

NEWTON
QUANT
牛 顿 量 化

量化投资方法丛书
主编 刘振亚

30%

50%

20%

刘振亚 李伟 / 编著

金融数据 挖掘

以生动简明的理论陈述和案例演示
细致地介绍新颖的金融数据挖掘方法
抛砖引玉，推动金融数据挖掘系统化的研究



中国经济出版社
CHINA ECONOMIC PUBLISHING HOUSE

NEWTON
QUANT
牛 顿 量 化

量化投资方法丛书
主编 刘振亚

刘振亚 李伟 / 编著

金融数据 挖掘

以生动简明的理论陈述和案例演示
细致地介绍新颖的金融数据挖掘方法
抛砖引玉，推动金融数据挖掘系统化的研究



中国经济出版社
CHINA ECONOMIC PUBLISHING HOUSE

• 北 京 •

图书在版编目 (CIP) 数据

金融数据挖掘 / 刘振亚, 李伟编著. —北京: 中国经济出版社, 2016. 9

ISBN 978 - 7 - 5136 - 4248 - 4

I. ①金… II. ①刘… ②李… III. ①金融—数据处理 IV. ①F830. 41

中国版本图书馆 CIP 数据核字 (2016) 第 109334 号

责任编辑 杨 莹

责任审读 贺 静

责任印制 巢新强

封面设计 任燕飞

出版发行 中国经济出版社

印 刷 者 北京科信印刷有限公司

经 销 者 各地新华书店

开 本 710mm × 1000mm 1/16

印 张 15. 5

字 数 206 千字

版 次 2016 年 9 月第 1 版

印 次 2016 年 9 月第 1 次

定 价 68. 00 元

广告经营许可证 京西工商广字第 8179 号

中国经济出版社 网址 www.economyph.com 社址 北京市西城区百万庄北街 3 号 邮编 100037

本版图书如存在印装质量问题, 请与本社发行中心联系调换 (联系电话: 010 - 68330607)

版权所有 盗版必究 (举报电话: 010 - 68355416 010 - 68319282)

国家版权局反盗版举报中心 (举报电话: 12390)

服务热线: 010 - 88386794

前 言

数据挖掘 (Data Mining) 是人们在早期对该学科的称呼, 现在, 学界将其称为 “统计学习” (Statistical Learning)。这可能主要是由于早期很多不当的应用, 使得数据挖掘的声誉受到很大的损害。但鉴于国内目前的流行叫法, 本书仍采用数据挖掘之名。

统计学习是指对复杂数据集合进行分析和建模的工具。它结合了统计学和计算机科学 (尤其是机器学习) 的最新发展。这个领域的覆盖面很广, 包括很多方面, 比如岭回归 (LASSO Regression)、分类回归、回归树以及支持向量机 (SVM) 等。

近年来, 随着互联网的迅速发展, 如何处理大数据 (Big Data) 是人们急需解决的问题。因此, 统计学习近年来成为人们十分关注的研究领域。基于这些大数据, 统计学习很快在市场营销、传统金融行业 (银行、保险) 等诸多领域得到广泛应用, 并取得了很好的效果。因此, 很多人需要了解和掌握统计学习领域的一些基本方法和模型。

有关统计学习最早的一本书是由斯坦福大学统计系的 Hastie、Tibshirani 和 Friedman 三位教授合写的《The Elements of Statistical Learning》(简称: ELS), 我曾指导我的博士生认真研读此书。该书对数学

的要求高，主要是写给那些有较好数学背景的人士。另外一本是 2014 年底 James 等人合写的《An Introduction to Statistical Learning》（简称：IEL）一书。当然，还有很多其他很好的有关统计学习的参考书，这里就不一一列出。

本书是根据我这些年为中国人民大学财金学院硕士研究生授课的讲稿整理、精简而成。我的博士生李博为此书的编辑出版做了很多工作。

本书是一本入门性读物，书中的例子也十分简单，目的是向广大读者介绍统计学习的主要方法。这里需要特别指出的是，由于投资行为受政策、环境、市场结构等诸多因素的影响，因此，对此类数据进行建模分析的难度极高，希望本书能够起到抛砖引玉的效果。

由于作者水平所限，本书的缺点和错误之处在所难免，恳请广大读者批评指正。

刘振亚

于英国伯明翰大学 J. G. Smith 楼

2015 年 6 月

量化投资与现代科技

——量化投资系列图书付梓之际

量化投资被广泛应用在国际对冲基金中。在过去的三十年里，由于计算机技术和统计分析技术的进步，量化投资得到了迅猛的发展。

作为量化投资基金中的杰出代表，数学家西蒙斯（Jim Simons）所领导的复兴技术（Renaissance Technology）可谓独树一帜——他的大奖章基金在1988—2008年的20年时间里创造了年均收益35.6%的奇迹。1958年，20岁的西蒙斯从麻省理工大学数学专业本科毕业后，转入加州大学伯克莱分校攻读数学博士，1961年博士毕业后回母校麻省理工大学任教。一年后，他跳槽到哈佛大学任教，1964年进入美国国防分析研究院工作。1967年，西蒙斯出任纽约大学石溪分校数学系系主任。在此期间，他与著名华裔数学家陈省身合作，创造了著名的陈—西蒙斯理论。1976年，他获得美国数学学会的威布伦奖。1978年，他离开石溪大学，成为职业投资人。1988年3月，西蒙斯成立复兴技术公司。

除西蒙斯外，复兴技术的三位元老：Leonard Baum、Henry Laufer和James Ax都是一流的数学家，对复兴技术的长期发展产生了很大的

影响——Baum 是西蒙斯在国防分析研究院的同事，统计学中著名的 Baum - Welsh 算法的发明者，该算法被广泛应用于隐蔽的马尔可夫链、语音识别、生物和应用统计中。James Ax 也曾任石溪大学数学系主任，在数论和几何学领域造诣颇深；他曾在 1967 年获得美国数学学会数论方面的科伦奖。Henry Laufer 也曾是普林斯顿大学数学教授，退休后在石溪大学创办了量化投资专业，专业课程包括：概率论与统计方法、线性规划、线性几何、数据分析、随机微积分、金融计量、最优化算法、资产定价、投资组合、金融市场、衍生品定价、固定收益产品等。

当然成功的量化投资基金，除数学家西蒙斯所领导的复兴技术外，还有计算机教授肖尔（David Shaw）领导的肖尔公司，物理学家哈丁（David Harding）领导的元盛资本（Winton Capital），经济学家格里芬（Kenneth Griffin）领导的大本营投资（Citedal）等。

一些国际知名的对冲基金对人才的要求也可归纳为以下内容：很强的电脑编程能力（C，C++；API，FIX；R，Matlab，SAS 等）；很强的数学或统计学分析技能（线性和非线性时序分析，数据挖掘，隐蔽马尔可夫链，随机分析等）；很强的大型数据库处理能力；对衍生品、资产定价、市场微结构等有深入了解。

由此可以看出，量化投资是现代金融理论、现代数理方法和现代信息技术的综合体。所涉及的金融理论主要包括：资产定价，投资组合，衍生品定价，市场微结构，行为金融学等；信息技术主要包括：编程技术（C，C++），数据库技术，交易底层通信技术（API/FIX）；数理方法主要包括：经济计量分析基础，线性和非线性时序分析，数据挖掘，隐蔽马尔可夫链，随机分析等。

现代金融理论为量化投资提供了科学的理论基础。资产定价、衍生品定价、行为金融学和市场微结构是量化投资策略设计科学基础；资产组合理论是量化投资降低和控制风险的有利工具。

现代信息技术和计算机技术为量化投资提供了坚实可靠的工具。编

程技术和数据库技术是量化投资策略程序化的基础；API/FIX 是交易底层自动化的基础。以上两者的结合，使得量化投资能够实现完全自动化交易。有人曾说，即使西蒙斯将其所用策略公众于世，能够把公式变成钱的人在全球范围仍然是屈指可数。可见电脑技术和通信技术在量化投资中的重要作用。

数理方法是量化投资最为核心的部分，也是数据分析和交易策略设计以及评估的基础。其所涉及的领域也是五花八门、仁者见仁、智者见智，主要包括：经济计量分析基础，线性和非线性时序分析，数据挖掘，隐蔽马尔可夫链，随机分析，小波分析等。西蒙斯曾经说过，如果抛弃了数学模型，将对我们没有任何好处。

在过去近十年的时间里，作者本人作为摩根大通期货公司（JP Morgan Futures Co.）的董事，有机会接触到一些国际知名的对冲基金，并对其进行深入的研究和合作，元盛资本（Winton Capital）就是其中的一个。作为全球最大的管理期货（期货对冲基金），元盛资本管理着约 300 亿美元的资产，在其创始人科学家哈丁先生的领导下，在 1987 至今这 28 年的时间里创造了 16% ~ 17% 的年均收益。仅从年均收益的数字来看，元盛资本不如复兴技术可观，但大家知道，复兴技术的大奖章基金的规模只有约 50 亿美元，而元盛资本的规模是 200 多亿美元。

元盛资本现有员工 200 多人，近 50% 为研究人员，远离热闹非凡的一线业务，专心从事研究工作。若简单的按照国际对冲基金 2% ~ 20% 的收费标准，元盛资本一线人员创造了人均产出近亿元人民币的奇迹。可见，当今现代科技与金融的结合能够创造出巨大的财富。量化投资行业绝对是高科技产业中的佼佼者。

在过去的三十多年里，我一直从事计量经济方法和信息技术方面的学习和研究，本丛书也算是本人过去多年学习和研究的小结。这里，首先要感谢我的大学老师们：财金学院的黄达教授（我的博士生导师，前校长）、陈共教授（前财政系主任）、王传伦教授，经济学院的杜厚文

教授（前国际经济系主任，前副校长），信息学院的陈禹教授（前院长）、方美琪教授（我的本科导师）、魏权龄教授、严颖教授、王鄂芳教授（北京大学数学系）。感谢我的美国老师们：George Horwich 教授（美国普渡大学，我的博士后导师），Roger Gordon 教授夫妇（加州大学圣地亚哥分校），Mark Machina 教授（加州大学圣地亚哥分校），Kajal Lahiri 教授（纽约大学阿尔巴尼分校），Harold Watts 教授（哥伦比亚大学）。他们的教育使我领略了数学的美妙、计算机的精巧、计量经济的严谨以及金融市场的奥妙，为日后的量化投资研究奠定了坚实的金融、经济、数理和计算机基础。

非常感谢中国人民大学校长陈雨露教授，研究生院常务副院长吴晓求教授，财金学院院长郭庆旺教授，副院长赵锡军教授，他们多年的友情、宽容和理解使得我能够有充足的时间，在中国和英国两地从事我所感兴趣的项目研究。我也非常感谢我在伯明翰大学商学院的同事们：David Dickinson（前院长）、Nicholas Housewood、Robert Elliott、Toby Kendall、Alessandra Guariglia 和 William Pouliot。

感谢摩根大通期货的同事们：周小雄先生、Nicholas Forgan、Richard Shen、Sandie O' Connor、Brandon S. Konigsberg、Emily Portney、Jerome Kemp、Richard Berliand、Joseph Regan、Zili Shao。尤其感谢摩根大通期货公司董事长周小雄先生多年无微不至如兄长般的关照。感谢元盛资本的创始人 David Harding，元盛亚洲 CEO Charles Allard、Kurt Settle 和田野先生。与元盛资本的合作，尤其是元盛资本每年的年会使我眼界大开，受益匪浅；感谢 AHL——牛津量化金融研究院（AHL - Oxford Institute），使得我有机会参加他们所举办的具有世界一流水平的有关量化投资的研讨会。

最后，感谢我在中国人民大学和英国伯明翰大学所指导的学生们，尤其是人民大学财金学院硕士生和实验班的学生们。根据我的讲稿，他们帮助我整理了这套丛书中的很多内容。尤其是我的博士生杨武（现

任教于中央财经大学)、唐涛(现任职于中国人民银行研究所)、罗涛(现任职于国家审计署)、邓磊(现任职于北京工商大学)、陈宇(现任职于中央财经大学)、张庆雪、李伟和李博。我在伯明翰大学的博士生曹瑞玫、王仕炫,以及张昆(阿斯顿大学),他们对量化投资的深入研究督促着我不断学习新的知识。这些学生们对新知识的追求,使得我多年来不敢懈怠,不断学习。从他们身上,我看到了中国量化投资业的未来。

出版量化投资系列图书的主要目的是为国内的投资者系统地介绍有关量化投资理论、技术和方法,国际上成功的量化对冲基金公司,以及量化投资的研究成果和量化投资产业的发展动向。本系列图书内容主要包括:经济计量分析基础,投资组合理论与实践,时序分析与神经网络,金融数据挖掘,基于隐蔽马尔可夫链的时序分析方法,随机分析,小波分析等。

我深切地希望,此套丛书能够为中国的证券、基金、期货、私募以及个人投资者提高量化投资水平起到抛砖引玉的作用。

写这个总序前后花了我三四个月的时间,从北京写到香港,从中国香港写到英国,从英国写到法国,最后又从法国写回北京。每次提笔都是千言万语涌上心头,本人最大的心愿就是在未来20—30年时间里,中国能够出现复兴技术和元盛资本这样世界一流的对冲基金。

为此,我愿奉献一生!

刘振亚

2013年9月于北京中国人民大学

导 言

数据挖掘虽然的确是和数据打交道的，但这种挖掘过程倒不见得一定是枯燥无味的，相反，它可能是一个能令人感到回味无穷、点石成金的奇妙过程。

“二战”前夕，德军秘密逮捕了一位作家^①，因为他们怀疑他是外国的谍报人员，证据就是该作家出版的一本书。这本书详细披露了关于德军组织结构的许多信息，甚至包含军队各级组织的番号、驻地、人员姓名、年龄、履历和装备部署，而且其准确程度也非常高。在德军对其审讯时，该作家的回答却令审讯人员大跌眼镜，他说：“这些都是从你们每天公开发行的报纸上抄来的。”审讯人员不信，遂找来了一大堆旧报纸，对照书中披露的内容一一进行查对，结果从报纸上找出了全部原始数据来源（当然，有些信息是报纸上披露出来的，但更多的详细信息则是他以报纸的公开信息为基础“挖掘”出来的）。最终，德军只能将他无罪释放。在这个案例中，德军公开报纸上的信息是数据源，是数

^① 这位作家名叫伯尔托尔德·雅各布，是一位瑞士作家，其于1935年3月出版的《战斗情报》一书详细披露了当时德军的组织结构状况。

据挖掘的基础，而该作家对公开信息的分析整合便是“挖掘”，他使用的分析方法便是“挖掘方法”，而得出的德军组织结构和军事部署的相关详细信息便是“挖掘结果”，是非常有价值的信息。

唐朝的刘晏是一位古代财经实务方面的专家，他利用几十年间储备粮食的收购价格和数量的统计资料，“挖掘”出了名为“常平、均输”的一整套方法来指导各州县制订购粮或卖粮计划，从而革除了以往存在的政府购粮或卖粮总落后于市场上的富商巨贾的弊病，使得政府可以主动调控市场粮价，而不必再受制于富商巨贾们的“投机炒作”。这个案例中，几十年间的粮食价格和粮食数量的数据是“挖掘”的基础，而刘晏制定出的政府购粮或卖粮的规则便是“挖掘成果”，而挖掘的方法则是一种函数映射，即每个粮价对应一个购粮或卖粮数量。而这种函数映射背后所隐含的则是粮食市场实际价格和均衡价格的差距引起粮食实际供需数量之差的经济学基本规律。这种“数据挖掘”收到了极好的效果，那就是：政府对粮价的调控取得了主动权，并具有科学依据，从而维护了市场粮食价格的稳定，真正做到了“谷贱则粢，谷贵则粢”，使“天下无甚贵贱，而物常平”。

在文学方面，曾经有关于《红楼梦》这本书作者的一个争论。《红楼梦》前八十回是由曹雪芹所著这一点，几乎可以说是在学者中已经达成了共识，但对于后四十回是由谁所著却产生了广泛的争论。有些人认为，《红楼梦》后四十回是由高鹗创作的，但有些人则认为后四十回是由曹雪芹的遗留书稿编纂而成。一时间，两种观点的支持者们各自搜寻着相关证据，以求证明自己的观点才是对的。大部分的争论者，无论持有哪种观点，几乎都是从文学或历史的角度来进行考证，从而往往因文学或历史资料的某些缺陷而出现“公说公有理，婆说婆有理”的局面。然而，有位学者则突发奇想，他摒弃了常用的文学或历史角度，而是从数据挖掘的角度入手，试图从《红楼梦》书稿中挖掘出其作者之谜的谜底。

他是怎么“挖掘”的呢？首先，将《红楼梦》权威古籍版本的书稿作为研究对象；接着，选择书稿中最常用的若干（如10个）语气词（字），如之、乎、者、也，使这些语气词或字的使用与小说内容的展开几乎没有联系，而只与作者的使用习惯有关系；然后，以这些语气词（字）在每章回中出现的、标准化后的频数作为指标，使用数据挖掘方法中的聚类分析方法，对全书一百二十回进行聚类分析；最后，得到的结果是，书稿被分为了两类，一类由前八十回组成，另一类由后四十回组成。这一挖掘结果表明，前八十回与后四十回应当是由两位不同的作者所写，从而间接地支持了后四十回是由高鹗所写的观点。

通过以上这些例子，我们可以直观地认识到：无论是在战争年代，还是和平时期，数据挖掘都是十分重要的；无论是在军事方面，还是在经济、文化方面，数据挖掘技术都具有很高的实用价值。

一、数据挖掘的定义、特征及方法

（一）数据挖掘的定义与特征

关于“数据挖掘”这个概念，一个广泛被接受的定义是：“数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的，但又潜在有用的信息和知识的过程^①。”但是，这种定义没有阐释挖掘的过程及方法。下面我们从数据、挖掘过程及方法、挖掘目的三个方面来综合阐释“数据挖掘”的定义。

1. 数据

数据挖掘所依赖的数据多种多样，既包含常见的数值型数据，也包

^① Lan H. Witten, Eibe Frank. 数据挖掘实用机器学习技术 [M]. 董琳, 译. 北京: 机械工业出版社, 2006.

含文本、图形、音频和视频、图像等类型的数据，甚至还包含其他种类的异型数据。可以这么说，信息就是数据，有信息的地方就有数据，有数据的地方就用得着数据挖掘，而且数据挖掘的重要性随着数据量的增长而不断增加。

因为大多数原始信息是多种多样且不规则的，所以原始数据大多也是不完全、有噪声、模糊和随机的。对这样的数据，即使搜集到，也未必能直接进行挖掘，这就需要在数据准备阶段对这些原始数据进行初步的整理归类，俗称“数据清洗”。之后，便可以利用常用的数据挖掘方法对这些准备好的数据做进一步的深加工，以期从中提炼、发掘出客观存在，但不为人知的潜在规律并据此进行预测。

2. 挖掘过程及方法

数据挖掘的过程大致可分为三个部分：数据准备、数据处理，以及解释或预测。其中，数据准备是基础，数据处理、解释或预测是在此基础上的一种数据加工过程。数据挖掘的方法有很多，如聚类分析、决策树分析、支持向量分类法等，也会涉及模型选择方法、样本选择方法和估计方法等。

3. 挖掘目的

数据挖掘的主要目的是解释或预测。“解释”是指解释现象，找出现象背后的原因；“预测”是指利用规律对事物未来的发展进行预测。

需要指出的是，有些研究认为数据挖掘仅针对大数据而言，只有处理大数据的方法才算是数据挖掘方法。但是我们认为，这是一种狭义的定义，因为一方面将数据挖掘限制在大数据处理范围之内，无疑将削弱数据挖掘的应用价值，另一方面，多大数量的数据能算作“大数据”并无统一标准，这样模糊了数据挖掘的定义。所以，我们给出的数据挖掘的定义是一种广义的定义，并不限定数量的多少，即数据挖掘既能应用于大规模数据，也能应用于中小规模数据。

4. 特征

数据挖掘有三个重要的特征：①挖掘方法多种多样，且预测效果的好坏并不取决于挖掘方法的复杂程度，而取决于是否能将挖掘方法与对象数据的特征巧妙地结合起来。②并非每次挖掘都能得到令人满意的预测结果。这与人类在探索未知客观世界常常遇到曲折是一个道理。这一点也常令数据挖掘面临两种尴尬：一方面，如果挖掘出一个显而易见的“规律”，则投入与产出不成比例；另一方面，如果挖掘不出有价值的东西，费时费力。③数据量越大，数据挖掘的重要性越高。特别是当遇到海量的、看似无规律可言的数据时，数据挖掘才能更好地体现其社会价值。因为哪怕是能挖掘出一条有价值的规律，都有可能创造出巨大的社会效益。

（二）数据挖掘方法概述

虽然数据挖掘的应用越来越广泛，也越来越得到人们的重视和认可，数据挖掘的相关方法也日渐丰富，但是却很少有人对这些方法进行过系统的分类，为此，本书希望能对数据挖掘方法进行一个框架性的概述。

一般认为，数据挖掘方法主要可以分为有监督的数据挖掘方法和无监督的数据挖掘方法两类，也被称为“有指导的和无指导的数据挖掘方法”。这主要是因为对“Supervising”一词的翻译方法不同而已，我们这里翻译为“监督”。其中，有监督的数据挖掘方法一般是指探讨自变量与因变量之间关系的方法，而无监督的数据挖掘方法一般只探讨变量之间的关系，而不对变量做自变量和因变量的区分，其实“监督”就是靠因变量的存在来实现的。有监督的方法如 KNN 方法、决策树方法等，无监督的方法如聚类分析方法等。通常来说，以上方法被认为是属于统计学领域的方法，换句话说，一般认为数据挖掘方法就是属于统计学领域的方法在数据上的应用。

这种观点虽然没有被正式提出过，但是在很多有关数据挖掘的书籍

的字里行间都能表现出来。不可否认，数据挖掘是经常与数据打交道的人提出来的概念，也是基于数据来分析其中蕴藏着的规律的方法，因此特别擅长处理数据的统计方法就当仁不让地承担起了数据挖掘方法的主要责任。我们也必须承认统计方法在数据挖掘中的重要作用，但是数据挖掘方法不应仅被局限在统计学方法的框架内，而应当有更为广阔的内涵。

我们认为，凡是与数据分析有关的方法都可以被视为数据挖掘的方法，而且数据挖掘方法绝不是仅仅针对大数据而言的，其处理的对象包含各种规模和类型的数据，此外，数据挖掘方法绝不应该仅是单纯从数据中找规律，而是可以允许在理论模型的指导下通过数据来发现规律，理论模型可以被视为与因变量一样起到“监督”作用的因素。只有这样来构建数据挖掘方法的理论框架，才能真正体现出数据挖掘的无限价值和魅力。我们可以将这种数据挖掘方法的理论框架称为“广义”数据挖掘方法理论框架。

在这种广义框架下，我们几乎可以将计量经济学中所有的计量方法都归为数据挖掘方法，即经济计量方法可以被看作是数据挖掘方法的一个分支；另外，经济学领域中的数值模拟方法也都可以被纳入到数据挖掘方法的框架中来，这一点很重要，因为常见的数据挖掘方法很少涉及有关数值模拟的方法，而在经济学领域，特别是宏观经济学领域，构建理论模型并且根据理论模型来进行数值模拟，以讨论某些经济变量对宏观经济的影响，是非常常用的方法。几乎所有研究宏观理论模型的学者都会用到数值模拟的方法。所以，将这一类方法纳入数据挖掘方法的范畴内能极大地拓展数据挖掘的内涵，并且能促进统计学方法与经济学方法的融合。理论模型可以被视为与因变量一样的“监督”因素，因而通过构建理论模型来引导数据的使用的方法可以被归入“有监督”的数据挖掘方法中。

这样一来，我们可以重构有关数据挖掘方法的框架，使其内涵空前

的丰富。这样，数据挖掘方法仍然可以被分为有监督的数据挖掘方法和无监督的数据挖掘方法，有监督的数据挖掘方法包含构建理论模型进行数值模拟的方法。而无论哪一类方法，都会有对应的模型或算法，因而数据挖掘方法又可被分为基于模型的数据挖掘方法和基于算法的数据挖掘方法。而基于模型的数据挖掘方法又可以被分为基于参数估计模型的数据挖掘方法、基于半参数估计模型的数据挖掘方法和基于非参数估计模型的数据挖掘方法。基于参数估计模型的数据挖掘方法最为常见，包含常用的经济计量方法，如线性回归模型方法、离散选择因变量回归模型方法、面板回归模型方法、时间序列模型方法等；基于非参数估计模型的数据挖掘方法，主要是指不事先指定模型的参数，而是通过函数的形式来发挥参数的功能。

二、金融数据挖掘

从应用上讲，只要有数据存在的领域，就可以进行数据挖掘，而根据本书对数据的广义定义，金融领域的各个部门都可以进行数据挖掘。但是，这些部门的数据并非都是直接和可获得的（“可获得的”是指普通个体都能够获得的）。比如，银行会有大量的信用数据，但这些数据是保密的，既不对外公布，也不会任意由普通人去调取。当然，银行内部人员可以利用其内部数据，结合数据挖掘的方法，进行挖掘，进而取得有用的规律，进一步指导其信用政策的制定和实施；银行相关的监管部门也可以轻易地获取其信贷数据，进行分析和使用；公安部门在追查经济类案件时，依法也可以调取相关的信贷记录数据，进而也可以利用相关的数据挖掘方法进行分析，指导破案；研究金融的学者们有些也可以获取到一些银行内部数据，使用数据挖掘的方法进行分析。但是，除了上面这些情形，对于一个普通的个体，是很难获取相关的银行数据的，即使其掌握了数据挖掘的方法，也只是纸上谈兵而已，如果金融数据挖掘的应用全部用银行信贷来举例，无疑不利于向公众普及金融数据