狗熊会

数据思维

从数据分析到商业价值

王汉生◎编著

联袂推荐

上海纽约大学杰出全球商学讲席教授 陈宇新
复旦大学大数据学院创始院长 范剑青
考拉征信服务有限公司执行总裁 葛伟平
狗熊会CEO 李广雨
百分点集团董事长兼CEO 苏 萌
伦敦政治经济学院统计学讲座教授 姚琦伟
北京大学数学学院教授 张志华

中国人民大学出版社

数据思维

从数据分析到商业价值

王汉生◎编著

中国人民大学出版社
·北 京·

图书在版编目（CIP）数据

数据思维：从数据分析到商业价值/王汉生编著. —北京：中国人民大学出版社，2017.9
ISBN 978-7-300-24856-1

Ⅰ.①数… Ⅱ.①王… Ⅲ.①统计数据-统计分析-应用-商业经营-研究 Ⅳ.①F715

中国版本图书馆 CIP 数据核字（2017）第 199746 号

数据思维
——从数据分析到商业价值
王汉生 编著
Shuju Siwei

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	邮政编码	100080
电　　话	010－62511242（总编室）		010－62511770（质管部）
	010－82501766（邮购部）		010－62514148（门市部）
	010－62515195（发行公司）		010－62515275（盗版举报）
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京鑫丰华彩印有限公司		
规　　格	170 mm×230 mm　16 开本	版　　次	2017 年 9 月第 1 版
印　　张	18 插页 1	印　　次	2017 年 9 月第 1 次印刷
字　　数	243 000	定　　价	69.00 元

序 一

与狗熊会的结缘始于五年前。2012 年，我在拉卡拉支付有限公司任集团高级副总裁，承蒙集团董事长兼总裁孙陶然先生和松禾资本厉伟先生的推荐，有幸考入北京大学光华管理学院就读工商管理硕士，在燕园结识了商学院多个领域的顶级专家和教授。

狗熊会的定位是致力于数据产业的高端智库。先和大家分享一下我和数据产业亲密接触的过程，或许和众多数据领域的创业者们有着一样的心路历程。

2015 年 1 月 5 日，中国人民银行印发《关于做好个人征信业务准备工作的通知》，要求八家机构做好个人征信业务的准备工作，考拉征信位列其中。受集团委托以及董事会任命，我出任考拉征信总裁。虽然我有十余年支付结算领域的工作经验，但是在个人征信方面几乎是一片空白，工作一时难有头绪。于是，在最初的几个月里，我把大部分时间和精力用于学习和交流。我陆续拜访了监管部门、征信业同行、金融机构以及多家大数据公司，发现三个问题：（1）很多机构对征信业务的方向、产品以及服务模式认识不清晰；（2）相当一部分大数据公司缺乏好的商业模式和盈利能力；（3）技术储备不足，数据统计模型设计普遍不强。前两个问题很难在

短期内解决，需要在长期的市场实践中逐步清晰完善。唯有第三个问题或许可以尽快解决，那就是产学研相结合。于是我找到了熊大，也就是王汉生教授。王教授是北京大学光华管理学院统计与经济计量系主任，在国内统计和数据科学领域具有极高的知名度。双方合作由此展开，并成立了联合研究组。由王教授带领的狗熊会团队定期来到公司，双方的数据和模型团队联合作业，对多个产品和评分模型进行了长期深入的研究，成果显著。

2016 年年底我投身于大数据领域的创业热潮。在机缘巧合下，受熊大的邀请有幸出任狗熊会 CEO。此时狗熊会已经与近十家机构开展了联合研究工作，涵盖征信、广告、车联网、消费金融、证券、汽车等多个领域。同时，狗熊会微信公众号聚集了大量粉丝，其中 70%是来自高校的老师和学生，30%是来自大数据企业的从业者。狗熊会团队出品的精品案例甚至已经走进课堂和企业内部的分享培训。

狗熊会的快速发展伴随着中国数据产业的蓬勃兴起，其使命是聚数据英才，助产业振兴。其文化内涵体现在三个方面：一是创造。首先是内容创造，无论是案例还是教材以及研究成果，始终坚持原创，均出自狗熊会成员的智慧。其次是价值创造，知识成果能够为合作伙伴带来数据价值和商业价值。二是分享。助力院校培养更多应用型的数据科学人才，帮助企业提升数据科学水平，共同分享育人的欣慰、科研的成果和智慧的结晶。三是陪伴。从点滴做起，或许是一个案例、一个模型，抑或是一本书、一堂课，还有可能是一个学科、一个专业，狗熊会将始终乐于与大家并肩而行，陪伴中国数据科学产业共同成长。

桃李不言，下自成蹊。欢迎数据科学领域的莘莘学子与从业者关注和加入狗熊会！

狗熊会 CEO 李广雨

序 二

我与王汉生教授相识于北京大学光华管理学院，作为共事多年的老同事，汉生对学术研究的执着、对教书育人的用心都给我留下了深刻印象，用“诲人不倦、古道热肠”来评价恰如其分。这些年，随着中国数据科学产业的蓬勃发展，汉生意识到数据科学人才的匮乏，遂发起成立了狗熊会，旨在聚数据英才，助产业振兴，在资本喧嚣繁华之下尤为难得。值其新书《数据思维》出版之际，汉生委托我写序。盛情难却，故将感慨之言以示支持。

2009 年，我有幸与几位小伙伴一起创立了一家大数据公司——百分点，身份也从一名大学教授转变成一个在商海中打拼的创业者，在大数据这个最热门的“风口”摸爬滚打七八年，接触几千家客户后感慨良多。中国经济经历了 30 多年的快速发展并取得了举世瞩目的成就，经济水平、市场规模、企业数量和质量都取得了飞跃式发展。但不可否认的是，在信息技术层面，我们是断层的，延续性也比较差，并未跟上国家的经济发展水平。西方国家能够比较容易从传统 IT 平稳延展到云计算、大数据，而我们在不同行业则呈现出千差万别的状况，我想这种情况跟思维有着密不可分的关系。

机械思维带来了工业革命，数据思维则引爆智能革命。传统机械思维的核心思想是确定性和因果关系，任何事情一旦发生，则必然会产生结果，一定有可用的模型来描述其发生的原因。而到了数据时代，这个世界正在变得越来越复杂，不确定性无处不在，强相关性则取代了过去的因果关系，数据中包含的信息以及数据之间的相关性则可以帮助我们消除不确定性。在中国大数据产业方兴未艾之际，需要更多人拥有数据思维，无论是政府机构的决策者、商业组织的管理者，还是普通员工、老百姓，都需要学习和了解数据思维。人们常说："思维决定命运。"对于即将到来的智能革命，将会是一个崭新的开始，大家都需要用数据思维来重新认识这个世界。相信汉生这本《数据思维》一定会给广大读者带来受益良多的启发。

王汉生教授也是百分点科学委员会的首席统计学家，在百分点的核心技术、产品研发、大数据项目中给予了大力帮助和支持。此外，百分点与狗熊会都意识到数据科学人才培养的重要性。近年来，百分点与狗熊会联合举办了多场数据科学培训活动，我们都希望涌现出更多的人才来推动国家数据科学产业的快速发展。

"21 世纪什么最贵？人才！"电影中黎叔这句话道出了这个时代的真理。人才的培养，首先体现在思维上，思维跟不上，则永远跟不上。在大数据一线奋斗多年，让我尤其感叹大数据人才在各个行业中的匮乏，也深深明白汉生所做工作的意义和价值。但愿有更多的人能够读到这本《数据思维》，从而为自己开启一个不一样的新世界。

百分点集团董事长兼 CEO 苏萌

序　三

我非常荣幸地阅读了王汉生教授撰写的《数据思维》一书。我首先要祝贺汉生教授和他的团队狗熊会，感谢他们的卓越工作。当今，大数据和人工智能是两大最有活力的热点领域，而现代人工智能的发展本质上也是应数据而驱动。数据思维展示了观念的转换，从而推动了技术的突破。

汉生教授是著名的统计学家，他早年主要从事统计学的理论研究，后来重点关注产业界实际问题的数据分析。特别是近几年，他以敏锐的眼光抓住了学科发展的态势，组建了狗熊会团队。他们从业界中寻找数据科学的实际问题，并帮助业界寻找解决问题的可行途径，由此积累了一批翔实的数据分析案例，这夯实和丰富了数据学科的内涵。《数据思维》一书正是他们实践的总结，蕴涵了汉生教授对数据科学的思考和探索，也体现了汉生教授及狗熊会的时代使命和科学情怀。他们是“聚数据英才，助产业振兴”的践行者，他们的具体行动对“皇帝的新装”给出了最有力的鞭挞。

该书不是仅仅基于文献的总结，也不是基于数学公式的堆砌，而是利

用作者自己完成的案例来对经典和现代的数据分析工具和方法进行重新认识。该书视角独特，语言活泼、风趣、幽默，处处闪烁着作者的思想光芒。我相信它将是一本非常好的数据科学通识读物，该书的出版对数据科学的普及和推广是及时的。我再次祝贺和感谢汉生教授！

北京大学数学学院教授张志华

前　言

市场上已经有那么多关于数据科学（或者大数据）的书了，为什么还要再写一本呢？这是一个很好的问题，我也问过自己八百遍。说老实话，有点稀里糊涂，有点说不清楚。直到有一天，狗熊会公众号（微信 ID：CluBear）上发了一篇题为《关于应用型高校“数据科学与大数据技术”专业建设的一些思考》的文章，探讨产业实践之于数据科学教育的重要性。文章发表后，一位热心读者的留言吸引了我的注意力。这位朋友的留言大意是产业实践可以通过参加类似 Kaggle 的数据建模比赛获得。支撑这个观点的一个原因是这种类型的比赛所使用的数据都来自真实的数据产业，有定义清晰的业务问题，所以，通过参加此类比赛，或者接受类似的训练，就可以获得不错的产业实践经验。但是，我的看法有所不同。我对数据产业实践的理解可能更丰富一些。

我认为数据产业实践的核心任务是：让数据产生价值。更准确地说，是在真实的产业环境中，让数据产生可被产品化的商业价值。这个商业价值是一个广义的商业价值，既包括企业的价值，也包括政府的价值。从这个角度看，数据产业实践至少涉及三个关键环节：数据业务定义（把一个具体业务问题定义成一个数据可分析问题）、数据分析与建模（描述统计、

数据可视化、回归分析、机器学习）、数据业务实施（流程改造、产品设计、标准制定等）。这三个环节缺一不可。而各种数据建模比赛主要关注的是第二个环节（数据分析与建模）。对于第一个环节（数据业务定义）与第三个环节（数据业务实施）能够提供给大家的训练很少。原因很简单，第一个和第三个环节属于赛事主办方的思考范畴，不需要参赛者再操心。参赛者只要对第二个环节发力就可以了。当然，能够对第二个环节提供优质的训练，这仍然是非常值得称赞的事情。

带着对第二个环节无限的尊重，我想说，其实另外两个环节可能更加重要，而且极具挑战性。如果不能把一个业务问题（例如客户价值提升）定义成数据可分析问题，那么任何数据分析都是胡说八道。只有把业务问题准确定义成一个数据可分析问题，数据分析与建模才能有用武之地。最后，即使数据分析得再好、模型建立得再漂亮，如果无法落地成为可被执行的数据产品，那所有的努力也都是白费的。因此，从这个角度看，这两方面更加重要。而这就是狗熊会的核心理念，可能会和很多书籍文章中的看法有所不同。为了方便起见，我称之为朴素的数据价值观。

朴素的数据价值观认为，数据产业实践不是单纯的数据分析与建模，而是要在一个产业环境下，让数据产生价值。为此，前面提到的三个环节都非常重要，尤其是第一个和第三个。而写作本书的目的就是要同大家分享狗熊会朴素的数据价值观。

为了更好地分享，本书大量采用了狗熊会的精品案例。章节内容都是从狗熊会发布的精品案例的微信推文直接润色修改形成的。因此，这些内容继承了狗熊会精品案例的一些有趣的基因：（1）尽最大的努力把业务问题定义清晰；（2）尽最大的努力让数据分析与建模瞄准业务问题；（3）尽最大的努力让最终分析结果有产品化的可能。这三个基因也正好对应了数据产业实践的三个重要环节。为了增加阅读的趣味性，所有案例的写作风格都诙谐幽默，但努力不失科学的严谨。当然，由于各个案例的作者不尽相同，不同章节的写作风格也有所不同，这可能会在一定程度上影响阅读

体验，对此，我表示深深的歉意，请大家原谅。同时为了方便读者利用碎片化时间进行阅读，所有案例之间基本上互相独立，因此，大量章节可以独立阅读，而不受制于前后内容的逻辑顺序。此外，特别值得强调的是，为了降低阅读难度，本书几乎不涉及任何数学符号和计算机代码。但是，这并不代表这些案例是虚构的或者肤浅的。事实上，狗熊会精品案例的生产是一个非常艰辛的过程。一个非常有经验的精品案例 Leader，带领自己的团队，一年最多生产 5 个精品案例。不敢说这些案例多么了不起，但确实是创作团队的心血之作。

在内容组织方面，本书从基本理念入手，按照不同的数据分析方法，由浅入深，组织成不同的章节。其中，第一章系统阐述狗熊会朴素的数据价值观。第二章对经典的统计图表做了系统幽默的阐述。其原型来自狗熊会公号的“丑图百讲”系列。第三章系统阐述我们对于回归分析的理解。在“道”的层面，回归分析是一种重要的思想，是一种将业务问题定义成数据可分析问题的能力；而在“术”的层面，回归分析才是我们常见的各种模型。第四章主要讨论传统的机器学习方法，以及最近很火爆的深度学习。最后一章分享了狗熊会这些年来积累的众多非结构化数据分析的有趣案例，其中涉及中文文本、网络结构、图像分析等不同领域。

本书由狗熊会的核心创作团队，在熊大的“压迫剥削”下，齐心协力，经过多次讨论、修改而成。参与创作的成员有（按姓名拼音排序）：常象宇（政委）、陈昱（昱姐）、黄丹阳（小丫）、刘婧媛（媛子）、罗荣华（康爸）、潘蕊（水妈）、王菲菲（灰灰）、王汉生（熊大）、周静（静静）、朱雪宁（布丁）。创作团队付出了巨大的心血和努力。其中特别要感谢两位朋友：一位是百分点集团的董事长兼 CEO 苏萌博士，是他的启发与鼓励坚定了我们写作的决心；另一位是中国人民大学出版社的李文重编辑，他为书稿的形成付出了巨大的努力，帮助本书选择书名、安排章节、修改文字。大家为什么愿意做出如此辛苦的努力与付出呢？我想都是基于狗熊会的理念：聚数据英才，助产业振兴。这是狗熊会从创立之初到现在从未

改变的理念。

- 聚数据英才说明狗熊会关注数据科学相关的基础教育，并愿意为之付出卓绝的努力。狗熊会希望通过提供优质的教育素材，帮助年轻人成长，享受数据分析的快乐，而不是痛苦，并在这个过程中实现个人职业的幸福成长。
- 助产业振兴说明狗熊会看重产业实践，并认为这才是产生数据科学知识的唯一源泉。狗熊会立志要通过自己微薄的努力，陪伴数据产业一起成长。狗熊会感激每一位曾经合作过的企业伙伴，是他们的鼓励支持让狗熊会站在了中国数据产业实践的第一线，并因此产生了接地气的研究课题，以及高质量的教学产品。

另外，本书中的引用的图片除特别标注的之外均来自网络，鉴于编者在引用这些图片时无法获知原创作者及出处，在此统一对原创作者表示感谢。

最后，把本书献给所有培养过我们的老师，谢谢你们的辛苦栽培。献给我们所有的企业合作伙伴，站在你们的肩膀上，才能看得更远。献给我们的学生，是你们渴望知识的双眼，还有那最美丽的青春年华，让我们重任在肩。献给我们的家人，感谢你们的理解支持，我们才能够努力拼搏，一往无前。祝福我国的数据产业，祝福数据科学教育事业，愿它的每一天都更加美好。祝福狗熊会，愿有更多志同道合的小伙伴，跟我们一起拼搏，“熊”赳赳向前！由于本书写作仓促，疏漏之处难免，请大家多多批评指正！

王汉生（熊大）

狗熊会简介

前言中提到，本书是狗熊会（微信 ID：CluBear）的核心创作集体创作的。相信很多朋友对狗熊会并不了解，因此需要简单向大家介绍一下狗熊会。这是一个什么样的组织？它的名字是怎么来的？它的定位和使命是什么？

几年前，我在美国的一所大学的统计系访问一位很杰出的统计学家。期间我能够比较近距离地观察他的研究团队，那是一个非常棒的、跨学科的科学家团队。我从中学到了很多东西，受到很多的启发。其中最重要的启发就是：也许未来的统计学研究，或者数据科学研究，会跟工程类学科越来越相似。单打独斗，是没有前途的，需要“打群架”才行！因此需要一个强大的、多学科、相互支撑的团队。为此，我下了一个决心：回国后也要好好组织一个强有力的研究团队。要彻底改变过去“小分队作战”的风格，转为“集团军联合作战”。想想当时还是非常兴奋的！

但是，回国以后，这个“集团军”到底应该怎么组织？我没有经验，因此一头雾水，毫无想法。正在这个时候，微信群开始流行起来。于是，我把学生，还有数据领域相关的朋友，整合在一个微信群里，大家经常东拉西扯，也聊和数据相关的话题。这时，问题来了，这个微信群取个什么

名字呢？我想了好久，决定叫“大数据讨论班”。结果没多久，统计之都论坛的二代目魏太云同学就跳出来说：“王老师，这个名字太土了。”原话我记不得了，大意就是：现在啊，到处都在说大数据，但大数据是啥？有清晰、统一的定义吗？还有什么不是大数据吗？这个名字太 low 了！想想也是，于是我说：“那请你给取个名字呗！”太云同学估计受武侠小说荼毒不浅，笑着说：“王老师，咱们叫‘英雄会’怎么样？”我听了，差点没晕过去！这个名字不是更土吗？还英雄会？谁认为我们是英雄啊？我觉得“狗熊会”还差不多！

当时，就是一句逗乐的气话。结果过了几周，我自己也没想出更好的名字来。相反，我越来越觉得“狗熊会”这个名字挺好。狗熊多可爱啊，很多动画片的主角都是狗熊：小熊维尼就是一只熊；《熊出没》里的熊大、熊二也是熊；还有《奇幻森林》里也有一只非常可爱的熊。于是，我在微信群里说了一下这个想法，没想到没人反对！“狗熊会”就这样叫开了，一直延用到现在的微信公众号。由于本书大量的原始素材（例如，原文、音频、数据、程序）都在微信公众号上。因此，要充分享受本书的乐趣，请大家关注狗熊会公众号（ID：CluBear），或者直接扫描二维码。

其实当时也没有什么特别的想法，就是觉得好玩。接下来，意想不到的事情发生了！我意外地发现，“狗熊会”的品牌传播效果出奇得好。为什么？因为这个名字太奇葩了，人们忍不住要问：狗熊会是什么？为什么

要取这么一个奇葩的名字？这名字跟数据分析有什么关系呢？就是这一问一答的过程，让很多朋友记住了这个名字。因此，“狗熊会”成了我们团队的称号，也成了我特别珍惜的品牌。从此在数据的江湖上，王老师开始以“熊大”自称。

作为一个高大上的品牌，狗熊会需要一个自己的 logo。在我的百般恳求下，我家小朋友用铅笔在素描纸上，画了一个大大的熊脑袋。他画出了小朋友心中憨态可掬的熊大。这张草图后来在一位名为冯璟烁的大朋友的帮助下，去掉了一些不必要的线条和背景，再无任何其他修改，成为了狗熊会的 logo。我对这个 logo 超级满意！他画出了我心中狗熊那种傻傻的但是很可爱的样子！这个 logo 也时刻提醒我两件事情：第一，傻傻的狗熊提醒我自己是无知的——对这个世界，对数据相关的学科，自己都是无知的，要保持好奇心，督促自己持续学习。第二，可爱的狗熊提醒我要善良、要快乐，为这个社会多创造一点欢乐的正能量。这两点构成了狗熊会的品牌内涵。

如今的狗熊会是一个致力于数据产业的高端智库。狗熊会帮助合作伙伴制定数据战略，培养数据人才，研究数据业务，发现数据价值，推动产业进步！狗熊会给自己确定的使命是：聚数据英才，助产业振兴！

第一，聚数据英才。这说明狗熊会关注数据科学基础教育，希望通过

生产优质的数据科学科普教育内容（例如本书），提供卓越的研究、实践、就业机会，帮助相关专业的老师、同学、从业者，充分享受数据分析的快乐，促进个人职业的终身幸福与成长。

第二，助产业振兴。狗熊会认为优质的数据科学教育一定不能脱离数据产业实践。狗熊会的任务就是通过联合研究、高端咨询等多种形式，陪伴中国的数据产业一起成长。在此过程中，通过多种形式（例如本书），致力成为连接产学研的桥梁。

温馨提醒：进入狗熊会公众号（CluBear）输入文字："前世今生"，听熊大音频！