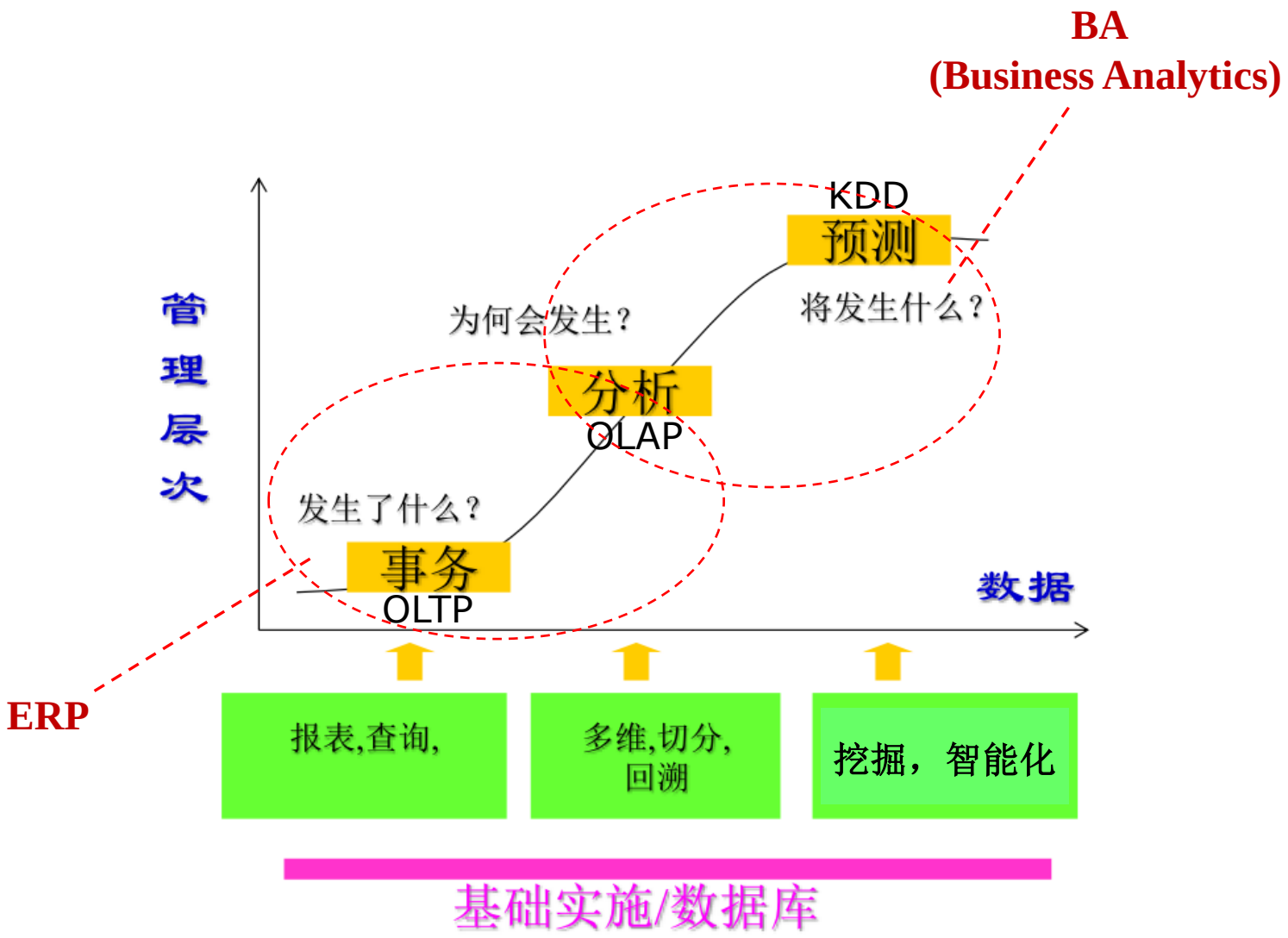


数据挖掘/商务智能分析方法 参考资料

- 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

课程回顾1：企业数据分析的管理层次



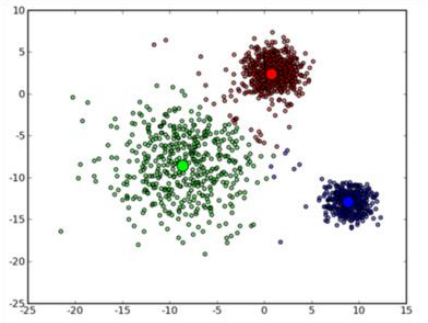
课程回顾2：深度业务分析——原方法

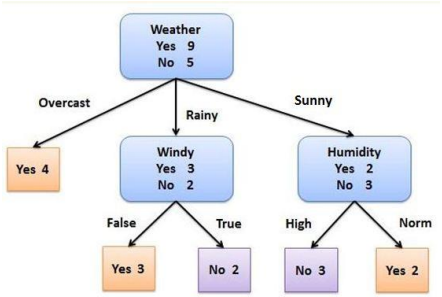
- 聚类（Clustering）
- 分类（Classification）
- 关联（Association）
- 模式（Pattern）
- ……

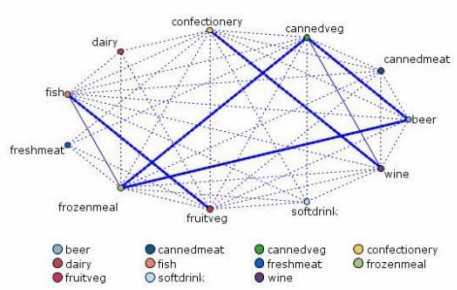
类别

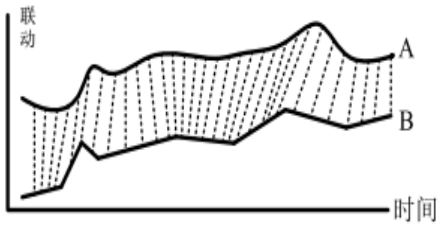
联系

轨迹







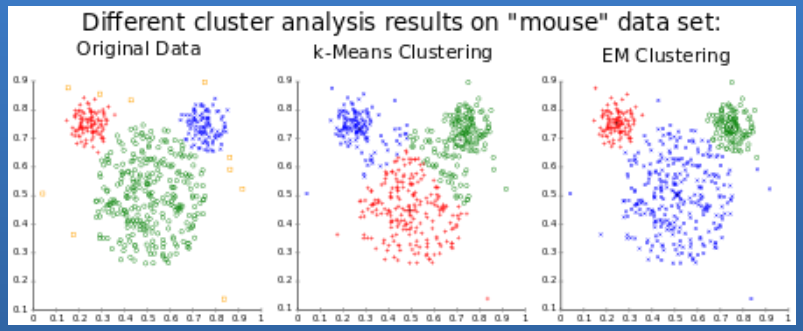


课程回顾3：聚类方法的种类

- **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
- **……**

聚类分析方法

划分式聚类方法的代表：K-means

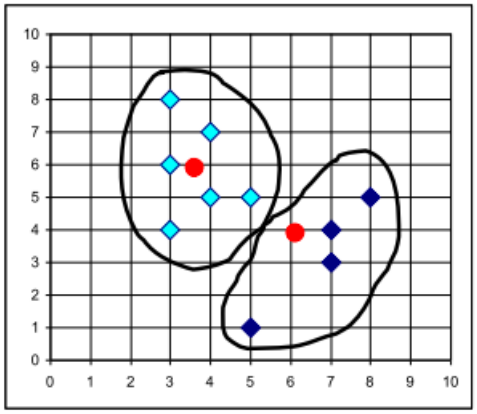


划分聚类方法（Partitioned clustering method）

- **目标：将数据库*D*中的*n*个对象（向量、元组）划分到*k*个类$C_1, C_2, \ldots, C_k$中，使得如下表达式表达的距离平方和最小：**

$$\sum_{m=1}^{k}\sum_{t_{mi}\in C_m}[d(C_m, t_{mi})]^2$$

- **给定*k*，找到一个最优的划分**
  - 全局最优：穷尽所有的分区



典型代表方法：K-means

- **划分式聚类方法中最简单、应用最广**
  - 1967年由MacQueen提出
  - 给定类别数：*k*

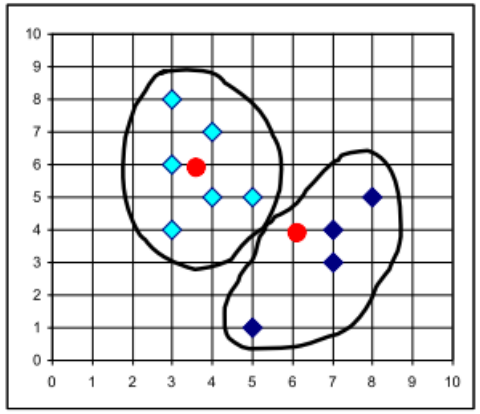
- **基本思路和步骤**
  - 1. 随机或按某种策略从*n*个对象中选择*k*个对象作为初始的类中心（Centriod，Mean Point）；
  - 2. 计算每个对象与这*k*个类中心的距离；
  - 3. 将每个对象划分/分配到与其距离最近的类中心所在的类中；
  - 4. 回到第2步，直到和前一次划分/分配结果无差异，停止。

Centroid of a Cluster

- Centroid: the "middle" of a cluster

$$Centroid_m = \frac{\sum_{i=1}^{N} t_{ip}}{N}$$

- $t_{1p}=(1,5)$
- $t_{2p}=(3,4)$
- $t_{3p}=(2,6)$

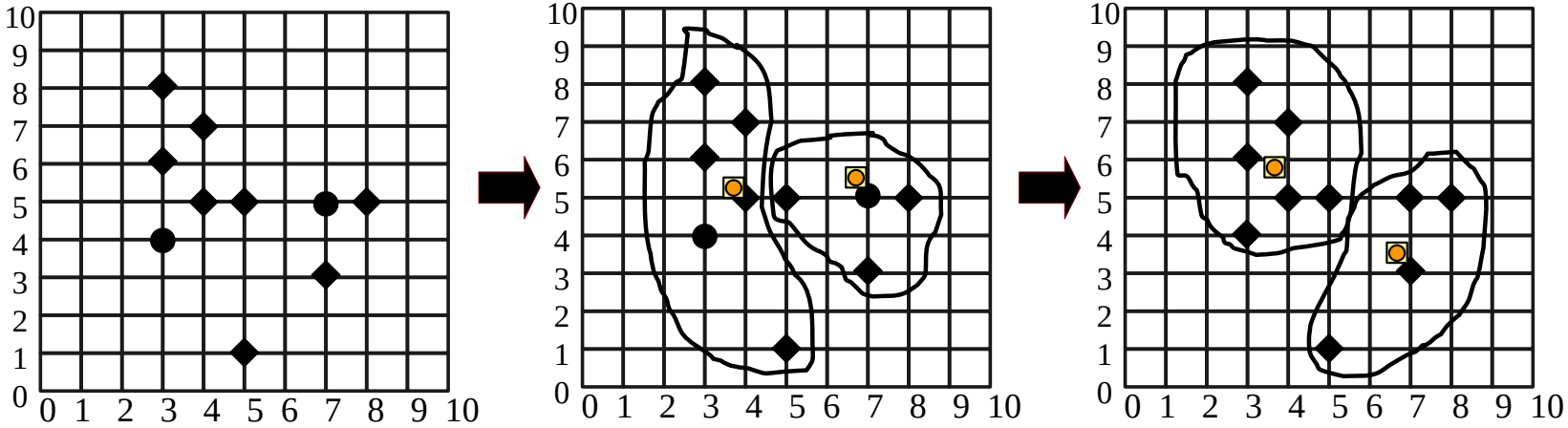


❖ Centroid:

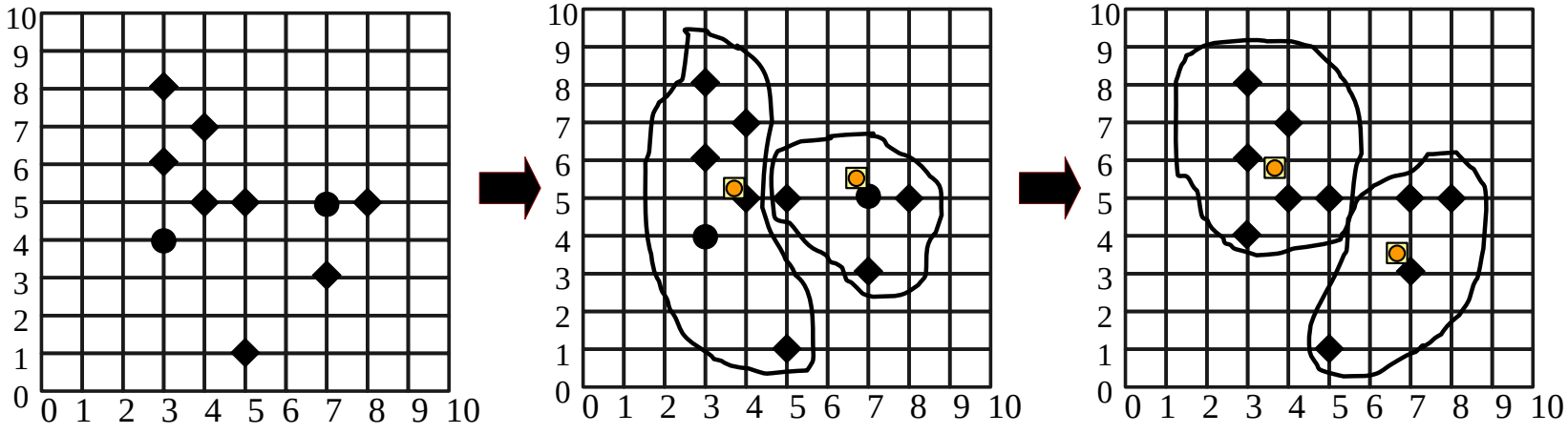
((1+3+2)/3, (5+4+6)/3) 即c=(2,5)

K-means实例

序号	坐标	序号	坐标
1	(3, 4)	**6**	(5, 1)
2	(3, 6)	**7**	(5, 5)
3	(3, 8)	**8**	(7, 3)
4	(4, 5)	**9**	(7, 5)
5	(4, 7)	**10**	(8, 5)



K-means实例（续）



（1）　（2）　（3）

序号	坐标	到 K_1 类中心(3, 4)的距离	到 K_2 类中心(7, 5)的距离	所属聚类
1	(3, 4)	0	$\sqrt{17}$	K_1
2	(3, 6)	2	$\sqrt{17}$	K_1
3	(3, 8)	4	5	K_1
4	(4, 5)	$\sqrt{2}$	3	K_1
5	(4, 7)	$\sqrt{10}$	$\sqrt{13}$	K_1
6	(5, 1)	$\sqrt{13}$	$\sqrt{20}$	K_1
7	(5, 5)	$\sqrt{5}$	2	K_2
8	(7, 3)	$\sqrt{17}$	2	K_2
9	(7, 5)	$\sqrt{17}$	0	K_2
10	(8, 5)	$\sqrt{26}$	1	K_2

序号	坐标	到 K_1 类中心 $(3\frac{2}{3},5\frac{1}{6})$ 的距离	到 K_2 类中心 (6.75,4.5) 的距离	所属聚类
1	(3, 4)	1.344	3.816	K_1
2	(3, 6)	1.067	4.039	K_1
3	(3, 8)	2.911	5.130	K_1
4	(4, 5)	0.373	2.795	K_1
5	(4, 7)	1.863	3.717	K_1
6	(5, 1)	4.375	3.913	K_2
7	(5, 5)	1.344	1.820	K_1
8	(7, 3)	3.976	1.521	K_2
9	(7, 5)	3.337	0.559	K_2
10	(8, 5)	4.337	1.346	K_2

序号	坐标	到 K_1 重心 $(3\frac{2}{3},5\frac{5}{6})$ 的距离	到 K_2 重心 (6.75,3.5) 的距离	所属聚类
1	(3, 4)	1.951	3.783	K_1
2	(3, 6)	0.687	4.507	K_1
3	(3, 8)	2.267	5.858	K_1
4	(4, 5)	0.898	3.132	K_1
5	(4, 7)	1.213	4.451	K_1
6	(5, 1)	5.014	3.052	K_2
7	(5, 5)	1.572	2.305	K_1
8	(7, 3)	4.375	0.559	K_2
9	(7, 5)	3.436	1.521	K_2
10	(8, 5)	4.413	1.953	K_2

K-means方法的评价

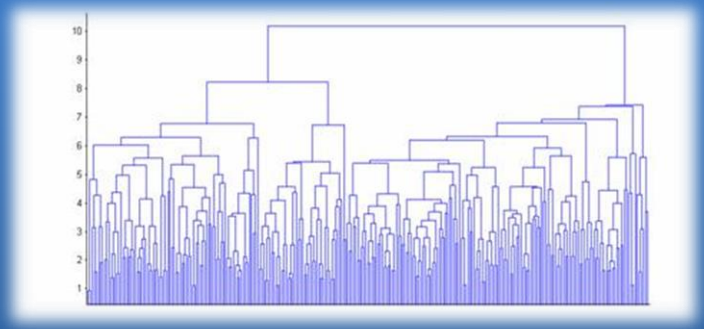
- **优点**
  - 算法简单、快速、应用广泛
  - 计算复杂度与对象个数n、类别数目k成正比，处理大规模数据时相对有效
  - 当结果聚类是密集的，且类别与类别之间的区别明显时，聚类效果较好

K-modes clustering
（1998）

- **缺陷和不足**
  - 需要用户事先指定类别个数k
  - 聚类结果对初始选择的类中心数据点较为敏感：对于不同的初始值，可能会导致不同的聚类结果
  - 由于需要计算聚类对象的均值，所以只适用于聚类均值有意义的情况
  - 不适合用于发现非凸形状的聚类，或具有各种不同大小的聚类
  - 对噪声和孤立点数据也很敏感，因为这类数据可能会影响到各聚类的均值

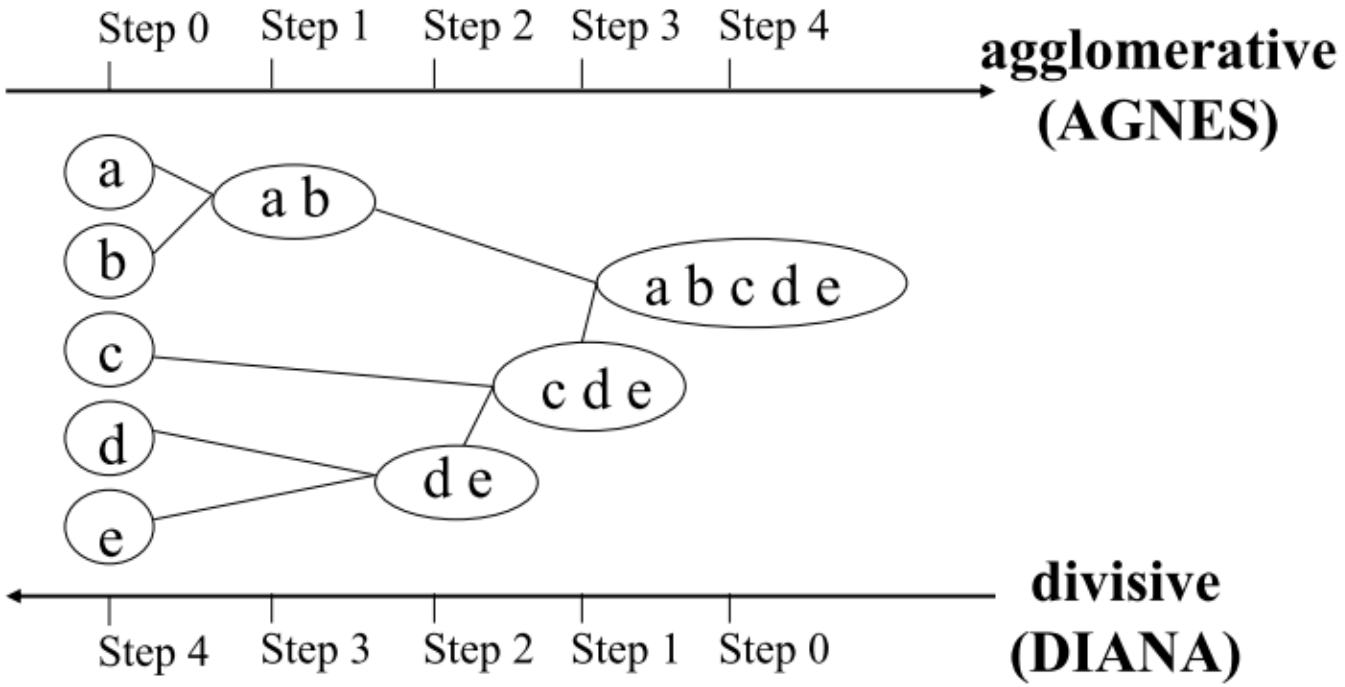
聚类分析方法

层次聚类方法（Hierarchical Clustering）



层次聚类方法（Hierarchical clustering method）

- 层次聚类方法是将数据对象分为若干组并形成一个组的树形结构来进行聚类的，不需要事先给定类别数量*k*，但需要聚类终止条件
- 可以分为自下而上的聚合(agglomerative)和自上而下的分解(divisive)的层次聚类两种



层次聚类的核心：类别之间的距离衡量

- 类间最小距离

$$d_{\min}\left(C_i, C_j\right) = \min_{p,q}\left|t_{ip}, t_{jq}\right|$$

- 类间最大距离

$$d_{\max}\left(C_i, C_j\right) = \max_{p,q}\left|t_{ip}, t_{jq}\right|$$

- 类均值之间的距离

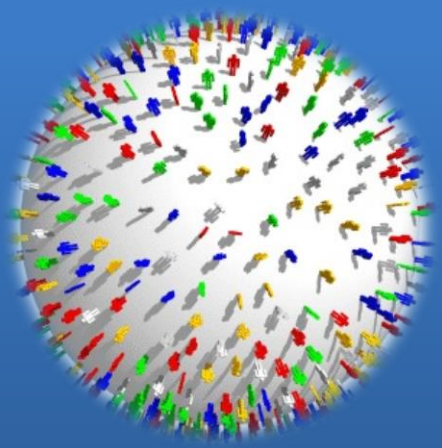
$$d_{\text{mean}}\left(C_i, C_j\right) = \left|centroid\left(C_i\right) - centroid\left(C_j\right)\right|$$

- 类间对象的平均距离

$$d_{\text{avg}}\left(C_i, C_j\right) = \text{avg}_{p,q}\left|t_{ip}, t_{jq}\right|$$

聚类分析方法

聚类分析小案例：客户细分

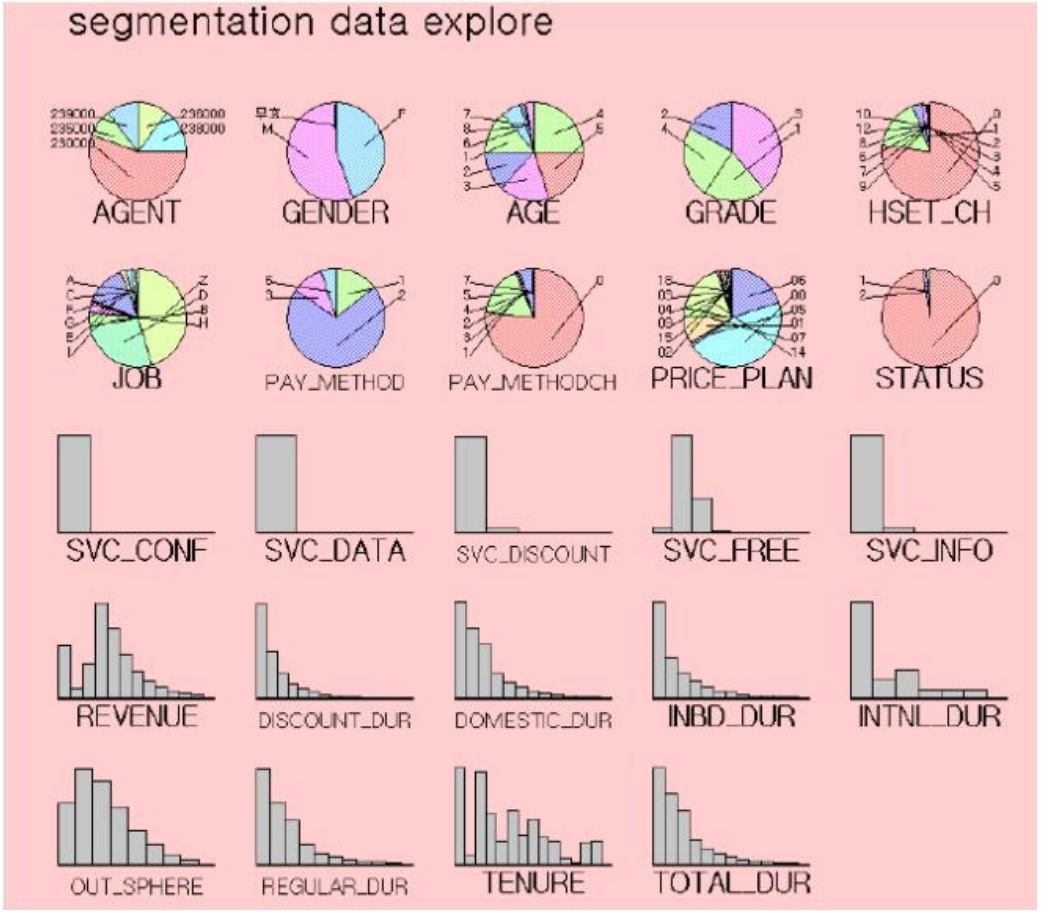


客户细分（Customer Segmentation）：数据描述

	Variable name	Description
	CALLING BEHAVIOR	
1	Regular_dur	Minutes of call in regular time frame
2	Discount_dur	Minutes of call in discount time frame
3	Night_dur	Minutes of call in night time frame
4	Domestic_dur	Minutes of domestic call
5	Intnl_dur	Minutes of international call
6	Total_dur	Minutes of total call
	SERVICEBEHAVIOR	
7	Svc_call	Number of call forwarding, call waiting type services
8	Svc_conf	Number of conference call type services
9	Svc_data	Number of internet services
10	Svc_discount	Number of friend-and family type services
11	Svc_info	Number of information services
14	Svc_free	Number of free services
15	Svc_nonfree	Number of charged services
	DEMOGRAPHIC DATA	
16	Age	Customer age
17	Gender	Customer gender
18	Job	Customer job
	ADDITIONAL DATA	
19	Agent	Agent office where the phone was initially bought
20	Hset_type	Type of handset
21	Hset_ch	Number of times customer upgrade the handset
22	Price_plan	Price plan
23	Pay_method	Payment method
24	Pay_methodch	Number of times customer change payment method
25	Inbd_dur	Minutes of inbound calls
26	Out_sphere	Number of different outbound telephone numbers
27	Grade	Customer grade
28	Status	Status of current contract
29	Revenue	Revenue
30	Tenure	Account age (how many month customer with your company)

数据的预处理

- **可视化观察**
  - 发现异常的分布
- **空缺数据的处理**
- **相关变量的提取**
- **变量的离散化**



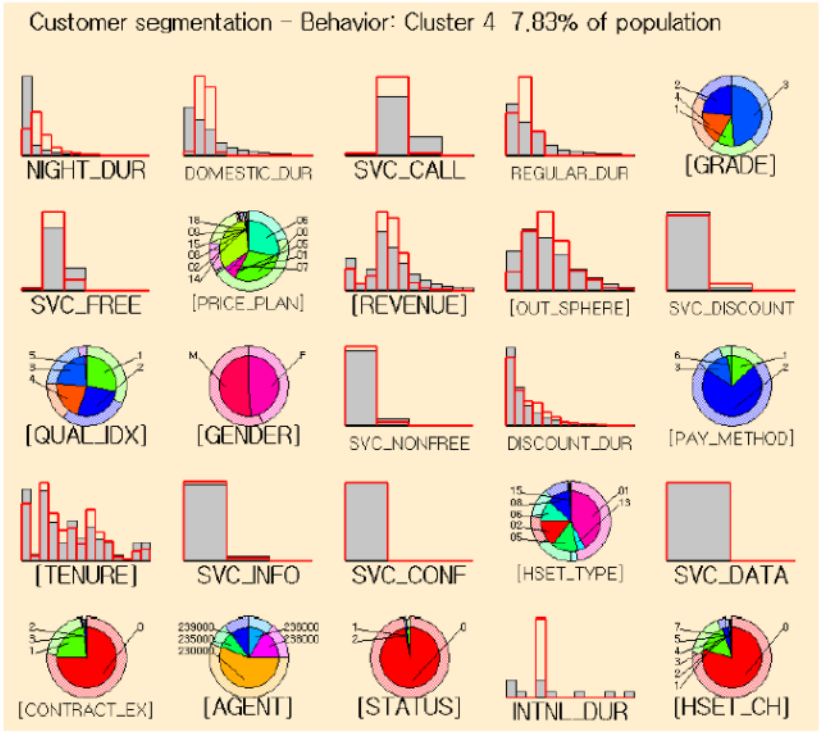
发现主要的客户类别群体

- **选择聚类算法划分客户群体**
- **对单个群体作具体分析**
  - 注意自身特征与整体特征区别大的聚类
  - 占整体百分比小的聚类

- **例如**
  - **拥有高额呼叫但少量服务的客户 (Segment 8)**
  - **拥有少量呼叫但多种服务的客户(Segment 5)**
  - **在某些时段具有高额呼叫的客户(Segment 4)**

类别4：“午夜话友”

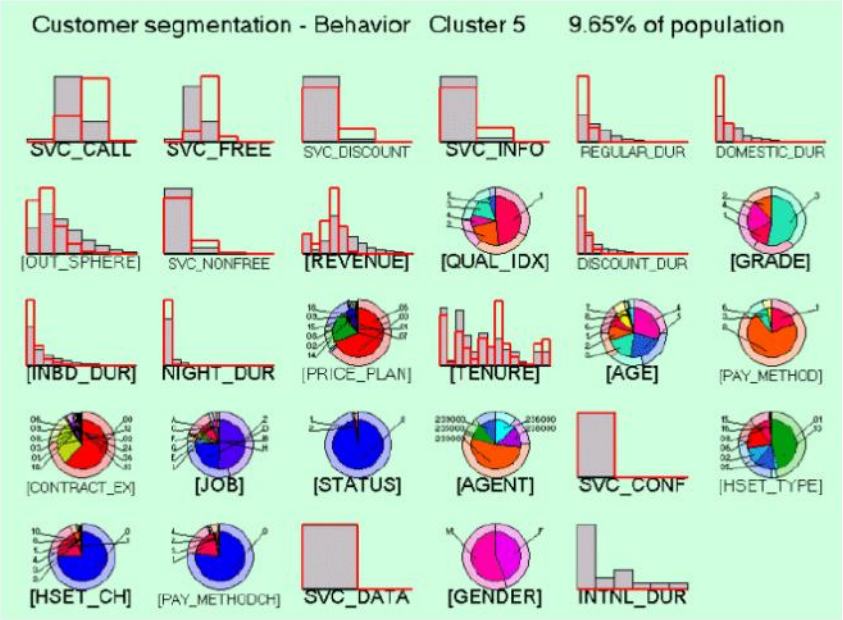
- 绝大部分通话在夜间(NIGHT_DUR)
- 除了朋友和家庭优惠(SVC_DISCOUNT)外，对其它服务不感兴趣(SVC_CALL,SVC_FREE)
- 向外拨出多个不同的号码(OUTSPHERE)：大多数人每月有30~40不同号码拨出
- 在折扣时段通话时长明显增加(DISCOUNT_DUR)
- 被叫通话时长更长(INBD_DUR)
- 青少年学生(AGE, JOB)



总人数 7.83%

类别5：“面向服务”

- 除电话会议外，他们对所有服务都很感兴趣 (SVC_CALL,SVC_FREE, SVS_NONFREE 及 SVC_CONF)
- 在所有的时段他们的通话量都很小 (REGULAR_DUR,DOMESTIC_DUR)
- 他们向外拨打电话的号码较少 (OUTSPHERE)
- 他们在本移动公司有很长时间 (TENURE)
- 他们年龄在 40~50之间 (AGE), 小业主居多(JOB)



总人数 9.65%

客户分类总结

- 根据业务特征为每个聚类命名

Segment	Segment name	Relative size	Call behavior	Service behavior	Revenue
8	Basic	18.1%	VeryLow	Low	10.6
7	Economic	13.6%	Low	Low	16.0
0	Premium-Young	12.0%	VeryHigh	Medium	41.8
6	Conservative	11.9%	Medium	Low	21.3
3	Inbound	11.1%	High	Low	31.1
5	Service-oriented	9.7%	Low	High	15.0
1	True mobile	8.5%	VeryHigh	High	37.7
4	Night friends	7.8%	Medium	Low	20.2
2	Regular teen	7.5%	Medium	High	21.8

人数和收入的排序

Rank	Segment size	Revenue share
1	**Segment 8 (18.1%)**	**Segment 0 (21.8%)**
2	**Segment 7 (13.6%)**	**Segment 3 (15.0%)**
3	**Segment 0 (12.0%)**	**Segment 1 (13.9%)**
4	Segment 6 (11.9%)	Segment 6 (11.1%)
5	Segment 3 (11.1%)	Segment 7 (9.5%)
6	Segment 5 (9.7%)	Segment 8 (8.4%)
7	Segment 1 (8.5%)	Segment 2 (7.1%)
8	Segment 4 (7.8%)	Segment 4 (6.9%)
9	Segment 2 (7.5%)	Segment 5 (6.3%)

将聚类结果部署到业务应用中

- **将聚类标识赋给任意一个客户**
  - 根据分类特征形成公式
- **客户类别部署到业务应用**
  - 呼叫中心：分类用户区别对待
  - 市场分析：客户聚类结果作为“维”
  - 驱动个性化服务机制类
- **客户聚类部署到市场竞争过程**
  - 确定目标客户群
  - 确定竞争策略
  - 实施竞争手段并评估结果

本次课程内容小结

- 数据挖掘/商务智能分析方法 参考资料
- 划分式聚类方法的代表：K-means
- 层次聚类方法（Hierarchical Clustering）
- 聚类分析小案例：客户细分

期末课程论文说明

- **主题要求**
  - 必须与“大数据管理”相关
  - 建议围绕所学专业背景下的“大数据管理问题”展开
- **内容要求**
  - 不少于4000字，版式：word中正文小四字体，1.5倍行距
  - 独立完成，不得大段拷贝或直接引用网上、书上及他人已发布内容，需要适当引用时请在引用位置注明参考文献来源（查重）
  - 论文内容框架（建议）：
    - 1. 学习本课程的心得体会、感受，对本课程教学的建议和意见（必有）
    - 2. 论文背景介绍
    - 3. 论文涉及的大数据问题及管理需求、策略和意义（可举实例说明）
    - 4. 本人对该大数据问题的看法、观点及讨论
    - 5. 总结
    - 6. 参考文献和资料

期末课程论文说明（续）

- **论文提交要求**
  - 需要以电子版提交，建议提交word版本
  - 作业提交邮箱：bigdata_homework@163.com
  - 作业提交截止时间：**第19周周日（2015.01.11）24时**
- **其他说明**
  - 请在截止时间之前提交论文（不要在截止时间附近，以避免系统原因过期），过期将不再接收论文提交，成绩为0，请务必注意；
  - 每次提交论文后，作业邮箱都会有“已收到邮件”的自动回复，如未收到自动回复，表示发送不成功，请在截止时间内重新提交；
  - 论文评分的关注重点
    - 有效的课程建议和意见
    - 关注问题的新颖度
    - 个人分析和讨论的深度
    - 论文的整体工作量