全国地方高校UOOC联盟指定参考书
深圳大学教材出版基金资助

新媒体数据挖掘

NEW MEDIA DATA MINING: BASED ON THE R LANGUAGE

基于R语言

王小峰　方　捷　编著

清华大学出版社

新媒体数据挖掘

——基于R语言

王小峰　方　捷　编著

清华大学出版社
北　京

内 容 简 介

计算传播领域尤其是新媒体数据挖掘方向一直缺乏系统的教材，本书旨在为计算传播和计算社会科学领域的读者提供学习 R 编程语言和开发平台的捷径，希望能够填补这方面的空白。“让学习层次变得更宏观，让学习过程变得更轻松，让学习所获变得更通用”是本书的编写理念与特色。本书首先剖析了社会科学研究范式的革新，介绍了 R 语言的作用和特点；然后系统讲解了编程语言的通用学习方法和 R 语言的基本组成；最后展开实战应用，包括网络数据采集、文本挖掘和情感分析、社会网络分析、社交编程平台协作等非常有趣且有意义的内容。

本书适合作为计算传播和计算社会科学领域相关专业本科和研究生教材。高职高专学校也可以选用部分内容开展教学。本书还适合作为计算传播学和计算社会科学科研人员的自学书籍。

本书课件可通过网站 http://www.tupwk.com.cn/downpage 免费下载。

图书在版编目(CIP)数据

新媒体数据挖掘：基于 R 语言 / 王小峰，方捷　编著. —北京：清华大学出版社，2018
ISBN 978-7-302-49322-8

Ⅰ. ①新…　Ⅱ. ①王… ②方…　Ⅲ. ①数据采集－研究　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2018)第 004249 号

责任编辑：王　定
封面设计：周晓亮
版式设计：思创景点
责任校对：曹　阳
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：13.25　　**字　　数**：315 千字
版　　次：2018 年 2 月第 1 版　　**印　　次**：2018 年 2 月第 1 次印刷
印　　数：1～2000
定　　价：58.00 元

产品编号：071992-01

前言

PREFACE

随着互联网、大数据、人工智能等技术的发展，科学技术已经不再只是人类社会的生活背景，而是真正关系到人类整体的生存与发展。多学科相结合、以各学科的视角和专业背景促成人类的自由与科学的发展，是科研工作者在当今时代最重要的研究论题，这反映在人文社会科学领域，正是“哲学社会科学”。

如果说自然科学的使命是研究和预测宇宙，那么哲学社会科学则是研究和预测人类社会。如今“计算范式”已经开始引发社会科学领域的科学范式革命，社会科学的实证研究已经形成“计算范式”与“计量范式”并驾齐驱的格局。

在这样的时代与科研背景下，近些年“人人都要学编程”“人人都要会数据统计”在人文社会科学领域显得越来越重要。由于具有开源、强大的网络扩展功能，广泛的社区支持，强大的数据处理/统计分析和可视化功能，R 语言和 Python 语言俨然已成为当前人文社会科学领域的师生们必须掌握的学习和科研工具。该如何选择这些软件工具？如何真正地高效学习编程语言？如何以最简单但又最标准、最正确的姿态选择和学习一门网络编程语言？这些对人文社会科学领域的师生们来说并不是一件简单的事情。

笔者由于跨学科的背景：十年从事计算机领域的教学与开发工作，转型并进入深圳大学新闻与传播学院(人文社会科学领域)任教，在梳理人文社科、自然哲学的脉络关系中得到了“让世界在内心中逐渐合理起来”的哲学愉悦；为满足教学科研的需要，现将“十年来对计算机编程语言教与学的方法”和“对人文社科、自然哲学的统一观察”一并写成《新媒体数据挖掘——基于 R 语言》，作为这些年来工作与学习的总结。

本书的全部章节安排如下：

第 1 章首先从计算社会科学、计算传播学在国内学术圈中的兴起入题，介绍 R 语言的诞生、功能和在科研工作中的作用，对比几种科研工具的优缺点，并强调“R 是一种自带编程环境的统计软件，Python 是一种自带统计功能的编程语言”，以供读者做出符合自己实际情况的正确选择。

第 2 章以 R 为选择对象，先引入 R 的核心软件 RGui，然后介绍 R 的综合 IDE 开发环境 RStudio 的下载、安装和基本使用。

第 3 章和第 4 章本着“程序=数据+代码”的宏观架构，本着将“编程语言作为语言来学习”的核心主线，选择大家熟悉的英语语法为参照物，对 R 语言的语法进行平缓、细致、精确的讲解，让读者能够真正掌握一种学习任何编程语言的“万能通用方法”：英语被称为

动词的语言，名词(相当于数据类型)和以动词(相当于运算符)为核心的谓语构成简单句(相当于表达式语句)，为表达更复杂的逻辑，英语语法又扩充出了并列句、复合句等语法结构(相当于流程控制)。读者会发现，几乎所有计算机语言的图书，其前几章必然是“数据类型”“运算符”“流程控制”，这其实就是本书提出的“编程语言通用学习主线”；主线之外其他语法项目无外乎锦上添花，例如函数是为了提高代码复用率，软件包是为了引入第三方扩充。

第5章至第8章分别讲述如何用R的核心功能包和扩展功能包实现可视化绘图、互联网数据采集、文本挖掘与情感分析、社会网络分析等功能，这些内容不仅非常有趣，而且非常有应用和科研意义。

第9章引入一个案例学习社交编程平台GitHub，不仅加深对第6章中网络数据采集的学习和应用，还可结合附录中的R软件包的制作、发布与引入方法，真正认识到GitHub作为社交编程平台的重要性和意义：程序员世界的重建巴别塔。

在本书的编写分工上，深圳大学传播学院网络与新媒体系的王小峰老师负责全书的规划、主编与统稿，并参与撰写了第1、第3、第4、第6、第9章；福建师范大学福清分校电子与信息工程学院的方捷老师撰写了第2、第5、第7、第8章和全部附录内容。

由于时间仓促、作者水平有限，本书难免存在遗漏与不足，编者敬请读者批评与指正，我们将会在后续的工作中不断地调整、改进。

深圳大学 王小峰
2017年10月30日夜
于深圳市福田区安托山

目录

CONTENTS

第1章

为什么学习R语言

近年来，由互联网、移动互联网、物联网等平台汇聚的海量数据层出不穷，这为科学研究提供了前所未有的机遇，并在全球范围内兴起了一种不同于通过实验、抽样调查等方法采集结构化数据进行实证研究的“计算范式”。随着大数据时代的到来，“计算范式”的兴起不仅在自然科学领域已经如火如荼，也必定会引发社会科学领域的科学范式革命。社会科学的实证研究已经从“计量范式”的一统天下到“计量范式”与“计算范式”的并驾齐驱、相得益彰。

对科研人员尤其是社会科学科研工作者来说，问卷调查、质化分析等越来越不能满足当今技术时代和研究新范式的要求，人们对结合了“网络存取、科学计算、统计分析、图形可视化”等功能的综合实验平台的需求日益迫切。

本章的学习目标是了解优秀的现代科研软件 R 的概念与功能，及其在社会科学研究新范式——计算社会科学(以计算传播学为例)中的作用。

1.1 R 是什么

近些年，许多新的计算机语言层出不穷，R、Python、Node.js、Go、Ruby……让人感觉眼花缭乱，那么 R 语言到底是什么呢？下面将从 R 的诞生、创造目的、优势与不足、R 与 Python 的区别等方面进行阐述。

1.1.1 R 是一款优秀的现代科研软件

S 语言是由 AT&T 贝尔实验室(AT&T Bell Laboratories，如图 1-1 所示)开发的一种用来进行数据探索、统计分析和作图的解释型语言。最初 S 语言的实现版本S-PLUS是一款商业软件，于 20 世纪 80 年代后被广泛应用于统计领域，并由 MathSoft 公司(世界领先的统计计算软件开发商和供应商)的统计科学部进一步完善。后来新西兰奥克兰大学统计系的 Robert Gentleman 和 Ross Ihaka 及其他志愿人员基于 S-PLUS 完善并开发了 R 系统。综上，可以认为 R 是 S 语言的一种实现。

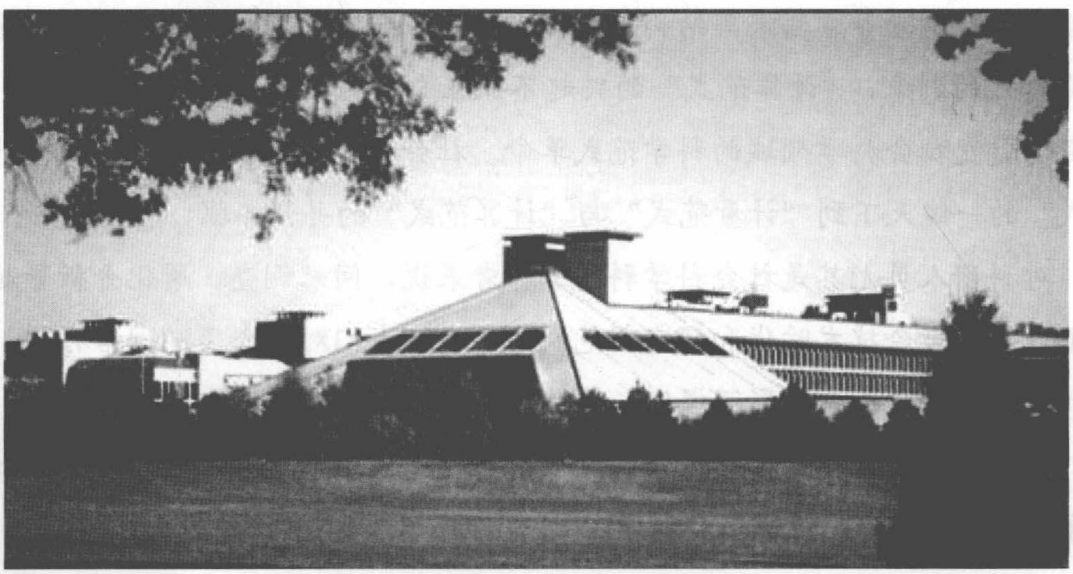

图 1-1　R 语言的起源与诞生地：AT&T 贝尔实验室美国总部

简单来说，R 是：

(1) 一种统计计算编程语言，源自 S 语言(由 AT&T 贝尔实验室的 Rick Becker、John Chambers 和 Allan Wilks 开发的一种统计编程语言)。

(2) 开放[遵循 GPL(General Public License，通用性公共许可证)协议]的统计编程环境。

(3) 一种综合科研软件平台——集科学计算、统计分析、图形可视化等功能于一体的科研软件。

目前 R 由 R 核心开发小组(R Development Core Team，RDCT)维护，他们完全自愿，工作努力负责，并将全球优秀的科学计算、统计应用、图形可视化功能打包提供给我们免费使用。我们可以通过 R 的官方网站(http://www.r-project.org)了解有关 R 的最新信息和使用说明，得到最新版本的 R 软件和基于 R 的扩展功能软件包。

1.1.2　R 的优势与不足

R 以其强大的计算和统计能力、突出的可视化能力绘图，受到越来越多数据分析科研工作者的喜爱。R 具备容易学习、跨平台、自由/免费、源代码开放、强大的社区支持等优势。表 1-1 所示为 R 与 Python、SPSS、Excel、MATLAB 等软件平台的对比。

表 1-1　R 和其他软件的对比

软件平台	编程支持度	开源/免费	跨平台	社区	大规模数据支持	是否通用语言
R	高	是	是	有	支持	否
Python	高	是	是	有	支持	是
MATLAB	低	否	是	无	支持	否
SPSS	低	否	是	无	不支持	否
Excel	低	否	否	无	不支持	否

当然 R 也并非十全十美，比如不能用 R 开发 Web 类或互联网类应用程序；数据必须保存在物理内存中——但随着计算机内存容量的不断提升，这个问题已经在很大程度上得到解决；缺少充足的交互元素，但以 JavaScript 为代表的客户端脚本语言介入并填补了这项空白，虽然我们仍然需要利用 R 语言处理分析任务，但最终结果的具体显示方式则可以由 JavaScript 等其他语言来完成。

总之，R 语言并非只适用于程序员，它非常适合那些面向数据并试图解决相关问题与科研工作的用户——无论他们的实际编程能力如何。

1.1.3　R 和 Python 的区别

许多人在选择统计分析、数据科学科研平台时有所困惑，尤其是在 R 和 Python 之间进行选择时，这里简单剖析如下：

- R 是一种带有编程环境的统计软件，由统计学家发起并主导开发。
- Python 是一种带有统计环境的编程语言，由数据科学家发起并主导开发。
- 结合自身的实际情况，例如工作过程是以统计分析为主还是以数据科学研究为主？工作过程是以科研报告为主还是以系统开发为主？只有明白了自己真正的需求，才能进行科学、合理的选择。

1.2 计算社会科学的兴起——以计算传播学为例

在历经数个世纪、从史料分析到统计和数学模型的发展之后，社会科学也变得“可计算”了。由于计算所发挥的核心作用，计算社会科学在方法论上是十分显然的。从科学的角度来看，计算范式的提出非常重要。同时，计算社会科学也依赖复杂适应系统的跨学科理论，以及信息处理在人类和社会行为各个尺度上所起的作用。但社会系统本身具有高维属性，对社会复杂系统进行研究需要跨学科合作。大数据的重要特征就是数据的超高维品质，这为跨学科研究提供了合作平台。大数据时代计算社会科学的跨学科研究——计算社会科学是研究社会科学问题的新思潮和新方法。

1.2.1 什么是计算社会科学

1. 自然科学的终极使命：预测宇宙

2017年10月3日，雷纳·韦斯等三位教授关于引力波的研究获诺贝尔奖，该研究被评价为“完成了广义相对论最后一块拼图”。事实上，20世纪30年代以后，量子物理、相对论物理都有了较大的发展。随着实验条件的提高，特别是大能量粒子对撞机的建成，物理学家们发现了六十多种都可被称为“基本”的粒子。本着追求统一性的思想，他们居然找到了一种被称为标准模型的理论。把这些粒子归纳进去，同时统一了三种力。他们通过标准模型预言了一些粒子的存在，有些真的在实验室中找到了，这更加鼓舞了物理学家们的信心，很多物理学家一直在做着把引力像其他三种力那样也统一到一起的努力。如果这样的工作完成了，建立起来的大统一理论，就是一种能够解释宇宙全部现象的终极理论。

更早以前，牛顿在其最重要的著作《自然哲学之数学原理》中构建了科学有史以来第一个完整的、科学的宇宙论和科学理论体系，并试图用统一的力学原因解释宇宙所有的运动和现象，它所造成的影响极其深远。当科学从牛顿时代走来，辉煌的成功使得自然科学家终将具有预测一切的能力：万事万物都已经由物理定律规定下来，一个细节都不能更改。过去、现在和未来都像已经写好的剧本，宇宙的发展只能严格地按照这个剧本进行，不允许加以任何的发挥。这就是牛顿力学的可预见性假设，是决定论或者说确定论的牛顿力学送给我们的礼物。18世纪，法国著名科学家拉普拉斯曾以此为依据豪迈地宣告：“如果已知宇宙中每一粒子的位置和速度，我就能够预测整个宇宙的未来。”

这一美好的图景在20世纪变得黯淡许多，先是遭遇量子论的严重挑战，随后又被兴起的混沌学彻底击碎。1986年，这部划时代巨著——《自然哲学之数学原理》出版整整三百周年，英国皇家学会专门举办了隆重的纪念大会。在这次大会上，著名的流体力学权威詹姆士·莱特希尔爵士发表了令人震惊的道歉宣言。他说：“今天，我们深深意识到，

我们的前辈对牛顿力学惊人成就的崇拜，促使他们认为世界具有可预见性。的确，我们在 1960 年以前大都倾向于相信这个说法，但现在我们知道这是错误的。我们曾经误导了公众，向他们宣传说满足牛顿运动定律的系统是决定论的，可在 1960 年后这已被证明不是真的，为此，我们愿意向公众表示道歉。”这种看似出乎意料的举动，却是发展着的科学的一种正常的表现，实际上是科学对自身发展的一种反思。综上所述，无论实现与否，可以认为自然科学的终极目标是“预测宇宙”，如图 1-2 所示。

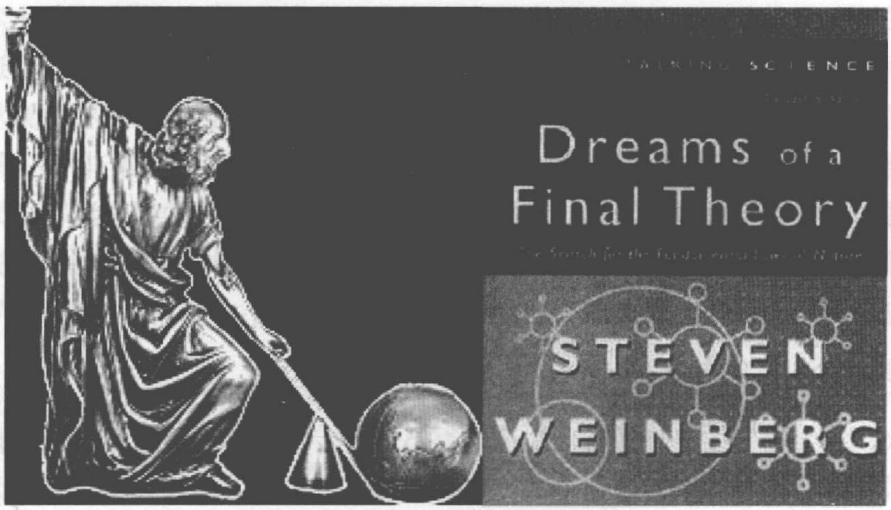

图 1-2　自然科学的目标是“预测宇宙”

2. 社会科学的终极使命：预测人类社会

社会科学(social science)是用科学的方法，以社会客体为对象，研究人类社会现象与事物的本质及其规律的系统性科学，包括法学、经济学、政治学、社会学、历史学等学科的庞大知识体系，广义的哲学社会科学还包括人文科学。简单来说，社会科学是用科学的方法，研究人类社会现象的学科；如果自然科学的终极目标是预测宇宙，那么社会科学的终极目标则是预测人类社会。图 1-3 所示是 2017 年 10 月南京大学召开的全国第二次计算传播年会上，香港城市大学祝建华老师展示的 PPT 中的内容。

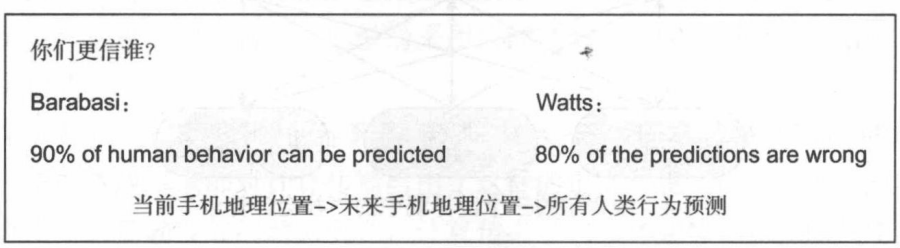

图 1-3　社会科学的目标是“预测人类社会”

马克思曾预言：“科学只有从自然科学出发，才是现实的科学。历史本身是自然史的，即自然界成为人这一过程的现实部分。自然科学往后将包括关于人的科学，正像人的科学

包括自然科学一样：这将是一门统一的科学。”(《马克思恩格斯全集》第42卷第128页)一百多年来，人类在自然科学和社会科学两个方面均已取得突飞猛进的发展，特别是自然科学的发展尤为突出，不仅深入到自然界的宏观领域、宇观领域和微观领域，还深入发展到人类机体(包括大脑)和人类社会领域，而且这两个方面仍在不断地相互渗透、相互整合。社会科学(包括思维科学)越来越多地受到自然科学的影响，许多自然科学的研究方法已经卓有成效地应用于社会科学的研究过程之中，有力地推动了社会科学的发展。社会科学的这种不断采用自然科学研究方法的发展趋势，被称为社会科学的“自然科学化”。

其实原因很简单，社会科学的认识论往往同描述、解释和预测有关，如果这种描述、解释和预测要具有系统性、可测性和可重复性等重要科学特征，则需要将计算与社会科学进行有机结合。计算社会科学(Computational Social Science，CSS)是近10年内兴起的一种采用互联网、大数据、机器学习等计算技术来研究社会科学问题的新思潮和新方法。作为“研究自然界物质的类型、状态、属性及运动形式的科学”，计算是自然科学的核心手段，如将计算应用于社会科学，就提出了社会科学新的研究范式：“计算社会科学”。

3. 计算社会科学的提出

近年来，社会科学在“计量时代”已取得不小进步，但因为受研究方法和技术手段限制，社会科学还有不少基本问题至今都没有得到解决，对于人类社会与人类行为规律的总结与发现等核心社会科学研究问题，还处于非常初级的水平。大数据时代的到来，不仅为社会科学研究获得了全新的数据来源，更为社会科学范式革命提供了基础数据，使得人们对复杂社会系统的信息收集与分析能力取得突破性进展。如图1-4所示，西方当代著名哲学家卡尔·波普尔在其“三个世界”的理论中提出了“社会计算学”。

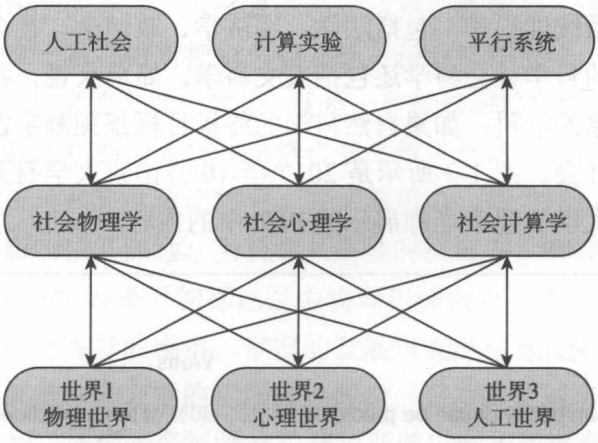

图1-4　波普尔的三个世界观中的“社会计算学”

2012年，R. Conte(意大利国家科研委员会)、N. Gilbert(英国萨里大学)、G. Bonelli(意大利国家科研委员会)、C. Cioffi-Revilla(美国乔治梅森大学)等14位欧美学者在 *The European Physical Journal Special Topics* 第1期上联合发布了一份《计算社会科学宣言》(后文简称《宣言》，见附录1)。《宣言》从机遇、技术发展、方法创新、面临的挑战和预期

的影响五个方面，全景式地说明了计算社会科学发展现状及其未来的方向。《宣言》认为信息与通信技术可以为社会科学提供帮助，它不仅有助于获得、分析和构建大数据(Big Data)以探讨大问题(Big Problems)，还有助于提供大思考(Big Thinking)的方法，力图呼唤一场社会科学革命。

无论是创建新的理论和仿真，还是对社会系统及其过程的假说进行测试和建模，我们都需要涉及大量的信息。在 2016 年中国社会学年会的“大数据与计算社会科学论坛”上，天津工业大学阎耀军教授从控制论角度，展示了大数据对于预测社会复杂系统和实现前馈控制的重要意义。阎教授的报告给人印象特别深刻的是，他展示了一种来自 110 报警地点的空间信息数据，据此可以分析不同类型犯罪行为在城市空间的分布特征，进而为城市警力的布置提供依据和向导。也就是说，可以根据这种类型的大数据，预测城市不同空间位置、不同类型犯罪行为的发生概率。他们根据这样一种数据所获得的启示进行犯罪干预，结果使得天津某区域的犯罪率下降 50%以上。阎教授的报告展示了大数据在社会治理领域的强大威力。

1.2.2　计算传播学的起源和概念

计算社会科学以计算范式为辅，以社会科学为主，仍属于社会科学。这一思想完全类似于计算天文学、计算生物学和计算语言学分别属于天文学、生物学和语言学一样。根据定义，任何一门计算 X 科学都属于 X 科学的一部分。结合笔者目前的研究经历和工作平台，本书将以计算传播学为应用场景，讲解 R 软件的实际应用。

计算传播是指数据驱动的、借助可计算方法进行的传播过程，而分析计算传播现象的研究领域就是计算传播学(王成军，2015)。计算传播的应用有很多，例如数据新闻、计算广告、媒体推荐系统等，在过去的几年里，产生了深远的影响。数据新闻风靡全球，重要的国际媒体和国内媒体纷纷采用数据新闻，以开放数据、数据挖掘、可视化的方式提供信息；计算广告备受瞩目，不管是门户网站、搜索引擎，还是社交媒体，纷纷将计算广告当作数据变现的重要渠道，以可计算的方法对广告进行拍卖，实现媒体、内容和用户三方的匹配；媒体推荐系统成为个性化信息获取的重要途径，既包括传统的社交新闻网站，也包括“今日头条”这种后起之秀，它们纷纷采用协同过滤的方法为用户提供信息，建立了新的信息把关模式。

如图 1-5 所示，在国内最大的计算传播研究社区——“计算传播网”(南京大学)的主页上，南京大学的王成军老师对计算传播给出了这样的定义与思考：“我们一直在寻找可以支撑这个分支学科蓬勃发展的动力所在，计算传播学的整体架构仍在酝酿当中。我觉得其精髓在于可计算性。如何才能具有可计算性？测量是第一步。货币使得经济学具有可计算性，实验使得心理学获得可计算性，字节使得计算机科学具有可计算性，基因使得生物学具有可计算性。什么可以使得人类传播行为具有可计算性？寻找传播学的货币和基因是计算传播学的首要任务。按照我的个人理解，网络化的大规模数据的 digital traces 第一次使

得传播行为获得了计算性，而 document、collect、analyze、visualize 这些传播行为成了计算传播学的主要工作。按照这个设想，传播学必须走出传统的研究套路，获得在网络上保存、抓取、分析、可视化大规模电子化数据的能力，也需要支持这些工作的工具。毫无疑问，传播学因此将和计算机科学开始交汇，至少需要程序员投入到这种大规模数据的挖掘工作中来。”从中，我们可以简化出关于计算传播的三段论：

(1) 计算传播寻找人类传播行为可计算化的基因。

(2) 如果基因是生物学飞跃的原因，货币是经济学发展的关键，那么人类传播行为所隐藏的计算化“基因”是什么？

(3) 计算传播学致力于寻找传播学可计算化的基因、学习和传播可计算化思维/方法(电子化数据收集能力、编程能力、数学建模能力、网络分析、文本挖掘)、了解和训练计算传播学的社会化应用方法(数据新闻、计算广告、可视化等)。

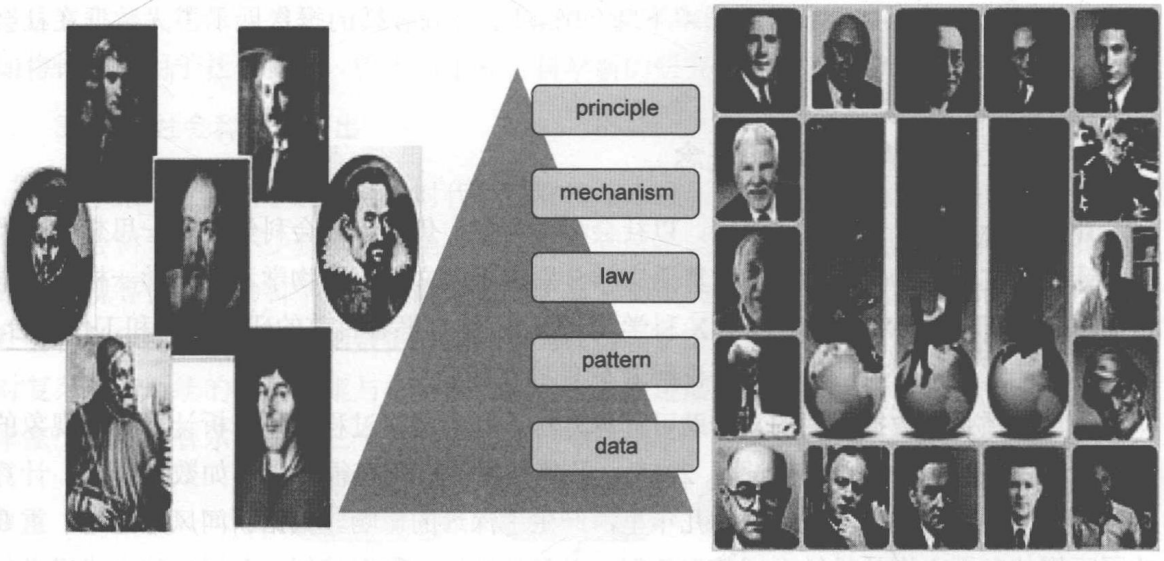

图 1-5　科学研究的五个层次

简单来说，计算传播(computational communication)是计算社会科学(computational social science)的重要分支。国内最早进行计算传播研究的祝建华老师(香港城市大学)认为，计算传播主要有如下几个特征：

- 主要关注人类传播行为的可计算性基础。
- 以传播网络分析、传播文本挖掘、数学建模等为主要分析工具。
- (以非介入地方式)大规模地收集并分析人类传播行为数据。
- 挖掘人类传播行为背后的模式和法则。
- 分析模式背后的生成机制与基本原理。
- 可以被广泛地应用于数据新闻和计算广告等场景。
- 注重编程训练、数学建模、可计算思维。

计算传播作为一个崭新的研究领域，需要研究者投入更多注意力。分析计算传播应用、

从传播学的角度研究计算传播的实际问题，具有不可忽略的意义，反过来讲，分析和总结计算传播学的研究方式，对于传播学自身的发展而言也具有重要意义。

目前国内主要的计算传播科研社区支持有：

- 计算传播网 https://computational-communication.com/
- 计算传播学豆瓣小站 http://site.douban.com/146782/

1.3 R 在计算传播学中的典型应用

下面以几个简单、有趣、可重复试验的案例，展示 R 语言在计算传播学中的典型应用。注意，全书所有代码都正确通过运行测试，运行环境为：Windows 7 64 位旗舰版，R x64 3.4.2，RStudio-1.0.153。

1.3.1 用 R 进行文本分析初探

由于语言的特殊性，中文在进行文本挖掘时需要进行分词，这里以《金庸-天龙八部》作为离线文本数据，使用 jiebaR 包进行中文分词和去停用词，构建词频统计表，最后利用 wordcloud 进行词云可视化展现。

#R 代码清单，在 R-3.4.2、RStudio-1.0.153 环境下运行通过

#例 1-1：《金庸-天龙八部》文本基础分析，运行结果如图 1-6 所示

```
#加载 jiebaR 分词包和 wordcloud 词云图包
if(!require("jiebaR")){install.packages("jiebaR")} ; library("jiebaR")
if(!require("wordcloud")){install.packages("wordcloud")};
library("wordcloud")
engine<-worker()#根据默认参数建立分词引擎
```

##下面读取 txt 文本

```
xajh<-read.table("金庸-天龙八部.txt",
       sep="\t",header=F,colClasses="character")      #读取《金庸-天龙八部》txt 文件
head(xajh)                                           #查看表头
xajh$V1[c(1:5,5000:5005,20000:20005)]                #随机查看部分内容
```

##下面进行文本分词

```
words<-engine<=xajh$V1                               #分词
words1<-unlist(words)
words1<-words[words!=""]
words2<-words1[nchar(words1)>1 & nchar(words1)<7]    #取字符长度介于 2 和 6 的词
wordFreq25=sort(table(words2),decreasing=T)[1:25];wordFreq25 #输出前 25 个高频词
pal2 <- brewer.pal(8,"Dark2")
```