01

第 1 章

大数据在各行各业

FROM

DATA

TO

INSIGHTS

1.1 什么是大数据?

我们的生活正由大数据（big data）掌握着。

然而，大数据究竟该如何定义呢?

或许你会认为，只要知道数据分析的方法就可以了，何必细究大数据的概念呢？其实深入了解概念及概念的演变一定可以加深你对它的认识，从而帮助更好地实现应用。不过在进入大数据的应用世界之前，先来看看各种关于大数据的定义。

1.1.1 非常流行的大数据概念

在当下流行的各种概念中，听起来最简单明了、细想起来却最难以解释通顺的，恐怕就是“大数据”了。

如果你发现自己还不知道什么是大数据，别担心，其实别人也不清楚。

例如，从一份面向全球 154 位顶尖企业高管的调查报告中（见图 1-1）可以看出：对大数据的定义迄今仍没有达成共识。

如图 1-1 所示，28% 的高管认为，“大数据”主要是指数据量

洞见数据价值.indd 2 2018/1/15 16:14:14

的急剧增加；24% 的受访者认为大数据是指新的数据处理技术；19% 的受访者认为大数据的重点在于对数据存储的新要求；18% 的受访者则认为各类数据新来源才是最重要的问题；当然，还有 11% 的高管另有见解。

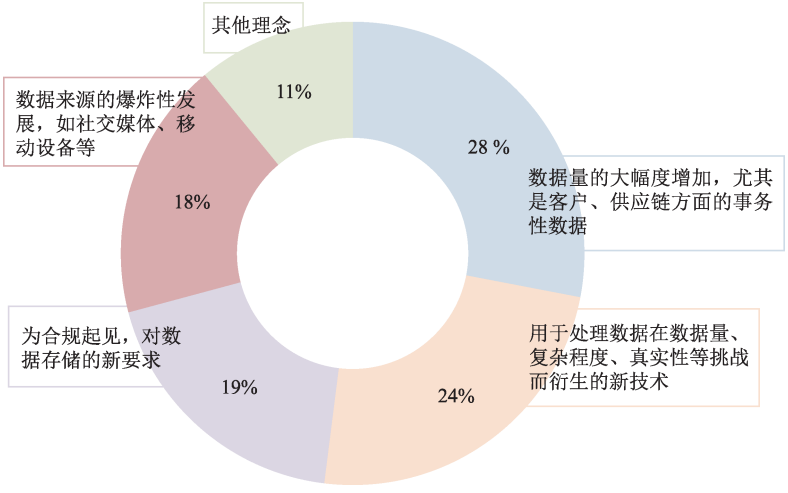

图 1-1　面向全球 154 位顶尖企业高管的调查数据汇总

数据来源：2012 年 4 月，Harris Interactive 公司代表 SAP 针对 154 位公司高管开展的在线调查结果

不仅仅是高管，关于大数据，好像人人都能说出最基本也是最著名的“3 个 V”（Volume,Velocity，Variety）定义：海量、高速、多样性。

“3 个 V”的概念可以追溯到 2001 年，当时 Gartner 公司的 Doug Laney 首次阐明了这一定义。“海量”是指数据量的急速增长；“高速”是指数据流前所未有的产生速度和得到处理的紧迫性；“多样性”则强调当今数据格式的复杂多样，从前占据主流的是结构性

洞见数据价值.indd 3　2018/1/15 16:14:14

数据，现在更需面要对非结构性的文本、视频、音频、金融交易等数据。

不过，你可能也知道，IBM 又添加了一个 V（Veracity，真实性），强调某些数据来源的不可靠。例如，从社交媒体上捕捉的顾客反馈可能因为包含主观判断而不精确，但也包含了有价值的信息，需要进行处理、挖掘。从此，“3 个 V”进化成“4 个 V”。

著名的数据服务器公司 Oracle 则对第 4 个 V 有着不同见解，他们认为 V 代表价值（Value）。在 Oracle 看来，大数据是传统的结构性数据 + 新的非结构数据中为业务决断衍生的价值。同时，Oracle 格外强调与大数据相关的技术手段，例如 NoSQL、Hadoop、HDFS 和 R 等。他们希望同时提出大数据的定义，以及解决方法。

但这还不是终点。如果你看过 SAS 公司的报告，就会知道他们又给大数据增添了两个维度：多变（Variability，又一个 V）和复杂性（Complexity）。前者是说，随着数据产生速度和多样性的飞速提升，数据流可能变得格外不稳定，例如，社交网络上就有高峰时段和低谷时段，从而对数据处理提出了挑战；后者指数据来源丰富，难以溯源、匹配、清洗或者转换。

还有很多公司并不想加入以上所说的“V 字数据队”。在“数据控们”看来，英特尔给出的大数据定义真是简洁明了：每周产生的数据量中位数不低于 300TB 的即为大数据。事实上，这是我们找到的唯一一个量化概念，当然，“300TB”这一数字恐怕也会在某一天被日新月异的数据流改写。

另外，你可能还对 Google Trends 的统计结果感兴趣，它能在很大程度上反映人们对大数据的认识，如图 1-2 所示。

在图 1-2 中，在说起大数据时，首先想到的还是数据分析（data

洞见数据价值.indd 4 2018/1/15 16:14:14

analytics），其次是近年大热的 Hadoop，而以数据为根基的 Google 也占有一席之地。顺带说一句，不远处还有个热门搜索组合即“big data dangerous”——很多人觉得大数据很危险哦。

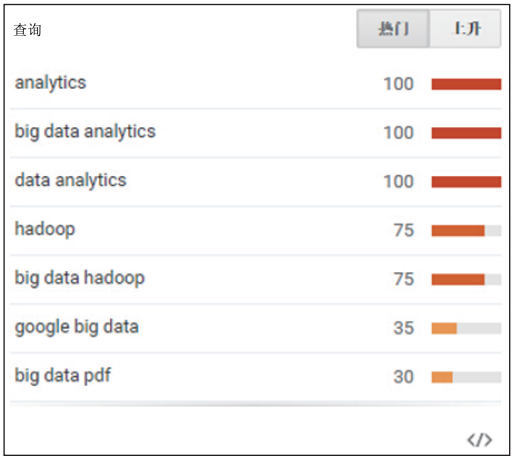

图 1-2　2016 年 12 月 Google Trends 结果

好吧，无论 Gartner、IBM、SAS、Oracle 或者英特尔，毫无疑问它们都是数据业界顶尖的权威公司，他们在定义大数据这一工作上都不甘落后，所提出的观点也十分具有启发意义，只可惜不尽相同。

这并不稀奇。无论 3V 还是 4V，这些属性词都很难全面概括大数据在各个领域的应用重点。就像每位高管的观念不同一样，每个真正参与到大数据工作中的人，心中必然都有属于自己的“大数据定义”，由于各自承担着不同的工作角色而有所区分。

当然，深入理解大数据的概念，也一定有助于在日常业务中应用它。

洞见数据价值.indd 5　2018/1/15 16:14:15

1.1.2 不那么流行的大数据概念

既然一千个人心中可以有一千种大数据，接下来看看一些没那么流行，但也自有启发意义的大数据定义。

1. 以分类定义：大数据 = 交易 + 互动 + 观察

这一概念来自业内领先的数据平台开发公司 Hortonworks。提出者 Shaun Connolly 认为，ERP、SCM、CRM 等系统都是处理交易的典型系统，它们可以提供高度结构化的数据，并储存在 SQL 数据库中。而“互动”是指人物与事件相互或与你的业务互动产生的数据，例如网络日志、用户单击、社交网络的互动与订阅或者用户产生的内容，这些都是典型的“互动数据”。

“观察数据”则是指来自物联网（Internet of Things）的数据，在万物皆联网的时代，各种信息传感设备，如射频识别装置、红外感应器、全球定位系统等种种装置与互联网结合起来，隐藏在移动设备、ATM 机甚至飞行器引擎中，形成了一个巨大的网络，并可以提供大量数据。

这个概念提出之后，也有人将这三层数据分别称为流程介导数据、来源于人的信息和机器产生的数据。这种三分类法从来源方面较为全面地概括了大数据的定义，其关注点在于数据的清晰分类。

2. 以用途定义：大数据 = 用于实时预测的数据

SAP 公司的 Steve Lucas 认为，大数据与传统数据最根本的区别在于，以往人们惯常收集的各类数据仅仅用来记录与总结，或者用于识别企业存在的问题，但大数据的核心在于，以实时获得的数据流时刻监测并预测可能发生的情况，并据此合适应对。

洞见数据价值.indd 6 2018/1/15 16:14:15

例如，以企业上一年度的交易数据归纳其财务状况属于数据的传统应用，而在社交网络上实时监测人们对品牌的观念或者预测某设备的使用周期，则是大数据的工作范畴。

1.1.3　也许会带给你灵感的大数据概念

既然已经介绍了这么多前人的观点，现在是时候亮出我们自己的概念了。

先来看如图 1-3 所示的一张趋势图。

你可能还记得前面介绍过：大数据概念的首次提出是 2001 年。然而，从图 1-3 可以看出，起初人们对这一概念并不是特别感兴趣，直到 2011 年后，它的搜索量才急剧攀升，成为热门搜索词汇。大数据并不是一开始就流行起来的。由于新技术的支持，尤其是各类先进的开源存储或处理工具的迅速发展，big data 的概念才得以再次浮现。

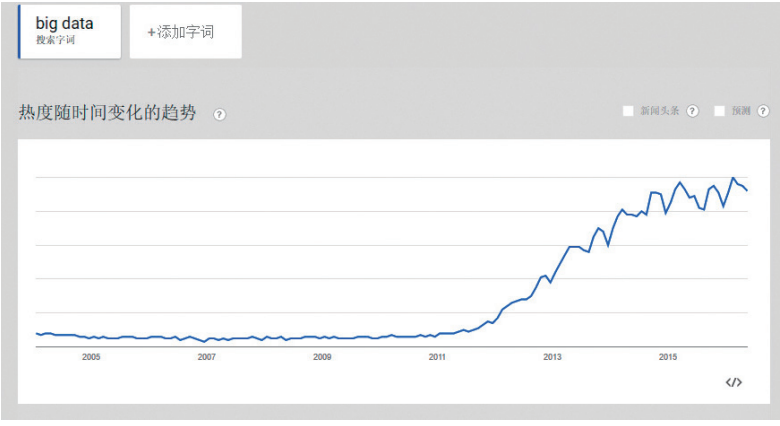

图 1-3　截至 2016 年年末 Google Trends 结果

洞见数据价值.indd 7　2018/1/15 16:14:15

在这样的发展过程中，大数据的概念难免不断变化。任何定义都是有一定时间和技术的局限性的，2017 年面对的大数据也绝不是 2001 年时的模样。因此，这里从不同角度尝试给出多个概念，希望能给大家带来帮助。

1. 技术角度

大数据带来的挑战首先是技术性的。当数据的数量或种类达到一个限度时，传统的数据采集、存储、处理、分析技术已经不再适用，须采用分布式存储、分布式并行计算、流处理等技术。在这些技术得到应用之前，海量的非结构数据、飞速变幻的数据模式或者巨大的数据存储任务难以解决，更无法从中提取价值，“大数据”自然也就无从谈起。

因此，技术能力决定大数据的理念极限。

2. 业务角度

大数据归根结底是需要服务业务、创造价值的。因此，利用新的采集、处理、存储和分析技术，能为组织带来新的业务洞察，并转化成可执行的业务策略或改变现有业务模式，最终为组织创造价值的一切数据就是我们认为的大数据。

3. 数据本身角度

其实“大数据”也是数据，很难说数据与大数据之间的分野何在。可以把数据比喻为一个小孩子，他从出生到长大成人，为社会创造价值，需要上学受教育，就像数据需要治理、完善、标准化，但不一定何时可以着手挖掘、带来价值。从数据到大数据代表了人们在

洞见数据价值.indd 8 2018/1/15 16:14:15

认知上的深入和广泛。

4. 娱乐角度

大数据像一场“贪吃蛇”游戏，蛇头是数据的采集和整理，蛇身是数据的分析和应用，如果不注意分析和应用的目的，大数据就只会越积累越多，限制自己，无法前行。只有把两者结合起来才能充分消化，越玩越好。

有关数据的十大基本概念和常见误区

大数据本身作为一个行业热点，一直是大家日常工作乃至茶余饭后讨论的重点话题。过去一段时间以来，笔者与国内许多金融机构（银行、农合机构、基金公司、资产管理公司、保险公司等）交流和探讨了数据相关的工作事项。

通过交流，笔者发现一个特别有意思的现象，那就是朋友们聊起大数据来都是一副畅所欲言的架势，侃侃而谈、各抒己见，毕竟数据和自己日常的工作与生活是分不开的，而且能看出来都饱受数据的困扰。期间所引用的一些措辞和描述都不尽相同，虽然大概都能明白是什么意思，但是也引起了一些困扰，特别是对于管理层以及业务人员，甚至包括一些专业数据和技术人员。

借此机会，笔者想抛开一些官方能查到的正式定义和术语解释，结合自身的实际经验和体会，把一些关键的数据基本概念以及大家容易混淆的常见误区进行释疑，希望能够对朋友们有所帮助。

洞见数据价值.indd 9　2018/1/15 16:14:15

以下是笔者整理的数据十大基础概念组合，与大家一一辨析。

数据、信息、大数据

既然说到数据，首先从最基本的概念入手，那就是数据、信息和大数据，这三者之间到底有什么关系呢?

可以把这三个词分成两组来解释，先说数据和信息。数据，顾名思义，是数字化的凭据，例如 1234、ABCD 等，以二进制来处理，所有的数据都可以用 0 和 1 来记录，形成数据化的凭据。之所以说它是凭据（或单据），就是因为数据本身没有任何含义，它是一份记录，只是人们把想表达的意思记录下来所形成的凭据。而信息则不同，信息所体现的正是人们想表达的这层含义。举个例子，常见的手机号码，139XXXXXXXX，单就这 11 位数字而言只是一组数据，没有应用场景的话其本身不体现任何有价值的信息，只有在打电话的时候，人们用这组数字拨号，此时才体现出它的价值，即某个联系人的电话号码。因此，数据是信息的载体，信息多是复杂的、不规范的，但是数据也可以是简单的、规范的，简单到能够用标准化的二进制语言 0 和 1 来表示。因此，在一般情况下，计算机处理的是标准化的数据，而人脑处理的是复杂的信息。

马云提出的从信息技术（IT）转变为数据技术（DT）的发展战略，反映的是去繁从简的理念，从本身没有含义的数据出发，挖掘出有意义的信息，并支持业务经营和管理决策。简单总结一下，就是人先闭嘴（“先”字很重要，不要误会），让数据说话!

再来说数据和大数据。大数据这个词可以说最近几年被滥

洞见数据价值.indd 10 2018/1/15 16:14:15

用了，全世界都在谈论大数据，每个企业都说应用大数据。相对数据而言，大数据到底大在哪里？可以这样理解，首先是体现在数据量上，大数据是海量数据，传统的技术处理不了，好比在普通个人计算机用 Excel 处理超过多少万条的数据就会死机，当然海量可远不止这个量级。海量具体是多少，有没有标准的说法？你想得有多远、多广，这个海量就有多大。其次是数据的范围和类型，不仅仅是机构内部的数据，还有好多外部的数据；不仅仅是结构化的数据，还有非结构化的数据。说到结构化，一般情况下计算机处理的是标准化的数据，这就是结构化数据。但是，还有好多文本、视频、语音等并非标准化的，这就是非结构化数据。当然，通过一定的技术手段，还是可以把非结构化数据转化为结构化数据并进行处理，看来计算机的二进制仍然是王道啊。

说明一下，大家千万不要被大数据的“大”所迷惑，非得追求大数据应用，非要用海量的、外部的、非结构化的数据等。引用一位伟人说过的，管它黑猫白猫，能抓住老鼠的就是好猫。翻译过来，管它是大数据还是一般数据，能支持经营管理决策的就是好数据。说到这里，笔者不禁又想起金庸小说《天龙八部》里的桥段，鸠摩智向大理天龙寺一众高僧展示各种炫目的武功，并表示希望以此交换六脉神剑秘籍，在众高僧被鸠顾问忽悠的眼红、心痒及矛盾纠结的时候，住持一句话就点醒众人：就你们这样连本派一阳指（都不说六脉神剑）都没练到位的，还有脸想学其他门派的武功！大数据应用也是一样的道理，先把企业自身积累的数据好好用起来，修炼自身内功，把以数据为驱动的管理理念和应用模式

洞见数据价值.indd 11　2018/1/15 16:14:15

建立起来后，再迭代进行大数据的应用提升吧。

数据源、数据元、元数据

标题中的这三个术语算是业内比较容易把人弄晕的，相信很多朋友都亲身经历过，更别提后来不知道哪位同仁又造出来一个“源数据”，这四个词的关系就更乱了，各种“YUAN”，应该怎么才能把它们说得圆啊。

先说数据源，字面上理解就很容易明白，指的是数据的来源，例如数据来源的信息系统、表格等。举个实际例子大家就更容易理解了，“合约信息的数据源是核心系统”，核心系统就是合约信息的数据源了。数据源这个词在企业中使用的频率很高，数据源不一致（或不唯一）是企业数据质量低、数据打架的重要原因，所以统一数据源是企业的基础性数据工作。

而“源数据”这个词是后来不知道谁造出来的，可以认为这是一个口语化的用词，实际上不应该作为一个正式的术语。它的产生和“数据源”是有很大关系的，即源数据就是来自于特定数据源的原始数据，这么解释还是很绕吧。还是沿用上面的例子，“合约信息的数据源是核心系统”，那么来自于核心系统的合约信息就是源数据，或者称为原始数据，是后续数据加工处理的源头数据，这样大家就容易理解了。数据源的主语是在“源”上，指的是来源，其本身并非数据，而源数据的主语是“数据”，来自特定源的数据。当然，在实际应用中，不建议用这个口语化的词，因为确实比较不容易理解并引起混淆，建议大家直接说具体的数据。

再来说说数据元。数据元的“元”指的是元素，即数据元

洞见数据价值.indd 12 2018/1/15 16:14:15

素，你可以简单理解为数据项。例如“贷款余额”就是一个具体的数据项，把它抽象起来就形成一个数据元素。那么为什么要进行抽象并形成数据元？其目的是对这些数据元素进行标准化和规范化，以便统一使用。财政部 XBRL 准则中引用的就是数据元（数据元素）的概念，把一项项数据进行抽象、定义和规范，形成基础元素，以便在财务报表中组合使用。在其他外部监管机构发布的各类标准中，数据元也是基本的要素，形成了数据元目录，并提供统一和标准化的定义，作为行业标准。

元数据中的“元”含义则和数据元的“元”不一样，指的不是元素，而是，怎么说，暂时还想不到一个比较合适的词来解释，因为这个“元”太高大上了，和“元始天尊”“天元”中的“元”是一个意思，而且你还不好解释为“原始”，否则就成原始天尊，变成了原始人，这级别和地位一下就拉下来了。可能“本质”“本源”这些词更能用来解释，但似乎也不是那么准确。所以，这里试着这么来解释，“元”是很高大上的，元数据就是数据中的数据，是最大的！那么如何理解数据中的数据？那就是，用来解释、定义数据的数据，称为元数据。如果大家还是被绕晕的话，那么通过一个例子来理解。例如，上面说到把“贷款余额”抽象为一个数据元，那么“贷款余额”的业务定义、统计口径和计算规则、管理属性和其他数据的关联关系等描述性的数据就是“贷款余额”这个数据元的元数据。这么看来，元数据的重要性就显而易见了，连数据标准都属于元数据的范畴，元数据管理也就成为保障、提升数据质量的重要手段。

洞见数据价值.indd 13　　2018/1/15 16:14:15

数据治理、数据管理、数据管控

既然前面已经提到元数据以及元数据管理，那么接下来聊聊和数据管理相关的几个概念。

数据治理、数据管理、数据管控是目前最容易被互相替代使用，且不太影响其表达含义的三个词。在实际使用中，大家确实也常在各种场景下“随机”使用这三个词，不过最近数据治理被使用的频率相对高一些。本着精益求精的精神，笔者还是试着来解释这三者之间细微的区别。

首先来说数据治理，相信大家看到这个词后会很快联想到“公司治理”。确实，数据治理本身属于一种公司治理活动，而且区别于一般的管理和管控活动，数据治理强调的是从企业的高级管理层及组织架构与职责入手，建立企业级的数据治理体系，自上而下推动数据相关工作在全企业范围的开展。可以说，数据治理是数据工作的顶层架构设计。

相对地，数据管理则更多地偏重于管理流程方面，涵盖不同领域的数据管理流程和内容，包括数据需求管理、数据认责管理、数据标准管理、元数据管理、数据安全管理、数据质量管理、数据评价管理等各个领域，这些是数据工作的核心内容。

而数据管控就更偏执行层面了，其重点在于如何执行和落地实施，涉及具体的管控措施和手段。

因此，数据治理、数据管理和数据管控体现了自上而下的管理层级，治理的重点在于管理架构和体系，管理重点在于流程和机制，管控重点在于具体措施和手段。这三者之间是相辅相成的，缺一不可。前面提到了最近业界常用到“数据治理”，启动的专项工作也多以数据治理来命名，这主要是因为在过去

洞见数据价值.indd 14 2018/1/15 16:14:15

几年中，许多金融机构实际上都已经开展了一系列数据管理和数据管控的具体工作，但是主要都是以信息科技部门牵头，配合信息系统建设为主要目的。这种自下而上的推进方式，其实际成效往往不是特别显著，很难解决企业在业务经营和管理上存在的实际用数（即使用数据）困难。这也是为什么现阶段大多数金融机构逐渐意识到在企业战略层面推动数据治理的重要性和必要性，并启动数据治理相关项目。对于数据治理工作具体如何开展，有何切实、有效的实施策略，这属于一个专项课题，本书后面还会专门讨论。

数据标准、数据规范、数据字典

聊完数据治理，接下来很自然地就要谈到数据治理的核心内容之一，那就是数据标准。

对于数据标准相信各位业内人士都经常接触到，包括外部的金标、银标等行业标准以及企业内部的标准，都属于数据标准的范畴。然而，在很多场景下，特别是早些年企业所制定的所谓数据标准，其实往往更偏向于数据字典的概念，其内容还没有达到数据标准的要求。因此，接下来先来说明一下数据标准和数据字典的区别。

首先数据标准的完整内容应该包括：业务属性标准、管理属性标准、技术属性标准。业务属性标准指的是数据元的业务相关属性，包括名称、业务定义、统计规则和逻辑等，这都是需要数据的业务归属部门负责进行定义的；管理属性标准指的是数据元的管理过程属性，包括归属部门、使用部门、管理部门、加工系统、存储系统、应用系统以及数据的生命周期关系等内

洞见数据价值.indd 15　2018/1/15 16:14:15

容，需要业务部门和技术部门共同确定；技术属性标准则偏重于数据元的技术规范，包括数据格式、编码规则、代码取值、库表字段名称等，一般由信息科技部门进行定义。

数据标准是企业级的标准化语言，既统一规范了部门间沟通的业务语言，又规范了系统间交互的技术语言。相对于数据标准，数据字典更偏重于某个或某类系统的技术属性标准，解决的主要是系统层面的开发和交互语言。

虽然在实际应用中，数据字典也可以是企业级的，可以统一规范企业所有信息系统的数据字典（往往比较难），但一般情况下只作为单一系统的数据字典，即多个系统有多套数据字典。究其主要原因，是由于数据字典对应的是系统的实际数据库表设计，但是往往很少有企业能够在企业级实现所有信息系统均按照同样的数据字典进行库表设计，这里有历史遗留的原因，也有外购成熟软件（难以调整）的原因。

而数据标准就不一样了，它更关注于不同信息系统之间进行数据交互或数据整合时，需要遵循的统一标准。不同信息系统在进行数据交互时，如果相应数据字典的规则互不一致，则要按照统一的标准进行数据映射和转换，例如 A 系统的客户性别字段名是“Customer_sex”，取值为 1（男）和 2（女）；B 系统的客户性别字段名为“Client_gender”，取值为 M（男）和 F（女），两个系统的数据字段名和取值均不一致。那么两个系统在交互或进行企业级数据整合时，就需要按照统一的标准命名进行数据项的映射，并按照统一的代码取值进行转换方可实现。

此外，数据标准还是业务部门之间的标准化语言，例如 X

洞见数据价值.indd 16 2018/1/15 16:14:15

部门的统计报表中“贷款余额”与 Y 部门统计报表中的“各项贷款余额”实际的业务含义和规则是一致的，那么应统一命名，避免管理层和部门之间使用报表数据时产生歧义。

因此，从内容范围上来说，数据字典必须涵盖系统的所有数据项，但数据标准主要针对跨部门、跨系统的共享数据项。

讲完数据标准和数据字典的区别后，相信大家就理解为什么前面提到以往很多企业做的数据标准实际都还是停留在数据字典的层面，那是因为一方面当时做的标准不是企业级的，没有得到所有业务部门的认可，另一方面更多的是科技部门主导的信息系统级的标准。另外标准的内容也多偏技术属性标准，业务属性和管理属性标准普遍缺失。再有就是标准的应用主要为了指导信息系统开发和建设，而不是为了规范业务部门用数。

当然，在与很多金融机构的同事聊起这个问题时，大多数科技部门的同事都表示当时也是没办法，科技部门自身很难（无论是行政管理还是业务能力上）推动企业级数据标准的建设，而且聘请的外部机构往往是系统实施厂商，项目过程中又没有和管理层及业务部门充分沟通，主要靠科技部门和系统厂商一起闭关修炼做出一套标准，那自然而然就做成了偏系统和技术层面的标准了。而与业务部门同事沟通时，却发现许多业务人员都表示自己不知道企业原来还有这样的数据标准。在了解完情况后，业务人员表示：“噢，那是科技部做的吧，我们不太清楚。”是啊，连业务部门都没有推广并使用的数据标准，还能称之为企业级的数据标准吗？但是，最终的项目成果还是很“显著”的，为什么呢？因为标准在系统上落地实施了！这也是目前业界存在的一个重大误区，即检验数据项目成败的关

洞见数据价值.indd 17　　2018/1/15 16:14:15

键不在于业务应用和管理上是否有成效，而是在于系统上是否落地实施？！还是回到数据治理的本质，我们不是为了治理数据而开展数据治理工作，最终还是为了服务于企业的业务经营和管理用数需求，如果过分关注（当然还是需要关注）技术和系统层面的实施，而忽略了业务应用和管理，那么真是本末倒置了。

再来讲数据规范。数据规范是一顶大帽子，既包括前面说到的数据标准、数据字典，也包括现在很多金融机构在做的业务术语规范、指标体系规范、数据模型规范等。

数据集市、数据仓库、应用数据库、数据工厂

笔者此前在一家金融机构交流数据集市的时候，了解到许多同事对数据集市以及相关的几个概念理解得不是特别准确，这也给项目工作目标和计划推进造成了一定的困扰。业务人员纷纷提出，为什么要建立数据集市，业务部门才不关心科技部是通过数据仓库、数据集市还是什么其他数据库来实现，要的就是简简单单地能够满足日常业务用数需求的一个平台！

是的，笔者百分之百支持业务用户的这个观点。对于用户而言，只要能够满足其用数需求，通过什么方式那是技术的问题。接下来通过一个比喻帮助大家理解数据集市和业务用户之间还是有一定关系的。

首先，把企业看成一个家庭，把用户看成家庭成员，把数据看成一种商品（以食品为例）。

在家里，孩子想喝饮料、吃水果等，可以直接去冰箱里拿。妈妈煮饭炒菜煲汤，也去冰箱里拿蔬菜、肉类等原材料。用户

洞见数据价值.indd 18 2018/1/15 16:14:15

什么时候想吃东西了，能够很方便地从冰箱里直接取来食用，或进行加工后再食用。那么，这个冰箱就像是应用数据库（或叫应用系统的数据库），冰箱里的食品就是数据。应用数据库能够快速地解决业务用户日常的用数需求，想要时直接拿就行，快速便捷。

但是，冰箱本身存储的东西是有限的，如果冰箱里的东西没了怎么办？那么当然是重新去超市采购喽，不过这个活儿不一定要所有的人来干，在家里买菜这样的活一般还是由妈妈负责，这样看起来妈妈是管理员（当然，同时也是用户），爸爸和孩子才是真正的业务用户啊！

好了，妈妈去超市买菜了，这里的超市就可以比作数据集市。超市里分海鲜区、蔬果区、肉类区、饮料区等，东西应有尽有，且按照客户的需要进行分门别类管理，便于大家进入相应的区域采购。那么数据集市也是如此，按照业务应用需求，把数据进行分类，形成不同的主题区，例如，风险主题集市存储了风险管理所需的数据，管会主题集市存储了管理会计所需的数据，监管主题集市存储了外部监管统计信息披露所需的数据等，便于管理和使用。

妈妈买完东西后，再把食品存放到冰箱里，以供所有家庭成员取用。同理，业务部门的管理员会定期从主题集市中获取所需的数据，并返给应用数据库以支持日常用数，只不过往往这个动作不是人工来执行，而是定制好后通过系统接口自动批量处理。这样，数据集市就完成了对应用数据库的供数了！

噢，那么大家自然而然就想到了，超市的货仓一定就是数

洞见数据价值.indd 19　2018/1/15 16:14:15

据仓库啦！是的，这个货仓可以存储的东西非常多，虽然也是按照不同的商品分类进行管理（好比数据仓库的分区），但是用户一般不会直接到仓库里买东西，所以和用户是相对隔离的。能接触仓库的就是库管员和送货员，这些人往往都是有技术背景的，不属于业务用户的范畴。当然，事无绝对，不排除有些商家采取批发直销的模式，让用户直接去仓库里取货（国外有好多批发超市），不过这种情况毕竟较少。业务用户直接从数据仓库取数的情况也不是完全不存在的。

通过以上的生活实例，相信大家很容易看出来数据集市与数据仓库、应用数据库之间的区别和联系。数据集市对业务部门来说还是很有意义的，因为至少同时作为管理员和用户的妈妈还是需要到超市里采购东西的，而且不排除作为真正业务用户的爸爸和孩子也愿意去逛逛超市，直接选择自己想要的产品。大家想想，如果家里附近没有超市，那是一件多么不方便的事啊，这就是建立数据集市的意义所在！

最近业界还出现一个很新、很火的概念，叫数据工厂，很多人不理解它的含义和定位。没关系，还是拿上面的例子来解释。大家都知道，去超市买完鱼后回来得自己去鳞、取鳔等，一系列动作特别烦琐且不便，现在的年轻人一般都懒得做。如果说有这样的数据工厂，能够按照用户的需求进行数据的定制加工，帮你把这一系列麻烦事做了，甚至还可以再进一步把鱼清炖或红烧了，用户可以直接获得加工好的数据产品，何乐而不为？不过，这个加工厂如何选址、按什么工序加工、如何配送那是科技部的事，用户可以不用管，提出对最终产品的要求即可。

洞见数据价值.indd 20 2018/1/15 16:14:15

数据平台、大数据平台、元数据平台、数据服务平台

数据平台是一个很大的概念，但是这个概念在日常工作中往往被用窄了。例如，很多业务用户认为数据平台就是报表平台或者数据仓库，这和大家的关注点、出发点不同有很大的关系。

数据和信息是对等的，平台和系统是对等的，那么数据平台和信息系统也就是对等的。信息系统是多么大的概念，那么数据平台也是如此。有些金融机构说自己计划建个数据平台，想问问专家有什么建议。这就好比说建个信息系统，请你告诉应该怎么做一样让人难以具体回答，因为还没说清楚要建什么样的信息系统。所以，首先要清楚数据平台包括的范围和类型。

一般来说，数据平台大致可以分为数据整合平台、数据管理平台以及数据应用平台。当然，这只是第一层级的划分，不过目前业界的数据相关平台都可以归到这三大类中。

数据整合平台指的是把数据从源系统进行收集和整合、加工处理、存储和共享的平台。数据整合平台的首要目标是形成企业级统一、共享的整合数据，为后续的数据应用提供基础。当然，整合平台也不仅仅是简单整合，还包括一定的数据加工。前面提到的数据集市、数据仓库、数据工厂都属于数据整合平台的范围。

那么大数据平台是否属于数据整合平台的范围？从严格意义上说不完全是，就看这大数据平台是广义的还是狭义的。广义的大数据平台和数据平台是在一个层级的，它也包括大数据的整合平台、管理平台和应用平台。不过目前业界所说的大数据平台一般还是偏狭义的，即只包括大数据的整合平台。即狭

洞见数据价值.indd 21　2018/1/15 16:14:15

义的大数据平台指的就是把海量的大数据进行整合和存储的平台。

区别于传统的数据整合平台，大数据平台架构由于其处理的数据类型和性质而与传统平台存在较大不同。前面已经和大家介绍过大数据的特点，那么大数据平台比较特别的地方就在于它可以高效处理海量的历史数据、外部数据，特别是非结构化数据，一般采用基于 Hadoop 的分布式架构，具有高效性、高容错性和高扩展性的特点。而传统的数据整合平台（如数据仓库）多为关系型数据库，仅比较擅长处理结构化的数据。

再来说数据管理平台。数据管理平台的目标是实现数据全生命周期的管理，确保前面提到的数据治理、管理和管控措施的系统落地，保障数据质量。元数据平台就属于数据管理平台的范畴，通过元数据的管理最终实现对数据的管理。

数据管理平台的用户主要是数据管理部门和技术部门，当然也包括数据的业务用户。不过对于业务用户而言，数据管理平台最重要的意义主要体现在，一是可以查询并了解企业级的数据规范，包括数据的业务定义、统计规则等规范语言；二是在发生数据变化后，能够进行上下游的提示。例如，数据源发生变化后影响到后端的报表应用，报表用户能够及时获取变化信息，评估对自身用数的影响并制定应对措施。这一点其实非常重要，在与很多业务部门同事沟通的过程中，许多人提出了对此类上下游数据问题的不满，好比财务部的同事会抱怨前台部门修改数据前后都没有及时和财务部进行沟通，最后报表已经对外报送了，但是前台数据修改后导致与业主的数据不一致。这类问题可通过数据管理平台解决。

洞见数据价值.indd 22 2018/1/15 16:14:15

最后说说数据应用平台，这也是业务用户最关心的、直接使用的用数平台。企业的报表平台、统计报送平台、指标系统、管理驾驶舱系统、决策支持平台以及高阶的数据分析和挖掘工具等都属于数据应用平台的范畴。一句话概括，就是支持最终业务用户日常用数的应用交付平台。无论是数据整合平台还是大数据平台，最终都是服务于业务用户的用数需求，虽然在应用模式上有所区别，有初阶的固定报表、数据模板应用，也有中阶的指标多维灵活查询、决策驾驶舱应用，还有高阶的大数据分析和挖掘应用，但是都需要通过数据应用平台进行实现数据服务的最终交付。

很多人还会问，那么数据服务平台又是什么？和数据应用平台有何区别？笔者个人认为，数据服务和应用都属于一个范畴，即服务于用户的视角，实在没必要进行如此独立的平台区分，许多金融机构也只建设统一的数据应用平台。如果非要区分两者，数据应用更偏重于最终的数据结果交付，例如，用户通过应用界面获得想要的报表或指标；而数据服务更偏重于过程的调度，例如，进行用户用数偏好分析以便选择用户可能需要的数据进行主动推送等。

基础数据、主数据

前面介绍了有关系统和平台的两组概念后，接下来又要回到比较抽象的几组概念，不过大家不用担心，这里会尽量用简单的语言快速说明，大家别打瞌睡。

先来聊聊基础数据。基础数据是相对于衍生数据而言的，是从业务前端直接产生和采集的，未进行过加工计算的基础性

洞见数据价值.indd 23　　2018/1/15 16:14:15

数据。例如，业务交易产生的交易流水数据、采集的客户基础数据、签署的合约数据等，都属于基础数据的范畴。基础数据是企业开展日常业务经营所直接产生的，为中台、后台的风险管理、财务核算、管理分析和统计应用等提供了数据基础。

基础数据的缺失和质量不高是目前业界碰到的最大难题。许多企业在开展数据统计和分析时，都或多或少遇到由于基础数据缺失而导致难以进行准确地分析和预测，无法更好地支持经营管理与决策，这个困难直接体现了企业的业务前台和中台、后台管理之间的先天矛盾。

从前台的角度来讲，其主要的任务在于提高业绩、提升客户体验。一方面，基础数据的采集和维护一般都需要前台业务部门人员来执行，而这项工作却没有纳入其绩效考核的内容，因此引不起前台业务人员的重视。另一方面，在业务交易过程中进行数据的采集和维护也是需要一定的时间成本，如果在客户办理业务时过分强调信息的高质量采集，也会影响业务的办理效率和客户体验。而中台、后台人员为了满足其管理目的，则希望前台能够尽可能地的采集所需的基础数据。看起来，这还真是一个难以调和的矛盾。曾遇到这样一个情况，企业的后台部门为了获取统计报送所需数据，提出了前台业务系统改造的数据需求，但是前台业务部门以影响业务效率为由拒绝接受，导致相关工作难以顺利推进。相信许多中台、后台管理人员都有类似的体会吧！

所以，为了更好地协调这个先天矛盾，还是要站在前台业务的视角来解决问题，而不是一味互相推卸和指责。首先要找到一个契合点，即哪些数据是前、中、后台部门，特别是前台部门关心的数据，这些数据又如何解决前台部门的用数需求，

洞见数据价值.indd 24 2018/1/15 16:14:15

那么能够求同存异，将关键数据范围作为切入点，并以应用迭代的方式逐步完善基础数据的采集。例如，团队曾帮助企业建立客户交叉销售、客户维挽等数据分析模型以便更好地支持业务经营，通过这类应用项目，前台部门深刻意识到客户数据采集的重要性，并主动开展客户基础数据质量的提升工作。

在基础数据中，有一类数据非常重要，作为不同系统之间交互和共享的关键数据，把它称为“主数据”。例如，不同业务系统都会有客户的数据、产品的数据，如果每个部门、系统分别维护这些数据，必然会造成不一致。因此，需要在企业级建立统一的主数据管理，如客户主数据、产品主数据、机构主数据等，在全公司范围内都应是统一且唯一的，并确定部门和系统间交互识别这些主数据的唯一标识，如客户编号、产品编号、机构编号等。那么，这些唯一标识也成为主数据中的主数据，是开展主数据管理首先要统一、整合的对象。虽然这是项非常基础的工作，但是许多企业不见得做得好，例如，有些金融机构的客户编号就没有在企业级范围进行统一，系统间的客户信息无法唯一识别，同一客户可能在机构内存在多条记录却并未整合等。解决问题就要抓住问题的主要矛盾，那么主数据问题就是所有数据问题中急需解决的关键性问题。

衍生数据、指标

基础数据和衍生数据是相对的两个概念，讲完基础数据，那自然就要说到衍生数据。

衍生数据是指按照一定的规则和逻辑，对已有的数据进行计算和加工后形成的数据。相对于基础数据可直接由前台部门采集

洞见数据价值.indd 25　2018/1/15 16:14:15

而言，衍生数据一般是基于基础数据或其他衍生数据的再加工。

衍生数据常见的质量问题和基础数据不同，基础数据难在采集环节，而衍生数据却难在统一口径规范上。一般来说，金融机构的前、中、后台部门由于视角和目的不同，对特定衍生数据会有不同的计算规则和口径，这也导致产生数据的不一致问题。但是，这类问题解决起来相对基础数据还是简单一些，只要企业能够自上而下统一进行标准化规范，明确衍生数据的唯一定义部门和加工系统，那么很多问题都会迎刃而解。不过说起来容易做起来难，这里面仍然需要协调各部门的不同利益和需求。

相对于基础数据中的主数据，衍生数据中也有一些非常关键的数据，用于满足管理层日常的经营管理和决策支持，这些衍生数据称为指标。

在和业务人员交流的过程中，其实很多人都不太清楚衍生数据的概念，但是只要提到指标，大家一下子就明白了，毕竟日常工作中用到了各种指标。如果非要区分衍生数据和指标，可以这样理解，衍生数据是一个广泛的技术概念，但是指标更偏重于业务应用，之所以许多金融机构近期致力于建设企业级指标体系，这是因为只有指标才是管理层和业务部门关心的，衍生数据离业务太远了。

接下来简单说明一下有关指标体系的几个基本概念。

报表统计项、指标、维度、度量

指标是反映对象特征属性的、可衡量的单位或方法，是具有（业务）意义的指向和标杆。

指标可分为基础指标和衍生指标，其中基础指标是具备宽

洞见数据价值.indd 26 2018/1/15 16:14:15

泛定义的统计对象，从而为指标的灵活组合与多维分析提供基础。衍生指标是在基础指标的基础上，通过添加一个或多个统计维度形成新的指标或通过不同指标进行运算而形成新的指标。一般来说，指标都属于衍生数据。

维度（或称统计维度、筛选条件）是对指标进行描述的不同视角，用于标识指标的不同方面的属性。例如，针对贷款余额这个指标，可以按照产品类型维度（如个人消费贷款、个人住房贷款等）来分析，可按照资金投放行业维度（如农、林、牧、渔等）来分析，可按照贷款期限维度（如长期、中期、短期等）来分析，也可以按照贷款状态维度（如正常、不良等）来分析等。一般来说，维度都属于基础数据。

度量是针对指标而言，和维度并没有关系，是指标的度量衡，本身并无实际统计意义，如金额、余额、比率、笔数、个数等。例如，贷款余额这个指标的度量就是余额。而报表项、统计项等无非指标的应用，企业所抽象出来的指标都是来自于各类报表的具体报表项或统计项，并回过头来应用于不同的报表。

有关具体如何建设企业级指标体系，受限于篇幅，在这里就不展开介绍了。这是很有意思的一个主题，特别是应用指标体系，能够有效解决企业的用数需求。

数据模型、数据分析模型、统计模型、数据建模

近期大数据分析特别火，很多朋友都在谈论数据分析和建模，这就引出了许多有关模型的概念。

其实大家谈的数据分析模型是一种统计模型（或者叫数学模型），是以数学统计的算法构建的复杂模型，满足于一定的

洞见数据价值.indd 27　2018/1/15 16:14:15

应用场景。目前数据分析模型的应用场景较常用于两个方面，一是客户营销，二是风险管理。在风险管理领域，许多银行金融机构为了满足新资本协议的合规要求，建立了相应的风险计量模型，实际上这也属于统计模型的一种应用。

数据分析模型一般需要输入大量的历史数据，按照既定的参数和算法进行模型计算后，输出对业务有意义的分析结果，例如，有潜在产品需求的目标客户清单、特定价格策略下客户响应预测、存量贷款未来逾期的概率预测、最低资本要求等。数据分析模型最终还是要服务于特定的业务经营和管理要求。

相对于数据分析模型，数据模型的含义可就不同了。数据模型指的是数据的结构和关系。最顶层的模型一般是概念模型，主要是从业务视角提出的对数据的概念性的需求表述。具体的数据模型一般分为数据的逻辑模型和物理模型。逻辑模型指的是数据之间的逻辑关系，一般通过实体关系图来体现，而物理模型大家可以简单理解为数据库的表结构。

数据分析模型更偏重于业务应用和决策支持，数据模型则更偏重于系统设计和实施，两者的目标是不同的。

解释完这几个模型的区别，那么模型的构建（简称为“建模”）就很清晰了。业界说的数据建模一般指的是数据模型的构建，但是描述不够准确，称为“数据模型建模”更加准确。而数据分析模型的构建则可以称为“分析模型建模”或“统计模型建模”。当然，在日常用语中，大家只说建模就可以了。

希望以上的概念解释和经验分享对朋友们理解相关术语有所帮助，许多观点也都是笔者个人的见解，不一定准确，也算是抛砖引玉，欢迎大家批评指正！

洞见数据价值.indd 28 2018/1/15 16:14:15

1.2　大数据在银行业

1.2.1　业界展望：大数据，银行业未来的核心动力

由于数据的相对高质量和行业性质，银行业已经是大数据工作涉猎最多的领域之一。银行 3.0 时代，客户的行为正在网络、通讯、智能设备等技术的影响下快速变化，客户使用银行服务的方式，正在从柜台走向网络和智能手机。银行和客户的联系更多依赖于线上渠道，对客户的了解越发依赖数据。数据本身成为银行的重要资产，数据分析则成为一次重要的、革命性的思维变革，正在驱动银行业务模式的转型，实现智能化营销、智能化运营、智能化风控，推动银行整体展业和运营模式的变革。

1. 数据将成为核心竞争力

如今，越来越多的客户开始使用电子银行和移动终端，银行已经记录下极为丰富的位置、行为偏好、需求偏好等信息，大量信息等待分析和挖掘。对银行来说，人员、资金、技术在未来一段时间内均是可替代的，只有数据是长期积累、不可替代的关键因素，如果不能将数据作为银行的战略性资产予以开发利用，未来在激烈的市场竞争中将处于落后地位，甚至失去银行的核心竞争力。

2. 典型的数据应用领域

当前，在整个银行业务中，数据分析与挖掘主要可以作用于以下三个领域。

- 客户领域：通过全面收集与整合客户内外部数据，形成完整的客户分析视图，预测客户需求，整合全行产品和服务资源，

洞见数据价值.indd 29　2018/1/15 16:14:15

为客户提供全面的综合服务。客户领域是数据挖掘作用与研究的主要方向。

- **风控领域：**使用客户交易行为、行业趋势、区域环境、企业主行为等信息，更加及时、准确地发现银行面临的潜在信用风险。
- **运营领域：**利用大数据优化运营管理，识别流程缺陷和改进之处，推进全产品线经营。

3. 常用的数据处理方法

总体而言，常用的五大类数据处理方法是：统计、数据挖掘、预报、文本挖掘和优化。

统计（statistics）是指常见的传统方法，如线性回归、假设检验、logistic 回归等。

数据挖掘（data mining）的算法包括决策树、神经网络、知识向量机、随机森林等，如今它与统计方法已经深度交叉，很难区别看待。

在预报（forecast）方法中，需要强调的是，“forecast”与“predict”有所不同，后者重点在于预测某事件是否发生，例如，客户是否会响应营销？其风险是高是低？而预报则是指利用时间序列方法，与计量经济学相结合，预测未来走势。例如，GDP 的发展、失业率的变化、人口走势、贷款不良率的走势以及销售预测、业务量发展规模预测、小企业客户数发展预测、按周预测理财余额等。

文本挖掘（text mining）在银行业务中同样有用。银行的文本数据主要有两个来源。第一是呼叫中心的投诉和建议记录。从前这类投诉等解决之后都归于闲置，但其中许多信息其实可以反映出客户

洞见数据价值.indd 30 2018/1/15 16:14:15

对银行到底哪里不够满意、哪些方面需要改进：是理财利率太低？还是贷款流程太长？借此，可以为银行找到业务痛点。第二则是银行的网银平台中，后台服务器会记录许多信息，例如客户的踪迹、在某页面的停留时间、接下来的跳转……这些行为会形成网络日志，供挖掘与分析。如今社交媒体发展迅猛，通过微博、微信公众号等搜集到的文本信息也很有价值。

优化（optimization）则是利用运筹学得到最优结果，这一方法从前一直为人忽视，但发展前景不可小视。例如，银行的好客户在各个业务条线上都有好的效应，储蓄方面表现好的客户，在理财、基金等方面也表现良好，各业务条线都希望针对他们进行营销。但不能对客户形成骚扰，总不能在一周之内让客户接到 10 个营销电话后产生负面情绪吧？因此，银行可能会设定规则，例如三个月内针对一位客户只能营销一个条线。

然而，该对谁营销哪个条线呢？依据又是什么？毕竟银行的每个部门都需要争取好客户。解决方法就是优化，关键点在于总体目标和约束条件。例如，总体目标是“带给银行的整体利润最高”，而约束条件是“共有 5 个营销条线，一位客户只能参与一个条线，每个条线最多接触若干位客户”，最终以此得到最优解，指导各业务条线发展客户。

优化方法还可以帮助改进风险管理的策略，因为高风险客户也是银行盈利的重要部分。例如，银行信用卡业务的收入来源有 70% 来自循环利息收入和分期收入，其余来自年费收入、罚金等，而循环利息收入显然依靠引入风险客户获得。由于近来商户的回佣降低，如果所有客户都信用极好、从不拖欠、全额还款，那么银行光靠手续费很难赚到钱。因此，风险客户的引入是必需的，而这个过程中

洞见数据价值.indd 31　2018/1/15 16:14:15

也必然需要优化手段分析，在保障利润最大的前提下应该吸引多少风险客户为最佳，让赚到的钱可以覆盖风险成本。

4. 国内银行的数据分析现状

目前，国内许多银行都已经认识到数据分析的重要性，特别是在零售、信用卡、风险等条线均开展了有益尝试，也取得了一定的效果。然而，数据作为全行资产，其价值还远未发挥，这一点在各个领域都有所缺憾。

- **客户领域：**银行已经建设并使用了客户信息库 ECIF，但客户行为、价值、偏好等分析属性仍然缺失，为支持展业模式从以产品为中心向以客户为中心转变，银行对客户的洞察还需要大幅提升。
- **风控领域：**已经开展评分卡、风险计量等一系列应用，但在预测预警方面还有更为广阔的空间。传统的风险计量更多依靠客户财务数据，而这些数据往往是滞后的，特别针对小微企业客户，利用财务数据来管控风险的问题更为突出。
- **运营领域：**在传统业务模式下，国内银行已经形成了固定的运营和流程体系，而互联网公司通过运用大数据技术，颠覆性地改变了如营销和授信这类核心流程，从而将银行远远甩在身后。

相比于国外的领先银行，国内银行之所以尚未开展有效的大数据应用，其深层次的原因包括以下四方面。

- **员工意识：**大数据应用仅仅是对现有业务模式的补充和支持，还是业务转型的关键驱动因素？不同人员对大数据应用和价值的理解参差不齐。

洞见数据价值.indd 32 2018/1/15 16:14:15

- **数据孤岛：**银行在发展过程中，形成了数以百计的信息系统，但数据分散、没有整合，为了不同应用反复开展数据提取、清理与整合，数据基础成为大数据应用的制约因素。
- **专业能力：**为了从数据中提取有价值的信息，并转化成可行的业务策略，需要集中技术人才、数据人才、分析人才和业务骨干，专业的数据分析人才是市场稀缺资源。
- **企业文化：**在应用大数据的过程中，会对现有业务策略和流程模式提出挑战，甚至会牵涉到各业务部门的绩效，难免遭遇阻力。没有高层的大力支持，有效开展大数据应用几乎是不可能的事。

5. 数据分析的创新尝试

国内某些大型银行通过借鉴海外银行以及互联网公司的领先实践，已经在数据分析领域开展了大胆创新尝试，将大数据上升为全行战略，将数据分析结果作为制定业务策略和指导日常操作的关键输入，这些变革举措包括。

- **整合数据基础：**制订全行范围内的数据标准，在现有数据仓库的基础上，进一步补充不同来源的内外部数据，也包括结构化和非结构化数据，构建大数据分析平台，确保数据分析所要求的准实时性和数据的全面性，数据分析结果会统一发布和实时推送到业务前端。
- **变革组织架构：**设立大数据分析中心，将分散在科技部门、业务部门、分支行机构以及管理信息部门的专业人才进行资源整合，同时引入外部专业数据分析和建模人才。将释放业务价值作为分析中心的核心考核指标，同时在全行范围内开展

洞见数据价值.indd 33　　2018/1/15 16:14:15

大数据宣讲，培育企业的大数据文化。

- **探索合作模式：**通过一系列的短平快项目，首先从创收而非风险和运营的角度开展大数据应用，迅速获取业务部门特别是前台的支持，探索出一套行之有效的分析中心与业务部门的合作模式。在合作过程中，坚持以价值为导向，注重维护与业务部门的良性互动，而非传统的需求 + 被动开发模式，利用数据分析对业务部门的各种假设进行讨论验证和迭代优化，转化成切实可行且基于事实的业务建议。
- **构建知识体系：**通过数据分析所获得的洞察会作为知识在全行传播和共享，改变过去某个业务部门分析能力较强导致洞察仅限于某个部门的情况。为最大化银行的收益，制订整合的业务策略，推行客户整合营销和全方位服务提供事实依据。

6. 数据分析的效果示例

从银行整体视角开展数据分析，会突破业务部门各自为战所带来的局限性，例如在下面这些领域的应用。

在客户领域的应用。

- **营销活动管理：**对各业务部门的营销活动进行整体管理，提升营销活动的效率和效果。例如，避免出现过多或缺乏联系客户，避免重复开展不盈利的营销活动，自动推广盈利的营销活动，集中追踪和评估营销活动。确保业务部门所建议的营销活动能满足最小 ROI 的要求。
- **跨产品线营销：**通过分析银行卡商户数据，识别到大量商户将销售款转入本行后又转入他行同名账户，本行账户仅仅作为中转账户，这类个人商户不符合私人银行高端客户 AUM 标

洞见数据价值.indd 34 2018/1/15 16:14:16

准因此未开展营销，但却是私人银行部的潜在客户。还有，通过分析银行卡交易数据，识别到某些行业例如儿童用品行业连续 6 个月出现交易井喷现象，公司业务部可以及时跟进这些行业客户，开展现金管理或融资服务等营销。

- **价格弹性管理：**分析发现，某些客户追求服务质量和客户体验，对价格并不敏感，在强化客户服务的基础上，在不同业务品种上适当且一致地提高手续费不会影响客户忠诚度。
- **资产负债分析：**在客户分群层面开展，对负债端来说（如活期存款），识别比平均行为更长的期限，付出较短期限的利率，但投资到较长期限上。这将改进资产负债的资金供给策略，增加贷款产品在定价方面的吸引力。

在风险领域的应用。

- **小企业风险管理：**除了传统的财务数据之外，更加依赖小企业和小企业主的行为数据。数据分析表明，在中国人民银行征信系统中的征信记录查询次数越多，小企业贷款逾期或违约的概率更高。
- **贷款申请和发放舞弊：**通过社会网络分析（Social Network Analysis，SNA），例如共享电话号码、钱款的转入转出等，将客户、员工、账户等进行关联，考虑交易方向的联通图，识别潜在的关联人和非法中介，防止多头授信、重复放贷和有组织的骗贷。

在运营领域的应用。

- **销售网络激励：**对客户经理激励机制和奖金发放设计、计算和跟踪，减少销售激励成本，增加销售收入，促进交叉营销，减少客户服务问题。识别激励机制对客户经理行为的影响，将

洞见数据价值.indd 35　2018/1/15 16:14:16

好的销售行为推广到全国。

- ATM 现金加仓：结合 ATM 物理位置、交易记录、客户分布、周边设施，对运钞路线和加仓进行优化，降低运营成本。
- 网上银行优化：通过分析系统日志，了解网上银行各页面访问次数、停留时长、跳转次数、成功交易和失败交易频率，优化网上银行设计和交易流程。

通过对各种业务场景开展反复和深入挖掘，银行的数据分析团队会迅速培养分析能力，积累业务洞察，从整体上推动银行向智能化发展。

1.2.2 创新方向：大数据助力银行网点实现转型

假设今天你手头恰巧缺点现金，按照通常的思维，你有三种选择。

- 去公司楼下的银行网点柜台取款。
- 去公司楼下的 ATM 机取款。
- 跟你的同事借这笔现金，以微信支付或支付宝支付的方式把钱还给他。

大概没有谁会选择第一种方式。在过去若干年中，随着互联网的迅猛发展和消费观念的改变，消费行为也发生了翻天覆地的变化。无论是交水电费、存取款还是购买理财产品，都不再是银行实体网点的专利，多数人也不肯去遭受排队等待之苦。

银行网点在生活中究竟还有多重要？或许一些盈利数据可以揭晓部分答案。根据《亚洲银行家》的相关研究，目前在亚太地区的新兴市场和成熟市场中，分别有多达 43% 和 32% 的银行网点无法实现盈利。就国内而言，银行业整体净利润增速持续下滑至个位数，中国银监会此前公布的数据显示，截至 2015 年四季度末，商业银行当年累计实现净利润 1.59 万亿元，较 2014 年仅增长 2.43%。

洞见数据价值.indd 36 2018/1/15 16:14:16

银行业盈利增速的下降导致巨大的成本压力，而成本压力已成为亚太地区银行业推动网点转型的最主要动力。严峻的外部环境和高昂的经营成本令许多网点的盈利能力出现问题，传统银行网点冗余的人工支出若不能带动更多收益，则必将成为银行沉重的包袱，而“**网点转型**”正是目前国内商业银行面对这一问题提出最多的一个概念。

目前市面上较成熟的银行网点转型思路为**硬转型 + 软转型**，从交易结算型网点向服务销售型网点转变，注重优化服务流程，提升客户体验。然而，这样的思路未免过于泛泛而论，缺乏数据支撑。网点究竟应该保留原样，进行转型，还是考虑撤销？如果要转型，该朝什么方向转型？这些问题都有待提供更具体的方案。

整体看来，多数银行仍缺少一套完整的网点竞争力评估系统，这样的系统可以用来评价网点的发展水平、同业态层级，指导网点转型思路、市场定位和目标客户等。通过对国内商业银行网点转型现状分析，为了更加系统化、数据化地支撑网点转型，设计了如图 1-4 所示的物理渠道评价闭环体系。

图 1-4　物理渠道评价闭环体系

洞见数据价值.indd 37　2018/1/15 16:14:16

如图 1-4 所示，通过持续进行、不断完善物理渠道评估和改进，基于准确可靠的评价数据，可以对银行网点进行较为系统地评估。而在评价网点绩效时，也应注重以下四大原则。

- **平衡性：**注重财务指标和非财务指标结合运用。
- **导向性：**绩效评价的主要目的在于合理、及时地引导被评价网点遵循上级银行的经营管理战略、理念，不断调整其经营管理策略、思路和行为。
- **适用性：**绩效评价指标的设置应力求体现适用性，即可操作性。
- **代表性：**在指标设置上，应选择具有代表性的核心指标，凭借这些指标对网点的经营、管理、发展各方面进行评价。

设计评价指标不但要依据以上四大选取原则，注重指标的全覆盖性也非常重要。在实践中，财务指标、客户发展指标、业务量指标、风险控制指标和优质服务指标五大指标类型最能全面反映网点绩效。

当然，对于所有网点也不能一概而论，可以利用网点差距分析法，将网点划分为旗舰网点、综合网点、轻型网点三大类型，为它们分别建立不同权重的体系，计算网点竞争力得分，逐个对网点进行分类诊断指导和精准督导，基于网点进行业务发展和数据分布情况，依据网点绩效评价体系，对网点经营现状进行多角度分析。

重点收集的数据包括外部环境、业务种类和客户分析三方面。

1. 外部环境

- **网点周边环境是什么？**人口流动性怎么样？周边 2 公里范围内人均可支配收入如何？
- **潜在的个人客户群体分布如何？**潜在的对公客户类型包括哪些，个人工商户、企事业单位？客户规模多大？

洞见数据价值.indd 38 2018/1/15 16:14:16

- **网点是否针对特定客户群体或特定业务产品？**每个具体网点是否有所区别？

2. 业务种类

- **网点支持哪些业务产品交易？**每种业务种类的业务量占比多少？
- **每天客流量分布如何？**客流高峰期业务种类分布情况如何？
- **针对不同的客户，是否有不同的业务 / 服务流程？**如何提升客户体验？

3. 客户分析

- 网点客户如何分群，客户量如何划分？
- AUM 各级分段客户年龄段分布如何？客户到访网点频率？客户对银行忠诚度，产品认可度如何？

综合以上数据，可以确立增强网点营销能力、提升客户服务体验、提高业务处理效率、降低业务运营成本、加强业务运营管理这五大网点转型提升目标。根据上述部分数据分析方法，结合网点转型目标，可以给出“网点诊断书”，对影响每个网点竞争力得分、业态分类等级和全省排名的重要经营指标进行提示，引导网点扬长避短，认清经营短板，分析失分原因，制定改进措施，明确努力方向，并对网点转型效果进行持续监控。

接下来以两个典型网点为例，进一步说明工作设想。网点情况均为实际生活中提取的情景，不代表真实情况。

根据网点绩效评价体系，网点 A 的网点绩效评价总得分处于区域内中等偏上水平，网点业务量指标评价得分处于区域内中等偏下水平，但网点财务指标评价得分处于区域内优秀水平。经过对数据

洞见数据价值.indd 39　2018/1/15 16:14:16

的分析，可以发现。

- **周边环境：**网点位于区域内的中高端居民住宅区，住户受教育程度普遍较高，人均可支配收入较高，基本没有对公业务。
- **业务分析：**网点日常主要业务种类为个人理财资讯、理财销售，业务高峰期分布较为平均。
- **客户分布：**网点客户年龄段分布均衡，高价值新客户获取难度较大，但存量客户的贡献度较其他网点显著提升。

综合以上分析，可以将网点 A 定位为轻型网点，提出的布局优化建议包括两点：第一，网点布局应兼顾私密性与开放性，提供高价值客户的私密交易空间，开放空间给予客户轻松享受的交流咨询环境；第二，尽可能收集和记录每个客户的附属信息，如家庭人口、工作性质、家庭成员等信息，便于针对客户提供进一步的精准营销等。

再来看看网点 B。该网点的网点绩效评价总得分处于区域内中下水平，并且网点财务指标评价得分和网点业务量指标评价得分也均处于区域内中下水平。数据方面，网点位于区域内的普通街道上，车流、人流量均属一般；网点日常主要业务种类为个人存取款，业务高峰期分布较为平均；网点客户年龄段分布均衡，客户流动性较大，资产持续保持在 10 万元以上客户较少。

图 1-5 展示了网点 B 的个人业务种类业务量分布图，可见，该网点以经营简单结算型业务为主，理财类、资产类等增值性业务乏善可陈，而根据网点的环境、客户等分析，这方面的潜力也十分欠缺。网点 B 恐怕难以实现有效增值、盈利，或许撤并或迁址是更为明智的选择。

洞见数据价值.indd 40 2018/1/15 16:14:16

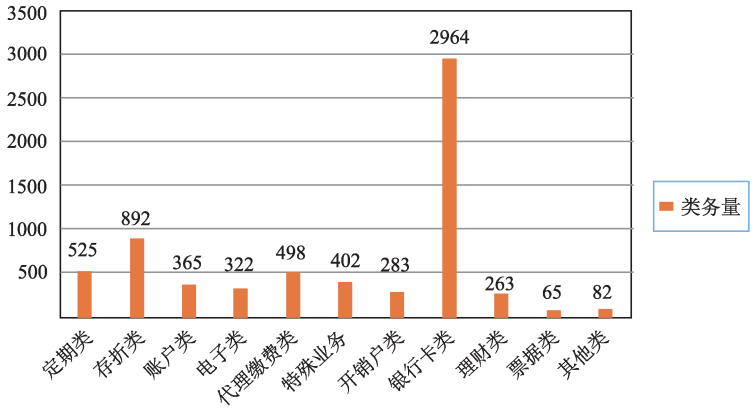

图 1-5　网点 B 的个人业务种类业务量分布

数据来源：团队模拟

制定网点转型诊断书，需配套考虑行内战略目标、客户服务体系、绩效考核体系、联动营销体系，以及同业竞争、市场占有率等维度，并不是一刀切，合适的才是最好的。不过，在当今瞬息万变的环境之下，银行只有积极寻找新的经营模式和盈利模式，才能顺应潮流、满足消费者的需求。“变”才是唯一不变的应对之道。

1.3　大数据在征信业

数据，一直以来就是征信业的基础：征信活动要采集、整理、审核、加工数据，一切都围绕着数据进行。大数据时代的到来，又为征信业提供了新的契机。一方面，海量数据完善了征信业的信息资源；另一方面，大数据相关的概念也在逐步改变着征信业的工作理念。

洞见数据价值.indd 41　2018/1/15 16:14:16

1.3.1 业界展望：FICO 评分与芝麻信用，传统征信向大数据征信的转变

如果你身为“高富帅”或者“白富美”，同时又是个“剁手族”，那么你的阿里巴巴芝麻信用分一定很高。这样，当其他人还在焦急地等待签证消息时，你就可以潇洒地从限量版的爱马仕包里拿出 iPhone 7，单击支付宝，轻触芝麻信用，哇，你的信用分数高达 760！随后你任性地把护照和 iPhone 7 扔在签证官面前，看着他 / 她在你护照上盖章，你微笑着说：“Please charge visa fee directly from my Ant Credit Account（蚂蚁花呗）！”

签证办理方便是最吸引人的芝麻信用特权，不过其他许多生活方面的应用也离不开芝麻信用。如果你想免押金租车、租房，在酒店先入住后付款，都需要一定的芝麻信用分数，更不用说金融方面的信用贷款了……

FICO 评分在中国的推广一直说不上热火朝天，可是在中国有芝麻信用分的人不说上亿，起码也有几千万。这么广泛的客户来源，再加上支付宝的强大覆盖力，FICO 评分这样的传统信用评分还有多大价值呢？会不会在不远的将来被芝麻信用分、腾讯征信评分等取代，或者映衬得黯然失色呢？

其实也不能一概而论。先看看传统的 FICO 评分和芝麻信用分是如何计算出来的。

FICO 评分是 Fair Isaac 公司开发的信用评分系统，也是目前美国应用得最广泛的一种。FICO 评分系统得出的信用分数范围为 300 ~ 850，分数越高，说明客户的信用风险越小，它采集客户的人口统计学信息、历史贷款还款信息、历史金融交易信息、人民银行征信信息等，通过逻辑回归模型计算客户的还款能力，预测客户在未来一年违约的概率。

洞见数据价值.indd 42 2018/1/15 16:14:16

- **人口统计学信息：**如客户年龄、家庭结构、住房情况、工作类别及时间等。
- **历史贷款还款信息：**即过去 6 个月或 12 个月的付款方式、逾期次数等。
- **历史金融交易信息：**即过去 6 个月或 12 个月的平均月交易笔数、金额等。
- **银行征信信息：**如过去 12 个月中新开的账户总数、所有账户的总额度、账户是否逾期等。

以上这些信息都是 FICO 评分模型的自变量，最终会通过逻辑回归模型输出最终分数。不同的是，阿里巴巴推出的芝麻信用分则是以大数据分析技术为基础，采集多元化数据，包括传统的金融类交易、还款数据、第三方的非金融行为数据，以及互联网、移动网络和社交网络数据等，帮助贷款方从多个方面考察个体的还款能力、还款意愿，做出合理、全面的信用评分。

图 1-6 展现了基于大数据分析技术的机器集成学习法 Ensemble 的运行过程。不同于传统的逻辑回归模型，它采集了上万个数据项，从不同的层面（还款能力、还款意愿、欺诈可能性、稳定性等）对个体进行建模打分；再把这些单个层面的评分、结合个体的综合信息，给个体一个最终的信用评分。

两种评分模型采用数据量的不同体现了其评分思路的区别。通常，FICO 评分模型只有十几个评分项，每一个评分项对目标变量（即是否违约）的预测性和影响力都很高。但是，在机器集成学习法中，最终进入模型的评分项可能多达成千上万个，而且每一个这样的评分项对目标变量的单独预测性可能都很小。Ensemble 就是利用机器学习法，把这么多微小的预测性汇总成为

洞见数据价值.indd 43　2018/1/15 16:14:16

最终对个体的违约可能性有很强预测性的评分。

图 1-6 基于大数据分析技术的机器集成学习法 Ensemble 的运行过程

那么，芝麻信用有哪些局限性呢？不妨参照已有的实例进行横向对比分析。美国的互联网金融公司 ZestFinance 从 2009 年就开始研发基于大数据的信用评估模型：融合多源信息，采用机器学习的预测模型和集成学习策略，进行大数据挖掘。他们收集了上千种来源于第三方的数据，例如水、电、煤气账单，电话账单、房屋租赁信息，以及传统的金融借贷、还款信息等，通过机器学习的方法寻找数据间的关联性并对数据进行必要的转换。在关联性的基础上将数据重新整合成不同的测量指标。每一种指标反映个体的某一方面特征，例如诈骗概率、长期和短期的信用风险和偿还能力。最后，将所有指标按加权投票的原则，做成最终的信用评分。

但是 ZestFinance 的个体信用评分只适用于缺乏或没有信贷记录的人群，也就是说，这些人或者刚移民到美国，或者之前从来没有

洞见数据价值.indd 44 2018/1/15 16:14:16

过贷款行为。所以 ZestFinance 的大数据征信最终无法替换 FICO 评分，而只是用来补充 FICO 评分的不足。原因包括多个方面。

- ZestFinance 的大数据征信的体量不大，到现在也只为 10 万美国人提供服务，对模型的有效性、准确性还很难做出有效评价。
- ZestFinance 的大数据模型也给传统的风险管理带来挑战：传统的 FICO 评分需要处理的变量比较少，对模型结果可以给出合理的解释，方便金融机构不同部门之间、金融机构与客户之间的沟通。而 ZestFinance 的基于大数据的数以千计的变量规模和多模型应用，使得数据的处理和模型的解释变得很复杂，在实际应用中会带来许多麻烦。
- ZestFinance 在利用个体消费者的大数据进行信用评估时，很多数据会涉及个人隐私，例如个人社交网络数据（微信朋友圈）、电商交易数据、通话记录等，所以涉及个人隐私的保护和合规性。

阿里巴巴的芝麻信用和 ZestFinance 的大数据征信相似，也存在一定的问题。例如，芝麻信用覆盖的人群可能上亿，但是芝麻信用分的有效性和准确性还没有得到公认的评估；凭借高的芝麻信用分可以在支付宝开通蚂蚁花呗，类似信用卡的透支服务，但是芝麻信用在其他方面的应用还没有达到一定的规模。

当然，大数据信用评分终归是历史的趋势，目前 FICO 公司和国外三大征信机构都已经开始利用大数据分析技术来完善传统信用评估体系的前瞻性研究。例如，益百利 Experian 已经投入研究团队关注社交网络数据对信用评分的影响；FICO 公司也已经开始在线评估的信息工具和基于互联网的信用评估系统的项目研究。随着理论与方法的完善和实践的深入，基于大数据分析的信用评分终有一天将占据主流

洞见数据价值.indd 45　2018/1/15 16:14:16

地位，不过，市场上会不会出现有力的新竞争者，最终赢家究竟是芝麻还是西瓜，敬请拭目以待。

1.3.2 创新方向一：从拒绝推断看个人征信业的大有可为

1. 从“拒绝推断”说起

读过前面的内容之后，你一定已经对“信用评分”的工作方式有所了解。除了打开支付宝就可以看见的芝麻信用评分之外，目前，较为权威的第三方信用评分机构包括 FICO、Vantage、Experian 等，而提供信贷服务的大型银行等往往也会为自己的客户建立评分模型，以便信用管理工作更加高效且富有针对性。

对银行来说，建立有效的客户信用评分模型可能比你想象得更重要。例如，一位客户想从某银行申请一笔贷款，银行是否应该予以批准呢？为了确认客户的还款能力，银行可能需要查验他的个人信息，例如工作情况、收入水平、历史交易行为等。而信用评分模型则是一个信用风险量化的过程，通过可观察到的客户特征变量计算出一个信用分数，评估客户可能的信用风险，并据此将客户归类于不同的风险等级。无论是否已有合适的信用评分模型，最终，银行都会根据获取的信息对客户进行二分类，判断该客户是“好客户”还是“坏客户”，从而对客户申请的信贷等业务选择“批准”或者“拒绝”。

国内多数银行的信用评分模型仍处于建设或更新阶段，这就需要利用大量的历史数据，并寻找合适的预测模型，如 logit 模型、probit 模型等。显然，客户历史数据可以分为两大类：曾经被批准的申请人的信息和曾经被拒绝的申请人的信息。

被批准的申请人和被拒绝的申请人在信息方面的待遇可谓天壤

洞见数据价值.indd 46 2018/1/15 16:14:16

之别。对于曾经被批准的申请人，银行同时拥有这部分客户的申请信息和过往表现信息，可以充分利用这些数据，判断他们是好客户还是坏客户；然而，对于曾经被拒绝的申请人来说，银行仅仅保留了他们的申请信息，但这样的申请人并没有进一步的信用表现，也就无法判断究竟是好客户还是坏客户。因此，在创建信用评分模型时，这一部分客户样本的数据常常遭到忽略。

一个较为完善而有效的信用评分模型的应用对象，显然应该是未来所有的申请人构成的总体，而仅仅由所有被接受的申请人构成的建模样本显然并不是总体的代表样本，所创建的模型自然也存在缺陷。为了弥补这样的缺失，如何推断那些曾经被拒绝的申请人的好坏，并把他们也加入建模样本中，已经成为困扰信用评分领域多年的拒绝推断问题。

那么，为什么拒绝推断问题的研究有助于提升信用评分模型的预测准确性呢？

姑且将被接受的申请人组成的样本称为“接受样本”，将被拒绝的申请人组成的样本称为拒绝样本：不用说，接受样本和拒绝样本中都存在真正的好客户与坏客户。对于他们各自的好坏客户数量如表 1-1 所示。

表 1-1　接受样本和拒绝样本中真正的好客户与坏客户数量

	接受样本	拒绝样本
好客户	nag	nrg
坏客户	nab	nrb

如果使用的建模样本仅仅是接受样本，则该样本中的好坏客户比为 odds1 = nag/nab；相反，如果采用拒绝推断的技术，即同时采纳接

洞见数据价值.indd 47　2018/1/15 16:14:17

受样本和拒绝样本，则该样本中的好坏客户比是 odds2 = (nag+nrg)/(nab+nrb)。而这个模型的优比，显然是模型参数估计过程中必须重视的参数之一。

无论是接受样本还是拒绝样本，其中包含的客户都是通过一定的审批机制来决定到底是被接受还是被拒绝的。这个审批机制可能是一个专家评分模型，也可能是原本使用的信用评分模型。所以，拒绝推断是否必要和原先的审批机制如何运作显然不无干系。接下来模拟三种情景，看看拒绝推断究竟什么时候不可缺少。

情景 1：银行 A 原本在审查是否应该发放给某位客户贷款时，既不看客户信息，也不做深入调查，反而以扔硬币的方式来决定。好吧，不用说，该银行的审批机制完全无效，坏客户在接受样本和拒绝样本中肯定是以随机等可能出现，上面提到的 odds1=odds2，是否使用拒绝推断对模型的预测效果毫无影响。

拒绝推断在这种情景下并没有什么用，不过这种银行真的存在吗？

情景 2：银行 B 负责审批客户的是一位资深业务专家，这位专家这辈子没出过错，凡是接受的都一定是好客户，拒绝的都必然是坏客户。

显而易见，银行 B 的 odds1 趋近于正无穷，与 odds2 相差太远，引入拒绝推断很有必要。不过，这么神奇的业务专家大约还没出生，所以理想情景也很难在现实中出现。

情景 3：经历了以上两种超现实主义银行，常见的业务情景其实是：原有的审批机制有效，不如情景 2 中的完美，但也可以做到使接受样本中绝大多数为好客户，拒绝样本中绝大多数为坏客户。

在实际应用过程中，银行会把客户的信用评分从高到低排序分

洞见数据价值.indd 48 2018/1/15 16:14:17

成若干个评分池，如图 1-7 所示。

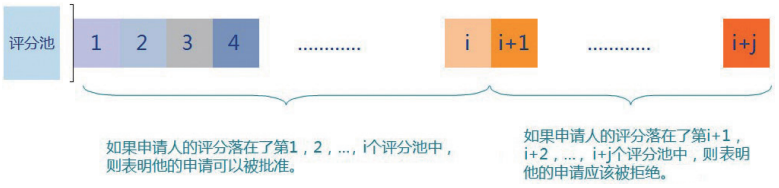

图 1-7　评分池

银行通常会接受落入前 i 个评分池的客户，而拒绝其余评分池中的客户。因此，一个有效的评分模型应该能够保证：分数高的评分池中的好坏客户比要大于分数低的评分池中的好坏客户比。这时，odds1 >odds2，仅使用接受样本的好坏比大于使用了拒绝推断的建模样本中的好坏比，引入拒绝推断是明智之举。

2. 拒绝推断的实证分析

在信用评分模型中引入拒绝推断的目的，是要推断出拒绝样本中那些被拒绝的申请人到底是好客户还是坏客户，从而将之用作建模样本数据。因此，可以认为，填补拒绝样本的数据就是一种缺失值处理问题。

对拒绝推断问题进行实证研究面临的最大困难是数据的获得性。在此，利用某银行信用卡业务的数据来做拒绝推断的实证研究，并通过实验设计的思路判断拒绝样本中的申请人到底是好客户还是坏客户，如图 1-8 所示。

这样，验证样本就相当于入门总体的有效代表。

为了充分论证、判定拒绝推断对信用评分模型的提升作用，首

洞见数据价值.indd 49　2018/1/15 16:14:17

先仅以接受样本创建信用评分模型，再采用核函数推断法、打包法（parceling）、硬截止法和外部数据法对拒绝样本进行推断，汇总接受样本和推断出因变量取值的拒绝样本，从而创建信用评分模型，并分别验证这 5 种信用评分模型的效果。对于需要用到先验信息的方法，统一假设被拒绝的申请人中好坏客户的比例为 8∶2。

图 1-8 某银行信用卡业务的数据

5 种方法展示如下。

方法一：不做拒绝推断，只是使用接受样本创建信用评分模型，并用验证样本做模型效果的验证，如图 1-9 所示。

图 1-9 创建信用评分模型（一）

方法二：用最相似、加权平均和 Q1 加权平均这三种核函数推断法对拒绝样本进行拒绝推断，并把推断出因变量取值的拒绝样本和接受样本汇总后，再创建信用评分模型，然后利用验证样本做模型效果的验证，如图 1-10 所示。

对于每一种核函数推断法又有三种子方案。

- 选择样本中的全部自变量（25 个）做核函数推断。
- 选择部分精选的自变量（18 个）做核函数推断。

洞见数据价值.indd 50 2018/1/15 16:14:17

- 选择全部自变量（25 个）做核函数推断。

图 1-10　创建信用评分模型（二）

方法三：用信用评分领域的**打包方法**对拒绝样本做拒绝推断，并把推断出因变量取值的拒绝样本和接受样本汇总后再创建信用评分模型，然后利用验证样本做模型效果验证。

打包方法首先利用接受样本创建初步的信用评分模型，并把预测概率排序分组，然后给拒绝样本中的申请人打分，并对打分得到的预测概率按照接受样本中的预测概率分组规则进行分组。该方法假设在同一概率组中，拒绝样本中的坏客户比例是相对应的接受样本中坏客户比例的若干倍，这个倍数就叫作事件增长率。事件增长率需要业务人员根据经验给出估计，是一种先验信息。

方法四：用信用评分领域的**硬截止方法**做拒绝推断，并把推断出因变量取值的拒绝样本和接受样本汇总后再创建信用评分模型，然后利验证样本做模型效果验证。

硬截止方法首先利用接受样本创建信用评分模型，并据此给拒绝样本中的申请人打分。该方法假设得分高于某个临界值的为好客户，低于临界值的为坏客户，这里的临界值也需要业务人员给出坏客户率的先验估计。

方法五：把**外部数据**（被拒绝申请人究竟是好客户还是坏客户）加入建模样本中，然后创建模型，最后利用验证样本做模型效果验证，

洞见数据价值.indd 51　2018/1/15 16:14:17

如图 1-11 所示。

图 1-11 创建信用评分模型（三）

实证结果如表 1-2 所示。

表 1-2 信用评分模型的实证结果

拒绝推断方法		AUC	K-S	GINI	LIFT
只使用接受样本，无拒绝推断		0.600	0.161	0.200	6.721
核函数推断	最相似（25 个自变量）	0.617	0.184	0.234	5.706
	最相似（18 个自变量）	0.618	0.167	0.236	6.043
	最相似（25 个自变量，分组）	0.612	0.167	0.224	6.149
	加权平均（25 个自变量）	0.611	0.184	0.222	8.202
	加权平均（18 个自变量）	0.630	0.218	0.260	6.673
	加权平均（25 个自变量，分组）	0.614	0.179	0.228	7.450
	Q1 加权平均（25 个自变量）	0.612	0.184	0.224	6.727
	Q1 加权平均（18 个自变量）	0.625	0.195	0.250	6.515
	Q1 加权平均（25 个自变量，分组）	0.617	0.188	0.234	6.604
打包法		0.626	0.255	0.252	5.125
硬截止法		0.627	0.307	0.254	2.952
外部数据法		0.710	0.301	0.420	3.247

从表 1-2 可以看出，由于样本数据的限制，拒绝推断的实证分析结果并不是很理想，除了 LIFT 值以外，AUC 统计量、K-S 统计量和 GIN 系数的值都偏小，这表明该模型的预测准确性相对偏低，但

洞见数据价值.indd 52 2018/1/15 16:14:17

仍可以看出，使用外部数据法来做拒绝推断的验证指标的值最大，它是提升信用评分模型预测准确性的最有效方法。然而，这种方法最耗费人力和物力，银行一般不愿意承担获取真实信息的风险和成本。

现在总结一下：在构建信用评分模型时，如果仅仅使用被接受的申请人构成的样本，则创建的信用评分模型必然存在样本偏差；为了获得更准确的预测结果，必须要做拒绝推断来校正这种偏差。

通常情况下，当对接受样本和拒绝样本所做的一些总体假设正确有效时，拒绝推断方法一般都能够提高模型的预测准确性，尤其是使用外部数据提供额外信息的方法，可以保证信用评分模型更加准确、合理；然而，受限于国内征信体系的发展不够充分，银行获取信息的风险和成本都较高，也就限制了模型准确性的提升幅度。

因此，如何解决当前信用评分模型面临的困境呢？建设完善的征信数据正是最优选择。它可以免除商业银行为了获得拒绝样本的信息而承担的风险。在国内大力推广征信体系意义重大。

3. 征信数据市场的巨大意义何在?

征信系统的逐步发展必将为信用评分体系带来整体提升：银行利用已有的客户信息，结合有权威性的外部征信信用报告，则一定可以得到更优的信用评分模型，从而降低信贷风险，提升效率。进一步说，完善的征信体系可以促进国民的良好还款行为习惯，增进国家的金融稳定性，并鼓励银行业的持续发展；对于个人而言，信用记录良好的客户也可以借此获得更公平的信贷机会，以及诸多便利甚至褒奖。

洞见数据价值.indd 53　2018/1/15 16:14:17

其实，“征信”已经不是个新鲜事了，从银行贷款、消费金融到租房租车，个人信用已经开始影响社会生活的各方面。早在 2013 年 3 月，国务院颁布了《征信业管理条例》，初步构成了征信体系的法律基础；截至 2014 年年底，中国人民银行征信中心已经接入了 1811 家数据机构，并覆盖了 3.5 亿有信贷记录的人口。然而，由于我国经济体量巨大，目前的征信数据规模和细化程度远远不够完善，并且已经直接影响了社会融资成本、放贷效率和行业的抗风险能力，抑制了经济活力。无论未来的征信业仍由央行主导，还是会催生更多具有一定权威性的商业机构如 FICO 或 Experian 一般自成体系，征信业短期之内必将大有可为，而数据资源的供应与数据分析的能力一定会在其中发挥自己的专长。

1.3.3 创新方向二：论大中型客户数字化授信的可行性

前面已经介绍了不少关于个人客户征信的工作：银行可以依赖信用记录和风险评价模型，对个人展开授信。在互联网金融的大趋势下，众多小贷公司开始使用大数据授信的方式，例如，打开微信钱包里的微粒贷，就能看到自己的可借额度。

授信当然不止于个人。目前，数字化授信的适用范围已经延伸到小微企业，银行通过企业自身的生产经营数据，例如第三方获取的水电、公检法、银税，以及企业主个人征信、交易、资产、关系信息等，给小微企业发放闪电贷。在毕马威 FINTECH50 强的调研中，发现个人和小微企业都可以采用数字化评审的方式，最快 5 分钟就能完成 50 万元以下的贷款审批。

那么，大中型企业是否可以同样采用数字化授信的方式呢？近年来，企业贷款不良率不断攀升，逐渐成为各大银行的业务痛点，

洞见数据价值.indd 54 2018/1/15 16:14:17

数字化授信是否正是解决问题的答案呢?

先来看看国内大行通用的大中型客户授信审批方式：

（1）客户经理找客户拿资料，跟管理层聊聊天找感觉，回来写上数十页的尽调报告。

（2）风险经理平行作业，一定程度上对客户资料的真实性负责。

（3）评审经理对尽调情况展开风险分析，酌情让客户经理补充调查，最终确定授信方案。

（4）各位老大审批再审批。

显然，整个链条中最为重要的环节是客户经理和评审经理，但双方依然面临着信息不对称的问题，例如，客户是否有欺诈行为？客户经理是否有隐瞒行为？评审的分析视角是否合理？这些问题都是传统授信审批方式无法解决的。

下面论证一下大中型客户，尤其是一般企业的数字化授信审批的可行性。

大中型企业的授信重点要看财务数据，而目前财务报表掺水和造假的问题层出不穷，但财务欺诈风险可以通过挖掘分析和业务规则验证在一定程度上解决。首先，可以采用逾期、不良、重组等问题客户的财务报告作为数据开展挖掘分析，重点关注企业出问题之前的一段时间内客户财务指标的异常表现，例如客户财务指标自身同比 / 环比与行业平均值和行业变化情况相比的差异或特征等，从而发现疑似欺诈行为和信用风险。如果发现该客户的毛利率与行业反向变化且与行业平均值的差异达到 2 倍及以上这样的问题，就可能识别其中的风险。

而业务规则归纳总结的重要意义，在于将专业评审多年积累的经验进行提炼，并转化为系统可以识别的规则，同样可用于识别欺

洞见数据价值.indd 55　2018/1/15 16:14:17

诈和信用风险。例如，可以发现企业固定资产投资方面的异常，如果企业把经营活动和筹资活动所产生的现金大量投入到固定资产，却没有伴随主营业务收入的增长（或者固定资产的增长率大幅度大于主营业务收入的增长幅度），就可能存在虚假问题或者经营异常。

目前，大中型企业的授信往往是由个人的经验决定的，由于评审经理看过的企业多了，积累了很多行业对标值，可以做出较为准确的判断，但一个新人却很难在短期内做到这一点。而大数据擅长的就是将行业数据进行加工汇总，形成各类对标值。例如，可以利用历年上市公司的财务数据形成按行业细分的各类指标的对标值，其中，A 股 H 股可以作为一类加工对标值，新三板可以作为一类加工。各行各业还有自己独特的指标，例如风电行业的单位总装机容量、并网装机容量，这些指标有些需要从行业协会获取，有些需要从企业年报获取。获取并分析这类指标的目的是验证企业描述的愿景和蓝图，让我们可以知道，当申请贷款企业的设备运转规模达到最大值的时候，企业的销售收入可能会是什么样的。

以上介绍的是大数据和挖掘分析手段在财务分析与经营分析中可以解决的部分问题，不过，大中型企业的授信决策不只是欺诈识别和对标分析，还涉及关联分析、舆情分析、征信分析、还款来源及综合收益分析等，这些方面的分析都可以不同程度地应用数据挖掘方法。

然而，大中型企业的贷款规模都在千万级，没有哪个银行敢于单纯使用数字化审批的方式完成，从授权体系和责权利对等的角度来说，管理上无法实现。而且，当今企业混业经营趋势明显，商贸型企业的涉猎范围更广，量化模型理论上可以实现细化到产品的对标及分析，但实际操作中样本量会有一定限制，系统维护成本又非常高，从成本效益原则讲，银行基本没有开发意愿。所以，结论是，

洞见数据价值.indd 56 2018/1/15 16:14:17

大中型企业客户的数字化审批无法完全实现，不过可以给出风险提示和决策支持的建议，最终交由专业评审进行审批。这样，评审中人员的经验缺失将不再成为短板，评审的效率也可以大幅度提升。

1.4　大数据在审计业

审计是一个已经发展了上百年的行业，而每次经济和科技的进步，都在不断推动着审计的发展和演变。随着企业海量数据的增长以及大数据技术的逐步成熟及推广应用，大数据和大数据分析及业务流程自动化也正为传统审计提供新的思路、技术和方法。

1.4.1　业界展望：大数据分析如何支撑审计工作

大数据和大数据分析作为近年来火遍全球的热门词汇，无疑正在对全球经济及社会发展带来巨大的影响，每个行业都在此背景下开始重新审视面临的机遇与挑战。

大数据分析可以为审计带来更多、更大的价值，这已经是业界的普遍共识。通过数据分析工具与分析方法，可以帮助审计人员更高效地识别审计中的舞弊和风险，开展更加具有针对性的深入检查。数据分析不仅是一种工具，更是一种基于数据的问题分析和思维方式。从实践操作中来看，数据分析技术为审计带来的帮助大致上可以概括为几个方面：提高审计效率，降低审计风险，提升审计价值；在工作内容、展示和传递三方面，数据分析都能提供巨大的价值提升；现有的 RPA（Robotic Process Automation，机器人流程自动化）能够将大数据分析过程或结果嵌入工作流程中以提高审计工作效率，减少人工劳作。

洞见数据价值.indd 57　2018/1/15 16:14:17

1. 工作内容：全量数据覆盖

在传统审计中，对于大样本量的审计，往往依赖于风险评估后的随机抽样分析。抽样过程中，为提升抽样效率，往往需要对风险、重要性水平、样本特性等多种要素进行综合分析，但抽样分析是由于信息缺乏条件下采用的一种方法，伴随其中的抽样风险本身是无法避免的。

大数据环境和技术下，可以采用全量分析，对审计的意义不单是抽样方式的变化，也将审计视角从局部提升至全局的巨大改变。并且，基于特定审计目标和分析场景可以开展数据分析建模。例如，通过模型对数百万笔贷款逾期概率进行测算，将其中的高风险贷款作为样本进行重点审计与测试。

2. 展示：可视化与交互式展示

多维钻取和可视化也可以为审计提供帮助，通过 BI 展示，企业的经营异常可以很容易被揭示。对于分析种类繁多、关系复杂的情况，依靠传统的 Excel 透视方法很难展现，大量的数据表、繁乱的关系图，加大了审计人员对信息读取的难度。因此，无论是对于审计师还是被审计公司，可视化分析能够直观地呈现大数据特点，同时非常容易被接受，看图说话，简单明了。

例如，在采购交易分析中，通过对销售与采购订单价格与市场价格绘制比较曲线，并通过选择不同分析维度，诸如地域、人员、岗位、时间、品类等过滤筛选条件，可更加直观地对异常波动和交易进行识别，如图 1-12 所示。

图 1-12 展示了某采购公司提供的价格与万得市场价格之间的对比。将采购货品的市场价格数据与客户实际采购价格进行匹配与比

洞见数据价值.indd 58 2018/1/15 16:14:17

对，获取价格波动异常，以及与市价差异较大的交易订单。

material	ddate	price	meihua_p	winddata_p	sub_p2	currency	amount	market	location	buyer	sub_type
褐煤_燃料	2014年1月24日	76.0684	.	.	.	CNY	87.58	内蒙古XXXXXX	锡林郭勒盟	邓XX, 03	0.00%
褐煤_燃料	2014年1月24日	76.0684	.	.	.	CNY	128.68	内蒙古XXXXXX	锡林郭勒盟	邓XX, 03	10.00%
褐煤_燃料	2014年1月24日	76.0684	76.07	122	45.93	CNY	400000	扎鲁特XXXXXX	锡林郭勒盟	邓XX, 03	20.00%
褐煤_燃料	2014年1月24日	76.0684	.	.	.	CNY	1082.54	内蒙古XXXXXX	锡林郭勒盟	邓XX, 03	30.00%
褐煤_燃料	2014年1月24日	76.0684	.	.	.	CNY	3545.9	内蒙古XXXXXX	锡林郭勒盟	邓XX, 03	40.00%
褐煤_燃料	2014年1月24日	76.0684	.	.	.	CNY	3896.8	扎鲁特XXXXXX	锡林郭勒盟	邓XX, 03	50.00%
褐煤_燃料	2014年1月24日	76.0684	.	.	.	CNY	799.46	内蒙古XXXXXX	锡林郭勒盟	邓XX, 03	60.00%
褐煤_燃料	2014年1月24日	76.0684	.	.	.	CNY	3742.82	内蒙古XXXXXX	锡林郭勒盟	邓XX, 03	70.00%
褐煤_燃料	2014年1月24日	76.0684	.	.	.	CNY	2548.54	扎鲁特XXXXXX	锡林郭勒盟	邓XX, 03	
褐煤_燃料	2014年1月24日	76.0684	.	.	.	CNY	220.62	扎鲁特XXXXXX	锡林郭勒盟	邓XX, 03	
褐煤_燃料	2014年1月24日	76.0684	.	.	.	CNY	200000	内蒙古XXXXXX	锡林郭勒盟	邓XX, 03	
褐煤_燃料	2014年1月24日	76.0684	.	.	.	CNY	2371	内蒙古XXXXXX	锡林郭勒盟	邓XX, 03	
褐煤_燃料	2014年1月24日	76.0684	.	.	.	CNY	1446.64	扎鲁特XXXXXX	锡林郭勒盟	邓XX, 03	

图 1-12　在采购交易分析中通过选择不同分析维度直观地对异常波动和交易进行识别

数据来源：团队采集 / 万得平台

多维度下的趋势分析、异常分析以及占比分析是经常使用的基础分析方法。在实际操作中，异常分析通常采用如下分析方法。

第一，同业比较，此类适用于同质化和标准化较高的业务场景，例如大宗商品采购、销售等，此类市场价格数据也较为容易获得。

第二，企业内部数据均值比较和时序性比较。通过业务范围内的横向比较和基于时间序列的纵向比较，识别不同要素的异常值表现。

3. 传递：传输审计价值

在审计全流程中，数据分析可以应用在不同的工作阶段：在审计计划阶段，可以基于数据分析的风险评估，锁定高风险领域制定

洞见数据价值.indd 59　2018/1/15 16:14:17

审计计划，并提供灵活、交互式的可视化工具，实现高效沟通；在内部控制测试阶段，数据分析可帮助识别潜在的控制薄弱环节，对特定控制进行全量数据的自动化测试；在实质性测试阶段，数据分析支持多维度的分析性复核，筛选高风险测试样本；在报告与完成阶段，数据分析更可构建关键预测模型，为业务的经营及风险管理提供增值建议。总之，数据分析已经渗透进审计工作的方方面面，可担当审计工作的长期伴侣与指导，如图 1-13 所示。

图 1-13 数据分析在审计全流程中的应用

除了对于审计师带来的帮助外，数据分析也完全可以为审计客户带来增值。借助数据分析，审计师在审计过程中可以识别更多的管理及经营问题，协助企业进行改善。例如，通过对国内某大型企业的数据关联分析，可以发现项目投资、实际价格、供求关系存在一定程度的不匹配，审计团队可基于数据分析发现，与企业就项目投资事项进行进一步询问与调研，这无疑协助企业更好地节约成本。

洞见数据价值.indd 60 2018/1/15 16:14:18

4．流程：智能化控制提高效率

大数据分析已经拓展到人工智能领域并逐步代替手工操作。那么审计领域的智能化现已全面展开，以一系列 RPA 工具为应用实现的方法已经在客户的实际场景中开展。场景一：智慧风控。通过引入人工智能学习算法，持续优化财务管理分析指标及管控体系。在快速支撑总体业务发展的同时，对高风险领域进行有效识别及智慧管控。场景二：智能报表。在“智动核算”的帮助下，财务审计人员能够有效实现人员转型，将主要精力聚焦于管理决策上。在数据挖掘技术的帮助下，财务审计分析视角将更具智能化。场景三：智动核算。通过 OCR 及语音识别等技术数字化流转单据及会计档案，归集大数据。机器人技术的出现能够快速串联不同的财务系统及业务流程，在快速落地的同时满足财务信息系统的统一管理要求。

笔者所在团队曾多次利用大数据分析帮助审计工作，并取得可观的成果及客户的认同。团队已经逐步开始最新的人工智能技术研发工作，将大数据分析智能化，未来将能够执行只有人类智能才能够做到的任务，包括但不限于视觉感知、语音识别、自然语言处理、不确定条件下的决策、学习和语言之间的翻译等任务。人工智能技术完全可以为审计与审计客户的合作提供更深入、更全面的创新与变革。

1.4.2　创新方向：大数据能否代替传统审计?

对于传统行业来说，每一种新的思想与技术往往既是挑战。RPA 大量使用了大数据技术，而 RPA 应用场景目前主要是财务和审计等领域，这里通过一个与传统审计工作结合的例子来说明。审计的目的是从正常中发现异常，数据类型的复杂化与数据量的急剧增加，

洞见数据价值.indd 61　2018/1/15 16:14:18

增加了审计工作的难度。在这样的背景下，传统审计工作必然需要寻求新的方法来优化传统审计工作。RPA 应用大数据技术，善于对企业的正常运作状态进行描绘、分析和预测，运用模型从数据中识别模式、关系、趋势和波动。在大数据的帮助下，审计工作者更容易发现可疑的异常点，提升审计工作的效率，可以说，大数据分析与传统审计工作的结合真是一拍即合。如今，越来越多的企业开始接受并重视大数据的理念，也越来越欢迎 RPA 流程及数据分析对审计工作的支持。

然而，蜜月期后，审计工作者也开始担忧：传统审计是否完全会被 RPA 和大数据技术完全替代？作为大数据分析领域的专业团队，曾经帮助数十个审计团队攻关大数据审计课题，根据经验，大数据分析不可能完全替代传统审计，两者是有益的互补。

下面以银行的信贷资产质量检查为例来阐述观点。近年来，由于经济处于下行，企业客户出现大面积亏损、停产甚至倒闭，银行的信贷业务承受着巨大的压力，对于信贷资产质量的检查也成为对银行业审计工作的重中之重。

大数据分析可以用哪些方式协助对信贷资产质量的检查工作呢？最常用的一个落脚点就是用假设检验来分析不良资产产生的原因。也就是说，大数据分析师往往会假设一些可能会造成信贷资产不良的原因，然后针对每个假设分别搜集证据，检查这些证据的充分性。如果证据确凿，那就可以认为这个假设是造成不良的原因，如果证据不充分，那么这个假设就不成立：这个过程与法院断案相似。

哪些因素可能成为假设检验的假设呢？这些假设可以是宏观层面的，也可以是微观层面的。举例来说，从宏观层面来看，是不是

洞见数据价值.indd 62 2018/1/15 16:14:18

2008 年全球经济不景气以后国家采取的经济刺激计划可能导致近几年大面积的信贷资产不良？或者是不是由于 2013 年开始 GDP 增速减缓，而银行的新贷款量却未减少，这样的产能过剩带来了大量不良贷款？从微观层面来看，是不是因为在当前的经济环境下企业的经营遭遇了困境？是不是客户的信贷审批存在欺诈行为？例如，企业是否刻意掩盖其经营状况、贷款用途，寻求更多贷款，分行客户经理是否存在为完成 KPI 或其他个人原因而为客户作假的情况？可能造成不良的假设很多，如何形成这些完整的假设呢？可以多问问几个“W”：Why——明确最关注什么，痛点在哪里；When——明确要分析哪个时间段的不良客户，且客户的信息和行为需要追溯到什么时间；What——明确分析什么内容，分析的维度有哪些，是机构、行业、规模还是产品；Where——梳理信贷全生命周期，明确问题点从哪里来。对于每一个假设点，大数据分析都需要逐个展开分析。

为了更好地阐述观点，就以评级结果的更改次数和客户的资产质量之间的关系来具体说。银行制度要求录入客户基本信息、关联信息、财务信息等进行客户信息评级，但客户经理可能通过反复测试的方法使客户达到最高评级。不妨假设反复更改评级输入数据的客户资产不良比例高于正常客户。如果假设成立，则说明在执行制度的过程中存在缺陷，应该进一步规范客户经理的行为，或在系统中做一些操作限制来达到降低风险的目的。

针对这个假设，需要搜集的证据就是评级阶段客户信息的修改日志，将客户分为两类，一类是该阶段没有修改过信息的或修改次数小于两次的客户，另一类为被修改信息两次及以上的客户，再将以上两类按正常客户和不良客户分组，交叉后最终一共得到四类。在此基础上，可做的分析有：不良客户中曾反复修改信息的和无反

洞见数据价值.indd 63　2018/1/15 16:14:18

复修改的有无明显差别，曾反复修改信息的客户正常的与不良的比例如何等。假设检验的方法可以采用卡方检验。如果数据可以证明在有反复修改的客户中，不良的比率远高于正常客户，则可以论证信用评级阶段反复修改客户信息对不良的成因有显著影响，并建议对信用评级阶段的操作细则做进一步梳理细化，或彻底禁止对同一客户多次发起信用评级，在制度设计层面动手术。

再举一个例子，如果评级模型的有效性对资产质量高低有影响，搜集的证据包括是否存在一些对区分信贷资产质量有显著能力却没有纳入评级模型的指标。例如，来自企业财务的报表可以用于评价企业回收现金能力的“资产现金回收率（= 经营现金净流量 / 平均资产总额）”和“收入变现比率（= 经营活动产生的现金净流量 / 营业收入）”这两个指标。如果数据分析可以证明不良客户在这两个指标上的均值与正常企业有明显差别，甚至这样的差别显著强于现有的评级模型，那就说明须考虑在信用评级模型中增加此方面因素。

从上面这些具体的案例可以看出，假设检验的结论是需要基于统计显著性的，而不是基于个案的。换句话说，如果某个原因的确导致信贷资产不良，但是出现这样问题的案例并不多，那就是证据不充分、不具备统计显著性，大数据分析无法将之判断为一个导致不良的原因；而传统审计是基于个案来研究的，只要存在这样的问题，哪怕只有一个案例，就可以认定是导致不良的原因，需要注意避免。这样就可以很好地回答题目中的问题：大数据分析能完全替代传统审计吗？答案就是不能。借助大数据可以关注宏观的趋势和动态，避免审计工作中只见树叶、不见树林的效率缺陷，但审计的对象毕竟仍是林中一片片具体的叶子。以上面所述的评级工作来

洞见数据价值.indd 64 2018/1/15 16:14:18

说，审计人员可能会发现，当企业客户所处行业面临压缩，甚至退出的危机时，少数客户经理会帮助篡改客户信息，擅自变更客户所属行业，为申请授信和最终放款打开方便之门，但在大数据分析中，只有当这种行为蔓延到大多数客户经理身上并在数据统计呈现的结果中出现明显偏离时，才可能被识别，对个体行为的识察是相对乏力的。

因此，传统审计工作与大数据分析之间无法相互替代，需要很好地互补。审计工作聚焦微观层面，能够对每一个评级审批人员进行询问，对每一份贷后检查报告的真实性进行核对，对每一个押品的估值和存在性发询证函；而 RPA 的应用可以通过预测模型分析出哪类资产质量可能偏低，为审计人员指明方向，使他们有的放矢、提高效率。大数据分析可以评价信贷客户的获取能力、外部环境变化对资产质量的影响大小、内部某个制度设计缺陷与资产质量之间的相关性，从而以整体特征和趋势的角度发现造成资产质量低下的内外部影响因素。审计则从个体角度，在高风险领域开展核查工作，利用抽样的方式对合同、借据、押品、申请人、审批人等逐一排查，发现不合规的行为……总之，大数据分析不能完全替代传统审计，两种方法将形成一种更有效率的工作机制，相辅相成。

1.5　大数据在传统制造业

对于大数据来说，国内的制造行业还是一片较为新鲜的领域，但与众多新兴行业相比，它也同样不容忽视。制造业不但具有丰富的案例和浩瀚的数据，更拥有极为成熟的体系和广阔的市场，这一切都有待大数据和大数据技术逐步发掘，为传统行业带来新的提升点。

洞见数据价值.indd 65　2018/1/15 16:14:18

业界展望：数字化企业进阶指南[①]

1. 什么？你出门带手机就够了？

2016 年 4 月，从慕尼黑来的小伙伴吵着要去吃火锅，笔者用滴滴打车叫车、支付，到海底捞支付，路上还帮他订了一部小米手机，这全靠手里的 iPhone，相比于 2 月在慕尼黑，3 月在维也纳没带足现金又不能移动支付，只能求救于小伙伴的窘境，嗯，君子报仇，一月不晚。数字化时代，就是这么任性，随身不带现金！

近年来，技术革新层出不穷，各种概念和解决方案带领我们进入全新的数字化时代，如制造业的“工业 4.0”（Industry 4.0）、物联网（The Internet of Things）、云计算（Cloud Computing）、大数据（Big Data）、移动（Mobility）、社交（Social）以及在数据传输过程中的安全（Security）等，令人眼花缭乱。不过，等一下，透过现象看本质，无论是基础层面的软件驱动硬件、系统互联、虚拟化，还是应用层面的移动支付、社交网络，不管怎么玩，还是在处理数据。

一起回顾一下之前的数据处理关键历史时刻：

1890 年，赫尔曼·霍尔瑞斯[②]应用卡片制表系统为美国人口普查局做人口普查，从此开启了系统处理数据的时代。

1951 年，莫奇利（Mauchly, John William）和埃克特（Eckert, John Presper, Jr.）设计的第一台通用自动计算机 Univac I，第一次采用磁带机作外存储器，数据处理进入了快车道。

1977 年，埃里森以 IBM 发表“关系数据库”的论文研发出甲骨

① 本节由西门子公司系统专家栗耀平提供。

② 赫尔曼·霍尔瑞斯（Herman Hollerith，1860—1929），美国统计学家，被誉为数据处理业的创造者，在人口统计局工作时发明了卡片制表系统，被认为是现代计算机的先驱。参见http://www-03.ibm.com/ibm/history/exhibits/builders/builders_hollerith.html。

洞见数据价值.indd 66 2018/1/15 16:14:18

文数据库管理系统（Oracle DBMS），大规模数据处理从此一发不可收拾。

2. 少年，还没信息化你就想数字化?

数据处理的关键是要有数据，所以数据的来源就成了首先要解决的问题。各位看官，你见过一个企业一出生什么数据都没有，就已经是数字化企业吗?

数据的产生来源于信息化，从最开始企业上 OA、ERP 等系统，把线下手工流程转变为信息系统控制流程，信息由纸张记录转变为计算机存储，以及各个系统的集成，如 OA 与 ERP 集成实现用户账号、权限审批自动化，ERP 与 MES 集成实现生产自动化等。

要信息化，就不得不提到理查德·诺兰[①]在 1974 年提出的信息系统进化的阶段模型，即诺兰模型。他认为，从手工时代向以信息技术为基础的数据处理发展时，包括四个阶段：初始阶段、普及阶段、控制阶段和成熟阶段。1979 年，诺兰又在此基础上额外添加了两个阶段，成为六个阶段：初始阶段、普及阶段、控制阶段、集成阶段、数据管理阶段和成熟阶段。

3. 信息化进阶指南

阶段一：初始阶段

企业开始意识到采用信息技术处理数据的重要性，有些部门开

① 查德·诺兰（Richard L. Nolan），曾任哈佛商学院教授，现任华盛顿商学院的管理与组织学教授。
以下两篇文章可帮助你进一步了解他的理念：
Managing the Four Stages of EDP Growth
https://hbr.org/1974/01/managing-the-four-stages-of-edp-growth
Managing the Crises in Data Processing
https://hbr.org/1979/03/managing-the-crises-in-data-processing

洞见数据价值.indd 67　2018/1/15 16:14:18

始采用信息技术开发专用系统，特别是财务部门会自己开发或采用成熟组件，实现部分业务的自动化，在这个阶段，对数据处理费用缺乏控制，信息系统的建立往往不讲究经济效益，部分用户对信息系统还处于观望态度。

阶段二：普及阶段

在看到一些部门采用信息技术后带来效率的提高，如财务部门采用信息技术明显提高了月结速度，准时准确关账，如人事部门采用信息技术实现将加密的工资条以邮件形式自动发送给员工等。随着信息技术应用在企业的一步步深化，各个部门都开始尝试着将信息技术应用到自己的日常工作中。这个时候，企业的管理者开始关注信息系统的投资收益，各个部门分别管理自己的信息化，没有统一的管理。

阶段三：控制阶段

当企业的系统越来越多，信息孤岛就会出现，相同的信息以不同形式存储在不同系统，相互不兼容，重复建设。出于统一规划、统一管理，信息技术的管理就会成为一个部门，以实现信息技术的管理以支持企业业务的运行，会采用项目管理计划和系统发展方法。信息技术开始被认为是企业发展重要的支持力量，为接下来的信息系统发展打下坚实基础。

阶段四：集成阶段

信息不对称，造成了信息孤岛的出现，相同的信息在不同部门的不同系统里重复出现，并且常常以不同的格式出现，造成了信息本质上的冗余。在这个阶段，打通企业各个部门的业务处理环节，按照数据的类型，统一数据存储，如主数据单独存储，共享使用；集成 ERP 和 MES 实现生产、库存以及采购等部门的协同工作，效

洞见数据价值.indd 68 2018/1/15 16:14:18

率大大提升。

阶段五：数据管理阶段

企业内部资源，如人力资源、财务和物流，都以数字形式存储在信息系统中，在这个阶段，企业开始考虑信息化建设的成本与收益、效率与效果。

阶段六：成熟阶段

企业开始将信息化作为企业不可或缺的一部分，开始采用成熟企业架构（如 TOGAF、Zachman 框架），实现信息技术和业务的整合，企业的信息技术战略实现业务战略。

各位看官，通过以上分析，您对所在企业的信息化现在处在哪个阶段、下一步该怎么走心里一定有数了。前期充分的信息化准备，将为下一步的数字化打下坚实的基础。只有实现数字化，才能让企业走得更远。

1.6　大数据在互联网行业[①]

在这个万物皆联网的时代，各类互联网企业风生水起：他们原本就依托数据而生，也需要大数据的进一步支持。

创新方向：从滴滴收购优步看垄断企业的马太效应

周末聚会之后，笔者与朋友叫了一辆优步拼车回家，因为比不拼车便宜 6 块钱，虽然这意味着要绕路接人，20 分钟的路也许要走

① 本节部分内容来源于已发表文章：Wei Qiuping, Zhang Bo, Matthew Effect in Mononpoly Industry, *Recent Advance in Statistics Application and Related Areas – Conference Proceedings of 2008 International Institute of Applied Statistics Studies*, 2008, 1313—1316.

洞见数据价值.indd 69　2018/1/15 16:14:18

40 分钟。优步 APP 显示，司机会先来接我们，然后从一个崎岖的地方接一位“丽”。

车来了，笔者与朋友两个坐在后排，然而路上出了点差错：快到接人地点时，“丽”的手机忽然打不通了。不是占线，也不是没人接，而是长达 10 分钟内完全打不通。车停在路边，三个人都在低头玩手机。

“等还是不等？”司机问道。

“听您的。”

最终司机决定放弃这位失踪的拼车客，我们又花了 10 分钟从这条街里拐到原路上。显然，此时司机已经非常烦躁：他没能赚到那份拼车费，又浪费了将近半小时，于是决定倒倒苦水。像世界上每一个男性一样，他先从大事说起，例如优步被滴滴收购。

曾经既没有滴滴，也没有优步的时候，网约车还被叫作黑车，那时笔者很喜欢跟黑车司机聊天，因为司机往往认为笔者似乎能理解他们不愉快的生活，然后抹去车费的零头。奇怪的是，网约车司机从来不会这么做。不管怎么说，这次优步司机提到：据说滴滴马上就会取消给司机的全部奖励，而且最近他接到的单子越来越少，“我都不想干了”。

“因为滴滴形成垄断了。”他说，“我们不管怎么选也只能加入他们家。你们也一样，以后车费肯定涨起来，不像以前那么便宜了。”

近些年来，关于“垄断”的新闻层出不穷，似乎每个行业巨头企业都曾惹上相关的官司。以 IT 领域来说，1969 年，IBM 曾深陷其中；2001 年，微软被指控垄断；不久前，欧盟也起诉 Google 垄断。根据相关资料来看，滴滴换股收购优步中国后，新的滴滴出行将占据中国专车市场份额的 93.1%，这个数字确实听

洞见数据价值.indd 70 2018/1/15 16:14:18

起来很危险。

垄断是个从古至今的概念，垄断企业的产生通常由于两种情况，一种是凭借技术上的优势，形成行业上的垄断，如社交领域的微信、出行领域的滴滴、网购领域的淘宝，都是凭借一定的技术优势成为某一细分领域的垄断巨头；另一种则是由国家控制下的垄断行业，例如烟草、电力、铁路等行业，一直由国营垄断，规定私人不可经营。

先不讨论第二种情况，无论某企业在技术上有多大优势，对市场份额的吞并仍不是一夜之间可以完成的，何况其竞争对手们也在进步。为什么最终企业 A 吞并了企业 B，而不是相反呢？是因为 A 太高明，B 不努力？还是因为 A 顺应了时代潮流，B 没有抓住机遇？答案也许是其中一个，也许都不是。也许马太效应是一个答案。

“马太效应”典出《马太福音》：凡有的，还要加给他，叫他有余；凡没有的，连他所有的也要夺去。而这一理论最早由美国科学史作者罗伯特·莫顿提出，他认为，相对于那些不知名的研究者，声名显赫的科学家通常得到更多的声望，即使两者有着相似的成就。在科学界，马太效应的确可以解释许多不公平的现象：某项重要成就的研究中，常常有些做出了重要贡献的科学家由于国籍、性别等原因，被人忽视甚至遗忘。不过，“马太效应”最为显著的领域还是当属经济学：富者愈富，穷者愈穷。

垄断企业和马太效应就是天生一对，垄断企业肯定存在马太效应，马太效应则必然带来垄断企业。不妨用模拟数据来证实一下。

一个行业体系中，企业的市场份额波动可以用一个特殊的随机过程来模拟，这就是马尔可夫链。马尔可夫链是一种无记忆系统随机演变的数学模型，这意味着系统未来的行为仅依赖于系统当前的

洞见数据价值.indd 71　　2018/1/15 16:14:18

状态，而不考虑其他方面的影响因素。

马尔可夫链有两个特征，一个是稳定性，另一个是马尔可夫性，也称为无记忆性。在一个初步成型的行业体系中，如果没有重大的改革，就可以认为企业市场份额的波动过程满足稳定性的要求。由于现在产品和服务的可选择空间越来越大，客户基本只会根据各个企业最新的服务水平来做出选择，而不会借鉴很久以前的服务水平，因此，企业市场份额的波动过程满足稳定性的要求是完全成立的。

首先令 $W(t)$ 表示在时刻 t 的市场份额状态，P 表示市场份额的转移概率矩阵。可以推导出在时刻 $t+1$ 的市场份额为

$$W(t+1)=W(t)\times P$$

假定一个行业内有 5 个主要的企业，第 t 年的市场份额向量为 $W(t)=(W_1(t),W_2(t),W_3(t),W_4(t),W_5(t))\times P$，其中 $W_1(t),W_2(t),W_3(t),W_4(t),W_5(t)$ 分别为第 t 年中每个企业的市场份额。转移概率矩阵 P 的元素 p_{ij} 表示市场份额从第 i 个企业流向第 j 个企业的概率。

为了找到转移的趋势，需要用二次规划来估计矩阵 P，该矩阵解释了这 5 个企业之间的关系。在设定目标函数时，希望市场份额的实际值越接近于预测值越好，即估计出 p_{ij} 的值可以使得总的误差达到最小。除了目标函数，还必须考虑约束函数：每一个转移概率都必须在 0 到 1 之间，每一行上所有的元素之和必须等于 1。可以得到以下限制条件：

$$\min\sum_{j=1}^{5}\sum_{t=1}^{n}(W_j(t)-\sum_{i=1}^{5}W_i(t-1)p_{ij})$$

$$s.t.\sum_{j=1}^{5}p_{ij}=1,i=1,2,3,4,5$$

洞见数据价值.indd 72 2018/1/15 16:14:18

$$0 \leqslant p_{ij} \leqslant 1$$

表 1-3 展示了该行业中 5 个企业的主营业务收入。

表 1-3　5 个企业的主营业务收入

	企业 A	企业 B	企业 C	企业 D	企业 E	合计
2015Q1	398.9	211.6	534.7	188.5	31	1364.7
2015Q2	413.9	223.9	574.8	199.4	34.1	1446.1
2015Q3	422.1	230.2	611.5	196.6	37.5	1497.9
2015Q4	409.7	222	625.6	193.3	34.3	1484.9
2016Q1	425.2	230.7	638.9	203.6	35.2	1533.6
2016Q2	434.5	222.4	700.4	210.6	35.9	1603.8
2016Q3	436	228.4	751.4	202.8	41.2	1659.8
2016Q4	437.8	235.2	762	205.4	42.8	1683.2
2017Q1	434.3	224.7	744.3	207.5	39	1649.8

显然，可以根据 2015 年第一季度的营业情况将该行业中的企业分为三等：第一等级包含企业 C，它的主营业务收入远远超过其他企业的；第二等级包含了企业 A、B 和 D，其中企业 B 和企业 D 的收入之和都要小于企业 C。第三等级包含企业 E，它的份额太小，几乎很难影响到其他企业。在这个行业中，最大和最小的企业之间有很大的差距。

那么，这 5 个企业的后续发展如何呢？从表 1-3 所示的数据中，可以看出，只有最大的企业 C 保持了迅速的主营业务收入增长。或许用市场份额来看会更加明显，如表 1-4 所示。

表 1-4　5 个企业的市场份额

	企业 A	企业 B	企业 C	企业 D	企业 E
2015Q1	29.23%	15.51%	39.18%	13.81%	2.27%
2015Q2	28.62%	15.48%	39.75%	13.79%	2.36%

洞见数据价值.indd 73　2018/1/15 16:14:18

续表

	企业 A	企业 B	企业 C	企业 D	企业 E
2015Q3	28.18%	15.37%	40.82%	13.13%	2.50%
2015Q4	27.59%	14.95%	42.13%	13.02%	2.31%
2016Q1	27.73%	15.04%	41.66%	13.28%	2.30%
2016Q2	27.09%	13.87%	43.67%	13.13%	2.24%
2016Q3	26.27%	13.76%	45.27%	12.22%	2.48%
2016Q4	26.01%	13.97%	45.27%	12.20%	2.54%
2017Q1	26.32%	13.62%	45.11%	12.58%	2.36%

企业 C 的市场份额在一开始就是最高的，而后来又几乎保持了连续的增长：从 2015 年第一季度到 2017 年第一季度，它的市场份额足足增加了 6 个百分点，从 39.18% 增长到 45.11%。与此同时，第二等级中 3 个企业的市场份额几乎一直在减少，减少的总量大约等于企业 C 的增长量。最小的企业 E 的市场份额非常小，只是在其他企业的竞争中得到一点喘息之机。

图 1-14 代表 5 个企业的营业收入走势，不用说，只有企业 C 一直有增长的趋势。

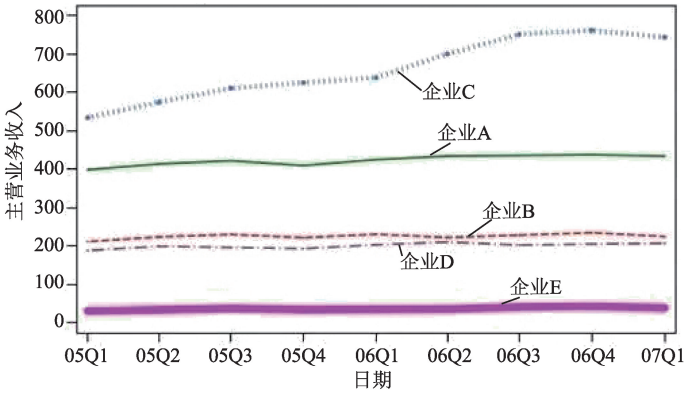

图 1-14 5 个企业的营业收入走势

洞见数据价值.indd 74 2018/1/15 16:14:19

利用市场份额的历史数据，可以在 SAS OR 中估计出转移概率矩阵。结果如下。

$$\boldsymbol{P}=\begin{bmatrix} 0.51 & 0.35 & 0 & 0.14 & 0 \\ 0.38 & 0.26 & 0 & 0.36 & 0 \\ 0.02 & 0 & 0.94 & 0.01 & 0.03 \\ 0.4 & 0 & 0.26 & 0.26 & 0.08 \\ 0.51 & 0.49 & 0 & 0 & 0 \end{bmatrix}$$

观察矩阵 $\boldsymbol{P}$ 的对角线元素可以发现，它们表现了保持市场份额的能力：这个能力值越接近于 1，说明保持市场份额的能力越强，反之说明保持的能力越弱。毫无悬念的是，企业 C 的保持能力最强，企业 A 的能力次之，企业 B 和 D 的能力并列第三，而企业 E 几乎没有保持市场份额的能力，这种顺序几乎和市场份额的大小完全一致。或许，在不久的将来，企业 C 即将成为该行业中的滴滴。不用说，该行业中存在着显著的马太效应。

需要注意的是，利用马尔可夫链进行的模拟过程只考虑当前状态，即各企业当前的市场份额，而其他可能的影响因素都不在考虑范围之内，这就是说，不会假定企业 C 比其他企业经营得更高明、技术发展更快等，企业 C 在系统中唯一的优势，就是超出其他企业的市场份额。因此，在一个行业体系中，如果没有重大的政策导向或技术突破，大企业会由于先发优势一直保持这样的增长势头，直到成为垄断寡头。在马太效应的作用下，垄断是一种趋势，甚至是一种必然。

除了垄断企业本身，恐怕没有人会觉得垄断是一种好事，司机和我们都会因为滴滴的垄断遭到若干损失：他少一些收入，笔者要

洞见数据价值.indd 75　2018/1/15 16:14:19

多付一些车费；他会跟接下来的每一位乘客抱怨这种悲观的可能，而笔者会把这件事写在文章里。甚至，司机与笔者之间也存在着一种不可避免的马太效应，他当前拥有的资产比笔者多，因此他可以买辆车，然后从笔者的资产中赚取车费；当然，指出这一点应该并不会让他感到多少安慰。

人们看到一种现象、一个结果时，常常会刻意从中寻找一些原因。人们可能会理所当然地认为，行业中的垄断者具有其他企业所没有的品质，例如经营得当、决策超前、老板特别有魅力……然而，如果以数据分析的视角看待问题，就可能会发现，所谓原因也许根本不存在或者并没有发挥那么大的作用，人们不过是被一种趋势（或称“历史的车轮”）携裹着前进。这样的视角可能会让你悲观，但也可能会让你更坦然、更有准备。

因此，很多时候，数据工作者所做的事情，就是一直在观察数据、调和模型，静静地等待着“趋势”的出现，并且提醒你：不要被抛在后面。

1.7 大数据在舆情行业

在只有报纸、广播、电视广告的年代，企业与客户之间的交互与了解都是非常有限的，即使企业或产品有若干负面消息，客户也很难即时获知，这就给了企业相关部门充分的反应和应对时间。

不过，到了互联网时代，信息传播速度堪比光速，有时候这一秒刚刚发现问题，下一秒就上微博热门了……于是，舆情管理的重要性呈几何倍数激增，数据分析也开始帮助企业掌握海量的舆论信息。

洞见数据价值.indd 76 2018/1/15 16:14:19

创新方向：数据分析帮你掌握话语权

如果你不相信舆情的重要性，就去看看 2016 年年中闹得沸沸扬扬的离婚案主角——某男星前妻的微博评论吧。

也许你认为，寻常网友的评论于当事人无关痛痒，但假设将来当事人想要拍电影，却一无投资人二无观众，恐怕也是顺理成章的结果。

舆情，是“舆论情况”的简称，指在一定社会空间内，围绕社会事件的发生、发展和变化，民众对社会管理者、企业、个人及其他各类组织及其政治、社会、道德等方面的取向产生和持有的社会态度。它是较多群众关于社会中各种现象、问题所表达的信念、态度、意见和情绪等表现的总和。

信息化时代，人们越来越深地感受到舆情，尤其是网络舆情的影响力，尽管可能未必意识到。外出聚餐之前，你会打开大众点评，参考过往食客的评价，而 APP 上一条条打分和评论正是一种舆情；买车、家电之前，你也会浏览相关网站，或者在微博上搜索一番，看看某品牌或款式的评价如何，负面反馈多不多，这些评价和反馈也正是一种舆情。有时，一个低分、一条关于售后服务的消极评价，不但可能彻底打消你个人对品牌或企业的全部好感，甚至可能经过发酵、大量传播，成为众人关注的一时热点，以至于企业不得不花费巨大成本，采用公关手段力求挽回。

当然，公众形象始终是一件与其亡羊补牢、不如防患于未然的事情，因此舆情产业已经深入政府、媒体、教育科研、软件等行业。大量舆情软件公司和市场调查公司高速发展，以抓取网络舆情数据为基础技术，成为舆情服务业重要的技术型方阵。截至 2012 年 1 月 16 日，全国共有约 68 款经过工信部软件司认定登记颁证的“舆情”软件，市场上还存在大量未经认证的同类软件。

洞见数据价值.indd 77　2018/1/15 16:14:19

目前，舆情分析机构主要提供的服务覆盖了从舆情监控分析到研判、预警，再到提出应对策略。舆情管理已经从简单粗暴的媒体宣传，发展成更为精细化、全面化和专业化的运作。网络舆情分析包含舆情信息采集、舆情事件预警、舆情智能分析与应对策略以及舆情报告等多方面综合分析。

舆情信息采集是舆情分析的基础工作，需要对网络上的舆情信息进行分类和扩展，从深度和广度两个方向采集，满足后续分析要求。**舆情事件预警**则需要对热点事件进行深度剖析，持续跟踪热点事件，分析事件发展趋势，对事件进行监控并及时告警。**舆情智能分析与应对策略**可以帮助企业从企业形象、产品口碑和舆情事件进行深入分析，对影响舆情的因素进行多角度判断，倾向性分析与统计，提供应对策略方向以及策略收益分析。

1. 舆情分析可以实现什么?

当前的舆情分析市场中，现有软件厂商和传媒科研机构各有所长，他们在政府部门的舆情分析中发挥着重要的角色。然而，这些厂商或机构往往注重信息采集，但收集了海量信息后，却在利用大数据技术进行信息分析的思维方面有所欠缺。

作为由数据专家、建模专家和业务专家组成的团队，我们认为，对海量信息的处理和分析才是舆情管理中的重点，而在掌握信息的基础上，可以完成的工作包括舆情事件预警和口碑管理。

2. 舆情事件预警

舆情事件预警主要包含四个重要功能。

其一，对整个网络舆情进行监控，及时甄别最新的舆情事件，

洞见数据价值.indd 78 2018/1/15 16:14:19

有助于事件及早发现，及早处置。

其二，对热点事件进行实时跟踪，对事件热度进行综合评分，设置预警线，实时告警。

其三，探查事件热度拐点，分析拐点背后原因，寻找事件意见领袖，研究意见领袖对舆情的影响。

其四，分析事件热词，识别热词倾向性。

3. 口碑管理

口碑管理的关键在于，通过网络舆情数据，分析展示市场对于该品牌的形象描述，对比企业自身的期望，分析差异原因。在掌握数据的基础上，应对企业现有的宣传策略进行诊断，对比各大宣传平台 PV 访问量、UV 访问量等信息，针对各个宣传平台中企业形象信息，分析宣传收益。以期指导企业根据分析结果，适当调整宣传策略，以达到最好的宣传效果。从产品、时间、地区等维度分析口碑变动情况及热词分布，对热词进行情感判断，识别口碑及热词倾向性。企业可针对分析结果，采取针对性的销售策略，也可以针对反馈集中的问题对产品进行优化。

那么，在了解概念的基础上，要如何实现有效的舆情管理呢?

4. 技术解决方案：如何实现舆情管理?

无论采用什么方法或技术，最高形态都是完整的系统，因此搭建舆情分析系统是最有效率的舆情管理手段。

舆情分析系统分为三个主要模块：舆情信息采集、舆情分析引擎和舆情应用。舆情信息通过舆情采集模块流入舆情分析引擎，分析引擎负责数据加工处理、分析舆情信息，最终分析结果通过舆情

洞见数据价值.indd 79　2018/1/15 16:14:19

应用展示，分析过程中的数据分别存在舆情数据库和查询索引数据库，如图 1-15 所示。

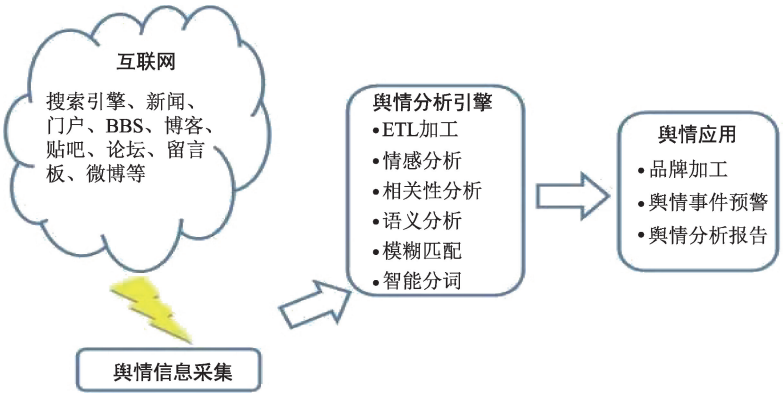

图 1-15 舆情分析系统示意

5. 舆情信息采集：爬虫技术

舆情信息采集的基础在于，使用当前十分成熟的网络爬虫技术，对指定网站的数据进行爬取。目前，网页信息的抓取策略主要分为广度优先、最佳优先和深度优先三种。由于深度优先经常会受到限制，在团队实践中，主要采用广度优先和最佳优先，部分数据采取深度优先策略。市面上的开源网络爬虫工具多达 100 多种，它们各有所长，而团队以 Python 为主，以其他开源软件为辅，从网络上抓取所需信息。另外，需要注意的是，现在大多数网站都采取了反爬虫技术，在搜集数据时，还需要使用多种技术手段应付反爬虫技术。

6. 舆情分析引擎

应用爬虫技术抓取的数据，通过一系列 ETL 过程，存储在本地

洞见数据价值.indd 80 2018/1/15 16:14:19

系统中，使用 SAS、R 等工具对数据进行分析挖掘。对关键词进行情感分析和综合评分，标注关键词的事件、时间、地区、关联用户、关联词等信息，计算该词的频率，在同一事件中的占比等数据。

7. 舆情应用

下面以汽车行业为例简单介绍舆情系统的应用界面。

1）口碑管理

图 1-16 是通过舆情分析得到的品牌形象，可见该品牌留给消费者的印象主要是“安全”和“适合成功者”。假如企业期望的品牌核心形象为安全和时尚，可见两个形象存在一定偏差，企业需在宣传策略上做出一定调整。

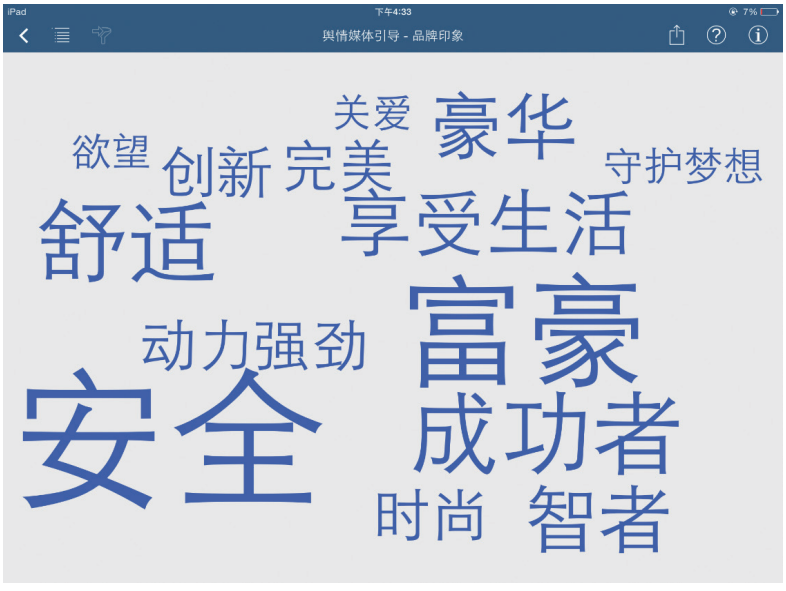

图 1-16　通过舆情分析得到的品牌形象

洞见数据价值.indd 81　2018/1/15 16:14:19

“声量诊断”也具有重要的参考价值。可以分析其宣传品牌文化的渠道和平台，比较各个平台的 PV 访问量（页面浏览量）和 UV 访问量（通过互联网访问、浏览这个网页的自然人数量），以及赞同数（见图 1-17），然后再比较各个平台间品牌形象的差异。

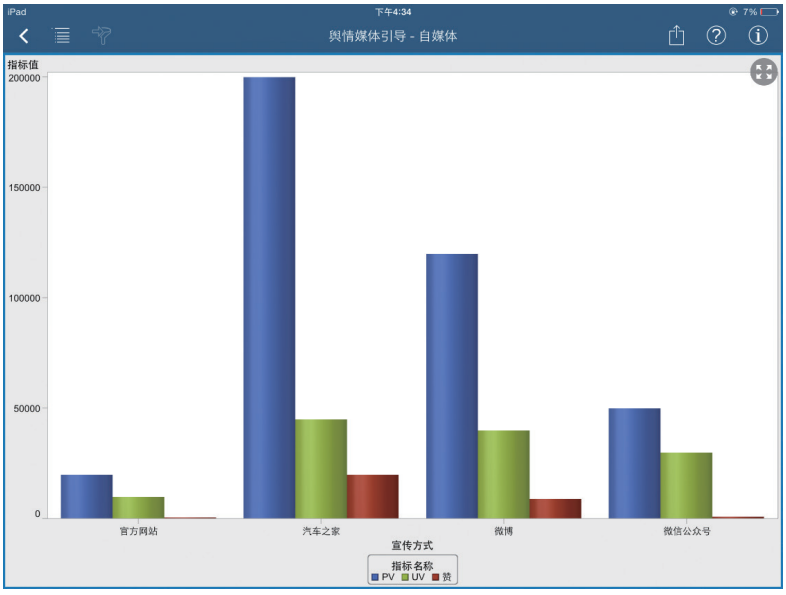

图 1-17 比较各个平台间品牌形象的差异

“口碑变化”可以帮助分析各类产品的口碑变化趋势（见图 1-18），其中深色代表较上期下降，浅色代表较上期增长，可以单击任意立柱链接到后面的一幅图，如图 1-19 所示。图 1-19 展示当期影响口碑的热词，热词大小代表权重，如频度、浏览人数和支持人数等；颜色代表倾向性，红色代表负面倾向，蓝色代表正面倾向。

洞见数据价值.indd 82 2018/1/15 16:14:19

图 1-18　口碑变化趋势

图 1-19　当期影响口碑的热词

洞见数据价值.indd 83　2018/1/15 16:14:19

另外，系统还支持地区间口碑对比，如图 1-20 所示。

图 1-20　地区间口碑对比

2）舆情事件预警

舆情监控系统采用准实时跑批和手动跑批两种方式，实时把握网络热点事件动向，如图 1-21 所示。针对 2012 款速腾进行全网搜索，通过模型算法计算当前热度值，并展现热词情况，仪表盘指向红色部分，热度值已超预警值，系统告警。热度值计算方法采用声量加权重的方式，微博的声量为 V，权重系数为 v，微信的声量为 W，权重系数为 w，新闻网页的声量为 D，权重为 d，搜索引擎的声量为 S，权重为 s，热度值为 $H=v\times V+w\times W+d\times D+s\times S$，权重系数由客户和咨询方商议协定。市场主流的权重系数为，v=0.15，w=0.4，s=0.25，d=0.2。其中，热词的大小代表热词频度，根据语义分析判断倾向性并以颜色标出，蓝色代表正面倾向，红色代表负面倾向，灰色代表无倾向。

洞见数据价值.indd 84　2018/1/15 16:14:19

图 1-21 中折线代表该产品热度随时间变化趋势。

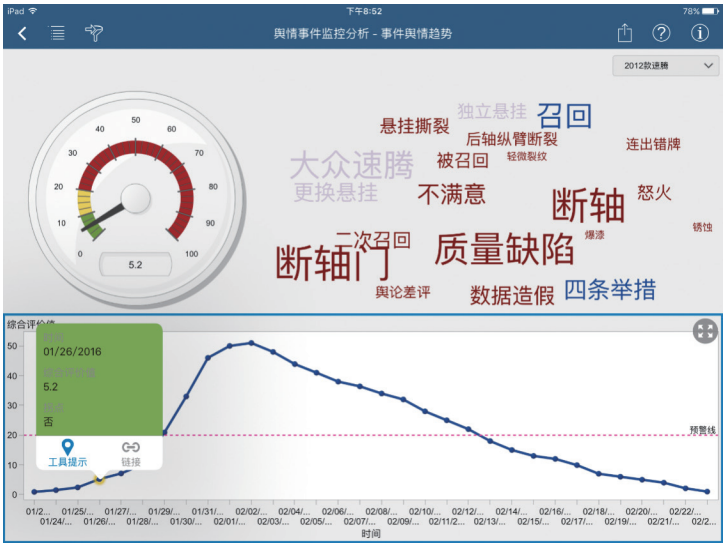

图 1-21　舆情事件预警

如图 1-22 所示，对舆情事件趋势中任意时间点进行钻取，分别对其倾向性、职业、收入、网站、平台、性别、年龄段、学历和城市类型等不同维度，展示各个人群对此事件的情感倾向性。就这样，舆情管理系统的主要功能已经搭建成功。

舆情管理是一项新兴产业，但舆情绝非一种新鲜事物，每个人一直都有很多话要说，很多观点要发布。也许就在当下，打开微博、微信或任何较为流行的媒体平台和论坛时，明星、品牌、企业、热点……仍在一刻不停地经受舆情的考验，而成功或毁灭往往就蕴含在每一段文字中间，可能无法预测，可能难以控制。

然而，归根结底，舆情仍是一种信息、一种数据。

洞见数据价值.indd 85　2018/1/15 16:14:19

在网络上，无数未能掌握舆情管理的品牌或企业眼看着自己一步一步被无序的舆论之海淹没，最终窒息。然而，只要成功运用了舆情管理的思维和方法，舆情的数据流就是有规律的，是可以预测，甚至调控的：每一条舆情言论，都可能是企业塑造形象的契机。

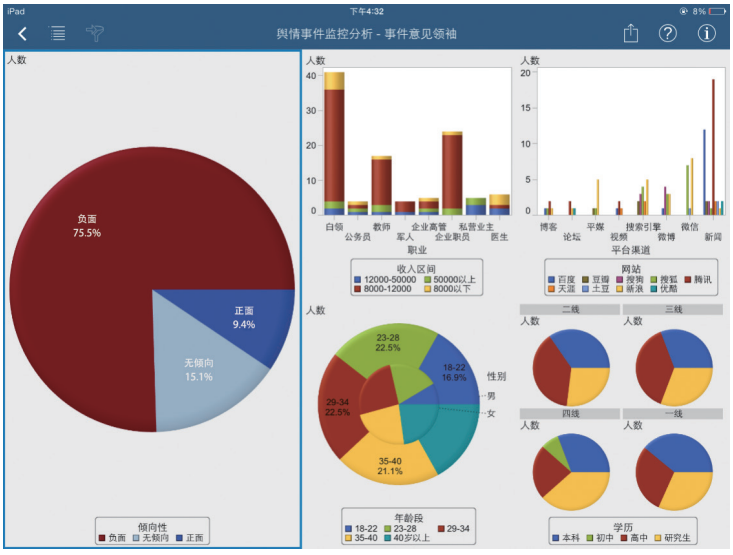

图 1-22　通过不同维度展示人群对于事件的情感倾向性

1.8　大数据在汽车行业

业界展望：征服汽车后市场，大数据与你同行

随着私家汽车销售在各线城市井喷式增长，我国汽车行业无疑正享受着黄金时代。汽车保有量的急速提升，也要求市场提供相匹配的高品质售后服务，让汽车后市场变成了充满机会与诱惑的领地。

洞见数据价值.indd 86　2018/1/15 16:14:20

所谓汽车后市场，是指汽车销售以后、为消费者提供买车后所需要的一切服务的市场。你驾驶的每一辆汽车，从售出到报废的生命周期中，都离不开配件、维修、装饰、金融保险、IT 甚至文化方面的服务，而这些服务自然引发了大量交易活动，也带来了无数盈利的机会。

影响汽车后市场需求的两大因素是：汽车保有量和汽车产业链利润结构。我国的汽车保有量已经非常可观，到 2018 年，汽车产量预计将达到甚至超过 2 亿辆。不过，与之相反的是，汽车产业链却远远不够成熟，在成熟国家的汽车产业链中，汽车后市场的利润可以占到 50% ~ 60%，而我国目前却只占到 10% 左右，这恰恰提供了巨大的提升空间。

市场空间显而易见，随之而来的就是各汽车厂商面临与日俱增的挑战和压力：成本压力、行业竞争、客户管理以及市场变化和波动。汽车新增量每年以 20% 的幅度高速增长，汽车后市场的提升空间比例更高，利润也更加可观，但各家厂商究竟能从其中分得多少，完全取决于你对于客户洞察是否清晰、产品结构是否合理、服务流程是否最优化、市场竞争分析是否到位等方面。最新的大数据分析为汽车厂商带来了前所未有的可能，以客户为中心，从新车销售到品牌服务，让各厂商全方位了解市场、挖掘市场潜力、开拓客户基盘后市场空间。

汽车行业想要将数据分析融入业务运营，成为战略资产，需要从客户洞察、运营洞察、产品洞察、市场分析洞察四个维度出发，通过大数据分析建模技术，汇总多种来源的海量数据，进行深入分析并解决问题，从而最终服务于汽车销售管理层、4S 店的售后经理、维保服务人员等各层级。

洞见数据价值.indd 87　2018/1/15 16:14:20

1. 客户洞察

- **客户基盘分析：**多维度分析经销商集团客户基盘总数及其分布情况，以及纵向变化。
- **客户活跃度及流失：**多维度分析经销商集团的客户流失情况及流失类型，并通过与其他集团的比较，发现差异及产生差异的主要原因。
- **客户赢回：**二手车客户行为分析，包括二手车置换分析、重构分析和跨集团维修历史信息推送。
- **客户获取：**分析客户获取及维系成本。

2. 运营洞察

- **销售流程分析：**多维度分析销售各个环节的执行情况，以及时间间隔，并通过与其他集团的比较，发现差异及产生差异的主要原因。
- **售后流程分析：**多维度分析首保服务和维修服务各个环节的执行情况，以及时间间隔，并通过与其他集团的比较，发现差异及产生差异的主要原因。
- **资源效率分析：**人力资源分析，包括人员资历资质、人员生产力、人员流失；整车资源分析，包括库存深度、库存结构和运转效率；零部件资源分析，包括零部件的库存周期和零件满足情况；金融资源分析，包括融资成本和融资结构。

3. 产品洞察

- **销售阶段：**多维度分析配套产品的购买情况，并通过与其他集团的比较，发现差异及产生差异的主要原因。

洞见数据价值.indd 88 2018/1/15 16:14:20

- **售后阶段：**维修服务项目优化，针对不同需求客户提供个性化产品。
- **客户关怀产品：**多维度分析客户关怀产品（活动）的参与情况和客户体验。

4. 市场分析洞察

- **市场环境分析：**借助网络爬虫和文本分析技术，从社会舆论的角度出发，对汽车市场的整体发展情况，以及公司整体或经销商集团的口碑进行分析。
- **同业竞争分析：**借助网络爬虫和文本分析技术，从社会舆论的角度出发，对竞争对手的发展情况进行分析。

为了充分实现汽车行业的数据化，首先，不妨通过管理仪表盘、高级可视化技术等，对业务发展现状进行数据统计和汇总，反映汽车厂商的历史发展趋势和正在变化及存在的问题。然后，发掘潜在的业务动因和模式，辅助业务决策运营，进而驱动业务变革与创新，通过数据分析来改进客户体验，带来更有效的市场推广和客户互动，帮助汽车厂商提升品牌形象。

1.9　大数据在影视业

如火如荼的影视业不但为从业者带来巨大的利润，也为数据工作者提供了思考的新方向。有时，数据得出的结论看似与人的经验判断相反——但那可能是因为，在做出判断之前，并没有进行深入思考。

洞见数据价值.indd 89　2018/1/15 16:14:20

创新方向：星期几上映的电影最具有票房号召力?

如果你问，现在最佩服国内什么人，笔者一定会毫不犹豫地告诉你：做电影的。

不知从何时起，进电影院似乎成了一种刚需，质量高、口碑好的电影自然不愁票房，连那些烂片往往也都能回本，使得影视界成了投资热门。一个行业一方面遭到大家纷纷吐槽，另一方面又能从大家口袋里把钱掏出来，这种行业的境界之高让人不服不行——市场的数据趋势如洪流携裹，影片本身的质量反而不是决定性因素。

那么，电影市场到底有多火呢?

姑且把这股数据流截至 2016 年 3 月 31 日，此时，2016 年 1 月 29 日上映的《功夫熊猫 3》票房已经达到 10 亿元，2016 年 2 月 8 日上映的《美人鱼》票房已经达到 33.9 亿元，2016 年 3 月 4 日上映的《疯狂动物城》已经达到 13.4 亿元……图 1-23 和图 1-24 是 2011—2015 年全国票房的趋势和影片平均票房。从中可以看出，尽管影片数量的变化不明显，但总票房增长速度很快，平均票房自然也急速飞升，大家对电影的消费需求在显著增加。

面对市场的高速扩张，作为数据工作者必须承认，这样的市场一定已经形成了自己的数据走势，也肯定想要从中探寻若干规律。本节就来解决一个看似显而易见的问题：

一部电影要想获得有利票房，应该选择在一星期中的哪天上映?

为严谨起见，请先接受假定：电影票房的**走势**和星期几上映还是有关系的。为了更好地判断变化趋势，这里引入一个概念：**票房相对值** = 当天票房值 / 上映第一天票房值，该值通常会先在 1 左右浮动，再随着时间而下降。

洞见数据价值.indd 90 2018/1/15 16:14:20

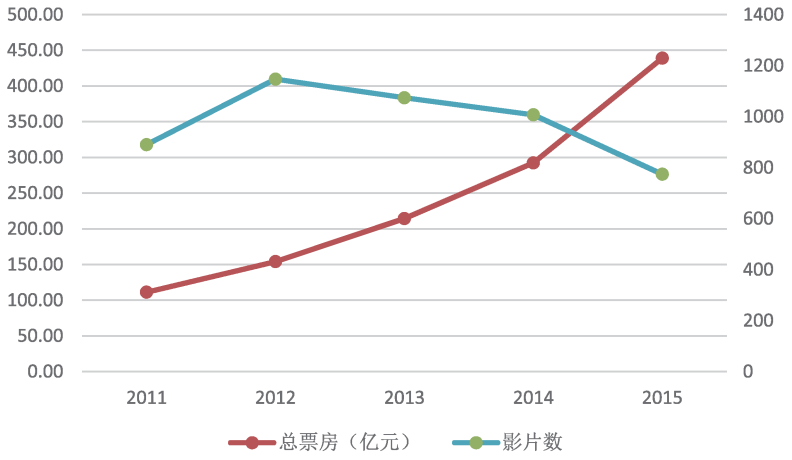

图 1-23　2011—2015 年全国票房趋势

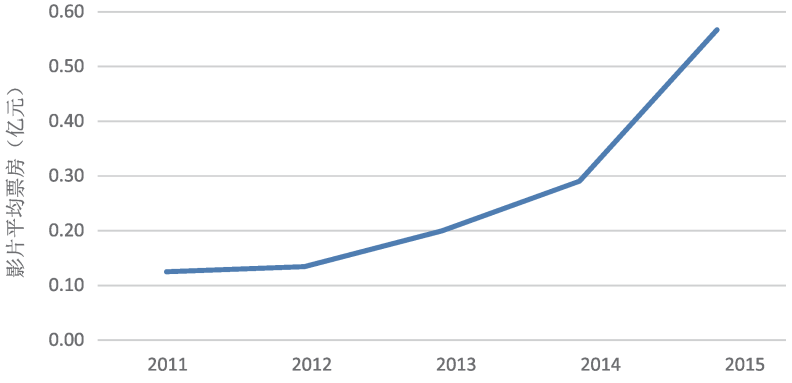

图 1-24　2011—2015 年影片平均票房

数据来源：实时电影票房网（www.piaofang168.com）

一部总票房 10 亿元的电影和一部总票房 5000 万元的电影，每天的票房数额差距可能都很大，但考虑到影片成本、定位等，不能简单说前者成功，后者失败。同样，这两部电影的票房走势曲线可

洞见数据价值.indd 91　2018/1/15 16:14:20

能是相似的。为了寻找抛开电影的总票房，单纯探讨其票房趋势的变化，这里引入了上述的“票房相对值”。这里选取 2011—2015 年上映的 3281 部影片进行分析，根据上映日期的不同绘制如图 1-25 所示的相对值曲线。

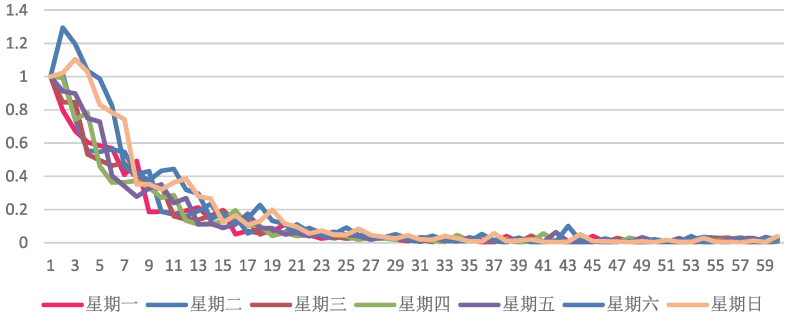

图 1-25　根据上映日期绘制的 2011—2015 年上映的 3281 部影片的相对值曲线

以星期五上映的若干电影为例，它们的票房发展趋势如图 1-26 所示。

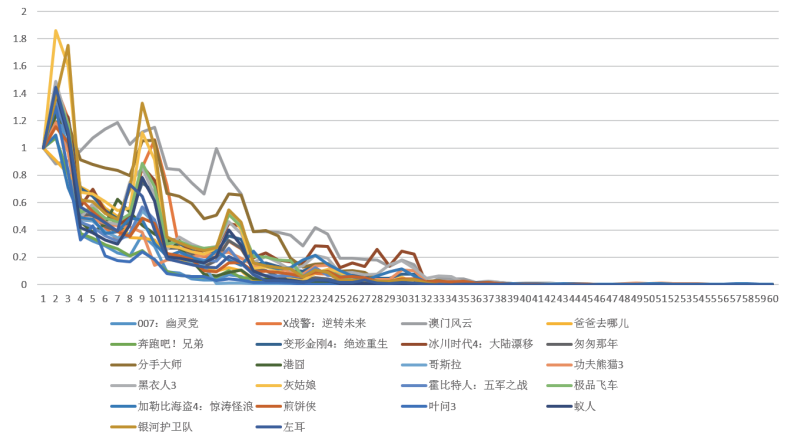

图 1-26　星期五上映的若干电影的票房发展趋势

洞见数据价值.indd 92　2018/1/15 16:14:20

很容易看出，一部电影最重要的就是首周票房，此后的票房发展趋势几乎等于一个缩小版的首周票房重演，而首周票房的模式很大程度上受到上映日期的影响。

现在研究一周 7 天上映的电影总票房规律，用 Y 代表总票房，Y_i 代表影片上映第 i 日的票房，X_i 代表上映第 i 天的票房相对值，则总票房为：

$$Y=\sum_{i=0}^{+\infty}Y_i=Y_1\times\sum_{i=0}^{+\infty}x_i$$

以在星期六上映的电影为例，星期六上映电影的票房相对值趋势如图 1-27 所示。

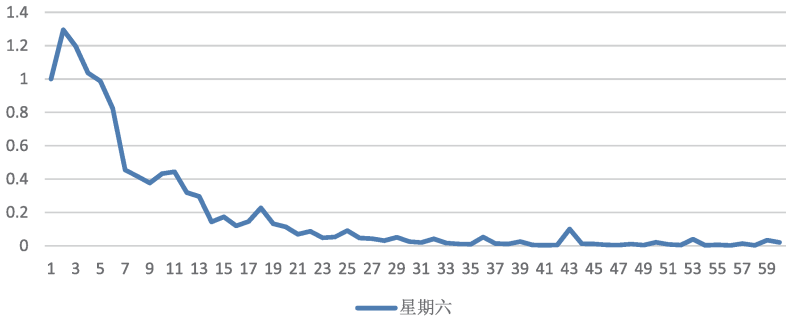

图 1-27　星期六上映电影的票房相对值趋势

从图 1-27 中可以看出，票房相对值随着上映天数呈递减趋势并逐渐趋近于 0。取 i 为 300 时，可计算出星期六上映电影的总票房为 $13.60Y_1$。同理可计算出一周其他 6 天上映的电影总票房。最后，可得出一周 7 天上映的电影总票房规律如表 1-5 所示。

请注意，以上结果是一个相对值，这里所探讨的并不是电影上映日期如何对总票房造成关键性影响，事实上，最关键的影响因子

洞见数据价值.indd 93　2018/1/15 16:14:20

当然仍要属排片、宣传、阵容等。我们想知道的是，**一部万事俱备、只欠上映的电影，应该选择在星期几首映才能获得最好的票房走势**。

表 1-5 一周 7 天上映的电影总票房规律

上映时间	总票房	排名
星期五	$14.56Y_0$	1
星期四	$13.96Y_0$	2
星期六	$13.60Y_0$	3
星期日	$13.60Y_0$	4
星期二	$13.35Y_0$	5
星期三	$12.70Y_0$	6
星期一	$12.07Y_0$	7

答案是：星期五，星期四位居其次。

考虑到首映时间往往是午夜、凌晨，人们理解中的“星期五首映”往往其实是星期五 0 点，我们认为最合适的时间就是星期四。

这样的结果是不是让你有点吃惊呢?

当然，由于票房方面的数据无论如何也说不上像金融业那么完善，以上得出的结果只是尽己所能的分析，未必完美。如果有更完整的数据，可能结果又会不同。

其实，当初步完成这一数据分析时，顺带让团队里的同事们猜了猜，结果大家普遍相信电影在星期六首映对票房最有利，因为周末看电影的人数最多。结果一出，同事们都不太敢相信。然而，随后查询了若干善于营销的票房大片的首映时间如《小时代》等，发现它们的确都是在星期四、星期五首映的（当然，要排除《美人鱼》这种抓紧大年初一首映的片子）。

看来，数据的结论又一次颠覆了人们自以为是的直觉和经验。若再仔细想想：想要星期六约会时看电影，难道真的会到电影院后

洞见数据价值.indd 94 2018/1/15 16:14:21

再做出选择吗？不，多数人会在星期五看看新上映电影的口碑，提前决定，甚至提前买票。因此，片方希望在星期四到星期五之间首映电影，先赢得一两天的口碑预热期，以便最大限度地争取首个周末的票房爆发。

有时，数据得出的结论看似与人们的经验判断相反，那可能是因为人们在做出判断之前，并没有进行深入思考。

票房预测尝试

一部电影在星期几上映的特性，可以用来将全部影片分为 7 种类型，每种类型的票房走势有所差异，这是在前面论证过的。据此，如果给每种类型单独建模，就可以更准确地预测其票房。

票房时间序列建模流程如图 1-28 所示。

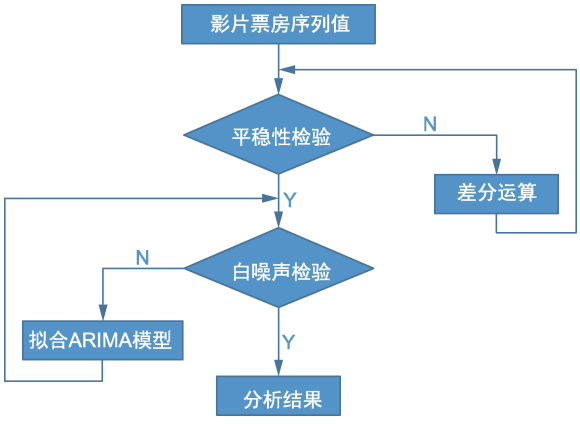

图 1-28　票房时间序列建模流程

这种模型的意义在于，通过差异化的建模，从电影票房初期的表现判断其后走势。模型拟合结果如表 1-6 所示。

洞见数据价值.indd 95　2018/1/15 16:14:21

表 1-6　模型拟合结果

上映星期	差分		自回归系数	移动平均系数	AIC	SBC	自回归因子	移动平均因子	常数项
	阶数	步数							
星期一	1	0	(1,3)	1	−168.906	−160.596	1 + 1.10943 B**(1) − 0.30693 B**(3)	1 + 0.91605 B**(1)	−0.01682
星期二	0	0	1	0	−159.937	−155.749	1 − 1 B**(1)	0	0.92349
星期三	1	0	2	(2)	−175.918	−173.84	0	1 + 0.33858 B**(2)	0
星期四	0	0	1	(2)	−164.887	−158.604	1 − 1 B**(1)	1 + 0.41643 B**(2)	1.019
星期五	1	0	(1,3)	2	−162.247	−158.058	1 + 0.91889 B**(1) − 0.3922 B**(3)	1 + 1.17831 B**(1) + 0.8615 B**(2)	0
星期六	0	0	2	0	−127.438	−121.155	1 − 1.23403 B**(1) + 0.23403 B**(2)	0	0.9116
星期日	1	0	1	1	−146.599	−142.444	1 − 0.92071 B**(1)	1 − 0.83848 B**(1)	0

洞见数据价值.indd 96 2018/1/15 16:14:21

对于非专业人士来说，这些数字比较难以理解，但变成曲线图就是图 1-29 所示的样子——以星期五为例，蓝色星线为票房相对值的实际值，中间的红色线为票房相对值的拟合值，上下的绿色线为票房相对值拟合值的 95% 置信限值。

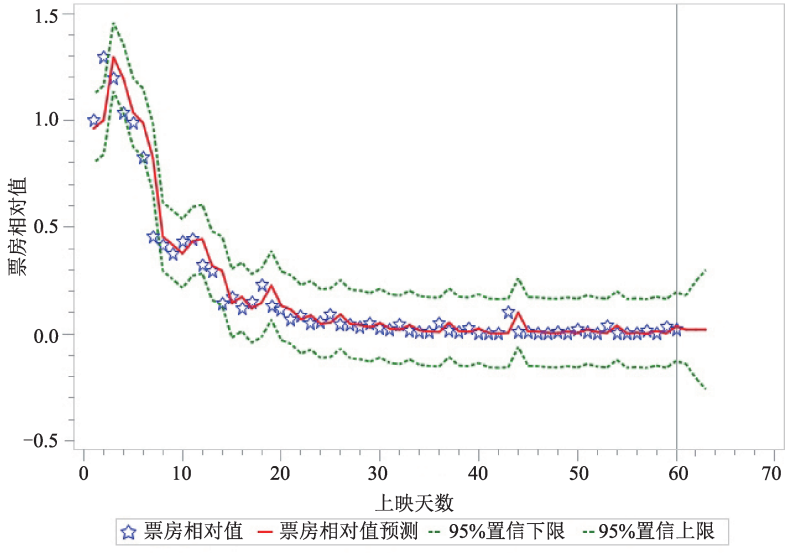

图 1-29　模型拟合结果转换成的曲线

接下来以《伦敦陷落》这部电影来验证一下模型的效果，如图 1-30 所示。

可见，电影实际票房落在预测值上下置信限之内。可以看出，本次的票房预测模型是成功的，而根据上映日期分类建模的思想正是成功的基础。看来，下次安排首映日期之前，可要好好考虑一下星期几。

洞见数据价值.indd 97　2018/1/15 16:14:21

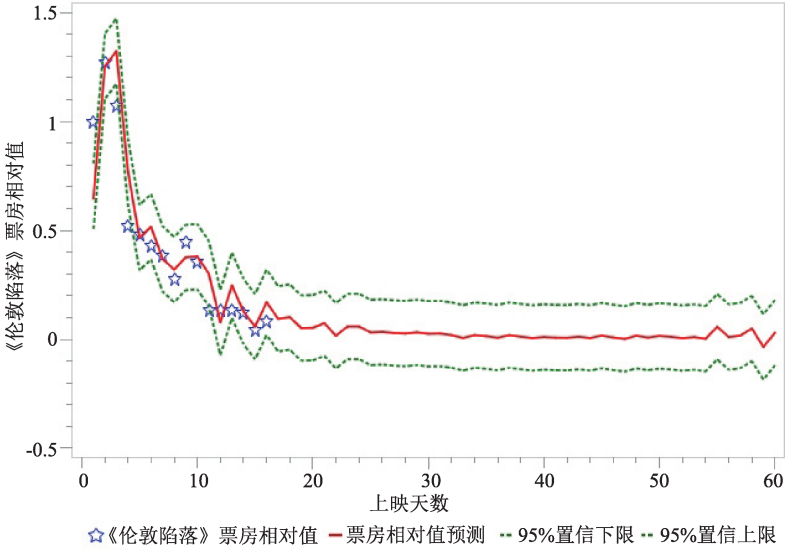

图 1-30 《伦敦陷落》票房分析曲线

1.10 大数据在环保产业

北京雾霾一直是人们关注的热点话题，环境保护也是关系民生的当务之急。数据分析又可以为环保产业做点什么呢？

创新方向：北京治霾，能为你做点什么？

北京，3 月 18 日，初春，风力二级，清晨气温大约 11℃，天气持续回暖，笔者却忽然听见宝贝连着咳嗽了几声。

笔者急忙过去摸他的额头，问他有没有哪里不舒服。他抬起头张大眼睛看着笔者，还没来得及说话，又打了个喷嚏。

换季时孩子容易生病，为稳妥起见，笔者决定跟老师请个假。

洞见数据价值.indd 98 2018/1/15 16:14:21

没想到，一大早幼儿园班级群里已经有了许多留言，好多家长都在替孩子请假，小朋友们普遍鼻塞、流鼻涕、咳嗽，甚至有几个还发起了低烧。笔者粗略算了一下，这一波中招的孩子有 11 个，比例已经超过了 60%。

看着窗外灰蒙蒙一片，笔者心里基本有数了，宝贝估计又是因为雾霾引发了过敏性鼻炎，这样的情景在 2015 年冬天也出现了几次，有一次甚至引起了很严重的咳嗽。笔者给宝贝喂了点控制鼻炎的药，把空气净化器开到最大挡位，就迎着雾霾出门去上班了。

从公司窗口向外望去，空气一片昏黄，看不见太阳。同事说：“现在只要闻闻味道，就能猜出 PM2.5 是多少了。”北京的天气，对于成年人的生活来说，似乎只是一个微不足道的偶然因素：要么忽视它，要么远离它，没有折中的办法。但笔者想起儿子班上 11 个生病的小朋友，他们要脆弱得多。

3 月 18 日的北京，PM2.5 的浓度最高达到 434 微克每立方米。

如果笔者身在制造业，一定要努力减少污染排放；如果笔者是气象工作者，会深入研究气候与污染之间的关系；甚至，如果笔者习惯了开车，可能也想转头去坐地铁……但，作为一名数据工作者，又该怎样做才能保护自己的孩子呢？

其实雾霾相关的数据一直都存在。2008 年 4 月，位于北京建国门外的美国驻华大使馆开始测量 PM2.5 每小时浓度数据，并在之后不久对外公布。2013 年 9 月 10 日，国务院印发《大气污染防治行动计划》，其中有一个具体目标：2017 年，全国地级及以上城市可吸入颗粒物浓度（PM10）比 2012 年下降 10% 以上；北京市 PM2.5 年均浓度控制在 60 微克每立方米左右。

打开美国驻华使馆 2013—2015 年的数据，想从统计分析的角度

洞见数据价值.indd 99　2018/1/15 16:14:21

研究一下北京市的雾霾。

关于 PM2.5 浓度的分类标准，我国与美国的标准有所差异，对比如表 1-7 所示。

表 1-7　我国与美国关于 PM2.5 浓度的分类标准

空气质量指数类别	空气质量分指数（IAQI）	24 小时平均 PM2.5 浓度范围（中国）	24 小时平均 PM2.5 浓度范围（美国）
优	0~50	0~35	0.0~12.0
良	51~100	35~75	12.1~35.4
轻度污染	101~150	75~115	35.5~55.4
中度污染	151~200	115~150	55.5~150.4
重度污染	201~300	150~250	150.5~250.4
严重污染	301~400	250~350	250.5~350.4
	401~500	350~500	350.5~500.0

令人感到欣慰的是，从图 1-31 所示的北京市 PM2.5 年均浓度和每年 PM2.5 日均浓度大于 60 微克每立方米的天数占比这两方面

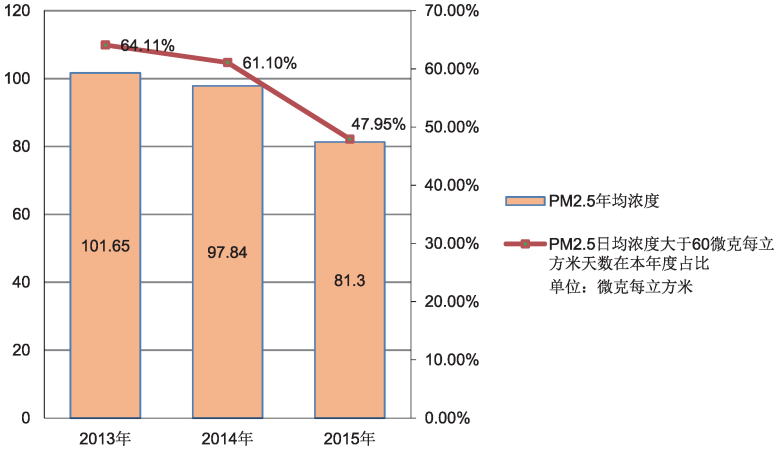

图 1-31　2013—2015 年北京市 PM2.5 平均浓度变化

数据来源：http://beijing.usembassy-china.org.cn/070109air.html

洞见数据价值.indd 100　2018/1/15 16:14:22

来看，这三年期间呈现逐年下降的趋势。但是 2015 年的 PM2.5 年均浓度为 81.3 微克每立方米，仍然高于 60 微克每立方米，还有很长的路要走。

若采用美国的标准对这三年的数据进行分析（见图 1-32），可以看出，令人伤心的是这三年中，北京市空气质量为优的比例最低，中度污染的比例最高；稍许安慰的是与 2013 年和 2014 年相比，2015 年的中度、严重、重度污染天数减少了，而优、良、轻度污染天数增加了（从这个角度来看，2015 年比往年要好一些）。

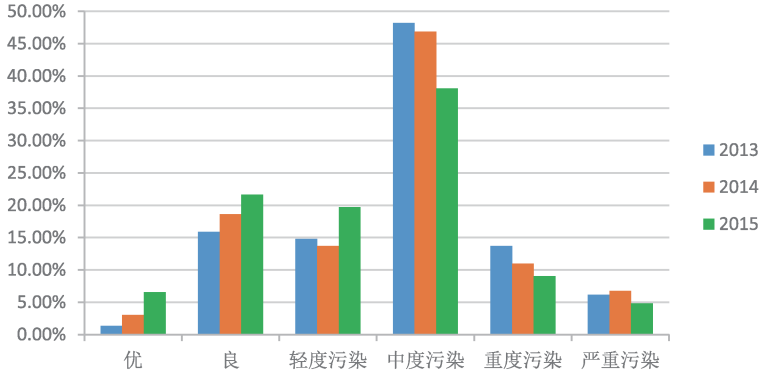

图 1-32　采用美国的标准对 2013—2015 年北京市的空气质量状况进行汇总

结合国内标准，把美国标准的 6 个档次合并为 3 个档次，其中优、良、轻度污染统称为良，重度和严重污染统称为重污染，在这个基础上，分别来看 2013 年、2014 年和 2015 年的四季情况，如图 1-33 所示。

从图 1-33 中可以清晰地看出，在春季和夏季，重污染较少，但是中度污染天气比例较高，达到 50% 左右（2015 年夏季情况稍好），可能是由于夏季温度和湿度较高，北风较少。冬季的良好天气和重

洞见数据价值.indd 101　2018/1/15 16:14:22

污染天气占比基本持平，约 30%，好天气可能是由于冬季北风较多，有利于空气扩散；重污染天气可能是由于冬季北京取暖造成的。

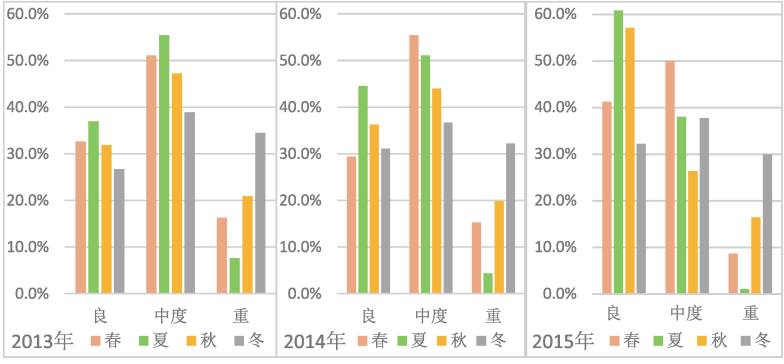

图 1-33 2013—2015 年每年四季的空气质量情况

当然，描述只是数据工作的初始阶段。数据工作究竟能在哪些方面对治理污染有所帮助呢？一方面，大数据可以预测污染，让人们及时做出反应、减少损害，例如在雾霾严重的时候减少户外活动，甚至改变生活作息等；另一方面，利用数据分析结果切实减少污染。

雾霾天气是否可以预测呢？笔者决定尝试一下。把 2013—2015 年的数据处理为月度数据，利用时间序列的 ARIMA 模型来展开分析。

从图 1-34 所示的散点图来看，雾霾的严重程度存在明显的周期。计算出春、夏、秋、冬四季的季节指数依次为：0.93、0.72、1.00、1.36，可以看出冬季的平均 PM2.5 浓度明显高于全年的平均值。

然后，结合 BIC 信息选择合适的时间序列模型，使用条件最小二乘法估计参数，参数全部显著，可以得到拟合的模型为：

$$x_t=(1+0.79714B+0.62765B^2)\varepsilon_t$$

根据该模型，可以对未来几个月 PM2.5 的浓度做出预测。例如，

洞见数据价值.indd 102 2018/1/15 16:14:23

2016 年 1 月的平均 PM2.5 浓度预测值为 60.146，再乘以冬季的季节指数 1.36，预测值为 81.79856。这个预测值与美国大使馆官网上公布的数据计算出来月平均 PM2.5 浓度为 70.707 非常接近。

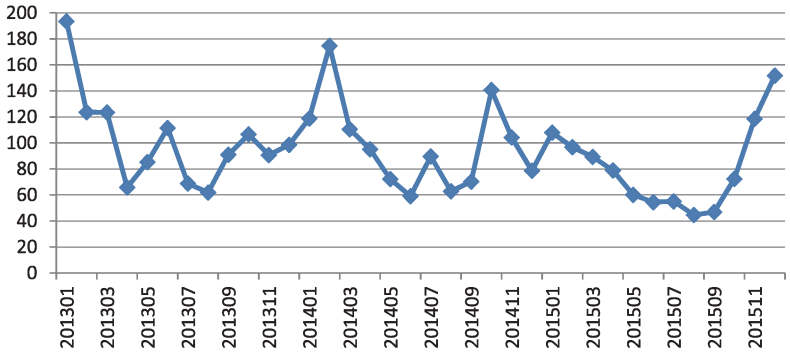

图 1-34　雾霾的严重程度存在明显的周期

不过，单纯使用时间序列模型来预测雾霾比较粗糙，还是存在一定偏差的，因为在使用模型之前，很难考虑风、能源、冬季取暖等复杂因素的影响，尤其是北风对雾霾的影响非常大，这是理论模型在应用中遇到的最大问题：理论模型都将一定合理的前提作为假设条件，而获取的数据一般是包含各种复杂信息的综合性数据。如果首先把对雾霾影响较大的因素进行量化，然后对 PM2.5 的数据进行修正，再对修正之后的数据应用时间序列模型，或许效果会更好一些。

此外，在预测方面，所使用的数据还不够精细。目前全北京市的空气污染监测点只有 300 处左右，如果能增加监测点，获得每一条街道、每一个小区的污染数据，就应该能够搭建更全面的动态模型，不但能够预测雾霾的整体趋势，还能监控污染来源，例如，此处的

洞见数据价值.indd 103　2018/1/15 16:14:23

生产是否严重影响了空气质量？同样，如果在北京周边省份建立同样的监测体系、加强数据获取，就可能找到多省市联合抵抗污染的方法。这样做应该可以为雾霾的防控提供更精准的参考依据。

傍晚下班回到家中，宝贝一见到笔者，用委屈的神情望着笔者，说："妈妈，我想下楼玩儿！"

"今天不行。"笔者安抚说，"你看外面都黑乎乎的。等周末妈妈带你出去玩儿？"

"可是我和他们说好要玩球的……"

笔者对他解释说，这样的天气，他的小伙伴们肯定也都待在家里；玩球今天不成，明天为时不晚。儿子听了笔者的话，想到明天、周末，似乎开心了一点。

明天、周末……其实孩子的未来的确很长，但童年总是短暂。笔者真的不希望孩子在最活泼的几年间，总有那么一些日子必须躲在家里，甚至要承受呼吸道疾病之痛；笔者更想记录下他在蓝天白云下自由奔跑、玩耍的年代。看着孩子对窗外期盼的眼神，笔者希望自己能为中国的治霾事业做点什么，例如白天思考过的雾霾预测预警体系。当然，治霾事业所需要的远不止一个预测预警体系。任重道远，时间紧迫，需要各行各业的共同努力，也包括大数据工作的支持。

1.11 大数据在体育产业

毕马威与体育界一直有着不解之缘。近来，相关同事利用云计算技术为 NBA 联盟制定赛程安排计划的新闻已经火遍全网。篮球需要大数据，足球亦然，那么，大数据可以为足球做些什么呢？

洞见数据价值.indd 104 2018/1/15 16:14:23

创新方向：欧洲杯，跟着西班牙队学数据挖掘！

2016 年的欧洲杯可谓是集冷门之大成，在笔者看来，冷门中的冷门当属 6 月 28 日凌晨的一场重量级对决——西班牙队对意大利队。虽然卫冕的西班牙队在 2014 年世界杯之后似乎已不复当年，但依然是夺冠大热门；反观意大利队，大家都觉得他们过不了西班牙队这一关。结果，西班牙队反而完败在意大利队手中。

现在看来，西班牙队中场统治力持续下滑，锋线上莫拉塔诺利托这样的组合听起来又远远没有比利亚托雷斯有震慑力，如果想要有所成就，战术上的优势格外重要。不妨扮演一下事后诸葛亮：对阵意大利队，西班牙队前场该上谁？

并不是每个球队的前场组合都像巴萨的 MSN 或皇马的 BBC 一样显而易见，对传控球队西班牙队来说更是如此。不过，或许事实隐藏在数据之间，下面利用数据挖掘中的关联算法来挖掘一下小组赛上西班牙队最有威胁的前场三人组合。

> **注意**
>
> 本节宗旨在于借用西班牙队的事例展示关联规则算法，其结果并不具有足球意义上的参考性。

小组赛中西班牙队对阵土耳其队的第三个进球过程。这个进球鲜明地体现出了西班牙队的打法特点，连续 21 脚传球，9 名队员参与其中，球员之间的配合对球队而言极为重要。那么，前场哪些球员的配合最为有效呢？

三场小组赛中，除去拉莫斯罚失的争议点球外，西班牙队一共打入 5 球，具体数据如表 1-8 所示。

洞见数据价值.indd 105　2018/1/15 16:14:23

表 1-8 三场小组赛中西班牙的进球情况

倒数第二传	最后一传	进球者	对应比赛
佩德罗	伊涅斯塔	皮克	西班牙队 vs 捷克队
阿尔巴	诺利托	莫拉塔	西班牙队 vs 土耳其队
莫拉塔	法布雷加斯	诺利托	
伊涅斯塔	阿尔巴	莫拉塔	
席尔瓦	法布雷加斯	莫拉塔	西班牙队 vs 克罗地亚队

这 5 个进球的球员名字的集合称为总项集：{ 莫拉塔，法布雷加斯，伊涅斯塔，阿尔巴，诺利托，席尔瓦，佩德罗，皮克 }。

一些基本概念

1. 关联规则

如果两个不相交的非空集合 M、N，如果有 $M \rightarrow N$，就说 $M \rightarrow N$ 是一条关联规则。西班牙对阵捷克时，{ 佩德罗 } → { 伊涅斯塔 } 就是一条关联规则。关联规则的强度用支持度 (support) 和置信度 (confidence) 来表示。

2. 支持度

支持度的定义：support($M \rightarrow N$)=M 与 N 同时出现的次数 / 数据记录的个数。例如：support({ 莫拉塔 } → { 阿尔巴 })= 莫拉塔和阿尔巴同时出现的次数 / 总进球数，即 2/5=0.4。

3. 置信度

置信度的定义：confidence($M \rightarrow N$)=M 与 N 同时出现的次数 /M 出现的次数。例如，confidence({ 莫拉塔 } → { 阿尔巴 })= 莫拉塔和阿尔巴同时出现的次数 / 莫拉塔出现的次数，即 2/4=0.5。

这里定义的支持度和置信度都是相对的支持度和置信度，不是绝对支持度，绝对支持度 = 数据记录数 N × support。支持度和置信

洞见数据价值.indd 106 2018/1/15 16:14:23

度越高，说明规则越强，关联规则挖掘就是挖掘出满足一定强度的关联。

关联规则挖掘

关联规则挖掘实际就是在给定的一个场景中，找出其中所有支持度 support 大于等于 cut_support、置信度 confidence 大于等于 cut_confidence 的关联规则。这里的 cut_support 和 cut_confidence 是阈值，需要人为设定。最直观的想法，就是穷举所有的组合，逐一测试这些组合的支持度和置信度是否满足相应的条件。但是通常情况下，备选的组合会是一个非常庞大的数字，巨大的时间成本促使人们必须寻找更高效的解决思路。关联规则挖掘大体上可以分为两步：

（1）这一阶段找出所有满足最小支持度的项集，找出的这些项集称为频繁项集。

（2）在频繁项集的基础上生成满足最小置信度的规则，产生的规则称为强规则。

其中第（1）步是比较耗费时间的，因为频繁项集的生成往往需要测试很多的备选项集，但是最终找出的频繁项集的数目一般不会太大，据此生成的满足最小置信度的规则也就不会花费太多的时间。下面介绍两个 Apriori 定律，利用这两个定律可以极大地减少生成频繁项集的时间。

定律 1：如果一个集合是频繁项集，则它的所有子集都是频繁项集。举例来讲，假设 $\{X,Y\}$ 是频繁项集，即 X 和 Y 同时出现的次数大于等于最小支持度 cut_support，则它的子集 $\{X\}$、$\{Y\}$ 出现次数必定大于等于 cut_support，即它的子集都是频繁项集。

定律 2：如果一个集合不是频繁项集，则它的所有超集都不是

洞见数据价值.indd 107　　2018/1/15 16:14:23

频繁项集。举例，假设集合 {*X*} 不是频繁项集，即 *X* 出现的次数小于 cut_support，则它的任何超集如 {*X*、*Y*} 出现的次数必定小于 cut_support，因此其超集必定不是频繁项集。

这两个定律显然是成立的，通过它们可以抛掉很多的候选项集，快速找到频繁项集。

威胁三人组挖掘

下面利用前面介绍的关联规则算法，寻找小组赛中西班牙队最有威胁的前场三人组。

第一步，生成一级频繁项集。汇总出总项集中所有球员的出现次数，如表 1-9 所示。

表 1-9 总项集中所有球员的出现次数

球员名字	出现次数	球员名字	出现次数
莫拉塔	4	诺利托	2
法布雷加斯	2	席尔瓦	1
伊涅斯塔	2	佩德罗	1
阿尔巴	2	皮克	1

将出现次数大于等于 2 的球员保留作为一级频繁项集，即 { 莫拉塔，法布雷加斯、伊涅斯塔、阿尔巴，诺利托 } 是一级频繁项集。

第二步，在一级频繁项集的基础上生成二级频繁项集。根据 Apriori 定律，生成所有可能的组合，如表 1-10 所示。

其中，次数表示在比赛中两人共同出现的次数，例如 { 阿尔巴，诺利托 } 只有在对阵土耳其队时两人有过一次进球配合。这里将次数大于等于 1 的组合保留下来作为二级频繁项集，那么只有 7 个组

洞见数据价值.indd 108 2018/1/15 16:14:23

合保留下来，这 7 个组合就是表 1-10 中灰色底纹对应的球员组合。

表 1-10　所有可能的组合

球员组合	出现次数	球员组合	出现次数
莫拉塔，法布雷加斯	2	法布雷加斯，阿尔巴	0
莫拉塔，伊涅斯塔	1	法布雷加斯，诺利托	1
莫拉塔，阿尔巴	2	伊涅斯塔，阿尔巴	1
莫拉塔，诺利托	2	伊涅斯塔，诺利托	0
法布雷加斯，伊涅斯塔	0	阿尔巴，诺利托	1

第三步，生成三级频繁项集。根据 Apriori 定律，在二级频繁项集的基础之上，只有如下 5 个组合才是三级频繁项集，如表 1-11 所示。

表 1-11　三级频繁项集中的组合

球员组合	
莫拉塔，阿尔巴，伊涅斯塔	莫拉塔，诺利托，阿尔巴
莫拉塔，诺利托，法布雷加斯	法布雷加斯，诺利托，阿尔巴
伊涅斯塔，阿尔巴，诺利托	

例如｛莫拉塔，诺利托，伊涅斯塔｝中的｛诺利托，伊涅斯塔｝不是二级频繁项集，因此这一组合也就不可能是三级频繁项集。

小结

从上面的挖掘结果来看，共有 5 个三级频繁项集，但是由于阿尔巴是左后卫，不属于前场球员，因此最后只有｛莫拉塔，法布雷加斯，诺利托｝构成了西班牙队小组赛的最有威胁的前场三人组合，这一组合中的任意一个人或任意两个组合都是频繁项集。

洞见数据价值.indd 109　2018/1/15 16:14:23