

# 为科学服务的大数据<sup>\*</sup>



郭毅可 潘 为 于思淼 吴 超 王世才

伦敦帝国理工学院数据科学研究所 伦敦 SW7 2AZ

**摘要** 数据驱动的科研活动已蔚为大观，然而厘清关于数据研究的基本问题仍是数据科学的首要任务。文章根据伦敦帝国理工学院建设数据科学研究院的经验，将数据科学聚焦于交叉研究上，讨论从数据整合与理解，到数据感知与交互，再到数据学习与认知，最后到数据交换与经济的完整链条，并结合开展的科研实践工作，分析了其中的基本研究问题。

**关键词** 大数据，数据科学，数据驱动的科学研究

DOI 10.16418/j.issn.1000-3045.2016.06.002

## 1 研究背景

微软研究院以 Jim Gray 曾经对科学研究方法的历史作了一个精辟的总结<sup>[1]</sup>：几百年前，科学研究是完全通过实验来观察自然、理解自然；到了近代数百年，科学才开始注重理论研究，通过建模和抽象来总结揭示自然的规律；近几十年来，计算机的广泛使用，使得计算模拟成了科学研究的一个重要手段。到了今天，计算技术已经完全普适化。科学仪器已经成为高通量数据采集的工具，由模拟和仪器采集的数据经过计算机的处理分析形成信息和知识。数据驱动已成为今天科学研究的新的方法。

如今，海量数据源源不断地被产生出来。科学家和工程师通过对数据的观察、整合、分析和解释，不断创造知识，推动着科学技术的进步和社会的发展。在这种背景下，在中国乃至世界各地，各类以数据为驱动或以数据科学为目标的研究单位如雨后春笋般涌现，在可预见的未来，数据驱动的科学研究必将得到蓬勃发展，蔚为大观。然而，在目前的探索阶段，厘清关于数据科学的基本问题仍然是首要任务，例如数据科学应该研究什么？它与传统计算机研究和统计分析到底有什么区别？它在学科交叉中应该扮演什么角色？本文根据伦敦帝国理工学院建设数据研究院（Data Science Institute）的实际经验，提出对如何建设一个支持以数据作为驱动为己任的数据研究院的见解，试图从我们的研究脉络中寻找共性问题，抛砖引玉，希望能在更广大范围内引起对这些基本问题的思考和讨论。

伦敦帝国理工学院是一所专注于科学技术、医学和商学的世界顶级名校。从事的科学

<sup>\*</sup>修改稿收到日期：2016 年 5 月 19 日

研究和数据紧密相关：从个人医疗数据到科学实验数据，从公共数据到商业数据。这样一个大学必须有一个数据研究所作为支撑学校数据驱动研究的科研机构。于2014年4月成立，其建所宗旨是：“研究先进的大数据管理和分析技术，并以此来促进数据驱动的科学研究及技术发展，造福人类社会。”它把自己的任务定义为：（1）作为学校交叉学科发展的枢纽，组织并推进以大数据为基础的多学科合作；（2）培养新一代有创新能力的科学家；（3）为学校的数据驱动的科学提供技术与设施的支持；（4）作为学校对外合作的窗口，与全世界工业界及学术界广泛开展大数据科研合作；（5）向政府、公共管理机构及全社会提供有关大数据的政策与技术咨询。

研究所自成立以来，秉承其宗旨，在上述5个方向上做出了许多努力，取得了令人瞩目的成果，得到了学界和社会的广泛关注和肯定，很多研究成果产生了国际影响力。因此，习近平主席2015年对英国进行国事访问期间专门参观了数据科学研究所，听取了一些研究成果汇报，包括：和浙江大学合作的对中国人口迁移的分析；和维也纳国际应用系统分析研究所、美国大气研究中心和上海大学合作有关“一带一路”战略国际影响力分析；和英国国家基因组计划、欧盟创新制药计划合作的有关精准医学的合作研究；以及和上海地铁在交通监测和预测方面的合作。习近平主席认为用大数据做交叉学科研究很有意义，和实际应用相结合是个好方向。习近平主席对我们的这些工作表示赞赏，肯定了研究所对大数据研究方向的思考和策略，使研究所倍受鼓舞。

## 2 数据驱动的交叉科学研究

科学技术的伟大进步往往需要多学科的交叉融合，数据科学的交叉同样会驱动产生重大的科学发现。而且我们认为数据科学无法作为独立学科存在，必须和特定领域结合在一起；如不对交叉学科领域知识有深入的理解，而设计脱离实际的数据分析方法是很难有发展前途的。

途的。

以目前热门的“精准医疗”为例，其涉及到生理学、分子生物学、药理学、化学、营养学、环境学、生物物理学等众多学科，很多学科在各自的领域对相关问题已经有了很长的研究历史，然而只有当交叉出现，特别是针对生物医学的大数据分析方法和工具出现之后，结合患者生活环境、生物信息、临床和药物等各种数据，实现精准医疗才有可能。

由此可见，数据科学是一个组合体，它在明确的应用目标下，驱动和连接各种学科，形成有机统一。把数据科学作为统计学和计算机科学的分支应用，把机器学习和大数据管理技术等数据科学的具体技术作为数据科学的主要内涵的思路与做法，未免是太狭隘了。

进而言之，数据科学的许多方法也来自于不同领域的科学研究，以今天非常流行的深度学习技术为例，它的许多进步是基于神经生物学和信号处理技术的研究。从数据驱动的领域科学研究中获取养料和动力，是数据科学研究的一个重要途径。

数据科学有自己的学科内涵，即基于数据的获取，清理、建模、分析等方法，从这个角度说，数据科学与数学及计算机科学一脉相承；它也有自己的外延，即面向各种应用问题，从这个角度说，数据科学又是各个交叉科学的载体。在后文中，我们将结合数据科学的内涵，即其研究问题，以及外延，即其应用领域，谈谈我们的理解。

## 3 伦敦帝国理工学院数据科学研究所研究方向

数据科学研究是一条完整的链条，由4个关键环节串联在一起。我们将这4个环节定义为数据整合与理解（Data Integration and Understanding）、数据感知与交互（Data Sensing and Interaction）、数据学习与认知（Data Learning and Cognition）、数据交换与经济（Data Exchange and Economy）。伦敦帝国理工学院数据科学研究所在这4个方面同时开展研究，并且将几方面的研究紧密地整合到一起。下面具体地阐释每部分的研究内容。

### 3.1 数据整合与理解

一份数据，从采集到分析，需要经历一系列的处理、理解和整合，这部分的工作，毫不夸张地说，可以占到整个数据研究工作量的80%。

(1) 在数据整合与理解方面，数据集成是大数据研究的关键。众所周知，数据的多样性和复杂性往往使得无法将所有数据进行整合，并为领域内的所有研究人员所共同使用。很多拥有相同实验目的的结果数据无法相互兼容。例如，在生命科学领域，在利用mRNA分析基因表达的过程中，基因芯片产生的表达程度数据通常用CEL格式存取，而如果使用mRNA测序技术则会产生大量基因序列的原始片段。两种数据都可以通过各自的计算方法得到基因表达的程度，但数据的格式天差地别，专业的分析人员也需要借助多种不同的技术分析汇总其中的结果，让计算机对此做出统一正确的理解可以说是困难重重。随着信息需求不断发展和增长，数据一体化的需求也不断增长。适当的标准化方法可以有效帮助数据的集成，标准化方法往往取决于数据集和特定领域的惯例，标准分数和T-统计量是转换医学研究中常用的标准化方法。

(2) 现有的数据集成技术，如本体论，语义Web可以起到关键的作用。这些现有语义框架和技术可以被用来建立各种数据之间的联系，并通过已有的映射关系拓展并建立新的联系。例如，对于医疗数据，可以通过预定义的、映射一致的本体森林模型来为临床数据和分子分析数据提供一个更加统一的数据表示，每一棵子树都表示一个研究项目，通过拓展子树节点之间的语义关系建立联系，获得新的语义知识。新的知识可以是拥有相同或相似病理特征的人的集合，或是治愈某种疾病的治疗方法的集合。

(3) 对数据标注，整理和ETL (Extract, Transform, Load) 自动化的研究是大数据研究的重要课题。ETL，用来描述将数据从来源端经过提取(extract)、转换(transform)、加载(load)至目的端的过程，也是对数据集成各个过程的集成和自动化过程<sup>[2]</sup>。ETL通过提取和转换完成数据清洗、标准化和语义建模的过程，使原始数

据转换成人、机都能理解的有效信息。ETL的核心在于减少繁复的数据预处理中的人工干预，自动化完成数据整合的各个步骤。其难点在于通过人工智能的方法对原始数据进行自动化标注，并利用语义分析的方法将被标注的对象加入语义网络。

(4) 对于数据的标准化和统一化，质量控制是关键技术。在标准化的过程中，需要特别重视数据质量控制。仍以mRNA分析基因表达为例，相对于基因芯片产生的少量高质量数据，mRNA测序技术产生基因序列数据量较大，但可靠性较差。通常的基因表达分析结果中都需要加注每个基因序列片段分析结果的质量，对于质量较差的片段，通常的分析中一般不予采用。

我们主持的“欧洲转化医学信息与知识管理服务”(European Translational Information & Knowledge Management Services, eTRIKS)项目就是以数据标准化和质量控制为目标的一个典型的数据质量工程。eTRIKS是由欧洲创新药物计划(Innovative Medicines Initiative)发起的5年科研总经费达2300万欧元的研发项目，由世界12大制药厂参与，旨在建设基于云计算的全欧洲范围内的医学研究标准大数据平台，成为欧盟医学临床研究的大数据标准。由全球性非盈利性组织 tranSMART 基金会主导开发的知识管理平台是eTRIKS平台的核心系统。它以系统级的方法来解决数据集成和理解的问题，其具体架构如图1所示。

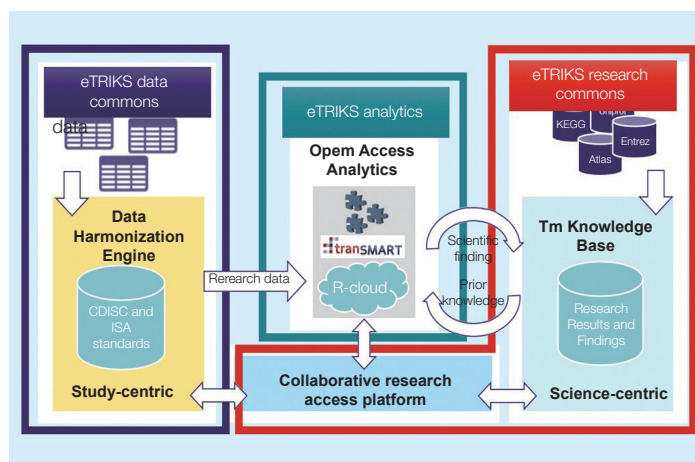


图1 欧洲转化医学信息与知识管理服务(eTRIKS)项目技术框架图



此平台上的研究主要包括生物信息数据联邦、高效数据存储架构设计以及相关数据索引技术。生物信息数据联邦主要用于解决生物信息的多元化带来的异构信息抽象和整合等问题，使得各种数据源可以依据其自身特点，以各自特有的模式进行低成本、高效率存储和处理。例如，基因芯片所产生的数据主要存储在 CEL 格式的元信息矩阵和数据信息矩阵中，高通量测序数据结果多存储在 FASTA 或 FASTQ 文件中，而单核苷酸多态性统计数据多以关系型数据库模型存储。一个复杂的病理研究通常需要综合多种来源的各种信息共同计算，而数据联邦通过抽象和整合这些多元数据，使得这种基于混合数据结构的高效海量数据计算成为可能。

在考虑多种信息集中处理的同时，我们也关注于对各类数据存储结构的优化<sup>[3,4]</sup>，通过引入先进的存储技术提高数据的存取效率。例如，数据科学中心设计实现的 CGC 索引（Collaborating Global Clustering Index）是针对遗传信息的高效数据存储和检索方法。

### 3.2 数据感知与交互

随着传感器技术及其产业的发展，传感网络大规模地被应用于收集不同领域的的数据<sup>[5]</sup>，其进一步所带来的普适感测促进了物联网这个新兴领域的发展<sup>[6]</sup>，带来了广阔的未来潜在应用，包括产品追踪、智慧环境、社会感知、智能设备、灾害预测等等<sup>[7]</sup>。面对感知大数据，如何构建针对物联网的通用高性能数据处理平台，及研究针对物联网和大数据感测的高性能数据管理方法成为关键。

在这方面，数据科学研究所提出了“认知感知”的方法论，认为感知数据的作用在于建立、验证和纠正模型。一旦一个目标感知对象被建模之后，其模型预测将与感知数据进行比对，如果模型正确，则无需进一步数据采集和模型修正；如果模型失效，说明目标对象出现新的行为或原模型粗糙，这时才需要进一步采集数据并修正模型。这种方法被叫做“认知感知”是因为它契合智能生物感知世界的方法，智能生物包括人类能在有限认知计算资源的限制下实现与动态环境的均衡，

其目标可以说是优化自由能量（Free energy）或最小化惊奇（Minimize surprise）<sup>[8]</sup>。基于这种认识，我们在感知系统中，将认知定义为优化主观认知分布和客观分布之间 KL 距离的建模行为，而感知行为被看做是减小此 KL 距离与实际 KL 距离的措施。为了实现这种感知和认知，我们解决了两方面的问题：如何调整模型和模型空间来适应感知对象的变化；如何减少感知维度。

感知之后的数据除了分析建模之外，一个重要的研究方法是数据可视化。数据可视化是研究如何将数据以形象化的方式展现出来的一门科学。它主要专注于分析，以连贯和简短的形式把大量的信息展现出来，而抽取何种数据进行形象化的抽象，本身就蕴含了对数据如何应用的科学思维。在大数据背景下，大规模的多维的数据正在被快速地产生和积累。如何更有效地探索数据、理解数据以及表达数据成为一项重要的研究课题。

通过图形化地表达数据，人可以利用自身复杂的视觉系统直接参与到数据探索和交流的过程中。这使得很多复杂的数据可以更有效地被分析和理解。数据可视化成为数据科学的重要组成部分的主要原因有两个：第一，由于人类视觉系统十分擅长模式识别，通过图形可视化数据以及相关的分析结果，可以更容易更准确地理解数据中的有效信息。第二，数据可视化技术可以很大程度地帮助人们交流和传播大数据所蕴含的有效信息和重要发现。

由此可见，可视化不是数据分析的结果，而是数据分析的过程。如何建立一个能支持发现科学直观的可视化环境是非常重要的，在这方面我们做了大量的工作，建立了全球最大的数据可视化设施“全球数据观察站”（图2），几十个电脑屏幕组成的动态数据图像准确衔接，其背后蕴含的是并行运算、多项目管理、编程，以及对数据的深刻理解。在数据观察站中实现了各种实时交互的可视化应用，比如全球比特币交易的实时数据可视化，个性化医疗系统可视化，上海地铁运行分析的数据可视化等，实时处理和展示随时间变化的各种类型的数据。

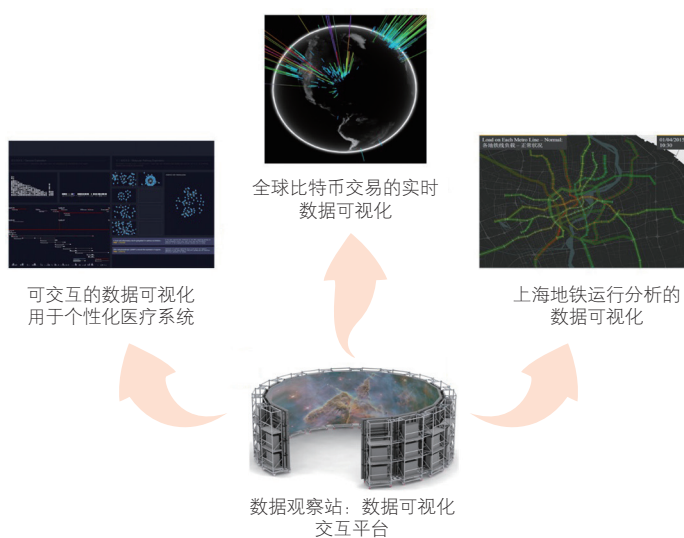


图2 数据可视化平台及应用

我们所处理的数据源不再是静态的，它会随着数据的实时变化进行即时接受、处理并更新可视化数据。这种方式的数据可视化可以帮助人们观察到最新的即时数据并理解其对旧数据产生的影响。可交互的数据可视化分析可以让人利用可视化信息与系统进行交互，并在此过程中进一步得到相关信息提取和挖掘的结果。在这个交互过程中，人可以在充分理解数据可视化信息的基础上，根据不同的目的主动探索和发现所需要的数据结果。这可以极大地提升数据探索和挖掘的效率。

与此同时，人在与数据可视化进行交互的过程中也可以产生新的数据。这些数据可以被收集和分析，以学习人与系统交互的具体情况。例如，在数据观察站我们可以利用眼动追踪设备来实时记录人眼在屏幕上注视点的位置。这些位置点形成的热点图可以清晰地展示出人对于数据可视化最关注的部分。这些数据可以帮助我们设计和创造出更有效的或者更吸引人的数据可视化系统。同时，人的眼动情况也提供了一种新的与系统交互的途径。系统可以通过探测人关注点的具体位置进行实时更新，例如展示额外信息等。新的交互方式毫无疑问会对人与数据可视化系统的交互产生积极的影响。

### 3.3 数据学习与认知

研究所进行数据学习与认知研究是从实际出发，为

了应用服务的机器学习。没有应用背景的数据科学研究会缺乏影响力，没有对数据科学理论的扎实研究也做不出好的应用。我们关注的实际问题包括功能核磁共振或者脑电图推断有效连接（effective connectivity）脑网络；基于微流控技术得到的蛋白质荧光表达推断基因网络结构；印度西北部平原地下水水位趋势变化预测；中国省会城市交通网络车辆速度和流量的预测；计量经济学中经济变量的因果性推断。这些科学问题都是由数据驱动的研究，而这些问题中的数据都可以用时间序列来描述。时间序列模型的主要目的是对系统的物理本质有洞察力的解释和根据已有的历史数据对未来进行预测。

基于贝叶斯理论和数据同化理论，我们团队致力于开发贝叶斯学习引擎（Bayesian Learning Engine）进行时间序列数据建模。贝叶斯学习引擎由两部分构成：大数据建模（Data Modelling）引擎和大数据同化（Data Assimilation）引擎。数据建模和数据同化用来做模型筛选的思想可以总结为同化学习理论（Assimilated Learning）。

大数据建模引擎基于贝叶斯理论构建，其实现分为如下步骤：（1）确定数据的似然函数。（2）选取适当的模型结构。一方面由于所研究的科学问题所在领域的特点不同，选取的模型结构往往具有很大的差别，而且往往是非线性的。比如在生化网络和基因网络中，模型中方程必须要遵循化学反应动力学原理，也就是模型的形式只能用多项式和有理函数来描述；比如在描述天气系统、生态系统的混沌震荡系统中，模型也往往是具有多项式形式；而在描述电力系统、通讯网络系统时，模型一般具有三角函数形式；在脑网络的动力因果模型（Dynamic Causal Model）描述中<sup>[8]</sup>，函数的形式限制于一阶和二阶多项式形式。即便是具备了一定的领域知识，由于非线性函数形式的无穷性，模型空间维数仍然极高。另一方面，如果系统具有高维的状态变量，比如基因网络中的基因数目，那么情形会更加严重，模型选择将面临很大的挑战。（3）根据先验知识和系统的特点构造先验概率，用于刻画模型中隐藏变量的不确定程

度。而这个不确定程度往往由超参数刻画。值得注意的是，超参数的个数往往小于或者等于候选模型中的隐含变量个数。

接下来我们对后验概率积分获得边缘似然函数，通过对其分析，一个令人喜悦的发现是对于不同的先验概率构造，我们只需求解一系列的平滑函数加变权重L1范数规则化优化问题<sup>[9]</sup>。而这类优化问题的集中化解法或者分布式解法已经被广泛地研究，基于不同的分布式计算平台与计算架构，比如 MapReduce、Hadoop、Spark/Shark 可以比较直观地实现并行化。

除此之外，模型选择依然面临着其他问题。首先，这类优化问题的一个问题是对规则参数的调试，不同的规则参数下会得到不同的模型。另外，如果起始选择了不同的候选模型，最后优化得到的模型往往更加不唯一。而且模型选择原则，比如赤池信息量准则（AIC）和贝叶斯信息量准则（BIC）往往相差不大，导致模型很难区分。

数据同化技术<sup>[10]</sup>可以对数据引擎得到的模型集合进行在线筛选。它能帮助一个动态模型不断地将观测数据的有用信息反馈进入原有的模型中，一方面能改良无法观测的物理量，从而不断地把模型的（预测）输出逼近现实，另一方面可以不断地修正模型，在线做出模型选择。

### 3.4 数据交换与经济

大数据时代的到来，不仅仅意味着更多数据被收集和被处理，更为重要的是，数据实实在在成为改变个人和社会的力量。众多案例<sup>[11-13]</sup>已向我们展示了大数据的应用价值，然而一个技术要深刻地推进社会发展，它需要从具有应用价值发展为具有“应用+经济”的双重价值。

从经济价值的眼光来看大数据，我们可以看到所谓的“数据”在整条价值链上处在起点的位置。数据从一开始作为原材料，到最后成为产品提供给用户，其中经历了一系列的加工和增值过程，包括清理<sup>[14]</sup>、语义化<sup>[15]</sup>、融合<sup>[16]</sup>、分析<sup>[17]</sup>、建模<sup>[18]</sup>、知识提取<sup>[19]</sup>、应用<sup>[20]</sup>、分发<sup>[21]</sup>等关键步骤，如同一个工业产品，从原材

料到最终产品形态再到市场，是一个复杂的价值链，需要精巧的协同工作。而在目前大部分的大数据研究中，关注点还在于这些具体过程的技术基础，我们相信随着整个大数据生态环境的建立，每个步骤背后的经济因素将成为最大的推动力量。

要推动从数据到数据产品的价值链，有很多关键的经济问题需要考虑，其中一个核心的问题是数据作为资产的定价问题。数据与其他原材料在4个方面有很大不同：（1）数据的使用不会带来数据的消耗，它的开发不是排他的，甚至反而是利他的；（2）聚合后的数据比单独的数据更有价值，也应该具有更高的价格；（3）同样种类的数据，不同来源的数据具有不同的价值，这点在医疗数据中尤为突出；（4）同样的数据在不同的使用者看来，也是价值各异。在这些特殊的条件中，如何对数据资产定价是一个很难的问题，我们认为采用一种基于市场协商的价格或许更为现实可行。

有了定价，还需要交易。目前很多概念仍需考察，例如交易是代表了数据所有权的转移？还是仅仅出让了使用权？数据作为一种容易复制和分发的资产，如何控制其再交易？另外一方面，定价和交易的问题同样存在于整个数据价值链上，例如对数据产品如何定价？目前基于app的交易模式是否是最合理的？

解决这些核心问题，有利于找到适合大数据产品和大数据经济的商业模式。目前很多商业模式初现雏形，例如基于众包的数据收集和基于用户数据收集的精确广告等。然而很多商业模式其经济模型暧昧不明，在数据定价、用户隐私等方面缺乏明晰思考和监管。总体来说，整个价值链上的商业模式尚处起步，大有研究和发展空间。

由大数据经济推动的各个参与者（数据提供者、加工者、产品开发者、发布商、用户等）最终会形成一个生态环境。一个好的生态环境会促进各个参与者的效益和效率，并提高从技术到效能再到效益的转化。目前此生态环境初见雏形，但在很多方面缺乏体系支持。以隐私为例，目前在用户和数据收集者之间缺乏一个有效的



隐私保护机制。针对这个问题，我们提出了一种新的移动隐私保护模型（Pay-by-Data, PbD模型<sup>[22]</sup>），用于控制以下这类常见问题：在目前的机制下，手机应用可以在用户不知情或无力控制的情况下，获取用户大量移动端数据。在 PbD 模型中，定义了一种新的应用价格，即数据；并建立了一种新的开发者与用户之间的关系，使得用户可以对他们的数据有更强的控制。模型让用户知道他们哪些数据被收集，而这些用户数据的使用也被显式地告知用户，并通过新的粒度更低的认证机制来控制。此模型同时使得用户可以从数据交易中获得奖励。这种显式的数据-服务交换使得我们可以建立一种以市场机制为调节手段的数据定价和交易方法。在过去的两年中，我们团队完成了 PbD 的计算模型并完成了其原型系统，包括 PbD 市场、数据交易价格、PbD 开发 SDK 和一个定制的 PbD Android 操作系统。

其他的支持体系包括法律、知识产权等方面，其中一个有意思的方向是科学领域的的数据知识产权，或者说数据出版。这个问题涉及到科学数据如何被开发利用，尤其是在学界之外的开发利用。这其中同样有经济模型的问题，例如科研经费如何对数据获取、处理和发布进行支持，以及如何建立对数据科学家的声誉和激励，从而在科学数据领域形成良好生态。我们在此领域做了一些初步工作，进行了一个大规模的数据出版调查，并出版了第一期的数据出版调查报告，调查围绕数据出版话题，侧重从数据出版动机、数据出版方式、数据出版运营模式以及数据出版质量评价 4 大维度出发，来了解世界范围内科学研究领域科学家对于科学数据出版相关内容的看法和态度，并针对数据出版的意义价值及其操作层面的诸多问题予以探讨，以期全面了解数据出版发展现状，并试图探索推进数据出版事业未来发展、为促进科学数据交流共享提供积极建议。

## 4 结语

大数据为人类社会提供了又一次新的资源机遇，其

具有已有自然资源所不具备的许多特征。如它的超可再生性——数据的使用本身并不消耗数据，相反，还会产生新的数据；它的非竞争性使用——一方对数据的占有并不限制其他人对这份数据的拥有。这些特征使得数据资源的使用不仅可以像其他的自然资源一样产生能量与财富，而且可以完全改变人类的社会组织结构和行为方式。所以，对数据科学研究必须站在社会发展、新的经济模式、新的工业体系、新的创新产品、新的生活方式以及新的科学研究的方法等宏观角度来进行系统化的科学研究。

## 参考文献

- 1 Hey T, Tansley S, Tolle K, et al. The fourth paradigm: data-intensive scientific discovery. General Collection, 2009, 317(8): 1.
- 2 Vassiliadis Panos. A survey of Extract-transform-Load technology. International Journal of Data Warehousing and Mining, 2009, 5(3): 1-27.
- 3 Wang S, Pandis I, Wu C, et al. High dimensional Biological data retrieval optimization with NoSQL technology. BMC Genomics, 2014, 15 (8): 1.
- 4 Wang S, Pandis I, Johnson D, Emam, et al. Optimising parallel R correlation matrix calculations on gene expression data using MapReduce. BMC Bioinformatics, 2014, 15 (1): 351.
- 5 Zhu T, Xiao S, Zhang Q, et al. Emergent Technologies in Big Data Sensing: A Survey. International Journal of Distributed Sensor Networks 2015, 2015: 1-13.
- 6 Zaslavsky A, Perera C, Georgakopoulos D. Sensing as a Service and Big Data. Proc. Int. Conf. Adv. Cloud Comput. Doi: arXiv: 1301.0159.
- 7 Aggarwal C C, Ashish N, Sheth A. The Internet of Thinys :A Surrey from the Data Centric Perspective, Managing and Mining Sensor Data. 383-428 (2014). Doi:10.1007/978-1-4614-6309-2\_12.
- 8 Friston K J, Harrison L W. Dynamic causal modelling. Neuroimage, 2010, 5 (4): 1273-1302.

- 9 Pan W, Yuan Y, Goncalves J, et al. A sparse bayesian approach to the identification of nonlinear state-space systems. IEEE Transaction on Automatic Control, 2015, 61 (1): 1.
- 10 Evensen G. Data assimilation: the ensemble Kalman filter. Springer Science & Business Media, 2009.
- 11 Ahnn J H. Big data computing for the personalization of services and its applicaiton to speech recognition. International Symposium on Big Data Computing, London, 2015.
- 12 Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity. Analytics, 2011.
- 13 Andrew M A, Erik B., et al. Big data: the management revolution. Harvard Business Review, 2012, 90 (10): 60-67.
- 14 Rahm E, Hong H D. Data cleaning: Problems and current approaches. IEEE Data Engineerhy Bulletin. 2000, 23 (23): 3-13.
- 15 Auer S, Bizerc, Kobilaror G, et al. Dbpedia: A nucleus for a web of open data. The Semantic web. Springer Berlin Heidelberg, 2007, 4825: 722-735.
- 16 Hall D L, James L. An introduction to multisensor data fusion. Proceedings of the IEEE, 1997, 85 (1): 6-23.
- 17 Trnka A. Big data analysis. European Journal of Science and Theology, 2014, 10 (1): 143-148.
- 18 Wu X D, Zhu X, Wu G Q, et al. Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 2014, 26 (1): 97-107.
- 19 Chen H, Chiang R HL, Storey V C, Business Intelligence and Analytics: From Big Data to Big Impact. MIS quarterly 2012, 36 (4): 1165-1188.
- 20 Murdoch T B, Detsky A S. The inevitable application of big data to health care. Jama the Journal of American Mediael Association. 2013, 309 (13): 1351-1352.
- 21 Naimi A I, Westreich D J. Big data: A revolution that will transform how we live, work, and think. Information Commwnicotion & Society, 2013, 17 (1): 181-183..
- 22 Wu C, Guo Y K. Enhanced user data privacy with pay-by-data model. 2013 IEEE International Conference on Big Data, 2013: 53-57.

## Big Data for Better Science

Guo Yike Pan Wei Yu Simiao Wu Chao Wang Shicai

( Data Science Institute, Imperial College London, London SW7 2AZ, UK )

**Abstract** Data driven scientific research has now gain great prosperity. However, we believe that the principle task of data science is to understand the basic problems within data research. In this paper, based on our experience in building the Data Science Institute in Imperial College London, we consider data science as the core of interdisciplinary research, and discuss the whole pipeline of data science research, including data integration and understanding, data sensing and interaction, data learning and cognition, and data exchange and economy. We discuss these basic scientific problems based on our practices in practice. We hope the work presented in this paper can bring thinking and discussion in a larger scale.

**Keywords** big data, data science, data-driven scientific research

**郭毅可** 英国帝国理工学院计算系教授，帝国理工学院数据科学研究所所长。1985年毕业于清华大学计算机系，获工学学士学位。1993年在英国帝国理工学院获得计算机博士学位，博士期间研究方向为计算逻辑及陈述性语言编程，其毕业论文获1994年英国帝国理工学院最佳博士毕业论文。2002年被聘为帝国理工学院计算机系终身正职教授，在当时



是英国最年轻的教授之一。其主要研究领域包括大数据管理与分析、分布式数据挖掘、网格计算、云计算、传感器网络及生物信息学等。1999年创立了帝国理工计算系的第一个衍生公司InforSense，并于1999年至2008任该公司首席执行官。InforSense有限公司于2009年6月为国际知名科学数据管理公司英国IDBS公司并购，迄今他一直担任IDBS公司首任首席创新官。2012出任全球性非盈利性组织tranSMART基金会的首席技术官。2011年至2013年担任清华大学信息科学与技术国家实验室讲席教授。2012年成为首批上海市千人计划入选者、上海特聘专家，并为北京市人民政府“海外人才工作顾问”。现任上海市产业研究院大数据首席科学家，中科院深圳先进技术院健康大数据中心主任，及上海大学计算机学院院长。E-mail: y.guo@imperial.ac.uk

**Yike Guo** Professor of Computing Science in the Department of Computing at Imperial College London. He is the founding Director of the Data Science Institute at Imperial College, as well as leading the Discovery Science Group in the department. Professor Guo also holds the position of CTO of the tranSMART Foundation, a global open source community using and developing data sharing and analytics technology for translational medicine. Professor Guo received a first-class honours degree in Computing Science from Tsinghua University, China, in 1985 and received his PhD in Computational Logic from Imperial College in 1993 under the supervision of Professor John Darlington. He founded InforSense, a software company for life science and health care data analysis, and served as CEO for several years before the company's merger with IDBS, a global advanced R&D software provider, in 2009. He has been working on technology and platforms for scientific data analysis since the mid-1990s, where his research focuses on knowledge discovery, data mining and large-scale data management. He has contributed to numerous major research projects including: the UK EPSRC platform project, Discovery Net; the Wellcome Trust-funded Biological Atlas of Insulin Resistance (BAIR); and the European Commission U-BIOPRED project. He is currently the Principal Investigator of the European Innovative Medicines Initiative (IMI) eTRIKS project, a €23M project that is building a cloud-based informatics platform, in which tranSMART is a core component for clinico-genomic medical research, and co-Investigator of Digital City Exchange, a £5.9M research programme exploring ways to digitally link utilities and services within smart cities. Professor Guo has published over 200 articles, papers and reports. Projects he has contributed to have been internationally recognised, including winning the “Most Innovative Data Intensive Application Award” at the Supercomputing 2002 conference for Discovery Net and the Bio-IT World “Best Practices Award for U-BIOPRED in 2014. He is a Senior Member of the IEEE and is a Fellow of the British Computer Society. E-mail: y.guo@imperial.ac.uk