科学研究的第四范式

吉姆·格雷的报告“e-Science: 一种科研模式的变革”简介

微软公司于2009年10月发布了《e-Science：科学研究的第四种范式》论文集，首次全面的描述了快速兴起的数据密集型科学研究。论文集分为四大部分，包括：地球与环境、健康与生活、科学基础设施、学术交流等。这些论文扩展了计算机科学图灵奖获得者——吉姆·格雷的思想，提出基于数据密集型的第四范式发现，提供了如何将其全面实现的见解。论文集开篇的文章“吉姆·格雷论e-Science:一种科研模式的变革”是根据吉姆·格雷于2007年1月11日在美国国家研究理事会计算机科学与通讯分会（NRC-CSTB）的报告整理而成的，我们摘译其重要观点，以供读者参阅。

文章明确提出：支持科研数据全生命周期管理的工具亟待完善。一个完整的科学研究周期包含四个部分：数据采集、数据整理、数据分析及数据可视化。现代科学研究可以通过多种方式收集和生成数据，对于大量收集到的数据，却缺乏好的整理与分析工具。信息技术的发展促进各学科信息化进程，e-Science为科学研究提供了一种全新的思维与科研模式，各种工具的应用旨在解决现代科研中的海量数据问题，促进各学科更快地发展。

1. 科学研究第四范式的提出

吉姆·格雷给我们展示了科学研究模式的发展历程（如图1）。科学研究最初只有实验科学，随后出现了理论科学，这种科学运用了各种定律和定理，比如开普勒定律，牛顿运动定律等。后来，对于许多问题，理论分析方法变得非常复杂以至于难以解决难题，人们开始寻求模拟的方法，这就产生了计算科学。科学无疑是不断向前发展的，模拟连同实验产生了大量的数据，针对海量数据问题，一种新科研模式产生了：由软件处理由各种仪器或模拟实验产生的大量数据，并将得到信息或知识存储在计算机中，科研人员只需从这些计算机中查找数据。比如在天文学研究中，科研人员并不直接通过天文望远镜进行研究，而是从数据中心查找所需数据进行分析研究，数据中心存有海量的、由各种天文设备收集到的数据。

鉴于数据密集型科学研究独特的技术支持需求和鲜明特点，因此有必要将数据密集型科学从计算科学中单独区分开来，我们称之为第四范式，一种新的科研模式。

2. 科学研究的发展需求

吉姆·格雷在报告中用大量篇幅分析了现在科学家对收集、管理、分析以及可视化数据工具的需求。主要包括以下几个方面：

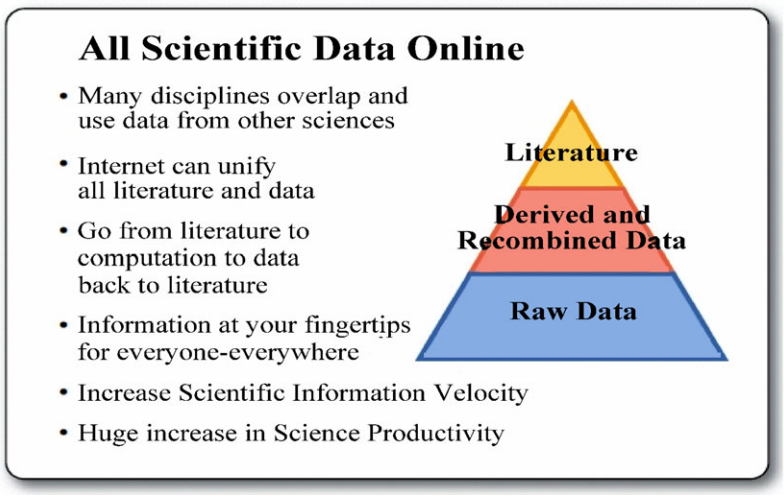


图 1 科学研究模式的发展历程

（1）信息交流需要编码处理

如图2所示，每个科学领域都逐步演变成两大分支，一支是负责从实验中收集和分析数据的；另外一支则是模拟该领域系统的运转的。例如，生态学，现在就分为生态信息学和计算生态学；类似的，生物学领域有生物信息学和计算生物学。这是因为，数据必须以一种规则的方式存储与表示，通过电脑程序可以处理的数据信息我们才能读懂和传播交流、共享使用。因此，为了与其他科学家交流，越来越多的科学家们正在努力进行信息编码处理。

（2）软件成本成为实验经费支出中的重头戏

事实上，科研在模拟工具上已经取得了长足的发展，但数据分析工具却建树颇少。无论是在大规模的科研活动中，例如，天文领域，还是在“小规模数据”科学领域，科学家们用以进行信息分析的经历、成本投入都要高出信息收集成本很多。而且他们进行信息分析的软件还具有十分典型的异构特征，缺乏通用工具软件对数据进行收集、分析和处理。

（3）多数科研项目得不到足够的软件和硬件预算支持

作者根据科研项目的规模将科研项目归纳为一个包括三层架构的“科研项目金字塔”。（图3）

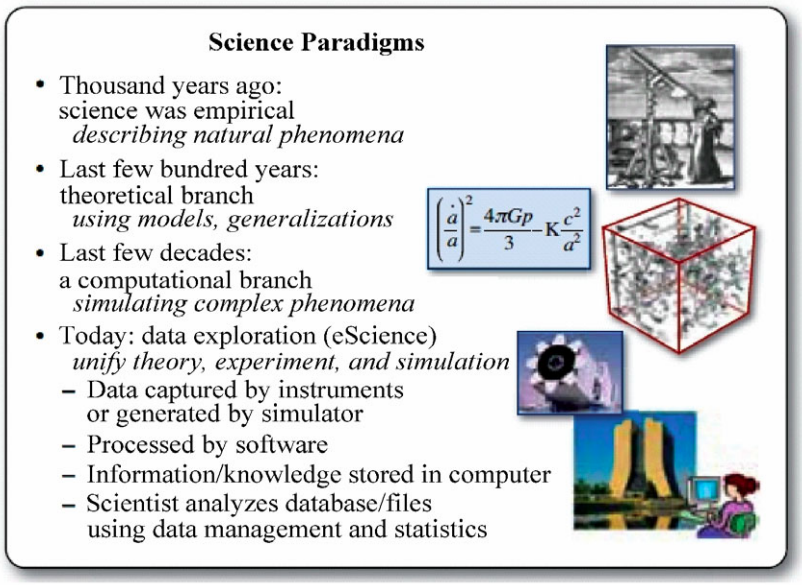
通常，第一、二层项目为国际型项目或者是大学间合作项目，这种项目一般都能得到系统的组织和管理，并且有足够的软硬件支持，但数量较小；第三层为一般的独立实验室项目，这部分项目数量众多，但是缺少相应的支持和管理。因此，资助机构需要充分地资助大型项目外，还应当拨出他们资助信息化基础设施的另一半用于支持更小的项目，以促进科学的发展。

（4）缺乏有效的实验室信息管理系统

目前，大多数的科研项目和实验室还缺乏有效的实验室信息管理系统。这样的软件系统可以提供一个从仪器或模拟数据进入数据归档的通道，对数据进行有效的描述和管理，以便进行分析处理，并最终把它放进一个方便发布到互联网上的数据库，让人们获取相关信息。在趋势分析、统计聚类和发现数据总体模式方面，我们需要更好的方法进行聚类和数据挖掘，并定义规范的数据库模式（schema）。而随着数据集的不断变大，数据传输也已成为了瓶颈问题。

（5）对数据工具的需求：亟待百花齐放

作者认为，现阶段应用于大多数科学学科的数据管理工具都不够成熟，缺少数据可视化和分析工具。一些科学群体使用例如MATLAB这样的工具，实际上美国以及其他地方的资助机构在扶植构建管理工具方面还有很多工作，才能促进科学家们产生出更多成果。作者建议管理工具多元化发展，并且其中的一些能够得到发展壮大。



3. 学术交流革命性的发展

本报告中，吉姆·格雷提出的另一个重要问题就是有关学术交流的问题。他提到，互联网的数据共享范围并不局限于研究论文全文的可获取性，原则上，它能把所有的科学数据与文献统一，创建数据和文献的交互操作。你可以正在阅读某人的一篇论文，然后去看原始数据，甚至可以重做数据分析。或者可以一边查看数据，一边查阅这一数据的所有相关文献。这样可以提高科学

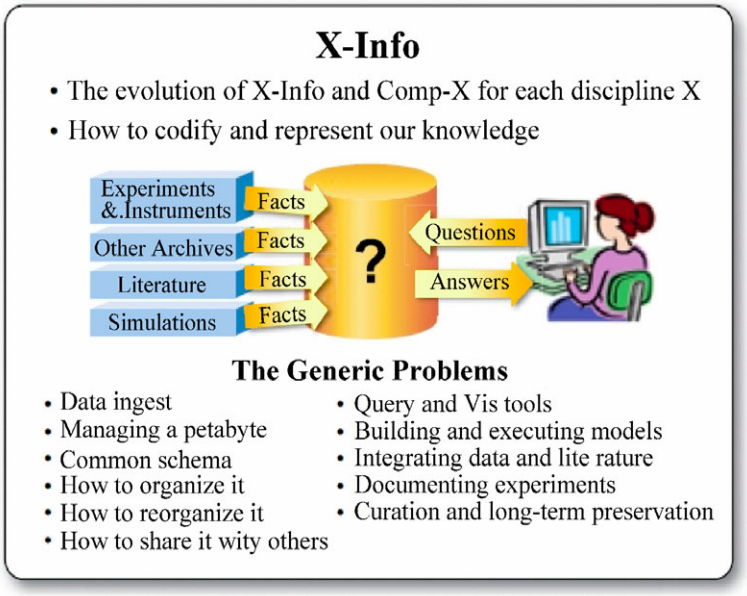


图 3 科研项目金字塔

的“信息速率”，并促进研究人员的科学生产力。

作者建议培育数字化的数据图书馆，即实际的、既为了数据也为了文献的真正意义上的数字图书馆；并建立开放存取的学术期刊（Overlay Journal），该观点认为：有数据档案库，也有文献档案库。文章就在文献档案库中存储，而数据则进入数据档案库中，从而促进学术交流。

4. 结语

吉姆·格雷认为，受信息技术的影响，几乎有关科学的所有事物都在变化。实证、理论和计算科学都受到数据泛滥的影响，因而出现了“数据密集型”科学模式。其目标是使世界上所有的科学文献联机，所有的科学数据联机，并且他们能实现可互操作。

编　译　本刊编辑部 郎杨琴、孔丽华

编译自　http：//research.Microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

吉姆·格雷在2007年时扩展了其对数据密集型科学的看法，提出七个重要行动领域：

- 支持、促进软件工具的开发；
- 从多个渠道获得资金，资助工具研发；
- 促进信息管理系统实验室的发展；
- 促进有关科学数据管理、数据分析、数据可视化和新型算法、工具的研究；
- 如同国家医学图书馆支持生物科学一样，建立更多数字图书馆以支持其他科学；
- 促进新的文档署名工具和出版模型的发展；
- 促进包含科学数据、已出版论著的数字数据图书馆的发展。