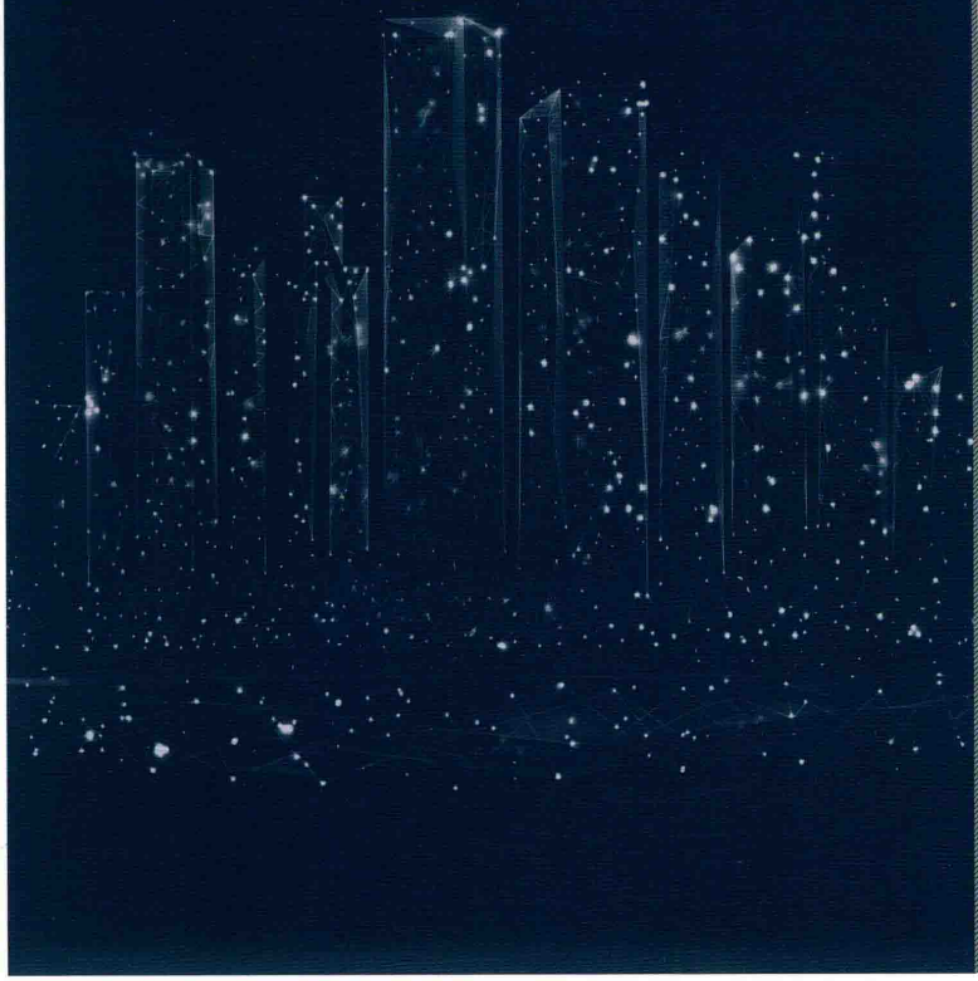
徐笛◎著

数据新闻的兴起

场域视角的解读



中国传媒大学出版社

数据新闻的兴起

场域视角的解读

徐 笛◎著

中国传媒大学出版社
·北京·

图书在版编目(CIP)数据

数据新闻的兴起：场域视角的解读/徐笛著．--北京：中国传媒大学出版社，2019.5
ISBN 978-7-5657-2424-4

Ⅰ.①数… Ⅱ.①徐… Ⅲ.①数据处理—应用—新闻报道—研究—中国
Ⅳ.①G219.2

中国版本图书馆 CIP 数据核字(2018)第 267875 号

数据新闻的兴起：场域视角的解读
SHUJU XINWEN DE XINGQI：CHANGYU SHIJIAO DE JIEDU

著　　者	徐　笛
策划编辑	姜颖昳
责任编辑	姜颖昳
装帧设计	拓美设计
责任印制	阳金洲

出版发行	中国传媒大学出版社	
社　　址	北京市朝阳区定福庄东街 1 号	邮编：100024
电　　话	86-10-65450532　65450528	传真：010-65779405
网　　址	http://www.cucp.com.cn	
经　　销	全国新华书店	

印　　刷	北京中科印刷有限公司
开　　本	787mm×1092mm　1/16
印　　张	11.75
字　　数	150 千字
版　　次	2019 年 5 月第 1 版
印　　次	2019 年 5 月第 1 次印刷

书　　号 ISBN 978-7-5657-2424-4/G・2424　**定　　价** 48.00 元

前　言

本书付梓成书之时，2018年全球数据新闻奖评选结果亦出炉。中国财新传媒数据与可视化实验室荣膺全球最佳大型数据团队奖，击败了英国广播公司、美联社、法新社、《卫报》等精英媒体的数据新闻团队。

数据新闻作为舶来品，进入中国内地之时，恰逢内地媒体深陷危机的时刻。纸媒关停并转，客户端、移动端导流乏力，而作为新闻业场域内的后来者，数据新闻经历了迅猛发展，从业者感叹，“只有在数据新闻领域，中国媒体能与国外媒体并驾齐驱。”本书即聚焦于这段历史时刻，从场域框架视角出发，通过问卷调查、深度访谈和行动研究所获得的一手实证性材料，审视数据新闻的兴起，谁在推动场域的生成，分析场域内的作用力量以及主导规则。

书中详细描画了数据新闻从业者的构成、价值认知、工作状态，以及数据新闻的生产流程，这些内容可作为数据新闻爱好者的入门参考。作者亦同时承担数据新闻教学工作，在写作本书过程中，受到研究结果的启发，几番修订教学内容，本书如能为教育者批判之借鉴，作者自当欣喜。同时本书基于场域理论框架分析数据新闻，也可供研究者作为进一步研究的参考文献，但求以一得之见抛砖引玉。

在本书写作过程中，作者得到诸多师友鼓励启发，此处无法一一言谢。特别感谢研究助理、复旦大学新闻学院本科生欧杨洲在写作过程中的帮助，她承担了资料搜集、整理的大量工作。书中所有纰漏都由作者承担，恳请读者谅解。

本书得到上海市哲学与社会科学基金（WBH3353016）以及复旦大学“网络数据挖掘研究”创新团队经费支持。

目 录

绪　论

本章，我们从英国《卫报》对伦敦骚乱的数据新闻报道出发，介绍数据新闻产生的时代背景，即我们所处的大数据时代。接着通过对“大数据”一词的溯源及意涵的解释，审视大数据给人们的思维方式和新闻业带来的影响。本章结尾简要概述了本书的架构及各章主要内容。

一、大数据时代的新闻业

2011 年 8 月，英国伦敦发生骚乱，持续五天的骚乱震惊了世界，英国《卫报》数据博客对骚乱的报道也震撼了新闻界。

骚乱起因于警察击毙了一名 29 岁的黑人男青年，这位名为马克·达根(Mark Duggan)的青年被怀疑非法持有枪械。2011 年 8 月 6 日，他的家人到警局门口举行和平抗议活动，要求警方公布执法详情。随后抗议升级为大规模骚乱，骚乱从伦敦北部街区扩散至整个伦敦，接着蔓延至曼彻斯特、利物浦等地。骚乱中，有 2 278 家商店被破坏或洗劫，共发生了 5 112 起与骚乱相关的犯罪事件，超过 4 000 人被捕。①

各媒体实时跟进报道了骚乱进展。按照传统的新闻操作方式，媒体一方面会派记者前往现场，借由“在场”彰显报道的客观性；另一方面，也会采访官方发言人，由此强调报道的权威性。《卫报》数据博客在推出实时报道的同时，

① ROGERS S. Data journalism reading the riots: what we know, and what we don't [EB/OL]. The Guardian, (2011 - 12 - 09)[2017 - 12 - 15].

独辟蹊径，通过数据新闻，展现出了与官方话语迥异的另外一番图景。

骚乱在蔓延，而官方无法提供骚乱发生地点的完整信息，于是《卫报》数据博客栏目推出了骚乱地图报道，在谷歌地图上标注已知的骚乱发生地点，并提供现场详情报道，同时邀请读者查缺补漏。读者可以上传骚乱信息，还可以校正骚乱地图中的错误，所有的数据都可以自行下载。骚乱发生第 2 天，数据博客又推出了在线问卷，请读者选填骚乱的原因。

骚乱逐渐平息后，1 984 人接受了法庭审讯，法庭审讯记录里包含了全部 18 岁以上嫌疑人的身份信息、家庭住址、涉及罪名等，法庭记者很快拿到了庭审记录摘要。但跑口记者只关心那些特殊、典型的案例，而数据博客却更想知道全部信息：受审者是谁？他们从哪里来？为什么参与其中？于是数据博客向法院提起信息公开申请，并在几经周折后终于拿到了以 PDF 格式存储的开庭记录。这种格式的文档无法直接分析，经过人工转录，数据博客最终自制了一份含有一千多条庭审记录的数据库。数据博客，对数据库进行数据分析后发现，法庭倾向于重判参加骚乱的犯罪嫌疑人，他们的刑期比其他类似罪行的人刑期平均长了四分之一，[①]而这些都是仅靠采访无法获得的信息。

骚乱报道最核心的问题，无外乎探究原因，政客和评论家们对此众说纷纭。时任首相卡梅隆认为，骚乱与贫穷无关。而数据博客把涉嫌骚乱的犯罪嫌疑人的家庭住址叠加在反应贫穷程度的地图上，显而易见，二者有较强的相关性。骚乱中，警方指责 Twitter 和 Facebook 等社交媒体散播谣言，助长了骚乱的蔓延，警方甚至一度考虑短时关闭社交媒体。那么社交媒体在事件中到底起到了何种作用？数据博客与伦敦政治经济学院的学者们合作，分析了几条典型谣言在 Twitter 上的传播过程，他们抓取了 Twitter

① ROGERS S. Data journalism reading the riots: what we know, and what we don't [EB/OL]. The Guardian, (2011 - 12 - 09)[2017 - 12 - 15].

上有关骚乱的257万条推文进行分析，其发现与警方的断定南辕北辙，实际上，大量的推文在澄清谣言，并提供了应对骚乱和洗劫的建议(图1)。[①]

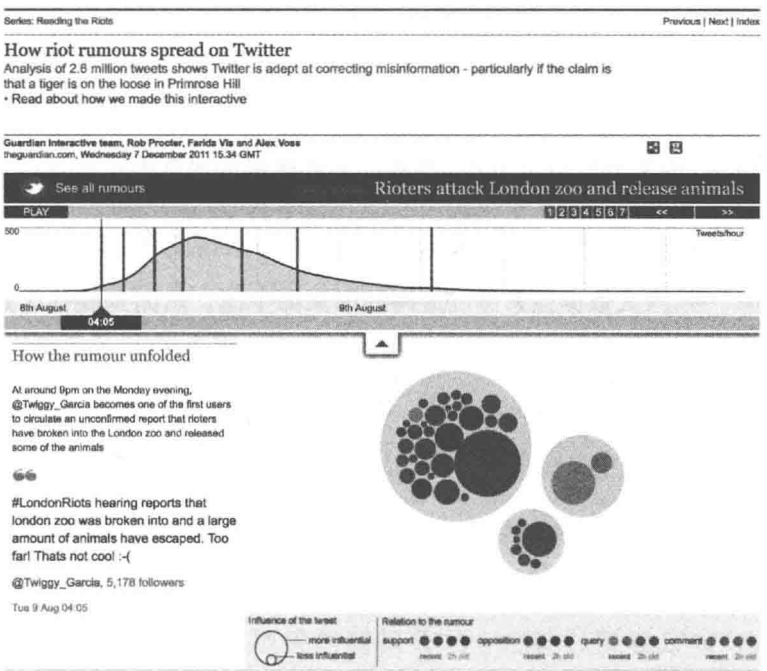


图1 《卫报》数据博客刊载的Twitter上有关骚乱的谣言传播路径图[②]

1981年，英国同样发生了类似的骚乱，斯卡曼勋爵(Lord Scarman)领衔的调查委员会对骚乱原因做了深入剖析，其调查结果对社会政策产生了深远影响。而2011年骚乱后，官方并未进行系统调查，但有关骚乱的数据新闻报道为我们揭示了更为丰富、全面的事件信息，也提供了有别于官方话语的图景，这些报道基于可供验证的数据而非个人判定。[③] 20年前，没有社交媒体，无法获取地理位置信息，也没有丰富的数据，今天，

① ROGERS S. Data journalism reading the riots: what we know, and what we don't [EB/OL]. The Guardian, (2011-12-09)[2017-12-15].

② Reading the riots [EB/OL]. https://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter(2011-12-07)[2018-03-02].

③ ROGERS S. Data journalism reading the riots: what we know, and what we don't [EB/OL]. The Guardian, (2011-12-09)[2017-12-16].

我们所处的“大数据时代”孕育了新闻生产更多的可能性，数据新闻正是其中之一。理解“大数据时代”是理解数据新闻的起点，以下简要阐述何为大数据。

（一）何为大数据

大数据”（big data）频繁出现在各种新闻报道、行业报告以及学术论文中。谷歌搜索趋势显示，“大数据”一词的搜索热度从 2011 年开始提升，2017 年，对大数据搜索热度最高的国家即中国。然而到底什么是“大数据”，学术界至今莫衷一是，我们先从追溯大数据的早期使用开始，探究何为大数据（图 2、图 3）。

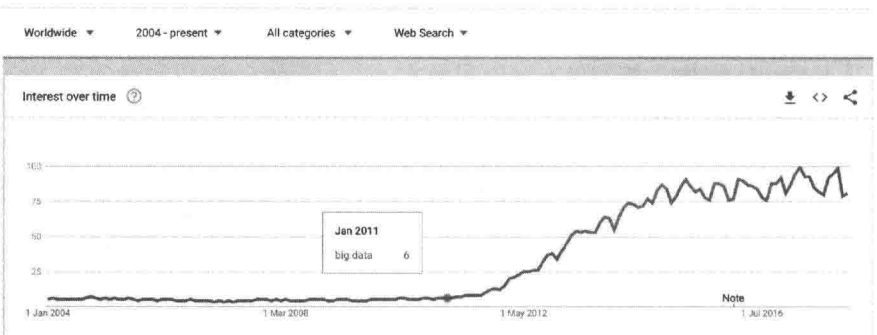


图 2 Google 趋势显示的大数据（big data）一词的搜索趋势①

图 3 Google 趋势显示大数据一词在不同国家的搜索热度②

①② Google trends[EB/OL]. https://trends.google.com/trends/explore?q=big%20data [2018-01-30].

1. 大数据溯源

常见的定义方法认为，大数据是“大小已经超出了传统意义上的尺度，一般的软件工具难以捕捉、存储、管理和分析的数据”[①]。这个定义被广泛使用，但它抹杀了大数据作为一种社会现象所凝结的人类社会文化、经济、科技等方面发生的深刻变革。诚然，在社会科学研究中，术语的定义通常充满争议，更有价值的路径是探究这个概念的生成历史，研究它如何被使用，追本溯源，可厘清社会现象背后不同力量的交织作用。

有学者认为，大数据一词最早源自经济学领域，华尔街的商业分析和经济学中的建模孕育出大数据的概念。宾夕法尼亚大学的经济学家迪耶伯德(Francis X. Diebold)在一篇宏观经济分析的论文中使用了“大数据”一词，这篇文章成文于 2000 年，发表于 2003 年，文中提出使用大数据分析的方法衡量和预测宏观经济。[②] 跳出经济领域，实际上，早在 20 世纪 90 年代中期，硅谷一家顶尖的科技公司——硅谷图表公司(Silicon Graphics Inc.，SGI)——的午餐会上就曾反复讨论过“大数据”。[③] 该公司的首席科学家约翰·R. 马西(John R. Mashey)在多个场合发表演讲，提出随着数据量飞速增长，以及数据类型日益多样，用户的期待也水涨船高，科技公司需要提升基础设施以应对爆炸性增长的数据。[④] 随后，1998

① 涂子沛．大数据：正在到来的数据革命，以及它如何改变政府、商业与我们的生活[M]．桂林：广西师范大学出版社，2015：57.

② DIEBOLD F. “Big Data”Dynamic factor models for macroeconomic measurement and forecasting：A discussion of the papers by Lucrezia Reichlin and by Mark W. Watson [G]// Dewatripont M.，Hansen L.，& Turnovsky S. Advances in economics and econometrics：theory and applications，eighth world congress (Econometric Society Monographs). Cambridge：Cambridge University Press，2003：115 - 122.

③ LOHR S. The origins of ‘Big Data’：an etymological detective story [EB/OL]. (2013 - 02 - 01)[2017 - 12 - 18]. https：//bits. blogs. nytimes. com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/.

④ MASHEY R J. Big data and the next wave of infrastress [EB/OL]. (1998 - 04 - 25)[2017 - 12 - 18]. http：//static. usenix. org/event/usenix99/invited_talks/mashey. pdf.

年，两位计算机科学学者在他们出版的图书《预测性数据挖掘：实践指南》（*Predictive Data Mining：A Practical Guide*）中也提出了大数据的概念。这正是大数据一词的早期使用，迪耶伯德认为，大数据最早便横跨计算机科学、统计学与计量经济学等多个学科领域。①

2. 大数据的意涵

大数据一词的流行离不开不同主体围绕它生产的大量话语，其中最为积极的主体莫过于经济和商业力量。2011 年，麦肯锡全球研究院推出了 156 页的报告——《大数据：下一个创新、竞争与提高生产力的前沿》，提出数据的爆炸性增长已将人类社会带入到大数据时代。② 根据统计，“1998 年全球网民平均每月使用流量是 1MB(兆字节)，2000 年是 10MB，2003 年是 100MB，2008 年是 1GB(1GB 等于 1 024MB)，2014 年将是 10GB。全网流量累计达到 1EB(即 10 亿 GB)的时间在 2001 年是一年，在 2004 年是一个月，在 2007 年是一周，而在 2013 年仅需一天，即一天产生的信息量可刻满 1.88 亿张 DVD 光盘。”③数据已渗透到每一个行业和多个业务领域，成为重要的生产力要素，并将成为未来商业竞争的基础。麦肯锡的报告提出，大数据将从 5 个方面创造新价值：透明性；发现需求，提升服务；人群细分，精准定制个体需求；通过算法替代或支持人类决策；创新商业模式、产品和服务。④ 越来越多的公司将数据业务作为其最新的增长点。与商业力量并驾齐驱，科技力量是另一股积极推广大数据的主体，物联网、微型传感器、云存储等技术的飞速发展，使得大数据的储存

① DIEBOLD F. A personal perspective on the origin(s)and development of“Big Data”：the phenomenon，the term，and the discipline [EB/OL]. (2012－11－26)[2017－12－18]. http://www. ssc. upenn. edu/～fdiebold/papers/paper112/Diebold_Big_Data. pdf.

② McKinsey Global Institute. Big data：the next frontier for innovation，competition，and productivity[R]. Seoul，San Francisco，London & Washington，2011.

③ 邬贺铨．大数据时代的机遇与挑战[J]. 求是．2013(4)：47.

④ McKinsey Global Institute. Big data：the next frontier for innovation，competition，and productivity[R]. Seoul，San Francisco，London & Washington，2011.

和分析成为可能。大数据将改变社会运行方式，基于海量数据，计算机可以帮助人类做出更好的决定，近年来，类似的论调不绝于耳。

大数据不仅是经济、科技现象，也是一种社会文化现象。路易斯(Seth Lewis)和韦斯特兰(Oscar Westlund)提出，作为社会文化现象的大数据主要受到三种动力机制的相互形塑影响：

● 技术层面：大数据的运用可以最大限度地提升计算能力和算法精度，以收集、分析、链接和比较大型的数据集。

● 分析层面：利用大型数据集挖掘模式，以做出经济、社会、技术或法律上的判断。

● 产生迷思：人们普遍认为，大数据带有真理、客观与准确的光环，它能够提供更高层级的智能和知识，能够产生此前人类无法获知的洞见。[①]

围绕大数据产生的迷思尤为值得社会科学研究者关注。这意味着研究者应跳出大数据有多“大”的迷思，转而去关注它如何宣称“大”以及如何使用“大”。吉莱斯皮对“平台”一词的研究也同样适用于大数据研究。有学者研究认为，这些语汇并非凭空产生，“它实际是由抱有明确目的的利益相关方，在可资选用的文化词汇表中精心雕琢后生成的词汇，它可以面向特定的群体产生特定的回响。这些话语努力不仅是为推销、说服、劝说、保护、战胜或谴责，而是要宣称这些技术是什么或不是什么，以及从技术中应该期待什么和不该期待什么。换句话说，这些话语意图确立的是衡量技术的标尺”[②]。而这个标尺会限制我们对社会现象的理解，会让我们按照话语生产者所期待的方向来理解社会现象，而拆解这个标尺的确立过程也是祛魅的过程。秉承这样的立场，本书在分析中始终谨慎地使用“大数据”一词，避免成为鼓吹者，而是更多关注大数据背后的驱动力量

---

① LEWIS C S & WESTLUND O. Big data and journalism: epistemology, expertise, economics, and ethics[J]. Digital Journalism, 2015, 3(3): 449.

② GILLESPIE T. The politics of 'platforms'[J]. New Media & Society, 2010, 12(3): 359.

以及话语努力。下文我们将从更具操作性的层面探讨大数据的特点，并据此理解大数据对方式方式带来的变革。

3. 大数据与思维变革

为将大数据概念与较大量的数据区分开来，计算机专业人士用 5 个“V”来界定大数据的特征：

- 海量(Volume)：数据量巨大，超过了传统软件的储存处理能力；
- 多样(Variety)：数据类型多样，包含文本、视频、音频、地理位置信息等多种类型的数据；
- 高速(Velocity)：高速产生甚至近乎实时产生的数据；
- 真实(Veracity)：数据质量较高，确保真实性；
- 价值(Value)：大数据需要跨学科、跨领域协作处理，以挖掘数据的多样价值。①

这套定义标准最早由 IBM 公司数据分析团队依据电子商务的特点加工后提出，随后计算机行业人士又进行了讨论修改，目前以 5 个“V”来定义大数据的特征已被广泛接受，这也说明科技、商业力量在阐释大数据现象时具有较多话语权。

在上述特征的限定下，大数据不仅指数量较大的数据，更意味着数据处理和分析方式的变革，这也意味着我们需要变革思维方式以挖掘数据的多样价值。②

首先，大数据不再需要随机抽样的方法，而是采用全部数据来做分析，也就是说样本＝总体。

其次，大数据不再也无法追求精确性，精确性是小数据时代的产物，

---

① YIN S & KAYNAK O. Big data for modern industry: challenges and trends [C]. Proceedings of the IEEE, February 2015, 103(2): 144 - 145.

② 迈尔-舍恩伯格，库克耶．大数据时代：生活、工作与思维的大变革[M]．盛杨燕，周涛，译．杭州：浙江人民出版社，2013：28 - 94.

在信息匮乏的时代，缺失任意一点的数据，都可能会导致结果的偏差。在信息爆炸的时代，数据量达到某个值以后，其边际效用会降低，由此大数据可以不再追求精确性。而且海量、高速和多样的数据也带来混杂性，接受大数据的混杂性反而能获得更多信息。

最后，大数据不再强调因果关系，转而探究相关关系。通过大数据分析，找出关联物，通过监测关联物就可以预测未来，所以“建立在相关关系分析法基础上的预测是大数据的核心”[①]。作为社会子系统之一的新闻业不可避免地受到大数据的影响与冲击，以下将就此展开论述。

（二）大数据与新闻业

使用数据对新闻业来说并不是什么新鲜事，《卫报》数据博客的前主编西蒙·罗杰斯（Simon Rogers）曾将《卫报》使用数据的历史追溯至1821年5月5日该报创刊号上刊登了有关学校教育的统计数据。[②] 虽然彼时的数据与今日之大数据相去甚远，但这至少说明新闻业从未离开过数据。不同的是，在大数据时代，新闻与数据间的联结更为紧密和复杂，有学者提出，大数据对新闻业最根本的改变来源于改变了新闻的认识论基础。

新闻业是现代社会重要的知识生产机构，新闻业的认识论是指新闻业如何认定何为知识，什么又是真实、合法的知识。一般来说，新闻业依据一定规则、成规和制度化程序展开知识生产实践，[③]比如新闻价值判断决定了什么是新闻业认可的知识，而遵循客观性和信源交叉检验规则的

---

① 迈尔-舍恩伯格，库克耶. 大数据时代：生活、工作与思维的大变革[M]. 盛杨燕，周涛，译. 杭州：浙江人民出版社，2013：75.

② 罗杰斯. 数据新闻大趋势：释放可视化报道的力量[M]. 岳跃，译. 北京：中国人民大学出版社，2015：52.

③ EKSTRÖM M. Epistemologies of TV journalism：a theoretical framework [J]. Journalism，2002，3(3)：259－282.

新闻业可以宣称这些知识是真实的，由此主张自己作为知识生产机构的合法性。价值判断、规范、成规等构成了新闻业的认识论基础，而大数据正在重塑这个基础。我们借用麦茨·艾克斯特罗姆（Mats Ekström）的研究成果，从三个维度分析本书开篇提到的《卫报》伦敦骚乱报道，以审视大数据对新闻业的改造。

艾克斯特罗姆在研究电视新闻时，提出新闻业的认识论可分为三个组成部分：[①]

- 知识的形式：即与媒介类型相关的知识的形式，以及这种知识的特征；
- 知识的生产：生产知识所遵循的专业规范或常规；
- 知识的接收：知识被公众接受或拒绝的决定性条件。

我们将这个理论框架简化为更具可操作性的衡量标准。其中，知识的形式简化为生产资料，即用来生产新闻的原始材料；知识的生产简化为生产方式，即对生产材料的处理方法；知识的接收简化为受众的接收方式。并据此分析了有关骚乱的数据新闻报道，详见表1。

表1 《卫报》骚乱报道分析

报道主题	生产资料	生产方式	接收方式
骚乱地图	骚乱发生地点的地理位置信息	原始数据众筹、开放	参与
骚乱原因	邀请读者填答的问卷结果	自制数据、数据开放	参与
谣言传播	社交媒体数据	与大学协作	获取
法庭审判	转化为电子格式的法庭审判记录	自制数据、数据开放	获取
骚乱与贫穷	法庭审判数据、反映贫穷程度的官方统计数据	自制数据、连接外部数据、数据开放	获取

从表1中可以看出，数据已成为生产资料的核心，并且数据被认为更

① EKSTRÖM M. Epistemologies of TV journalism: a theoretical framework [J]. Journalism, 2002, 3(3): 259.