“十三五”规划教材
13th Five-Year Plan Textbook

“十三五”规划全媒体人才培养丛书·数据科学系列

数据分析与数据挖掘实用教程

INTRODUCTION TO BIG DATA TECHNOLOGY

殷复莲 编著

中国传媒大学出版社

“十三五”规划教材
13th Five-Year Plan Textbook

“十三五”规划全媒体人才培养丛书 · 数据科学系列

数据分析与数据挖掘实用教程

INTRODUCTION TO BIG DATA TECHNOLOGY

殷复莲 编著

中国传媒大学出版社
· 北京 ·

图书在版编目(CIP)数据

数据分析与数据挖掘实用教程 / 殷复莲编著 . -- 北京：
中国传媒大学出版社，2017.9
（“十三五”规划全媒体人才培养丛书·数据科学系列）
ISBN 978-7-5657-2160-1

Ⅰ . ①数… Ⅱ . ①殷… Ⅲ . ①数据处理 ②数据采集
Ⅳ . ① TP274

中国版本图书馆 CIP 数据核字（2017）第 244128 号

数据分析与数据挖掘实用教程
SHUJU FENXI YU SHUJU WAJUE SHIYONG JIAOCHENG

编　　著　殷复莲
策划编辑　阳金洲
责任编辑　黄松毅
特约编辑　李克俭
责任印制　日　新
封面设计　风得信设计·阿东

出版发行　中国传媒大学出版社
社　　址　北京市朝阳区定福庄东街 1 号　邮编：100024
电　　话　86-10-65450528　65450532　传真：65779405
网　　址　http://www.cucp.com.cn
经　　销　全国新华书店

印　　刷　北京艺堂印刷有限公司
开　　本　787mm×1092mm　1/16
印　　张　18.25
字　　数　357 千字
版　　次　2017 年 9 月第 1 版　　2017 年 9 月第 1 次印刷

书　　号　ISBN 978-7-5657-2160-1/TP·2160　　定　　价　49.00 元

前 言

人类的智慧使文明不断地从陈旧桎梏中破壳而出，21世纪是大数据的时代，以数字形态存储的数据中蕴藏着巨大的信息和智慧，正如人们早已对“啤酒和尿不湿”的故事耳熟能详，在如今大数据的浪潮之下，数据分析和数据挖掘技术作为大数据的核心技术基础，其理论和应用价值不言而喻。本书从实际应用的角度，深入浅出地介绍了数据分析和数据挖掘的基本概念和典型技术，以案例的形式进行讲授，并配以基于R语言的实验仿真，帮助读者了解数据挖掘的基本理论体系、掌握数据分析和数据挖掘的基本方法。本书共8章：

第1章为绪论，首先介绍了数据和大数据的基本概念，以明晰数据和大数据各自的特点，继而讲述数据分析和数据挖掘的区别，同时指明笔者非常赞同的证析的观点：“无论是数据分析还是数据挖掘，无论采用的分析手段是简单还是复杂，只要能够达到指导决策的效果就是非常优秀的方法。”第1章还重点介绍了数据挖掘的作用、标准流程和工具，最后对R语言的基本操作进行了描述。

第2章为初识数据，作为数据分析和数据挖掘的主体，本章首先对数据类型进行了定义，包括数据的定义和数据集的类型。接下来介绍了包括中心趋势度量和数据离散程度度量的数据统计特性以及数据的相似性和相异性度量。最后为读入数据与列联分析和图形显示的案例分析。

第3章为初始数据获取，本部分首先介绍数据获取的方式以及信息搜索方式，并对爬虫程序的基本原理和网络爬虫的分类进行了介绍。第3章重点介绍了简单HTML网页页面爬取、HTML网页中复杂表格爬取和非规整多页网页数据爬取的实际操作。

第4章为数据预处理，本部分在明确为什么进行数据预处理的基础上，介绍了数据清理（包括处理缺失值和处理噪声数据）、数据集成、数据变换（包括光滑、聚集、数据泛化、规范化、特征构造和数据离散化）、数据归约（包括数据立方体聚集、属性子集选择、维度归约、数值归约、离散化和概念分层）。第4章给出了数据预处理中非常重要的缺失值处理和主成分分析的案例讲解。

第5章为关联分析，关联分析以“啤酒和尿不湿”的实际案例开篇，引出关联分析的基本概念，对关联分析的基本术语、频繁项集等预备知识进行介绍，重点介绍了频繁项集的产生和规则的产生，并对关联模式的评估进行介绍，包括兴趣度度量、支持度和置信度度量、基于统计的度量。最后以案例的形式进行了Apriori算法和FP-

growth算法的应用分析。

第6章为回归，该部分首先介绍了回归、分类和聚类的关系，重点介绍了回归的基本概念。此外，对线性回归、非线性回归方法也进行了介绍，同时给出回归模型的评估，最后以案例的形式进行了线性回归和非线性回归算法的应用分析。

第7章为分类，该部分首先介绍了分类的基本概念，重点介绍了决策树分类，包括ID3算法、C4.5算法、CART算法，而后介绍了其他一些分类算法，包括k-最近邻分类、贝叶斯分类、人工神经网络分类、支持向量机分类和组合方法分类，同时给出分类模型的评估，最后以案例的形式给出了以上算法的应用分析。

第8章为聚类，首先介绍了聚类的基本概念，对基于划分的方法，如k-means和k-medoids、基于层次的方法、基于密度的方法和基于聚类的方法进行了介绍，同时给出了聚类方法的评估，最后以案例的形式给出了以上算法的应用分析。

本书的修订受到了中国传媒大学理工学部和中传大数据分析挖掘研究所全体师生的大力支持，编者在此表示诚挚的谢意。由于编者水平有限，书中难免存在一些缺点和错误，因此殷切期望广大读者批评指正。

目 录

第 1 章 绪论

1.1 数据和大数据

1.1.1 数据

数据是我们耳熟能详的一个名词，百度百科给出的定义是：

“数据（Data）是事实或观察的结果，是对客观事物的逻辑归纳，是用于表示客观事物的、未经加工的原始素材。

数据是信息的表现形式和载体，可以是符号、文字、数字、语音、图像、视频等。数据和信息是不可分离的，数据是信息的表达，信息是数据的内涵。数据本身没有意义，数据只有对实体行为产生影响时才成为信息。

数据可以是连续的值，比如声音、图像，称为模拟数据；也可以是离散的，如符号、文字，称为数字数据。

在计算机系统中，数据以二进制信息单元0和1的形式表示。”

由此可见，数据本身是没有价值的，本书将从数据、信息、知识、智慧四者的定义和关系出发进行阐述，如图1-1所示。

图 1-1 数据、信息、知识、智慧

数据：是信息和知识的符号表示。

数据是用来记录、描述和识别事物的按一定规律排列组合的物理符号，是一组表示数量、行动和目标的非随机的可鉴别的符号，是客观事物的属性、数量、位置及其

相互关系等的抽象表示，以适合用人工或自然的方式进行保存、传递和处理。它既可以是数字、文字、图形、图像、声音或者味道，也可以是计算机代码。在计算机科学中，数据是指所有能输入到计算机中具有一定意义的数字、字母、符号和模拟量等并能够被计算机程序处理的符号的介质的总称，同时也具有能被计算机识别的二进制数的形式。

数据本身是孤立的、互不关联的客观事实、文字、数字或符号，没有上下文和解释，数据表达的仅仅是一种描述，如19491001只是一串数字。

数据用属性描述，属性也称变量、特征、字段或维。数据经过处理仍然是数据，但只有经过解释，数据才有意义，才能成为信息。

信息：数据的内涵意义。

信息是人们对数据进行系统地收集、整理、管理和分析的结果，是经过一系列的提炼、加工和集成后的数据。信息是对客观世界各种事物的特征的反映。数据是信息的符号表示，或称载体，数据不经过加工只是一种原始材料，其价值只是在于记录了客观数据的事实。信息是数据的内涵，是对数据的解释。如对某人来说，19491001可以指示他的生日，也可以指示中华人民共和国成立的日期。

信息来源于数据，是对数据进行加工处理的产物，信息对决策或者行动是有价值的，其价值在于人类认识世界和改造世界活动的现实意义。

数据资源中所有信息量的多少是由消除事物认识的不确定程度来决定的，数据资料消除的人们认识上的不确定性的大小也就是数据资料中所含信息量的大小。

知识：是具有前因后果的信息，是人们在长期的实践中总结出来的正确的内容。

所谓知识，就它反映的内容而言，是客观事物的属性与联系的反映，是客观世界在人脑中的相对正确的反映。就它反映的活动形式而言，有时表现为主体对事物的感性直觉或表象，属于感性知识，有时表现为关于事物的概念或规律，属于理性知识。知识是人们在实践活动中获得的有关世界的最本质的认识，是对信息的提炼、比较、挖掘、分析、概括、判断和推论。

一般而言，知识具有共享性、传递性、非损耗性（可以反复使用，其价值不会减小）及再生性等特点。

按知识的复杂性可将其分为显性知识和隐性知识。显性知识是用系统的、正式的语言传递的知识，可以编码和度量，可以清晰地表达出来，易于传播，可以在人与人之间进行直接交流，通常以语言文字的形式存在；显性知识的处理可以用计算机实现。隐性知识是存在于人脑中的、非结构化的、与特定语境相关的知识，很难编码和度量。隐性知识是人们在实践中不断摸索和反复体验形成的，通常以直觉、价值观、推断、经验、技能等形式表现出来。它难以描述，但却是个人能力的直接表现且更为宝贵。

隐性知识的处理只能通过人脑实现，一般要通过言传身教和师传徒授等形式传播。

智慧：是指富有洞察力的知识。

智慧是富有洞察力的知识，指人在了解多方面的知识后，能够预见一些事情的发生并主动地采取行动。智慧是人类特有的解决问题的一种能力，是人类基于已有的知识和信息，针对物质世界运动过程中产生的问题，根据获得的信息进行分析、对比、演绎、推理从而找出解决方案的能力。这种能力运用的结果是将信息的有价值部分挖掘出来并使之成为已有知识架构的一部分。

比如大家都知道国庆假期去北京旅游的车票非常紧张（知识），若是你已经非常有预见性地提前购买了车票，那么你就先人一步（智慧）。

由此可见，数据≠信息≠知识，可以从数据中提取信息，从信息中挖掘知识，而智慧是一种高层次的知识。

1.1.2 大数据

大数据是近年来新兴的一个名词，也由此引发了大数据浪潮，率先给出大数据定义的是麦肯锡全球研究所报告《大数据：创新、竞争和生产力的下一个前沿》：

“大数据是指大小超出了传统数据库软件工具的抓取、存储、管理和分析能力的数据群。”

这个定义有意地带有主观性，对于“究竟多大才算是大数据”，其标准是可以调整的，即，我们不以超过多少TB（1TB=1024GB）为大数据的标准，我们假设随着时间的推移和技术的进步，大数据的“量”仍会增加。还应注意到，该定义可以因部门的不同而有所差异，这取决于什么类型的软件工具是通用的以及某个特定行业的数据集通常的大小。因此，今天众多行业的大数据可以从几十TB到数千TB不等。

麦肯锡全球研究所报告从数量级的角度给出了大数据的概念，下面是量级的转化规则：

- **b和B**

b：bit，位，一个位代表一个0或1。

B：字节，8个位组成一个字节。

- **内存（B，b）**

1B＝8b，相当于一个英文字母。

1KB（千）=2^{10}B=1024B，相当于一则短篇故事的内容。

1MB（兆）=2^{20}B=1,048,576B，相当于一则短篇小说的内容。

1GB（吉）=2^{30}B=1,073,741,824B，相当于贝多芬第五乐章交响曲的乐谱内容。

1TB（太）=2^{40}B=1,099,511,627,776B，相当于一家大型医院中所有的X光图片

的信息量。

1PB（拍）=2^{50}B=1,125,899,906,842,624B，相当于50%全美学术研究图书馆藏书的信息内容。

1EB（艾）=2^{60}B=1,152,921,504,606,846,976B，5EB相当于全世界人类所讲过的话语。

1ZB（泽）=2^{70}B，如同全世界海滩上的沙子数量总和，目前正在进入。

1YB=2^{80}B，人类尚未进入的数字时代，但已并不遥远。

1NB=2^{90}B。

1DB=2^{100}B。

以超过TB的数量级作为大数据和普通数据的界线是目前较为广泛的一种区分标准，除此之外，在维克托·迈尔-舍恩伯格和肯尼斯·库克耶编写的《大数据时代》中给出了不从数量级出发的其他特征定义：

更多：不是随机样本，而是全体数据。

当数据处理技术发生了翻天覆地的变化时，在大数据时代进行抽样分析就像是在汽车时代骑马一样。一切都改变了，我们需要的是所有的数据，“样本=总体”。

更杂：不是精确性，而是混杂性。

执迷于精确性是信息匮乏时代和模拟时代的产物。只有5%的数据是结构化且能适用于传统数据库的。如果不接受混乱，剩下的95%的非结构化数据都无法被利用，只有接受不精确性，我们才能打开一扇从未涉足的世界的窗户。

更好：不是因果关系，而是相关关系。

知道“是什么”就够了，没必要知道“为什么”。在大数据时代，我们不必非得知道现象背后的原因，而是要让数据自己“发声”。

由以上定义可知，大数据的精髓在于分析信息时代的三个转变，这些转变将改变理解和组建社会的方法。

第一个转变就是，在大数据时代，可以分析更多的数据，有时候甚至可以处理和某个特别现象相关的所有数据，而不再依赖于随机采样。

19世纪以来，当面对大量数据时，社会都依赖于采样分析。但是采样分析是信息匮乏时代和信息流通受限制的模拟时代的产物。以前我们通常把这看成是理所当然的限制，但高性能数字技术的流行让我们意识到，这其实是一种人为的限制。与局限在小数据的范围内相比，使用一切数据为我们带来了更高的精确性，也让我们看到了一些以前无法发现的细节——大数据让我们更清楚地看到了样本无法揭示的细节信息。

第二个转变就是，研究数据如此之多，以至于不再热衷于追求精确度。

当测量事物的能力受限时，关注最重要的事情和获取最精确的结果是可取的。如果购买者不知道牛群里有80头牛还是100头牛，那么交易就无法进行。直到今天，数字技术依然建立在精准的基础上。假设只要电子数据表格把数据排序，那么数据库引擎就可以找出和检索的内容完全一致的检索记录。

这种思维方式适合于掌握“小数据量”的情况，因为需要分析的数据很少，所以必须尽可能精准地量化记录。在某些方面，人们已经意识到了差别。例如，一个小商店在晚上打烊的时候要把收银台里的每分钱都数清楚，但是人们不会、也不可能用“分”这个单位去精确度量国民生产总值。随着数据规模的扩大，人们对精确度的痴迷将减弱。

达到精确需要有专业的数据库。针对小数据量和特定事情，追求精确性依然是可行的，比如一个人的银行账户上是否有足够的钱开具支票。但是，在这个大数据时代，很多时候，追求精确度已经变得不可行，甚至不受欢迎了。当拥有海量即时数据时，绝对的精确不再是追求的主要目标。

大数据纷繁多样，优劣掺杂，分布在全球多个服务器上。拥有了大数据，人们不再需要对一个现象刨根究底，只要掌握了大体的发展方向即可。当然，人们也不是完全放弃了精确度，只是不再沉迷于此。适当忽略微观层面上的精确度会让人们在宏观层面拥有更好的洞察力。

第三个转变因前两个转变而促成，即不再热衷于寻找因果关系。

寻找因果关系是人类长久以来的习惯。即使确定因果关系很困难而且用途不大，人类还是习惯性地寻找缘由。相反，在大数据时代，人们无须再紧盯事物之间的因果关系，而应该寻找事物之间的相关关系，这会给人们提供非常新颖且有价值的观点。相关关系也许不能准确地告诉某件事情为何发生，但是它会提醒这件事情正在发生。在许多情况下，这种提醒的帮助已经足够大了。

如果数百万条电子医疗记录显示橙汁和阿司匹林的特定组合可以治疗癌症，那么找出具体的药理机制就没有这种治疗方法本身来得重要。同样，只要知道什么时候是买机票的最佳时机，就算不知道机票价格疯狂变动的原因也无所谓。大数据告诉人们“是什么”，而不是“为什么”。在大数据时代，不必知道现象背后的原因，也不再需要在还没有搜集数据之前，就把分析建立在早已设立的少量假设的基础上，只要让数据自己发声，人们就会注意到很多以前从来没有意识到的联系的存在。

例如：对冲基金通过剖析社交网络Twitter上的数据信息来预测股市的表现，亚马逊和奈飞（Netflix）根据用户在其网站上的类似查询来进行产品推荐，Twitter、Facebook和LinkedIn通过用户的社交网络图来得知用户的喜好。

除此之外，IBM提出的“三V”概念，即大量化（Volume）、多样化（Variety）和快速化（Velocity），是“大数据”时代的显著特征，这些特征正在给现在的IT企业带

来巨大挑战。而最近这两年，着眼于数据应用的专家们提出了大数据的“四V”概念。“四V”概念是在原有的“三V”概念基础上增加了第四个首字母为V的词——Value（价值），即企业要实现的是大数据的价值。第四个“V”才是关键，如果我们不能实现数据的价值，那么再多的数据也是没有意义的。

大量化（Volume）

在大数据的四“V”中，Volume是显而易见的。如果没有大量的数据，我们就无法称其为“大数据”。如今，各家企业的数据量正在从GB、TB向着PB、EB级大踏步迈进。

多样化（Variety）

Variety是指半结构化、非结构化数据的量和结构化数据一样在飞速增长。全世界40亿手机用户已经将自己变成了数据流的提供者，同时手机制造商在他们的产品中嵌入了3千万个传感器，而且这一装机量正以每年30%的速度增长。各个企业采集的数据并不限于传统的数据格式，非结构化数据的增长速度超过了结构化数据的增长速率。所谓半结构化，是指数据有一定结构，但又没有固定的模型描述。结构化和半结构化数据通常能够用普通的XML模式来描述，但是非结构化数据就需要特殊处理了。

快速化（Velocity）

Velocity主要是指商业和各种相关领域处理的交易以及数据在以越来越高的速度和频率产生。每一分钟都有大量的数据在商业环境和互联网环境中产生。

价值（Value）

“四V”中的Value，则是指数据运营和应用的重要性。如果没有数据分析和数据挖掘，数据就只是数据。只有通过处理和分析的数据才能转化成信息，归纳成知识。

除了这四个“V”之外，业内也有学者和从业者提出不少其他关于大数据的“V”概念，这值得我们关注。在这之前恐怕很少有人能意识到有这么多有趣的英文词是以“V”为首字母的：数据的可验证性（Verification）、可变性（Variability）、真实性（Veracity）和近邻性（Vicinity）。

可验证性（Verification）

Verification指的是数据需要经过验证，因为数据量大了之后，带来的一个后果必然是数据质量的良莠不齐以及不同级别的用户介入而产生的数据安全问题。

可变性（Variability）

Variability指的主要是数据格式的可变性，着重于非关系型数据。

真实性（Veracity）

Veracity指的是因为数据来自不同的源头，而有些数据的来源（比如Facebook上的评论和Twitter上的跟帖）的可信度是需要被考虑在内的。

近邻性（Vicinity）

Vicinity和大数据的存储相关，处理数据的程序和服务器需要能够就近获取资源，否则会造成大量的浪费和效率的降低。

21世纪是大数据的时代，而关于大数据的故事，才刚刚开始。新兴的大数据学科有自己特有的基础架构、计算和应用体系，也有自己特有的价值链，作为数据分析和数据挖掘的初级教程，本书并不专注于大数据的分析和挖掘，而仅介绍大数据的基本概念，并将更多的关注点集中在普通数据的分析和挖掘上。

1.2 数据分析和数据挖掘

1.2.1 数据分析和数据挖掘的定义

数据分析和数据挖掘实际上很难有一个严格意义上的分界线，百度百科分别给出的定义是：

“数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息、形成结论并对数据加以详细研究和概括总结的过程。这一过程也是质量管理体系的支持过程。在实用中，数据分析可帮助人们做出判断，以便适当采取行动。”

“数据挖掘（Data mining），又译为资料探勘、数据采矿。它是数据库知识发现（KDD，Knowledge-Discovery in Databases）中的一个步骤。数据挖掘一般是指从大量数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。”

另有学者将知识发现看作数据分析的一个方面，而数据挖掘作为知识发现的一个步骤，便自然而然地被归为其中，如表1-1所示。

表1-1 数据分析的四个方面

名称	功 能	描 述	分析场景
报表	实现预定义和用户自定义报表功能。	通过报表工具实现预定义报表的自动生成和分发，并能够灵活地实现用户自定义报表功能。	静态数据 预定义报表 受限数据交互
即席查询	进行准实时的业务查询。	通常即席查询的功能会涉及准实时的业务信息，可以由DOS区提供此类业务，通过即席查询，不需要非常专业的SQL知识即可完成信息的即席查看。	事实发现 查询 报表
联机分析	利用联机分析处理（OLAP，On-Line Analytical Processing）分析手段实现多维度的交叉分析。	利用OLAP分析工具，配合设计良好的OLAP数据模型，可以完成业务人员对业务的分析需求。联机分析的手段包括各种图形和表格的表现以及在其上进行的多维度的交叉分析，帮助用户快速定位和解决问题。	多维分析 例外管理 问题发现 What-if分析

（续表 1-1）

名称	功　能	描　述	分析场景
知识发现	利用数据挖掘、统计建模等知识发现技术实现特定的专题分析。	用户获取有用信息的能力体现了数据仓库系统的价值，通过数据挖掘等高级统计分析技术，企业能够将数据源中有价值的信息（知识）识别出来并建立模型，同时通过自动化或半自动化的工具进行分析。	规则发现 方案验证 交互图表 方案识别 相关性分析 聚类分析

从上面的各种定义可以看出，数据分析和数据挖掘都用到了统计分析等技术手段，数据分析强调对数据的概括总结，而数据挖掘强调的是搜索隐藏信息。本书将能否发现先前未知的信息作为数据分析和数据挖掘的主要区别，但不过分强调数据分析和数据挖掘的界线和包含关系，而是将它们同视为对数据进行处理、得到我们所需要的信息、提炼成知识和智慧的方法。

1.2.2　证析

《证析》一书的英文书名为*Analytics*，其中文翻译是一个熟悉的“新词”——证析。证，是证据的证，这个证据更多地强调定量的证据，也就是数据；析，仍然是分析的析，“析万物之理”，分析数据以产生新的洞察，以此影响决策，从而提升绩效。证析就是指对量化证据进行分析以影响决策的实践。当人们想到使用数据指导商业决策时，往往过于强调证析中“析”的一面，强调使用数理统计模型、数据挖掘工具等数学手段分析数据，这是一个相对被动的过程。“证”的一面同样重要，也就是需要主动地搜集数据和证据以指导决策。并且，“分析”一词中的“分”字强调的是分解的手段和还原论的方法论。而在证析的具体实践中，采用还原论还是整体论的方法论并不重要，重要的是找到能够指导决策的、证明什么样的做法是有效的证据。

这是本书非常看重的一个概念，无论是数据分析还是数据挖掘，无论采用的分析手段是简单还是复杂，只要能够达到指导决策的效果就是非常优秀的方法，如图 1-2 所示。

图 1-2　数据分析和数据挖掘的误区

证析的目的是使用数学手段、利用客观证据影响业务决策，在实践过程中它可能会涉及企业管理、数学与统计学、计算机科学与技术等诸多领域的知识和技能。下面

对证析过程中可能用到的技能、所需进行的工作按顺序进行一个简单的罗列。

1．需求分析

证析是为解决业务问题、提升业务决策服务的，所以分析师需要理解业务人员的问题与需求是什么，需理解业务人员所处的业务背景、通用的业务术语、所面临的挑战、不足及痛点。需求分析不仅仅是证析项目需要完成的工作，它是任何项目的起点。当很多人强调分析师应“以客户为中心”时，更好的提法是“以客户的价值为中心”，分析师应该考虑客户（即决策者）如何实现其价值，而不应受困于客户说了什么。

2．决策流程分析

企业通过其价值链实现客户价值，企业为实现企业价值、获取利润，需优化价值链中各环节的决策。提升企业业务流程中决策流程的决策效果是证析项目的主要目标。若不能从流程的观点考虑问题，证析将只能提供一些相互割裂的独立应用与优化，这些优化为局部的目标服务，只能达到局部优化的目的，甚至这些局部优化的结果是以损害其他环节的绩效或损害全局绩效为代价的。而如果能以流程的观点考虑问题，那么证析就只是流程中一些黑盒子，是整合在全部流程中的一部分。

3．数据管理

数据的极大丰富是当前社会的重要特征，是证析在当前日益受到关注与普及的基础。数据的来源多种多样。例如：企业运营系统自然而然地产生了大量的电子化数据，射频识别（RFID，Radio Frequency Identification）等感知技术的日益普及，在博客、微博、Facebook发表各种意见等。随着数据源的丰富，企业的数据管理工作面临着更艰巨的挑战。从各个来源抽取与搜集数据、建立数据仓库、管理数据是证析项目的基础和重要组成部分，并且这部分动辄需要购买昂贵的软硬件系统，占用大量投资。

4．度量

数据是度量的基础，但数据不等同于度量。度量除了收集数字之外还需要知道这个数字的含义是什么，所处的语境是什么。度量指标不仅仅描述了企业运行的状况，也指引着企业运行的目标与方向。一方面，度量指标决定了证析项目所需要优化的决策的目标，有缺陷的度量指标有可能得出偏颇、歪曲、有缺陷的结论。正确的度量是成功证析项目的基础；另一方面，作为企业内部量化沟通的重要手段，度量指标是证析影响企业各个层次决策的有力工具。发现并实施新的、有洞察力的、合理的度量指标是证析项目的重要工作。

5．探索性数据分析与数据可视化

在数据的分析和处理过程中，人类的模式识别能力仍然占有重要的地位。图形以及表格是有效组织数据、协助研究人员对数据进行探索的重要手段。数据可视化不仅用于探索性数据分析，也是传递分析结果的重要手段。可视化的方式使得分析师能够有效地将分析结论传递给消费数据的人，并与之高效沟通。在证析项目中，常常需要

由分析师设计图表、仪表盘或者信息图来向业务人员传递分析结论、绩效指标等信息，这就要求分析师不仅要对数字有深刻的理解，还应具备一定的审美和设计能力。

6．提出假设，发现模型、关联与模式

为了获得对世界的认识并对环境施加控制，人们在决策前希望发现外部世界存在的模式并做出关于环境的假设。这些假设可能来自人们的经验与直觉，可能来自已有知识的演绎，也可能来自探索性数据分析或对图表解读过程中形成的认识。随着海量数据的出现，“假设驱动”这种传统的研究方法受到了挑战，有人认为传统的假设没有足够的能力描述海量数据中蕴含的外界环境中存在的复杂关系。以数据挖掘和模式识别为代表的、在海量数据中自动发现关系和模式的机械化数据处理工具为人们分析海量数据提供了可能。这些关系和模式可能是以算法或计算机语言的形式存储在计算机中，而不以传统的假设中所使用的自然语言、数学语言及其他形式化语言显式表现。商业领域的一些特性也决定了数据驱动的数据挖掘算法对机械化数据分析和模式识别有着独特的优势。

7．检验与评估

假设可能成立，也可能不成立，假设成立与否需要使用数据统计的方法进行检验。另一方面，对于不同的数据挖掘模型也有不同的检验标准，例如，预测类模型的预测准确率就是一个对模型的检验指标。分析师可以用建模数据之外的另外一部分数据验证这个模型的预测是否准确。这种从数字的角度对模型进行的检验是在检验模型做得怎么样。另一方面，因为模型都是为了解决特定的业务问题而建立的，所以也需要从模型是否能够满足业务目标的角度对其进行改进型检验，也就是检验模型是否在做正确的事情。检验与评估是保证证析项目质量、确保证析项目的资源朝着正确方向努力的重要手段。

8．形成理论与洞察

人们在观察和分析数据的过程中会进一步加深对现象的认识，然而人们不满足于只是描述观测到的现象与数据，而是更希望利用自己的归纳和推理能力，对数据的产生机理做出猜测，从而形成理论。人们拥有理论之后将不满足于只是利用理论对已观察到的现象进行描述，而是希望将其外推到未知领域，并对其进行预测。分析师需要跳出日常商业运营的细节，在对经验总结的基础之上获得新的认知，从而形成更有普遍意义的理论。这需要分析师具有足够的洞察力与创造力，然而这样的分析师可遇而不可求。

9．推理与优化

有时虽然我们掌握了可靠的理论和事实，但如果要得出有用的结论还需经过一定的推理工作。分析师就是证析项目中的福尔摩斯，虽然了解了很多业务知识、构建了很多理论、观察到很多事实，但如果他不具备推理的能力，还是不能从这些知识、理论、事实中抽取出对解决问题有帮助的信息。