CSCE 42703-001
Big Data Analytics and Management

Spring 2025

Course Overview

- Catalog Description
  - Introduction to distributed data computing and management, MapReduce, Hadoop, Spark, NoSQL, and NewSQL systems, big data analytics, and real-time streaming data analysis.

- Goals
  - The goal of the class is for students to understand the technologies used in manipulating, storing, and analyzing very large amounts of data, and to use distributed computing platforms to conduct practical big data analytics and management tasks.

Class Hours

- Class hour 2:00 - 3:15 PM, Tue. & Thur.
  - Location: JBHT 239
- Office hour 3:00 - 4:00 PM, Fri. or by appointment
  - Location: JBHT 522
- Instructor – Lu Zhang
  - Email: lz006@uark.edu
  - Office: JBHT 522
  - Webpage: http://csce.uark.edu/~lz006/
- Course Website: https://csce.uark.edu/~lz006/course/2025spring/42703/42703.html

Course Materials

- Textbook
  - Mining of Massive Datasets by A. Rajaraman, J. Ullman, and J. Leskovec
    - Available for free at http://mmds.org
- Reference Materials
  - Big Data: Principles and Paradigms (1st Edition) By R. Buyya, R. N. Calheiros, A. V. Dastjerdi. (2016).
    - Available on course website
- Other reading materials will be posted on the course website

Course Prerequisite

- CSCE 3193 or CSCE 3193H Programming Paradigms or DASC 2103 Data Structures & Algorithms, each with a grade of C or better.
  - Good programming skills in Java or Python.
- Probability and statistics basic concept
- Knowledge of data mining will be a plus

Grading

- Composition
  - Homework 48%
  - Midterm 20%
  - Final 30%
  - Attendance 2%

- The final class grade will be assigned according to the 10-point scale shown below. The grades may or may not be curved.
  - A 90 – 100%
  - B 80 – 89.9%
  - C 70 – 79.9%
  - D 60 – 69.9%
  - F < 60%

Assignment

- There will be 7+1 assignments that will enhance understanding of the material taught in the course.
  - Electronic submission through Blackboard

- Late policy
  - 10% penalty for each day after the due date for up to 3 days late.
  - Submission more than 3 days late should be submitted together with an explanation.
  - Weekends count as 1 day.

Mid-term and Final Exams

- Date and time TBD
- Closed book: students ARE allowed one 8.5x11 page of handwritten notes (double-sided) and a calculator, but they are NOT allowed any other materials or other electric devices such as cell phones, smart watches, tablets, or computers.

University Policies

- Academic Integrity
  - Refer to https://honesty.uark.edu/policy/
- Emergency Preparedness
  - Refer to http://emergency.uark.edu/
- Inclement Weather
  - Refer to http://safety.uark.edu/inclement-weather/
- RazALERT
  - Refer to http://safety.uark.edu/emergency-preparedness/emergency-notification-system/
- Academic Support
  - Refer to http://www.uark.edu/academics/academic-support.php

Academic Dishonesty Policy

- As a core part of its mission, the University of Arkansas provides students with the opportunity to further their educational goals through programs of study and research in an environment that promotes freedom of inquiry and academic responsibility. Accomplishing this mission is only possible when intellectual honesty and individual integrity prevail. Each University of Arkansas student is required to be familiar with and abide by the University's ‘Academic Integrity Policy’ at honesty.uark.edu. Students with questions about how these policies apply to a particular course or assignment should immediately contact their instructor.

About Use of Generative AI (e.g., ChatGPT, GitHub Copilot, etc.)

- Students are NOT allowed to use generative AI in all assignments.

Briefly About Me

- B.S. – University of Science and Technology of China
- Ph.D. – Nanyang Technological University, Singapore

- Research interests: data mining, machine learning, artificial intelligence, particularly in fair machine learning and AI, causal modeling and inference, natural language processing

Introduction to Big Data

Adopted from slides by Xintao Wu,

Jure Leskovec, Anand Rajaraman, Jeff Ullman, http://www.mmds.org

Big Data Era

- Google: in 2010 every 2 days we create as much data as we did up to 2003.
- Facebook: 500+ TB of new data every day including
  - 2.5 billion items shared
  - 2.7 billion Likes
  - 300 million photos
- Twitter: 500 million tweets per day

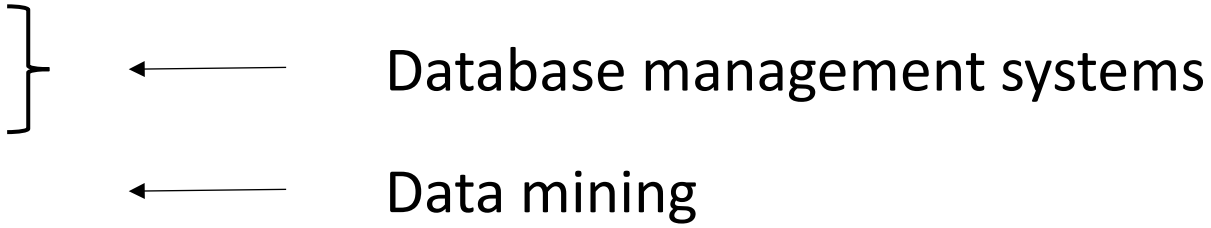
The Big Data Comes from

- Activity data
- Conversation data
- Photo and video image data
- Sensor data
- The Internet of Things (IoT) data
- …

Data contains value and knowledge

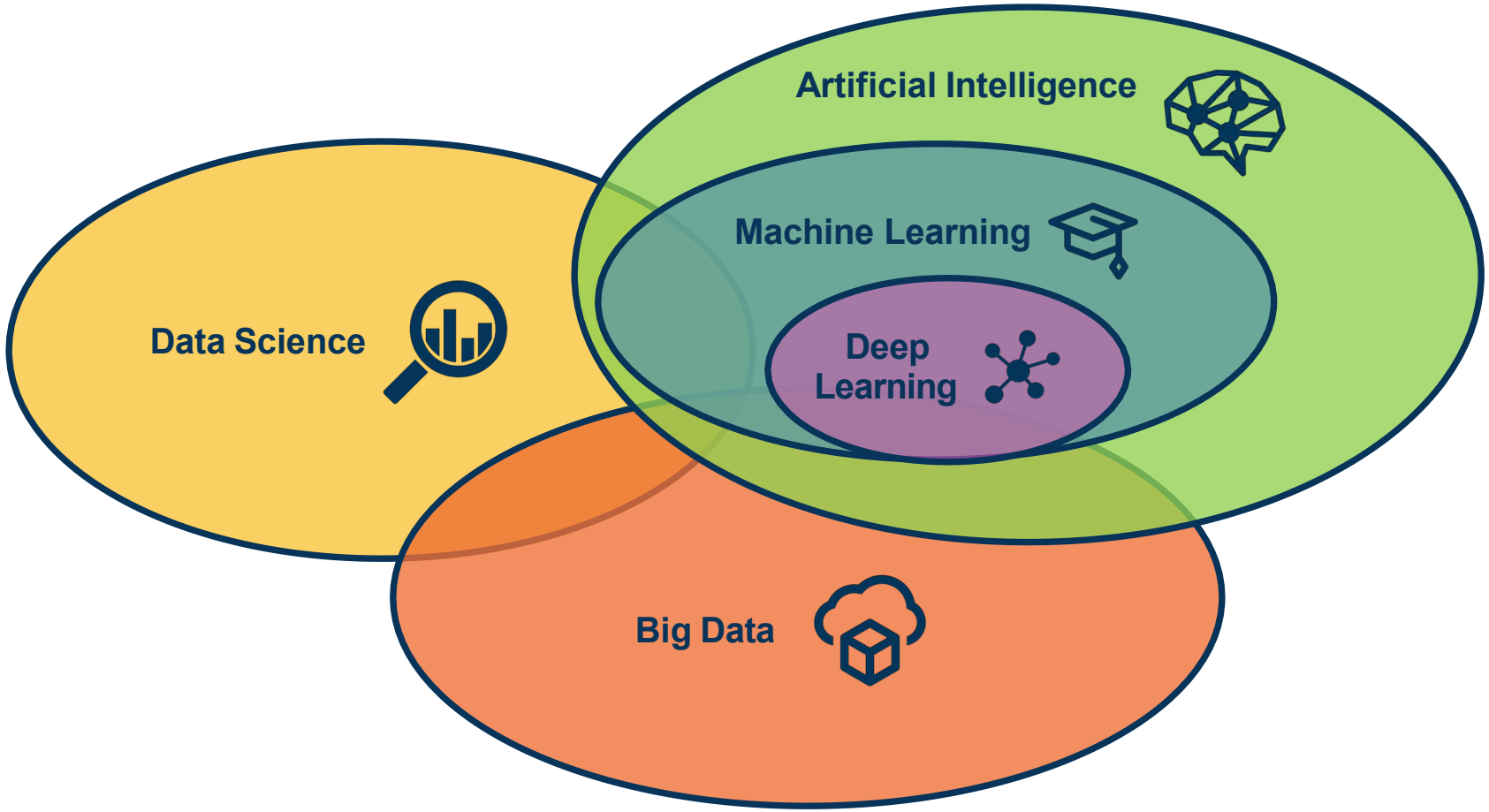
From Big Data to Knowledge/Value

- To extract the knowledge data needs to be
  - Stored
  - Managed
  - Analyzed



- Big data: traditional techniques cannot handle

Notes on terminology



Related Courses

- CSCE 41403. Data Mining
- CSCE 47803. Cloud Computing and Security
- CSCE 46103. Artificial Intelligence

- CSCE 52003. Advanced Database Systems
- CSCE 50603. Machine Learning
- CSCE 55603. Introduction to Deep Learning
- CSCE 57003. Computer Vision

Characteristics of Big Data

- Big data is usually characterized with 4V’s
  - Volume
  - Velocity
  - Variety
  - Veracity

Volume

- We need to deal with terabytes (10^{12} bytes) but zettabytes (10^{21} bytes) or even larger amounts of data
- Simple problem can become difficult at this magnitude
- Example: matrix-vector multiplication

$$\boldsymbol{v}' = \boldsymbol{M} \cdot \boldsymbol{v}$$

where $\boldsymbol{M}$ is an $n \times n$ matrix
and $\boldsymbol{v}$ is a vector of length n

In PageRank, n is in the billions

Volume

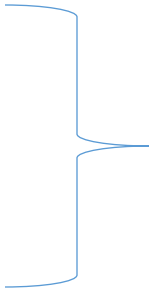
- Data cannot be stored in a single storage node
- The database system need to partition the data into one or more parts, which may be located at different locations
- The transmission of the data from the storage node to the compute node is time consuming

Velocity

- Data arrives so rapidly that it is not feasible to store it all
- Need ways to analyze the data without putting it all into databases

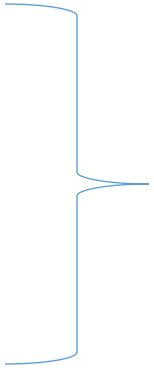
Variety

- Traditional data
  - Documents
  - Records
  - Files



Structured data that neatly fit into tables

- Big data
  - Photographs
  - Audio & video
  - Location data
  - Conversations
  - Sensor data

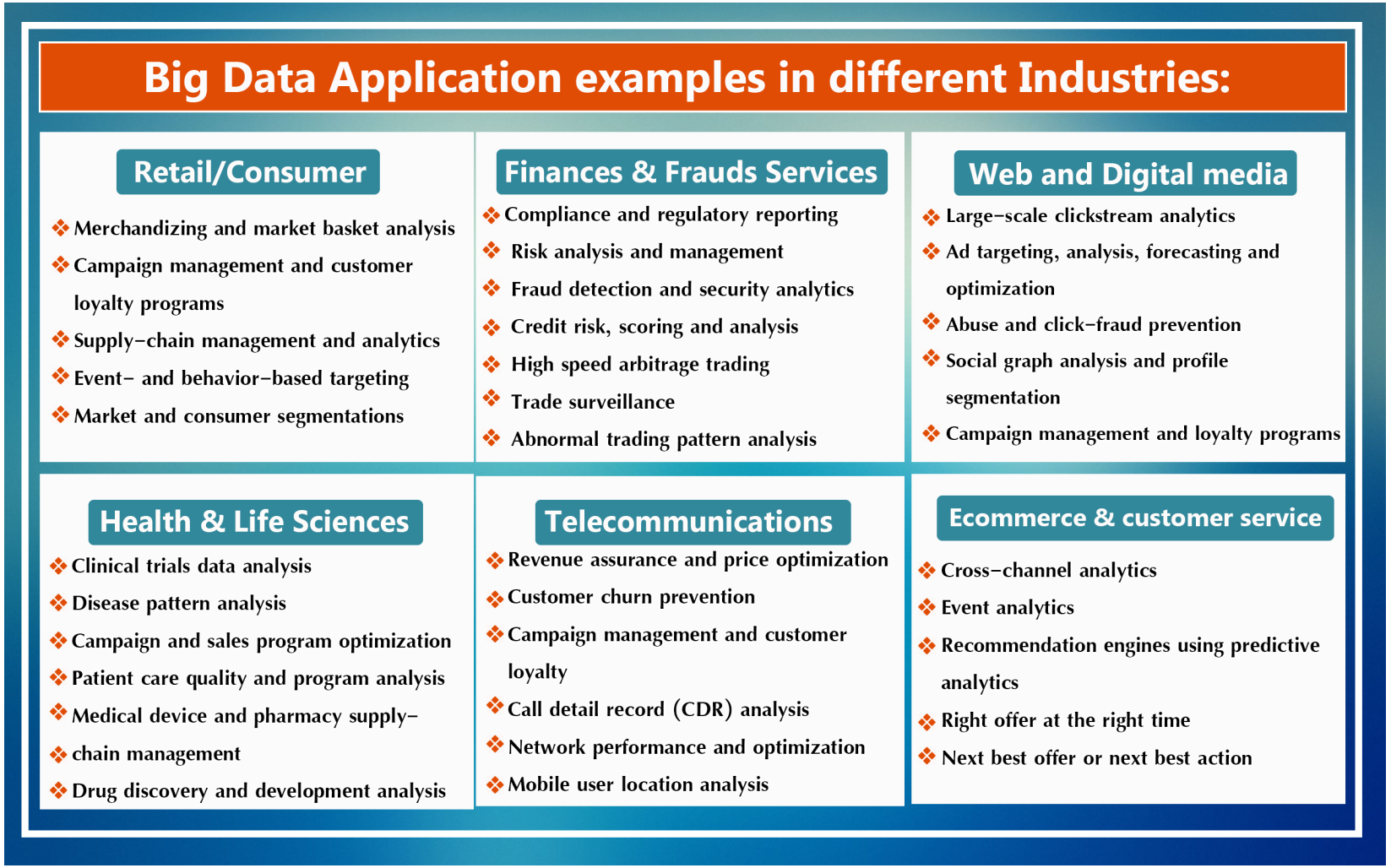


Unstructured data

Veracity

- Trustworthiness
- Privacy
- Fairness
- Accountability

Applications of Big Data



Topic Outline

- MapReduce, Hadoop and Spark
- Frequent itemset mining & association rules
- Finding similar items & locality sensitive hashing
- Large-scale clustering
- PageRank and link analysis
- Mining large data streams
- Database management systems
- NoSQL databases
- NewSQL databases

What Will We Learn

- Models of computation
  - MapReduce
  - Streams and online algorithms
- Platforms for big data computing
  - Hadoop
  - Spark
- NoSQL and NewSQL systems
  - BigTable, Hbase, Cassandra, Hive, Pig, MongoDB, Neo4j
  - H-Store, VoltDB, Amazon RDS, Microsoft SQL Azure, Google Spanner, SAP HANA

What Will We Learn

- Data mining algorithms and how they handle big data
  - Association rules
  - Locality Sensitive Hashing
  - PageRank
  - k-means
  - Bloom filters

Programming with Clusters

- Make use of clusters at AHPCC
- Your account is already created.
  - To verify your account, open a web browser, on the address bar type in
  - hpc-portal2.hpc.uark.edu
  - On the popup window, enter your uark ID and password
  - You should be on the dashboard page
- Off-campus access to the hpc portal
  - Install and run GlobalProtect VPN first
  - https://its.uark.edu/network-access/vpn/index.php
- Email me if you cannot access the hpc portal.