概述

1 大家好，我是来自北京理工大学计算机学院数据科学与知识工程研究所的车海莺
北京理工大学计算机科学学院，
大数据是指传统数据处理应用程序无法处理的庞大或复杂的数据集。大数据也可以定义为来自各种来源的大量非结构化或结构化数据。
从学术角度来看，大数据的出现导致了对很多主题的创新性的研究。这也导致了大数据的各种统计方法的发展。大数据没有统计抽样的方法;它只是通过记录数据来观察和跟踪正在发生的事情。
因此，大数据的规模往往超过传统软件在可接受的时间内处理的能力。
由于最近的技术进步，大数据分析在现代研究中取得了很大进展，比如发布新数据的容易程度，以及世界上大多数政府要求的高透明度。
大数据分析是运用不同方法从海量数据中发现价值的关键。
在这门课程中，我将介绍一些大数据分析的相关概念、理论、技术、算法、平台以及一些应用实验。
在第一节课中，让我简单介绍一下本课程的内容。

6 本课程根据大数据的总体架构进行，包括数据存储系统、数据处理系统和数据应用系统。
在数据存储系统部分，包括数据采集与建模、分布式文件系统、分布式数据库与数据仓库、统一数据访问接口四个部分。
在数据处理系统部分，包括计算算法、计算模型、计算引擎和计算平台。
数据应用部分包括大数据可视化、大数据产品与服务、大数据应用。

7 现在让我们更具体地看看数据存储系统。
数据采集与建模涉及数据源、数据预处理、数据清理、数据提取、标准存储格式和建模。数据模型包括概念模型、逻辑模型和物理模型，三者之间应相互独立，保持一致。
数据物理存储是分布式文件系统，主要包括 Hadoop 分布式文件系统和谷歌文件系统。我们将解释什么是 HDFS，它是如何工作的，包括读写数据和容错机制。
数据逻辑存储分为数据库和数据仓库，本文将对 NoSQL 数据库进行说明，并将 NoSQL 数据库与传统 DBMS 进行比较，包括各自的优缺点、适用场景等。

8 数据存储后，需要对其进行正确的处理。
数据处理需要算法，包括机器学习算法和数据挖掘算法。
机器学习算法包括有监督、无监督、半监督和强化机器学习。
数据挖掘算法包括分类。聚类回归与关联分析。
对于不同类型的数据，我们应该使用不同的计算模型，例如，
海量数据的 MapReduce 批处理模型
动态数据流的流计算模型，
大规模并发处理(如结构化数据的 MPP 模型)
大规模物理内存模型、内存计算模型;
数据流图模型;
算法和模型可以由不同的计算引擎和平台提供，如 Hadoop、Spark、storm、Hana、Green plum、Pregel、MLlib、TensorFlow 等。

9
在对数据进行存储和处理后，根据需求构建数据应用系统。

在数据应用系统中，我们可以嵌入一些大数据产品和服务，为了直接直观地理解数据和数据分析结果，可以应用数据可视化。
通过大数据产品、服务和可视化，可以构建一个可理解的、有意义的大数据应用程序。
在本课程中，我们将解释两个典型的大数据应用，推荐系统和社交网络分析，以深入理解大数据分析。
我们使用一些工具来构建应用程序并将其可视化。

10
为了让学生更深入地理解这些概念和理论，本课程设计了大量的实验并提供给学生。提供的实验材料包括源代码、实验手册和相关说明。
实验涵盖以下主题。
1 .网络爬虫，包括静态网络爬虫和动态网络爬虫，
2 Mllib，一个 Spark 机器学习平台
3 TensorFlow: TensorFlow 由谷歌的人工智能团队谷歌 Brain 开发和维护，广泛应用于各种机器学习算法的编程实现。
还有两个典型的大数据应用项目，
4)推荐系统
5)社交网络分析。

11 这一节我概述了大数据分析这门课程计划讲述的内容，如果有任何问题，请随时与我联系。