第 5 章

数据产品及开发

如何开始学习

【学习目的】

- 【掌握】数据产品的类型、主要特征及开发方法。
- 【理解】数据能力的评估方法、数据治理的主要内容、数据柔术的基本思想。
- 【了解】数据战略的制定要求。

【学习重点】

- 数据产品的开发方法。
- 数据能力的评估方法。
- 数据治理的主要内容。
- 数据柔术的基本思想。

【学习难点】

- 数据产品的设计。
- 数据柔术的基本思想。
- DMM 模型的应用。

【学习问答】

序号	我的提问	本章中的答案
1	数据产品是什么？与传统产品之间的区别是什么？	定义(5.1节)、主要特征(5.2节)
2	如何开发数据产品？	关键活动(5.3节)、数据柔术(5.4节)、数据能力(5.5节)、数据战略(5.6节)、数据治理(5.7节)
3	数据产品开发需要具备哪些基本功？	数据柔术(5.4节)、数据能力评估(5.5节)、数据战略制定(5.6节)、数据治理(5.7节)
4	数据管理与数据治理的区别是什么？	数据治理与数据管理的区别(5.7节)
5	数据柔术是什么？如何掌握数据柔术？	数据柔术(5.4节)
6	如何评估一个组织机构的数据管理能力？	数据能力(5.5节)
7	如何制定一个组织机构的大数据战略？	数据战略(5.6节)

数据产品开发是数据科学的重要研究任务之一，也是数据科学区别于其他科学的重要研究任务。与传统产品开发不同的是，数据产品开发具有以数据为中心、多样性、层次性和增值性等特征。数据产品开发是数据科学的主要抓手，也是传统产品的下一轮创新和更新换代的关键所在。

数据产品开发案例1——Metromile项目及保险产品的创新

Metromile是2011年在美国旧金山成立的一家汽车保险机构。在传统汽车保险中，无论行车多或少，所缴的汽车保费是固定不变的，这对于那些行车少的人明显不够公平。

根据Metromile提供的数据，65%的车主支付了过高的保费以补贴少数行车最多的人。Metromile提供的是按里程收费的汽车保险，以改变传统的固定收费模式，让行车少的人支付更少的保费，实现里程维度上的个性化定价。

Metromile提供的车险由基础费用和按里程变动费用两部分组成，其计算公式为：每月保费总额＝每月基础保费＋每月行车里程×单位里程保费。其中，每月基础保费和单位里程保费会根据不同车主的情况有所不同(例如年龄、车型、驾车历史等)，每月基础保费一般为15～40美元，按里程计费的部分一般是2～6美分每英里(1英里＝1.609 344千米)。Metromile还设置了保费上限，当日里程数超过150英里(华盛顿地区是250英里)时，超过的部分不需要再多缴保费。

之所以能够实现按里程计算保费，源于物联网等信息技术的应用。车主需要安装一个由Metromile免费提供的OBD设备——Metromile Pulse，以计算每次出行的里程数。配合手机App，Metromile还能为车主提供更多的智能服务，例如最优的导航线路、查看

油耗情况、检测汽车健康状况、汽车定位、一键寻找附近修车公司、贴条警示等服务，并且每月会通过短信或者邮件对车主的相关数据进行总结[①]。

数据产品开发案例 2——Amazon 专利及电商产品的创新

在顾客购买之前，电商已经知道顾客近期会买什么并把货物送到您家附近。本案例为读者解读亚马逊的一项重要发明——Amazon's Anticipatory Shipping（预期送货），具有很强的开创性，是数据科学领域的经典实践之一。

1. 提出者

提出者是 Amazon Technologies Inc 的 Joel R. Spiegel 等。

2. 提出时间

2004 年首次申请专利，后全文并入新专利中，于 2013 年底发布。

3. 提出目的

提出目的是降低物流成本，缩短顾客收货时间。

4. 基本思路

这项专利采用的是大数据预测性分析技术，属于数据科学中的数据产品开发范畴。其基本思路为预测顾客需求，提前运送商品到目的地区域，在运输中匹配订单，确定最终送货地址。主要创新之处在于提出预期运输的方法和计算机系统，并应用于预测先前物品状态，确定包裹的位置、成本、风险、重定向及顾客动机。

5. Amazon 应用

据美国国家公共电台报道，自亚马逊取得“预期送货”专利之后，它在全国各地建立了庞大的仓储业务，并且持续在靠近市中心的地方增加小型仓库。后推出 Prime Now 超快速交付选项。Prime Now 会员可以享受免费 2 小时到货。

5.1　定义

数据产品（Data Product）是指“能够通过数据来帮助用户实现其某一个（些）目标的产品”。数据产品是在数据科学项目中形成，能够被人、计算机以及其他软硬件系统消费、调用或使用，并满足他们（它们）某种需求的任何产品，包括数据集、文档、知识库、应用系统、硬件系统、服务、洞见、决策及其各种组合。需要注意的是：

- 数据产品开发涉及数据科学项目流程的全部活动。数据产品不仅包括数据科学项目的最终产品，而且也包括其中间产品以及副产品。例如，图 3-1 所示的数据科学的基本流程中的每个活动产生的中间产品均可称为“数据产品”。

① 晓保. Metromile：更公平的车险[J]. 金融经济，2018(17).

- 与传统物质产品不同的是,数据产品的消费者不仅包括人类用户,还包括计算机以及其他软硬件系统。其实,数据产品被计算机以及其他软硬件系统调用和“消费”的过程是“数据转换为能源和材料的过程”,进而可以推动信息化和工业化深度融合。
- 数据产品的存在形式有多种,不仅包括数据集,还包括文档、知识库、应用系统、硬件系统、服务、洞见、决策或它们的组合。

从数据流的视角看,“数据产品的开发过程”是一个“数据加工”(Data Wrangling 或 Data Munging)的过程。通常,数据产品开发需要一种特殊的方法和技术——数据柔术(Data Jujitsu),如图 5-1 所示。

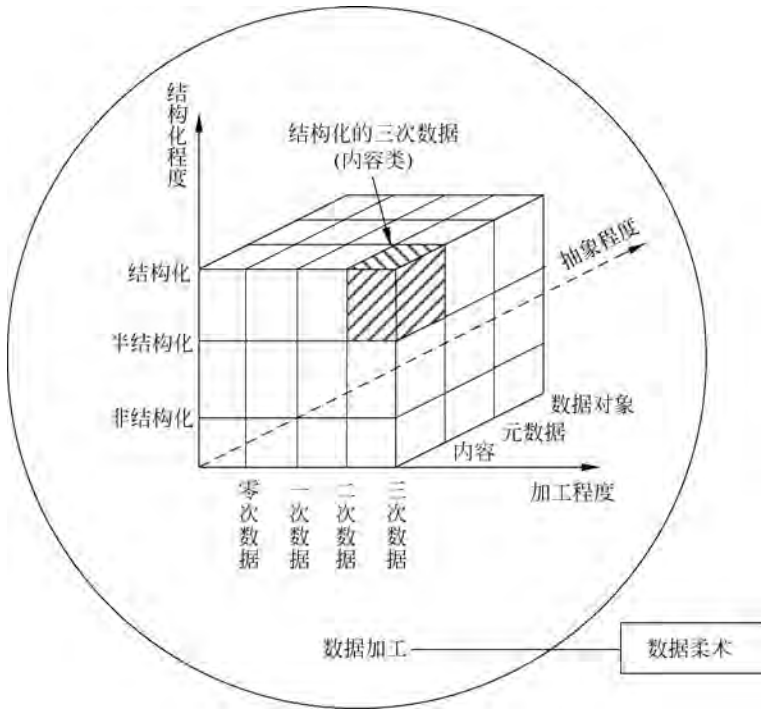

图 5-1 数据产品开发中的数据与数据柔术

1. 数据加工

数据产品开发的关键环节是数据加工。从实现方式看,数据加工是一种数据转换过程,可分为单维度转换和多维度转换。

- **单维度转换**。在数据加工过程中,从结构化程度、加工程度和抽象程度等多个维度(见图 5-1)中选择某一维度,并在此维度上进行数据转换。例如,将非结构化数据转换为结构化数据。
- **多维度转换**。数据加工的工作中也可以在不同维度之间进行转换,例如将零次半结构化数据转换为二次结构化数据。

需要注意的是,数据科学中的数据加工不完全等同于传统意义上的数据转换。二者的

主要区别在于：数据加工过程更强调的是将数据科学家的3C精神融入数据转换过程，追求的是数据处理过程的创新与增值，如表5-1所示。

表5-1　数据转换与数据加工的区别

原始数据	以学生基本信息为例
传统意义上的数据转换	删除脏数据、合并冗余数据，并存入关系数据库
数据科学中强调的数据加工	通过学生基本信息与社交信息的互联，进行关联分析，从而实现创新与增值，如在学生基本信息中增加新的字段——社交能力

2. 数据柔术

数据产品开发的关键技术是数据柔术。从目标与对象看，数据柔术属于数据处理方法。但是，与传统意义上的数据处理方法不同的是，数据柔术更加强调的是数据科学家的主观能动性、创造性思维和艺术设计能力。关于数据柔术的详细介绍，请参见5.4节。

Google公司的数据产品开发

有统计显示，Google公司的产品和服务已经超过200余种（Mahesh Mohan，2016）。据Alexa Traffic的统计结果显示，Google公司的十大产品和服务如表5-2所示。不难发现，Google公司的产品和服务，尤其是这十大产品和服务的开发具有一个共同的数据基础——Google搜索引擎爬取的原始数据。

表5-2　Google公司的十大产品和服务

Where do visitors go on google.com?	
Subdomain	Percent of Visitors
google.com	72.14%
mail.google.com	49.44%
accounts.google.com	34.27%
docs.google.com	13.49%
plus.google.com	10.69%
drive.google.com	7.07%
transtate.google.com	6.63%
support.google.com	5.65%
maps.google.com	5.28%
adwords.google.com	3.46%
play.google.com	2.88%
news.google.com	2.30%
developers.google.com	1.09%
productforums.google.com	1.02%

续表

Where do visitors go on google. com?	
Subdomain	Percent of Visitors
sites. google. com	1.01%
code. google. com	0.94%
urt. google. com	0.81%
groups. google. com	0.77%

(来源：Alexa Traffic stats of Google. com)

5.2 主要特征

相对于传统意义上的其他产品，数据产品的主要特征如下。

1. 以数据为中心

“以数据为中心”是数据产品区别于其他类型产品的本质特征。数据产品“以数据为中心”的特征不仅体现在“以数据为核心生产要素”，而且还表现在以下三个方面。

- **数据驱动**。数据产品开发的目的、方法、技术与工具的选择往往是由数据驱动的，不再是传统产品开发中的常用驱动方式，如目标、决策或任务驱动。
- **数据密集型**。数据产品开发的瓶颈和难点往往源自数据，而不再是计算和存储。也就是说数据产品开发具备较为显著的数据密集型的特点。
- **数据范式**。数据产品的开发往往采用“基于数据的研究范式”(即数据范式)，其方法论往往属于历史经验主义的范畴。然而，传统产品开发往往依赖“基于知识的研究范式”(即知识范式)，其方法论通常属于理论完美主义的范畴，如图 5-2 所示。

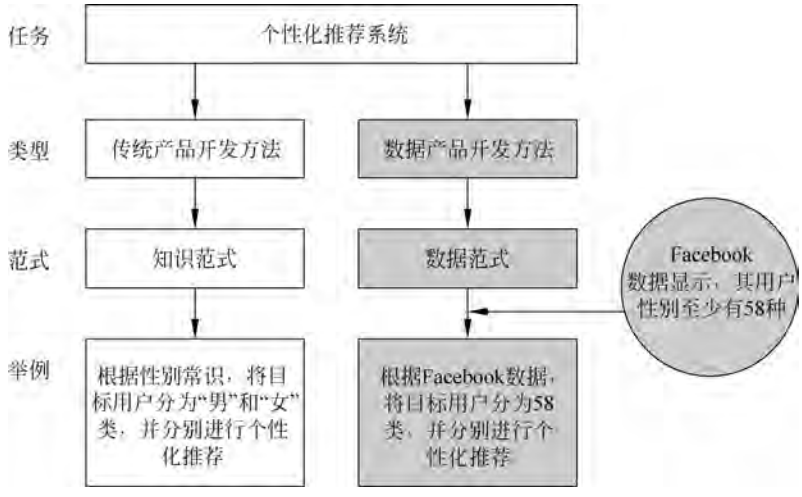

图 5-2 知识范式与数据范式

Facebook 中的 70 多种性别——数据范式与知识范式的差异

传统知识告诉我们，在理论上，人类的性别只有两种——“男”与“女”。与此不同的是，在实践中，数据表明人类的性别选择可能有多种。例如，在 Facebook 性别选择中给出了超过 70 多种性别，例如：

- Androgyne(雌性同体)。
- Bigender(双性别)。
- Cis Female(Cis 女)。
- Cis Male(Cis 男)。
- Female to Male(女性到男性)。
- Male to Female(男性到女性)。
- Transsexual Female(变性女)。
- Transsexual Male(变性男)。
- Transsexual Person(变性人)。

……

值得思考的是，假如想基于性别推荐某一产品，如何做呢？可能有两种选择：

- 如果采用知识范式，则推荐策略有两种——男和女。
- 如果采用数据范式，则推荐策略可能与知识范式不同——不再是两种，可能多达 70 种。

2. 多样性

从产品形态看，数据科学中的“数据产品”并不是特指某一类型的产品，如数据集、知识库或应用系统。相反，数据产品的存在或(和)表现形式可以有多种，如图 5-3 所示。

图 5-3　数据产品的多样性

- **数据类产品**。对输入数据进行清洗、脱敏、集成、归约、标准化和标准等处理后形成的,以数据形式输出的产品或服务,如干净数据。
- **信息类产品**。将数据转换成信息之后,以信息形式输出的产品或服务,如数据新闻,数据订阅,报告、快报、摘录和定题服务等。
- **知识类产品**。将数据转换成知识之后,以知识形式输出的产品或服务,如百科全书、语料库、领域本体、知识库、规则库等。
- **智慧类产品**。将数据转换成智慧之后,以智慧形式输出的产品或服务,如决策支持、数据洞见、数据业务化、数据驱动等。

3. 层次性

从加工程度看,“数据产品”的另一个特征是层次性。可以将数据产品分为内容、应用、服务和决策四个不同层次,如图 5-4 所示。

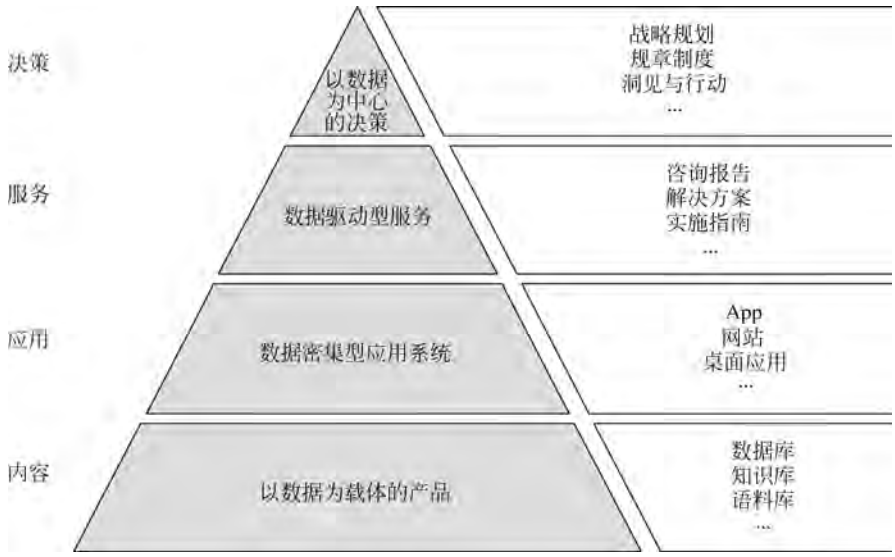

图 5-4 数据产品的层次性

- **内容类产品**。以数据为载体的产品,即对输入数据进行一定的数据加工处理之后得到的结果,如新的数据库、知识库和语料库等。
- **应用类产品**。以数据密集型应用系统为载体的产品,如 App、网站或桌面应用等。如图 5-5 给出的 Google 全球商机洞察(Google Global Market Finder)是一种典型的应用类产品,可以帮助客户找到所需的商机,并将客户广告推送给全球用户。
- **服务类产品**。以数据驱动型服务为主的产品,如咨询报告、解决方案及实施指南等。
- **决策类产品**。以数据为中心的决策,主要指数据视角下的战略规划、规章制度、洞见与行动等。

图 5-5　Google 全球商机洞察

4. 增值性

从价值维度看，数据产品的开发过程应是“增值”过程，将数据科学家的 3C 精神融入数据产品开发活动之中，进而实现数据产品的增值。

- **创造性地工作**。数据产品的设计应将数据科学家的创造性思维加入数据产品的主要创新与增值活动，数据产品的设计具有较高的原创性、艺术性和突破性。
- **批判性地思考**。数据产品的研发过程需要采用批判性思考方式，对于已有相关产品及新产品的开发过程均应采取批判性思考方式，不断改进产品的质量。
- **好奇性地提问**。提出一个好问题是成功开发一个数据产品的重要前提。通常，数据产品开发的难点是如何提出一个好问题。

可见，增值性是数据产品开发与传统意义上的数据处理的主要区别之一。需要注意的是，增值活动贯穿于数据产品开发的全过程，包括数据对象的封装、集成与服务的所有环节，如图 5-6 所示。

- **数据对象的封装**。将数据内容及其元数据封装成“数据对象”。例如，Google 将网络爬虫收集的数据内容、来源、点击率、用户评价等元数据封装成一个“数据对象”，并以搜索结果的形式提供给用户。
- **数据系统的研发**。在数据对象的封装基础上，开发出特定的软件系统（如 Google 翻译）、硬件系统（如 Google 眼镜）或基础设施（如 Google MapReduce、BigTable、GFS 等）。
- **集成应用**。在开发特定数据产品的基础上，将多个数据产品（如软件系统、硬件系统、基础设施）进一步集成为新产品。例如，Google 结合自己的搜索数据及软硬件

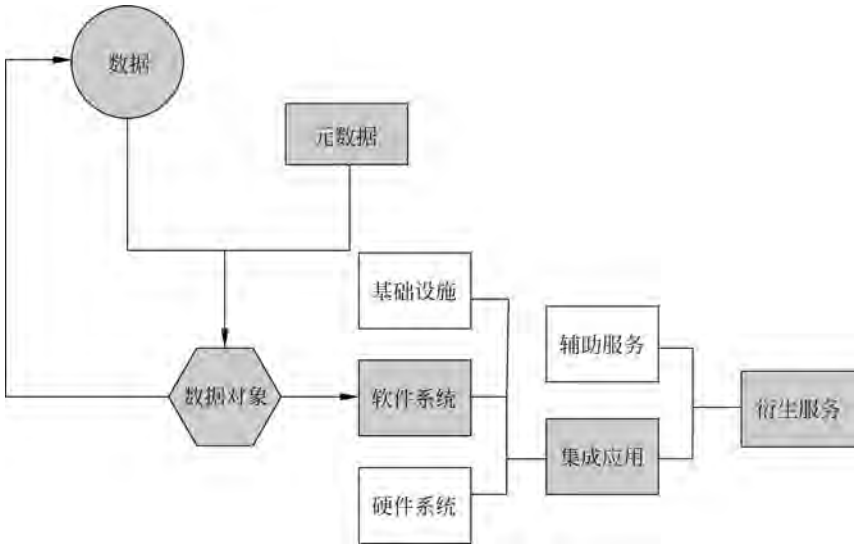

图 5-6 数据产品链

系统,提供集成应用 Google App Engine。

- **辅助服务**。在数据、软件系统、硬件系统、基础设施的基础上,还可以提供辅助服务类数据产品。例如,Google 基于自己的大数据以及 BigTable、服务器设备等软硬件系统提供一些辅助服务,如 Google Docs、委托开发、委托维护、外包等。
- **衍生服务**。在集成应用和辅助服务的基础上,数据产品开发还可以提供一些衍生服务。例如,第三方机构针对 Google 的集成服务和辅助服务,提供的市场咨询、决策支持、数据的深度开发等衍生服务。

5.3 关键活动

为了确保数据产品的基本特征,数据产品开发应遵循的基本原则和应特别予以重视的主要活动要素如下。

1. 基本原则

通常,传统 IT 产品的开发遵循的是“三分技术、七分管理和十二分数据”的原则——技术固然很重要,管理比技术还重要,但更重要的是数据,因为数据比“技术+管理”还关键。与此不同的是,数据产品开发中首先关注到的是“数据”,也就是说,数据产品开发中“数据”当然很重要;但是,智慧(开发数据产品的艺术)比“数据”还重要;然而最重要的是“用户体验”,即“三分数据、七分智慧和十二分体验”原则,如图 5-7 所示。

图 5-7　传统产品开发与数据产品开发的区别

“三分数据、七分智慧和十二分体验”原则反映了数据产品开发中应予以重视的三个基本问题：

- 数据是数据产品开发的原材料。
- (数据科学家的)智慧是数据产品开发的主要增值来源。
- (用户的)体验是数据产品的主要评价指标。

2. 活动要素

数据产品开发工作之中需要特别注意的基本活动有以下几项：

- 创造性设计。
- 数据洞见。
- 可视化。
- 故事化描述。
- 虚拟化。
- 按需服务。
- 个性化服务。
- 安全与隐私保护。
- 用户体验。
- 政策分析。

5.4　数据柔术

数据柔术是指将“数据”转换为“产品”的艺术。数据柔术是由D. J. Patil(见图 5-8)提出的一个新术语。在他看来，数据产品开发与古代柔术有很多相似之处——“借助对方的力量(而不是自己的力量)获得成功的艺术”。因此，数据产品开发的难点在于“如何借助目标用户的力量来解决数据产品中的难题”。数据柔术强调两个基本问题：一是产品开发要有较高的艺术性；二是以目标用户为中心的产品开发。

图 5-8　D. J. Patil

D. J. Patil

著名数据科学家,1974 年生,毕业于马里兰大学帕克分校,获得应用数学博士学位。D. J. Patil 是美国白宫第一任首席数据科学家(2015—2017),曾任 LinkedIn 首席科学家和数据产品团队负责人,并在 Greylock Partners、Skype、PayPal 和 eBay 等多家企业担任过重要角色。

他撰写(或参与撰写)的专著《如何构建数据科学家团队》(*Building Data Science Teams*)、《数据柔术——将数据转换成产品的艺术》(*Data Jujitsu: The art of turning data into product*)、《数据驱动——培育数据文化》(*Data Driven: Creating a Data Cultur*)(与著名数据科学家 Hilary Mason 合著)以及论文《数据科学家——21 世纪最性感的职业》(*Data Scientist: The Sexiest Job of the 21st Century*)(与著名管理学家 Thomas H. Davenport 合著)在数据科学领域产生了重要影响。Tim O'Reilly 曾提到,D. J. Patil 和 Jeff Hammerbacher 一起创造了术语"数据科学"(Data Science)。

1. 引入设计思维

产品设计是数据产品开发中不可忽略的重要活动。设计质量的好坏将直接影响数据产品的服务质量与用户体验。以某个数据产品中的输入框——用户的毕业院校为例,用户填写自己的毕业院校信息时,可能将同一个学校名称写成多种形式(如图 5-9(a)),具体如下:

- 中国人民大学。
- 人民大学。
- 人大。
- 陕北公学。
- 华北大学。
- Renmin University of China。
- RUC。

……

显然,这些输入数据的多样性会导致后续数据处理的复杂性。为此,可能采取的设计方案有多种,如:

- 下拉列表。通过下拉列表给出所有可能的院校信息,并要求用户"只能以选择方式提交自己的毕业院校",这当然可以,但会导致另一个问题——当院校个数较多时,用户体验很差(图 5-9(b))。

- 单选按钮。通过单选按钮形式给出院校名称，虽然看起来解决了复杂数据的输入问题，但同样会导致另一个问题——当院校个数较多时，用户体验很差(图 5-9(c))。
- 智能提醒。当用户开始输入时系统智能地动态提醒相关学校名称，如输入“中国”二字时，系统自动提示以“中国”二字开头的高等院校名称(图 5-9(d)。
- 其他解决方案，如提供“您是否指的是 *** 学校?”或提供候选学校的校徽。

图 5-9　UI(User Interface)设计方案与设计思维

显然，从功能角度看，上述设计方案都可以实现所需功能，但是，其用户体验却不同。根据设计式思维的观点，数据产品设计应重视充分发挥目标用户通过“前台”界面做出的贡献——在方便用户操作的同时，借助用户力量，有“艺术性”地解决数据产品的难题，而不是数据科学家通过自己设计的复杂算法，在“后台”解决这些难题。有统计数据显示，对于同一个问题而言，数据科学家在“后台”解决的成本往往是目标用户在“前台”解决此问题的 100～1000 倍。

设计思维是数据产品开发的要素。Google 搜索关键字智能提示的主要依据为关键字的出现频率、搜索用户的地理位置及历史记录。以关键字 Renmin University 为例，同一个用户在不同地理位置上用 Google 搜索时，系统给出的智能提示有所不同，如图 5-10 所示。

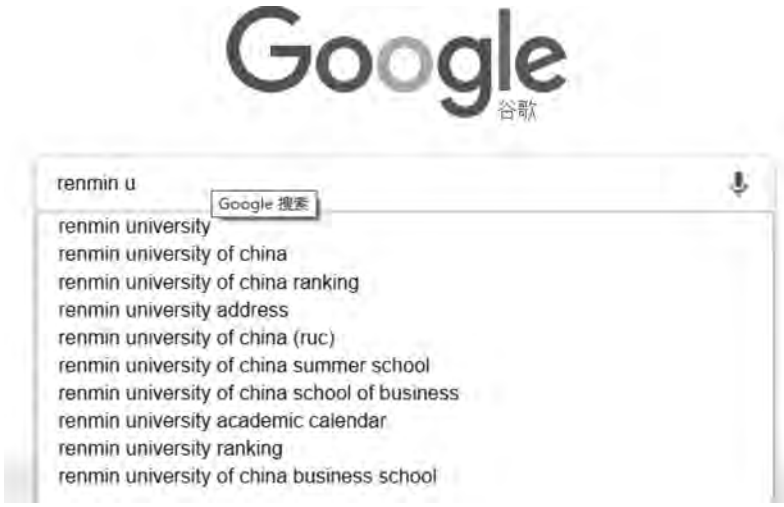

图 5-10 Google 搜索的用户体验

值得一提的是,数据产品的设计中不能忽略可能出现的错误或副产品。以 Google 搜索中的智能提示为例,所给出的智能提示可能造成性别歧视、变相广告甚至造成种族或宗教偏见。

图 5-11 人与计算机图像内容识别能力的不同

2. 支持人机协同

数据产品开发中应正确认识人与计算机在数据处理中的不同优势。以图 5-11 所示的两张照片的内容识别为例,对于人类来说非常容易,很小的孩子就能看出照片中是“长发人”还是“长毛狗”。但是,对于计算机来说却不那么容易。因此,在数据产品的开发中应重视人与计算机的不同优劣势,必要时采取人机协同方式进行数据处理。

亚马逊的一款数据产品——Amazon Mechanical Turk 发挥人与计算机的不同优势,较好地实现了提升数据产品的服务质量与用户体验的目的。

Amazon Mechanical Turk 的数据处理

Amazon Mechanical Turk 是 2005 年由亚马逊公司研发的数据处理平台。其名称来自于 1769 年匈牙利发明家 Wolfgang von Kempelen 研制的会下棋的机器人——Mechanical Turk[①]。其实,Mechanical Turk 机器人只是个道具,由躲在机器里面的下棋

① Amazon Web Services LLC or its affiliates. Amazon Mechanical Turk FAQs[OL]. http://aws.amazon.com/mturk/faqs/.

高手操纵，并没有依靠目前机器人中普遍采用的人工智能技术。Amazon Mechanical Turk[①] 为数据处理中的数据提供方（Workers，又称 Turkers 或“供方”）和数据需求方（Requesters，或“需方”）之间提供了一个合作平台。

Amazon Mechanical Turk 数据处理的主体是人，而不是计算机。需方将数据处理需求分解成较具体的、易于完成的“小任务”——Human Intelligence Tasks（HITs）——之后，通过此平台向供方发布。供方选择自己擅长的 HITs，并完成指定操作。供方提交的结果经需方确认后将得到一定数额的资金回报。Amazon Mechanical Turk 平台还提供了编程接口，软件开发者也可以通过调用平台提供的接口构建自己的应用程序。

相对于传统数据处理平台，Amazon Mechanical Turk 平台具有如下的特殊性。

（1）参与者的长尾性。Amazon Mechanical Turk 平台的需方和供方均具备长尾性。

一方面，此平台对数据处理任务的需求发布者不做任何限制，任何长尾主体注册和登录之后均可通过互联网在此平台上发布自己的数据处理需求。

另一方面，此平台中的数据处理任务由人工完成，而且参与完成的长尾主体是通过互联网选择数据处理任务和提交数据处理结果，一般对数据处理主体的身份和职业不做限制。

因此，参与者的长尾性保证了 Amazon Mechanical Turk 平台的灵活性和低成本性。

（2）获取劳动力的弹性。Amazon Mechanical Turk 平台中劳动力的规模具备弹性特点。在传统数据处理模式中，数据处理的劳动力的获得需要经过一系列的常规过程，如公布招聘信息、简历挑选、组织面试、岗位培训、职责分配、绩效考核等。因此，传统数据处理模式中获取劳动力的即时性较差，对劳动力的利用率较低。

但是，Amazon Mechanical Turk 平台（见图 5-12）改变了这种做法，其劳动力获取是按需的、弹性的[②]，不仅可以很容易获得与特定数据处理任务相对应的劳动力，而且也可以根据任务量和完成情况，调整劳动力的数量和范围。

（3）小任务性。Amazon Mechanical Turk 平台发布的任务粒度都比较小，当需方的数据处理任务粒度较大时，需要进一步分解成一批更小的、更容易完成的数据处理任务，即 HITs。小任务性是此平台的主要特点之一，较好地吻合了长尾主体的数据处理特征和规律，可以充分利用自己的业务时间，在不花费太多精力的前提下，轻松完成供方的任务。一个 HIT 的生命周期包括可指派状态（Assignable）、不可指派状态（Unassignable）、可评审状态（Reviewable）、正在评审状态（Reviewing）和已处置状态（Disposed）等主要阶段，如图 5-13 所示。

---

① Amazon Web Services LLC or its affiliates. Amazon Mechanical Turk[OL]. http://aws. amazon. com/mturk/.

② Amazon Web Services LLC or its affiliates. Overview of Mechanical Turk[OL]. http://docs. amazonwebservices. com/AWSMechTurk/latest/RequesterUI/OverviewofMturk. html.

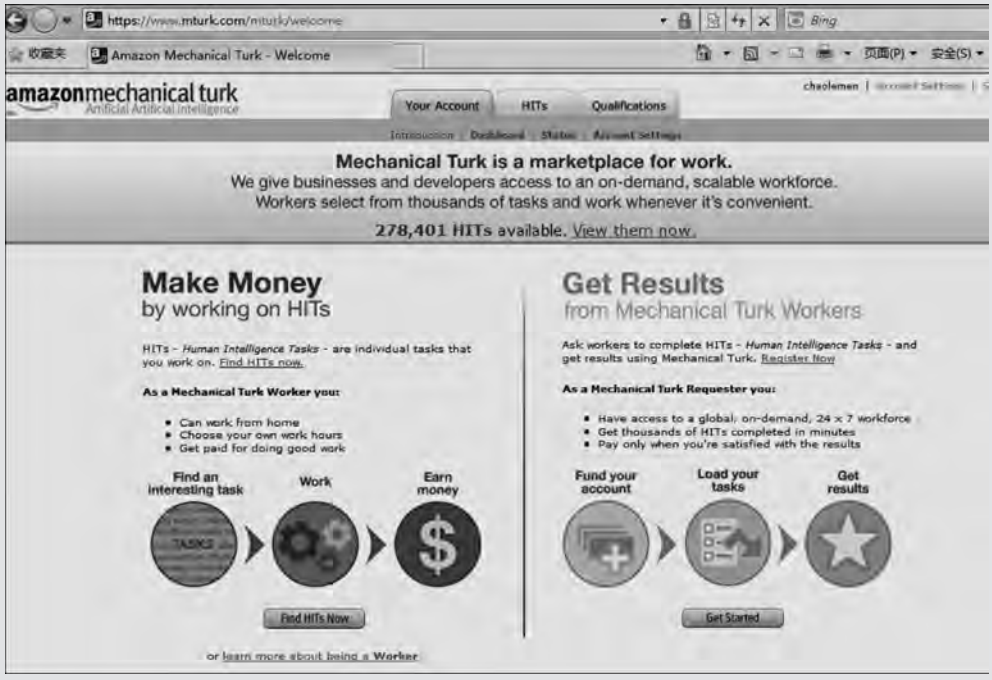

图 5-12　Amazon Mechanical Turk 平台

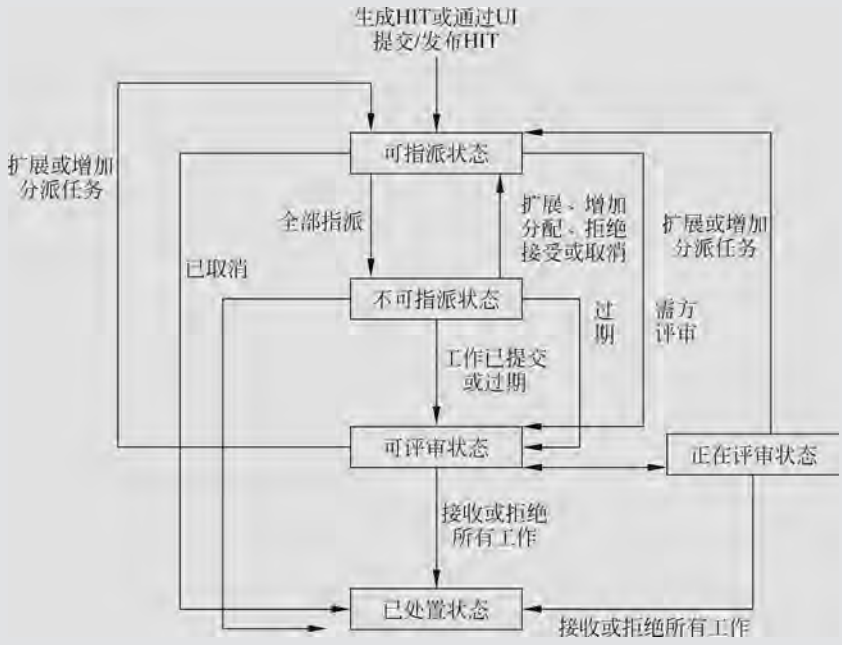

图 5-13　一个 HIT 的生命周期

此外，小任务性也有助于保证供方工作的原始性，便于收集供方的第一感觉或原始想法，避免供方进行过多的修饰和过滤自己的观点。

（4）后支付模式。Amazon Mechanical Turk 平台采用的是“先劳动后支付”的模式，需方在发布任务的同时公布报酬金额。供方完成的小任务，经需方确认后，方可获得相应的报酬。“先劳动后支付”较好地避免了不认真用户的参与，提高了用户在完成任务时的积极性。

（5）资格审查。平台还提供了设置供方的资格条件（Qualification），如地域、领域和诚信度。供方可以采用资格条件选择劳动者类型。

（6）数据处理成本低。通过 Amazon Mechanical Turk 平台进行数据处理时，不需要聘请固定员工和日常管理成本，而是利用长尾主体的力量和网络平台，采取“先劳动后支付”模式，省去了传统数据处理中的员工管理的成本。此平台建议需要对一个 HIT 承诺的最小报酬可以低至 0.005 美元。

因此，当任务量不是太大、复杂度不够大时，通过此平台数据处理成本小于传统数据处理成本。但是，当需方数据处理任务非常复杂、工作量很大、参与完成的供方过多时，此平台上的数据处理成本可能超过传统模式。

可见，如何发挥人与计算机的不同优势是数据产品开发的难点之一。数据科学家往往关注的是基于人或计算机的力量进行数据产品开发时在成本上的差异性。图 5-14 给出了二者的成本曲线。从长远看，基于机器的数据产品的处理成本低于基于人的数据产品。因此，在数据产品的开发初期，可以采取基于人的数据处理模式，当数据产品相对成熟或获得用户认可时，逐渐引入计算机自动化处理技术。

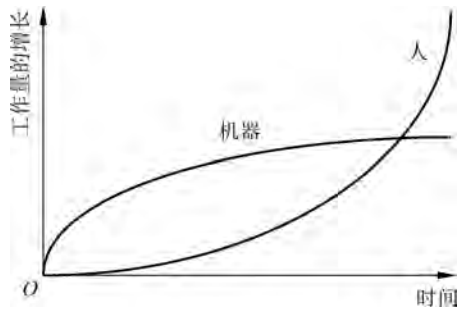

图 5-14　基于人与计算机的数据处理成本曲线

3. 善于留住用户

用户的“中途离开”是数据产品消费中最常见的问题之一。因此，如何留住用户是数据产品开发中值得重视的问题。以亚马逊数据产品——“其他商家”（Other Sellers）为例，该

平台在显示某个图书的详细信息(如书名、作者、价格、用户评论)的同时,还提供了一个比较有创意的功能,即"其他商家",如图 5-15 所示。在此 Other Sellers 选项卡中,列出了正在出售该图书的其他商家及最低市场价格,其用意在于用户不会为了收集其他商家的数据而离开该产品的页面。

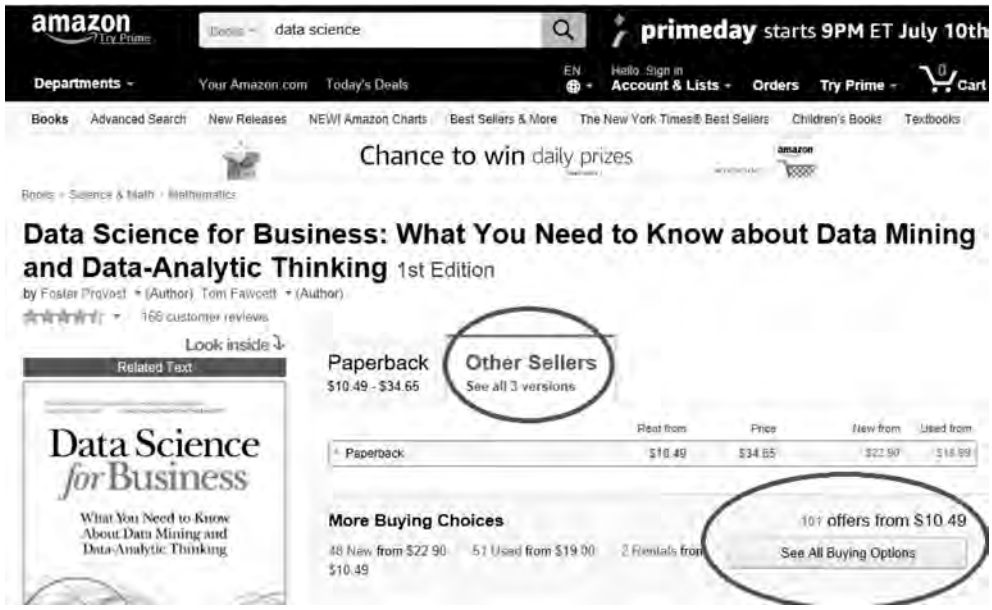

图 5-15 亚马逊的数据产品——其他商家(Other Sellers)

4. "顶天立地"的产品设计

数据产品的设计必须"顶天立地"——既需要一定的创造性、引导用户行为和引领未来的特点,又要结合用户的实际需要,满足用户的实际需求。相对于数据产品的"顶天",数据科学家往往忽略其"立地"。例如,亚马逊的一款数据产品——"你看过的产品,还有谁看过"的思想来自于我们在现实生活中的购物体验——往往在朋友的陪同下购物和/或听取朋友的购买建议。

LinkedIn 也有一款数据产品——"你可能认识的人们"(People you may know)(见图 5-16)的设计思想也源自于现实生活的实际场景——当人们在会议接待处报到时,往往喜欢去寻找自己可能认识的参会者是否也在报到处或已经报到。

5. 数据,取之于民,用之于民

用户不仅是数据的消费者,也是提供者。数据产品开发中应遵循一个基本的原则——"取之于民,用之于民",将用户产生或留下的数据,"以恰当的方式馈赠给用户"。也就是说,数据产品中的数据流并不是单向的,而是数据产品与目标用户之间的双向流动,进而达到数据柔术中提倡的"借助用户力量来解决数据产品中的难题"。

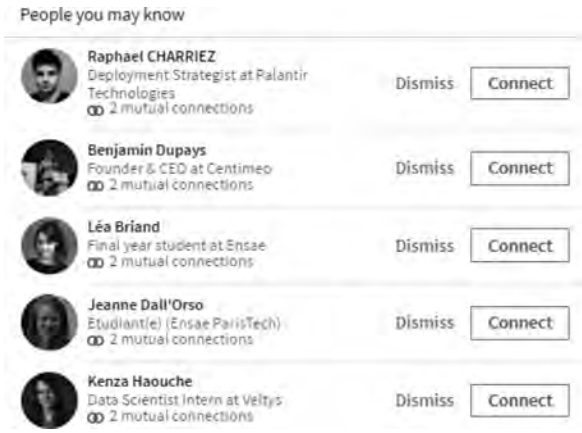

图 5-16　LinkedIn 的数据产品——你可能认识的人们

数据产品开发中实现“取之于民，用之于民”的难点也正是如何找到一个“恰当的方式”馈赠给用户。如果数据产品简单地将用户产生的数据反馈给他们，很容易造成另一个问题——“数据恶心”。

那么，如何实现数据的“取之于民，用之于民”？需要通过数据加工将数据转换成产品，具体做法可以有很多种。例如：

- LinkedIn 以一款数据产品——“你的观众是谁”(Who's viewed your profile)的形式将用户产生或留下的数据返还给用户，进而确保较高的用户体验，如图 5-17 所示。

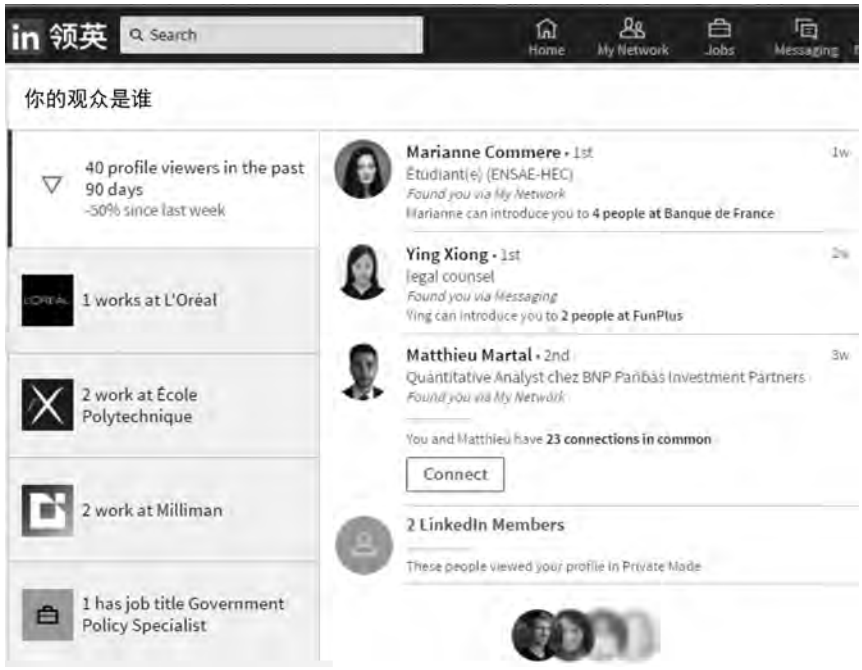

图 5-17　LinkedIn 的数据产品——你的观众是谁

- Xobni 收集和分析用户的 E-mail 信息,并以“收件箱管理功能”的方式返还给用户。
- Mint 收集和分析用户的信用卡信息,并以“帮助目标用户理解自己的消费习惯”的形式返还给用户。
- 智能电表类数据产品往往以“分析你的电力消费习惯”的形式将数据反馈给用户。

6. 避免导致“数据恶心”

数据科学家应避免所开发出的“数据产品”在目标用户群中产生“数据恶心”。也就是说,数据产品的开发必须有效结合目标用户的需求与体验,不能仅仅以数据科学家自己的兴趣爱好或工作需要作为设计基准。因此,数据产品开发应特别注意目标用户与数据科学家对同一个数据产品可能产生的不同体验。例如,数据科学家喜爱的产品,目标用户不一定喜欢,甚至感到“恶心”。

“取之于民,用之于民”是数据产品开发的重要原则。但是,数据的“用之于民”环节很容易导致“数据恶心”——提供过多的数据或过于复杂的人机交互往往会导致目标用户的反感。LinkedIn 在其数据产品“你的观众是谁”的原型系统之中曾提供过“多次单击链接即可查看更详细的内容”的功能,但其产品开发团队的测试结果发现“几乎没有人通过多次单击的方式查看更详细的内容”。

“逆向交互定律”(Inverse Interaction Law)可以解释 LinkedIn 的这款数据产品原型的设计中存在的“数据恶心”的现象。所谓“逆向交互定律”就是“平台提供的数据超过一定规模后,产生的用户交互会越少”,如图 5-18 所示。

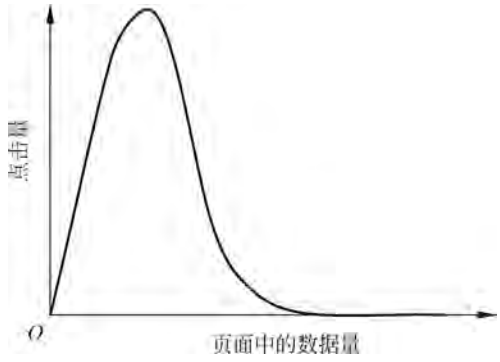

图 5-18 逆向交互定律

避免“数据恶心”的有效方法之一是使数据产品开发活动尽量聚焦在“数据的可操作性”——需要给用户提供哪些操作?这些操作是否是用户真正需要的?用户的操作体验如何?

7. 预估可能产生的“副产品”或“负面影响”

为了更好地实现某一功能与服务,数据科学家往往专注于特定算法的设计,但很容易

忽略可能出现的“副产品”或“负面影响”——在个别情况下得出错误结果，或者产生社会、法律、道德、宗教、舆论等问题。因此，数据产品的开发需要识别各类风险，进行风险评估和风险应对策略，并积极制定应急预案。

德国最高法院判决 Facebook“查找好友”功能违法

北京时间 2016 年 1 月 16 日早间消息，德国最高法院本周维持了两家低级别法院的判决，即 Facebook 帮助用户向联系人推荐该服务的功能违法。

德国联邦最高法院的一个委员会判决，Facebook 的“查找好友”功能构成广告骚扰。这一诉讼由德国消费者组织联盟（VZBZ）于 2010 年提起。

Facebook 的这项功能要求用户提供授权，从而向用户的电子邮件联系人发送邀请注册邮件。这意味着 Facebook 可以向非该公司用户发送推广信息。

法庭认为，这是一种带欺诈性质的营销手段。2012 年和 2014 年，柏林的两家低级别法院也做出了类似判决。当时的判决认为，Facebook 违反了德国的数据保护法，并存在不公平的贸易行为。

德国最高法院还表示，Facebook 未能适当地告知用户，该公司将如何利用联系人数据。对此，Facebook 驻德国发言人表示，正在等待正式判决，并将研究这一判决对该公司的服务有何影响。VZBZ 对这一判决表示欢迎。该组织在公告中表示，对于在德国进行类似广告宣传的其他公司，这一判决具有参考意义。VZBZ 负责人克劳斯·穆勒（Klaus Mueller）表示：“我们需要研究，对于当前的‘查找好友’功能，这一判决意味着什么。除 Facebook 之外，其他服务也会用类似的广告形式吸引新用户。他们现在要重新思考自己的做法……”

（来源：新浪科技）

8. 正确处理查全率、查准率和响应时间之间的关系

数据产品开发中需要综合考虑三个不同的指标——查全率、查准率和响应时间。需要注意的是，这三个指标往往是相互限制，难以同时确保三个指标的最高值。因此，数据科学家在数据产品开发中往往妥协或放弃其中的一个或两个指标，进而确保另一个指标。

- 搜索引擎中的返回结果。可以采取“响应时间优先”策略，做到快速显示搜索结果的目的。
- 搜索引擎中的餐饮类广告信息。采取“查准率优先”策略，根据用户搜索的关键字和地理位置推荐有针对性的广告。
- 搜索引擎中的图书类广告信息。可以采取“查全率优先”策略，尽可能地提供与目标用户输入的关键字相同的图书。

用户体验的重要性

- Aberdeen Group 的调查发现“页面的显示速度每延迟 1s,网站访问量就会降低 11%,从而导致营业额或者注册量减少 7%,顾客满意度下降 16%”。
- Google 公司认为“响应时间每延迟 0.5s,查询数将会减少 20%”。
- Amazon 公司认为“响应时间每延迟 0.1s,营业额下降 1%”。

9. 重视用户认知行为的主观性

数据产品的开发中,应注意用户认知行为的主观性——错误或负面信息往往更容易被目标用户感知,并对整个数据产品产生错误的认知。以 LinkedIn 的数据产品——“岗位推荐”(Jobs you may be interested in)为例(见图 5-19),如果所推荐的 10 个工作岗位中,只要有一个“不良岗位”,多数用户会对整个推荐目录产生不好的印象,工作岗位的推荐会以失败告终。

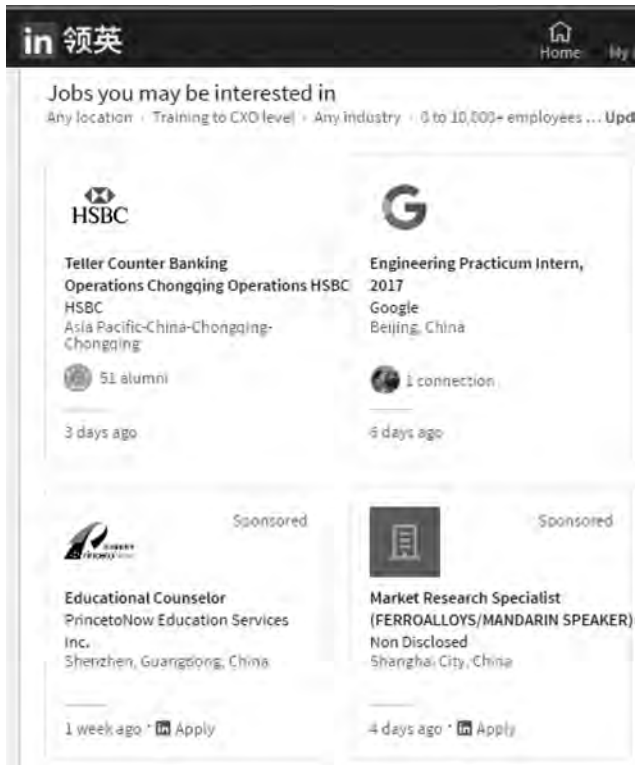

图 5-19 LinkedIn 的数据产品——岗位推荐

因此,数据产品的设计中应重视“最坏的结果”对整个产品的影响——“最坏的结果”对目标用户的主观认识所产生的消极作用往往大于“最好的结果”的积极作用。

10. 招募更多的用户，获得有效的数据

在数据产品的开发中，应重视招募更多的用户，并挖掘用户之间的社交关系。数据科学家可以通过目标用户的“朋友”的数据或响应来训练推荐算法，实现精准推荐或协同过滤的目的，进而避免数据产品中“最坏的结果”所导致的颠覆性负面影响。以 LinkedIn 的职位推荐系统（见图 5-20）为例，在对某一个用户推荐职位列表之前，可以将候选职位发送给目标用户的若干朋友，并根据这些朋友的反馈数据或历史数据来优化推荐结果。因此，数据产品的开发中还需要注意两个问题。

- 需要让用户提供哪些信息以及这些信息是否满足数据产品开发的需求。例如，让用户输入自己的邮政编码和自己的工作单位的邮政编码会对后续数据产品开发产生不同的影响——显然后者更便于处理和分析。
- 在要求用户提供个人信息时，应明确告知收集范围、目的、承诺、利用方式以及未来返还给用户的服务。

图 5-20　LinkedIn 的数据产品——帮助你的朋友找到工作

11. 预见失败及确保良好的用户体验

数据产品开发工作难以避免“失败的结果”。以基于协同过滤的推荐类数据产品为例，在个别情况下，推荐系统可能向某个用户推荐错误的产品。那么，当产生错误或失败的推荐时，数据产品的应对策略尤为重要——是给用户一个“关闭窗口”的功能，还是给用户“重新推荐”的按钮？不同的策略对目标用户产生的体验可能不同。

Facebook 的广告系统较好地解决了“如何在产生失败的推荐时还能确保较好的用户体验”的问题。当用户认为 Facebook 推荐的广告为“失败”的广告时，用户不仅可以隐藏该广告，而且还可以填写“为什么这个广告是失败的广告”，如图 5-21 所示。可见，为用户提供更多的“控制权”和“主动性”是提升用户体验的重要保障。

图 5-21　Facebook 的良好用户体验

5.5 数据能力

在数据管理和数据治理领域,常见容易混淆的术语及其含义如下。

- 数据管理(Data Management):数据获取、存储、整合、分析、应用、呈现、归档和销毁等各种生存形态演变的过程(来源:国家标准《信息技术服务 治理 第5部分:数据治理规范》(GB/T 34960.5—2018))。
- 数据治理(Data Governance):数据资源及其应用过程中相关管控活动、绩效和风险管理的集合(来源:国家标准《信息技术服务 治理 第5部分:数据治理规范》(GB/T 34960.5—2018))。
- 数据处理(Data Processing):数据操作的系统执行(来源:国家标准《信息技术 大数据 术语》(GB/T 35295—2017))。
- 数据战略(Data Strategy):组织开展数据工作的愿景、目的、目标和原则(来源:国家标准《数据管理能力成熟度评估模型》(GB/T 36073.5—2018))。
- 数据架构(Data Architecture):数据要素、结构和接口等抽象及其相互关系的框架(来源:国家标准《信息技术服务 治理 第5部分:数据治理规范》(GB/T 34960.5—2018))。
- 数据生存周期(Data Lifecycle):将原始数据转换为适用于行动的知识的一组过程(来源:国家标准《信息技术 大数据 术语》(GB/T 35295—2017))。
- 元数据:关于数据或数据元素的数据(可能包括其数据描述),以及关于数据拥有权、存储路径、访问权和数据易变性的数据。
- 数据元(Data Element):由一组属性规定其定义、标识、表示和允许值的数据单元(来源:国家标准《信息技术 元数据注册系统(MDR) 第1部分:框架》(GB/T 18391.1—2009))。
- 主数据(Master Data):组织中需要跨系统、跨部门进行共享的核心业务实体数据(来源:国家标准《数据管理能力成熟度评估模型》(GB/T 36073.5—2018))。

从理论上讲,数据能力的评价方法有两种:评价结果(结果派)和评价过程(过程派)。根据软件工程等领域的经验,质量评价和能力评估中通常采用过程派的思想。在数据科学中,数据能力的评价也采取过程评价方法。

数据管理原则

与其他形式的资产管理不同，数据管理需要遵循一些特殊原则，如表 5-3 所示。

表 5-3　数据管理原则

有效的数据管理需要领导层的承诺及参与	
数据是有价值的	数据是具有独特属性的资产
	数据的价值可以而且应该用经济术语来表达
数据管理需求符合业务需求	管理数据就是管理数据的质量
	需要元数据来管理数据
	管理数据需要计划
	数据管理要求必须推动信息技术决策
数据管理依赖于多种技能	数据管理应跨职能部门
	数据管理需要企业视角
	数据管理必须考虑一系列的视角
数据管理是生命周期管理	不同类型的数据具有不同的生命周期特征
	管理数据包括管理与数据相关的风险

（1）有效的数据管理需要领导者的承诺及参与。数据管理涉及一组复杂的过程，为了有效，需要协调、协作和承诺。要做到这一点，不仅需要管理技能，还需要坚定的领导能力带来的愿景和目标。

（2）数据管理需求符合业务需求。数据管理需要平衡战略和运营需求，要达到这种平衡，应重视：

- 数据是一种具有独特属性的资产。数据是一种资产，但它在影响其管理方式的重要方面不同于其他资产。这些属性中最明显的是数据在该使用时不被使用，金融资产和实物资产也是如此。
- 数据的价值可以而且应该用经济术语来表示。将数据称为资产意味着它有价值。虽然有一些技术可以衡量数据的定性和定量价值，但目前还没有这样做的标准。希望对数据做出更好决策的组织应该开发一致的方法来量化该价值。它们还应该衡量低质量数据的成本和高质量数据的好处。
- 管理数据意味着管理数据的质量。确保数据适合于目的是数据管理的主要目标。为了管理质量，组织必须确保它们理解涉众对质量的需求，并根据这些需求度量数据。
- 管理数据需要使用元数据。管理任何资产都需要拥有关于该资产的数据（员工数量、会计代码等）。用于管理和使用数据的数据称为元数据。由于数据无法保存或触摸，要理解它是什么以及如何使用它，需要元数据形式的定义和知识。元数据来源于与数据创建、处理和使用相关的一系列流程，包括体系结构、建模、管理、治理、数据质量管理、系统开发、IT 和业务运营以及分析。

- 管理数据需要计划。即使是小型组织也可能有复杂的技术和业务流程。数据在许多地方创建,并在不同地方之间移动以供使用。为了协调工作并保持最终结果的一致,需要从架构和过程的角度进行规划。

(3) 数据管理依赖于多种技能,包括:

- 数据管理应跨职能部门,它需要一系列的技能和专业知识。一个团队不能管理一个组织的所有数据。数据管理需要技术和非技术技能以及协作能力。
- 数据管理需要企业视角。数据管理具有本地应用程序,但必须在整个企业中应用,以便尽可能有效。这是数据管理和数据治理交织在一起的原因之一。
- 数据管理必须考虑一系列的角度:数据是流动的,数据管理必须不断发展,以跟上数据的创建和使用方式以及使用数据的数据使用者。

(4) 数据管理是生命周期管理。数据有生命周期,管理数据需要管理数据的生命周期。因为数据会产生更多的数据,所以数据生命周期本身可能非常复杂。

- 不同类型的数据有不同的生命周期特征,因此,它们有不同的管理需求。数据管理实践必须认识到这些差异,并足够灵活,以满足不同类型的数据生命周期需求。
- 管理数据包括管理与数据相关的风险。除了作为一种资产,数据还代表着组织的风险。数据可能会丢失、被盗或被滥用。组织必须考虑使用数据的伦理影响。与数据相关的风险必须作为数据生命周期的一部分进行管理。
- 数据管理需求必须驱动信息技术决策。数据和数据管理与信息技术和信息技术管理深深交织在一起。管理数据需要一种方法,确保技术服务于组织的战略数据需求,而不是驱动。

数据管理成熟度(Data Management Maturity,DMM)模型是最为典型的数据能力评价方法。该模型由 CMMI[7] 研究所于 2014 年推出,其设计沿用了能力成熟度模型集成(Capability Maturity Model Integration,CMMI)的基本原则、结构和证明方法。DMM 模型将机构数据管理能力定义为五个不同的成熟度等级,并给出了机构数据管理工作抽象成六类关键过程域共 25 个关键活动,如图 5-22 所示。

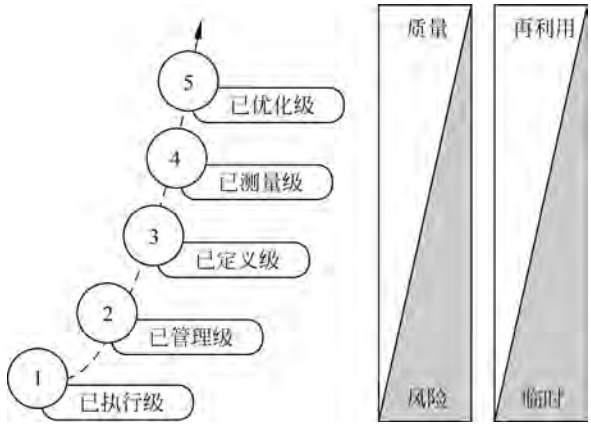

图 5-22 DMM 模型基本思路

CMM

CMM 是在“软件工业浪潮”和“软件过程运动”的背景下，由美国国防部（DoD）资助卡内基-梅隆大学（Carnegie Mellon University，CMU）的软件工程研究所（Software Engineering Institute，SEI）的 Watts Humphrey 等专家进行软件过程计划研究的代表性成果之一，主要用于软件质量评价，其基本思想如图 5-23 所示。

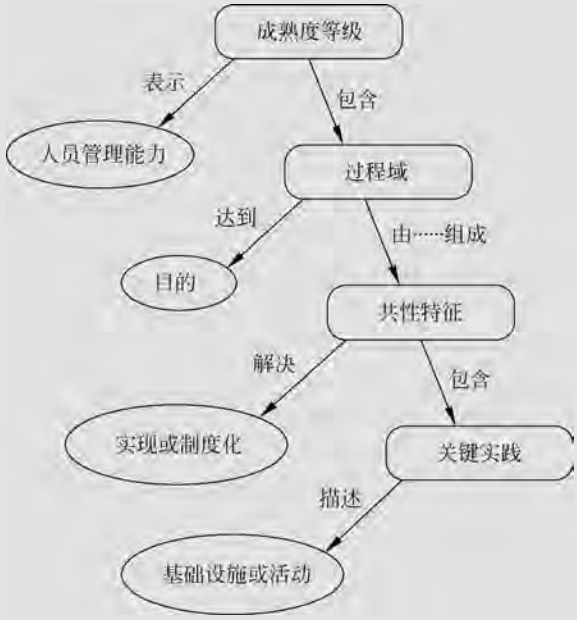

图 5-23　CMM 基本思想

CMM 的发展历程如下。

- CMM 的提出目的是应对 20 世纪 70 年代左右出现的“软件危机”。在此，所谓“软件危机”并不是“入侵威胁”或“病毒威胁”，而是由于当时的软件开发项目过于重视“结果”的好坏而忽略了“过程”的规范性，导致了软件的维护成本过高。软件危机之后，人们开始思考如何评价软件质量的问题——软件开发的“结果”重要还是“过程”重要？
- CMM 的重要贡献在于“看到了软件开发过程的成熟度在保证软件质量的重要地位”，是“软件工业浪潮”和“软件过程运动”的标志性成果之一。CMM 的出现标志着软件质量的评价从“结果派”转向“过程派”。
- 随着 CMM 在软件开发领域的成功应用，CMM 在其他相关学科领域得到了推广应用，出现了一些面向特定领域的模型，如 SE-CMM、SW-CMM、SA-CMM 和 IPPD-CMM 等，但导致了另一个问题——“框架泥潭”，即不同领域的 CMM 的差异性太大且难以集成。

- 为了解决当时的CMM“框架泥潭”问题,SEI又提出了CMMI。显然,CMMI的主要目的是为所有CMM类模型建立共同框架。CMMI项目组的研究目标分长期和短期两种。短期目标是集成SW-CMM、SE-CMM和IPD-CMM三个具体过程改进模型,在此基础上提出CMMI的初步框架。目前,该目标已经实现,其标志是2000年发布的CMMI 1.0;长期目标是为更多学科加入CMMI的工作奠定基础,提供一种可自动扩展的框架。

CMMI的主要内容包括CMMI框架、CMMI部件、制度化、表示方法(阶段式表示、连续式表示)以及CMMI的使用等。可以采用成熟度等级、关键过程域(Key Process Area,KPA)、共性特征(Common Feature,CF)和关键实践(Key Practice,KP)四个关键概念刻画CMM的核心思想。

- 成熟度等级:将组织机构的软件开发能力划分为五个成熟度等级,如图5-24所示。CMM的五个等级反映了从混乱无序的软件生产到有规律的开发过程,再到标准化、可视化和不断完善的开发过程的阶梯式结构。
- 关键过程域:每一级成熟度(除第一级外)由若干个关键过程域构成,每个KPA描述软件开发过程的某一个方面应达到的目标所必需的关键实践。CMMI评估结果分为五个等级,共由18个关键过程域和316个关键实践。
- 共性特征(Common Feature,CF):定义了每个关键过程域中应完成和达到的基本特征。
- 关键实践:达到共性特征需要完成的具体实践。需要注意的是,关键实践只提出了软件过程必须达到的标准而并未限定如何实现这些标准。因此,组织机构可根据自身具体情况采用不同的过程和方法完成同一个过程等级。

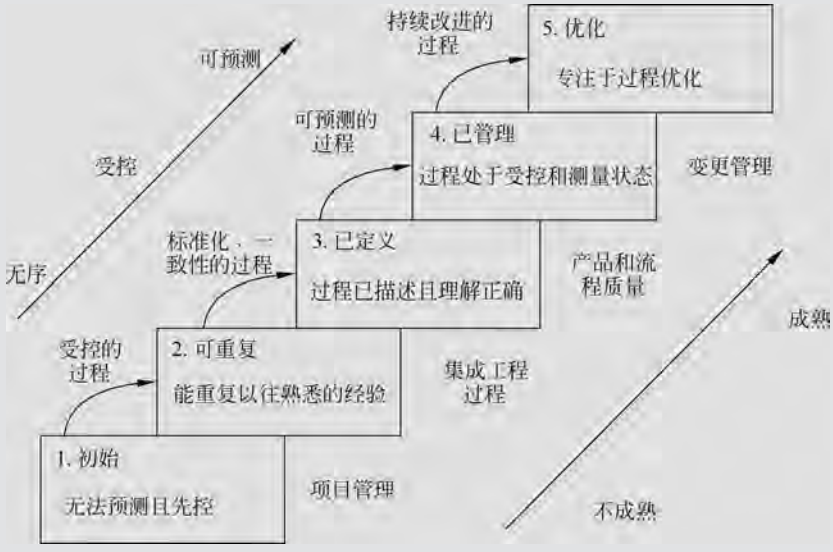

图5-24 CMM成熟度等级

CMMI 的实施应遵循以下指导原则：

- CMMI 组织要有代表性和广泛性。
- 使用系统工程过程。
- 保护业界已有的投资。
- 与 ISO 标准的兼容。
- 模型可剪裁性。
- 商业界参与。

1. 关键过程域

关键过程是一系列为达到某既定目标所需完成的实践，包括对应的工具、方法、资源和人。DMM 给出了组织机构数据管理所需的 25 个关键过程，并将其进一步聚类成六个关键过程域：数据战略(Data Strategy)、数据治理(Data Governance)、数据质量(Data Quality)、数据操作(Data Operation)、平台与架构(Platform & Architecture)和辅助性过程(Supporting Process)，如图 5-25 所示。

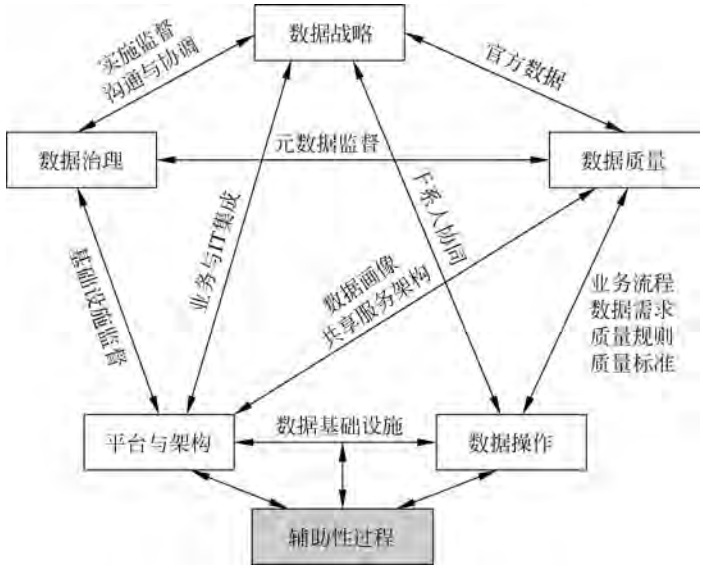

图 5-25　DMM 关键过程域

- **数据战略**。组织机构科学管理其数据资源的重要前提。数据管理工作需要在统一的顶层设计和战略规划的框架下进行，因此组织机构的数据管理往往以制定数据战略为起点。DMM 中的关键过程域“数据战略”包括五个关键过程：数据管理战略(Data Management Strategy)、有效沟通(Communication)、数据管理职责(Data Management Case)、业务案例(Business Case)和资金供给(Funding)。

- **数据治理**。确保数据战略顺利执行的必要手段。数据治理与数据管理的区别在于数据治理是“数据管理的管理”。DMM 中定义的关键过程域“数据治理”包括三个关键过程：治理管理(Governance Management)、业务术语表(Business Glossary)和元数据管理(Metadata Management)。
- **数据质量**。组织机构数据管理的主要关注点，要求数据管理中的输入数据和输出数据的质量必须达到当前业务需求与未来战略要求。DMM 中定义的关键过程域“数据质量”包括四个关键过程：数据质量战略(Data Quality Strategy)、数据画像(Data Profiling)、数据质量评估(Data Quality Assessment)、数据清洗(Data Cleaning)。
- **数据操作**。组织机构数据管理的具体表现形式，需要明确定义组织机构的数据操作规范，并予以监督和优化。DMM 中定义的关键过程域“数据操作”包括三个关键过程：数据需求定义(Data Requirement Definition)、数据生命周期管理(Data Lifecycle Management)、供方管理(Provider Management)。
- **平台与架构**。组织机构数据管理的必要条件，为数据战略的实现提供统一的架构设计和平台实现。DMM 中定义的关键过程域“平台与架构”包括五个关键过程：架构方法(Architectural Approach)、架构标准(Architectural Standard)、数据管理平台(Data Management Platform)、数据集成(Data Integration)以及历史数据、归档和保留(Historical Data，Archiving and Retention)。
- **辅助性过程**。虽不是数据管理的直接内容，但在组织机构数据管理工作，尤其是在其数据操作、平台和架构等关键过程域中扮演辅助性作用，具有不可或缺的地位。DMM 中定义的关键过程域“辅助性过程”包括五个关键过程：测量与分析(Measurement and Analysis)、过程管理(Process Management)、过程质量保证(Process Quality Assurance)、风险管理(Risk Management)和配置管理(Configuration Management)。

数据管理成熟度模型的过程域分类如表 5-4 所示。

表 5-4 数据管理成熟度模型的过程域分类

数据战略	数据治理	数据质量	数据操作	平台与架构	辅助性过程
• 数据管理战略 • 有效沟通 • 数据管理职责 • 业务案例 • 资金供给	• 治理管理 • 业务术语表 • 元数据管理	• 数据质量战略 • 数据画像 • 数据质量评估 • 数据清洗	• 数据需求定义 • 数据生命周期管理 • 供方管理	• 架构方法 • 架构标准 • 数据管理平台 • 数据集成 • 历史数据、归档和保留	• 测量与分析 • 过程管理 • 过程质量保证 • 风险管理 • 配置管理

2．成熟度等级

DMM 模型将组织机构的数据管理成熟度划分为五个等级，从低到高依次为：已执行级、已管理级、已定义级、已测量级、已优化级，并给出了每一层级的特点描述及其对数据重要性的基本认识，如图 5-26 所示。

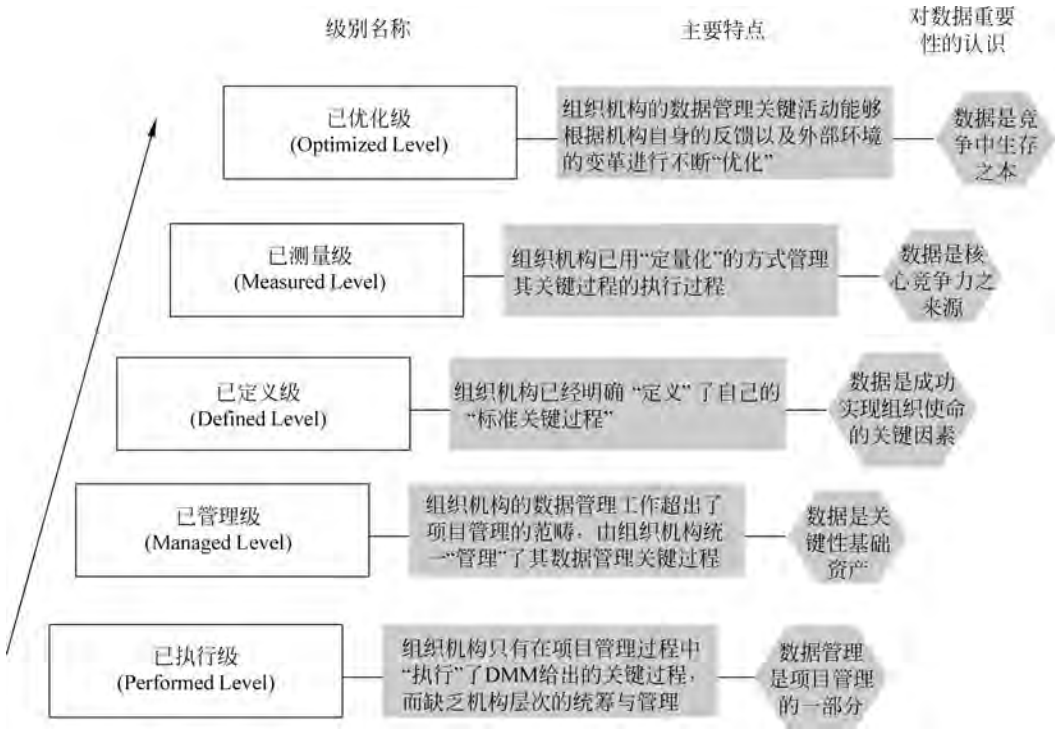

图 5-26　DMM 层级划分及描述

（1）**已执行级（Performed Level）**。组织机构只有个别项目的范围之内"执行"了 DMM 给出的关键过程，但缺乏机构层次的统筹与管理。其主要特点如下。

- 在具体项目中，DMM 关键过程域中给出的关键过程已被执行，但随意性和临时性较大。
- DMM 关键过程的执行往往仅限于特定业务范畴，很少存在跨越不同业务领域的关键过程。
- 缺少针对 DMM 关键过程的反馈与优化。以 DMM 关键过程中的"数据质量"为例，其数据管理工作可能过于集中在一个特定业务，如"数据修复活动"，并没有扩散到整个的业务范围或并没有开展对数据修复活动本身的反馈与优化工作。
- 虽然有可能在特定业务过程中进行了基础性改进，但没有进行持续跟进，也未拓展到整个组织机构。
- 组织机构没有统筹其数据管理工作，而数据管理活动局限在具体项目中，主要按照

其具体项目的实施需求进行,如果一个具体项目中需要进行数据管理,可能执行DMM中给出的相关过程,反之亦然。

(2) **已管理级(Managed Level)**。组织机构的数据管理工作超出了项目管理的范畴,由组织机构统一"管理"了其数据管理关键过程。其主要特点如下。

- 关键过程的定义与执行符合组织机构数据战略的要求。
- 组织机构聘请了数据管理相关的专业人士,员工的数据利用与数据生产行为有效。
- 关键过程已拓展至相关干系人。
- 对关键过程进行监督、控制和评估。
- 关键过程的评估依据为该DMM中对过程的具体描述。
- 组织机构已经意识到数据的重要性——数据是关键性基础资产,并开始对其实施"管理",但其管理往往并不规范。

(3) **已定义级(Defined Level)**。组织机构已经明确"定义"了自己的"标准关键过程"。其主要特点如下。

- 组织机构已明确给出了关键过程的"标准定义",并定期对其进行改进。
- 已提供了关键过程的测量与预测方法。
- 关键过程的执行过程并不是简单或死板地执行组织机构给出的"标准定义",而是根据具体业务进行了一定的"裁剪"工作。
- 数据的重要性已成为组织机构层次的共识,将数据当作成功实现组织机构使命的关键因素之一。

(4) **已测量级(Measured Level)**。组织机构已用"定量化"的方式管理其关键过程的执行过程。其主要特点如下。

- 已构建了关键过程矩阵。
- 已定义了变革管理的正式流程。
- 已实现用定量化方式计算关键过程的质量和效率。
- 关键过程的质量和效率的管理涉及整个生命周期。
- 数据被认为是组织机构核心竞争力的来源。

(5) **已优化级(Optimized Level)**。组织机构的数据管理关键活动能够根据组织机构自身的反馈以及外部环境的变革进行不断"优化"。其主要特点如下。

- 组织机构能够对其数据管理关键过程进行持续性拓展和创新。
- 充分利用各种反馈信息,推动关键过程的优化与业务成长。
- 与同行和整个产业共享最佳实践。
- 数据被认为是组织机构在不断变革的竞争市场环境中持续生存之本。

3. 成熟度评价

基于DMM模型的组织机构的数据管理能力成熟度水平的评价工作的实施可以借鉴

SEI 建议的 IDEAL 模型(见图 5-27)。

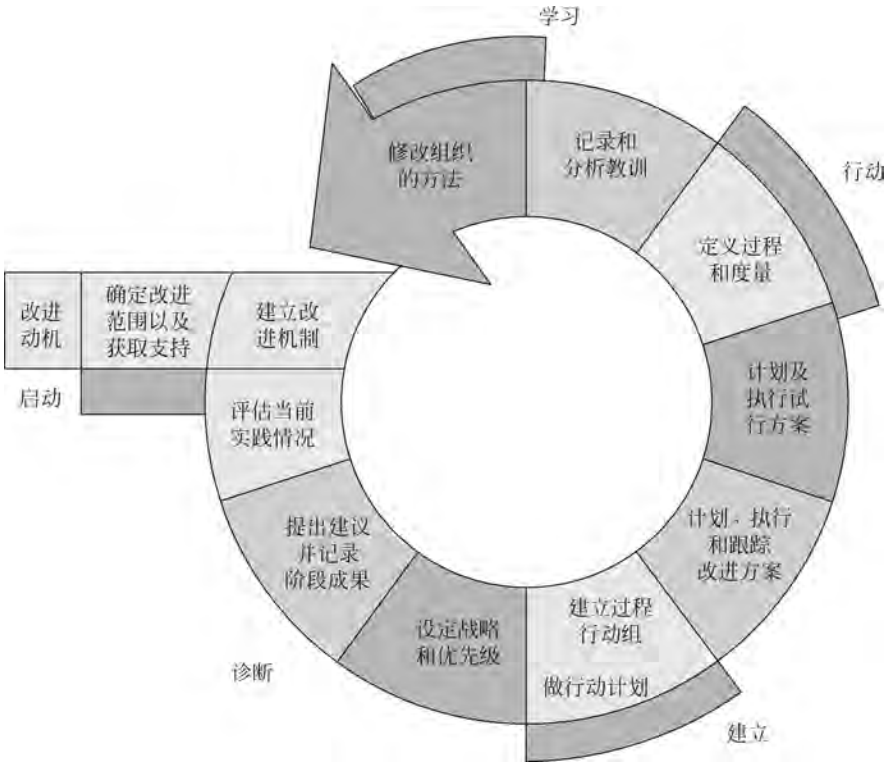

图 5-27　IDEAL 模型

- **启动(Initiating)**。组织机构应为 DMM 的引入做好准备工作,确定组织机构为数据管理目标所做的过程及其他内在联系。
- **诊断(Diagnosing)**。确定组织机构的数据管理过程成熟度等级。主要活动是确定组织机构的数据管理能力的当前和期望状态,并拟定建议稿。
- **建立(Establishing)**。构建实现改进目标的具体步骤。主要活动包括设定数据管理改进活动的优先级、开发方法和规划行动。
- **行动(Acting)**。实施上一阶段中设定计划的过程。主要活动包括创建和实现解决方案。
- **学习(Learning)**。改进数据管理能力的最后一个阶段,即分析数据管理过程改进中的经验教训,引入新的理论、方法和技术,进而增强自身的数据能力。

需要注意的是,能力成熟度评价的目的并不是给组织机构的数据管理现状进行“打分”,而是在于“如何帮助组织机构改进其数据能力”,因此,数据能力的成熟度评价过程是一个螺旋式推进的过程,需要进行多轮的“评估—改进—评估”的工作。另外,在数据能力的成熟度评估过程中,数据科学家应充分发挥“3C 精神”,综合运用数据科学的理念、理论、方法、技术、工具和最佳实践。例如,CMMI 采用雷达图的方式给出了组织机构数据管理能

力的成熟度评估结果,如图 5-28 所示。

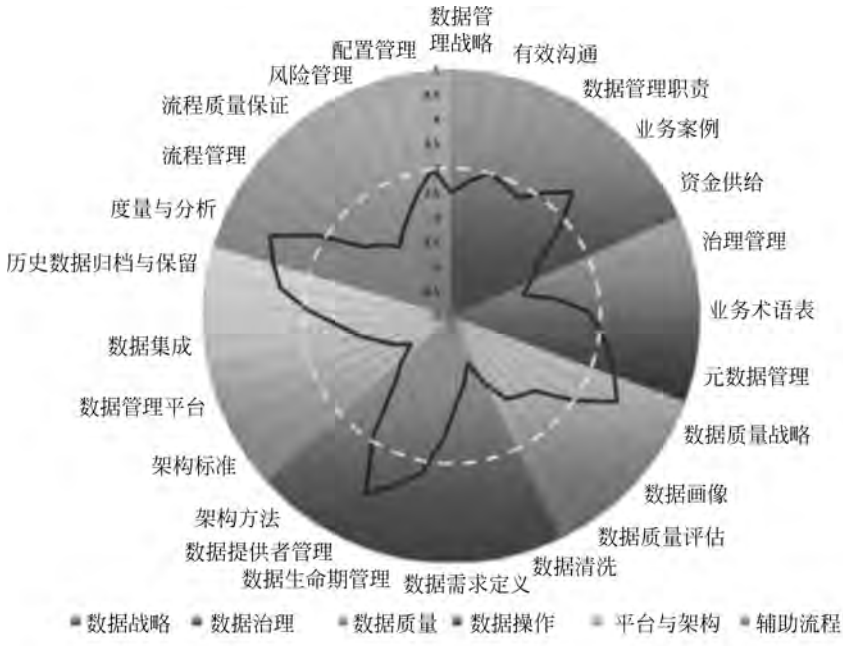

图 5-28 组织机构数据管理能力成熟度评估结果的可视化

国家标准《数据管理能力成熟度评估模型》(GB/T 36073.5—2018)是借鉴数据管理成熟度模型制定的国家标准,主要给出了数据管理能力成熟度模型及相应的成熟度等级,定义了数据战略、数据治理、数据架构、数据应用、数据安全、数据质量、数据标准和数据生存周期等八个能力域。

5.6 数据战略

数据战略是一个组织机构的数据管理的愿景、目标以及功能蓝图的统一管理。从 DMM 模型可以看出,数据战略是一个组织机构数据管理工作的重要前提。数据战略的制定需要注意以下基本问题。

美国国防部数据战略框架

美国国防部(DoD)的战略框架为 DoD 建立了与数据相关的远景、指导原则、基本能力、目标,图 5-29 显示了这些不同方面之间的关系。其中,愿景声明捕获数据的未来状态。DoD 将以指导原则为基础,以目标为重点,实现其愿景。基本功能跨越目标,枚举广泛的组织机构功能。

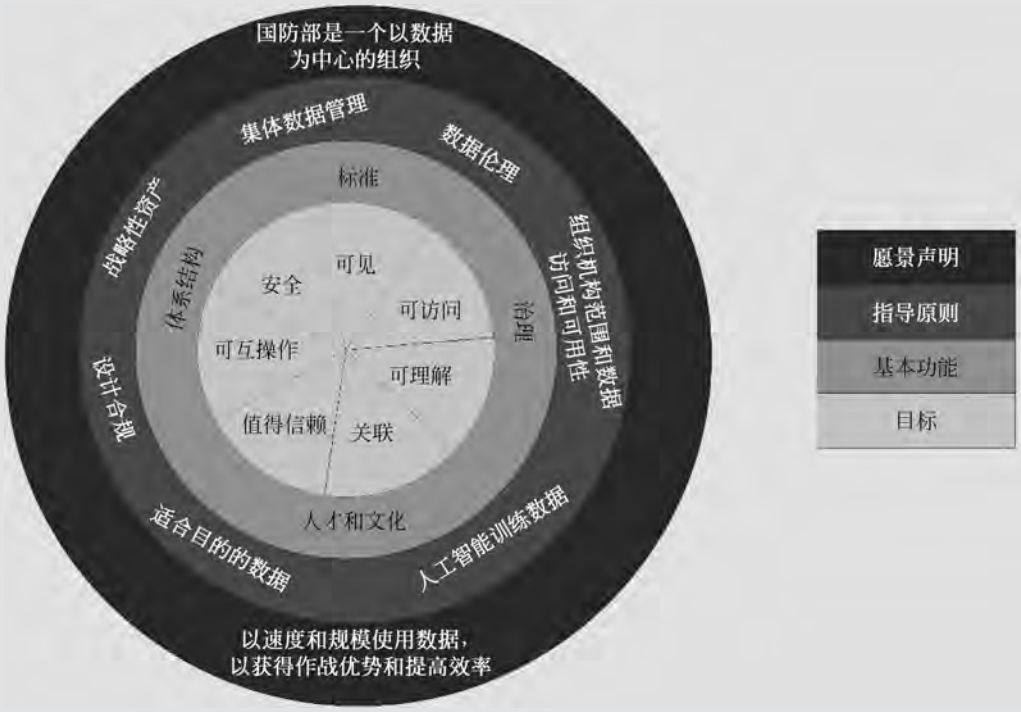

图 5-29　DoD 数据战略框架

其中，愿景声明为 DoD 是一个以数据为中心的组织，以速度和规模使用数据，以获得作战优势和提高效率。

DoD 利用八项指导原则来影响这一战略的愿景、目标和基本功能。这些指导原则是 DoD 所有数据工作的基础。具体的指导原则包括数据是战略性资产、集体数据管理、数据伦理、数据收集、组织机构范围的数据访问和可用性、人工智能训练数据、适合目的的数据、设计合规。

实现 DoD 数据目标需要四种基本能力——体系结构、标准、治理、人才和文化。这些能力并非特定于一个目标，但它们是实现所有目标所必需的。

DoD 数据战略的一个核心原则是理解数据不是 IT 资产，而是任务本身必不可少的组成部分。数据是无处不在的。DoD 武器平台、连接设备、传感器、训练设施、试验场和业务系统产生大量的数据，这些数据都保留并共享，以供更广泛的使用。数据的高质量、准确、完整、及时、受保护和值得信任是至关重要的。因此，该部通过制定以下目标，使数据成为一项战略资产。

- 可见——使用者可以定位所需的数据。
- 可访问——消费者可以检索数据。
- 可理解——消费者可以找到数据的描述来识别内容、上下文和适用性。

- 关联——消费者可以通过固有的关系开发互补的数据元素。
- 值得信赖——消费者可以对决策数据的各个方面充满信心。
- 可互操作——消费者和生产者对数据有共同的表示和理解。
- 安全——消费者知道数据是受保护的,不会被未经授权的使用和操纵。

1. 数据战略的定位

“数据战略”和“数据管理目标”是两个不同的概念。数据战略不仅需要定义数据管理的目标,更重要的是给出如何实现这些管理目标的具体行动方案以及如何动态调整数据管理目标的机制,如图 5-30 所示。

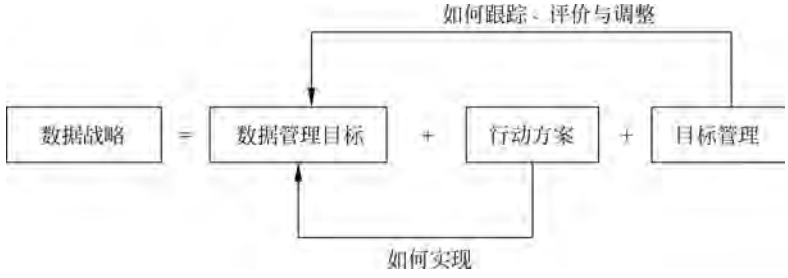

图 5-30 数据战略与数据管理目标的区别

2. 数据战略的目标

数据战略的根本目的是定义一个“数据驱动型组织”或培育“数据驱动型文化”,将数据作为组织机构决策活动的驱动因素,增强组织机构的敏捷性,进而提高组织机构的核心竞争力,如图 5-31 所示。

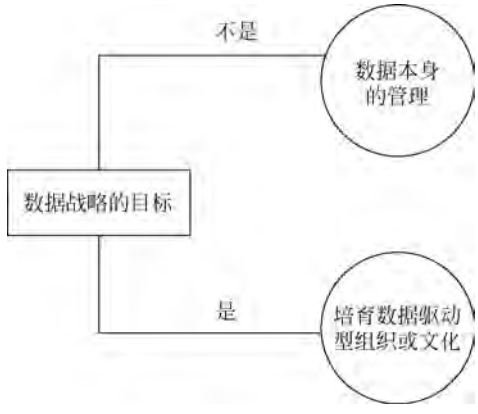

图 5-31 数据战略的目标

3. 数据战略的侧重点

数据战略应以解决数据密集型问题为主要关注点和责任，从数据视角分析组织机构业务活动中存在的瓶颈性问题，而不是过于强调计算密集型或人才密集型问题，如图 5-32 所示。

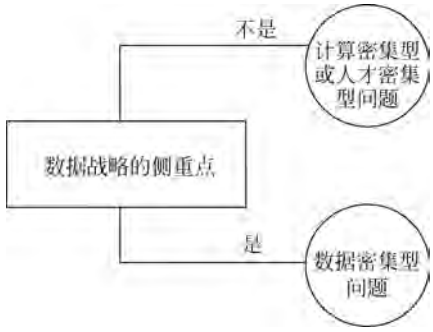

图 5-32　数据战略的侧重点

4. 数据战略的范畴

数据战略的制定不仅要考虑组织机构的当前业务需求，更重要的是综合考虑潜在风险与未来需求。数据的安全与质量风险是数据管理中的两个重要潜在风险，需要予以重视。另外，大数据的真正价值往往体现在未来，而组织机构的数据战略需要提前考虑企业未来需求的变化趋势，如图 5-33 所示。

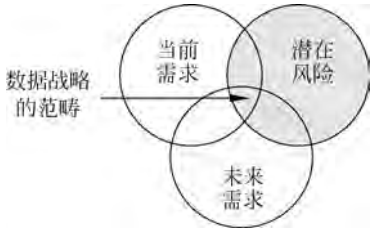

图 5-33　数据战略的范畴

数据战略可以针对国家、地区、机构、部门等不同层次制定。以国家或地区层次的数据战略为例，近年来很多国家或地区都纷纷制定其大数据相关的战略，如：

- **中国**。促进大数据发展行动纲要。
- **欧洲**。欧洲大数据价值战略研究与创新议程(European Big Data Value Strategic Research and Innovation Agenda，BDV SRIA)。
- **美国**。联邦大数据研究与发展计划(The Federal Big Data Research and Development Strategic Plan)。
- **英国**。英国数据能力战略(UK Data Capability Strategy)。

- **德国**。工业 4.0(Industrie 4.0)计划。
- **日本**。面向 2020 的 ICT 综合战略(2020 年頃に向けたICT 総合戦略)。

《促进大数据发展行动纲要》

- 发文字号：国发〔2015〕50 号。
- 发布日期：2015 年 9 月 5 日。
- 主要任务：加快政府数据开放共享，推动资源整合，提升治理能力。推动产业创新发展，培育新兴业态，助力经济转型。强化安全保障，提高管理水平，促进健康发展。
- 主要目标：立足我国国情和现实需要，推动大数据发展和应用在未来 5～10 年逐步实现。到 2020 年，形成一批具有国际竞争力的大数据处理、分析、可视化软件和硬件支撑平台等产品，培育 10 家国际领先的大数据核心龙头企业，500 家大数据应用、服务和产品制造企业。

5.7 数据治理

数据治理(Data Governance)可以理解为对数据管理的管理。从 DMM 模型可以看出，数据治理是实现数据战略的重要保障。需要注意的是，数据管理和数据治理是两个不同的概念，如图 5-34 所示。数据管理的是指通过管理“数据”实现组织机构的某种业务目的。然而，数据治理则指如何确保“数据管理”的顺利、科学、有效进行。

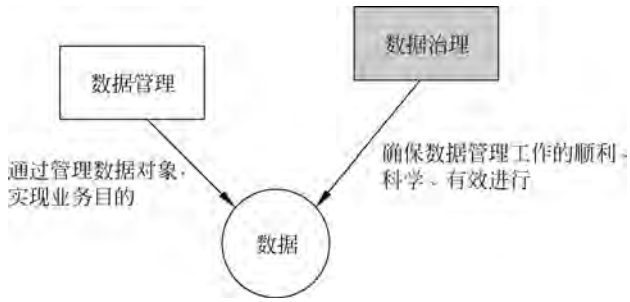

图 5-34 数据管理与数据治理的区别

1. 主要内容

数据治理工作涉及数据管理工作的每一个环节，是一项全员参与的常规性工作，主要工作重点如下。

- **理解自己的数据**。首先，需要理解组织机构自己的数据，并明确其特征、类型、趋势、风险及价值；其次，进行安全等级划分，定义组织机构的主数据管理。

IBM 提出的企业数据管理的范畴

图 5-35 是 IBM 提出的企业数据管理的范畴。从图中可以看出，企业数据主要包括以下四种类型。

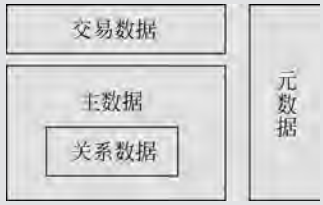

图 5-35　IBM 提出的企业数据管理的范畴

- **交易数据**。用于记录业务事件，如客户的订单、投诉记录、客服申请等，往往用于描述在某一个时间点上业务系统发生的行为。
- **主数据**。用于记录企业核心业务对象，如客户、产品、地址等。与交易流水信息不同，主数据一旦被记录到数据库中，需要经常对其进行维护，从而确保其时效性和准确性。主数据还包括关系数据，用以描述主数据之间的关系，如客户与产品的关系、产品与地域的关系、客户与客户的关系、产品与产品的关系等。
- **元数据**。用于记录数据的数据，用以描述数据类型、数据定义、约束、数据关系、数据所处的系统等信息。
- **关系数据**。用于描述主数据之间的关系，如客户与产品的关系、客户与客户的关系、产品与厂家的关系等。

- **数据干系人的识别与分析**。明确组织机构的数据管理中各干系人，包括数据的生产者、采集者、保管方、利用者及间接利益相关方。数据干系人的正确识别是数据治理的重要前提。
- **数据部门的设立**。需要设立专门的统一指挥部门，负责组织机构数据管理工作，并明确其职责，在不同数据干系人之间建立有效沟通渠道。
- **行为规范的制定**。需要针对组织机构的不同业务的特殊性，明确给出较为详细的数据管理规范，例如文档模板、数据词典、撰写文档要求等。主数据管理、商务智能、数据洞见是数据管理规范的重点内容。
- **数据管理方针和目标的确定**。数据治理工作应按照组织机构数据战略的要求，定期地制定和更新阶段性数据管理的方针与目标，确保组织数据管理的有效执行。
- **岗位职责的定义**。需要明确定义数据管理中的各参与方的岗位职责，预防各种潜在风险，并设立责任倒查机制和弥补措施。
- **应急预案与应急管理**。数据治理的重要组成部分之一，需要明确规定各种可能的紧

急事件及其具体应对方案。

- **等级保护与分类管理**。组织机构数据治理应对其数据、人员、技术、设备进行分类管理,并根据其安全和保密要求进行等级保护。
- **有效监督与动态优化**。组织机构数据工作必须建立有效监督机制,并根据监督中发现的问题与风险,不断优化其数据管理工作。

2. 基本过程

数据治理并不是一次性工作,而是一种循序渐进的过程,主要包含计划(Plan)、执行(Do)、检查(Check)和改进(Action)等基本活动,如图 5-36 所示。

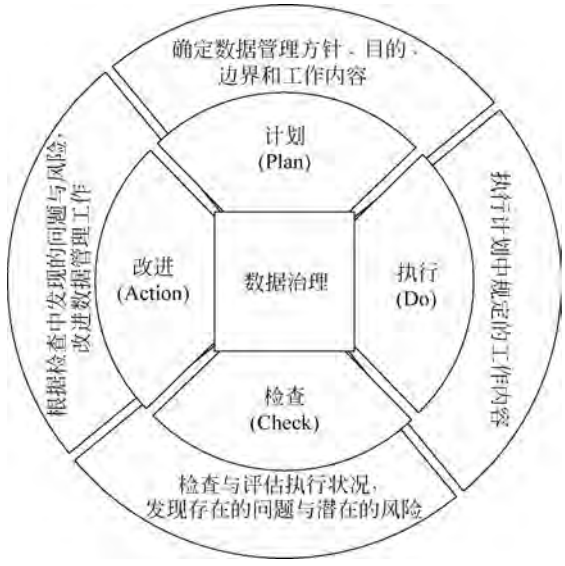

图 5-36 数据治理的 PDCA 模型

- **计划**。数据管理方针和目标的确定,明确组织机构的数据管理的目的、边界和工作内容。
- **执行**。根据数据管理计划,设计或选择具体的方法、技术、工具等解决方案,实现计划中的工作内容。
- **检查**。定期检查执行效果,进行绩效评估,并发现存在的问题与潜在的风险。
- **改进**。根据检查结果中发现的问题与风险,进一步改进自己的数据管理工作。

DGI 数据治理框架

DGI(The Data Governance Institute)成立于 2003 年,是世界上较早从事数据治理研究和实践方向,并且当今影响力较大的专业机构之一。该机构提出的数据治理框架(The DGI Data Governance Framework)在数据治理领域具有很大的影响。

DGI 认为数据治理是对数据相关的决策及数据使用权限控制的活动。它是一个信息处理过程中根据模型来执行的决策权和承担责任的系统，规定了谁、可以在什么情况下、对哪些信息做怎样的处理。图 5-37 给出了 DGI 数据治理框架。DGI 数据治理框架是用于分类、组织和传递复杂企业数据的逻辑框架。数据治理任务通常有如下三个部分。

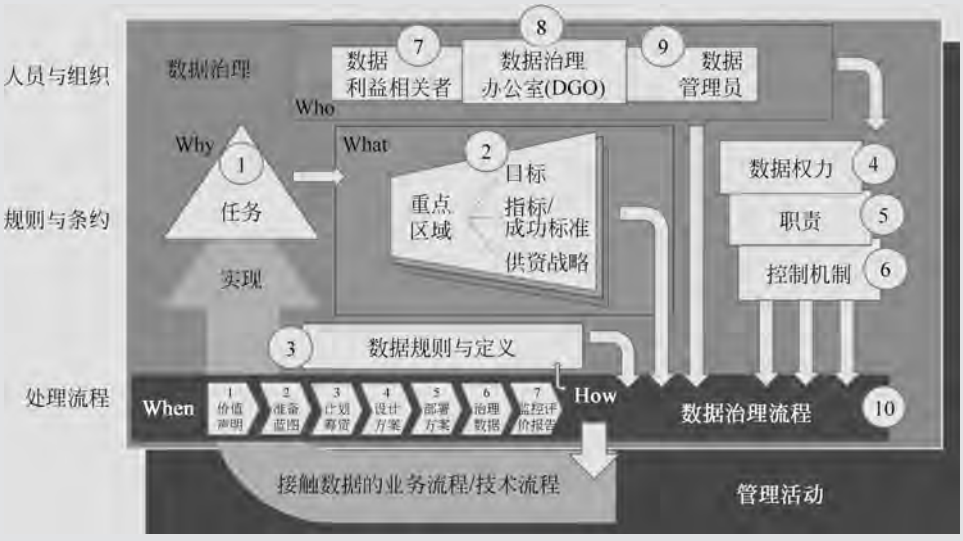

图 5-37　DGI 数据治理框架

- 主动定义或序化规则。
- 为数据利益相关者提供持续的，跨职能的保护和服务。
- 应对并解决因不遵守规则而产生的问题。

国家标准《信息技术服务 治理 第 5 部分：数据治理规范》(GB/T 34960.5—2018)是 GB/T 34960 系列标准的一部分。GB/T 34960 分为如下部分。

第 1 部分：通用要求。

第 2 部分：实施指南。

第 3 部分：绩效评估。

第 4 部分：审计导则。

第 5 部分：数据治理规范。

5.8　数据安全、隐私、道德与伦理

在数据产品开发中，不能忽视数据安全、隐私、道德与伦理问题，防止出现数据安全、数据偏见、算法歧视、数据攻击和隐私泄密。

1. 数据安全

目前,人们对大数据安全普遍存在两个曲解。一是数据安全只是技术问题。数据安全不仅是技术问题,而且还涉及管理问题。通常认为,数据安全事件中,70%来自管理上的漏洞,而30%才是来自技术上的缺陷。因此,管理是数据安全中不可忽略的重要问题,将数据安全放在组织机构的数据战略、数据治理和数据管理之中进行统一管理,应重视安全管理制度建设、安全机构设置、人员安全管理、系统建设管理和系统运维管理。二是数据安全的主要威胁是外部入侵。统计数据发现,70%左右的数据安全事件来自于内部人员,而30%左右是因为外部入侵造成的。例如,著名的斯诺登事件中斯诺登本人曾是一名美国中情局的职员,同时还曾负责美国国家安全局的一个秘密项目。因此,数据安全中不能忽略对内部人员的信息安全教育和管理,应提升其信息安全意识与能力。

需要注意的是,数据安全不等同于数据保密。通常,除了数据保密——数据的机密性(Confidentiality)之外,数据安全还包括完整性(Integrity)、可用性(Availability)、不可否认性(Non-repudiation)、鉴别(Authentication)、可审计性(Accountability)和可靠性(Reliability)等多个维度。在具体工作中,数据安全也并不是独立存在的,一般与其对应信息系统的安全密切相关。目前,信息系统的安全保护普遍采取等级保护策略,即针对不同的攻击来源和保护对象采取不同的应对策略。以国家标准《信息安全技术 信息系统安全等级保护基本要求》(GB/T 22239—2008)为例,其主要安全等级及保护基本要求如表5-5所示。

表5-5 信息系统安全等级及保护基本要求

等 级	攻击来源	保护对象	应对要求
第1级	个人的、拥有很少资源的威胁源发起的恶意攻击、一般的自然灾难	关键资源	在系统遭到损害后,能够恢复部分功能
第2级	外部小型组织的、拥有少量资源的威胁源发起的恶意攻击、一般的自然灾难	重要资源	能够发现重要的安全漏洞和安全事件;在系统遭到损害后,能够在一段时间内恢复部分功能
第3级	来自外部有组织的团体、拥有较为丰富资源的威胁源发起的恶意攻击、较严重的自然灾难	主要资源	能够发现安全漏洞和安全事件;在系统遭到损害后,能够较快恢复绝大部分功能
第4级	国家级别的、敌对组织的、拥有丰富资源的威胁源发起的恶意攻击、严重的自然灾难	全部资源	能够发现安全漏洞和安全事件;在系统遭到损害后,能够迅速恢复所有功能

P^2DR 模型

大数据很难做到(或不存在)无条件的绝对安全,人们追求的是有条件的相对安全,数据安全保障是数据的保护者和攻击者之间的一个动态博弈过程。当攻击(或入侵)的代价超出数据本身的价值或攻击(或入侵)所需要的时间超出数据的有效期时,入侵者一般不会采取攻击或入侵。

P^2DR 模型是美国 ISS 公司提出的一种动态网络安全体系,其认为网络安全是一种动态的、有条件的相对安全。P^2DR 模型包括四个主要部分:Policy(策略)、Protection(防护)、Detection(检测)和 Response(响应),如图 5-38 所示。其中,策略处于核心地位,为其他三个组成部分提供支持和指导,而保护、检测和响应为网络安全的三个基本活动。从相对安全角度看,P^2DR 模型可以用以下公式表示。

图 5-38　P^2DR 模型

(1) 当入侵所需时间大于 0,即 $P_t>0$ 时,

$$P_t>D_t+R_t$$

(2) 当入侵所需时间等于 0,即 $P_t=0$ 时,

$$E_t=D_t+R_t$$

其中,E_t 为数据的暴露时间。

2. 数据偏见

在数据科学项目中,避免出现 BIBO(Bias In,Bias Out, 偏见进则偏见出)现象的出现。数据偏见(Data Bias)的成因可能是有意的,也可能是无意的,但均会造成数据科学项目的失败。数据偏见可能出现在数据科学流程的任何一个活动之中,常见的数据偏见有:

- 数据来源选择偏见。有的数据工作者偏向于仅仅选择自己喜欢或熟悉的、对自己有利的数据来源进行数据化和数据分析工作,导致数据科学项目失败于其起点。在数据来源的选择上,如果不做预调研和试验研究,仅仅用自己的常识或直觉选择数据来源时,经常会出现此类偏见,比较著名的是幸存者偏见(Survivorship Bias)。

幸存者偏见

幸存者偏见指的是人往往会注意到某种经过筛选之后所产生的结果,同时忽略了这个筛选的过程,而被忽略的过程往往包含着关键性的信息。

1940 年左右,在英国和德国之间的空战中,双方都失去了很多轰炸机和飞行员。因此,当时英国军事部门研究的一个主要话题是:在飞机的哪一部分加厚装甲可以提高飞

机的防御能力并减少损失。当时的技术还不是很成熟,如果加厚一部分装甲,势必减少其他部分的装甲,否则就会影响飞行的平稳度。因此,研究人员需要做出选择,为飞机最脆弱的区域增加装甲。

当时的英国军方研究了从欧洲大陆的空战中返回的轰炸机。如图 5-39 所示,飞机上的弹孔主要集中在机身中央和机翼。因此研究人员提出,在这些部位添加装甲,以提高飞机的防御能力。

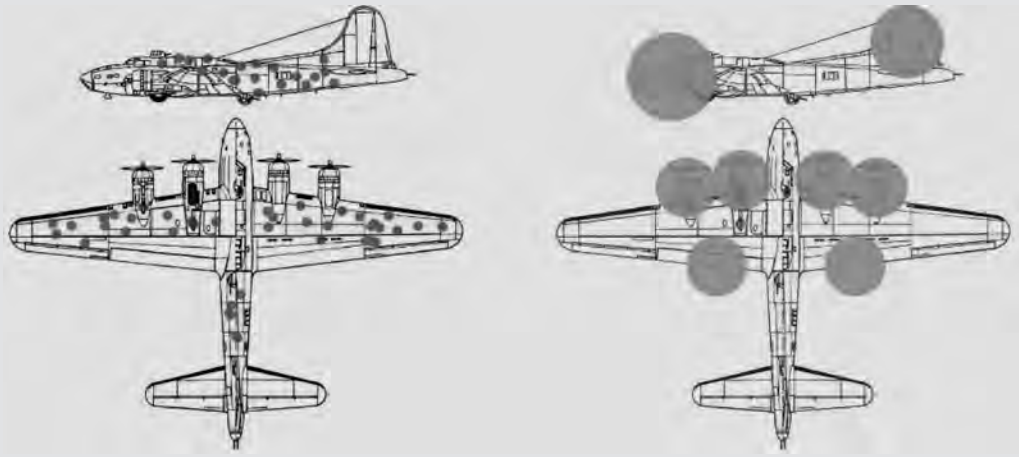

图 5-39 从欧洲大陆的空战中返回的轰炸机

注:左图的点代表的是机身上的弹孔,右图的圆圈代表的是"无弹孔区域"。

而统计学家沃德认为,应当加厚座舱和机尾,减弱机翼装甲。他提出,能够根据返航的飞机统计出机翼的损伤,这正说明机翼的受损对飞机的飞行并不致命。而大部分坠毁的轰炸机应当是座舱和机尾受到了严重损伤。想要减少坠毁率,必须加厚座舱和机尾的装甲。

由于战况紧急,空军部长决定接受沃德的建议,立即加厚座舱和机尾的装甲。不久之后,英国轰炸机的坠毁率显著下降。[①]

- 数据加工和准备偏见。在数据加工和准备过程中,有的数据工作者偏向于将数据加工成对自己的观点(或研究结论、研究假设)有利,过滤掉那些与自己的观点不一致的数据,表面上看在用数据证明自己的观点,实际上在找对自己观点有利的片段数据。

① http://bazyd.com/talk-about-survivorship-bias/。

伯克森悖论

伯克森悖论(Burkson's Paradox)是条件概率和统计的结果,即两个本来无关的变量之间体现出貌似强烈的相关关系。

该悖论由美国医生和统计学家约瑟夫·伯克森在1946年提出,以a和b两个事件为例,它们是完全独立的事件(例如肺癌和糖尿病)。如果一项研究同时检测a(肺癌)和b(糖尿病),那么糖尿病的存在将增加肺癌的发生概率,两个本来无关的变量之间体现出貌似强烈的相关关系,即伯克森悖论。

造成伯克森悖论的最主要的原因是统计偏差。例如事件a发生的概率在事件b发生的情况下更高概率出现的原因是将两者都没有发生的情况排除在外。如果对全体人群进行统计,就会发现a(肺癌)和b(糖尿病)之间并没有相关性。但是如果只对医院中的患者进行统计,就会出现问题。

- 算法和模型选择偏见。在数据分析中,有的数据工作者偏向于直接套用自己常用的、自己已知的算法和模型,而不是根据数据本身的特点选择和论证算法/模型的信度和效度。算法和模型选择偏见的存在使得数据工作者不去学习新的算法和模型,习惯于套用自己擅长的算法、模型,导致"以不变应万变"所带来的盲目性。

A/B 测试

A/B测试起源于Web测试,是为Web、App界面或流程制作两个(A/B)或多个(A/B/n)版本,在同一时间维度,分别让属性或组成成分相同(相似)的两个或多个访客群组(目标人群)访问,收集各群组的用户体验数据和业务数据,最后分析、评估出最好的版本,将其正式采用。

A/B测试是一种对比试验,准确地说是一种分离式组间试验,在试验过程中,从总体中随机抽取一些样本进行数据统计,进而得出对总体参数的多个评估。从统计学视角看,A/B测试是假设检验(显著性检验)的一种应用形式。在进行A/B测试时,首先需将问题形成一个假设,然后制定随机化策略、样本量以及测量方法。

A/B测试有效避免了数据加工和准备偏见以及算法/模型选择偏见,具有重要借鉴意义。例如,The Guardian(卫报)的约会网站Soulmates通过每月付费订阅实现盈利。产品经理Kerstin Exner通过A/B测试来优化Soulmates的关键绩效指标。Kerstin Exner注意到大多数登录到Soulmates入口页面的访客并没有转化为订阅者。基于研究她提出假设:提前展示更多现有用户的信息将增加订阅量。她做了A/B测试来验证这一点,测试包括一个添加了类似的个人资料、搜索功能和客户评价的变体登录页面,获胜的版本将订阅转化率提高了46%以上。

- 分析结果的解读和呈现上的偏见。在解读数据科学项目的最终结果时,数据工作者需要避免各种偏见的出现,如过拟合或欠拟合现象的出现、根据自己的爱好(而不是目标用户的爱好)进行数据可视化、根据自己的主观偏见(而不是忠于数据本身)进行数据解读与呈现,以及根据自己想要的结论修改数据或数据分析过程等。

辛普森悖论

辛普森悖论(Simpson's Paradox)是概率和统计学中的一种现象,即几组不同的数据中均存在一种趋势,但当这几组数据组合在一起后,这种趋势消失或反转。例如,在肾结石治疗数据分析中,比较了两种肾结石治疗的成功率。其中方案 A 包括所有开放式外科手术,方案 B 仅涉及小的穿刺。小肾结石和大肾结石的治疗的成功率和治疗案例数如表 5-6 所示(括号中的数字表示:成功案例数/治疗总案例数)。

表 5-6 肾结石治疗数据分析——两种治疗方案的分别统计

结石大小	治疗方案	
	方 案 A	方 案 B
小	93%(81/87)	87%(234/270)
大	73%(192/263)	69%(55/80)

从表 5-6 中可以发现治疗方案 A 的成功率更高,那是否就应该选择方案 A 呢?把两种治疗方案进行总计(见表 5-7),却发现方案 B 的成功率更高。

表 5-7 两种治疗方案的汇总统计

治疗方案	方 案 A	方 案 B
总 计	78%(273/350)	83%(289/350)

当数据中存在多个单独分布的隐藏变量,不当拆分时就会造成辛普森悖论。这种隐藏变量被称为潜伏变量,并且它们通常难以识别。而这种潜伏变量可能是由于采样错误或者数据领域本身属性造成的。如本例中,可能是采样方法存在误差导致加权结果出现问题,不同大小的结石中对于不同方法的应用数量有较大的差异,没有做到正确地控制变量。

3. 算法歧视

算法歧视是指算法设计、实现和投入使用过程中出现的各种“歧视”现象。根据 Reuters 的报道,某公司曾于 2014 年开发了一套“算法筛选系统”,用来自动筛选简历,开发小组开发出了 500 个模型,同时教算法识别 50 000 个曾经在简历中出现的术语让算法学习

在不同能力分配的权重。但是久而久之，开发团队发现算法对男性应聘者有着明显的偏好，当算法识别出“女性”(women and women's)相关词汇的时候，便会给简历相对较低的分数，如女子足球俱乐部等；算法甚至会直接给来自于两所女校的学生降级。

大数据杀熟

同样的商品或服务，老客户看到的价格反而比新客户要贵出许多，这在互联网行业被叫作“大数据杀熟”。调查发现，在机票、酒店、电影、电商、旅游等多个价格有波动的网络平台都存在类似情况，而在线旅游平台更为普遍。同时，还存在同一位用户在不同网站的数据被共享的问题，许多人遇到过在一个网站搜索或浏览的内容立刻被另一网站进行广告推荐的情况。

“大数据杀熟”是一个新近才热起来的词，不过这一现象或已持续多年。有数据显示，国外一些网站早就有之，而近日有媒体对2008名受访者进行的一项调查显示，51.3%的受访者遇到过互联网企业利用“大数据杀熟”的情况。

和任何新事物都会存在不同看法一样，“大数据杀熟”到底该如何定性，目前也面临争议。如上述调查中，59.2%的受访者认为在大数据面前，信息严重不对称，消费者处于弱势；59.1%的受访者希望价格主管部门进一步立法规范互联网企业歧视性定价行为。另外，也有专家表示，这一价格机制较为普遍，针对大数据下价格敏感人群，系统会自动提供更加优惠的策略，可以算作接受动态定价。

倘若搁置具体应如何定性的争议，“大数据杀熟”所表现出来的现象和逻辑还是存在相当大的问题。

“大数据杀熟”虽然可以说是商家的定价策略，但最终形成了“最懂你的人伤你最深”的局面，确实与人们习以为常的生活经验和固有的商业伦理形成了明显冲突。例如，一些线上商家和网站标明新客户享有专属优惠，从吸引新客户的角度完全可以理解，但在这一优惠政策的另一端，若老客户普遍要支付高于正常价格的金额，甚至越是老客户价格越高，就明显背离了朴素的诚信原则，也是对老客户信赖的一种辜负。由此还会引发商业伦理的扭曲，值得人们警惕。

有专家表示，与其称这种现象为“杀熟”，不如说是“杀对价格不敏感的人”：一罐可乐，在超市只卖2元，在五星级酒店能卖30元——这不能叫价格歧视，而是因为你能住得起五星级酒店，那么你就是要被“杀”，这样的例子在现实中比比皆是。但是，这个理论套用在“大数据杀熟”上却并不恰当。一个关键问题是，一罐可乐的正常价格是透明的，所以在五星级酒店的溢价是公开的。但“大数据杀熟”却处于隐蔽状态，多数消费者是在不知情的情况下“被溢价”了。此外，将老顾客等同于“对价格不敏感的人”，也有偷换概念之嫌。

（来源：光明日报）

4. 数据攻击

最有代表性的数据攻击为“谷歌炸弹”(Google Bomb)。谷歌炸弹[①]是指人为恶意构造锚文本,在搜索引擎中提升有关他人不利报道的文章或网页的点击率,即便这些文章或网站与搜索主题可能并不相关。谷歌炸弹大部分出于商业、政治或恶作剧等目的。其实现是基于搜索引擎排名算法中的两个事实:①外部链接是排名的重要因素之一。②链接文字很多时候比链接数量更重要。因此,当有大量包含特定关键词的链接指向某一个网页的时候,即使这个网页没提到这个关键词,排名也会非常靠前。需要注意的是,谷歌炸弹并非谷歌公司操控所为,而是人们利用谷歌算法漏洞产生的现象。

数据攻击及谷歌炸弹

2018 年 12 月 11 日,在美国国会听证会上,民主党国会议员 Zoe Lofgren 就“在谷歌图片上搜索 idiot(白痴)会出现某著名政治家的照片”一事,质问了时任谷歌公司 CEO 桑达尔·皮查伊(Sundar Pichai)——为何搜索 idiot 会出现特朗普总统的图片?谷歌搜索到底是如何运作的?桑达尔·皮查伊回答说:“每当您输入关键字,谷歌就会在其索引中抓取并存储几十亿个“网站”页面的副本。我们将关键字与其页面进行匹配,然后根据 200 多个因素对结果进行排名,如相关性、新鲜度、流行度、其他人如何使用它等。基于此,在任何给定时间内,我们尝试为该查询排序并找到最佳搜索结果。然后我们用外部评估员评估它们,他们根据客观指导进行评定。这就是我们确保(搜索)这个过程有效的方法。”Zoe Lofgren 讽刺地问道:“所以,不是你们有一些小人躲在窗帘后面操控要向用户展示什么吗?”。对此,桑达尔·皮查伊回答道:“这是大规模的运作,我们不会手动干预任何特定的搜索结果”。

5. 隐私保护

随着大数据时代的到来,隐私保护成为热门话题,得到社会各界的广泛关注。在数据科学项目中,需要注意保护用户隐私。隐私保护需要遵循相关的法律法规和伦理道德的要求。

剑桥分析公司数据丑闻

2013 年剑桥大学的研究员 Aleksandr Kogan 创建了一款名为 *This is Your Digital Life* 的应用,付费吸引 Facebook 用户做心理测试,它不仅可以收集参加测试的用户的数据,还可以在用户好友不知情的情况下获取他们的数据,然后把多达 8700 万用户的数据

① https://en.wikipedia.org/wiki/Google_bomb。

卖给了剑桥分析公司。2015 年，Facebook 曾要求剑桥分析公司删除上述数据，但 Facebook 接到的其他报告表明，这些被滥用的用户数据并未被销毁。2016 年总统大选，剑桥分析公司利用这些数据协助特朗普竞选。2018 年，剑桥分析公司的前员工 Christopher Wylie 公布了一系列文件，揭露了剑桥分析公司的数据丑闻。2018 年 5 月 2 日，剑桥分析公司正式关闭其运营业务并宣布破产。2018 年 3 月 19 日，Facebook 股价大跌 7%，市值蒸发 360 多亿美元。“卸载 Facebook”运动得到了许多网友的支持。之后，各大媒体对本事件的启示进行了如下报道。

- 卫报：用户数据，尤其是 Facebook 个人资料形式的数据，对黑客和营销人员等来说，一直是个诱人的目标。而用户（或他们的朋友）没有意识到这一点，错误地允许 *This is Your Digital Life* 应用程序获取了他们的数据。Facebook 对用户个人数据的不对称控制一直延伸到第三方应用程序。它们被允许从用户那里“窃取”Facebook 的个人资料，这些用户通常是被引诱去玩游戏或测试的，他们不仅同意交出自己的数据，还同意交出朋友的数据，而大多数人都不知道自己的数据被窃取了，从而为剑桥分析公司搜集用户数据并为特朗普竞选锁定选民提供了可乘之机。
- 纽约时报：大量的用户数据对 Facebook 的广告客户及其用户很有价值，使其能够只提供与用户相关的广告，从而搞清并操纵用户的情绪状态。
- 时代周刊：在这个连吃哪家店的东西、和谁保持联系、要去哪里都会告诉 Facebook、谷歌、亚马逊之类的公司的时代，用户自己也应该对平台提出要求，以可读并且可获取的方式了解他们的信息发给了谁，那些人会怎么用他们的信息。用户自己也应该在分享个人信息的时候更加谨慎。
- 哈佛商业评论：2018 年 5 月，世界上最严格的隐私法——欧盟（EU）出台的《一般数据保护条例》（General Data Protection Regulation）开始生效。到 2018 年年底，苹果公司和微软公司的首席执行官也呼吁在美国制定新的国家隐私标准。尤其是美国、中国等，作为数字大国，需要尽快对大数据的使用提供法律约束，对公民的个人隐私信息予以法律保障。

《中华人民共和国个人信息保护法》的目录

（《中华人民共和国个人信息保护法》于 2021 年 8 月 20 日第十三届全国人民代表大会常务委员会第三十次会议通过）

《中华人民共和国数据安全法》的目录

(《中华人民共和国数据安全法》于2021年6月10日第十三届全国人民代表大会常务委员会第二十九次会议通过)

如何继续学习

【学好本章的重要意义】

数据产品开发是数据科学家的核心竞争力之源,也是数据科学中独有的知识内容。因此,学好数据产品开发相关的知识是数据科学中不可忽略的核心内容之一。

【继续学习方法】

数据产品开发不仅涉及理论学习与实践经验,更重要的是数据科学家的3C精神的培

养(详见“1.6 基本原则”)。因此,建议在后续学习中不仅要重视基础理论和最佳实践的跟踪,而且也应重视与领域高端人才的合作与沟通,如参加开源项目、活跃于各大专业社区等。

【提醒及注意事项】

正确理解数据产品的本质特征是学习好本章知识的关键所在;培养自己的数据科学家精神与素质是我们继续学习本章的首要任务。

【与其他章节的关系】

本章是“第1章基础理论”的进一步深入讲解,系统讲解了数据产品开发的方法与内容。“第6章典型案例及实践”是本章的拓展,建议结合典型实践理解数据产品开发的知识。

习题

1. 结合自己的专业领域或研究兴趣,调研自己所属领域的数据产品开发方法、技术与工具。

2. 分析 DMM 与 DAMA-DMBOK(DAMA Guide to the Data Management Body of Knowledge)的区别和联系。

3. 调研常用数据产品开发工具软件(包括开源系统),并进行对比分析。

4. 阅读本章所列出的参考文献,并采用数据产品开发或故事化描述方式展示该领域的代表性文献数据。

参考文献

[1] ANDERSON C. Creating a data-drivenorganization [M]. Sebastopol: O'Reilly Media, Inc., 2015.

[2] CMMI. Data Management Maturity (DMM). http://cmmiinstitute.com/data-management-maturity.

[3] HURWITZ J, NUGENT A, HALPER F, et al. Big data for dummies[M]. Hoboken: John Wiley & Sons, 2013.

[4] KHATRI V, BROWN C V. Designing data governance[J]. Communications of the ACM, 2010, 53(1): 148-152.

[5] KNAFLIC C N. Storytelling with data: a data visualization guide for business professionals[M]. Hoboken: John Wiley & Sons, 2015.

[6] MARZ N, WARREN J. Big data: principles and best practices of scalable realtime data systems[M]. New York: Manning Publications Co., 2015.

[7] MAYER-SCHÖNBERGER V, CUKIER K. Big data: a revolution that will transform how we live, work, and think[M]. Boston: Houghton Mifflin Harcourt, 2013.

[8] PATIL D J. Data Jujitsu[M]. Sebastopol: O'Reilly Media, Inc., 2012.

[9] PAULK M C, WEBER C V, CURTIS B, et al. The capability maturity model: guidelines for improving the software process principal contributors and editors[M]. Boston: Addison-Wesley Pub. Co., 1995.

[10] 朝乐门. 数据科学[M]. 北京: 清华大学出版社, 2016.

[11] CUPOLI P, EARLEY S, HENDERSON D. Dama-dmbok2 framework[J]. Dama International, 2014 (3): 1-27.

[12] 朝乐门, 王锐. 数据科学平台: 特征、技术及趋势[J]. 计算机科学, 2021, 48(8): 1-12.