

# 数据挖掘导论

陈小军

大数据研究所

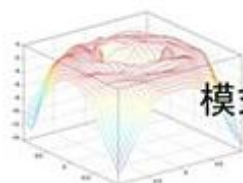
# 教师信息

- 姓名: 陈小军
- 所在研究所: 大数据研究所
- 研究方向: 数据挖掘、机器学习
- 办公室: 南区计算学院614
- 邮箱: [xjchen@szu.edu.cn](mailto:xjchen@szu.edu.cn)

# 机器学习 vs. 人类学习



# 相关学科



模式识别

计算机视觉



数据挖掘



机器学习

语音识别



统计学习

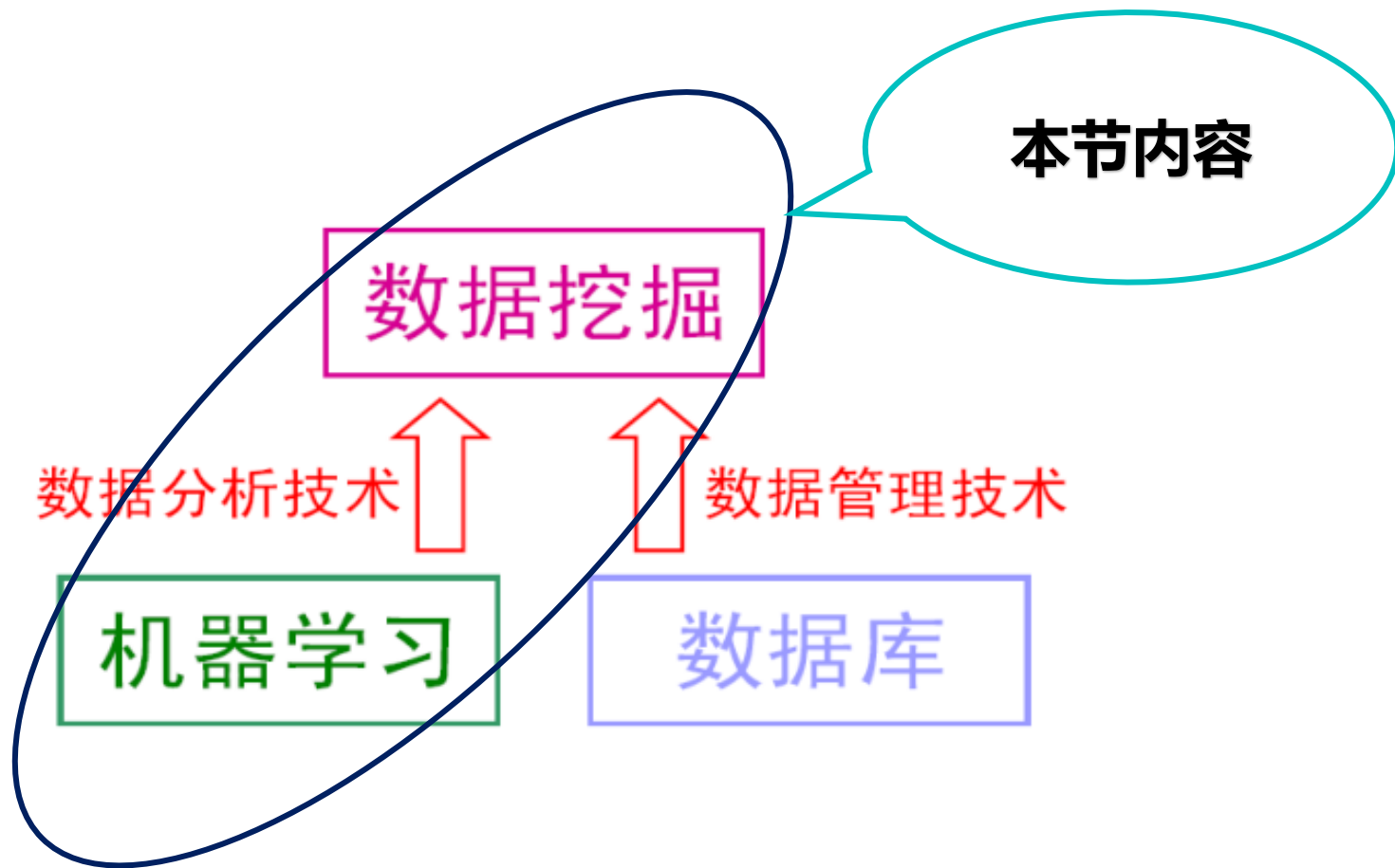


自然语言处理



# 数据挖掘

---



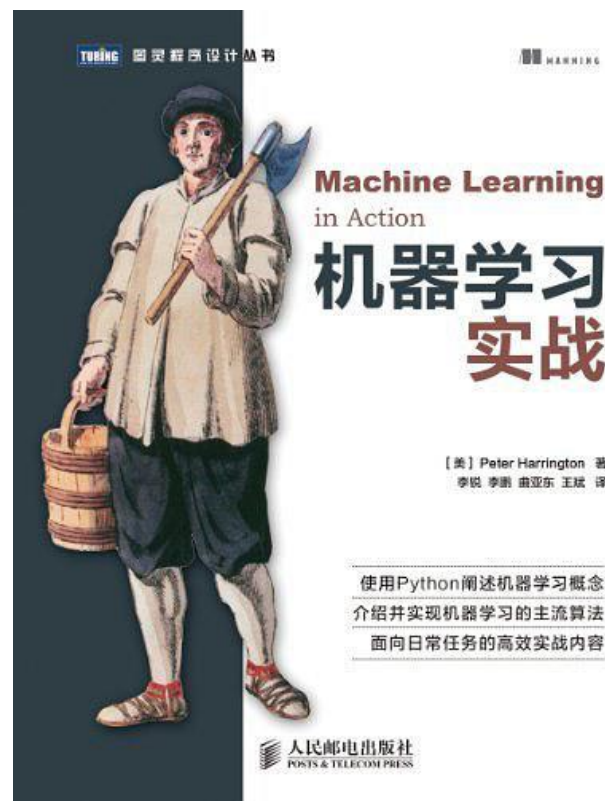
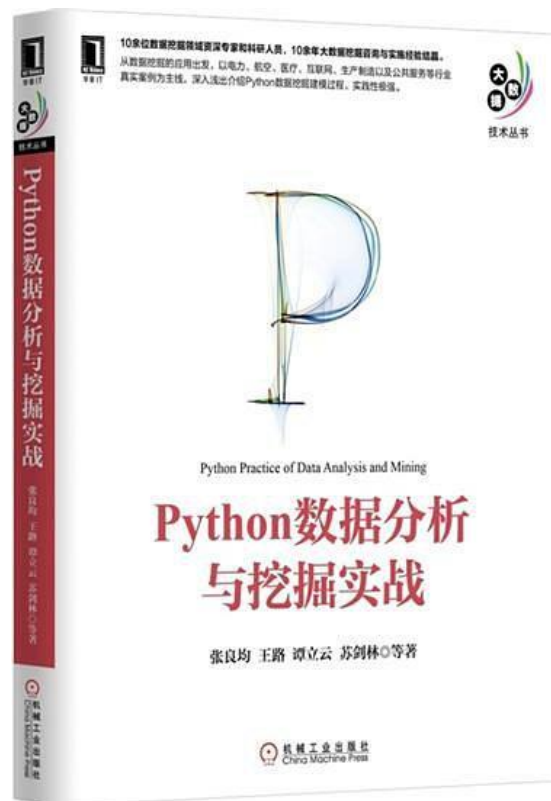
# 课程情况

序号	教学内容	学时	实验内容
1	绪论：课程简介、数据挖掘	2	
2	数据：数据类型、数据质量、相似度/相异度计算	2+2	Python安装及编程练习
3	数据预处理：数据清洗、数据集成、数据变换、数据规约、python主要预处理函数	2	
4	数据探索：数据质量、数据特征分析、python数据探索及统计作图函数	2+2	数据可视化实验
5	分类与预测：原理、回归、决策树、神经网络、集成学习等	8+2	分类与预测实验
6	聚类：原理、k-means、谱聚类	4+2	聚类算法实验
7	关联分析：原理、Apriori、FP-Growth	2	
8	时序分析：原理、平稳时间序列模型、非平稳时间序列模型、RNN	2	
9	异常检测：原理、基于模型的异常检测算法、基于聚类的异常检测算法	2	
10	案例讲解及实践	8+10	大作业
11	随堂考试	2	
合计		36+18	

# 考核标准

- 实验成绩占比: 30%
- 平时成绩占比: 10%
- 期末随堂考试占比: 30%
- 期末大作业占比: 30%

# 参考教材





# 第1章 绪论

# 目录

1	某知名连锁餐饮企业的困惑
2	从餐饮服务到数据挖掘
3	数据挖掘的基本任务
4	数据挖掘建模过程
5	常用数据挖掘建模工具

# T餐饮简介

- 国内某餐饮连锁有限公司（以下简称T餐饮）成立于1998年，主要经营粤菜，兼顾湘菜、川菜、中餐等综合菜系。至今已经发展成为在国内具有一定知名度、美誉度，多品牌、立体化的大型餐饮连锁企业。
- 属下员工1000多人，拥有16家直营分店，经营总面积近13000平方米，年营业额近亿元。
- 其旗下各分店均坐落在繁华市区主干道，雅致的装潢，配之以精致的饰品、灯具、器物，出品精美，服务规范。



# T餐饮简介

- 近年来餐饮行业面临较为复杂的市场环境，与其他行业一样餐饮企业都遇到了原材料成本升高、人力成本升高、房租成本升高等问题，这也使得整个行业的利润率急剧下降。人力成本和房租成本的上升是必然趋势，如何在保持产品质量同时提高企业效率，成为了T餐饮急需面对的问题。从2000年开始，T餐饮通过加强信息化管理来提高效率，目前已上线的管理系统包括：

1. 客户关系管理系统
2. 前厅管理系统
3. 后厨管理系统
4. 财务管理系统
5. 物资管理系统

# T餐饮的困惑

- 通过以上信息化的建设，T餐饮已经积累了大量的历史数据，有没有一种方法可帮助企业从这些数据中洞察商机，提取价值？在同质化的市场竞争中，找到一些市场以前并不存在的“捡漏”和“补缺”？

# 目录

1	某知名连锁餐饮企业的困惑
2	从餐饮服务到数据挖掘
3	数据挖掘的基本任务
4	数据挖掘建模过程
5	常用数据挖掘建模工具

# 餐饮如何盈利.....

- 企业经营最大的目的就是盈利，而餐饮业企业盈利的核心就是其菜品和顾客，也就是其提供的产品和服务对象。企业经营者每天都在想推出什么样的菜系和种类会吸引更多的顾客，究竟各种顾客各自的喜好是什么，在不同的时段是不是有不同的菜品畅销，当把几种不同的菜品组合在一起推出时是不是能够得到更好的效果，未来一段时间菜品原材应该采购多少.....

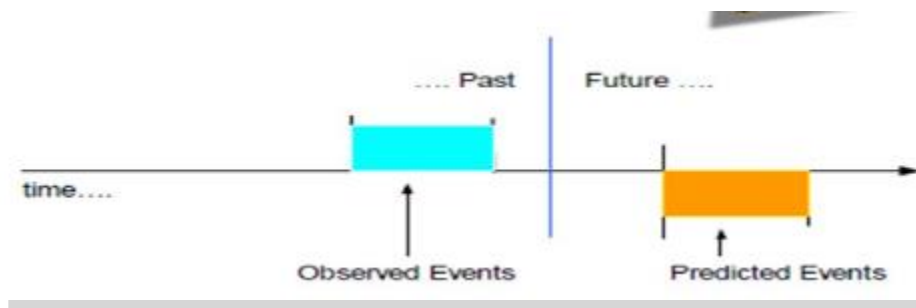
# T餐饮服务之经验.....

- T餐饮在经营过程中，通过分析历史数据，总结出一些行之有效的经验：
  1. 在点餐过程中，由有经验的服务员根据**顾客特点**进行菜品推荐，一方面可提高菜品的销量，另外一方面可减少客户点餐的时间和频率，提高用户体验；
  2. 根据菜品历史销售情况，综合考虑节假日、气候和竞争对手等影响因素，对菜品销量进行**预测**，以便餐饮企业提前准备原材料；
  3. 定期对菜品销售情况进行统计，**分类统计**出好评菜和差评菜，为促销活动和新品推出提供支持；
  4. 根据就餐频率和金额对顾客的就餐行为进行评分，**筛选出优质客户**，定期回访和送去关怀。



# 从餐饮服务到数据挖掘

- 以上经验从数据中获得有关产品和客户的特点以及能够产生价值的规律更多依赖于管理人员的个人经验。
- 如果有一套工具或系统，能够从业务数据中自动或半自动地发现相关的知识和解决方案，这将极大地提高企业的决策水平和竞争能力。
- 这种从数据中“淘金”，从大量数据（包括文本）中挖掘出隐含的、未知的、对决策有潜在价值的关系、模式和趋势，并用这些知识和规则建立用于决策支持的模型，提供预测性决策支持的方法、工具和过程，就是**数据挖掘**。



# 分析能力的八个等级



# 分析能力的八个等级



## 常规报表 STANDARD REPORTS

**回答:** 发生了什么? 什么时候发生的?

**示例:** 月度或季度财务报表

我们都见过报表, 它们一般是定期生成, 用来回答在某个特定的领域发生了什么。从某种程度上来说它们是有用的, 但无法用于制定长期决策。



## 即席查询 AD HOC REPORTS

**回答:** 有多少数量? 发生了多少次? 在哪里?

**示例:** 一周内各天各种门诊的病人数量报告。

即席查询的最大好处是, 让你不断提出问题并寻找答案。



## 多维分析 OLAP

**回答:** 问题到底出在哪里? 我该如何寻找答案?

**示例:** 对各种手机类型的用户进行排序, 探查他们的呼叫行为。

通过多维分析(OLAP)的钻取功能, 可以让您有初步的发现。钻取功能如同层层剥笋, 发现问题所在。



## 警报 ALERTS

**回答:** 我什么时候该有所反应? 现在该做什么?

**示例:** 当销售额落后于目标时, 销售总监将收到警报。

警报可以让您知道什么时候出了问题, 并当问题再次出现时及时告知您。警报可以通过电子邮件、RSS订阅、评分卡或仪表盘上的红色信号灯来展示。

# 分析能力的八个等级



## 统计分析 STATISTICAL ANALYSIS

**回答:** 为什么会出现这种情况? 我错失了什么机会?

**示例:** 银行可以弄清楚为什么重新申请房贷的客户在增多。

这时您已经可以进行一些复杂的分析, 比如频次分析模型或回归分析等等。统计分析是在历史数据中进行统计并总结规律。



## 预报 FORECASTING

**回答:** 如果持续这种发展趋势, 未来会怎么样? 还需要多少? 什么时候需要?

**示例:** 零售商可以预计特定商品未来一段时间在各个门店的需求量。

预报可以说是最热门的分析应用之一, 各行各业都用得到。特别对于供应商来说, 能够准确预报需求, 就可以让他们合理安排库存, 既不会缺货, 也不会积压。



## 预测型建模 PREDICTIVE MODELING

**回答:** 接下来会发生什么? 它对业务的影响程度如何?

**示例:** 酒店和娱乐行业可以预测哪些VIP客户会对特定度假产品有兴趣。

如果您拥有上千万的客户, 并希望展开一次市场营销活动, 那么哪些人会是最可能响应的客户呢? 如何划分出这些客户? 哪些客户会流失? 预测型建模能够给出解答。



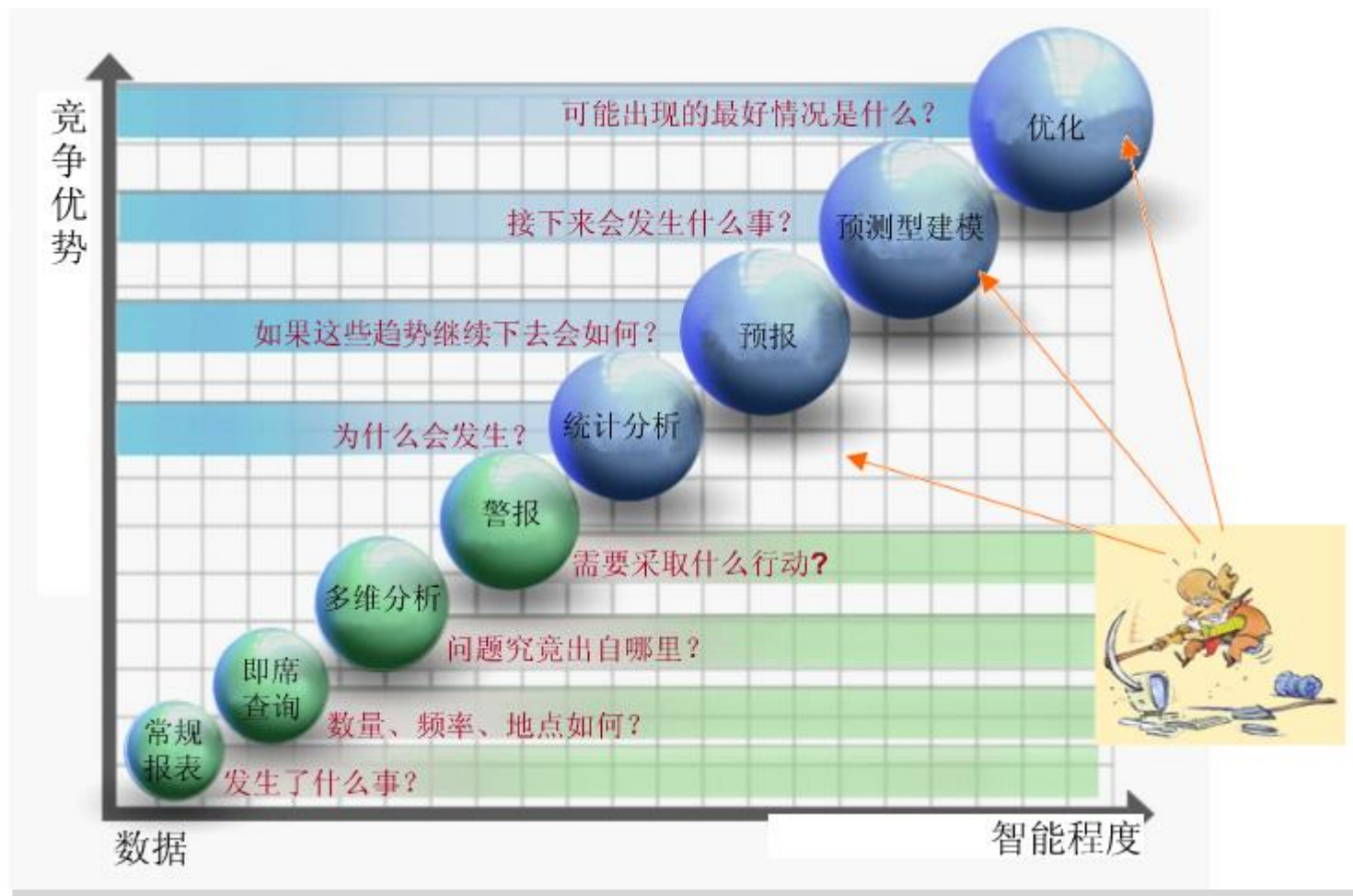
## 优化 OPTIMIZATION

**回答:** 如何把事情做得更好? 对于一个复杂问题来说, 那种决策是最优的?

**示例:** 在给定了业务上的优先级、资源调配的约束条件以及可用技术的情况下, 请您来给出IT平台优化的最佳方案, 以满足每个用户的需求。

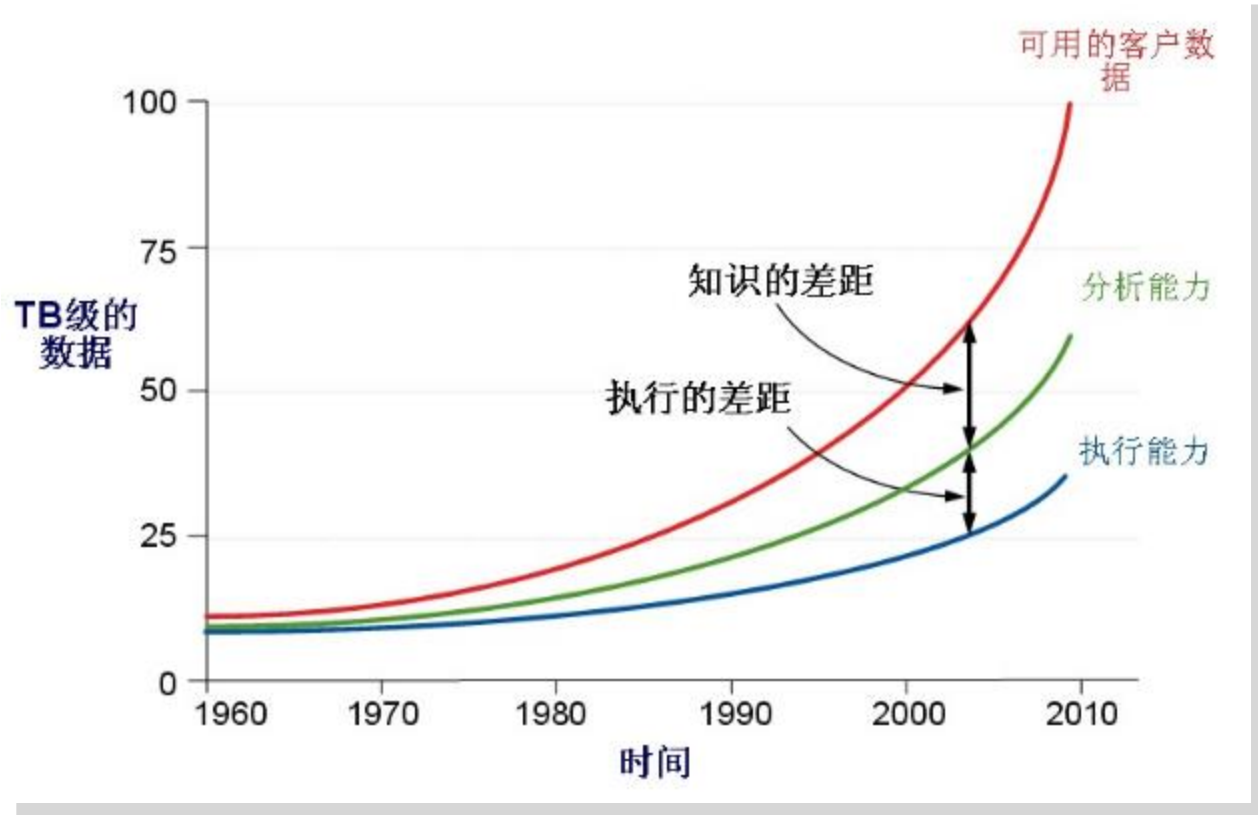
优化带来创新, 它同时考虑到资源与需求, 帮助您找到实现目标的最佳方式。

# 数据分析能力的演进





# 分析和执行能力远跟不上信息的增长



# 目录

1	某知名连锁餐饮企业的困惑
2	从餐饮服务到数据挖掘
3	数据挖掘的基本任务
4	数据挖掘建模过程
5	常用数据挖掘建模工具

# 数据挖掘的基本任务

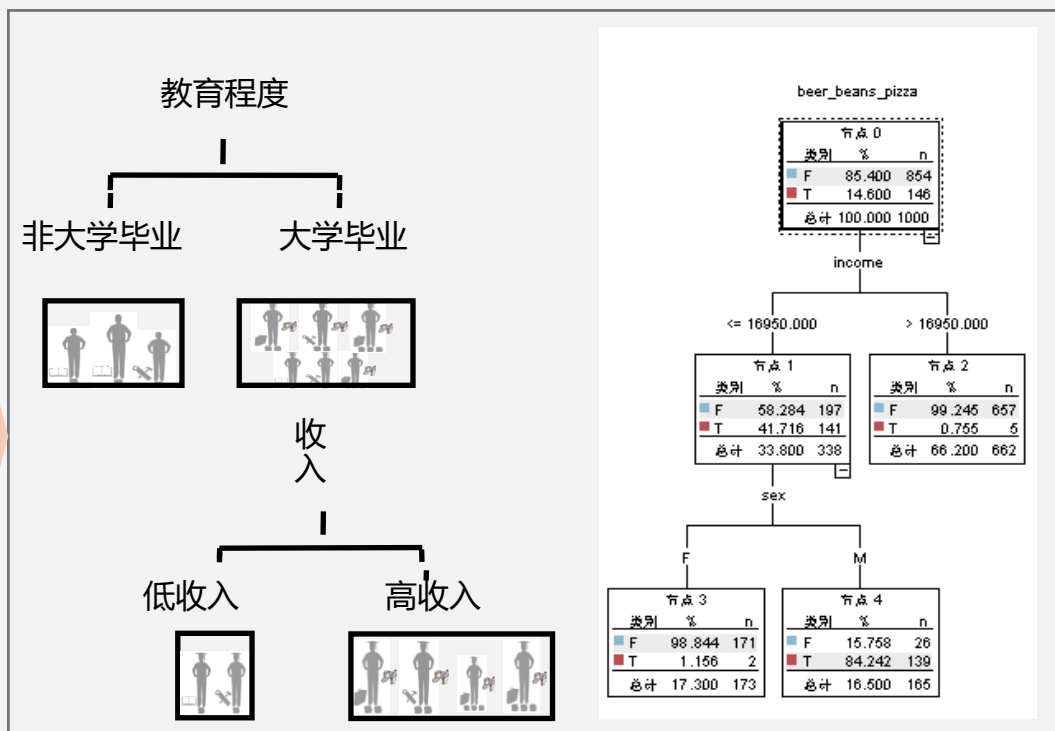
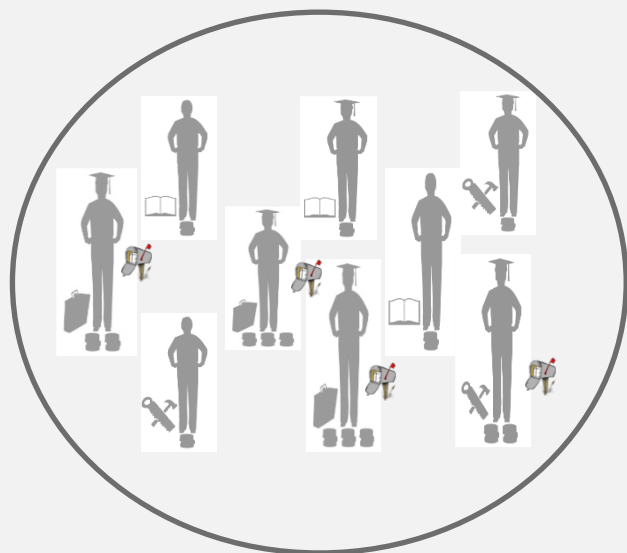
- 数据挖掘的**基本任务**包括利用分类与预测、聚类分析、关联规则、时序模式、偏差检测、智能推荐等方法，帮助企业提取数据中蕴含的商业价值，提高企业的竞争力。
- 对餐饮企业而言，数据挖掘的基本任务是从餐饮企业采集各类菜品销量、成本单价、会员消费、促销活动等内部数据，以及天气、节假日、竞争对手以及周边商业氛围等外部数据；之后利用数据分析手段，实现菜品智能推荐、促销效果分析、客户价值分析、新店选点优化、热销/滞销菜品分析和销量趋势预测；最后将这些分析结果推送给餐饮企业管理者及有关服务人员，为餐饮企业降低运营成本，增加盈利能力，实现精准营销，策划促销活动等提供智能服务支持。



# 数据挖掘的基本任务

## ➤ 分类与预测

有目标的对事物进行分类预测，如：客户流失预测、偷窃电用户识别等。



# 数据挖掘的基本任务

## ➤ 关联规则

关联模式挖掘旨在从大量的数据当中发现特征之间或数据之间的相互依赖关系。这种存在于给定数据集中的频繁出现的关联模式，又称为关联规则。

### Buying Pattern



前项(Antecedent)



蔬菜



鲜鱼

后项(Consequent)



红酒? 啤酒?

前项(Antecedent)



手机



配饰

后项(Consequent)



耳机? 内存?

前提(1) & 前提(2) & ... & 前提(m)

结论

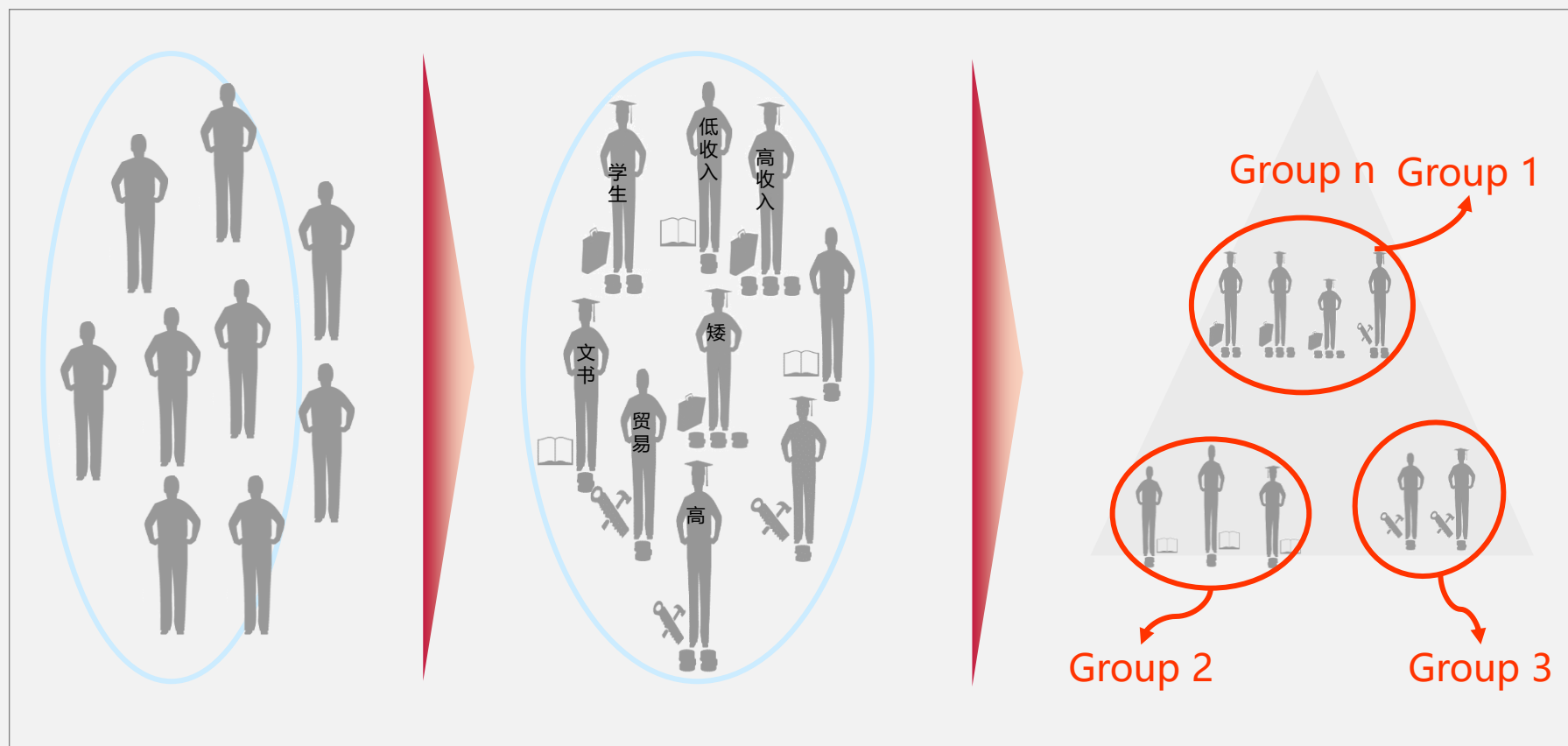
Antecedents

Consequent

# 数据挖掘的基本任务

## ➤ 聚类分析

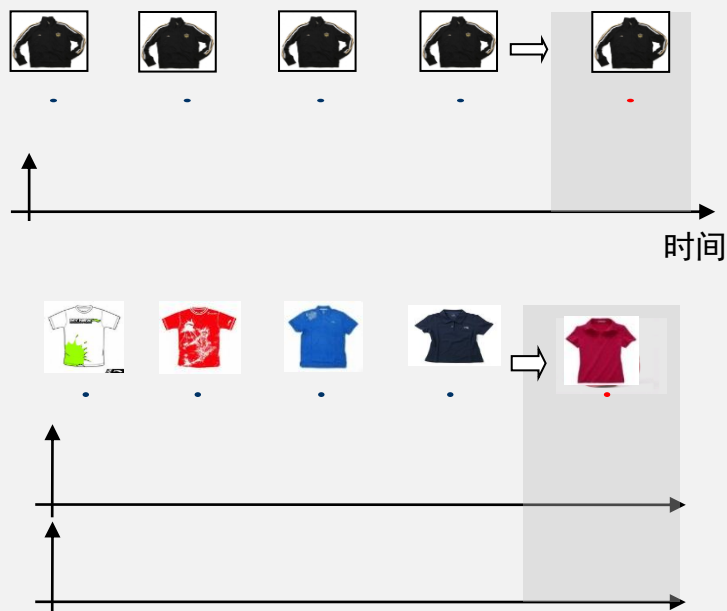
聚类分析是根据数据本身结构特征对数据点进行分类的方法。实质是按照彼此距离的远近将数据分为若干个类别，以使得类别内数据的“差异性”尽可能小(即“同质性”尽可能大)，类别间“差异性”尽可能大。



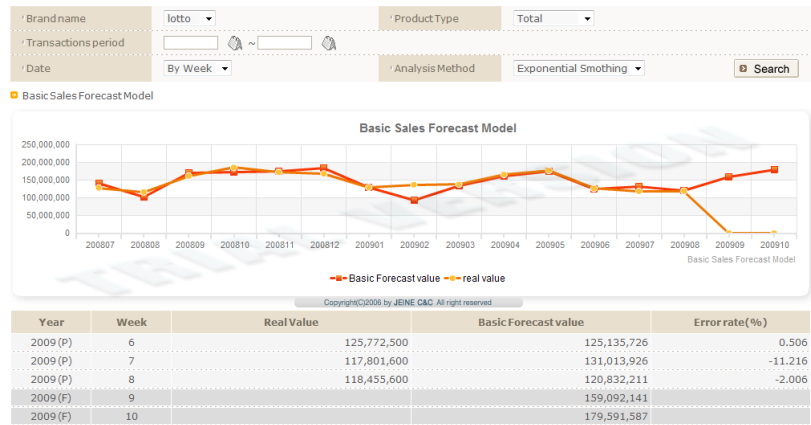
# 数据挖掘的基本任务

## ➤ 时间序列

基于事物发展的延续性和随机性预测事物未来的发展，如：销售量预测、天气预测等。



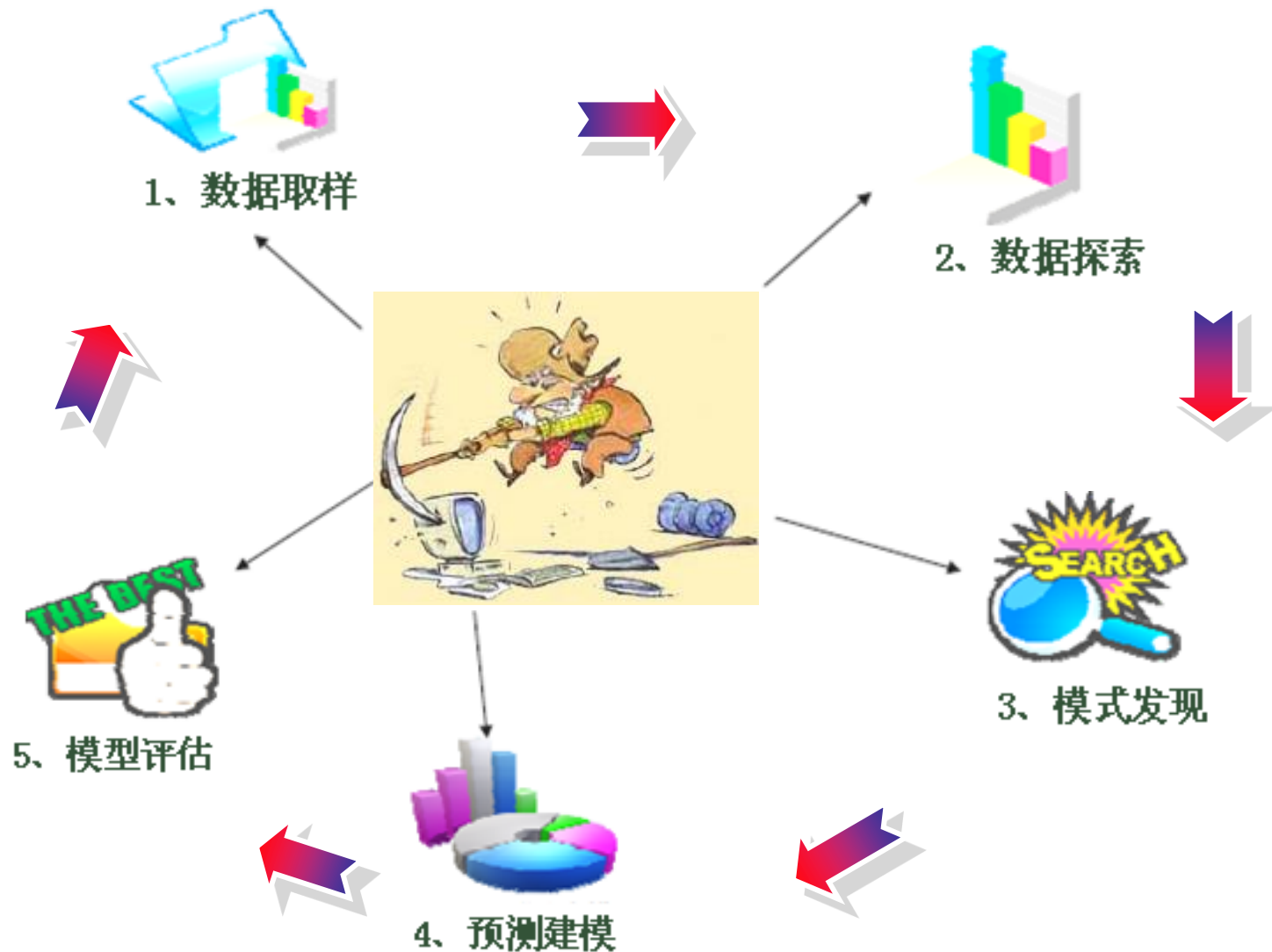
Basic Sales



# 目录

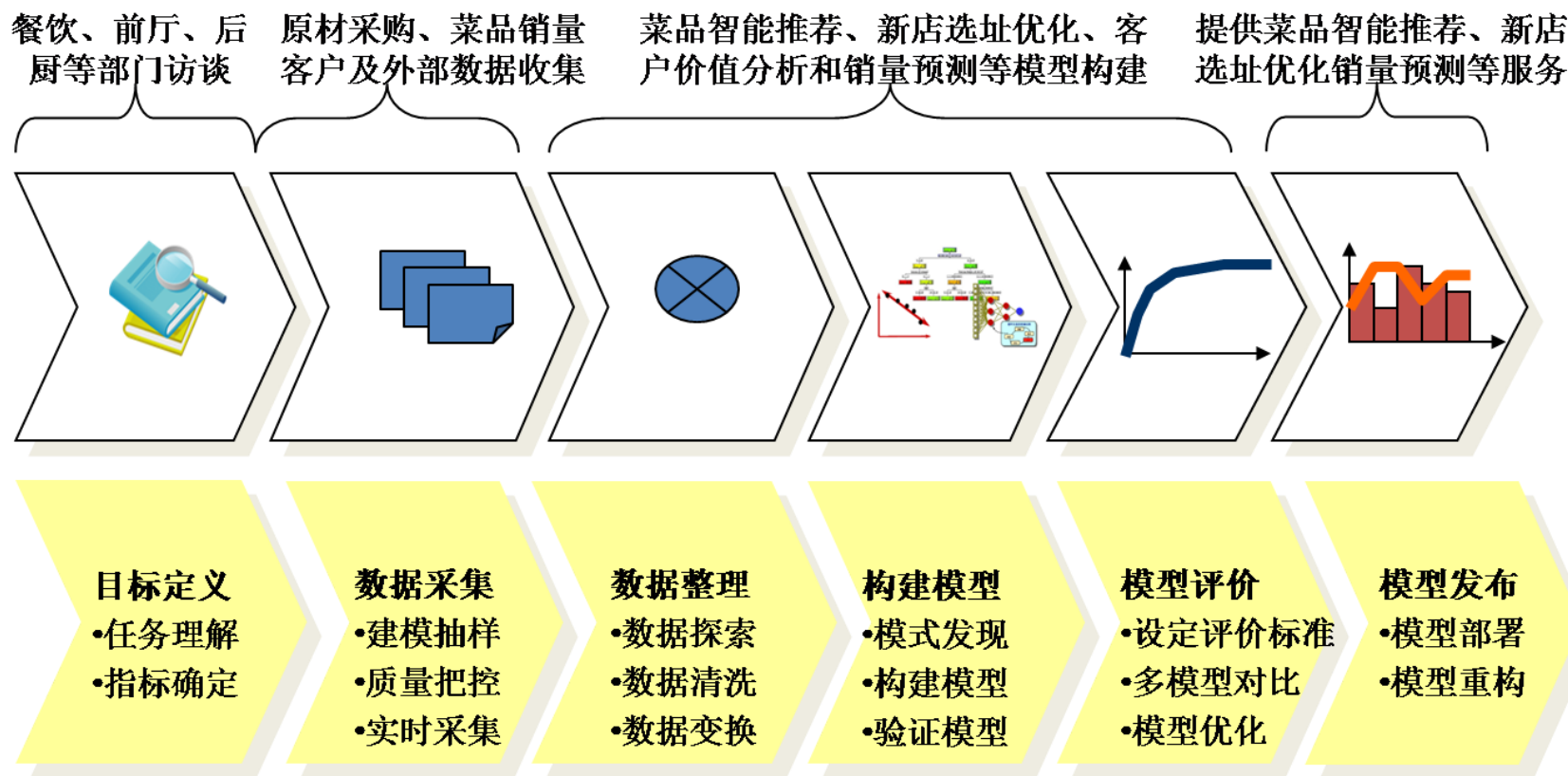
1	某知名连锁餐饮企业的困惑
2	从餐饮服务到数据挖掘
3	数据挖掘的基本任务
4	数据挖掘建模过程
5	常用数据挖掘建模工具

# 数据挖掘建模过程



# 数据挖掘建模过程

## ● 餐饮行业数据挖掘建模过程：



# 数据挖掘建模过程

## 第1步：定义挖掘目标

- 针对具体的数据挖掘应用需求，首先要明确本次的挖掘目标是什么？系统完成后能达到什么样的效果？因此我们必须分析应用领域，包括应用中的各种知识和应用目标，了解相关领域的有关情况，熟悉背景知识，弄清用户需求。要想充分发挥数据挖掘的价值，必须要对目标有一个清晰明确的定义，即决定到底想干什么。



# 数据挖掘建模过程

## 第1步：定义挖掘目标

- 针对餐饮行业的数据挖掘应用，可定义如下挖掘目标：
  1. 实现动态菜品智能推荐，帮助顾客快速发现自己感兴趣的菜品，同时确保推荐给顾客的菜品也是餐饮企业所期望的，实现餐饮消费者和餐饮企业的双赢；
  2. 对餐饮客户进行细分，了解不同客户的贡献度和消费特征，分析哪些客户是最有价值的，哪些是最需要关注的，对不同价值的客户采取不同的营销策略，将有限的资源投放到最有价值的客户身上，实现精准化营销；
  3. 基于菜品历史销售情况，综合考虑节假日、气候和竞争对手等影响因素，对菜品销量进行趋势预测，方便餐饮企业准备原材料；
  4. 基于餐饮大数据，优化新店选址，并对新店位置的潜在顾客口味偏好进行分析，以便及时进行菜式调整。

# 数据挖掘建模过程

## 第2步：数据取样

- 在明确了需要进行数据挖掘的目标后，接下来就需要从业务系统中抽取出一个与挖掘目标相关的样本数据子集。**抽取数据的标准**，一是相关性，二是可靠性，三是有效性，而不是动用全部企业数据。通过数据样本的精选，不仅能减少数据处理量，节省系统资源，而且使我们想要寻找的规律性更加突显出来。
- 进行数据取样，一定要严把质量关。因为数据挖掘是要探索企业运作的内在规律性，原始数据有误，就很难从中探索规律性。若真的从中还探索出来了什么“规律性”，再依此去指导工作，则很可能造成误导。若从正在运行的系统中进行数据取样，更要注意数据的完整性和有效性。

# 数据挖掘建模过程

## 第2步：数据取样

### ● 数据抽样方法：

1. 随机抽样：在采用随机抽样方式时，数据集中的每一组观测值都有相同的被抽样的概率。
2. 等距抽样：如按 5%的比例对一个有100 组观测值的数据集进行等距抽样，则有： $100 / 5 = 20$ ，等距抽样方式是取第20、40、60、80 和第100 五组观测值。
3. 分层抽样：在这种抽样操作时，首先将样本总体分成若干个子集。在每个层次中的观测值都具有相同的被选用的概率，但对不同的层次可设定不同的概率。这样的抽样结果通常具有更好的代表性，进而使模型具有更好的拟合精度。
4. 从起始顺序抽样：这种抽样方式是从输入数据集的起始处开始抽样。抽样的数量可以给定一个百分比，或者直接给定选取观测值的组数。
5. 分类抽样：在前述几种抽样方式中，并不考虑抽取样本的具体取值。分类抽样则依据某种属性的取值来选择数据子集。，如按客户名称分类、按地址区域分类等。分类抽样的选取方式就是前面所述的几种方式，只是抽样以类为单位。

# 数据挖掘建模过程

## 第2步：数据取样

### ● 餐饮建模数据取样：

1. 餐饮企业信息：名称、位置、规模、联系方式；部门、人员、角色等；
2. 餐饮客户信息：姓名、联系方式、消费时间、消费金额等；
3. 餐饮企业菜品信息：菜品名称、菜品单价、菜品成本、所属部门等；
4. 菜品销量数据：菜品名称、销售日期、销售金额、销售份数；
5. 原材料供应商资料及商品数据：供应商姓名、联系方式、商品名称；客户评价信息；
6. 促销活动数据：促销日期、促销内容、促销描述；
7. 外部数据，如天气、节假日、竞争对手以及周边商业氛围等数据。

# 数据挖掘建模过程

## 第3步：数据探索

- 数据取样，多少是带着人们对如何实现数据挖掘目的的先验认识进行操作的。当我们拿到了一个样本数据集后，它是否达到我们原来设想的要求；其中有没有什么明显的规律和趋势；有没有出现从未设想过的数据状态；属性之间有什么相关性；它们可区分成怎样一些类别.....
- 对所抽取的样本数据进行探索、审核和必要的加工处理，是保证最终的挖掘模型的质量所必需的。可以说，挖掘模型的质量不会超过抽取样本的质量。数据探索和预处理的目的是为了保证样本数据的质量，从而为保证模型质量打下基础。
- 数据探索主要包括：异常值分析、缺失值分析、相关分析、周期性分析等。

# 数据挖掘建模过程

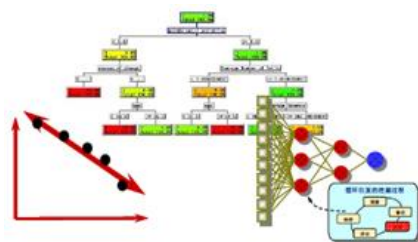
## 第3步：数据预处理

- 由于采样数据中常常包含许多含有噪声、不完整、甚至不一致的数据，对数据挖掘所涉及的数据对象必须进行预处理。那么如何对数据进行预处理以改善数据质量，并最终达到完善最终的数据挖掘结果的目的呢？
- 针对采集的餐饮数据，数据预处理主要包括：数据筛选、数据变量转换、缺失值处理、坏数据处理、数据标准化、主成分分析、属性选择、数据规约等。

# 数据挖掘建模过程

## 第4步：挖掘建模

- 样本抽取完成并经预处理后，接下来要考虑的问题是：本次建模属于数据挖掘应用中的哪类问题（分类、聚类、关联规则、时序模式或是智能推荐），选用哪种算法进行模型构建？
- 针对餐饮行业的数据挖掘应用，挖掘建模主要包括基于关联规则算法的动态菜品智能推荐、基于聚类算法的餐饮客户价值分析、基于分类与预测算法的菜品销量预测、基于整体优化的新店选址。
- 以菜品销量预测为例，模型构建是对菜品历史销量，综合考虑节假日、气候和竞争对手等采样数据轨迹的概括，它反映的是采样数据内部结构的一般特征，并与该采样数据的具体结构基本吻合。模型的具体化就是菜品销量预测公式，公式可以产生与观察值有相似结构的输出，这就是预测值。



# 数据挖掘建模过程

## 第5步：模型评价

- 模型评价的目的之一就是从这些模型中自动找出一个最好的模型出来，另外就是要根据业务对模型进行解释和应用。
- 对分类与预测模型和聚类分析模型的评价方法是不同的。
- 不管黑猫、白猫，抓到老鼠就是好猫。



# 目录

1	某知名连锁餐饮企业的困惑
2	从餐饮服务到数据挖掘
3	数据挖掘的基本任务
4	数据挖掘建模过程
5	常用数据挖掘建模工具

# 常用数据挖掘建模工具

- SAS Enterprise Miner
- IBM SPSS Modeler
- SQL Server
- Python
- WEKA
- KNIME
- RapidMiner
- TipDM

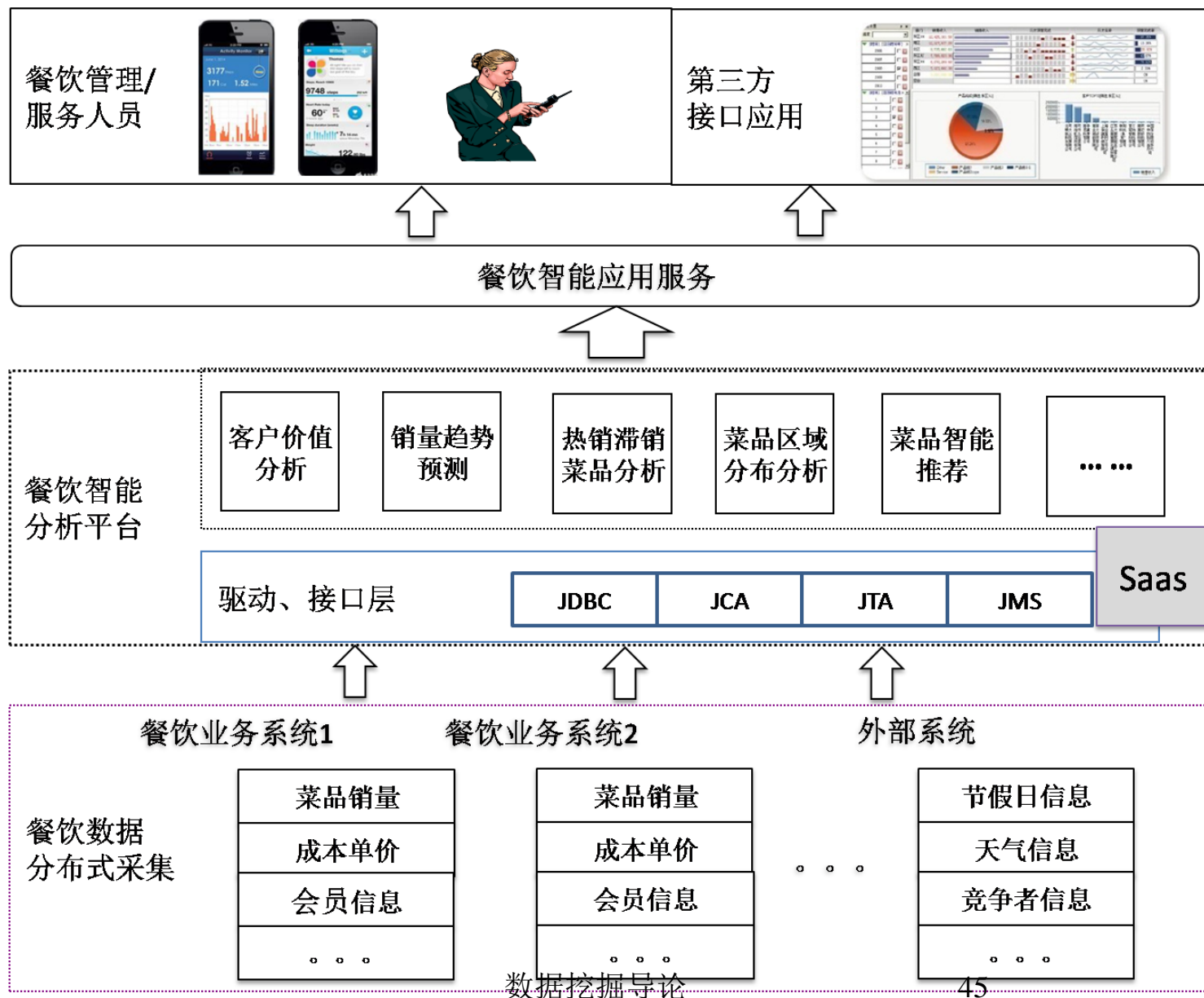
# 拓展思考

- 根据商务部餐饮行业统计数据测算，2012年，全社会提供正餐服务、快餐服务、饮料及冷饮服务、其他餐饮服务的餐饮企业单位共计236.7万个。上一定规模的餐饮管理软件企业超过500家。

# 拓展思考

- 如果你是大数据实践者，你想像的基于大数据的餐饮服务平台是怎样的？

# 拓展思考—餐饮智能服务平台



# • 餐饮智能服务平台—主要界面

餐饮智能服务平台

2014-06-26

泰迪餐饮

收入总额

38053.20

成本总额

22285.47

毛利/毛利率

15767.73/0.71

功能中心

周

最高

38053.20

2014-06-26

最低

24573.00

2014-06-23

月

最高

172627.50

2014-06-15

最低

23258.90

2014-06-19

年

最高

177920.90

2014-05-11

最低

8093.00

2014-03-03

时菜

节菜

原材

反馈

更多

←

泰迪餐饮

营业额: 105609.6

2014年04月27日

当日

当月

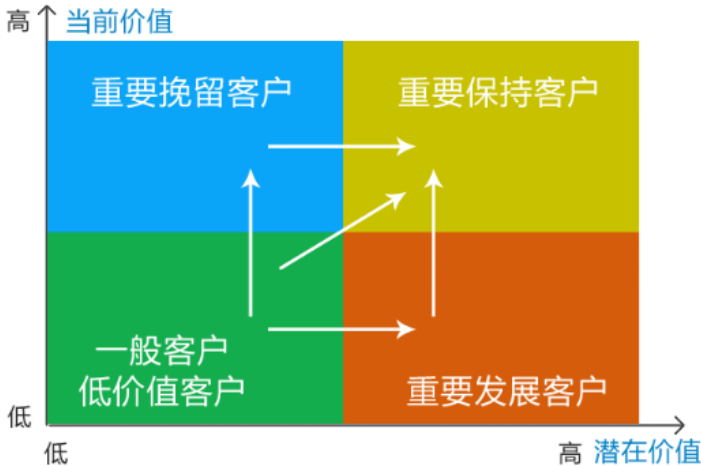
当年

门店	部门	销售额
东圃西店	中厨	13132.0
东圃西店	海鲜	5821.0
东圃西店	点心	7846.0
东圃西店	豆腐档	732.0
东圃西店	楼面	1259.0
东圃西店	汤水	1633.0
东圃西店	味部	5675.0
东圃西店	房间茶位	755.0

# • 餐饮智能服务平台——主要界面



会员价值矩阵



客户分群	人数	R	F	M
低价值客户	127	14.81	13.51	29.56
重要发展客户	281	14.38	2.16	466.33
重要挽留客户	42	60.17	7.83	358.24
重要保持客户	492	14.89	12.99	120.33



重要保持客户

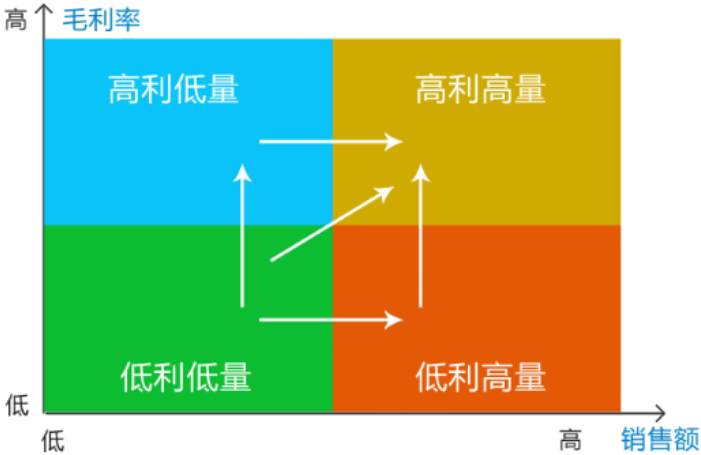
序号	姓名	R	F	M
2	林斯达	3	5	1507.11
3	林细君	4	16	817.62
5	林晓莲	14	7	1913.05
10	尊敬的会员：感谢您的支持，凭本短信，乐膳海鲜酒家6月17日中午5折优惠，本信息转发无效！			
13	刘荔	17		1744.55
14	刘荔	30	16	1957.44
15	刘沛欣	5	7	1713.79
16	刘琪	4	21	1768.11
23	卢坚明	12	13	1434.29

推送广告

# • 餐饮智能服务平台——主要界面

←

菜品利量矩阵 II



菜品分群	个数	均毛利	均销份数
低利低量	942	0.36	14.37
低利高量	152	0.35	579.91
高利低量	2001	0.92	16.64
高利高量	332	0.96	801.79

←

泰迪餐饮

🕒

菜品	份数^	单价^	毛利^
拆骨猪手/特价	42	1	0.92
原汁原味菜心	17	32	0.99
凉瓜焖三文鱼骨/特价	17	16.8	0.25
铁板酸菜豆腐	15	38	0.43
金牌烧鹅/例	14	45	0.31
番薯叶	12	28	0.4
上汤辣椒叶	11	32	0.26
干炒牛河	11	12	0.88
红烧乳鸽	10	25	0.55
东北黑木耳	10	8	0.7



# • 餐饮智能服务平台—主要界面

←

菜品智能推荐

全部

中厨

海鲜

豆腐档

味部

猜你喜欢

白切莲藕 20.0	- 2 +
金牌烧鹅/例 45.0	- 0 +
乐膳真味鸡/半只 45.0	- 2 +
东北黑木耳 8.0	- 0 +
柠香水晶汾肉 38.0	- 0 +
真味猪手/特价	- 0 +

4 ¥130.0

←

泰迪餐饮菜品维护

菜品	单价	成本
豉油皇吊桶仔	48.32	24.64
银芽肉丝炒台湾茨粉	23.5	10.58
元贝/斤	68.0	21.76
桂花鱼/斤	78.0	40.56
虫草花木耳蒸牛肉	38.0	15.96
芥兰炒腊味	38.0	19.76
农家三宝炒土猪肉	18.0	6.84
乐膳一桶骨	38.0	14.06
客家煎酿豆腐	26.0	10.92
黄豆酱炒云南通菜	49 28.0	10.08

# • 餐饮智能服务平台—主要界面

菜品智能推荐终端

店家推荐

主菜

特色菜

开胃菜


点心

小吃


菜品搜索...




东坡肉 45.0 元/份




小炒鸡杂 23.0 元/份




四季发财 24.0 元/份




香酥带鱼 35.0 元/份




口水鸡 45.0 元/份




蒜炒青豆 18.0 元/份




蟹香拔丝 25.0 元/份



清蒸花蟹 37.0 元/份



浓香鲍鱼 24.0 元/份



红烧猪脚 47.0 元/份



炸酱排骨 38.0 元/份



鸡肉豆米笋丁 23.0 元/份

数量： 0 份  
金额： 0.0 元  

进入订单

- 动态菜品智能推荐

- 菜品推荐的目的：

- 1) 帮助顾客快速发现自己感兴趣的菜品；
- 2) 推荐给顾客的菜品最好也是餐饮企业期望的。

*什么是餐饮企业期望的？*

**实现餐饮消费者和餐饮企业的双赢**

## • 动态菜品智能推荐

基于关联规则挖掘的个性化菜品智能推荐设计：

- 待点菜品与已选菜品是相关联的；
- 热销度总体能反映客户对不同菜品的喜好程度；
- 选择的菜品毛利率越高，对商家越有利；
- 商家有主推菜品；
- .....

## • 动态菜品智能推荐（效果）

T餐饮的某单店，日均营业收入2万~6万，2014年6月开始同时上线本平台，效果如下：

- 单店一次性新增平板终端20台，费用：2.7万；
- 点菜服务员减少3人，节约成本：1.2万/月；
- 菜品原材节省5~10%，节约成本：2.3万/月；
- 店家经常有主推菜品，需动态更新菜单，印刷成本0.1万/月。