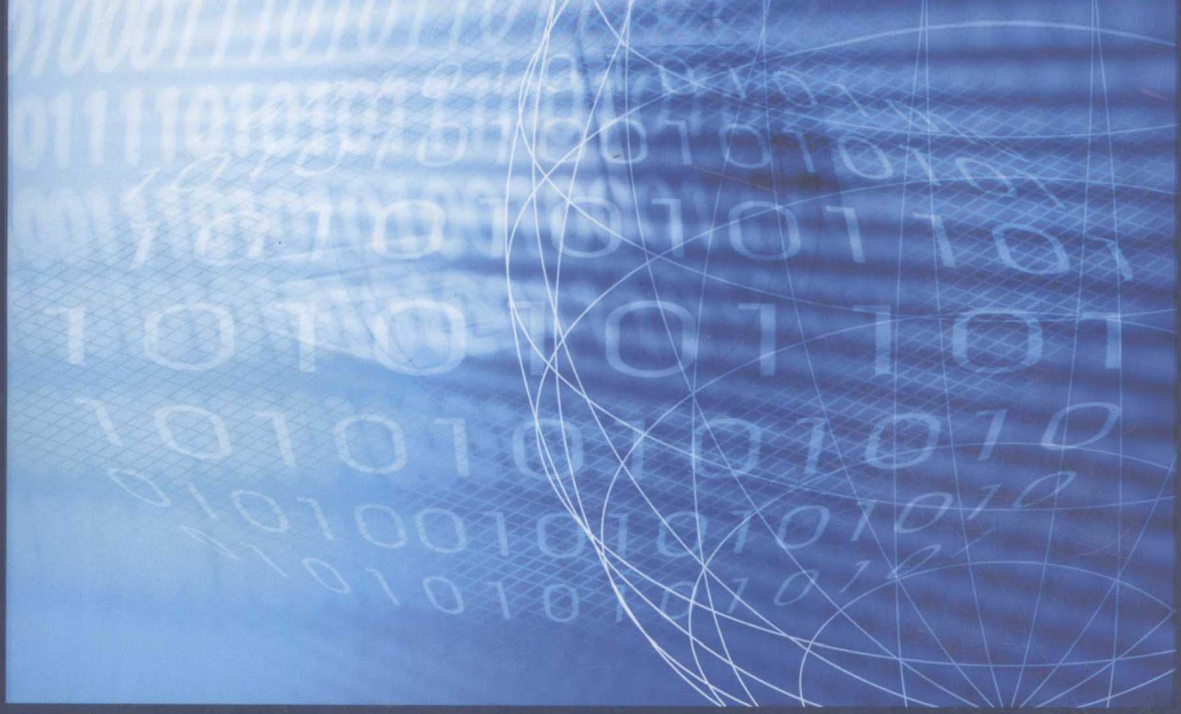


THE FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

第四范式：数据密集型科学发现

Tony Hey　Stewart Tansley　Kristin Tolle

潘教峰　张晓林　等 译

科学出版社

第四范式：数据密集型科学发现

Tony Hey　Stewart Tansley　Kristin Tolle
潘教峰　张晓林 等　译

科学出版社
北京

内 容 简 介

本书系统介绍了地球与环境科学、生命与健康科学、数字信息基础设施和数字化学术信息交流等方面基于海量数据的科研活动、过程、方法和基础设施，生动揭示了在海量数据和无处不在网络上发展起来的与实验科学、理论推演、计算机仿真这三种科研范式相辅相成的科学研究第四范式——数据密集型科学发现，进一步探讨了这种新范式的内涵和内容，包括利用多样化工具不间断采集科研数据、建立系统化工具和设施来管理整个数据生命周期、开发基于科学研究问题的数据分析及可视化工具与方法等，并深入探讨了这种新范式对科学研究、科学教育、学术信息交流及科学家群体的长远影响。

本书将帮助从事科学研究、科技研究规划、科技政策等领域的科研人员和管理者理解和把握科研环境与科研方法的革命性变化，也将为学术出版、文献情报、科学数据及其他从事信息与知识管理的人士提供未来的战略视角，同时也有助于有志于科学研究和学术信息交流管理的高层次学生了解未来的挑战和需求。

图书在版编目（CIP）数据

第四范式：数据密集型科学发现/Tony Hey等；潘教峰等译.—北京：科学出版社，2012.6

ISBN 978-7-03-034725-1

Ⅰ.①第… Ⅱ.①T… ②潘… Ⅲ.①数据采集—研究 Ⅳ.①TP274

中国版本图书馆CIP数据核字（2012）第124010号

责任编辑：孙 芳/责任校对：邹慧卿
责任印制：张 倩/封面设计：陈 敬

科学出版社 出版
北京东黄城根北街16号
邮政编码:100717
http://www.sciencep.com

北京通州皇家印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2012年6月第 一 版 开本：B5（720×1000）
2012年6月第一次印刷 印张：17 1/2
字数：247 000

定价：90.00元

（如有印装质量问题，我社负责调换）

《第四范式：数据密集型科学发现》翻译组

组织与统稿：潘教峰　张晓林

审校与协调：潘教峰　张晓林　张志强　张智雄　梁　娜

参 加 翻 译：（按姓氏拼音排序）

安培浚　中国科学院国家科学图书馆兰州分馆

陈　方　中国科学院国家科学图书馆成都分馆

邓　勇　中国科学院国家科学图书馆成都分馆

房俊民　中国科学院国家科学图书馆成都分馆

高　峰　中国科学院国家科学图书馆兰州分馆

高柳滨　中国科学院上海药物研究所

侯玉芳　中国科学院计算机网络信息中心

黄金霞　中国科学院国家科学图书馆

姜　禾　中国科学院国家科学图书馆成都分馆

乐小虬　中国科学院国家科学图书馆

李　麟　中国科学院国家科学图书馆

梁　娜　中国科学院国家科学图书馆

刘　磊　中国科学院计算机网络信息中心

刘　晓　中国科学院上海生命科学研究院生命科学信息中心

刘建华　中国科学院国家科学图书馆

刘细文　中国科学院国家科学图书馆

马廷灿　中国科学院国家科学图书馆武汉分馆

曲建升　中国科学院国家科学图书馆兰州分馆
阮梅花　中国科学院上海生命科学研究院生命科学信息中心
沈志宏　中国科学院计算机网络信息中心
宋　文　中国科学院国家科学图书馆
万　勇　中国科学院国家科学图书馆武汉分馆
王　玥　中国科学院上海生命科学研究院生命科学信息中心
王金平　中国科学院国家科学图书馆兰州分馆
王勤花　中国科学院国家科学图书馆兰州分馆
吴　慧　中国科学院上海药物研究所
吴振新　中国科学院国家科学图书馆
于建荣　中国科学院上海生命科学研究院生命科学信息中心
张　军　中国科学院国家科学图书馆武汉分馆
张　磊　中国科学院上海药物研究所
张晓林　中国科学院国家科学图书馆
张志强　中国科学院国家科学图书馆
张智雄　中国科学院国家科学图书馆
郑军卫　中国科学院国家科学图书馆兰州分馆

译者的话

科学正在进入一个崭新的阶段。在信息与网络技术迅速发展的推动下，大量从宏观到微观、从自然到社会的观察、感知、计算、仿真、模拟、传播等设施和活动产生出大量科学数据，形成被称为“大数据”(big data)的新的科学基础设施。科学家不仅通过对广泛的数据实时、动态地监测与分析来解决难以解决或不可触及的科学问题，更是把数据作为科学研究的对象和工具，基于数据来思考、设计和实施科学研究。数据不再仅仅是科学研究的结果，而是变成科学研究的活的基础；人们不仅关心数据建模、描述、组织、保存、访问、分析、复用和建立科学数据基础设施，更关心如何利用泛在网络及其内在的交互性、开放性，利用海量数据的可知识对象化、可计算化，构造基于数据的、开放协同的研究与创新模式，因此，诞生了数据密集型的知识发现，即科学研究的第四范式。

微软公司撰写的 *The Fourth Paradigm: Data-Intensive Scientific Discovery* 是第一本、也是至今为数不多的从研究模式变化角度来分析“大数据”及其革命性影响的著作。全书以吉姆·格雷提出科学研究第四范式的著名演讲开篇，邀请国际著名科学家对数据密集型科学发现的理念、应用和影响进行了全面分析。第一部分，Dan Fay 等人介绍了地球、环境、海洋、空间等领域的大数据环境与科学应用；第二部分，Simon Mercer 等人分析了医学、认知科学、生物系统、医疗服务等领域的数据密集型科学发现；第三部分，Daron Green 等人提出了适应大数据时代的科学信

息与科学计算基础设施面临的挑战；第四部分，Lee Dirks 等人对数据密集型科学发现给学术信息交流带来的深刻变化做了描述。全书视野开阔、思考深邃，既把握大势，又深入具体，为把握第四范式的要旨与含义提供了坚实的基础。

译者第一次接触到此书是在它刚刚出版后不久，并在此后有幸聆听了此书编者、作者之一的微软全球副总裁 Tony Hey 和微软学术合作部负责人 Lee Dirks 等人就此所做的专门报告，深感此书对于理解当今科学变革的重要。承蒙微软公司同意，我们将此书翻译为中文，以供中国读者学习。此书的翻译得到了微软亚洲研究院的支持，该院吴国斌先生多方协助，在此一并致谢！

前　言

Gordon Bell | 微软研究院
译校　张晓林

本书提出了一种新的科学研究范式：基于数据密集型计算的科学研究第四范式。这种科研范式就像当初印刷术的发明一样具有重大意义。印刷术历经一千多年才完善为今天我们所熟悉的许多形态，而新的范式——利用计算机从我们创建并存储在电子数据库中的数据中发现和理解自然与世界——也需要或多或少几十年才能成熟。本书的作者们通过非同寻常的努力，从不同领域帮助我们提炼和理解这个新范式。

在许多情况下，科学界在从数据中推演意义并基于此采取行动方面已经落后于商业领域。但是，商业推理相对比较简单，主要是关于那些可以用少数指标描述的产品及其生产、采购和销售。科学学科却很难被封装为少数可计量的指标或产品，而且多数科学数据缺乏足够高的经济价值来促使人们利用它们进行迅速的科学发现。

但是，正是第谷·布拉赫（Tycho Brahe）的助手约翰内斯·开普勒（Johannes Kepler）从布拉赫对天体运动的系统观察记录中发现了行星运动定律。在对所采集并仔细保存的实验数据进行挖掘和分析的基础上建立起新的理论，也正是第四范式的一个重要特征。

在 20 世纪，蕴藏着科学理论的科学数据经常被淹埋在零零散散的实验室记录本中，只是在少数“大科学”项目中才会被存储在磁介质里，而这些磁介质里的数据最后却无法读出。科学数据，尤其是来自单个、小型的实验室数据，一般都很难获得，很可能随着科学家的退休而被遗弃，如果

运气好的话会被藏到研究机构图书馆里直到最后被扔掉。大规模数据管理和支持科学群体获取分布保存的数据成为巨大的挑战。

幸运的是，类似于国家大气研究中心（NCAR[1]）这样的“数据空间”一直愿意帮助那些通过分析所保存的观察数据或计算数据来开展试验的地球科学家。至少在这里，我们拥有了一个提供整个学科的数据采集、保存和分析的完整链条的机构。

在21世纪，人们通过各种新工具不间断地采集着海量的科学数据，也通过计算机模型产生着大量的信息，其中大部分已经长期存储在各种在线的、可以公共获取的、得到有效管理的系统上，可以支持持续的分析，这些分析将引发许许多多新理论的发现。我相信，我们很快会进入这样的时代：数据会像纸本文献一样被长期保存，而且能够通过数据云被人和计算机公开获取。直到最近我们才敢想象，我们能像在图书馆和博物馆保存实体遗产一样长久地保存和利用数据。这种永久性并不是像有些人以为的那样“不靠谱”，其实图书馆一直就坚持和长期在做数据保存，如它们努力保存科学家的个人记录（甚至有时是与科学家有关的所有信息）。磁介质云所存储的数据和数字图书馆中的数字文献会变成致力于保存纸本及其印刷产出的图书馆书架的现代替代品。

2005年，国家科学基金会下的国家科学理事会发表了《长期保存的数字数据集合：支持21世纪的研究与教育》报告。报告一开始引用了一段关于数据保存重要性的对话，提出了如何培育和支持被称为数据科学家的新兴科学家群体的问题：“数据科学家包括信息学家、计算机科学家、数据库和软件工程师或程序员、学科专家、数据管理者、数据标引专家、图书馆学家、档案学家等一系列对科学数据资源的成功管理起着关键作用的人们，他们希望自己的创造性和智力贡献得到充分认可。”[1]

第四范式：聚焦于数据密集系统和科学交流

吉姆·格雷（Jim Gray）于2007年1月11日在计算机科学与电信委

1 http://www.ncar.ucar.edu.

员会上的最后一次演讲中描绘了自己关于科学研究第四范式的愿景[2]，呼吁资助开发用户数据采集、管理和分析的工具，以及交流与发布的基础设施，还强调要建立起与传统图书馆一样普及和强大的现代化数据与文件存储体系。本书收录了根据吉姆·格雷演讲稿和演示文档编辑的文章，此文章成为本书的基调。

数据密集型科学由三个基本活动组成：采集、管理和分析。数据从各种不同规模和性质的来源涌来，包括：①大型国际实验；②跨实验室、单一实验室或个人观察实验；③以后还可能来自个人生活之中[2]。各种实验涉及许多学科，涉及大规模数据，特别是它们的高数据通量，使得合适的采集、管理与分析工具成为巨大的挑战。澳大利亚的平方公里阵列射电望远镜项目[3]、欧洲粒子中心的大型强子对撞机[4]、天文学领域的泛 STARRS 天体望远镜阵列[5]等，每天都能产生好几个千万亿字节（PB）的数据，但现在却只能按照可管理的能力限制其数据速率。基因测序机器由于开销原因只能提供小的数据输出，所以，只有人们的某些编码区域才被测序（如 25KB 用于测几千个碱基对）。不过，这种情况只是暂时的。当千万美元的 X 竞赛（X PRIZE）[6]被人夺得时，我们就可以在 10 天内为 100 个人进行全面测序，每个人只花 1 万美元就可对 30 亿对碱基对测序。

我们需要资金来创建一系列通用的工具以支持从数据采集、验证到管理、分析和长期保存等整个流程。数据管理覆盖从确立合适的数据结构到将数据映射到各种存储系统之中的多种活动，包括支持跨工具、跨实验项目和跨实验室的长期可用和集成的数据模式及必要的元数据。没有这些明确的模式与元数据，对数据的理解只能是含糊不清的，必须依靠特殊的程序才能分析它。到最后，这种没有得到有效管理的数据肯定会丢失。我们必须仔细地考虑到底哪些数据应该长期可用，由此必须采集或创建哪些额

2 http://research.microsoft.com/en-us/projects/mylifebits.
3 http://www.ska.gov.au.
4 http://public.web.cern.ch/public/en/LHC/LHC-en.html.
5 http://pan-starrs.ifa.hawaii.edu/public.
6 http://genomics.xprize.org.

外的元数据来使“长期可用”变得可行。

数据分析也覆盖了整个工作流的所有活动，包括建立数据库（而不仅是建立一个可以用数据库系统去读取的文件集）、建模和分析，然后是数据可视化。吉姆·格雷就如何为一个学科建立数据库提出的要求是：这个数据库必须能够回答该领域科学家希望它回答的 20 个关键问题。在多数科学领域，科学家并没用数据库来实际存储数据，而只是用数据库来存储关于数据的有关信息，这是因为扫描所有数据需要花费太多时间，无法有效进行分析。到 2010 年，磁盘容量增加了 1000 倍，但磁盘读取时间仅仅提高了两倍。

数据和文件的数字图书馆：就像现代文献型的图书馆

科学交流，包括同行评议，正在经历根本的变革。由于价格、及时性和需要把实验数据与关于数据的文件存储在一起，公共的数字图书馆已经取代了传统图书馆而成为出版物的主要收藏系统。

在本书撰写时，数字图书馆仍然在形成阶段，规模不等，形态不一，性质也多样。当然，NCAR 是历史最久远的用于建模、采集和管理地球科学数据的系统之一。加州大学圣地亚哥分校的圣地亚哥超级计算中心（SDSC）在为科学界提供计算服务的同时，也是最早将数据管理纳入目标的机构之一。SDSC 建立了自己的数据中心[7]，目前在超过 100 个数据库中保存了 27PB 数据，包括生物信息学和水资源等数据。2009 年，它专门划出 400TB 存储空间用于为各类实验室、图书馆和博物馆等科学机构保存其私有或公共数据。

澳大利亚数据服务中心（ANDS[8]）已经启动了“登记我的数据”服务，它类似一个卡片目录，登记包括来自个人的各类数据库的标识、结构、名称和地点（IP 地址）。仅仅登记自己的数据还远不能算数据的长期保存。ANDS 的目的是要影响国家数据政策，宣传数据管理的最佳实践，从而把

7 http://datacentral.sdsc.edu/index.html.

8 http://www.ands.org.au.

当前研究数据收集管理的令人绝望的状态转变为一个协调统一的研究资源集合。在英国，联合信息系统委员会(JISC)资助成立了数据管理中心(DCC[9])来探索如何解决这些问题。随着时间的推进，我们可以期待，许多类似的数据中心会出现。国家科学基金会计算机与信息科学与工程部最近发布了一个项目征集公告，将对数据密集型计算和数据长期保存方面的研究者提供长期资助。

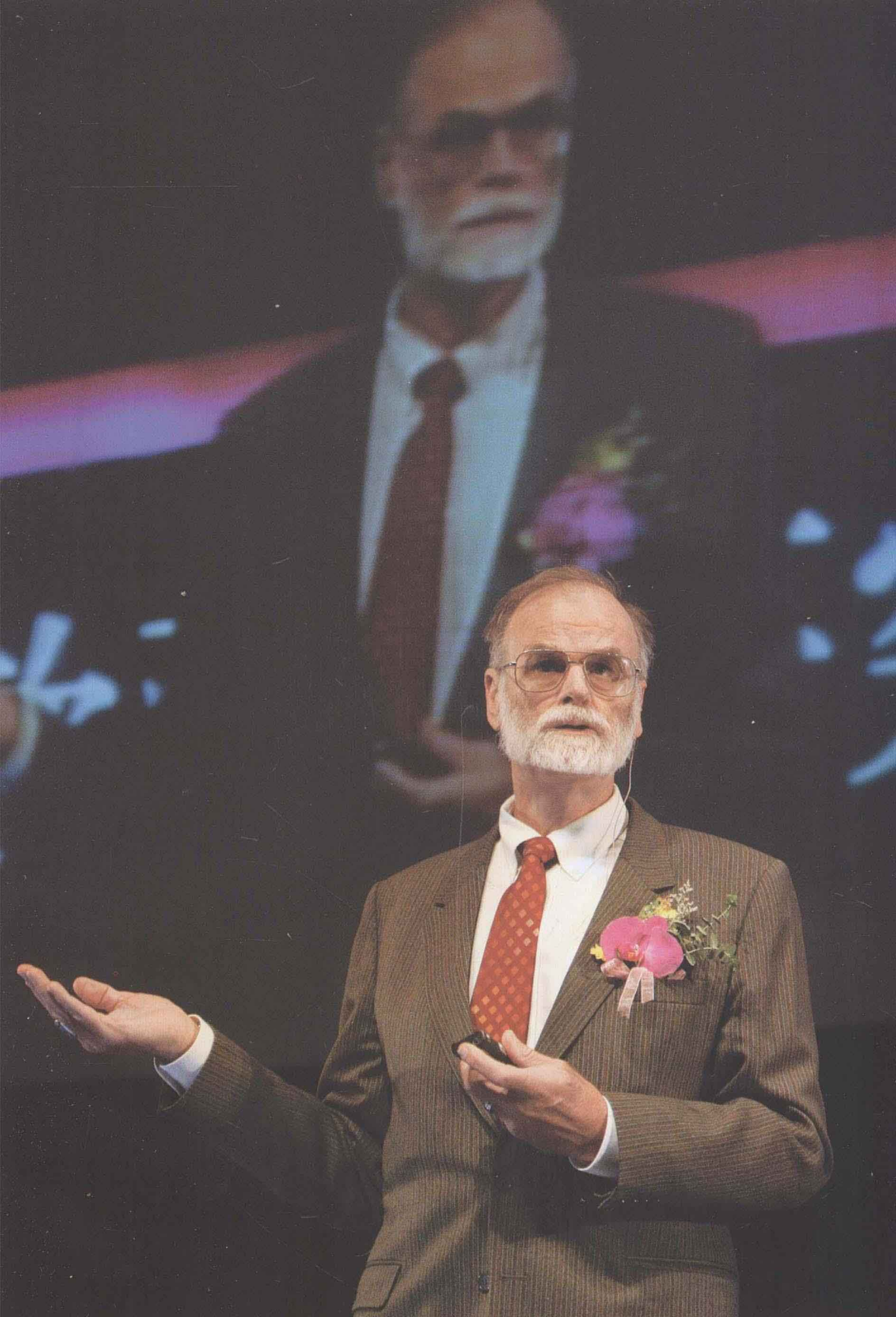
希望读者在阅读本书时能够了解数据密集型科学的许多机会和挑战，包括跨领域合作和培训，支持科学数据融汇的机构间数据共享，建立新的流程与渠道等，并提出研究议程来利用这些机会驾驭海量数据。应对这些挑战需要大量的建设和运行投资。要实现能够支持新的科技研究的、基于“无处不在的感知器”的数据基础设施的梦想，将需要资助机构、科学家和工程师之间的大规模合作，需要有充分的鼓励和资助。

参考文献

[1] National Science Board, “Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century,” Technical Report NSB-05-40, National Science Foundation, September 2005, http://www. nsf. gov/pubs/2005/nsb0540/nsb0540.pdf.

[2] Talk given by Jim Gray to the NRC-CSTB in Mountain View, CA, on January11, 2007, http: //research.microsoft.com/en-us/um/people/gray/JimGrayTalks.htm. (Edited transcript also in this volume.)

9 http://www.dcc.ac.uk.



吉姆·格雷论 eScience：科学方法的一次革命

本文基于吉姆·格雷于 2007 年 1 月 11 日
在加州山景城召开的 NRC-CSTB[1] 上的演讲记录整理而成[2]

Tony Hey, Stewart Tansley, Kristin Tolle | 微软研究院
翻译　刘磊 / 审校　张晓林

我们必须更加善于生产有关的工具，支撑从数据采集、数据管理到数据分析和数据可视化整个科研周期。如今，无论在超大规模（mega-scale）或在微细规模（milli-scale）的科学研究中，采集数据的工具都很糟糕。当你采集了数据后，需要在开始做任何数据分析之前妥善管理好数据，然而我们缺乏好的数据管理和分析工具。人们在分析数据后会发表研究成果，但发表的文献仅仅是数据冰山之一角。通过这个例子，我想指出，人们收集了大量数据，然后把这些数据缩减发表到 *Science* 或 *Nature* 的有限的专栏空间——如果由一个计算机科学人士撰写，最多或可达到 10 页篇幅。因此，我用数据的冰山一角来说明，我们有收集好的大量数据，但没有妥善管理或没有以任何系统的方式发表。也有一些例外，这些“例外”案例是我们寻找最佳实践的源泉。我下面会谈到同行评审的整个过程必须改变和它正在发生变化的方式，以及我认为 CSTB 能发挥什么作用来帮助每个人获取我们的研究成果。

1 National Research Council, http://sites.nationalacademies.org/NRC/index.htm; Computer Science and Telecommunications Board, http://sites.nationalacademies.org/cstb/index.htm.
2 令人痛惜的是，这篇演讲稿是 2007 年 1 月 28 日吉姆在海上失踪前发布在他的微软研究院网页上的最后一篇文章——http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt.

eScience 指的是什么？

信息技术与科学家的相遇催生了 eScience。科研人员利用许多不同的方法收集或产出数据——从传感器、CCD 到超级计算机、粒子对撞机等。当数据最终呈现在你电脑中的时候，你用现已存储于你电脑硬盘中的这些信息做什么呢？不断地有人找到我说："帮帮我！我有了所有的这些数据，我该利用它做些什么？我的电子表格正在失控！"而接下来将发生什么？当你有 10 000 个电子表格，每个电子表格里面都有 50 个工作簿的时候，将会发生什么？没错，我已经在系统地命名它们，但是现在我要做什么？

科学范式

每次讲话我都展示这个幻灯片（图 1）。如实地说，我是在 CSTB 资助的一个计算期货的研究项目中才逐渐明白它体现的这种洞察力。我们说："瞧，计算科学是第三条腿。"在科学研究中，最初只有实验科学，接着有理论科学，有了开普勒定律、牛顿运动定律、麦克斯韦方程式等，然后，对于许多问题，用这些理论模型来分析解决变得太复杂，人们只好开始进

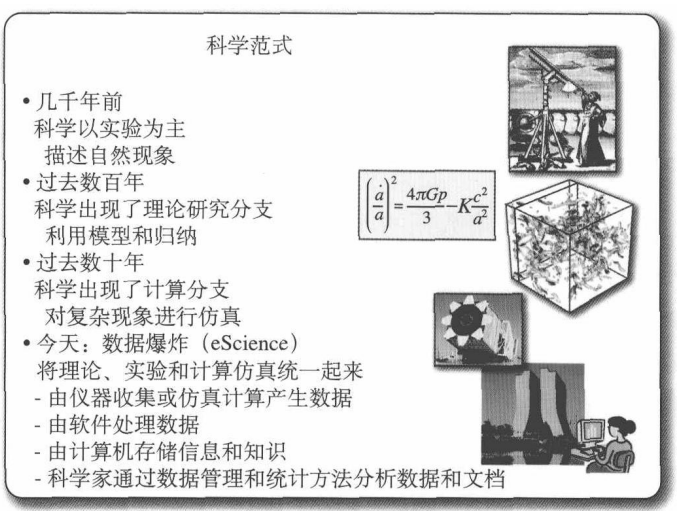


图 1

行模拟，这些模拟方法已经引领我们走过了上一个千年最后一半中的几乎全部时间。现在，这些模拟方法正在生成大量数据，同时实验科学也出现巨大的数据增长。人们事实上并不用望远镜来看东西了，取而代之的是通过把数据传递到数据中心的大规模复杂仪器来“看”，直到那时他们才开始研究在他们电脑上的信息。

毫无疑问，科学的世界发生了变化。新的研究模式是通过仪器收集数据或通过模拟方法产生数据，然后用软件进行处理，再将形成的信息和知识存储于计算机中。科学家们只是在这个工作流中相当靠后的步骤才开始审视他们的数据。用于这种数据密集型科学的技术和方法是如此迥然不同，所以，从计算科学中把数据密集型科学区分出来作为一个新的、科学探索的第四种范式颇有价值[1]。

X-Info 和 Comp-X

我们正在见证每个学科演变为两个分支，正如在如下幻灯片中显示的那样（图 2）。如果你看一下生态学，现在既有计算生态学——与模拟生

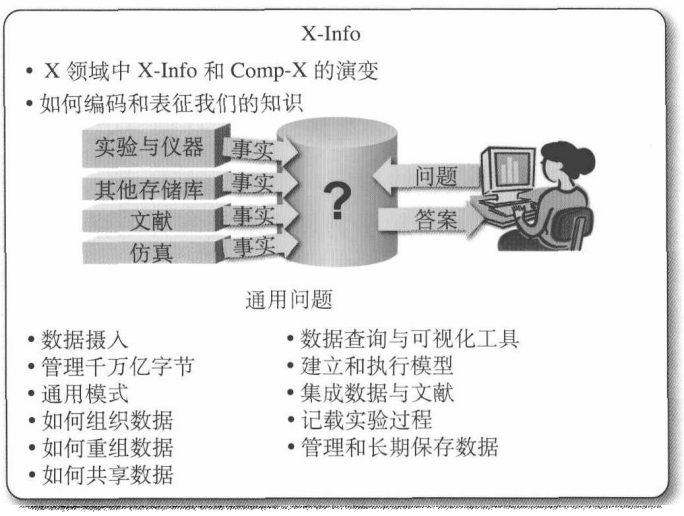


图 2

态学有关，又有生态信息学——与收集和分析生态信息有关。类似地，生物信息学从许多不同的实验中收集和分析信息；计算生物学模拟生物系统怎样运转，一个细胞的行为或代谢路径，又或一个蛋白质生成的方式。这与珍妮特·温（Jeannette Wing）的“计算思维”想法类似，计算机科学技术和方法被应用于不同的学科中[2]。

许多科学家的目标是要对他们的信息进行编码，这样，他们可以与其他科学家进行交流。为什么他们需要编码他们的信息？因为如果把一些信息放进计算机里，你能理解该信息的唯一方式是你的程序能否理解该信息，这意味着这些信息必须以某种算法的方式表达出来。为了做到这一点，需要给基因是什么、银河是什么或者温度测量是什么等提供一种标准的表达方式。

实验预算中软件应占 1/4~1/2

几乎在过去的十年期间，我经常和天文学家在一起，曾经到访过他们的一些台站。令我吃惊的一件事是，他们的望远镜简直令人难以置信，大概价值 1500 万或 2000 万美元的设备，有 20 ~ 50 人在操作。然后你开始领会到，需要成百上千的人写代码来处理这种仪器产生的信息，还需要数以百万计的代码行来分析所有这些信息。事实上，软件成本在资产开支中占主导地位！这一点在斯隆数字巡天计划（SDSS）中是不争的事实，并且在更大规模的巡天计划中，事实上对许多大规模实验来说，都是如此。虽然我不确定，对于粒子物理学界及其大型强子对撞机（LHC）来说软件成本的这种主导性是否属实，但我感觉，对于 LHC 实验来说肯定会是这样。

甚至在涉及“小规模数据”的科学活动中，人们收集信息，然后不得不投入比当初获得这些信息所付出的更多精力用于对这些信息的分析，这些软件有非常典型的异质性，因为实验室科学家几乎没有通用工具来收集、分析和处理这些数据。构建通用工具，这是我们计算机科学家能为他们做的事。