# 大数据时代的统计挑战与应对

文 / 程开明 陈龙

随着互联网、物联网、无线传感器、云计算等快速发展，全球数据量出现爆炸式增长，人类社会进入了一个以太字节为单位的大数据时代。每一天，无数的数据被收集、交换、分析和整合，数据如一股“洪流”注入这个世界。

大数据观点一经提出，便引起了全球广泛的反响，似乎所有的商业或组织活动都可视为大数据问题。譬如，当今的制造业，大多数机器上都安装有一个或多个微处理器，已进入大数据的状态；消费零售行业，无数顾客的交易触点和网上点击的流量，也构成了大数据；谷歌甚至认为无人驾驶汽车也是大数据的问题。

各国政府也逐渐认识到其拥有的海量数据有待于发掘，美国、加拿大、新西兰、德国和法国等国家先后推出自己的政府公共数据开放网站，为大数据敞开了大门。亚洲国家的政府也出现了基于大数据战略的数据分析方案和倡议，2011 年新加坡经济发展委员会资助成立了德勤数据分析研究所，目标是引领政府和企业对于大数据的研究和应用。

## 一、大数据的核心特征——数据之“大”

大数据是指那些大小超出了传统意义的尺度，一般软件工具难以捕捉、存储、管理和分析的数据。除了包括大量的结构化数据外，还涵盖所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频、视频等信息，即非结构化数据。

如今，只需要一些想象力，万千事物都可通过数据化工具转化为“数据”，进而带给人们意想不到的惊喜。数据化的构思是许多社交网络公司的脊梁，腾讯 QQ 平台不仅提供了寻找和维持朋友、同事关系的场所，而且将日常生活的无形元素提取出来，转化为数据。微博，让人们能轻易记录以及分享其零散的想法，“那些曾被遗忘在时光中的碎片，使情绪数据化得以实现”。

一般认为，大数据的数量级应该是“太字节”($2^{40}$)。为了理解“太字节”，举个例子：美国国会图书馆所有登记的印刷版书本的信息量为 15 太字节。但有一点必须清楚，随着技术的进步，这个衡量“什么是大”的数量级本身还在不断增长。大数据之“大”，并不仅仅在于其“容量之大”，更多的意义在于：人类可以“分析和使用”的数据在大量增加，通过这些数据的整合和分析，人类可以发现新的知识，创造新的价值，带来“大知识”、“大科学”和“大发展”。

通常，大数据具有以下几个特点：1. 数据量相当庞大。例如，淘宝目前每天的活跃数据量已经超过 50 太字节，共有 4 亿条产品信息和 2 亿多名注册用户在上面活动，每天超过 4000 万人次访问。2. 数据增长速度快。大数据的产生速率非常快，移动互联网、物联网无时无刻都在生成海量电子数据，例如亚马逊每天都在产生和发布数以百万计的商品和服务交易价格信息。3. 数据的多样性。数据的格式多种多样，包括网页、图片、视频、音频等。4. 不稳定性，即数据的来源和数据量不稳定。大数据生成和存储的管理者、数据的使用者五花八门，其生成和存储不一定具有法律约束力，今天有需求就有记录，明天没需求就可以停下来不记录，因而这些数据的数据量可能变化无常，容易产生时间序列或数据内容等方面的缺失，连续性或可持续性难以保证。

表 1　数据的存储单位

| 单位 | 英语标识 | 大 小 | 含义和例子 |
|---|---|---|---|
| 字 节 | Byte | 8 位 | 这是计算机存储信息的基本物理单位，存储一个英文字母在计算机上，其大小就是一个字节 |
| 千字节 | KB | 1024 字节，或 $2^{10}$ 个字节 | 一页纸上的文字大概是 5 千字节 |
| 兆字节 | MB | 1024 千字节，或 $2^{20}$ 个字节 | 一首普通 MP3 格式的流行歌曲大概是 4 兆字节 |
| 吉字节 | GB | 1024 兆字节，或 $2^{30}$ 个字节 | 一部电影大概是 1 吉字节 |
| 太字节 | TB | 1024 吉字节，或 $2^{40}$ 个字节 | 美国国会图书馆所有登记的印刷版书本的信息量为 15 太字节 |
| 拍字节 | PB | 1024 太字节，或 $2^{50}$ 个字节 | 美国邮政局一年处理的信件大约为 5 拍，谷歌每小时处理的数据为 1 拍 |
| 艾字节 | EB | 1024 拍字节，或 $2^{60}$ 个字节 | 相当于 13 亿中国人人手一本 500 页的书加起来的信息量 |
| 泽字节 | ZB | 1024 艾字节，或 $2^{70}$ 个字节 | 截至 2010 年，人类拥有的信息总量大概是 1.2 泽字节 |
| 尧字节 | YB | 1024 泽字节，或 $2^{80}$ 个字节 | 超出想象，难以描述 |

## 二、大数据的深刻影响——统计挑战

大数据时代的到来，必然对社会经济各个方面产生重大冲击，对与大数据紧密相关的“统计”又会产生什么样的影响呢？

（一）引致统计思维的转变

统计学是关于数据的科学，即研究如何收集、整理和分析数据的科学。数据是依据，是灵魂，是统计方法生命力的根源所在，大数据时代的统计首先要适应三个重大的思维转变。

1. 分析与事物相关的所有数据，而不是依靠分析少量的样本数据。统计往往希望用尽可能少的数据来证实可能重大的发现、假设等，小数据时代一般采用随机采样，用最少的数据获得最多的信息。统计抽样是在技术受限的条件下，解决当时存在的一些问题而产生的；如今的大数据时代，计算和制表不再像过去那样困难，感应器、手机导航、网站点击和微博等能够收集大量数据，而计算机也能够轻易处理。因此，在处理大数据时不再采用随机抽样的方法，而利用所有数据进行分析。例如：谷歌流感趋势预测并不是依赖于对随机抽样的分析，而是分析了整个美国几十亿条互联网检索记录而得到的结论。分析整个数据库，而不是对一个样本进行分析，能够提高微观层面分析的准确性，甚至能够推测出任何特定尺度的数据特征。

2. 乐于接受数据的纷繁复杂，而不再追求精确性。对小数据而言，最基本、最重要的要求是减少误差，保证数据质量。生活于信息时代的我们，掌握的数据越来越全面，不再只包括手头现象的一点点可怜数据，而是包括了与之相关的大量级数据甚至全部数据。人们不再那么担心某个数据点对整套分析的不利影响，要做的是接受纷繁的数据并从中受益。大数据要求人们能够接受混乱和允许不精确性，例如一个小商店晚上打烊的时候要把收银台里的每分钱都数清楚，但如果用“分”这个单位来精确计算国内生产总值显然不适用。大数据时代，随着数据规模的扩大，人们对数据精确度的痴迷将逐步减弱。

3. 不再探求难以捉摸的因果关系，转而关注事物的相关关系。在小数据时代，人们往往乐此不疲地想知道现象背后的原因。大数据时代，由于坐拥海量数据和良好的机器计算能力，相关关系分析为人们提供了一系列新的视野和有用的预测，能够找出新种类数据间的相互联系来解决日常需要。例如：如果电子医疗记录显示橙汁和阿司匹林的特定组合可以治疗癌症，那么找出具体的致病原因就没有通过相关关系而获得的这种治疗方法来得重要；亚马逊根据用户在其网站上的类似查询来进行产品推荐，也是大数据相关关系的典型应用。通过探求“是什么”而不是“为什么”，能够帮助人们更好地了解这个世界。

（二）推动统计业务的变革

大数据对统计业务及统计工作过程也会产生深远的影响。具体而言，大数据对统计业务的影响主要包括：

1. 统计数据收集。数据收集是政府统计的主要任务之一，数据来源包括业务工作的管理数据、民意社情的调查数据、社会经济监测数据等。政府统计数据的收集工作，需要公民和社会的配合，对公民和社会而言是一种负担，政府相关部门必须尽量控制这一社会负担，减少“信息扰民”。因此，政府统计部门应避免重复收集，确保收集信息的方式简洁有效，尽量减少普通公民和社会组织的信息填报负担。大数据时代，由于无线传感器的快速普及，记录性数据快速增长，很多数据在相当程度上不用再向公众或机构调查收集，能够极大地减少调查负担，必将成为政府数据的重要来源。

2. 统计数据整理。由于大数据具有数据量大和多样性的特点，除了传统格式的二维结构化数据外，还包括图片、音频、视频等非结构化数据，传统的统计分组、频数分布等数据整理方法显然难以完全适用。数据仓库是对海量数据进行分析的核心物理构架，成为大数据环境下数据整理和存储的依托。数据仓库能为大数据的整合提供有效途径，可以对运行平台不同、编制语言不同、所处物理位置不同的海量数据按统一定义的格式进行提取，再通过清洗、转换、集成，最后百流归海，加载进入数据仓库。

3. 统计数据分析。大数据的爆炸不局限于政府部门，还包括制造业、新闻业、银行业和证券投资业等。由于需要处理的数据量日益庞大，使用传统的统计数据分析方法往往效率低下，呼唤新的方法对大数据进行分析。应运而生的大数据处理技术主要包括联机分析、数据挖掘和数据可视化等，数据收集、编译、链接和分析系统等也不断发展和完善，由此产生的大量需求正在互相促进，互相影响，形成汹涌的“大数据”浪潮。

类似于小数据时代的面板数据，大数据时代的联机分析（也称多维分析）把分立的数据库“相连”进行多维度的分析。“维”是联机分析的核心概念，是人们观察事物、计算事物的特定角度，譬如沃尔玛如果要分析自己的销售量，可以按时间顺序、地区国别进行分析，也可以按进货渠道、客户群体等进行分析，不同的分析角度就构成“维度”。随着维度的不断增加，问题可能变得异常复杂，一旦超过三个维度，人类的思维和想象力往往受到很大限制。而联机分析的惊艳之美在于可根据统计分析的需要随时创建“万维”动态报表，从不同的维度、不同的粒度，对海量数据进行分析，从而获得全面、动态、可随时加总或细分的分析结果。

数据挖掘是针对海量数据通过分析和建模，发现数据背后隐藏的模式和微妙关系，以揭示过往规律、预测未来趋势的分析方法。之所以称为“挖掘”，是因为面临海量数据，要从中寻找信息和知识就像开矿掘金一样困难。传统的数据挖掘是指在结构化数据当中发现潜在关系和规律，在大数据时代如何把散布在各领域的非结构化数据整合起来，并从中挖掘出有价

值的信息和知识，成为当前数据挖掘面临的最大挑战之一。如果说联机分析是对数据的一种透视性探讨，数据挖掘则是对数据进行掘山凿矿式的开采，前者为描述性分析，后者为预测性分析。

数据的爆炸性增长使人们急需能够有效展示数据、理解数据、演绎数据的工具，由此刺激了数据可视化方法的创新。数据可视化是指以图形、图像、地图、动画等生动、易理解的方式来展示数据，诠释数据之间的关系和发展的趋势。数据可视化可以通过图像在逻辑思维的基础上进一步激发人的形象思维和空间想象能力，吸引、帮助数据分析者洞察数据之间隐藏的关系和规律。新兴数据可视化方法通过把复杂的数据转化为直观的图形，并呈现给最普通的大众，使复杂的海量数据不再是少数高级分析人员的专利，更加贴近大众生活。

4. 统计数据发布。政府部门收集的数据应无偿地与其他部门共享，如无法律明禁，还必须向全社会发布。欧洲一些国家利用发布的数据和信息收取费用，看起来是为政府开源创收，但由于其对信息的封闭和垄断，对整个社会的经济发展其实是弊大于利。数据发布最方便和最有利的方式是政府网站，由于这种方式的开放性，将有利于促进全社会的充分利用。对于政府所发布的数据，其获得的方式、产生的方法必须透明，所以同时要发布一系列文档来说明数据来源、产生的方法，以及用户复制过程中可能出现的问题和错误。

5. 数据质量。数据质量是一个涉及数据收集、整理、分析和发布等全过程的问题，重要性毋庸置疑。在大数据时代，政府于网上发布海量数据，所面对的不是一个特定群体，而是全体国民，传播时间大大缩短，对社会各个领域都可能产生重大影响，因此必须保证质量。大数据时代的数据质量除遵循传统的质量准则外，还需遵循以下几个指导原则：质量要有统一标准，即客观性、实用性和完整性；完善质量管理的流程，防止低质量的数据出现；建立数据质量应急及救助机制，应对社会公众对于数据质量的质疑和挑战，若数据质量确实存在问题，必须加以补救。

## 三、大数据的未来策略——统计应对

未来社会的竞争不是劳动生产率的竞争，而是知识生产率的竞争。数据是信息的载体，是知识的源泉，能够创造极大的价值和利润，未来基于知识的竞争将集中表现于数据竞争。而未来的数据竞争又将是大数据的竞争，大数据的收集、整理、分析和发布能力将成为关键之所在，某种程度上可以说是得大数据者得天下。

由于大数据具有数据量大、增长速度快、多样化和不稳定性等特征，数据质量可能较差，譬如不客观、存在分析错误或者具有误导性等。因此，立足于统计视角，首先要科学确定大数据需求。对大数据的需求取决于官方统计的核心目标，而不能为了赶时髦而泛化大数据，也不能不顾数据质量和投入产出效率。其次，要规范标准和统计口径。在使用“大数据”的过程中，要注意规范及其标准，指导“大数据”在收集时就尽可能与统计相关指标的口径和标准一致起来；还要完善大数据的统一编码系统和登记系统，推动大数据的分类使用等。最后，面对大数据对统计数据收集带来的挑战，不妨尝试以下应对措施：尽量推行具有自动计算功能的网络填报方式；在确保安全的情况下，尽量使用电子签名，以减少信息传递、投递的时间；降低信息收集的频度；数据收集过程，避免信息项的重复收集；减少小型企业的信息收集负担等。

大数据时代的数据整理和分析，可能才是最大的挑战。在大数据的洪流中，为提升数据带给我们的洞察力，从技术角度首先应该收集和开发特定工具，以管理大规模的结构与非结构化数据，而大数据处理技术一般包括大规模并行处理数据库、数据挖掘网络、分布式文件系统、云计算平台等。其次，每个数据处理组织，需选定数据分析软件来挖掘大数据的内在价值，除 SAS、R、Minitab 等大型统计分析软件外，还要适应和使用新型开发的软件，譬如 Hadoop、NoSQL 等。同时要注意到，由于大数据及数据处理技术一般只为政府以及像淘宝、百度等巨头公司所掌握，“大数据”发展的一个重要障碍就在于数据的“流动性”和“可获取性”。因此，应推动相关部门制定适应性强的反垄断条例等措施，以保护极具竞争力的大数据市场，防止数据大亨的崛起，促进大数据平台上的良性竞争。

另外，面对大数据时代的挑战，统计的一个重要职责是培养能够整理和分析大数据的人才，即“数据科学家”。数据科学家是指对数据的数字化重现与认识，并在数据领域有一定贡献的人，一般应具备统计分析、对数据的提取与综合以及数据的可视化表示等多种能力。

对大数据的掌握可以转化为经济价值的来源，大数据已开始撼动世界的方方面面，从商业科技到医疗、政府、教育、人文以及社会的其他各个领域。同时，大数据时代也向包括统计在内的众多领域提出了众多挑战，我们需要做好充足的准备及应对，以迎接大数据时代带来的各种改变。

参考文献：

[1] 陈如明 .《大数据时代的挑战、价值与应对策略》. 移动通信 .2012 年第 8 期

[2] 姜奇平 .《大数据时代到来》. 互联网周刊 .2012 年第 1 期

[3]Martin Klubeck.《量化大数据时代的企业管理》. 人民邮电出版社 .2013

[4] 徐子沛 .《大数据：正在到来的数据革命》. 广西师范大学出版社 .2012

[5] 维克托·迈尔·舍恩伯格，肯尼思·库克耶 .《大数据时代：生活、工作与思维的大变革》. 浙江人民出版社 .2013

[6] 郑京平，王全众 .《官方统计应如何面对 Big Data 的挑战》. 统计研究 .2012 年第 12 期

作者单位：浙江工商大学统计学院