实现数据科学的潜能 *

作者：弗朗辛·伯曼 (Francine Berman)
罗伯·鲁坦巴尔 (Rob Rutenbar)
亨里克·克里斯滕森 (Henrik Christensen) 等

关键词：数据科学研究　数据科学教育

译者：刘宇航

本文关键的深刻见解：

• 数据科学提供了新的洞见 (insights)[1]，有助于将信息转化为可以推动科学和工业发展的知识，有助于将之前分离的学科、社区和用户连接起来，为当前和未来的挑战提供更丰富更深刻的洞见。

• 数据科学涵盖了广泛的领域，例如以数据为中心的算法创新和机器学习，数据挖掘和利用数据进行科学发现，收集、组织、管理和保存数据，隐私挑战和与数据相关的政策，培养和训练数据专业人员的教育学科。

• 数据系统的商业实践和学术研究之间的差距越来越大，需要予以解决。

处理和理解数据的能力对于科学发现和创新越来越关键。因此，我们得以看到数据科学这个新领域的出现，它主要关注那些能够使我们从各种形式的数据中提取知识及洞见并把知识转化为行动的流程和系统。它在实践中已演变为一个跨学科的领域，整合了来自统计、数据挖掘和预测分析等数据分析领域的方法，而且包含了可扩展计算以及数据管理方面的进步。但作为一门学科，数据科学尚处于初级阶段。

以充分开发潜能的方式发展数据科学的挑战引发了研究和教育界需要面对的重要问题：应当如何发展数据科学领域，才能使它支持数据在各应用领域发挥日益重要的角色？我们应当如何培养可以使数据充分发挥最大优势的专业人才？我们应当教给他们什么？政府机构可以做什么以使数据科学的潜力最大化，进而驱动科学探索并且满足拥有数据科学技能的人才当下和未来的需求？作为由美国国家科学基金会 (NSF) 计算机和信息科学与工程局 (CISE) 召集的数据科学探索工作组 (https://www.nsf.gov/dir/index.jsp?org=CISE)，我们就这些问题提出一些看法，特别关注研发机构在支持和培养数据科学的发展与影响过程中所面临的机遇与挑战。关于本文所基于的完整报告，请参阅伯曼 (Berman) 等人的工作 [2]。

---

*本文译自 *Communications of the ACM*，“Realizing the Potential of Data Science”，2018, 61(4):67~72 一文。

[1] insights 是数据科学中的一个重要词汇，但在中文中目前还缺乏准确的对应，现有词典（比如牛津高阶英汉词典）译为“洞悉”或“了解”，可能是不恰当的。它有两种含义，一种含义是洞察力，即深刻理解事物的能力，为不可数名词，另一种含义是深刻的理解和见解，为可数名词。在数据科学中，我们认为译为“深刻见解”或“洞见”可能比较恰当。——译者注

数据科学的重要性与其中蕴含的机遇是显而易见的（参见 http://cra.org/data-science/）。如果美国国家科学基金会通过与其他机构、基金会或行业合作，在未来十年能够帮助促进数据科学学科以及数据科学家的进步与发展，我们的研究社区就能够更好地发挥数据科学的潜能，推动新的发现与创新，有助于将信息时代转变为知识时代。希望本文能够成为学术界、工业界、ACM 和相关的 ACM 特别兴趣组（如 SIGKDD 和 SIGHPC）内部对话的基础。

数据生命周期

数据从来就不存在于真空中。与生物有机体一样，数据具有生命周期，从诞生开始经历活跃的一生直到“永恒”或某种形式的消亡。它也像活着的智能有机体一样，生活在一个拥有物质支持、社交背景和生存意义的环境中。数据生命周期对于理解充分利用数据所带来的机遇和挑战至关重要。从图 1 可以了解数据生命周期的基本组成部分。

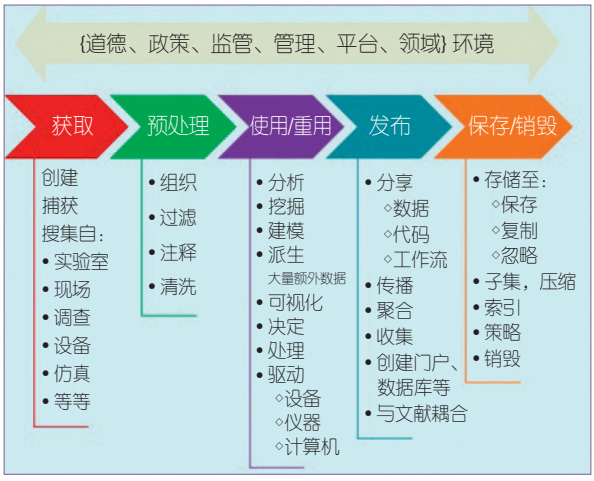

图1 数据生命周期及其所处的数据生态系统
（节选自《实现数据科学潜力科学报告》[2]）

大型强子对撞机 (the Large Hadron Collider, LHC) 实验输出的数据可以作为理解数据生命周期的一个例子。大型强子对撞机对物理学界极为重要，得到世界各国研究人员的支持。LHC 实验通过碰撞粒子以测试粒子物理和高能物理的各种理论的预测是否正确。2012 年 LHC 实验的数据为希格斯玻色子 (Higgs Boson)[2] 提供了有力证据，证实了物理学标准模型的正确性。这项科学发现被 *Science* 杂志评为 2012 年“年度进展 (Breakthrough of the Year)”[3]，并获得 2013 年诺贝尔物理学奖。

LHC 数据的生命周期很有趣。在“诞生”时，LHC 数据表示粒子在法国与瑞士两国边界的 17 英里隧道中的一台仪器内发生的碰撞的结果。绝大部分数据因在技术上“没有价值”而被弃掷一边，但仍有数目可观的“有意思的”数据有待分析和保存。据估计，到 2040 年，LHC 产生的“有意思的”数据将达到 10EB~100EB（1EB=2^{60} 字节）。留存的 LHC 数据会被标注后保留，保存在十几个不同的地方。数据会被公布并提供给各社区，供 100 多个研究机构分析和使用。对 LHC 数据整个生命周期的管理、使用和传播的重视，在推动实验带来的科学突破方面发挥了关键作用。

除了发展数据管理、传播和使用协议，LHC 数据生态系统还提供了可持续支持数据及其基础设施的经济模型。更大的生态系统与社区协议（关于数据如何组织）、政治经济支持相结合，发挥 LHC 数据的潜能来革新我们对物理学的认知，并使科学家能够充分利用对 LHC 物理仪器和设施的巨大投入。

图 1 中概述的数据生命周期和 LHC 示例展示了一组无缝衔接的数据操作和数据转换。但是在现在许多科研团体与学科中，这些步骤是相互隔离的。领域科学家专注于产生和使用数据，计算机科学家通常关注平台和性能问题，包括数据的挖掘、组织、建模和可视化，以及通过机器学习或其他方法从数据中提取出意义的机制。数据采集和仪器控制的物理过程，往往是工程领域的重点，即数据或者作为“脏信号”或者作为其他设备的控制输入。统计学家可能会关注风险和推理的数学模型。信息科学家和图书馆科学家可能会重点关注数据的管理、保存

[2] 希格斯玻色子是粒子物理学标准模型预言的一种自旋为零的玻色子。——编者注

以及流水线的后端，即在出版、保存、清洗领域中的数据采集、裁定、处理。

首先，建立计算机科学、信息科学、具体领域、物理科学与工程领域间的联系来填补学科间的间隙是一个重要机遇。其次，填补机器学习、数据分析以及相关学科（如统计学和运筹学）之间的间隙也是一个重要机遇。在此我们聚焦于某些机遇。

美国的数据科学研究

图 1 所示的数据生命周期几乎每个阶段都有深入研究的机遇。此外，如前所述，美国数据科学议程的首要机遇是弥合数据生命周期中的间隙，在计算机科学、信息科学、统计学、领域和物理科学与工程社区之间建立更强有力的联系。也就是说，依循旧制的研究议程也许能加强数据生命周期中单独步骤各自背后的个别技术，但不太可能培育出较为广泛的技术突破或跨越学科孤岛的范式转型。数据的一个重要本质属性是它能够连接之前分离的学科、社区、用户，从而为当前和未来的挑战提供更丰富和更深入的洞见。

鼓励用一种更宽广更统一的视角将数据视为整合科学、工程与应用领域中的研究契机是非常重要的。其中第一个机遇是对整个数据生命周期及其所处的环境投入精力研究——即将此机遇作为一个核心成果本身，而不是作为另一个理想成果的副产物或中间步骤。数据科学作为计算机科学核心组成部分在深入发展的同时，也应该在广度上有所发展，以满足计算机科学以外领域的需求。我们的社区有一个独特的机会来推动数据科学的发展，即可以将数据驱动战略应用于单独领域研究和跨领域研究。

第二个机遇涉及由大数据首次赋能的被称为“嵌入智能”的场景。最近一系列基础人工智能和“深度学习”技术方面取得的突破[1]，使创建能“智能地反应 (act intelligently)”的成熟软件工具成为可能。关键的创新在于数学化的模式识别技术，该技术以几百万能够正确反应的训练示例作为输入来创建各种软件系统（不久可能也涵盖硬件系统），能够更好地识别图像、解码人类语音、发现法律和商业文档中的关键模式，等等。作为工程人造物 (engineered artifacts)，这些人工智能系统体现为复杂的数学公式，这些公式使用数量惊人的数值参数（比如今天一个成熟的图像分类系统需 1000 万个参数）根据特定目的来定制或者“训练”。

这些经过训练、面向决策的模型正在成为为复杂问题提供解决方案的一系列新型软件的核心组成部分，形成了跨学科的挑战[6]。比如，当一个组件可能仅有 70% 的准确率时，它的“正确”意味着什么？用于训练和更新这些模型的数据生命周期应该是什么？针对这些数据进行训练的嵌入式智能体的行为造成负面影响时（例如，当汽车自动驾驶系统崩溃时，或者基于自动推理的客户账户被错误中止时）所涉及的政策细节（以及责任分摊）应如何设定？作为一门学科，软件工程面临着一系列挑战，如性能不精确，针对这些系统的大量数据组件（10^9 字节 ~10^{12} 字节规模的训练数据）需要进行版本控制和测试。现有的模型检测 / 验证的概念是远远不够的。政策、管理和监护问题在很大程度上没有得到解决，更没有答案。

值得注意的是，预测模型并不是机器学习所专有的。例如，统计模型已被用于流行病学，物理模型在天气预报和核模拟中也很常见。数据科学的“训练”在解决方案的软件工程背景下可能实属新颖，因为被训练的模型可能缺乏与统计功效和样本量计算相关的保证。

第三个机遇是弥补现有技术前沿中数据系统的商业实践和科研实践之间越来越大的差距。目前许多高水平的学术研究人员向拥有海量数据的企业（如脸书、谷歌和微软）“逆向迁移”。虽然这对近期的美国国民经济可能有好处，但对未来基于发现的开放式科学研究、教育与训练等学术部门来说则是非常令人担忧的。除了富有吸引力的研究经费之外，人才从学术研究界流向私营企业的另一个原因可能是基础设施支撑环境日益恶化，包括缺乏大数据集，缺乏足够的基础设施支持大规模数据科学研究。如果前沿研究的最佳基础设施环境始终在私营企业，

公共部门的创新机会就会弱化。政府支持建立具有战略高度与互信基础的公私合作伙伴关系，如建立大型基础设施促进学术研究的创新发展，最终也会以培养受过良好教育与训练的尖端人才的方式回馈给私营企业。

美国的数据科学教育与训练

美国的高等教育机构认识到，掌握数据科学是21世纪科研与劳动力所必需的一项关键技能。在高等教育中，数据科学课程有两类受众：数据科学的新专业人员以及将数据科学技能应用于其他领域的科学家和专业人士。高等教育中的数据科学课程往往侧重于这两者，就像计算机科学系的课程培养计算机科学系的学生一样，它同时也为来自其他学科的学生提供计算机技能方面的培训，以提高他们的计算机素养。

值得注意的是，目前还没有一个固定的模式关于由高校内的哪个系、学院或跨部门协作来负责数据科学的教育和训练。数据科学项目正在计算机科学、信息科学、统计学和管理学的部门和学院中进行。许多成功的案例，尤其在本科阶段，都是跨学科机构经常赞助的全校范围内的联盟，而非某一个特别的系或学院。因此，尽管此前已有很多有趣的尝试（参见附录**“数据科学教育的现状：百花齐放”**了解一些程序化的配置），但对于数据科学应该“落脚”的机构仍没有共识。请注意，当一所大学选择在现有的系或学院中设置“数据科学”时，它默认会采用现有组织的标准和文化。相比之下，当一所大学将“数据科学”以跨学科的职能引入时，它将面临引进新领域所带来的异质性，同时可能需要处理与跨机构实体相关的额外管理开销。我们接下来关注数据科学教育和训练的趋势。

数据科学的教育课程尚未“标准化”，目前出现了许多有趣的课程设置。一般而言，我们期望数据科学家能够使用统计技术分析大数据集，因此统计学和建模通常是所需课程的一部分。此外，一个综合的数据科学课程不仅包括机器学习和统计学，还包括程序设计、数据管理和伦理学等课程。数据科学家必须能够在非结构化数据中发现意义，因此程序设计、数据挖掘和机器学习等通常是核心课程的一部分。数据科学家还必须能够有效地交流他们的发现，所以可以提供关于数据可视化的课程，至少作为选修课。鉴于滥用数据和从数据中得出错误结论带来的挑战，伦理学也正在成为该领域核心课程的一部分。

研究设计、数据库、算法、并行计算和云计算等其他课程作为核心课程或选修课，都反映了学校对学生应当掌握技能的期许。许多课程还需要结合项目，让学生通过在特定领域的团队中解决实际问题来获取经验。数据科学课程正在成为高质量在线教学项目的主要内容。

强大的数据科学课程要求教师具备相应的专业知识并熟悉该领域。在数据科学和相关领域具有专业技能的人士从学术界转到工业界，对教育机构开展一系列教学计划带来了挑战。这也对数据科学发展为一门正式学科构成了潜在的挑战。

为了扭转这一趋势，摩尔和斯隆基金会(the Moore and Sloan Foundations)在2013年创建了一个总价值3800万美元的项目，即“摩尔-斯隆数据科学环境”，为创建“数据科学环境”的初期项目提供资金[7]，解决学术职业生涯、教育、训练、工具、软件、可重复性和开放科学、物理和智能空间以及数据科学研究面临的挑战。该基金是变革性的，为数据科学项目对现在与未来的工作提供了有利的关键“实例”。

从目前多样化的课程和计划来看，数据科学正在经历一个重要的健康的试验阶段。重要的是我们不应过快地将数据科学“标准化”，而应继续探索课程、领域、项目、师资和合作关系间的配置方式，从而为最优地培养新一代数据科学家获得关键经验。

除了将数据科学演化为一门学科的“数据科学”课程和专业之外，数据科学的技能训练对其他学科和专业也越来越重要，因为它们越来越多地被数据赋能。有效的训练使专业人员和领域科学家能够有效地利用数据，并在更广泛的数据驱动环境中操作数据，了解数据可以告诉我们什么和无法告诉我们

什么，获得有关如何处理数据的技术知识，意识到数据之间的相关性并不一定意味着因果关系，并养成对数据使用负责的态度，遵守道德准则。

在处理数据的细节方面进行更具体的培训对于数据驱动的职业是至关重要的。对于将在研究中使用数据驱动模拟和模型的学生来说，程序设计和软件工程方面的训练是有用的。应该向计算科学研究人员传授版本控制以及管理工作的细节，包括与数据库和软件存储库协作。数字学术能力和可重复性的最佳实践的训练也应纳入研究方法学课程。使用（和滥用）数据的道德标准应纳入所有训练计划，以促进有效且负责任地使用数据。从大学课程到在线课程，再到科学团体和社区可开发的专业课程，这些课程可以在各种场所获得。

数据科学研究与教育的基础设施

数据科学的研究和教育方面的任何创新议程都将取决于数据基础设施和有用数据集的启用。数据科学的研究需要访问海量数据集去阐明和验证结果。数据集也必须对可重复研究可用并由可靠的基础设施托管。

缺乏这样的基础设施和数据集，数据科学研究将会受到阻碍。数据科学中的教育和训练是最真实的，在这种环境中，学生可以处理代表他们将在专业领域中看到的数据集和环境数据；也就是说，数据既是规模化的，又嵌入在管理的基础设施中，使其成为分析、建模和挖掘过程中的有用工具。

在最理想的情况下，数据基础设施应支持对数据的访问以进行研究和教育，就像访问任何其他关键设备一样：它一定“永远在线”，必须足够健壮以支持广泛使用，而且质量必须良好。在数据领域，这归结为负责任的管理，这意味着必须有执行者、计划以及“社会学”意义和技术意义上的基础设施来确保以下内容：

数据被适当地跟踪、监控和识别。谁在创建、组织和使用数据？数据能一直被识别吗？是否有足够的隐私和安全控制？

数据被妥善保管。谁负责以什么格式、多长期限保存数据？谁负责为数据管理提供资金？数据如何存储并迁移到下一代媒体上？

数据可被发现且有用。数据以何种形式提供给哪一方？需要哪些服务才能充分利用它？需要哪些元数据和其他信息来提高可重复性？

数据管理符合政策和实践。管理权是否符合社区标准和有关报告、知识产权及其他问题的相关政策？确定适当使用的权利、许可和其他属性是否清晰？哪些数据和元数据将被保存？谁拥有数据及其衍生物？谁有权访问全部或部分数据及其元数据？

由于数据将成为许多学科研究和发展的核心，在合理的时间范围内以可用的形式访问数据，成为任何有效研究和教育议程的切入点。政府研发机构（如美国国家科学基金会）要确保不会因为缺乏足够的数据基础设施而阻碍创新研究与教育计划。

发展和维持基础设施，确保研究数据向公众开放并可重复使用及可重复生成，这都需要稳定的经济模型。美国联邦政府尽管对工具、技术、构建模块和数据共享方法的开发提供了大量支持，但很少直面数据管理资源短缺的挑战，也没有为图书馆、域名存储库和其他管理环境提供帮助，使其能够自我维系并满足公众访问的需求。

虽然美国联邦政府不能承担管理科研数据及基础设施的全部责任，但也不应该回避向组织机构提供种子资金或过渡资金，以便为美国的社区提供可持续管理的支持。我们鼓励政府内部和外部社区通过战略计划、指导方针以及跨机构公私合作伙伴关系，支持数据驱动研究和数据科学教育的可持续数据管理模型的发展和试点工作。像美国国家科学基金会这样以科学为中心的政府机构应该与专注于此类问题的同行机构（如美国国立卫生研究院）进行协作，以利用投资提供一定范围和规模的经济支持。

实现潜能

本文关于研究、教育、基础设施的探讨聚焦于

增加数据科学家和有数据科学素养的专业人员的数量，进而应对所有部门的数据驱动工作在当前和近期面临的挑战，以及满足将数据科学发展成为一门学科的需要，该学科能够应对未来数据驱动场景中的各种挑战。

数据无处不在，为各式各样的工作提供了日益重要的工具。随着系统变得“更智能”，并具备更多自主和决策能力，我们将越来越多地面临数据科学带来的技术挑战以及管理、道德、政策和隐私等问题引发的社会伦理挑战。解决这些问题至关重要，这将使得数据驱动的系统变得有用、有效和高效，而不是具有侵入性、限制性和破坏性。这种解决方案将在诸如物联网等高度数据驱动的环境中显得特别重要。此外，随着摩尔定律半导体扩展逐渐到达极限，计算平台将有重大改变[12]，未来将会有巨大的机遇来重新构建整个硬件 / 软件产业。

结论

我们的社区必须做好应对未知状况的准备，包括：鼓励那些为创新使用数据奠定基础的初步研究，维护以数据为中心的优质系统，建立有用的政策和保护机制，以及有效治理数据驱动的环境。借助计划性的资源和社区领导性的平台，联邦研发机构（如美国国家科学基金）在引导社区进行创新方面发挥着重要作用。需要更大的努力来拓展数据科学领域及其影响，推动数据科学的潜力发挥，以变革 21 世纪的科研、教育、商业和生活。 ■

致谢：

感谢美国国家科学基金会召集这个小组，并感谢作者们所在的机构和组织对本文的支持。

附录：**数据科学教育的现状：百花齐放**

为了支持数据科学领域的研究和人力资源的发展，我们必须确定应当如何教授数据科学，以及这门学科应该设置在什么部门。就像 20 世纪 60 年代计算机科学的出现使得现代大学创建了第一批致力于计算的组织单位和学位一样，数据科学的兴起正在推动一系列有趣的课程尝试。为了了解快速演化的形势，我们考察了以下五个单位：

加利福尼亚大学 加州大学伯克利分校数据科学教育项目[8]是其最近成立的数据科学部 (Division of Data Sciences) 的一部分，与伯克利的学院处于同一层级，并与之融合。入门课程为所有领域的学生提供了接触数据的基础，并为从事高级工作开辟了道路。基础课程包括核心的计算科学和统计学的概念，使学生能够处理各种领域的真实数据。高级课程包括一门名为《数据科学的 100 条原理与技术》的高级综合课。

密歇根大学 密歇根大学的数据科学[11]是电气工程系、计算机科学系与统计学系联合开设的一个新专业。数据科学专业是一个严谨的计划，聚焦于计算机科学、统计学和数学中与分析和处理大型数据集相关的方面。学生可从工程学院或文学、科学和艺术学院转入该专业学习。

哥伦比亚大学 哥伦比亚大学数据科学研究所的数据科学硕士点[4]为任何获得本科学位的学生提供专业的硕士学位，包括合适的量化的课程作业。它从四门基础课程开始（可以独立进行以获得数据科学证书），侧重于算法、概率 / 统计、机器学习和可视化。

伊利诺伊大学 伊利诺伊大学厄巴纳 - 香槟分校数据科学计算机科学学位[10]是大规模开放式在线课程平台上的在线专业硕士学位[5]。该学位致力于以综合全面的方式覆盖包括数学、计算、监护和管理的整个数据生命周期。

芝加哥大学 芝加哥大学计算分析与公共政策硕士学位[9]由计算机科学系和哈里斯公共政策学院联合建立。该学位侧重于政策与计算机科学的交叉。学生在两个领域都参加课程，让他们做好准备为公共部门的政策设计、实施和严格分析做出有意义的贡献。

作　者：

弗朗辛·伯曼 (Francine Berman)

美国伦斯勒理工学院杰出教授，科研数据联盟 (RDA) 主席。她曾担任 NSF CISE 咨询委员会数据科学工作组的联合主席。bermaf@rpi.edu

罗伯·鲁坦巴尔 (Rob Rutenbar)

美国匹兹堡大学教授。他曾担任 NSF CISE 咨询委员会数据科学工作组的联合主席。rutenbar@pitt.edu

亨里克·克里斯滕森 (Henrik Christensen)

美国加利福尼亚大学圣地亚哥分校计算机科学教授。hichristensen@ucsd.edu

其他作者：

Susan Davidson, Deborah Estrin, Michael Franklin, Brent Hailpern, Margaret Martonosi, Padma Raghavan, Victoria Stodden, Alex Szalay

译　者：

刘宇航

CCF 专业会员、CCCF 特邀译者。中国科学院计算技术研究所副研究员。主要研究方向为计算机体系结构、高性能计算、大数据、智能并发系统。

liuyuhang@ict.ac.cn

（本期译文责任编委：**姜波**）

参考文献

[1] Bengio Y, LeCun Y, Hinton G. Deep learning. *Nature* 2015,521 (7553): 436-444.

[2] Berman F, Rutenbar R, Christensen H, et al. Realizing the Potential of Data Science: Final Report from the National Science Foundation Computer and Information Science and Engineering Advisory Committee Data Science Working Group. National Science Foundation Computer and Information Science and Engineering Advisory Committee Report, Dec. 2016; https://www.nsf.gov/cise/ac-data-science-report/CISEACDataScienceReport1.19.17.pdf

[3] Cho A. The discovery of the Higgs Boson. *Science* 338, 6114 (Dec. 21, 2012), 1524-1525.

[4] Columbia University Data Science Institute. Master of Science in Data Science. http://datascience.columbia.edu/master-of-science-in-data-science.

[5] Coursera. Master of Computer Science in Data Science. https://www.coursera.org/universityprograms/masters-in-computer-data-science.

[6] Dhar V. When to trust robots with decisions, and when not to. Harvard Business Review (May 17, 2006). https://hbr.org/2016/05/when-to-trust-robots-withdecisions-and-when-not-to.

[7] Moore-Sloan Data Science Program. http://msdse.org/.

[8] University of California, Berkeley. Data Science Education Program. http://data.berkeley.edu/datascience-education-program.

[9] University of Chicago. Master of Science in Computational Analysis & Public Policy. https://capp.uchicago.edu/.

[10]University of Illinois, Urbana-Champaign, CS@ILLINOIS. Master of Computer Science in Data Science, Data Science Track. http://www.cs.uiuc.edu/academics/graduate/professional-mcs-program/mcsdata-science-track.

[11]University of Michigan. Undergraduate Program in Data Science. https://www.eecs.umich.edu/eecs/undergraduate/data-science/.

[12]Waldrop, M.M. The chips are down for Moore’s Law. Nature 530, 7589 (Feb. 11, 2016), 144-146.