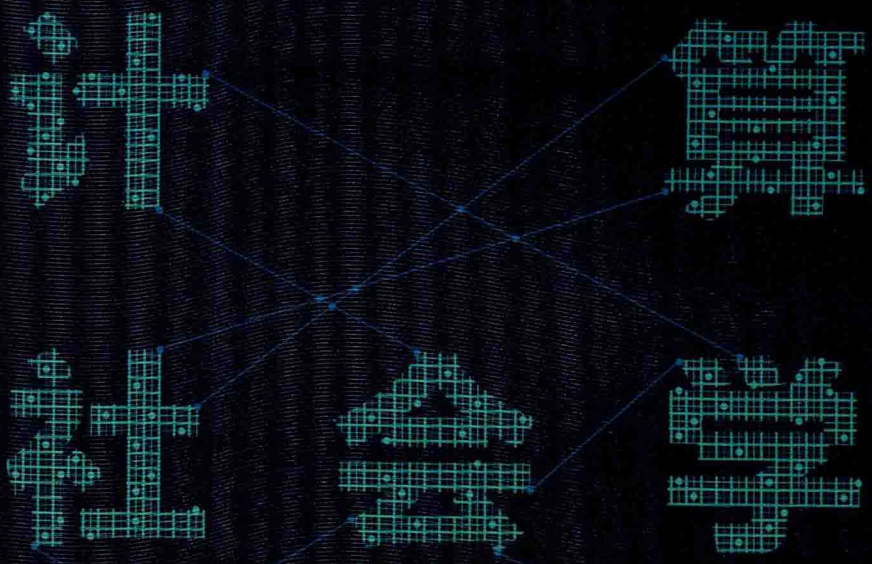

荣获美国出版商协会
2019年
专业与学术杰出出版奖

数据时代的社会研究

SOCIAL RESEARCH IN THE DIGITAL AGE

[美] 马修·萨尔加尼克（Matthew J. Salganik）◎著
赵红梅　赵婷 ◎译

BIT BY BIT

中信出版集团

计算社会学

[美] 马修·萨尔加尼克（Matthew J. Salganik）◎著
赵红梅　赵婷 ◎译

BIT BY BIT

SOCIAL RESEARCH IN
THE DIGITAL AGE

中信出版集团 | 北京

图书在版编目（CIP）数据

计算社会学 /（美）马修·萨尔加尼克著；赵红梅，赵婷译 . -- 北京：中信出版社，2019.5

书名原文 : Bit by Bit: Social Research in the Digital Age

ISBN 978-7-5217-0118-0

I. ①计… II. ①马… ②赵… ③赵… III. ①数据管理—应用—社会科学—研究 IV. ① C53

中国版本图书馆 CIP 数据核字（2019）第 033627 号

计算社会学

著　　者：[美] 马修 · 萨尔加尼克

译　　者：赵红梅　赵婷

出版发行：中信出版集团股份有限公司

（北京市朝阳区惠新东街甲 4 号富盛大厦 2 座　邮编　100029）

承 印 者：中国电影出版社印刷厂

开　　本：787mm × 1092mm　1/16　　印　　张：27.25　　字　　数：400 千字

版　　次：2019 年 5 月第 1 版　　印　　次：2019 年 5 月第 1 次印刷

京权图字：01-2018-7853　　广告经营许可证：京朝工商广字第 8087 号

书　　号：ISBN 978-7-5217-0118-0

定　　价：69.00 元

前 言

这本书源于2005年在哥伦比亚大学一个地下室中发生的事情。那时，我还是一名研究生，正在为最终的毕业论文做一项在线实验。有关这项实验的学术部分我将在第4章进行介绍，但现在，我将告诉你们一件我的毕业论文或其他任何论文都未涉及的事情。这件事从根本上改变了我对研究的看法。一天早晨，当我来到位于地下室的工作室时，我发现一夜之间竟有约100个来自巴西的人参与了我的实验。这一简单的经历对我产生了深远的影响。当时，我的几个朋友正醉心于传统的实验室实验。我知道他们必须很费心地有偿召集并指导人们来参与实验，如果一天能有10个人完成实验，就算是不错的进展了。但对我的在线实验来说，我在睡觉的同时，就有100个人参与了实验。也许，一边睡觉一边做研究听起来美好得令人难以置信，但这是事实。技术的变化，尤其是技术从模拟时代到数字时代的转变，意味着我们可以用新的方式搜集和分析社会数据。这本书就是关于如何利用这些新方式开展社会研究的。

这本书是写给那些“想研究更多数据科学的社会科学家”和“想研究更多社会科学的数据科学家”以及对这两个领域的结合感兴趣的人

的。因此，这本书的目标受众自然就不仅限于学生和教授了。尽管我目前在普林斯顿大学任职，但之前也在政府部门（美国人口调查局和技术产业领域的微软研究院）工作过，因此我知道，在大学之外同样存在着很多令人兴奋的研究。因此，只要你觉得自己正在做的是社会研究，那么无论在何处就职或使用何种技术，你都可以参阅此书。

你可能已经注意到了，这本书的语言风格和许多其他的学术著作稍有不同。这其实是我特意做的一个改变。这本书的部分内容源于我从 2007 年起在普林斯顿大学的社会学系带领的一个“计算社会学”（Computational Social Science）研究生研讨班，因此我希望它能反映这个研讨班的一些活力和激情。具体而言，我希望这本书能够具备以下三个特点：有帮助的、面向未来的以及乐观的。

有帮助的：我的目标是写一本对你们有帮助的书。因此，我将以坦诚的态度、非正式的写作风格，通过实例阐述我的观点。我最想传达的是一种特定的思考社会研究的方式，而经验告诉我，传达这一思考方式的最好的方法就是采取非正式的写作风格并列举大量例子。此外，在本书的参考文献中，有一个部分叫“拓展阅读”，它旨在帮你过渡到有关我所介绍的多个主题的更加详细、更加专业的内容上。最后，我希望此书能对你们自己开展研究以及评估别人的研究有所帮助。

面向未来的：我希望这本书能帮助你利用现有的以及未来将出现的数字系统开展社会研究。我是从 2004 年开始做这类研究的，这期间数字系统发生了诸多变化，我坚信在你们的职业生涯中，你们也会感受到数字系统的许多变化。因此，要想让这本书“以不变应万变”，就要做到“抽象”。例如，这本书不会教你如何使用现有的推特应用程序界面（Twitter API），相反，它会教你如何受益于大数据资源（第 2 章）。这本

书不会详细告诉你如何利用亚马逊土耳其机器人（Amazon Mechanical Turk，以下简称机器人 MTurk）开展实验步骤，相反，它将教你如何设计和解读依赖于数字时代基础设施的实验（第 4 章）。通过采用这种抽象化的手法，我希望这本书能够成为一本主题适时、经得起时间考验的书。

乐观的：本书涉及两个群体——社会科学家和数据科学家，他们有着截然不同的背景和兴趣。除了书中将要介绍的科学方面的差异以外，我还发现，这两个群体看待事物的态度也是不同的。数据科学家一般而言是满怀希望的，而社会科学家一般而言是更具批判性的。也就是说，同样是半杯水，数据科学家看到的是还有半杯水，而社会科学家看到的则是杯子有一半已经空了。在本书中，我将采取数据科学家的乐观态度。因此，在描述相关实例时，我将告诉你们在我看来这些例子的可取之处。当然，鉴于没有研究是完美的，我也会指出它们的问题所在，但我会尽力用乐观积极的方式指出。我不会为批判而批判，我的批判是为了能让你们设计出更好的研究。

我们仍处于数字时代社会研究的早期阶段，但我已经发现了一些普遍存在的误解，它们的普遍程度让我觉得有必要在前言中对其进行说明。就数据科学家而言，我发现他们有两个常见的误解。第一个是认为数据越多越有利于解决问题。但对社会研究来说，我的经验告诉我并不是这样的。事实上，对社会研究来说，好的数据似乎要比更多的数据更有帮助。第二个是数据科学家通常认为社会科学只不过是一堆围绕常识的花言巧语罢了。当然，作为一名社会科学家，更确切地说是社会学家，我不同意这样的观点。聪明的人长期以来一直在努力理解人类的行为，因此忽视这一努力所取得的成果似乎是不明智的。我希望通过这本

书，以一种易于理解的方式和你们分享其中的一些成果。

就社会科学家而言，我发现他们也有两个常见的误解。第一个是有些社会科学家会因为少数不真实的数据而彻底否定使用数字时代的工具开展社会研究这一观念。如果你正在读这本书，那你可能已经读过许多平庸地或错误地（或两种方式都有）使用社交媒体数据的论文。我也读过。但是如果因为这些论文就得出结论，说数字时代的社会研究都是不好的，这将是一个严重的错误。事实上，你可能也读过许多平庸地或错误地使用调查数据的论文，但你并没有因此而否定所有使用调查数据的论文。这是因为你知道，也有使用调查数据并且做得很不错的研究。而我将通过这本书告诉你们，使用数字时代的工具并且做得很不错的研究也是有的。

我所发现的社会科学家的第二个常见误解是容易将现在和未来混淆。当我们对数字时代的社会研究，即我在本书中将探讨的研究，进行评估时，思考以下两个截然不同的问题至关重要："这类研究现在做得怎么样"以及"这类研究将来会做得怎么样"。研究人员会被训练来回答第一个问题，但对这本书而言，我认为更重要的是第二个问题。也就是说，尽管数字时代的社会研究尚未做出巨大的、改变范式的贡献，但数字时代社会研究的进步速度快得惊人。因此，相比于其目前的发展水平，它的变化速度更让我感到兴奋不已。

尽管上一段似乎是在告诉你们，数字时代的社会研究可能会在未来的某个时间变得相当成功，但我的目标并不是向你们推销任何特定类型的研究。我个人并未持有推特（Twitter）、脸谱网（Facebook）、谷歌（Google）、微软（Microsoft）、苹果（Apple）或其他任何科技公司的股份。但是，为了做到充分披露，我应该告诉你们我曾在微软、谷歌和脸

谱网工作过或是接受过其研究经费赞助。因此，在整本书中，我的目标是让自己做一个可信的叙述者，告诉你们所有可能的令人兴奋不已的新事物，同时引导你们避开一些我曾看到有人掉进去的陷阱（有的我自己也曾掉进去过）。

社会科学和数据科学的交叉学科有时会被称为“计算社会学”。有些人认为这是一个技术领域，但这本书并不是传统意义上的技术图书。例如，这本书的正文中并没有公式。之所以选择这样的方式，是因为我想呈现对数字时代社会研究的一个全面的看法，其中包括大数据资源、调查、实验、大规模协作和道德伦理。但事实证明，涵盖所有这些主题并提供每个主题中详细的技术细节是不可能的。相反，我会在本书参考文献中的“拓展阅读”里推荐更多的技术资料。换句话说，这本书不是为了教你如何做某种特定的计算，而是为了改变你对社会研究的思考方式而写的。

如何在教学中使用这本书？正如前面所述，本书的部分内容来自我从 2007 年开始在普林斯顿大学带领的一个“计算社会学”研究生研讨班。你们可能想用这本书进行教学，所以我觉得有必要解释一下我是如何将源于课堂的素材写成这本书的，以及我想象的这本书在其他课堂中的使用方式。

有几年时间，我上课是没有指定教材的，我只是给学生指定一些文章。虽然他们能够从这些文章中学到东西，但只学习这些文章还不足以让他们发生我所期待的观念转变。所以我会用课堂大部分的时间讲述这些文章的背景，讲述应该采取怎样的视角以及给予他们建议，进而帮助学生获得更全面的认识。在这本书中，我试图以不涉及社会科学或数据科学专业知识的方式记录上述所有的背景、视角和建议。

对于为期一学期的课程，我建议将这本书与其他各种阅读材料配套使用。例如，课程可能会花两周时间来做实验，这时你可以使用第 4 章的内容，同时选取诸如以下主题的阅读材料：预处理信息在实验设计和分析中的作用；在公司大规模的 A/B 测试过程中所浮现出来的统计和计算问题；实验设计，尤其是原理方面，以及与通过机器人 MTurk 这样的在线劳动力市场招募实验参与者相关的实践、科学和伦理方面的问题。你也可结合编程方面的阅读材料或活动。至于如何从这些材料中选出合适的配套材料，就取决于你的学生（是本科、研究生还是博士）以及他们的背景和目标。

在一个为期一学期的课程中，你也可以每周给学生分配一些任务。这本书的每一章都会涉及各种各样的“活动”，我将把“活动”放在参考文献中，同时我也标注了它们的难度等级：简单（  ）、中等（）、困难（）以及非常困难（）。此外，我还标注了每个问题所需的技能：数学（  ）、编码（）以及数据采集（）。最后，对一些我个人比较喜欢的活动，我会备注心形图标（）。我希望在这么多的任务活动中，你能找到适合自己的。

为了帮助人们在教学中使用这本书，我已经开始搜集相关的教学资料了，例如教学大纲、幻灯片、每章推荐的配合材料以及一些任务活动的解决方案。你可以访问 http://www.bitbybitbook.com 查看或完善这些资料。

目　录

第3章 提问

第4章 开展实验

第 5 章 进行大规模协作

第 6 章 道德伦理

第1章
简介

1.1 一处墨迹

2009 年夏天，手机铃声响遍了整个卢旺达。除了来自家人、朋友和商业伙伴的数百万个电话之外，大约有 1 000 名卢旺达人还接到了由乔舒亚 · 布卢门斯托克（Joshua Blumenstock）及其同事打来的电话。研究人员从卢旺达最大手机供应商的数据库中随机抽样进行调查，以完成对财富与贫困的研究，这个数据库中有 150 万名客户。布卢门斯托克和他的同事会询问这些被随机选中的人是否愿意参与调查，然后向其解释这项研究的性质，接下来便会询问一系列有关他们的人口学特征、社会特征和经济特征方面的问题。

到目前为止，我所描述的一切都让这项研究听起来像是一项传统的社会科学调查。但接下来我要描述的就不再传统了，至少目前来说是这样的。除了调查而来的数据外，布卢门斯托克和同事还拥有这 150 万人的完整通话记录。他们将这两部分数据结合起来，利用调查数据训练了一个机器学习模型，使模型能根据一个人的通话记录预测其财富状况。接着，他们利用这个模型评估数据库中 150 万名客户的财富状况，还利用通话记录中包含的地理信息判断这 150 万名客户的居住位置。最后他们将所有这些信息——估算的财富状况以及居住位置，综合到一起，绘