DOI:10.1145/2611567

探索在挖掘大数据潜力时面临的内在技术挑战

作者：H.V. JAGADISH, JOHANNES GEHRKE，
ALEXANDROS LABRINIDIS，YANNIS PAPAKONSTANTINOU，
JIGNESH M. PATEL，RAGHU RAMAKRISHNAN，
CYRUS SHAHABI

大数据及其技术挑战

在广阔的应用领域中，数据正在以前所未有的规模增长。以前，决策基于猜测或人工构建的模型，费时费力；现在，人们使用数据驱动的数学模型做出决策。此类大数据分析现在几乎驱动了社会各领域的进步，包括移动服务、零售、制造、金融服务、生命科学和物理科学领域。

举例来说，科学研究因为大数据已经发生了根本变革。[1,12] 斯隆数字巡天（Sloan Digital Sky Survey）[23] 已经变革了天文学；之前天文学家的大部分工作是拍摄天空的图片；现在天文学家的工作是从数据库中找出感兴趣的对象和现象，因为照片已经存放在数据库中。在生物科学领域，把科学数据存放入公共的存储库现已成为一种约定俗成的习俗。该习俗也包括创建公共数据库供其他科学家使用。不仅如此，随着技术日益进步，特别是在下一代测序（NGS）出现后，可用的试验数据集的规模和数量均呈指数级增长。[13]

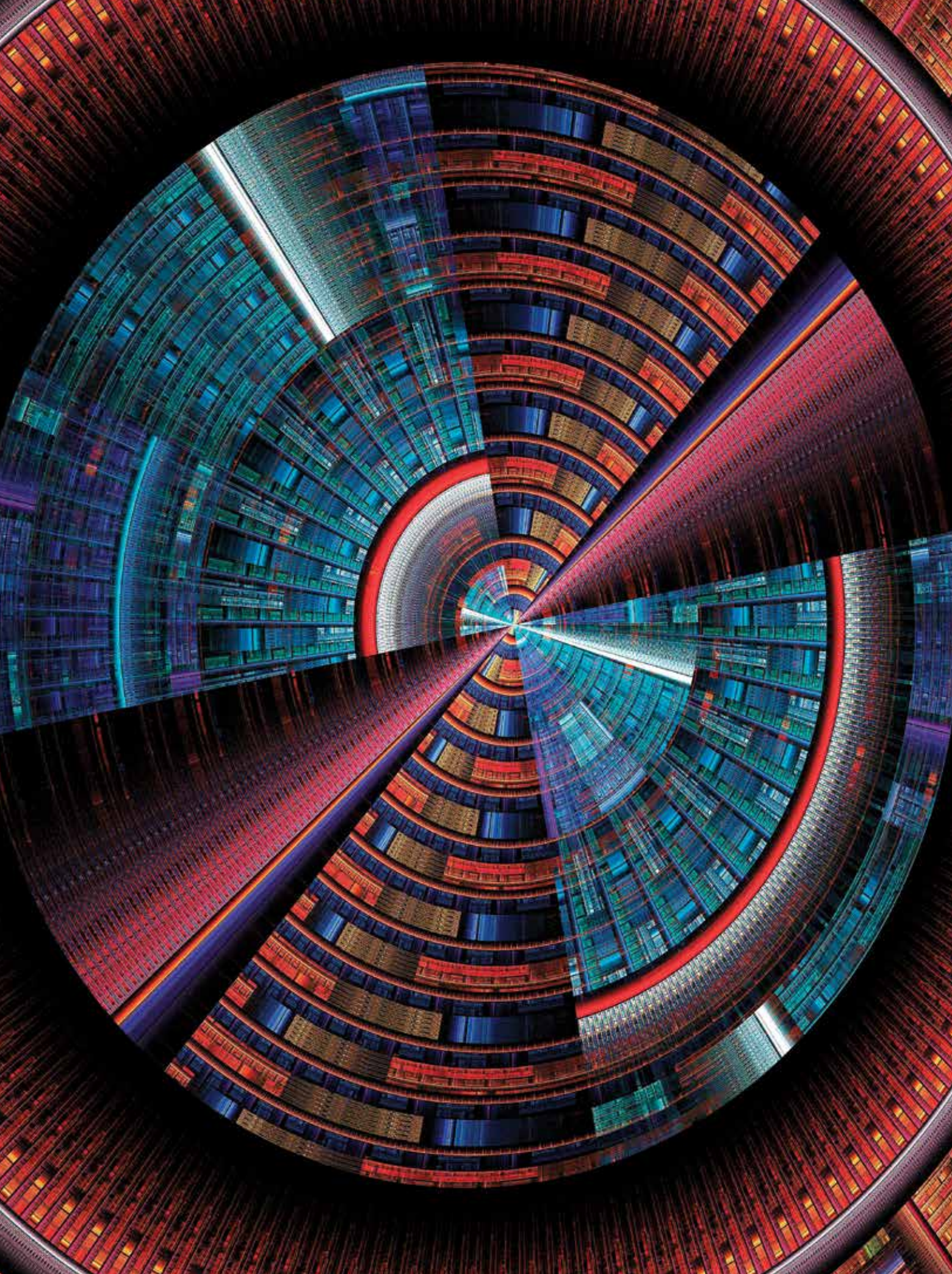
按每台*单独的* NGS 机器产生的原始序列数据计算，当前 NGS 方法的输出增长速度如图 1 所示，图中还描绘了 SPECint CPU 基准的性能增长情况。很明显，对于单线程应用（本文中的 SPECint）来说，NGS 序列数据的增长远远超过了摩尔定律提供的性能增长。请注意，图 1 中的序列数据大小为分析 NGS 仪器实际产生的原始图像后得出的输出结果。由于这些原始图像数据集本身规模太大（每天每个实验室产生数个 TB 的数据），即便是在现在，考虑保存他们也不切实际。而且，序列数据是在实时分析图像时产生并保存的。

大数据不仅对科学研究带来巨大的变化，而且有潜力在其它方面带来更大的变革。Google 对谷歌文件系统（Google File System）和 MapReduce 的研究以及随后出现的像 Hadoop 这样的开源系统已

» 重要见解

- **大数据正在彻底改变我们生活的所有方面，从企业到消费者，从科学到政府，均经历着根本性的变革。**
- **从大数据中创造价值是一个包含下列多个步骤的流程：采集，信息抽取和清理，数据集成，建模和分析，以及解释和部署。许多对大数据的论述只关注了一两个步骤，却忽视了其他的步骤。**
- **研究想光的挑战很多，范围包含从数据的异质性、不一致和不完整、及时性、隐私、可视化效果和协作到围绕大数据形成的工具生态系统等多个方面。**
- **很多案例表明，能够正确驾驭大数据的人将会赋予丰厚的奖赏。**

图片由 AMAZEME GRAPHICS 提供



经引发了业界对大数据技术最广泛的开发和应用。专注于 Web 的公司，如脸谱网（Facebook）、领英（LinkedIn）、微软、Quantcast、Twitter 和雅虎等公司引领了这一潮流。它们已经成了众多应用中不可或缺的基础，涵盖了从网络搜索到内容推荐和计算广告学等各种领域。在下列领域中，利用大数据价值的真实案例已然出现，且颇具说服力：医疗保健（通过基于家庭的连续性监测和跨供应商集成）、[3] 城市规划（通过融合高保真地理数据）、智能交通（通过分析和可视化展现实时的详细路网数据）、环境建模（通过无处不在的传感器网络收集数据）、[4] 节能（通过揭示使用模式）、智能材料（通过新材料基因组计划 [18]）、自然语言之间的机器翻译（通过分析大型语料库）、教育（特别是在线课程）、[2] 计算社会学（一种越来越热门的新方法论，因为获取数据的成本大大降低）、[14] 金融领域的系统性风险分析（通过集中分析大量的合同来找出金融实体之间的依赖关系）、[8] 国土安全（通过分析社交网络和潜在恐怖分子的金融交易）、计算机安全（通过分析日志记录的事件，也被称为安全信息与事件管理 SIEM）等。

图 1 下一代序列数据的规模与 SPECint 的对比

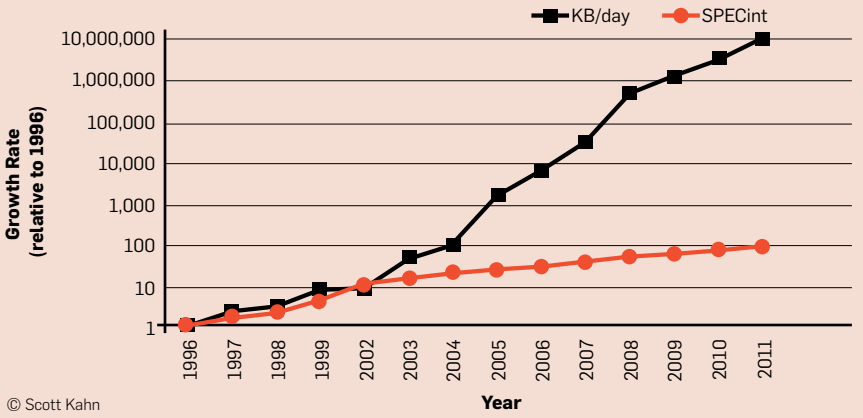
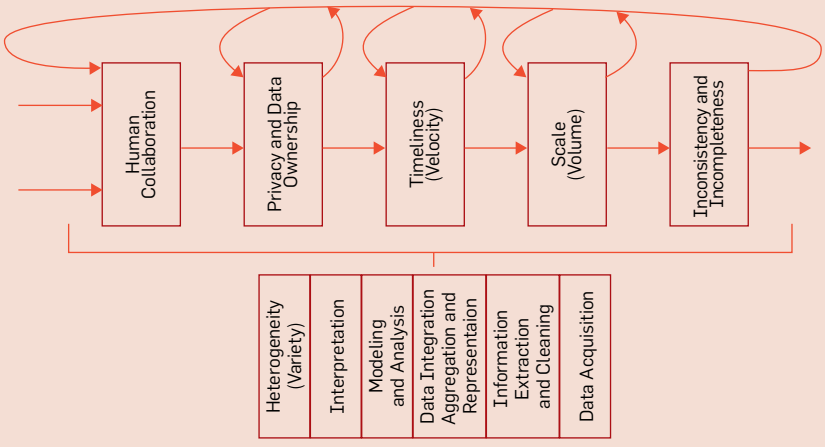


图 2 大数据分析管道图的上半部分说明了大数据分析的主要步骤。注意，在所有的阶段中均存在可能的反馈循环。图的下半部分说明了造成这些步骤颇具挑战的大数据特性。



2010 年，企业和用户存储了超过 13 百亿亿字节的新数据；这是美国国会图书馆的数据的 50,000 多倍。根据麦肯锡最近发布的报告，对于终端用户而言，全球个人位置数据的潜在价值估计有 7 千亿美元，它可以让产品开发和组装成本最多降低 50%。[17] 麦肯锡预测，大数据在就业方面也会造成同等规模的巨大影响，其中美国将需要 140,000 – 190,000 名拥有“深入分析”经验的员工；不仅如此，150 万名经理将需要精通数据。不令人吃惊的是，美国总统科技咨询委员会最近发布了一份有关网络化与 IT 研发的报告 [22]，其中把大数据确定为“前沿研究方向”，它能够 " 快速推动范围广泛的重点项目取得进展。” 现在，即使是大众新闻媒体也开始认识大数据的价值，*《经济学人》*、[7] *《纽约时报》*、[15,16] *美国全国公共广播电台*、[19,20] 和*《福布斯》*杂志的相关报道可以证明这点。[9]

虽然大数据的潜在利益是真实又巨大的，而且也取得了一些初步的成功（比如斯隆数字巡天），但为了全部实现这一潜力，必须解决很多尚存的技术挑战。海量的数据当然是一个主要的挑战，也是最容易识别的挑战之一。不过，还有其他挑战。行业分析公司会指出，不仅在*数量*上存在挑战，而且在*多样性*和*速度*上也存在挑战 [10]，所以公司不应该只重视其中的第一点。多样性指数据类型、表现形式和语义解释的异质性。速度有两种含义，一种是数据到达的速率，另一种是必须对数据处理的速率。虽然上述三点相当重要，但是这份简短的清单却未能覆盖其他重要的要求。各方已经提出了几项其他要求，比如*真实性*。其他的顾虑，如隐私和可用性，仍然继续存在。

大数据的分析是一个迭代的过程，每次迭代都会面临自己的挑战，也会涉及很多独特的阶段，如图 2

所示。在此，我们探讨一下端到端的大数据生命周期。

大数据生命周期的各阶段

不幸的是，很多人只关注分析 / 建模步骤 - 虽然这一步特别重要，但是如果缺少数据分析管道中的其他阶段，这一步几乎无用。例如，我们必须从数据在我们可能无法完全预料的方面存在价值这个角度来探讨应该记录哪些数据的问题，然后开发各种方法来从捕获的不完美和不完整的数据中推导出价值。这样做之后，便产生了追踪来源以及处理不确定性与错误的需求。让我们来看另一个例子，如果相同的信息以重复或重叠的方式呈现，那么它支持我们利用统计技术来应对数据集成和实体／关系抽取等各种挑战。这可能是成功利用多个数据来源数据的关键点（例如，不同的实验室报告的相关实验，众包的交通信息，从不同的网站提取出的特定领域的数据，比如娱乐数据）。上述方面对成功至关重要，但却极少得到像大数据那样的广泛论述。即使在分析阶段，虽然该阶段已经备受关注，但是研究人员对在多租户集群背景下多个用户程序并发运行时的复杂度却知之甚少。

在一处完成所有任务最重要的转变很可能是，随着相同的数据越来越多地经历生命周期中的所有五个阶段，对数据各阶段的处理任务分别独立进行将很难让人接收。我们如何才能提供一种数据管理和分析能力的综合集，为所有五个阶段提供足够的支持?

在本文的后续部分中，我们会首先讨论大数据管道中的五个阶段以及每个阶段中的特殊挑战。然后，我们会提供一个案例研究（见侧边条），用实例说明在不同阶段中面临的问题。在此，我们讨论一下六个相互交织的挑战。

数据采集。大数据并非从真空中产生：它是对潜在感兴趣的活动记录。例如，以我们感知和观测周围世界的能力为例，其中包括了从老年人的心率、呼吸的空气中存在的毒素到网站上的用户活动日志或软件系统中的事件日志等多种数据。如今，传感器、仿真器和科学实验能够产生大量的数据。例如，规划中的平方公里阵列望远镜每天能够产生高达一百万太字节（terabyte）的原始数据。

把概要推到边缘设备。我们可以筛选和压缩的东西往往与拟实现的分析紧密结合，所以固定的筛选策略效果不佳。我们能否提供灵活的复杂事件处理框架（这种框架会根据用户的分析把允许的筛选和压缩准则推向产生数据的边缘设备，进而实现数据采集的优化）?

在不影响我们对感兴趣的潜在活动进行推理的条件下，该数据中的很大部分将会以数分之一比率筛选和压缩。挑战之一是妥善定义这些“在线”筛选器，使得它们不会抛弃有用的数据，因为原始数据往往太大，实际上不允许我们全部存储。例如，收集的传感器数据在大多数情况下与空间和时间相关（比如相同路段上的交通传感器）。假设，某个传感器的读数与其他传感器存在明显区别。原因可能是传感器出故障了，但是我们怎么能确定这个读数没有实际意义呢?

不仅如此，加载大型的数据集通常也是一种挑战，特别是在结合使用在线筛选和数据精简的情况下；同时我们还需要有效的增量处理技术。即使如此，对于很多应用而言，这可能还不够，还需要设计高效的即时处理方法。

信息抽取和清理。收集的信息往往与分析时使用的格式不符。例如，以在医院里收集的电子健康记录为例，该记录包括多位医生的口述笔录，传感器和测量值的结构化数据（可能带有某种关联的不确定性），图片数据（如 X 光片）和探测仪的视频。我们不能让数据的格式保持现状，然后还要有效地对其进行分析。与此相反，我们需要一种信息抽取流程把所需的信息从底层数据源抽取出来，然后再以适于分析的结构化形式呈现。正确、完整地实现这一流程会一直成为技术挑战。此类抽取依赖应用的程度通常很高（例如，从一张核磁共振成像中抽取信息与从一张恒星的图片或监测照片中抽出有用信息截然不同）。生产率方面则要求创造新的声明性方法来精确地规定信息抽取任务，然后在处理新数据时对这些任务的执行进行优化。

众所周知，大多数数据源都不可靠；传感器可能会出现故障，人们可能会偏见的观点，远程的网站可能变旧（stale），等等。理解并为这些错误源建模是开发数据清理技术的第一步。不幸的是，工作的很多方面取决于数据源和应用。

数据集成、聚合和展现。有效的大规模分析往往需要从多个数据源收集异质数据。例如，获取病人（或人群）的 360 度健康视图时，可以通过整合和分析医疗健康记录获益，同时可以利用通过互联网获取的环境数据，随后甚至还能利用多种不同仪表的读数（例如，血糖仪、心率仪、加速度计以及其他仪表[3]）。数据转换和集成工具集帮助数据分析师解决数据结构和语义上的异质性问题。这种针对异质性的解决方案会促成集成后的数据在社区内部的统一性，因为它们符合标准化方案，满足分析要求。不过，由于完全集成的成本往往非常昂贵，而且分析需求变化迅速，所以最近的“按需付费”集成技术提供了一个诱人的“松弛”方法，它通过在线执行工作来支持临时的分析探索任务。

值得注意的是，互联网上大量可用的数据，若与支持生成派生数据的集成和分析工具结合，会导

致另一种类型的数据激增；此时不仅存在大数据量的问题，还要面对追溯此类派生数据的来源这种问题（随后我们会进行讨论）。

即使对于仅依赖一个数据集的较简分析而言，通常也会存在多种方法来保存相同的数据，每种方法均有利弊。作为证明，举例来说，生物信息学的数据库结构多种多样，但都包含了基本类似的实体（如基因）信息。如今，数据库设计是一种技能，在企业中由薪酬丰厚的专业人员谨慎地进行。我们必须让其他的专业人员（如专门领域科学家）具备创造有效数据存储的能力，或是通过设计工具帮助他们完成设计过程，或是完全放弃设计过程，开发在缺乏精细的数据库设计的情况下仍能有效使用数据的技术。

建模和分析。查询和挖掘大数据的方法与传统的小样本统计分析方法存在根本区别。大数据往往充满噪声，动态变化，异质，相关关联，且不可靠。然而，即使充满噪声的大数据也比微小的样本更有价值，因为从频繁模式和相关性分析得出的综合统计结果通常不仅能减弱个体数据的抖动和偏差，还能揭示更可靠的隐藏模式和知识。事实上，妥善利用统计方法后，研究人员能够使用近似分析来获取可靠的结果，而不会被数量难倒。

解释。决策者最终拿到分析结果后，必须解释这些结果。通常情况下，这需要审查做出的所有假设，并追溯分析。此外，还会存在很多可能的错误源：计算机系统可能存在错误，模型绝大多数时候都包含假设，结果可能基于错误的数据。因为上述原因，负责任的用户不会把他的职责转交给计算机系统。与此相反，她会想办法去理解和验证计算机生成的结果。计算机系统必须让上述工作变得容易。对于大数据而言，由于其相当复杂，这项工作尤为困难。在记录的数据背后，往往会存在一些至关重要的假设。分析管道可能包含多个步骤，并内置了各种假设。最近，抵押贷款对金融系统的冲击表明决策者迫切需要在这方面的尽职调查，而不是按面值接受金融机构所宣称的偿付能力，决策者必须认真审视在多个分析阶段中做出的各种假设。总之，只提供结果很难够用。更准确的说，不仅需要赋予用户解释所获取的分析结果的能力，还必须能让用户采用不同的假设、参数或数据集重复分析过程，以更好的支撑人类的思考过程和满足社会环境的要求。

虽然大数据的潜在利益是真实又重要的，而且也取得了一些初步的成功，但为了全部实现这一潜力，必须解决很多尚存的技术挑战。

数据解释是数据处理过程最后的环节，往往形成对基础数据带有观点性的注释。常见的情况是，这类观点彼此之间可能互相冲突，或者为它们做支撑的底层数据不够充分。在上述情况下，社区需要参与解决冲突的“编辑”过程（维基百科社区便是该过程的一个典型例子）。我们需要一种新型的数据工作平台。在该平台中，社区参与者能够对基础数据添加解释性的元数据注释，解决他们的分歧，并清理数据集；与此同时存在的部分干净或一致的数据仍然有待人工检查。

大数据分析中的挑战

在描述大数据分析管道中的多个阶段后，我们现在转向一些因大数据的特性而产生的共同挑战。在上述的很多阶段，有时是全部阶段中，人们都会面临这些挑战。图 2 的下部分用六个框说明了这些挑战。

异质性。当人们消费信息时，他们可以毫无困难地容忍大量的异质性。事实上，自然语言的细微差异和丰富性提供了颇具价值的研究深度。不过，机器分析算法却期望得到同质数据，因为它们理解细微差异的能力很差。因此，在数据分

案例研究

从 2010 年秋季开始，作为与洛杉矶大都会交通管理局（LA-Metro）的合同的一部分，南加州大学（USC）综合多媒体系统中心（IMSC）的研究人员获得了访问源于洛杉矶县路网的高分辨率时空交通数据。该数据以每分钟 46 兆字节的速度产生，迄今为止已经收集了超过 15 太字节的数据。IMSC 的研究人员开发了一套名为 *TransDec*（用于交通决策）的端到端系统来采集、存储、分析和可视化这些数据集（见附图）。在此，我们探讨一下与图 2 中描绘的大数据流程对应的多个 TransDec 组件。

采集：现有系统实时采集下列数据集：

- *交通用线圈探测器*：高速公路和交通干线上布置的约 8,900 个传感器收集交通参数，比如承载量、交通量和以及按读数 / 传感器 / 分钟的速率衡量的速度。
- *公共汽车和轨道交通*：包含的信息覆盖了在洛杉矶县运营的约 2,036 辆公共汽车和在 145 条不同的路线上行驶的 35 列火车。传感器数据包括每隔两分钟记录的每辆公共汽车的地理空间位置，相对于当前位置的下一个车站的信息以及相对于预先确定的时间表的延误信息。
- *匝道控制灯和 CMS（可变情报板）*：1851 组匝道控制灯根据当前的交通情况调节进入高速公路的车流量；160 个可变情报板（CMS）向旅行者说明了路况信息，比如延误、事故或道路施工区域。每组匝道控制灯和每个 CMS 传感器的更新速率是 75 秒。
- *事件*：从三个不同的机构中获取的有关特殊事件（比如撞车、交通危险等事件）的详细的自由文本格式信息（例如，伤亡人数，救护车到达时间）。

清理：数据清理算法使用微软的复杂事件处理（StreamInsight）删除多余的 XML 头，探测和删除多余的传感器读数等数据，让 46MB/ 分钟的输入数据缩减到 25MB/ 分钟。然后，结果被作为简单的表转储到微软 Azure 云平台上。

聚合 / 展现：使用 Oracle 11g 中的表集合聚合了数据，并建立了索引（利用 R- 树和 B- 树在空间和时间上建立了索引）。例如，聚合了数据以创建概要，以便支持预定的空间和时间查询集（例如，I-110 北部路段平均每小时的速度）。

分析：应用了数种机器学习技术来生成不同时间下（一天内不同时段（例如，高峰期），一周内的不同日期（例如，周末）以及不同的季节）洛杉矶县各路段的准确的交通模式 / 模型。使用了历史事故数据来为新的事故分类，以预测清道时间和事故引发的交通滞留的长度。

解释：在复杂的系统中，很多东西都可能出错，导致出现虚假的结果。例如，系统各（独立的）组件的失效可能无人关注，导致数据丢失。与此类似，有时候某个组织改变了数据格式后，却没有通知下游的组织，导致解析错误。为了解决此类问题，业界开发了多个监控用脚本，并配有相应的机制来获取用户的确认和纠正。

TransDec.

析的第一步（或之前），人们必须细心地把数据结构化。

相关的挑战之一是自动生成正确的元数据来描述记录的数据。例如，在科学试验中，为了正确地解释结果，可能需要与特定的试验条件和试验程序有关的大量细节。元数据获取系统能够尽可能地减少人们构建元数据的负担。在数据产生时就记录与数据有关的信息用处不大，除非这些信息能在数据分析管道中得到解释和传递。人们把它称为数据溯源。例如，某个步骤的处理错误可能会让后续分析毫无用处；有了合适的溯源机制后，我们可以轻松地定位依赖于该步骤的所有后续处理。因此，我们需要数据系统，以便在数据分析管道中传递数据源及其元数据。

不一致和不完整。大数据越来越多地包含来自多源的，不同可靠性的数据。不确定性、错误和遗失值很常见，必须对它们加以管理。从好的一面来看，往往可以利用大数据的数量和冗余信息来弥补遗失的数据，交叉验证冲突的情况，验证可信的关系，揭示内在的聚集，发现隐藏的关系和模型。

众包中也出现了类似的问题。虽然群体中的其他成员会发现并纠正大多数的此类错误，但是我们需

要各种技术来便利这一过程。作为人类，我们阅读某种产品的评论，其中有些让人动心，也有一些是负面的，然后我们会形成总结性的评价，并根据它来决定是否购买该产品。我们需要计算机具备完成上述任务的能力。在名为参与式感知的特殊众包中，不确定性和错误的问题更为明显。在这种众包中，携带手机的每个人都能作为一个即时收集多种类型数据的多模式传感器（比如照片、视频、音频、位置、时间、速度、方向和加速度）。此处存在的另一种挑战是数据收集设备固有的不确定性。事实上，收集的数据可能存在时空相关性，我们可以利用这点来更好地评估它们的正确性。如果众包数据通过雇佣的方式获取，比如 Mechanical Turks 的情况，那么工人的不同动机也会引发另一种错误模型。

即使在应用了纠错技术之后，数据中的某些不完整性和某些错误可能仍然存在。这种不完整性等错误必须在数据分析中加以管理。正确地实现这一任务也是一种挑战。处理和查询带概率和冲突数据的最新研究为这一方面的发展进步提供了有效的手段。

规模。当然，提到大数据时，所有人想到的第一个因素就是它的规模。在过去几十年里，管理迅速增长的大量数据一直是一个颇具挑战性的问题。过去，由于处理器遵守摩尔定律而变得越来越快， 这种挑战得以缓解。不过，如今却正在发生根本性转变：数据量比 CPU 速度和其他计算资源的增长速度要更快。

由于功率限制，时钟速度的增长已经基本停滞，现在的处理器开始构建数量越来越多的核。总之，人们必须处理单一节点内部的并行。不幸的是，过去应用于跨节点数据处理的并行数据处理技术并不能直接用于节点内部的并行处理，因为两者的架构看起来截然不同。例如，在单一的节点中，各核共享了更多的硬件资源，比如处理器高速缓存和处理器内存通道。

另一个正在发生的巨大转变是云计算；云计算现在把多种毫不相干、性能目标各异的工作负载聚合在一起，放入非常大的集群。在昂贵和大型的集群中实现上述级别的资源共享不仅需要着重利用以前的网格和集群计算技术，还需要新的手段来确定如何运行和执行数据处理作业，以便我们能够用划算的方式实现每个工作负载的目标，并确定如何处理系统故障（当我们操作的系统变得越来越大时，系统故障发生得越来越频繁）。

这带来了跨多个用户程序进行全局优化的需求，其中甚至也包括那些执行复杂机器学习任务的程序。对于依赖用户驱动的程序优化，通过虚拟化取得的集群使用率可能会相当低，因为用户不知道其他用户程序的情况。系统驱动的整体性优化需要程序足够透明。就像在关系数据库系统中，设计声明式查询语言时就考虑了这点。事实上，如果用户想要生成和构建用于大数据处理的复杂分析性管道，他们必须规定合适的高层级原语来明确说明他们的需求。

深入开发用于大数据分析的声明式方法除了技术上的原因外，还存在相当强的业务需要。组织一般会把大数据处理或是其中很多方面的工作外包出去。描述性的规范文档是必要的，用来制定可实施的服务协议，因为外包的目的就是精确的指定需执行任务，而不需要深入实现细节。

及时性。随着数据的增长，我们需要实时技术来汇总并筛选出需要存储的数据，因为在很多情况下，存储原始数据在经济上并不可行。这就产生了前文描述的采集率方面的挑战，以及后文我们将要讨论的及时性挑战。例如，如果怀疑出现了欺骗性质的信用卡交易，最理想的状况是能在该交易完成前予以标记 - 其有可能从根本上防止交易的发生。很明显，对用户购买历史进行实时全量分析不大可行。与此相反，我们需要提前计算出部分结果，以便在获取新数据时，可以使用少量的增量计算来实现快速判断。根本性的挑战是保证对高容量事件流进行大规模复杂查询时的交互响应时间。

另一个共同的模式是在极大的数据集中找出符合特定条件的元素。在数据分析的过程中，这种搜索可能会反复发生。通过扫描整个数据集来找出合适的元素明显不切实际。与此相反，人们提前创建了索引结构以便迅速找到满足条件的元素。例如，拿交通管理系统来说，其中包含了道路上行驶的数千辆车辆的信息以及局部的热点信息。该系统可能需要预测在用户选中的路线上存在的潜在拥堵点，并提出其他方案。实现此任务需要评估多个空间邻近查询，处理多个移动物体的运动轨迹。我们需要设计新的索引结构来支持多需求的查询。

隐私和数据所有权。数据的隐私是另一个不容忽视的因素，并且这一因素在大数据的背景下愈加明显。对于电子健康记录，存在严格的法规规定在何种的场景中可以披露何种数据。对于其他数据，特别是在美国，法规要宽松一点。然而，公众对于个人数据的不当使用，特别是将个人数据连接到多个数据源，抱有相当大的恐惧。有效的管理隐私不仅是技术问题，还是社会问题，必须从这两个方面同时着手才能实现大数据的远景。

例如，从基于位置的服务中收集的数据，需要用户向服务提供商分享他的 / 她的位置。其中明显有很多隐私方面的担忧，仅仅通过隐藏用户身份，而不隐藏她的位置无法平息这些担忧。攻击者或（可能恶意的）基于位置的服务器可以（后

续的）位置信息中推断出查询源的身份。例如，用户可能会留下一些能与其住所或办公室位置关联的“数据包碎屑踪迹”，从而可用于确定用户的身份。其他几种特殊的隐私信息，比如健康问题（例如癌症治疗中心的探访情况）或宗教偏好（例如，教堂的拜访情况），可以通过观察匿名用户在较长时间内的移动和使用模式揭示。总体来说，研究已经发现，人们的身份和他们的移动模式存在密切的相关性。[11] 但是，对于基于位置的服务而言，需要用户的位置用于成功的数据访问或数据收集，所以正确地完成该项工作颇具挑战。

另一个问题是，现在很多在线服务都需要我们分享隐私信息（想想 Facebook 应用），但是除了记录级的访问控制外，我们并不了解分享数据意味着什么，分享的数据如何被关联在一起，以及如何以符合直觉但却有效的方式允许用户对其分享数据进行细粒度的控制。另外，真实数据并不是静止不变的，而是会随时间变得越来越多；现在流行的技术中没有一种能在这种场景下揭示任何有用的内容。

隐私只是数据所有权的一个方面。总体来说，随着数据的价值逐渐得到认可，某个组织拥有的数据的价值会变成一种首要的战略考量。组织会关注如何在保持其独特数据优势的情况下利用这种数据；而且，如何在不失去控制的情况下分享或销售数据等问题会变得相当重要。在分发渠道从物理媒介（如 CD）的销售转向数字购买时，音乐界面临了数字版权管理（DRM）的问题，但上述问题与 DRM 不同；我们需要有效和灵活的*数据 DRM* 方法。

人类的视角：可视化和协作。 为了让大数据实现其全部潜力，我们不仅需要从系统角度考虑规模，还要从人类的角度考虑规模。我们必须保最终目标——人类——能够正确地“吸收”分析结果，而不会迷失在数据海洋中。例如，排序和推荐算法可结合用户的偏好为他 / 她识别最感兴趣的数据。然而，特别是当这些技术被用于科学发现和探索时，又必须采取特别的措施，从而避免把最终用户禁锢在仅由他们之前已经看过的类似数据构成的“过滤气泡”内[21]——很多有趣的发现源于探测和解释例外现象。

如果用户想要生成和构建用于大数据处理的复杂分析性管道，他们必须拥有合适的高层级原语来明确说明他们的需求。

尽管计算分析已经取得了长足的进度，但是仍然存在很多人类能轻松地发现，计算机算法却很难找出的模式。例如，验证码（CAPTCHA）精确地利用了这一事实来区分网络中的人类用户和计算机程序 . 理想情况下，大数据的分析不会全部都是计算性质的 —— 相反，设计时会明确在其环路中包含人类。视觉分析这一新分支正在尝试这样做，至少在管道的建模和分析阶段做成这样。在分析管道中的所有阶段，人类输入都有类似的对应值。

在今日的复杂世界中，通常需要多位来自不同领域的专家携手，才能真正了解正在发生的事情。大数据分析系统必须支持多位人类专家的输入以及共享探索结果。当把整个团队聚集在一间房间内的成本过高时，多位专家可能会被时空分隔。数据系统必须接受分布式的专家输入，并支持他们的协作。从技术层面讲，这需要我们考虑共享原始数据集之外的数据；我们还必须考虑如何支持这种共享算法和人工制品，比如实验结果（例如，在不断变化的数据集的某个指定快照上，采用特定的参数值后，应用一种算法所获得的实验结果）。

可视化效果可以迅速被创建，是向用户传递查询结果的最容易的理解方式和展示细节的手段。虽然早期的商业智能系统的用户满足于表格展现方式，但是今天的分析师需要用强大的可视化效果来包装和展现结果，因为这些可视化效果能帮

助他们解释结果，并支持用户协作。不仅如此，只要点击几次，用户应该就能向下钻取她观察到的每一种信息，并了解信息的源头。因为越来越多的人们拥有数据，也希望分析数据，所以这点变得尤为重要。

大数据合作实验室随着很多社区开始依赖基于云的数据管理以及大规模的共享数据储存库变成关键资源，使用共享数据开展协作的潜在价值得到了迅猛提升。我们如何允许用户创建一种结合自身数据与共享数据的数据分析，并（有选择性地）允许其他用户再次运行、再次改进和重新分配这些分析的人工制品，其范围可能包括从单一的查询到完整的建模和评分工作流等多项任务？这要求我们处理各种问题（比如，源头、访问控制或工作流），但处理时仍为不断增加的协作保留巨大的潜力，并提高协作工作的透明度等级（设想一下，您能用作者使用过的数据和代码重新运行论文中描述的所有分析，还能改进和发布结果！）。

众包是一种广为流行的，驾驭人类聪明才智用以解决问题的新方法。维基百科是一本在线百科全书，它可能是最有名的众包数据范例。采用社会化方法进行大数据分析具有广阔的前景。当我们把以各种各样以数据为中心的人工制品变成可共享的东西后，我们打开了一扇通往社会机制的门。这些社会机制包括人工制品的评级，领先榜（例如，透明地比较几种算法在相同数据集上的有效性）以及所形成的算法和专家的声誉。

结论

我们已经进入了大数据时代。经济中的很多领域现在正在转向以数据驱动的决策模型，其中核心业务依赖于对不断产生的大量繁杂数据进行分析。在这个数据驱动的世界里，存在某种可提升企业效率和提高生活质量的潜力。然而，在我们利用大数据的全部潜力之前，尚有若干必须应对的挑战。本文重点阐述了我们必须应对的关键技术挑战。同时本文承认还存在其他的挑战，比如经济、社会和政治方面的挑战，这些挑战在本文中并未论及，但也必须加以解决。另外，本文中探讨的所有技术挑战不一定会在所有的应用场景中出现。但是大多数挑战会出现。同理，某种挑战的解决方案可能无法在所有情况下通用。但是，同上面的情况一样，他们之间的相似之处不少，足以支持交叉学习。因此，本文中描述的多种挑战为计算机科学的跨领域研究提供了丰富的素材。我们还收集了一些建议供进一步阅读，请访问 http://db.cs.pitt.edu/bigdata/resources。我们依据广度和重要性选择了几十篇论文，而没有编制全面的参考书目。因为那样做的话，可能会包含几千篇论文。

鸣谢

本文根据由众多研究人员编著的白皮书[5]写就。我们在此向他们表示谢意。感谢 Divyakant Agrawal, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwar Dayal, Michael Franklin, Laura Haas, Alon Halevy, Sam Madden, Kenneth Ross, Dan Suciu, Shiv Vaithyanathan, 和 Jennifer Widom。

H.V.J. 的部分资金来源于美国国家科学基金会的资助（IIS1017296, IIS1017149 和 IIS1250880）。A.L. 的部分资金来源于美国国家科学基金会的资助（NSF IIS-0746696, NSFOIA-1028162 和 NSF CBET-1250171）。Y.P. 的部分资金来源于美国国家科学基金会的资助（IIS-1117527, SHB-1237174, DC-0910820）以及 Informatica 研究奖。J.M.P. 的部分资金来源于美国国家科学基金会的资助（III-0963993, IIS-1250886, IIS-1110948, CNS-1218432）以及谷歌、江森自控、微软、赛门铁克和 Oracle 的捐赠。C.S. 的部分资金来源于美国国家科学基金会的资助（IIS-1115153），与洛杉矶大都会交通管理局（LA Metro）的合同以及 Microsoft 和 Oracle 的无限制现金捐赠。

本文阐述的任何观点、发现、结论或建议仅属于其作者所有。

参考资料

1. Computing Community Consortium.Advancing Discovery in Science and Engineering.Spring 2011.
2. Computing Community Consortium.Advancing Personalized Education.Spring 2011.
3. Computing Community Consortium.Smart Health and Wellbeing.Spring 2011.
4. Computing Community Consortium.A Sustainable Future.Summer 2011.
5. Computer Research Association.Challenges and Opportunities with Big Data.Community white paper available at http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf
6. Dobbie, W. and Fryer, Jr.R.G.Getting Beneath the Veil of Effective Schools:Evidence from New York City.NBER Working Paper No.17632.Issued Dec. 2011.
7. *Economist*.Drowning in numbers:Digital data will flood the planet—and help us understand it better. (Nov 18, 2011); http://www.economist.com/blogs/dailychart/2011/11/big-data-0
8. Flood, M., Jagadish, H.V., Kyle, A., Olken, F. and Raschid, L. Using data for systemic financial risk management.In *Proc.5th Biennial Conf.Innovative Data Systems Research* (Jan. 2011).
9. *Forbes*.Data-driven:Improving business and society through data.(Feb. 10, 2012); http://www.forbes.com/special-report/data-driven.html
10. Gartner Group.Pattern-Based Strategy:Getting Value from Big Data.(July 2011 press release); http://www.gartner.com/it/page.jsp?id=1731916
11. González, M.C., Hidalgo, C.A. and Barabási, A-L. Understanding individual human mobility patterns. *Nature 453*, (June 5, 2008), 779–782.
12. Hey, T., Tansley, S. and Tolle, K., eds.*The Fourth Paradigm:Data-Intensive Scientific Discovery*. Microsoft Research, 2009.
13. Kahn, S.D. On the future of genomic data.*Science 331*, 6018 (Feb. 11, 2011), 728–729.
14. Lazar, D. et al.Computational social science.*Science 323*, 5915 (Feb. 6, 2009), 721–723.
15. Lohr, A. The age of Big Data.*New York Times* (Feb. 11, 2012); http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html
16. Lohr, S. How Big Data became so big.*New York Times* (Aug. 11, 2012); http://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html
17. Manyika, J. et al.Big Data:The next frontier for innovation, competition, and productivity.McKinsey Global Institute.May 2011.
18. National Science and Technology Council.*Materials Genome Initiative for Global Competitiveness*.June 2011.
19. Noguchi, Y. Following the Breadcrumbs to Big Data Gold.National Public Radio (Nov. 29, 2011); http://www.npr.org/2011/11/29/142521910/the-digital-breadcrumbs-that-lead-to-big-data
20. Noguchi, Y. The Search for Analysts to Make Sense of Big Data.National Public Radio, (Nov. 30, 2011); http://www.npr.org/2011/11/30/142893065/the-search-for-analysts-to-make-sense-of-big-data
21. Pariser, E. *The Filter Bubble:What the Internet Is Hiding From You*. Penguin Press, May 2011.
22. PCAST Report.*Designing a Digital Future:Federally Funded Research and Development in Networking and Information Technology* (Dec. 2010); http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf
23. SDSS-III:Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems (Jan. 2008); http://www.sdss3.org/collaboration/description.pdf/

H.V. Jagadish（jag@umich.edu）是安阿伯市密歇根大学电气工程和计算机科学专业 Bernard A Galler 大学教授。

Johannes Gehrke（johannes@cs.cornell.edu）是纽约州伊萨卡市康乃尔大学计算机科学系的 Tisch 大学教授。

Alexandros Labrinidis（labrinid@cs.pitt.edu）是匹兹堡大学计算机科学系副教授及高级数据管理技术实验室的联席主任。

Yannis Papakonstantinou（yannis@cs.ucsd.edu）是圣地亚哥市加利福尼亚大学计算机科学和工程教授。

Jignesh M. Patel（jignesh@cs.wisc.edu）是麦迪逊市威斯康星大学计算机科学教授。

Raghu Ramakrishnan（raghu@microsoft.com）是位于华盛顿州雷德蒙市的微软公司技术院士和信息服务技术总监。

Cyrus Shahabi（shahabi@usc.edu）是南加州大学计算机科学和电气工程教授兼信息实验室主任，以及美国国家科学基金会综合多媒体系统中心主任。

译文责任编辑：唐杰