大数据与软件工程

大数据概念最初起源于美国，是由思科、威睿、甲骨文、IBM等公司倡议发展起来的。大约从2009年始，“大数据”成为互联网信息技术行业的流行词汇。

大数据已成为继云计算等之后的全球性新热点，其关键词在全球范围内迅速升温，短短四年间，“大数据”关键词的搜索量指数双倍增长。由维基百科可知，“大数据是指无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合。”大数据，或者说海量数据，指的是在人类社会运用了现代信息技术之后，掌握了比以前数量大得多的各种数据。最早提出“大数据时代已经到来”的机构是全球知名咨询公司麦肯锡。它是一个不断演变的概念，当前的兴起，是因为从IT技术到数据积累，都已经发生重大变化。所谓大是相对的，随着信息技术的发展，在这20多年里，我们讨论的数据量从以K计算，发展到M、G、T、P等，每次都提高了三个数量级。

如今，大数据时代已经来临，它将在众多领域掀起变革的巨浪。但我们要冷静的看到，大数据的核心在于为客户挖掘数据中蕴藏的价值，而不是软硬件的堆砌。在政府公共服务、医疗服务、零售业、制造业、以及涉及个人位置服务等领域大数据将得到广泛应用，并产生巨大的社会价值和产业空间，预测2020年，大数据应用市场规模将达到近2600 亿美元。大数据是传统数据的延伸，是对传统数据在深度和广度上的补充。发展大数据是促进政务信息资源开发利用的必然要求，是提高政府决策科学化水平的必然要求，是提高城市管理精细化水平的必然要求，是促进现代服务业发展的必然要求。

2012年3月29日，美国奥巴马政府推出“大数据研究与开发计划”，提出“通过收集、处理庞大而复杂的数据信息，从中获得知识和洞见，提升能力，加快科学、工程领域的创新步伐，强化美国的国土安全，转变教育和学习模式”。2014年年初，英国商业、创新和技能部宣布，将注资6亿英镑发展8类高新技术，其中对大数据的投资即达1.89亿英镑。法国政府为促进大数据领域的发展，将以培养新兴企业、软件制造商、工程师、信息系统设计师等为目标，开展一系列的投资计划。法国政府在其发布的《数字化路线图》中表示，将大力支持“大数据”在内的战略性高新技术。大数据“可爱又可怕”。一方面，大数据给人们的生活带来了诸多方便，给企业提供了更多的商业机会；另一方面，大数据又增加了重要信息泄露的风险，家庭住址、密码、手机号码、支付密码等个人信息都存在被盗用的危险，有可能会给受害人造成较大的精神及财产损失。 数据的爆炸式增长出乎人们的想象，使得适应和应对数据增长成为整个社会关注的焦点，大数据的概念也在这一背景下诞生的。但是并不是所有的数据都可以称其为大数据，大数据的判断需要基于4个层面，即大数据的4V特性包括体量（Volume）、多样性（Variety）、价值密度（Value）、速度（Velocity），体量指非结构化数据的超大规模和增长，总数据量的80%～90%，比结构化数据增长快10倍到50倍，是传统数据仓库的10倍到50倍；多样性指大数据的异构和多样性，很多不同形式（文本、图像、视频、机器数据），模式或者模式不明显，不连贯的语法或句义；价值密度指大量的不相关信息，对未来趋势与模式的可预测分析，深度复杂分析（机器学习、人工智能Vs传统商务智能（咨询、报告等）；速度指实时分析而非批量式分析，数据输入、处理与丢弃，立竿见影而非事后见效。

自1968年“软件危机”以来，大规模软件工程常被类比为困住恐龙的史前焦油坑，其复杂性一直难以控制。软件开发是知识密集型活动。个体差异和社会化生产环境是影响效率的重要因素，但这些因素又都难以度量和理解，导致软件开发过程难以有效控制。度量难的首要原因是可用数据不多，一方面在实际项目开发中度量和评估并非首要任务，从而很难得到开发人员的重视，也就很难进行有针对性的数据收集；另一方面，即使开展了数据收集活动，被观察者也可能因为察觉到观察行为的存在而改变其真实的开发行为，这就是著名的霍桑(Hawthorne)现象[5]。因为开发活动的需要，近年来软件项目广泛使用版本控制系统、缺陷追踪系统、邮件列表和论坛等工具，大型软件工程的数字轨迹通常被保存起来。人们的每一次代码提交、每一个缺陷报告、对缺陷的每一次评论、每一个邮件及回复等，都被完整地保存在软件仓库中，记录了软件开发过程和代码的演变，以及开发者个体和交互的行为。以可自由访问的开源项目为例，仅 Gnome一个开源项目就有超过 70 万个缺陷数据，阿帕奇 (Apache) 开源社区拥有超过 600 万封邮件，Mozilla 社区有超过 2 亿条的代码提交日志。许多成功开源项目都拥有众多分布于全球的用户和贡献者，其软件工程数据时刻都在增长，例如 Gnome 每天有上百个新的缺陷提出，阿帕奇平均每天有 1000多封新邮件产生，Mozilla 每天有 2 万多条新的提交。而 Github.com拥有超过1000万个的项目，SourceForge.net和GoogleCode则分别有超过0 万个的项目。互联网上所有开源项目的版本控制系统数据总量估计在 70TB 的级别。这些数据分散在互联网上众多地方，每分每秒都在不断地产生和消亡，因此采集、存储和整理是首要面对的问题。因为软件工程面临大数据

的挑战，一种新的研究途径随之出现，如同人类考古学，人类祖先有意或无意留下的痕迹为考古学家提供了研究古代文化历史的可能，项目过程数据留下了软件的开发历史轨迹，使得人们能够从中挖掘有价值的内容，我们称之为软件的数字考古。经验软件工程挖掘软件库以及商业智能等都是学术界和产业界的一些尝试，采用数据驱动方法来理解软件的生产过程，并利用这些数据度量程序员效率、计算成本和预测软件质量等，为许多实践操作和业务决策提供了很大的帮助。值得一提的是，这些数据是因为程序员和项目开发自身需要而自然积累下来的，并非为了度量或评估而填写或保存的，这使得霍桑现象不再是问题。然而，如果只针对单个或若干个项目进行研究，研究结果的适用性相当有限。例如，文献指出，大部分发表的研究成果都不可复制，在新的数据集上研究结果经常不可再现。但是众多的软件仓库贡献了数量庞大的项目样本，对大样本将带来惊人洞察力的想象激励着人们用全新的观念审视大数据时代的软件工程：构建软件工程大数据池，更准确地计算成本，更精细地认识和改进过程，探讨以前无法度量的程序员的个体差异性和社会化生产的可知可控等难题。并且，在不同的时代条件下，软件开发模式在不断演化，我们从来不曾有机会纵横捭阖地比较各种项目的微过程，从广阔的时间维度上理解软件及软件开发的演化，获得对软件开发理论和方法的崭新认识。软件工程大数据为此创造了条件。综上所述，软件工程大数据是由众多软件开发过程中的工具自然产生和记录的软件演化和参与者活动的日志，它们分散在互联网的软件仓库、软件公司以及个体的各种环境中，而且软件项目样本多、日志数据种类和格式多（结构化的和非结构化的），总体规模巨大且时刻变化。基于大数据的软件工程研究是指在获取和组织软件过程大数据的基础上，采用新的视角和分析方法，一方面重新审视软件工程的基本问题，突破原有的认知瓶颈，另一方面探讨新的历史条件下新的开发模式，探寻新的软件过程规律，从而丰富人们对软件和软件开发复杂性的认识，建立新的软件理论和方法。

目前，大数据技术的运用仍存在一些困难与挑战，体现在大数据挖掘的四个环节中。首先在数据收集方面。要对来自网络包括物联网和机构信息系统的数据附上时空标志，去伪存真，尽可能收集异源甚至是异构的数据，必要时还可与历史数据对照，多角度验证数据的全面性和可信性。其次是数据存储。要达到低成本、低能耗、高可靠性目标，通常要用到冗余配置、分布化和云计算技术，在存储时要按照一定规则对数据进行分类，通过过滤和去重，减少存储量，同时加入便于日后检索的标签。第三是数据处理。有些行业的数据涉及上百个参数，其复杂性不仅体现在数据样本本身，更体现在多源异构、多实体和多空间之间的交互动态性，难以用传统的方法描述与度量，处理的复杂度很大，需要将高维图像等多媒体数据降维后度量与处理，利用上下文关联进行语义分析，从大量动态而且可能是模棱两可的数据中综合信息，并导出可理解的内容。第四是结果的可视化呈现，使结果更直观以便于洞察。目前，尽管计算机智能化有了很大进步，但还只能针对小规模、有结构或类结构的数据进行分析，谈不上深层次的数据挖掘，现有的数据挖掘算法在不同行业中难以通用。