

第 1 章 揭开数据挖掘的面纱

徐教授是某985院校的著名教授，国内数据挖掘专家、智能信息处理研究方向学术带头人，主持了20多项国家项目和国际合作项目，具有丰富的数据挖掘项目实施经验，获得过多项国家级大奖。数十年来，他潜心科研，除了给自己学院的本科生和研究生上课外，一直谢绝其他授课邀请。这次他破例了，欣然接受了本校管理学院第5届EMBA班的“数据挖掘及其应用”课程……



1.1 历史的使命

今天是第一节课，徐教授一跨进教室，迎接他的是学员们一阵热烈的掌声。他习惯性地扫视了一下学生，果然正像管理学院张院长介绍的那样，在座的学员不同寻常，年龄在35~50岁之间，个个西装革履，精神焕发，眼睛里放射出对新知识无比渴望的光芒。

徐教授走上讲台，先在黑板上写下了自己的名字和联系方式，然后微笑着说：“同学们，今天我能站在这儿给大家上课，不是因为你们管院张院长有面子，也不是因为你们这些学员地位有多高，说实在的，是党中央、国务院让我来的。”学员们个个目瞪口呆。

有人嘀咕道：“难道中央还关心我们这个EMBA班？。”

“关心，而且非常关心。”徐教授铿锵有力地回答。

大家更加疑惑了。

徐教授提高了嗓门：“2006年1月9日，在全国科技大会上，党中央、国务院作出了建设创新型国家的重大决策。大家都知道，创新型国家是指以技术创新为经济社会发展核心驱动力的国家。技术创新需要科学家和科技工作者的努力，更离不开政府和企业高层领导和管理人员的推动。张院长在邀请我来给你们上课时介绍说，在座各位都在政府部门或者企业地位显赫，所以我欣然地、破天荒地答应了你们院长的邀请。不过，别以为是你们的乌纱帽吸引了我，而是你们每一个人身上肩负的‘建设创新型国家’的历史使命召唤着我。”

徐教授越说越激动，喝了口水继续说：“我为科学事业奋斗了一辈子，深知‘象牙塔’里的发明、创造，需要与经济建设结合才更能体现出其价值，才更能为建设创新型国家做出贡献。理论创新的成果要真正转化为生产力，迫切需要一种推动力、催

化剂。而能起到这种作用的主体非你们这些人莫属，诚如是，你们就是建设创新型国家的排头兵。你们说，党中央能不关心你们吗？”



徐教授的话音刚落，教室里立刻响起长时间的掌声。

他双手从上向下慢慢挥动，示意大家停下，接着说：“近十年来数据挖掘技术飞速发展，在国外，数据挖掘正在变成整个信息技术的核心之一。尤其是世界500强企业均设立了数据挖掘研发与应用部门，数据挖掘技术已成为其业务成功的关键因素。2007年5月，《纽约时报》以‘数据挖掘正在进入主流’为题，介绍了数据挖掘技术，并指出这种新技术正在变成人们工作和生活中不可或缺的一个部分。”

徐教授停顿了一下，向大家问道：“在国内，数据挖掘应用的状况怎样？”

T钢铁公司的李部长抢先答道：“在我国，数据挖掘在互联网、金融、电信和商业等领域已经有一些成功的应用，而在其他行业如制造、航空、医药、反恐和刑侦等只有少量的尝试。”

“李部长的评价比较客观，但大家想过没有，为什么我们与发达国家的差距就这么大呢？”徐教授反问道。

教室里一阵沉默。

于是，徐教授坦率地表达了自己的看法：“其实我也一直在考虑这个问题，当然这里面的原因很多。直到你们管院张院长请我给你们上数据挖掘课时，我又发现了一个不可忽视的因素——政府和企业高层对数据挖掘不甚了解而导致他们对此不够重视或不能站在一定的高度提出有价值的需求。”

徐教授的一席话引起了李部长的共鸣，激动地说：“是的，徐教授讲得太对了。就拿我们钢铁公司来说吧，这几年，我们整天喊‘挺进世界500强’，忙于引进国外先进设备扩大生产规模，但却忽视与外界的技术交流而成为井底之蛙，就连数据挖掘这样在世界500强企业如雷贯耳的新技术我们却闻所未闻。由于自己不具备这方面的知识，生产管理中遇到了不能解决的问题，自然不会用数据挖掘的思想思考，甚至基层部门提出使用这样的方法，领导层却因对此不甚了解而不给力支持。”

李部长的话送到了其他学员的心坎上，他们个个首肯。

徐教授走下讲台，语重心长地说：“所以，我给你们上数据挖掘课来了，我期望从领导普及数据挖掘知识开始，唤起人们对数据的新认识，使你们告别基于简单统计分析的‘报表’决策时期，跨入使用数据挖掘技术的‘知识’决策时代。你们这些社会各界的精英们肩负的历史责任太大了，不管是政府部门的领导还是企业的老总，你们每天都在做各种各样的决策，稍有不慎就可能给国家和企业带来重大损失。我相信各位想为国家贡献自己的力量，但陷入‘心有余而力不足’的境地，正所谓‘我们沉浸在数据的海洋，渴望知识的淡水’！”



听完徐教授一席话，下面的各位老总感慨颇多，台下一片沉思。

徐教授鼓励大家道：“数据挖掘的最高境界就是‘从数据中获取知识，辅助科学决策’。希望通过我们的数据挖掘课程的学习，使你们了解到什么是数据挖掘？它能够干什么？有哪些数据挖掘技术？怎么应用？大家要认识到，数据挖掘不同于一般的管理软件，编好了拿来用就是了，数据挖掘在行业的成功应用也是一种创新。其实在数据挖掘算法方面，国内（也包括我）的研究团队也有一系列的国际水平的研究成果，但愿我们一起共同努力，推动数据挖掘技术在各行各业的应用，为建设创新型国家做出最大的贡献！”

教室里，又是一阵激动人心的掌声。

徐教授摆了摆手，接着说：“不过，给你们上这门课可让我费了不少脑筋，你们这些学员走向工作岗位都在10年以上了，大学所学的数学知识大都还给了老师，针对

研究生的讲法对你们不适用了。不过，我想出一种专门针对你们的案例教学法，通过典型的应用实例深入浅出地介绍数据挖掘的概念、功能、流程和算法。”

“太好了，徐老师。我曾经翻过几本数据挖掘的书籍，但理论性太强，满篇数学公式，真让人望而却步，而且应用实例甚少，让人难以理解。”李部长感慨地说。

徐教授接着说：“OK，言归正传，让我们开始数据挖掘之旅吧。我先给大家讲三个真实的故事，让你们感受一下数据挖掘到底是神马还是浮云？”

1.2 数据挖掘的故事

1.2.1 震撼业界的发现

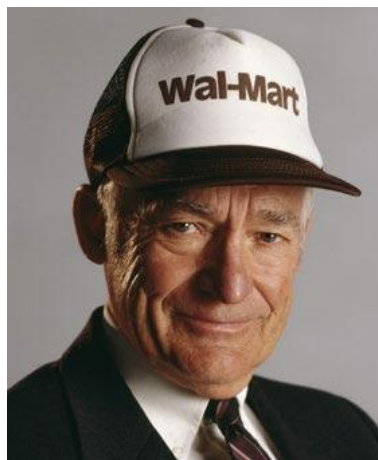
“有一个人叫萨姆·沃尔顿的人，大家认识吧？”徐教授问道。

教室里鸦雀无声。

“那沃尔玛，谁没听说过？”徐教授接着问。

“连三岁小孩都知道。”一学员小声说。

“哈哈，萨姆·沃尔顿是沃尔玛公司的创始人呀！”徐教授笑着说。

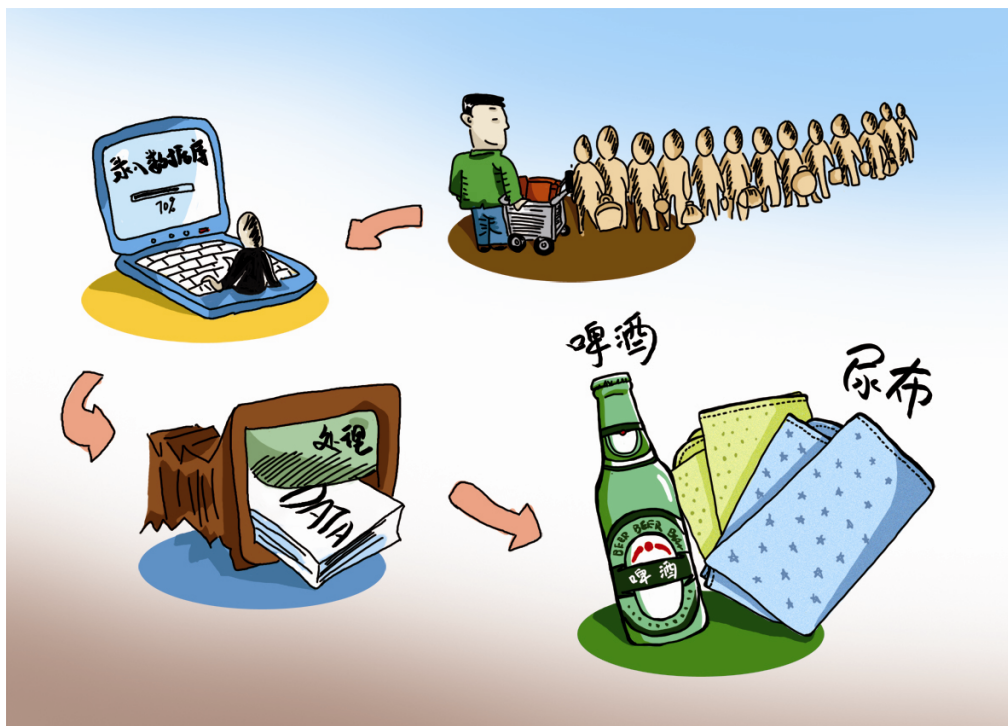


“对了，想起来了，萨姆·沃尔顿，是他将一个百货商店奇迹般地经营为全球最大的连锁零售企业，早在 1985 年 10 月就被《福布斯》杂志列为全美富豪排行榜的首位，连美国前总统布什都赞扬他是地道的美国人，展现了创业精神，是美国梦的缩影……”某超市的万总补充说。

“是的，勤奋、创新是这位智慧商人成功的法宝。他的‘日落原则’、‘十英尺

态度’和‘三米微笑’等服务理念以及营销策略‘女裤理论’和‘啤酒与尿布’至今在商业界令人津津乐道。更令人难忘的是，本世纪初‘啤酒与尿布’简直就成了‘数据挖掘’的代名词。”徐教授继续说。

“啤酒与尿布，这两个风马牛不相及的东西怎么与数据挖掘扯上了关系？徐老师，快给我们讲讲吧！”移动公司的梁总有点着急了。



“1983年，当一般零售商还在进行信息化建设的时候，沃尔玛已经开始与休斯公司合作，花费2400万美元发射了一颗人造卫星，此后先后投入6亿多美元建起了电脑与卫星系统，还发明了条形码、无线扫描枪、计算机跟踪存货等新技术。借助于整套的高科技信息网络，沃尔玛的各部门沟通、各业务流程可迅速、准确地运行，数据库系统很快积累了海量的经营数据，包括大量的顾客消费行为记录。一年一度的圣诞节快要到了，沃尔玛人按照惯例又一次筹划节日的营销策略。这一次他们使用了一种新的‘购物篮分析’软件，对海量的顾客消费行为进行分析，一个意外地

发现让他们瞠目结舌，‘跟尿布一起购买最多的商品竟然是啤酒！’”

“这怎么可能呢？”有学员也感到疑惑不解。

“经过反复计算、核实，结论没有错。”徐教授答道。

“不过，这个故事告诉我们什么？”又有人问道。

“告诉我们数据挖掘可以发掘埋藏在海量数据中有价值的信息。”徐教授答道。

突然，后排有人大声说：“也告诉大家如果想喝啤酒，老婆不让买，就说去买尿布吧！”惹得大家哄堂大笑。

接着，徐教授问：“这是数据挖掘技术对历史数据进行分析得出的知识，这个结果符合现实情况吗？是否有利用价值？”

“还利用价值，真是六月里穿皮袄——反常！”有学员不以为然。

“紧接着，沃尔玛派出市场调查人员和分析师对这一结果进行了深入研究，证实它揭示了一条隐藏在‘尿布与啤酒’背后的美国人的一种行为模式：一些年龄在25~35岁的年轻父亲下班后经常要到超市去给婴儿买尿布，而他们中有30%~40%的人会顺手为自己买几瓶啤酒。”

刚才那位学员想通了，小声说：“对了，这是在美国，老外的行为模式与中国人就是不一样！证实了这样的发现是符合实际的，沃尔玛会怎么办呢？”

徐教授挥动了一下电子教鞭，大声说：“沃尔玛立即采取了行动，将卖场内原来相隔很远的妇婴用品区与酒类饮料区的空间距离拉近，使顾客更加方便。然后对本地区新生育家庭的消费能力进行了调查，对这两个产品的价格也做了调整，并向一次购买达到一定金额的顾客赠送婴儿奶嘴及其他小礼品，结果是尿布与啤酒的销售量双双大增。”

某超市的万总激动地站了起来，情不自禁地说：“不愧为全球零售业巨头啊，高招，值得借鉴！”

徐教授一边示意她坐下，一边说：“是的，不仅在零售业值得借鉴，这种‘购物篮分析’后来演变为‘关联规则分析’，并在其他行业发挥重大应用，我们 EMBA 班的学员有很多来自于工业界，下面再给你们讲一个工业生产中利用数据挖掘技术节约成本的故事。”

1.2.2 降低成本的绝活

徐教授：“工业界的学员都知道，派克汉尼汾公司是一家世界一流的工业企业，总部位于美国，于 1918 年由 Arthur L.Parker 先生创立。早在上世纪 70 年代已发展为全球控制领域最广、产品种类最完备的公司，年销售额超过 100 亿美元。大家估计下派克公司的年维修费用是多少？”



“200 万美元？”

“500 万美元？”

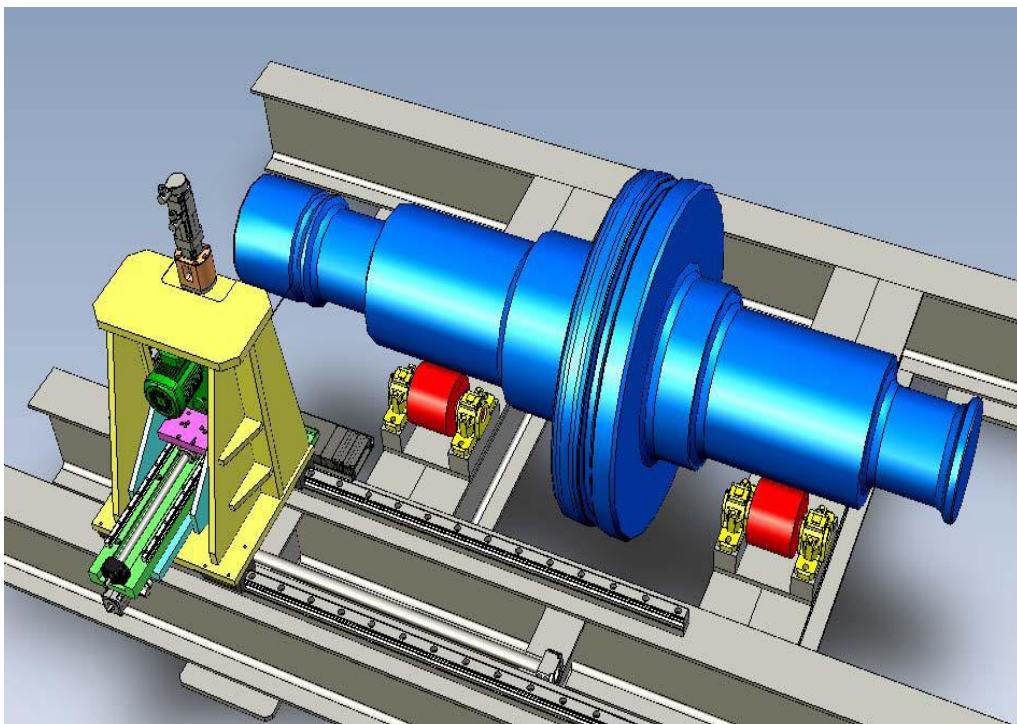
.....

“该公司产品出售后保修一年，年维修费用超过了一亿美元。” 徐老师说。

“我们鼓风机厂的年产值也比不上人家的年维修费。” 一学员喃喃自语。

“那怎样降低维修费用呢？” 徐教授问道。

“增加研发费用，提高产品质量！” 李部长抢先道。



“不错！但是如果我们假设在目前的技术条件下，产品质量已经达到了较高标准。还有没有其他办法？”

“这个……难道是数据挖掘？” 有一学员自语道，其他学员则低头沉思。

徐教授肯定地说：“是的，派克公司采用了数据挖掘方法。以一款干燥器为例，该机器 1200 多种零件中，常坏的贵重零件约 20 种。应用数据挖掘的关联规则分析发现这些价格昂贵的零件的寿命竟然大多数与少数几种便宜零件的磨损有关。”

李部长激动了：“妙，妙极了。采用常更换便宜部件，达到延长贵重部件的使用寿命，就可以大大地降低维修成本。我们怎么就想不到呢！”

徐教授看着李部长，说道：“对了，派克公司采用了这样的策略后，仅在这干燥器这种产品上，每年节省维修费高达上千万美元。”

李部长坐不住了，大声说：“我们公司的不锈钢生产线也有同样的问题。徐老师，您指导我们也挖掘挖掘吧！”

徐教授：“别着急，李部长。有很多数据挖掘方法能够解决你们公司生产管理、新产品设计、产品质量控制、能源分析、原料搭配、成本分析等许多问题，以后我们再进一步讨论。”

大家越来越坚信数据挖掘的巨大威力，精神也更加集中了。

1.2.3 出奇制胜的小纸条

徐老师接着说道：“我们在座的学员大部分喜欢看足球比赛，我再给大家讲个数据挖掘在体育方面应用的故事。”

这时，PPT 上出现了一个章鱼，光笔的红点在它身上晃动，徐教授问道：“上届世界杯的时候名噪一时的‘章鱼帝’大家还记得吧？”

“出道两年的章鱼保罗在 2008 欧洲杯和 2010 世界杯两届大赛中，预测 14 次猜对 13 次、成功率 92%，堪称不折不扣的‘章鱼帝’。”足球迷李部长先吐为快。

徐老师补充道：“从科学的角度来看，章鱼帝的预测仅是小概率事件在万众瞩目下发生了而已。但是 2006 世界杯同样是德国和阿根廷的赛场上，不是章鱼保罗救了

德国，而是一个神秘的小纸条。”

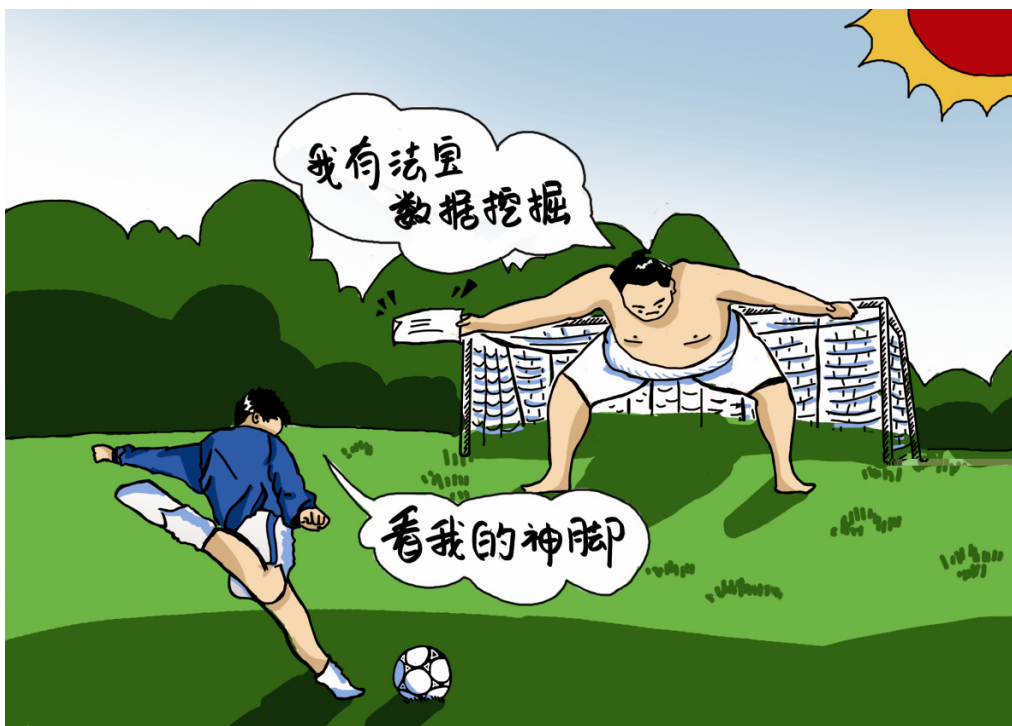
“一个小纸条有这么大的作用，到底是什么小纸条啊？徐老师您赶紧给我们讲讲吧！”有人急不可待。



徐教授不紧不慢地说：“2006 年世界杯上，阿根廷和德国在 1/4 决赛中 120 分钟难分高下，在点球大战之前，老门将卡恩将一张纸条递到莱曼手中。莱曼每次扑点球之前都要看一眼纸条。结果是，莱曼所有点球都判断对了方向，除了两个点球质量太高无力回天外，其他全部扑出，阿根廷只能黯然出局。”

“那纸条上到底写着什么锦囊妙计？”

“写着德国胜！哈哈，可惜章鱼保罗还没出生。”台下哄笑一堂。



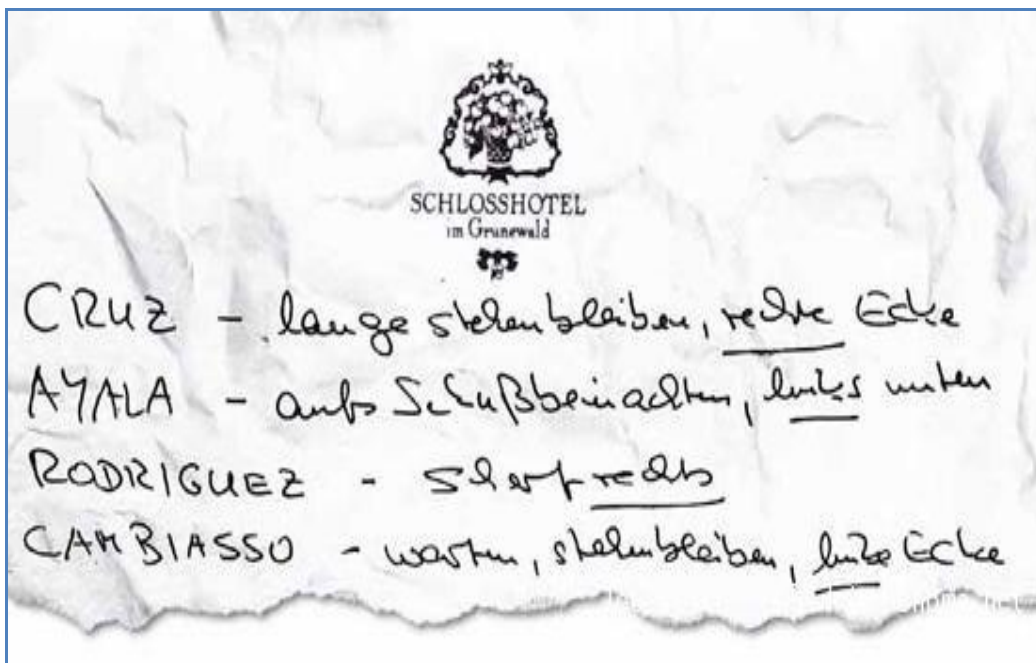
徐教授摆了手势，让大家安静，继续说道：“上面记录着阿根廷队的克鲁兹、阿亚拉、罗德里格斯以及坎比亚索习惯的脚法。德国队守门员教练科普克如此精确地预测出阿根廷球员射出的点球方向，并不是他有什么过人的占卜天才。那张草草写在格鲁内瓦尔德皇宫酒店便笺上的扑点球秘籍，来自于德国科隆体育学院数据分析小组夜以继日的努力。”

“点球就是点球了，纯技术问题，有什么可分析的嘛？”足球迷李部长不以为然。

徐教授：“这个问题问得好。分析小组的人员收集了阿根廷队 13000 个点球的录像，所有这些采集回来的点球数据被输入数据库中，并根据阿根廷射门练习的数据找出了一些可以描述射门动作的行为特征，最终从这些特征中提炼出很少的更具体特征。大家说说点球动作行为特征可以分为几类？”

“两类，进球和没进球！”某人的幽默引来全班大笑。

徐教授补充道：“这些特征被描述为：阿亚拉，短助跑，右下角；里克尔梅，斜向助跑，右下角；马克西，长距离助跑，左上角；坎比亚索，长距离助跑，右侧；索林，短助跑，右下角；特维斯，短助跑，中路……。这些特征描述了阿根廷队谁罚点球、怎样罚点球的规律。正是这张小纸条把大力神杯交到了德国队手中！小纸条上总结的这些规律是数据挖掘的结果！”



某省鼓风动力集团的王总快人快语：“数据挖掘可太有用了。徐老师，您快给我们讲讲什么是数据挖掘吧。”

这时，下课铃响了，徐教授示意大家休息。

1.3 什么是数据挖掘？

新的一节课开始了，徐教授走上了讲台，清了清嗓子，声音更加洪亮：“随着计算机技术、数据库技术、传感器技术和自动化技术的飞速发展，人们获取数据、存储数据变得越来越容易。这些数据不是人为产生的，是对我们所研究对象隐含的一定规律的反映。数据挖掘的目的就是要从所获取的数据中发现这种规律性的知识，从而帮助企业在他们的数据仓库中找到最重要的信息，预测未来趋势和行为，使得商务和生产活动具有前瞻性，并作出具有知识驱动的决策。”

徐教授将 PPT 翻回到数据挖掘的故事，继续说：“通过上节课所讲的三个故事，相信在座的同学对数据挖掘有了初步的认识。那么到底什么是数据挖掘呢？大家可以发表下自己的观点。”

学员们你一言，我一语，争先恐后。

“数据挖掘就是从数据中发现有价值的信息的技术。”

“数据挖掘是对数据建立模型，通过算法求解而发现隐藏在数据中的知识的一种手段。”

“.....”

徐教授总结道：“大家对数据挖掘的认识都值得表扬，不过表述得都不够全面。”说着，徐教授敲了一下键盘，说：“请看大屏幕，这才是最权威的数据挖掘的定义。”

数据挖掘（Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用信息和知识的过程。

大家认真地看着屏幕的内容。

片刻之后，有学员问道：“数据量小是不是就不能进行数据挖掘了？”

徐教授答道：“实际上数据挖掘的算法大都是建立在统计学**大数定律**基础上的。数据量太小，常常无法反映出真实世界中的普遍特性，这样挖掘算法得出的结论自然不可靠。但并非小数据量就不可以进行挖掘，近年来研究者也提出了一些对小样本进行挖掘的方法，如支撑向量机方法就是基于小样本学习理论的非常实用的方法。数据量虽小，但数据总是事物特性一定程度的反映，只要建立的模型和算法得当，当然也可以从这些数据中获取一定的信息。”

“那么是不是数据量越大越好？”有学员问。

“从理论上说，应该是这样。但随着数据量的增大，算法执行效率会越来越低，甚至无法计算。”徐教授回答说。

刚才提问的学员点了点头，接着问：“徐老师，数据挖掘的定义中，数据前面还有那么多的修饰，您还是给我们解释解释吧。”

“大家淡定点，‘不完全的、有噪声的、模糊的、随机的’确实有点绕口，现实中经常会碰到这种数据。例如，问卷调查时发现不少人不填婚姻状况和年龄，这些**不完全的或缺失的**数据会给数据挖掘带来一定的难度，我们要么干脆删除这些样本或记录，要么选择使用一定的方法将这些缺失数据补上，或者选择使用可以自动处理缺失数据的算法。”说到这儿，徐教授端起了茶杯，说自己也要补充一下水分了。

“那噪声是什么意思？”一个学员问。

徐教授合上茶杯盖子，一边狠狠地用杯子连续敲击着桌子，一边说：“对于我讲课的声音来说，敲桌子的声音就是噪音，我们的录音机录到的是我的讲话声和敲桌子声混杂在一起的混合声波数据。”

“我明白了，由于异常情况的干扰，使我们获得的数据偏离了真实值，这样的数据就是噪音数据。”刚才提问的学员说。

“不光是外界的干扰，测量仪器的故障、人工输入或抄写时的失误等都可能形成**噪音数据**，可见实际问题中噪音数据往往难以避免。”徐教授进一步解释说。

“徐老师，什么是模糊的、随机的数据？”又有一学员问。

“在数据挖掘过程中，我们不可避免地要涉及事物的不确定性。不确定性包括模糊性和随机性。**模糊性**则指事物本身从属概念的不确定性，**随机性**是指事件发生与否的不确定性。”

“太抽象了，徐老师，您给我们举个例子吧！”李部长建议说。

“好吧。其实模糊的数据大家平时经常见到，比如说张三个子很高，李四个子较矮，个子的高矮就是典型的模糊性概念，到底多高才算高，李部长 1 米 80，对一般人来说算高个子，但跟姚明比，就太矮了。随机数据也极为多见，比如说超市啤酒每天的销量显然是不确定的，大部分人买啤酒是在超市转悠时临时决定的。”徐教授回答道。

李部长扶了扶眼镜，支支吾吾地说：“我似乎明白了……”

本科应用数学专业毕业的王总快人快语：“李部长，我借给你《模糊集的应用》和《概率统计》两本书，看看你才会真正明白。我要问新的问题了，徐老师，数据挖掘的目的是从数据中发现新的信息和知识，那挖掘出来的知识是什么？”

徐教授回答道：“挖掘出来的知识就是‘散落的珍珠’，亦或是‘发光的金子’，它的实际决策价值非凡。知识是通过对数据进行深入地归纳、分析而获得的，是对所研究对象更深层次的认识。知识是隐藏在数据中的关于所研究对象的一种规律性，比如可以用来预测的数学模型、‘如果……那么……’这样的规则、描述事物的类别、有价值的模式、所研究对象的结构、研究对象与对象之间的关系等。”

1.4 历史的必然

EMBA 教室的座位是半弧形的，中间有通道，老师讲课时部分时间是站在学生中间的，课堂上师生交流非常方便。

“人类走过了石器时代，纸器时代，磁器时代，直至现在的网络技术时代和正在跨入的物联网时代，这些智慧、文明的结晶是怎么样代代相传，生生不息地保留和继承下来的呢？”徐教授问。

“信息获取……”

“信息存储……”

“信息查询……”

“信息的加工和应用……”

旁边的学员们陆陆续续地表达了自己的看法。

“对，确实是这样。人们通过信息的获取、存储与查询、加工和应用几个环节实现知识传播、继承和发展。”徐教授对学员们的回答很满意。

随后，徐教授通过 PPT 展示了一个图，并讲述了伴随着人类历史文明发展和进化的长河，人们对知识和信息的存储、加工应用的演化进程。



“从人类有了获取信息的能力开始，便不断对信息进行归纳总结。大家想想，有哪些谚语可以说明，古人就开始针对观察到的信息进行分析和归纳了？”徐教授刚问完，谚语大接龙便开始了。

“连发三日东北风，定有大水后面跟。”

“天上起了泡头云，不过三天雨淋淋。”

“星光闪闪如动摇，大雨下得没处逃。”

“……”

课堂气氛一下子变得十分活跃，大家在说起古人智慧的时候，都觉得万分光荣和自豪。

“通过祖祖辈辈的观察、积累与归纳，人们发现了自然现象与天气的‘关联规则’”，徐教授总结说。

突然，第一排的一个学员站起来说道：“对于一些简单的自然现象，可以通过归纳提取形成经验知识，但现实世界太多的复杂问题，数据量极大，已经远远超出了人

脑可处理的范围。”

他旁边的一位学员也感慨地说：“是的，现在获取数据非常容易，就拿我们钢铁公司来说，每日产生的数据超过3Gb，要是将这些数据放在我的脑子里，脑瓜肯定爆炸了，更不用说处理、归纳得到知识了。”

看看他憨憨的笑容，大家都被逗乐了，之后便都陷入了沉思。

“不是有计算机么，人就不用操那么多心了。”另外一个学员小声说。

于是，徐教授解释说：“上世纪60年代，尽管有了计算机，但对数据是以零散文件方式进行管理的。我们能够收集、存储、处理如此海量的数据，归功于20世纪70年代IBM发明的关系式数据库和SQL查询语言。在此基础上通过计算机和网络进行联机事务处理（OnLine Transaction Processing, OLTP）可以对管理信息进行日常操作并及时、安全、高效地存储数据，这样便引发了数据爆炸式地增长。”

电信公司冯总，计算机专业硕士，在单位负责数据仓库建设，听到这里，话匣子关不住了：“OLTP关心的只是业务操作，只对当前数据感兴趣。其实信息处理的目的是为人们提供决策支持，这就需要对历史数据进行大量地分析处理。对历史数据的分析，往往导致系统进行长时间运行，严重影响日常数据实时操作，这就要求把分析性操作及其相关数据从事务处理环境中提取出来，按照决策支持的需要进行重新组织，建立单独的分析环境。”

李部长这几年读了不少信息处理方面的书籍，他接上了话茬：“为了满足这种需求，W. H. Inmon 于1993年出版了‘Building the Data Warehouse’，从此数据仓库（Data Warehouse）隆重登场，W.H.Inmon也当之无愧地成为数据仓库之父。他给出了数据仓库定义：‘数据仓库是一个面向主题的、集成的、随时间变化的、持久的数据集合，用于支持管理层的决策过程’。在数据仓库产生的同时，联机在线分析（OnLine Analytical Processing, OLAP）出现了，它是一种具有对数据进行汇集、合并和聚集以及从不同角度观察信息的分析技术。”

电信公司冯总，在单位里被誉为数据仓库专家，继续说：“通过OLAP技术可以

对从数据库或数据仓库得到的经验、规则进行验证，当然也可以对数据挖掘结果的有效性、可行度进行检验、完善。然而，数据库和数据仓库越建越大，通过直观的感觉、简单的统计分析和OLAP技术并不能发现隐藏在数据中有价值的信息和知识。”

“上世纪80年代末到90年代初，广泛流传着一句耐人寻味的话‘我们沉浸在数据的海洋中，但却渴望着知识的淡水’，这句话生动地描绘了人们面对海量数据的迷惘和无奈。”徐教授深沉地说。

突然，徐教授抬高了嗓门：“一石激起千层浪，这时沃尔玛演绎了一场‘啤酒和尿布的故事’，它使人们看到了数据分析的希望，擂起了数据挖掘的战鼓，一场数据挖掘的风暴开始了……”

几个学员抑制不住内心的激动，你一言、我一语地表达自己的观点：

“商业界发现了沃尔玛迅猛发展的密招，纷纷效仿。”

“电信行业沸腾了，各公司纷纷争先恐后地利用数据挖掘这一锐利武器解决他们面临的最紧迫的问题，如客户分群、客户流失原因及预测、业务套餐及响应、关联消费等。”

“工业界也着急了，他们的数据堆积如山，期望从中挖掘出金子，指导生产和管理。”

“科学界大批科研工作者聚焦于数据挖掘，紧锣密鼓地投入到该新生领域的研究。”

“……”

徐教授走上讲台，总结道：“人常说，‘需求’是成功之源。商业管理、生产控制、市场分析到工程设计、科学探索等将堆积如山的数据资源转换为信息和知识的巨大需求，促使着数据挖掘技术的飞速发展。九十年代中期以后，基于数理统计、人工智能、机器学习、神经网络等多种技术，关于数据挖掘软件的开发和应用成为热点。”

徐教授的话音刚落，有学员便问道：“徐老师，您一会儿说数据库中的知识发现，

一会儿又用数据挖掘，我真不知道它们之间的关系。”

“2008 年我在李部长他们钢铁公司作数据挖掘报告，也有几个人问我同样的问题。在 1989 年 8 月第 11 届国际人工智能联合会议上，数据挖掘以数据库中的知识发现（Knowledge Discovery in Database, KDD）第一次正式亮相。从此以后，**数据挖掘（Data Mining）**和数据库中的知识发现（KDD）互为别名，但后来数据挖掘渐渐被多数人喜闻乐道。”徐教授回答道。

刚才那位提问的学员是 EMBA 班里有名的“问到底”，继续穷追不舍：“徐老师，数据挖掘是在什么时候被大家普遍接受的呢？”

李部长急了，站了起来：“‘问到底’同学，你是不是一定要考倒徐老师！”

徐教授赶紧解围：“这个问题已经难以考证，大约在上世纪 90 年代开始，数据挖掘占了上风，其中还有一段趣事。”

“问到底”顾不上理会李部长，高兴地说：“徐老师又要讲故事了。”

徐教授示意李部长坐下，笑着说：“其实学院派最初一直沿用数据库中的知识发现即 KDD。在一次 KDD 国际会议中，委员会曾经展开讨论，到底使用 KDD，还是 Data Mining。”

“问到底”急切地说：“肯定一致同意使用 Data Mining。”

徐教授摆了摆手道：“会议上大家争论不休，讨论了两个小时没有结果。要是你们是当时参会的专家，会怎么定这个名字？”

“抓阄”

“抛硬币”

“听会议主席的”

“……”

学员们也开起了玩笑。

徐教授说到：“呵呵，我们中国人喜欢举手表决，外国人也兴这一套。会议主席最后决定投票表决，结果很具有戏剧性，一共 16 名委员，其中 8 位投票赞成 KDD，另 8 位赞成 Data Mining。”

“问到底”露出一副为难的表情：“这可怎么办呢？”

徐教授答道：“事实上，根据当时会议的记录，最后一位元老站出来说‘数据挖掘这个术语太为土气，科学研究就是要获得新的知识’。”

“问到底”感到有些失望：“老奸巨猾，跟没说一样！”

“怎么跟没说一样？他作了双重肯定。于是在科研界便继续沿用 KDD 这个术语，而在商用领域，因为‘数据库中的知识发现’显得过于冗长，就普遍采用了更加通俗、简单的术语‘数据挖掘’。”

1.5 数据挖掘能干什么？

要讲数据挖掘的功能，大家都非常感兴趣。徐教授提高了嗓门：“前面我给大家讲了数据挖掘的三个故事，并给出了数据挖掘的定义，还简要地回顾了一下数据挖掘产生的过程，可数据挖掘到底能干些什么呢？”

“购物篮分析”

“用户分群”

“客户流失分析”

“服务套餐设计”

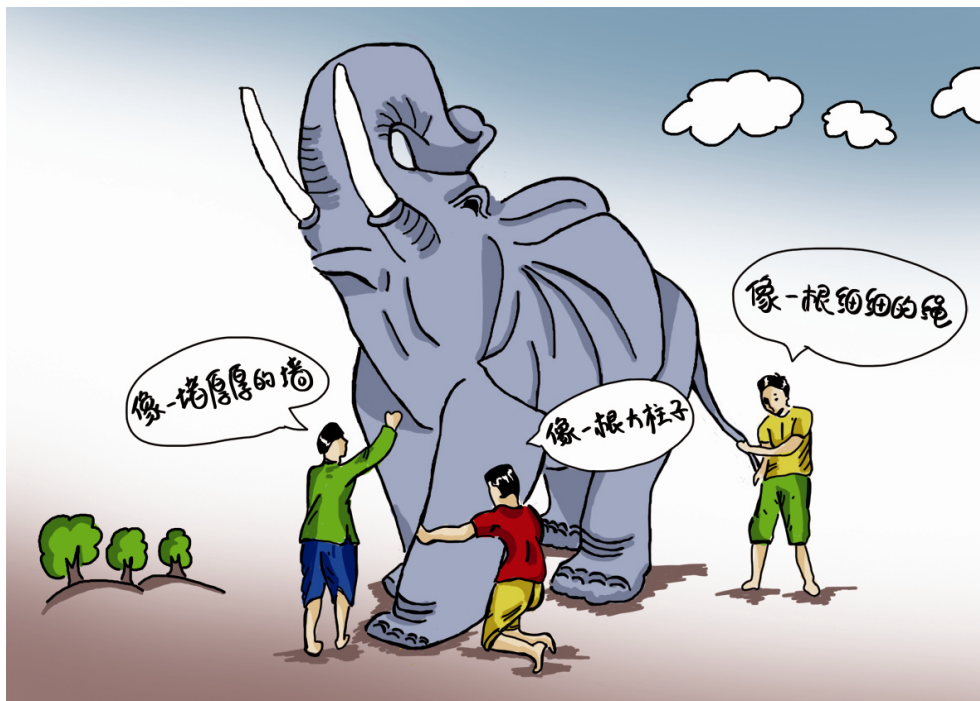
“预测”

“.....”

学员们纷纷根据自己的直观理解回答着。

“大家所说的只是根据我前面讲的内容概括了数据挖掘的一些功能。有个成语叫做‘盲人摸象’，我才领着大家摸了大象的一条腿而已，哈哈。”徐教授开玩笑。

“徐老师，在座的学员大部分是政府部门和大中型企业的头头脑脑，我们首先希望知道数据挖掘到底能够干什么，至于怎么干那就是工程师的事了。您就先概括一下数据挖掘的功能吧。”高新区的段主任建议说。



徐老师：“好吧。概括地说，数据挖掘的功能主要包括关联分析、聚类分析、分类、回归、时间序列分析和偏差甄别等，下面我们分别介绍这些功能。”

1.5.1 关联 (association) 规则挖掘

1.5.2 聚类

1.5.3 预测

1.5.4 序列和时间序列

1.6 数据挖掘工具