

O'REILLY®

TURING

“数据时代提高思辨能力和逻辑
推理能力之精选。”

— Andrew Theriault, 博士

美国民主党全国委员会数据科学部主任

[美]萨曼莎·克莱因伯格 著
郑亚亚 译

别拿相关 当因果!

因果关系简易入门

WHY

A GUIDE
TO FINDING
AND USING
CAUSES



中国工信出版集团



人民邮电出版社
POSTS & TELECOM PRESS

TURING

别拿相关当因果！ 因果关系简易入门

Why: A Guide to Finding and Using Causes

[美] 萨曼莎·克莱因伯格 著
郑亚亚 译



Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

授权人民邮电出版社出版

人民邮电出版社
北京

图书在版编目 (C I P) 数据

别拿相关当因果! : 因果关系简易入门 / (美) 萨曼莎·克莱因伯格 (Samantha Kleinberg) 著; 郑亚亚译. — 北京: 人民邮电出版社, 2018.7
ISBN 978-7-115-48518-2

I. ①别… II. ①萨… ②郑… III. ①因果性—研究
IV. ①B025.5

中国版本图书馆CIP数据核字(2018)第108657号

内 容 提 要

本书是写给普通人的因果逻辑入门书,旨在帮助读者培养严谨的思维方式,在不借助任何专业知识的前提下,准确定位问题。主要内容包括:认识原因,对原因的理解和运用,如何只通过观察找到原因,大数据集与原因的关系,因果关系相关实验,如何利用因果关系来制定有效的干预措施,研究因果关系的意义。

本书适合所有对探究事件真相感兴趣的读者,无须统计学等专业背景。

◆ 著 [美] 萨曼莎·克莱因伯格
译 郑亚亚
责任编辑 朱 巍
责任印制 周昇亮

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 <http://www.ptpress.com.cn>
北京隆昌伟业印刷有限公司印刷

◆ 开本: 880×1230 1/32
印张: 9.375
字数: 249千字 2018年7月第1版
印数: 1-3 000册 2018年7月北京第1次印刷

著作权合同登记号 图字: 01-2018-0946号

定价: 69.00元

读者服务热线: (010)51095186转600 印装质量热线: (010)81055316

反盗版热线: (010)81055315

广告经营许可证: 京东工商广登字 20170147号

站在巨人的肩上

Standing on Shoulders of Giants



iTuring.cn

版权声明

© 2016 by Samantha Kleinberg.

Simplified Chinese Edition, jointly published by O'Reilly Media, Inc. and Posts & Telecom Press, 2018. Authorized translation of the English edition, 2016 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 O'Reilly Media, Inc. 出版，2016。

简体中文版由人民邮电出版社出版，2018。英文原版的翻译得到 O'Reilly Media, Inc. 的授权。此简体中文版的出版和销售得到出版权和销售权的所有者——O'Reilly Media, Inc. 的许可。

版权所有，未得书面许可，本书的任何部分和全部不得以任何形式重制。

尽管用先进的计算工具很容易从数据中找到规律，但是最深刻的认识还是来自于对这些规律来源的把握，而这可不能只通过计算机来完成。克莱因伯格巧妙地向读者介绍了寻找因果关系过程中所用到的主要概念和方法，思路清晰且内容实用，使得这本书不同于这一领域的其他著作。书中内容全面又易于理解，是科研领域的新人、经验丰富的专家以及其他想要从数据中获取更多认知的读者的必读之物。

——Andrew Therriault, 博士, 美国民主党
全国委员会数据科学部主任

哲学、经济学、统计学以及逻辑学都有志于理清因果关系，克莱因伯格成功地将这些完全不同的方法以一种简单而又实用的方式综合在了一起。随着我们的生活越来越多地“为数据所驱动”，要想理解国家政策、健康问题以及我们周围的世界，就必须对“从观察中推理因果关系”有一个清晰的思考。

——Chris Wiggins, 博士, 《纽约时报》首席数据科学家,
哥伦比亚大学副教授

因果关系是生活中的一个重要特征，但是人们对它有着大量争议和误解。本书在未借助任何先验知识或专业技术的前提下，对因果关系做出了清晰的解释，并且通俗易懂、妙趣横生，用严密的逻辑和深刻的分析来帮助我们理解复杂的概念。

——David Lagnado, 博士, 伦敦大学学院高级讲师

前 言

喝咖啡会使人长寿吗？是谁把流感传染给你的？股票价格为什么会上涨？无论是做饮食安排还是投资选择，抑或是责怪某人毁了你的周末，你都需要不断去了解其中的原因。正是这种因果关系在帮助我们预测未来，解释过去，让我们能够介入其中并对事物的变化产生影响。与流感病人接触会让你在一段时间内也染上流感，知道了这一事实，你就能知道自己会在什么时候出现流感症状。针对性很强的游说可以为你筹集政治竞选资金，了解了这一关系，你就可以将这些游说活动视为增加竞选资金的一个可行方案。高强度运动会导致高血糖，意识到这一点，你就可以帮助糖尿病患者控制血糖。

尽管推断因果关系这一技能非常重要，但你之前可能没有接触过这方面的课程。事实上，你可能都未曾静下心来想过，为什么某件事会成为另一件事发生的原因。虽然这个问题的答案涉及很多因素，但从根本上看，原因可以提高一个事件发生的概率，是产生某种结果的前提，或者是让某件事情发生的策略。但是，不能因为某种药物会引发心脏病，就认为某人的心脏病发作是由这种药物引起的。也不能因为某个地区在缩小班级规模以后，学生的成绩都得到了提高，就认为同样的做法在其他地区也会产生同样的效果。本书不仅要讨论在一切进展顺利的情况下可能会出现哪些

结果，还要研究为什么成功看似很难被复制。除此之外，我们还要考察那些在理论研究中经常被忽视的实际问题。

研究因果关系的方法有很多种（有些是互补的，有些是对立的），而且涉及众多领域（包括哲学、计算机科学、心理学、经济学、医学等）。我无意在这些争辩中选择立场，只想为读者呈现各种观点，厘清各种观点之间的共识与分歧。除此之外，我们还将探讨关于因果关系的心理学（人们是如何了解原因的）、如何进行因果关系的实验（以及这些实验的局限性是什么），以及如何根据因果关系来制定相应的策略（我们是否应该减少食物的含盐量，以此来预防高血压）。

我们首先要弄清楚什么是原因，以及为什么我们常常会弄错事情发生的原因（第 1~3 章）。然后，要认识到在原因的理解和运用方面，“什么时候”与“为什么”同等重要（第 4 章）。接着，要学习如何只通过观察就找到事情发生的原因（第 5 章）。

大型数据集可以让我们找到事情发生的原因，而不是简单用来检验我们的假设。但是我们必须认识到，并不是所有的数据都能用来推理事件发生的原因。在第 6 章，我们将考察数据特征对推理的影响。第 7 章将探索在可以做实验的情况下，如何去克服这些数据特征给我们带来的挑战。这里所说的实验可能是复杂的临床试验，也可能只是某人对自己不同锻炼计划所做的对比实验。通常情况与个别情况之间的差异，正是我们需要使用专业性策略对各种事件做出解释的原因（第 8 章讨论的内容）。但是，要想利用因果关系来制定有效的干预措施，如在菜单上提供食品热量信息来降低肥胖症发生的概率，就需要有更多的信息，而且很多干预措施还可能带来意想不到的后果（第 9 章将详述这一点）。本书将会告诉你为什么因果关系如此难找（比报纸文章告诉你的要更细致、更复杂），以及为什么尽管如此，它仍是一个相当重要且广泛适用的话题。

虽然困难重重，但也并非毫无希望。你将会形成一套基于原因的思

考体系：要问的问题、应引起怀疑的危险信号以及证实因果关系的方法。除了帮你找到事情发生的原因以外，本书还能帮你基于因果关系来做出决策、制定策略，并通过进一步测试来验证你找到的原因。

这本书是为普通读者而写的，我并未假定这些读者具有相关的背景知识。我唯一假定的是读者对因果关系充满好奇，我要让复杂的因果关系变得通俗易懂、广为人知。读完之后，我们会更加关注人们的直觉以及如何从概念上理解因果关系，而不是数学细节（实际上，本书不会介绍任何数学知识）。如果你是计算机科学或统计学博士，也许会学到一些新的工具并且很享受在其他领域的工作之旅，也可能会向往更多方法论方面的知识。不过，本书要研究的只是普通人眼中的因果关系。

目 录

前言	ix
第 1 章 引子	1
因果关系的概念以及寻找因果关系的方法从何而来?	
1.1 何为原因	5
1.2 怎样才能找到原因	10
1.3 为什么需要原因	14
1.4 接下来	19
第 2 章 心理	21
人们是如何寻找原因的?	
2.1 原因的寻找与使用	23
2.1.1 感知	24
2.1.2 推理与论证	27
2.2 责任的划分	34
2.3 文化	37
2.4 人的局限性	40
第 3 章 相关性	45
为什么有那么多因果关系被搞错?	
3.1 相关性是什么	48
3.1.1 没有变化就没有相关性	49
3.1.2 相关性的测量与解释	51

3.2	相关性的用途	58
3.3	为什么相关性不是因果关系	61
3.4	多重测试与 P 值	64
3.5	没有相关性的因果关系	69
第 4 章	时间	73
	时间如何影响我们感知因果关系和进行因果关系推理的能力？	
4.1	因果关系的感知	75
4.2	时间的方向性	83
4.3	当事物随着时间变化的时候	85
4.4	原因运用中的时间因素	91
4.5	时间可能具有误导性	93
第 5 章	观察法	98
	如何仅通过观察事物的运行方式来把握事件发生的原因？	
5.1	规律性	100
5.1.1	穆勒五法	100
5.1.2	各种复杂的原因	108
5.2	概率	110
5.2.1	为什么要使用概率	110
5.2.2	从概率到原因	113
5.3	辛普森悖论	119
5.4	反事实推理	122
5.5	观察法的局限性	127
第 6 章	计算法	130
	如何自动实现寻找原因的过程？	
6.1	假设	132
6.1.1	无隐藏的共同原因	133
6.1.2	典型分布	135

6.1.3 正确的变量	139
6.2 图解模型	140
6.2.1 图解模型在什么情况下会表示因果关系	144
6.2.2 从数据到图形	148
6.3 衡量因果关系	152
6.3.1 概率与因果关系的显著性	153
6.3.2 格兰杰因果关系检验法	158
6.4 现在该怎么办	161
第 7 章 实验法	164
如何通过对人和系统进行干预来寻找原因?	
7.1 从干预措施中获取原因	166
7.2 随机对照试验	168
7.2.1 为什么要做随机试验	169
7.2.2 如何设置对照组	175
7.2.3 研究结果适用于哪些人	178
7.3 当参与者只有你自己时应该怎么办	181
7.4 可再现性	184
7.5 机制	187
7.6 实验法是否足以找到事件发生的原因	189
第 8 章 解释	194
“这件事引起了那件事”这句话意味着什么?	
8.1 寻找某个事件发生的原因	197
8.1.1 出现多重原因时	197
8.1.2 解释可能具有主观性	200
8.1.3 原因出现的时间	202
8.2 具有不确定性的解释	204
8.3 将类型层面和实体层面分离来看	208
8.4 使解释过程自动化	211
8.5 法律活动中的因果关系	213
8.5.1 要不是因为	214
8.5.2 近因	216
8.5.3 陪审团	218

第 9 章 行动	224
如何根据原因进行决策？	
9.1 对因果假设的评估	227
9.1.1 强度	229
9.1.2 一致性（可重复性）	230
9.1.3 特异性	232
9.1.4 时间性	233
9.1.5 生物梯度	235
9.1.6 可信度与连贯性	236
9.1.7 实验	237
9.1.8 类比性	238
9.2 根据原因制定政策	239
9.2.1 背景	242
9.2.2 效力和效果	243
9.2.3 意外的结果	245
第 10 章 展望	250
为什么要研究因果关系？	
10.1 人们需要因果关系	250
10.2 主要原理	256
10.2.1 因果关系和相关性不是同义词	256
10.2.2 针对偏差的批判性思考	258
10.2.3 时间的重要性	259
10.2.4 并不是所有实验研究都比观察性研究好	260
10.3 一个百宝箱	261
10.4 知识的重要性	264
致谢	268
参考文献	269
关于作者	288

第1章 引子

因果关系的概念以及寻找因果关系的方法从何而来？

1999年，一个名叫 Sally Clark 的英国律师被法庭判定谋杀了她的两个孩子。1996年12月，她的第一个儿子在11周大的时候突然死亡。当时，人们认为孩子是自然死亡。但是就在第一个孩子夭折一年多以后，Clark 的第二个儿子又在8周大的时候死亡了。在这两个案件中，两个孩子似乎都没有什么生理上的疾病。于是，他们的突然死亡引起了人们的怀疑。

这两个案件有很多共同之处：孩子们死的时候年龄差不多，他们的死都是由 Clark 发现的，当时家里只有 Clark 和孩子在一起，而且验尸报告表明两个孩子身上都有伤。一开始，人们认为第一个孩子的伤是抢救时造成的。但是，第二个孩子死后，人们对孩子的伤重新做了检查，而这一次他们认为这些伤很可疑。第二个孩子死亡四周后，警方逮捕了孩子的父母，随后 Clark 被指控为谋杀罪并获刑。

同一个家庭的两个婴儿都死于婴儿猝死综合征（SIDS）的概率有多大？据英国检察官称，发生这种情况的可能性微乎其微，所以这两起死亡一定是谋杀所致。这一论据（一个原因的可能性几乎为零，所以一定还有另外一个原因）造成了这桩著名的冤案。这也是因为统计不当和忽视因果关系而造成严重后果的一个重要案例。

统计学家和因果关系研究者都知道这个案例，其主要原因在于，检

方的论据本质上基于这样的逻辑：被告的辩词几乎不可能为真，所以一定是假的。检方为此请来了一位专家证人——Roy Meadow 博士。Meadow 称，同一个家庭发生两起 SIDS 事件的概率为七千三百万分之一。检方由此认为，因为这个概率非常低，所以这两起死亡事件不可能出于自然原因，一定是谋杀所致。

然而，这一统计数据完全是错误的。即使这个数据是正确的，也不应该这样用。有一份研究报告估算出发生 SIDS 的概率为 $1/8543$ 。Meadow 根据这个研究报告提出：同一个家庭发生两起 SIDS 事件的概率为 $1/(8543 \times 8543)$ ，即约七千三百万分之一。¹ 这种计算方法错误的原因在于，它假定这些事件是相互独立的。抛硬币的时候，无论硬币落地时是正面朝上还是反面朝上，都不会对下一次结果产生任何影响。因为每一次硬币正面朝上的概率都是 $1/2$ ，所以将第一次正面朝上的概率与第二次正面朝上的概率相乘所得出的结果就是连续两次正面朝上的概率，这从数学角度来讲是没有问题的。Meadow 当时也是这样计算的。

引发 SIDS 的原因还不确定，但一个很重要的影响因素是孩子所处的环境（比如家里是否有人吸烟和饮酒）。这意味着，如果一个家庭发生过一起 SIDS，那么这个家庭发生第二起 SIDS 的概率就会远大于 $1/8543$ ，因为这些孩子的生活环境和遗传基因都是相同的。也就是说，第一起死亡事件会向我们透露第二起死亡事件发生的概率。这个案例和一名演员获得两次奥斯卡金像奖的情况非常相似。金像奖并不是随机颁发的，演员第一次获奖时具备的品质（才华、知名度、人脉）会提高他再次获奖的可能性。

这就是 Clark 一案的症结所在。在这个案子中，两起事件并不是相互独立的，可能还是共同的原因引发的。因此，不能通过简单的乘法来计算这两起事件发生的概率。相反，在计算第二起事件发生的概率时，应该考虑到第一起事件的发生。所以我们需要知道的是，在一个已经发生过一起 SIDS 事件的家庭中，发生第二起 SIDS 事件的概率。本案中的概率在计

算和使用方面都存在十分严重的问题，为此，被告在第一次上诉时请来了一位统计学家作为专家证人，皇家统计学会还专门写了一封信表达了他们对这个案件的关心。²

不过此案的问题并不仅仅是误算概率那么简单。在整个案件中，检方试图将事件（即这两起 SIDS 死亡事件）发生的这七千三百万分之一的概率等同于 Clark 无罪的概率。这种错误的推理将事件发生的概率当成了被告有罪或者无罪的概率，这就是我们所说的检察官谬误。³

我们知道，一个几乎不可能发生的事件真的发生了。一个家庭中发生两起 SIDS 死亡事件的可能性很小，但是一个家庭中两个婴儿都夭折的可能性也很小。人们不单单会考虑 SIDS 这一解释是否合理，更重要的是，他们会将其与关于这个事件的其他解释进行比较。因此在这个案件中，最好将同一个家庭中两个孩子都被谋杀（检方的假设）的可能性与同一家庭中两个孩子都患了 SIDS 的可能性进行比较。

一个家庭中两个孩子都死于 SIDS 的概率与这两个孩子都感染的概率是不同的。关于这个案件我们还有其他的证据，比如物证和犯罪动机等。必须将这些证据与概率结合起来看（比如说，一个人如果没有犯罪动机、没有作案机会或者没有行凶武器，那么他杀人的概率肯定要低于总谋杀率）。⁴

最后，无论一件事情发生的概率有多低，只要尝试的次数足够多，最后一定会发生。Clark 一案中，那个误算出来的极低的概率（七千三百万分之一）比中百万大博彩的概率（二亿五千八百万分之一）还要高三倍多。一个人中大奖的概率是极低的，但是如果我们说某个地方的某个人会中大奖，这个概率又如何呢？那就高得多了。这就说明，仅通过概率来判断一个人的清白一定会导致一些冤案。这是因为虽然对某个特定的家庭来说，发生这种事件的可能性很小，但是世界上有两个孩子的家庭有上百万个，这种事件总会在某个地方的某个家庭发生。

2003年1月，Clark第二次上诉时终于翻案。然而，那时她已经在监狱服刑三年了。

为什么Clark案会成为因果推理失败的重要案例呢？尽管此案在计算概率的过程中存在很多问题，但最根本的原因是，此案试图用一个事件发生的概率来支撑某个特定的因果结论。“这只是巧合而已”“这个概率有多大”，当你在说服别人相信某个因果关系时，是否也说过这样的话呢？生活中经常有这样的推理：公司来了一名新员工，而同一天你的订书机不见了；一名巫师知道你最喜欢的女性亲属的名字以“M”开头；两名重要人证记得那名嫌犯穿的是一件红色法兰绒衬衫。但是，如果因某件事情不大可能发生，而说其唯一合理的解释就是因果关系，那一定是错误的。前面已经说过，一个不大可能发生的事件在某个人身上发生的概率也许极低，但是在某个地方发生的概率却不低。除了会造成冤案以外，错误的因果推理还可能会带来其他严重的后果，比如将大量的时间和精力浪费在绝不可能起作用的药品上，或者制定一些无用的、代价高昂的公共政策。

本书的目的是提高读者的因果推理能力。严谨的因果思维是指质疑假设、衡量证据、分析各种说辞，以及辨别我们无法得知事情发生原因的情况。有时我们可能无法获得足够的信息来建立因果联系，有时我们获得的信息可能并不是我们所需要的，但重要的是能够认识到这些问题，并与其他人就这些问题进行交流。通过阅读本书，我希望读者至少能够对他们所听到的各种因果推论多一些质疑（我们将讨论在因果推论中需要注意哪些危险信号，以及可以提出哪些问题进行衡量这些推论），但首先我们会教大家如何寻找事件发生的原因、如何为因果关系提供强有力的证据，以及如何使用因果关系来指导我们日后的行为。