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

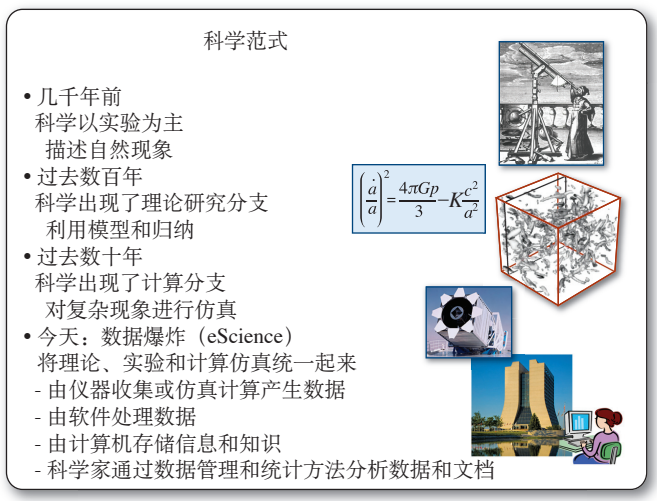

图 1

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

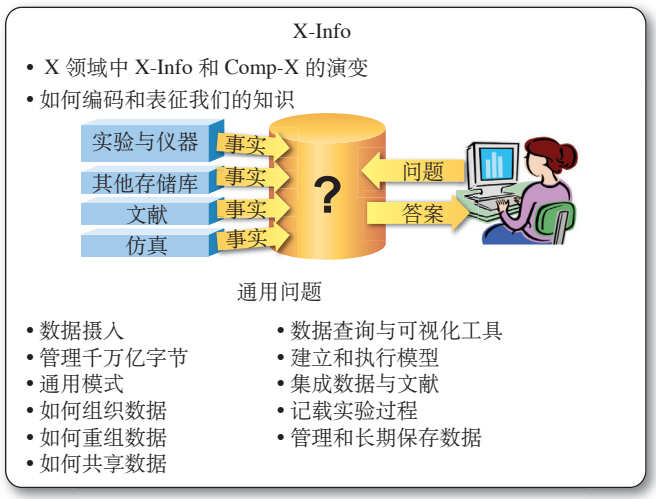

图 2

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。

我这里有一个给像 CSTB 这样的决策者提供的行动建议清单。第一个行动就是鼓励并资助构建工具。国家科学基金会现在有一个信息化基础设施资助部门，我并不想说关于它的任何坏话，但我们需要的不只是为万亿次网格（TeraGrid）和高性能计算提供支撑。我们现在知道如何构建 Beowulf 集群以获得便宜的高性能计算，但我们不知道如何构建一个真实的数据网格或如何构建由廉价的“数据砖石”建造的数据存储区，作为存放你所有数据并分析其中信息的地方。事实上，我们已经在模拟工具上取得相当的进展，但在数据分析工具上的进展不大。

项目金字塔和金字塔资助

这里谈谈关于大多数科研项目如何运转的一个观察。通常有少数几个国际项目，较多的多个大学间合作项目，许许多多的单个实验室项目，因此有这种第一层、第二层、第三层项目（设施）金字塔。你可以一再地在许多不同领域中发现这种情况。第一层和第二层项目通常都被系统地组织和管理，但只涉及相对很少的项目，这些大项目承受得起相应的软件和硬件预算，它们指定一些科学家团队为实验开发定制的软件。例如，我观察到，美加海洋气象台的海王星项目拨出大约 30% 的预算用于信息化基础设施 [3]，取其整数，就是 3 亿 5000 万美元的 30% 或大概 1 亿美元！类似地，LHC 实验有一个很大的软件预算，而这种提供大型软件预算的趋势在早先的 BaBar 实验中也显而易见 [4, 5]。但如果你是一个处于金字塔底部的实验室科学家，你会在软件预算上做些什么？你基本上会买 MATLAB[3] 和 Excel[4] 或一些类似软件，凑合着用这些现成的工具，再没有你能做的更多的其他事情了。

所以，千兆级别（giga）和更大级别（mega）的项目主要由大规模设施（如超级计算机、望远镜或其他大规模实验设施等）的需求驱动。这些设施通常被一个重要的科学家群体使用，并需要获得有关机构（如国家科

3 www.mathworks.com.

4 http://office.microsoft.com/en-us/excel/default.aspx.

学基金会或能源部等）的充分资助。较小规模的项目通常能从更多方面获得资助，资助机构通常得到一些其他机构（可以是该大学自身）的配套支持。在戈登·贝尔（Gordon Bell）、亚历克斯·萨雷（Alex Szalay）和我为 *IEEE Computer* 写的一篇论文里[6]，注意到，第一层设施（如 LHC）得到了一个国际财团机构的资助，但第二层的 LHC 实验和第三层设施所获得的资助通常是研究人员通过自己的资助源带来的。因此，资助机构需要充分地资助第一层的千兆级项目，但也需拨出他们信息化基础设施资助的其他一半用于支持更小的项目。

实验室信息管理系统

总结前面关于软件的讨论，我们需要的实际上是“实验室信息管理系统”。这样的软件系统提供一个从仪器或模拟数据进入数据档案馆的管道，在我一直致力于的许多案例中，我们已接近于实现这一点。我们基本上把从一堆仪器中获得的数据放入一个校准和“清洗”数据的管道中，包括必要时重新填补数据空缺，然后把这些信息“重新网格化”[5]，最终把它放进一个你愿意“发布”到互联网上的数据库中，让人们获取你的信息。

把数据从某个仪器转到一个网络浏览器的整个业务包含大量的技巧。其实需要做的事情很简单。我们应该能创建一个像 Beowulf 集群一样的程序包和模块，使做传统实验的人们有能力收集他们的数据，放入一个数据库，并且发表它。可以通过构建几个原型系统及文档化它们来实现。做这个会花费几年时间，但它会对科学研究的方式有很大影响。

正如我说过的，这一类软件管道被称为实验室信息管理系统或 LIMS。顺便说一句，已经存在商业化的系统，你能从货架上买到 LIMS 系统。问题是，它们事实上是为相对富裕的、有产业背景的人准备的，也常常是针对某特定群体的某个任务的，如从一个排序机或质谱仪中获取数据，通过该系统运行它，并输出结果。

5 这意味着将数据的组织规范化，直到每行的数据变量，类似于关系型数据库的规范化。

信息管理与数据分析

因此，典型的情况是：要么从仪器或传感器中、要么从运转的仿真模拟中收集数据，并且很快能得到数以百万计的文档，但我们没有简易的方式去管理或分析这些数据。我一直在实地调研，观察这些科学家是如何做的。通常，他们或是在（数据的）汪洋大海捞针，或是在寻找数据汪洋本身。大海捞针式的查询事实上很容易——要寻找数据中特定的异常现象，通常都对寻找何种类型的信号有所了解。粒子物理学家们就是在寻找 LHC 中的 Higgs 粒子，而他们非常了解这样一个重粒子的衰变在探测仪中看起来是什么样子。计算集群的网格对于这样一个大海捞针式查询是很棒的，但这样的网格计算机在趋势分析、统计聚类和发现数据中的总体模式上却很差劲。

事实上，我们需要用于聚类和数据挖掘的更好的算法。不幸的是，聚类算法不是线性（顺序 N）或对数曲线式（$N\log N$）的规模，而是典型的 N 立方规模，所以，当 N 变得太大，许多方法就会失效。因此，我们被迫发明新算法，而且还不得不忍受仅仅得到近似答案。例如，使用近似的中值结果被证明出奇的好。那么，谁会想到这点呢？不是我！

这些统计分析大多涉及创建统一样本、执行数据过滤、合并或对比蒙特卡罗模拟等，产生一大堆文档，而这些文档实际上是一串字节。假如我给你这一文档，你必须费很大工夫来弄明白该文档中的数据到底意味着什么。因此，文档的自我描述真的很重要。当人们使用“数据库”这个词时，从根本上说，他们所说的是：数据应该是自我描述的，并且应该有一个模式，那才是“数据库”一词的全部意思。所以，假如我给你一个特别的信息集合，你会看着该信息集说，“我要具有这种特性的所有基因”或“我要具有这种属性的所有恒星”或“我要具有这种属性的所有星系”。但若我只给你一串无法使用星系这样概念的文档，你将不得不到处搜寻去弄清什么才是文档中的数据有效模式。假如你有一个关于所针对问题或对象的模式，你就可以把该数据编入索引中，或聚集该数据，或对数据进行并行搜索，或对数据进行任意的查询，构建一些通用的可视化工具也会更容易。

说句公道话，科学界已经发明了一堆格式，我认为它们作为数据库格式是够格的。HDF（层次型资料格式）[6]就是这样的一种格式，NetCDF（网络公用数据形式）[7]是另一种。这些格式被用于数据交换并在传输过程中保存了数据模式。但整个科学领域需要比 HDF 和 NetCDF 更好的工具来支持数据的自定义。

数据传输：棘手的难题

另一个关键问题是：随着数据集变大，已不可能仅仅用文件传输协议（FTP）来传输数据或按给定模式搜索数据文件内容。1PB 的数据文件就很难用 FTP 来传输！因此，在某些点上需要目录及并行数据存取，这就是数据库能帮助你的地方。对于数据分析，一种可能性是把数据移到你手里，而另一种可能是把你的查询移到数据上。你可以移动问题或数据，结果常常是移动问题比移动数据更有效。

对数据工具的需求：鼓励百花齐放

我一直坚持的意见是：我们现在用于大多数科学学科的数据管理工具很糟糕。像沃尔玛这样的商业机构，负担得起构建他们自己的管理软件的费用，但在科学研究上我们就没那么奢侈了。现在，几乎没有任何数据可视化和分析工具。一些科研群体使用 MATLAB 这样的工具，而美国和其他地方的资助机构需要做更多事情来支持构建数据工具，以使得科学家们获得更多成果。当你去看看科学家们日复一日地在数据分析方面做的事情，那真是糟透了。我怀疑你们很多人和我处于同一种状态，基本上能支配的工具就只是 MATLAB 和 Excel!

我们的确有像 Beowulf 集群[8]那样的很好的工具，通过联合许多廉价计算机获得划算的高性能计算。有一款软件叫做秃鹫（Condor[9]），可使你

6 www.hdfgroup.org.
7 www.unidata.ucar.edu/software/netcdf.
8 www.beowulf.org.
9 www.cs.wisc.edu/condor.

从各部门的机器中收割周期性处理流程数据。类似地，伯克利开放式网络计算（BOINC[10]）软件允许像在 SETI@Home 项目“地外文明搜寻计划”那样收割个人计算机的周期性数据。我们有几个商业产品，如 MATLAB。所有这些工具产生于科研群体，我不明白的是，为何这些特别的工具是成功的。

我们也有 Linux 和 FreeBSD。FreeBSD 比 Linux 要早，但不知为何 Linux 成功了，而 FreeBSD 却没有。我认为这与科研群体、人员个性和时间选择有很大关系。所以，我的建议是应该开发更多的产品。我们有商业工具，如 LabVIEW[11]，但应该再创建几个这样的系统。我们只需要期望这些中的一些成功。为大量的项目提供种子资助应该不太昂贵。

即将到来的学术交流革命

以上是我的演讲的第一部分：关于帮助科学家收集、管理、分析及可视化数据的工具的需求。我演讲的第二部分是学术交流。大约三年前，国会通过了一项法案，如果你得到了国立健康研究院（NIH）的资助用于你的研究，你就应该把研究报告存放于国家医学图书馆（NLM），这样，你的论文全文就会成为公共知识。自愿遵守此法案的人只有 3%，所以需要有一些变化。现在，我们有可能看到所有公共资助的科学文献被资助机构强制地在线发布。当前，有一个由参议员柯尼恩（Cornyn）和李伯曼（Lieberman）发起的法案，将要求接受 NIH 资助的人必须把其发表的研究论文放进 NLM 的文献检索系统 PubMed Central[12]。在英国，维康信托基金会已经对接受其科研资助的人实施了一个类似的要求，并已创建了一个 PubMed Central 的镜像。

但互联网能做的事情不仅仅是使研究论文的全文可以获取，原则上，它能把所有的科学数据与文献联系在一起，创建一个数据和文献能够交互

10 http://boinc.berkeley.edu.

11 www.ni.com/labview.

12 截止 2011 年 11 月，遵守率已经超过 50%。参见 Peter Suber 的开放存取简报（Open Access newsletter）：www.earlham.edu/~peters/fos/newsletter/01-02-08.htm.

操作的世界（图 3）。到那时，你可以在阅读某人的一篇论文时查看他们的原始数据，甚至可以重做他们的分析；或者可以在查看某些数据时查出所有关于这一数据的文献。这样的能力会提高科学的“信息速率”，促进研究人员的科学生产力。我相信这将会是一个很好的发展！

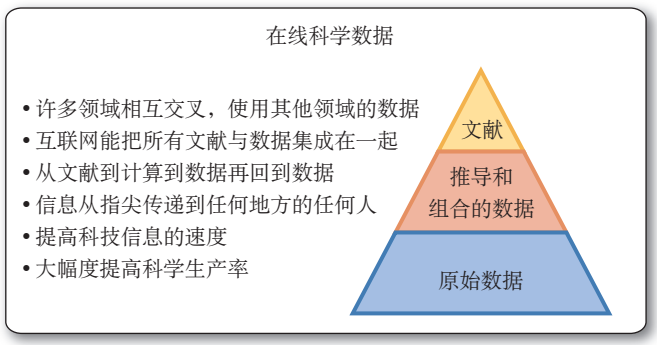

图 3

以某个在为 NIH 工作的人为例——这也是我们正在讨论的案例——他完成了一个报告，发现了关于疾病 X 的病因。你去找你的医生说：“医生，我感觉不太好。”他说：“安迪，我们会为你做一系列检查。”然后检查完第二天给你打电话说：“你什么问题都没有，服用两片阿司匹林，休假一下。”一年后你返回诊所，又是这一套。三年后，他给你打电话说：“安迪，你得了 X 病！我们把它查清楚了！”你说：“X 病是什么？”他说：“我不知道，它是一种罕见的疾病，但在纽约有一个人知道关于此病的所有事情。”于是你就到 Google[13] 上输入你的所有症状，在检索结果的第一页就出现了 X 病。你点击一下，它就把你带到 PubMed Central 文献检索系统并找到“关于 X 的所有问题”这篇文章的摘要。你点击它，它又把你带到 *New England Journal of Medicine*，但该期刊系统这时说：“请付 100 美元才可以阅读这篇文章的内容。”你发现，作者是在为 NIH 工作，用你这

13 或者，如吉姆今天建议的，用必应（Bing）。

个纳税人的钱做的工作。所以，李伯曼[14]和其他人说：“这太不合理了。科学信息经过同行评审后被公开发表了，但只有那些愿意付费的人才能阅读它。这是为什么呢？我们已经为它（的产生）付过费了！”

学术出版商提供服务来组织同行评审、印刷期刊，并把它们分发到图书馆。但现在互联网是我们的传播者，而且几乎是免费的。这个问题与社会正在进行的关于知识产权应该在哪里开始、在哪里结束的讨论是相关的。科学文献，尤其是同行评审的文献，很可能成为封闭知识产权的地方之一。如果你想找到关于疾病 X 的内容，你很可能会找到“桃核对疾病 X 有很好的疗效”的信息，但这个信息不是来自同行评审的科学文献，而是来自某个想向你售出桃核用于治疗疾病 X 的人。所以，那些一直倡导开放存取的人们主要来自医疗保健领域，因为他们看到好的医疗保健信息被封锁了，而坏的医疗保健信息却在互联网上传播。

新型数字图书馆

新型的图书馆应该怎样运作呢？它应该是免费的，因为把一页内容或一篇文章放在互联网上很容易。在座的每个人都能负担得起在 PubMed Central 上发布文章的成本，它只会花费你几千美元买台计算机——但你会发布多少我可不知道！不过，管理这些内容却不便宜。把那些内容输入计算机，使内容可交互参照（cross-indexed），所有这类事情，NLM 要花费大约每篇文章 100 美元的费用。据估计，如果每年收集 100 万篇文章，单是管理这些内容就需要 1 亿美元。这就是我们需要使整个管理过程自动化的原因。

现在正出现的情况是，NLM 的数字化部分——PubMed Central 已经可以移植，有各种版本的 PubMed Central 在英国、意大利、南非、日本和中国运转。在英国运转的那个镜像是上周才开始在线的。我猜你会理解，法国不想让它的 NLM“在马里兰州贝塞斯达运转”[15]或使用英语，而英国人不想让文本成为美语的，所以，英国的 PubMed Central 很可能会在网页界

14 The Federal Research Public Access Act of 2006 (Cornyn-Lieberman).
15 美国 NLM 所在地。

面中使用英式拼写法。但从根本上说，你可以在这些文献集中的任何一个里面存放文档，它将会被复制到所有其他的文献集。运行这些文献集之一会相当便宜，但更大的挑战是你怎样管理这些内容和怎样管理同行评审。

融汇型期刊

这里谈一下我认为这类基于内容融汇的期刊会怎么运作。基本想法是：你有数据档案库，也有文献档案库。文章存储在文献档案库中，而数据则进入数据档案库中。然后利用一个期刊管理系统，允许我们作为一个群体，来形成一个关于疾病 X 的期刊。人们通过把文章储存到文献档案库中来向这个期刊提交文章。我们对这些文章进行同行评审，对于那些符合要求的文章做一个目次标题页，说明“这些文章是符合要求的文章”，把这个目次标题页也放进文献档案库。现在，搜索引擎可以提高这些文章的网页排序值，因为它们已被这种很重要的标题页引用。这些文章当然也可以回溯指向相关的数据。在此基础上，会有一个协作系统出现，允许人们在期刊文章上加以注解或评论，这些评论并不被储存在同行评审的文献档案库中，因为它们还没被同行评审过——尽管这些评论可能是同行管理的。

NLM 将为生物医学界做这些事情，但这在其他科学领域尚未发生。对于 CSTB 成员而言，计算机科学界可以通过为其他科学领域提供适当的工具，促使它的发展。

我们在微软研究院设计了一款软件叫做会议管理工具（CMT），并且已经用它组织管理了大约 300 个会议。CMT 服务使你非常轻松地创建一个会议，支持组成议程委员会、发布网站、接收文稿、处理利益冲突或撤换稿件、进行评审、决定接受论文、形成会议日程、通知作者、修改论文等整个工作流。我们现在正在努力提供一个按钮，可把文章存到 arXiv.org 或 PubMed Central，并同时建立会议论文标题页，使人们很容易捕捉到研讨会和会议的信息，或管理一个在线期刊。这种机制会让创建融汇型期刊变得很容易。

有人曾问到，这是否会对学术出版商形成冲击，回答是肯定的。但是，

这是否也对电气与电子工程师协会（IEEE）和美国计算机协会（ACM）造成冲击？是的，假如他们没有任何独特的论文可发送给你，你就不会加入他们，因此，这些专业学术团体也对此感到很恐惧。我认为，他们必须面对这一问题，因为我认为开放获取必将来临。看看会场，我们中的大多数人年岁已老，不是 60 后至 80 后（X 代）的人。我们大多数人加入这些组织是因为我们认为这是在那些领域成为专业人士的一部分。麻烦的是，X 代的人不参加什么组织。

在同行评审上将发生什么？

这也许不是与你个人有关的问题，但许多人在问：“究竟为什么需要同行评审？为什么不只要一个维基？”我想，回答是同行评审是不同的，它非常有组织，有管理，而且人们讨论的问题有一定的保密度。维基是更加平等主义的。我认为对于收集论文发表后的评论，维基很有作用，当然我们需要某种支持结构，就像 CMT 为同行评审所提供的那样。

发表数据

我将简要讨论数据发表问题。我已经谈到了文献发表，但如果你得到一个答案，42，单位是什么？你把一些数据放进一个发布在网上的文件中，但我们又回到文件的各种问题上。要在文件中显示你的工作过程的重要记录被称作数据溯源：你是怎样得到 42 这个数据的？

这里有一个思维实验。你已做了一些科学研究，你想发表它。你怎样发表它，以至于他人能够在一百年后阅读它并复制你的结果？孟德尔和达尔文做到了这一点，但是很勉强。在技巧方面，我们现在比孟德尔和达尔文更加落后。尽管是一团糟，但我们必须着手解决这个问题。

数据、信息和知识：本体论和语义论

我们在努力使知识对象化。可以先做一些界定数据单位这样的基础事情，如什么是测量单位，谁做的测量，以及该测量是在什么时候做的。这些是适用于所有领域的工作。这里（微软研究院）我们做计算机科学。

我们说的行星、恒星和银河系是什么意思？那是天文学。基因是什么？那是生物学。那么什么是对象，什么是属性，什么是在面向对象的意义上的关于对象的方法？顺便说一句，互联网正在变成一个面向对象的系统，人们从中可以获取关于对象的知识。在企业界，他们在把顾客、发票等对象化。在自然科学中，例如，我们需要把基因对象化——这是基因数据库（GenBank[16]）所做的事情。

需要提醒大家，我们将走得更远些。我们用“O”这个字母来表示知识本体（ontology），“S”这个字母来表示知识模式（schema），和“受控词汇”（controlled vocabularies）这个词。这是说，沿着这条路往前走，我们将开始谈论语义学的东西，即“事情意味着什么？”当然，每个人对事情意味着什么有不同的看法，因此讨论可能是无休无止的。

这一点上的最好例子是 Entrez[17]，美国国家生物技术信息中心为 NLM 创建的生命科学搜索引擎。Entrez 可以搜索整个 PubMed Central 数据库，这里搜索的是文献，也有系统发育数据、核酸序列数据、蛋白序列和它们的三维结构数据，有基因数据库，它真是一个令人印象深刻的系统，并且它还建立了化学分子及其生物活性数据库（PubChem）和许多其他内容。这个系统完全允许数据和文献互操作。你可以正在读一篇文章，然后链接到相关的基因数据，随着该基因到相关疾病的数据，再回到与这个疾病相关的文献。它真的很棒！

所以，在传统世界里，我们已经有作者、出版者、文献管理者和消费者。在新的世界里，科学家们正在协同工作，期刊正变成包含数据和其他实验细节的网站，内容管理者们现在照管大型数字档案库。在传统世界和新的世界中，仍然相同的只有一点，就是一个个的科学家。我们从事科学的方法已经发生了根本的变化。

我们面临的问题是：所有项目都会在某个时间结束，但我们不清楚项目结束后它们的数据会怎么样。有不同规模的数据，如人类学家在外面收

16 www.ncbi.nlm.nih.gov/Genbank.
17 www.ncbi.nlm.nih.gov/Entrez.

集并放进他们的笔记本中的信息。LHC 实验中，粒子物理学家收集到的数据是海量字节，但大部分科学研究中的数据集是较小的数据量。我们现在开始看到数据的聚合应用，人们把从不同地方获得的数据集黏合在一起，形成新的数据集。所以，我们不仅需要期刊出版物的档案库，更需要数据的档案库。

因而这是我对 CSTB 的最后一个建议：培育数字化的数据图书馆。坦率地说，国家科学基金会的数字图书馆项目全部聚焦于传统图书馆的元数据，而非真正的数字图书馆。我们应该建设既支持数据也支持文献的数字图书馆。

总结

需要指出，由于信息技术的影响，关于科学的一切几乎都在变化中。实验的、理论的和计算的科学范式都正在被数据泛滥和正在出现的数据密集型科学范式——第四范式所影响，这一科学范式的目标是拥有一个所有科学文献和科学数据都在线且能彼此交互操作的世界，这需要许多新工具来促成。

编者注

吉姆·格雷演讲的全文和幻灯片可在第四范式网站上找到[18]，在该网站上可以看到本文中略去的演讲中的问答（注意，提问者没有标明身份）。这里的文本做了一点编辑处理以提高可读性，我们也添加了脚注和参考文献，但本文保证忠实于吉姆的演讲。

参考文献

[1] G. Bell, T. Hey, and A. Szalay, “Beyond the Data Deluge,” *Science*, vol. 323, no. 5919,pp. 1297–1298, 2009, doi: 10.1126/science.1170411.

[2] J. Wing, “Computational Thinking,” *Comm. ACM,* vol. 49, no. 3, Mar. 2006,doi:

18 www.fourthparadigm.org.

10.1145/1118178.1118215.

[3] NSF Regional Scale Nodes, http://rsn.apl.washington.edu.

[4] Large Hadron Collider (LHC) experiments, http://public.web.cern.ch/Public/en/LHC/LHCExperiments-en.html.

[5] BaBar, www.slac.stanford.edu/BFROOT.

[6] G. Bell, J. Gray, and A. Szalay, "Petascale Computational Systems," *IEEE Computer,* pp. 110–112,vol. 39, 2006, doi: 10.1109/MC.2006.29.

目录 CONTENTS

i 译者的话

iii 前言

ix 吉姆·格雷论eScience：科学方法的一次革命

第一章 地球与环境

3 一、引言

5 二、 格雷法则：以数据库为中心的科学计算

13 三、 正在兴起的环境应用科学

21 四、用数据重新定义生态科学

27 五、海洋科学2020年远景

39 六、拉近夜空：海量数据中的发现

45 七、装备地球：下一代传感器网络与环境科学

第二章 健康与幸福

55 一、引言

57 二、医疗奇点与语义医学时代

65 三、发展中国家的医疗服务：面临的挑战及可能的解决之道

75 四、大脑神经回路图谱探索

83 五、用于神经生物学研究的计算显微镜

91 六、数据密集型医疗保健的统一建模方法

99 七、生物系统进程代数模型的可视化

第三章 科学的基础框架

109 一、引言

111 二、科学新路径?

119 三、超越数据海啸：发展基础设施，处理生命科学数据

127 四、多核计算与科学发现

133 五、并行计算和云