大数据科学概述

摩尔定律

摩尔定律是指 IC 上可容纳的晶体管数目，约每隔 18 个月便会增加一倍，性能也将提升一倍。人们还发现这不光适用于对存储器芯片的描述，也精确地说明了处理机能力和磁盘驱动器存储容量的发展，该定律成为许多工业对于性能预测的基础。

大数据时代

随着计算机网络的普及，每个人每天都能产出大量的数据，如拍照，定外卖等。但更重要的是计算机产生的数据可能远远超过我们个人所产生的，机器日志、传感器网络和零售交易数据等——所有这些都将产生巨量的数据。

截止到 2012 年，数据量已经从 TB（1024GB=1TB）级别跃升到 PB（1024TB=1PB）、EB（1024PB=1EB）乃至 ZB(1024EB=1ZB)级别。

大数据引发人们思维和行为模式的变革：形成了继实验科学、理论科学、计算科学之后第四种研究范式 - 以数据为中心的科学。

大数据定义

IBM 认为大数据具有高容量、高速度、多类型等“3V” (Volume、Velocity、Variety)特点，具体为：数据体量巨大；数据的存在形式从过去结构化数据为主转换为形式多样，例如半结构化等；数据以非常高的速率到达系统内部。面对这么多数据，需要思考怎么样把它变成信息，把信息变成知识，把知识变成决策。

IDC 认为大数据处理技术代表了新一代的技术架构，这种架构通过高速获取数据并对其进行分析和挖掘，从海量且形式各异的数据源中更有效地抽取出富含价值的信息。

大数据分析

大数据分析可以应用在社会网络建模分析，生命科学，商业智能，智慧城市等。其中，大数据是发展智能城市的助推剂，是城市信息基础设施和智能应用的桥梁。

人工智能 60 年的发展经历了起步发展期，反思发展期，应用发展期，低迷发展期，稳步发展期，2010 年左右深度学习和大数据的兴起带来了人工智能的爆发。

存在多种从数据中产生智能的模式，主要有基于知识规则和逻辑推理、基于大数据和统计学习和在目标环境中利用问题引导等。 其中，基于知识规则和逻辑推理是一种较为基础的模式，该模式与人类的逻辑推理相似、易于理解，但难以构造完备的知识规则库，该模式要求人类设计初始的知识框架并不断提供新的知识。

而大数据智能是以人工智能手段对大数据进行深入分析，探析其隐含模式和规律的智能形态，实现从数据分析到知识分析、进而决策分析的理论方法和支撑技术，具体指：从各类数据源的定位和连接实现数据的采集和汇聚，到对加工数据的深度分析实现知识提取和利用，再到有效整合来源不同知识和数据进行服务。

数据密集型计算

数据密集型计算指面向大量数据的并行计算技术。其中，数据并行是指把数据划分成若干块并分别映像到不同的处理机上，每一台处理机运行同样的处理程序对所分派的数据进行处理。大部分并行处理均采用这种处理方式，尤其是对于计算复杂性很高的问题（如流体力学计算、图象处理）进行并行处理。在这种处理方式中，通常不同的处理机在计算过程中需要进行一定量的通信。因此，在这种并行处理方式中，也需要根据问题的特点设计合理的并行处理算法，以减小处理机间的通信对并行处理性能的影响。

而并行计算是指一个问题分解为多个计算任务，多个计算任务同时进行，并行计算可以在指令级、数据级和任务级进行实现。

高性能计算

高性能计算通常使用很多处理器(作为单个机器的一部分)或者某一集群中组织的几台计算机(作为单个计算资源操作)的计算系统和环境。高性能集群上运行的应用程序一般使用并行算法，把一个大的普通问题根据一定的规则分为许多小的子问题，在集群内的不同节点上进行计算，而这些小问题的处理结果，经过处理可合并为原问题的最终结果。由于这些小问

题的计算一般是可以并行完成的，从而可以缩短问题的处理时间，通常运用高性能计算去解决大规模的科学问题。

云计算

云计算是一种按使用量付费的模式，这种模式提供可用的、便捷的、按需的网络访问，其拥有可配置的计算资源共享池（包括网络，服务器，存储，应用软件，服务），这些资源能够被快速提供，只需投入很少的管理工作，或与服务供应商进行很少的交互。?

云计算服务是一种公共基础设施服务，从硬件的角度，云用户可以按需获取看似无限的计算资源，而不需要复杂的资源规划，消除云用户的事先投入，按需求增加硬件资源，以很短的时间为单位付费使用资源。