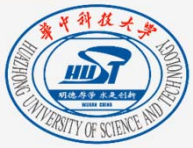


大数据分析

崔金华 副教授

邮箱：jhcui@hust.edu.cn
个人主页: https://csjhcui.github.io/
实验室主页:http://cpss.hust.edu.cn/index.htm
办公地址: 武汉市洪山区珞喻路1077号东湖广场柏景阁1单元1568 室

华中科技大学人机物系统与安全实验室

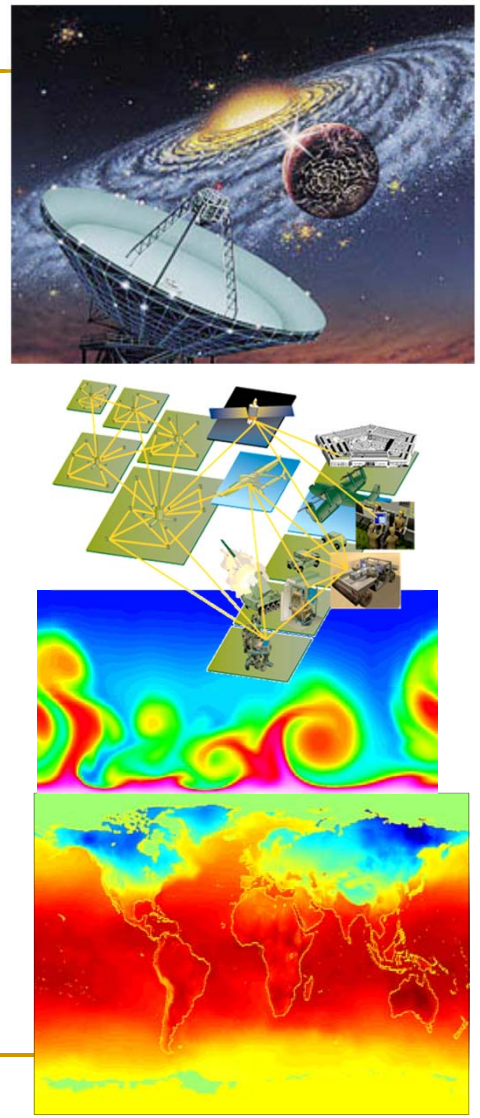
Course Introduction

- Data Analytics and Mining of Massive Datasets
- Why Data Mining? Commercial Viewpoint
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Course Introduction

- Why Data Mining? Scientific Viewpoint
- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation



Mining Large Datasets - Motivation

- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to extract the knowledge data
- Much of the data is never analyzed at all

Data contains value and knowledge!

Data Mining

- But to extract the knowledge data needs to be
  - Stored
  - Managed
  - And **ANALYZED** ← **this class**

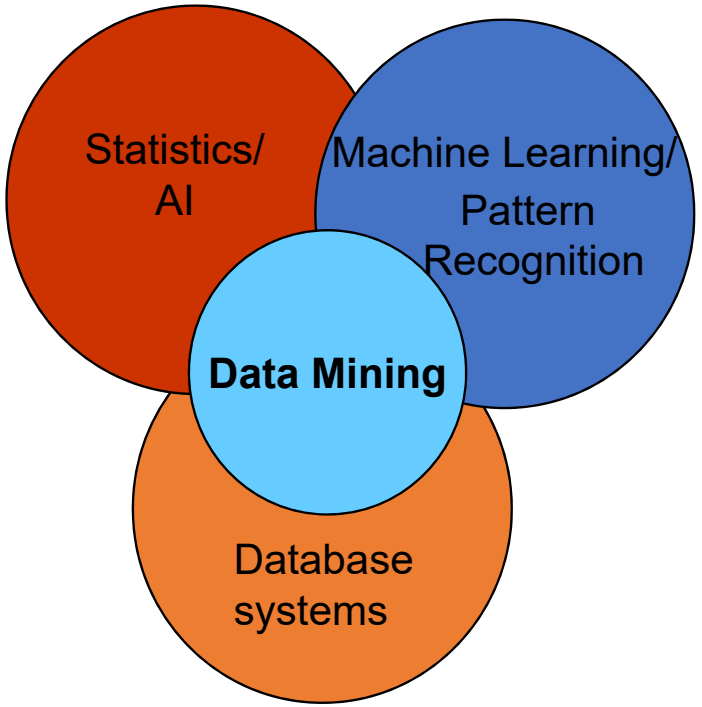
- Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science

What Is Data Mining?

- Given lots of data
- Discover patterns and models that are:
  - Valid: hold on new data with some certainty
  - Useful: should be possible to act on the item
  - Unexpected: non-obvious to the system
  - Understandable: humans should be able to interpret the pattern

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
  - **Example**: Clustering

- Description Methods
  - Find human-interpretable patterns that describe the data.
  - **Example**: Recommender systems

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

About the course

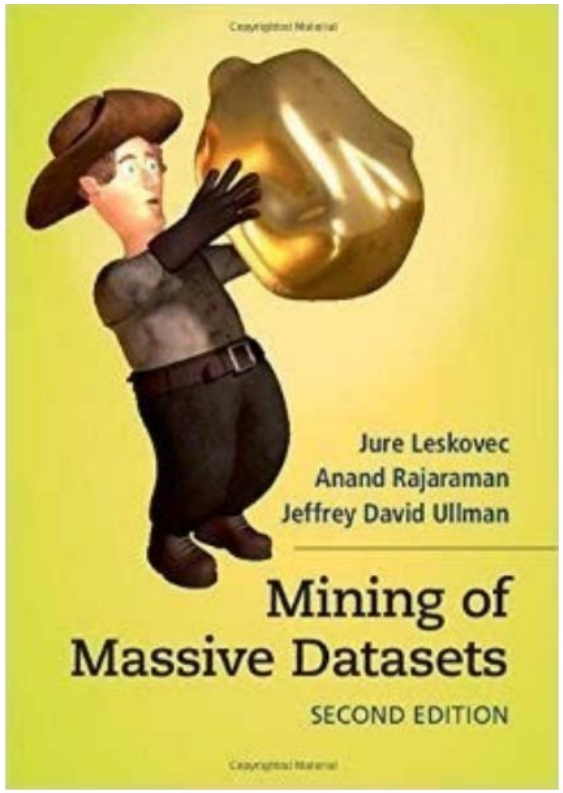
师资团队

崔金华 副教授
主讲教师、实验教师

杨驰 副教授
实验教师

助教团队:
曾志敏, 唐恺, 方思桐, 谭子熠等

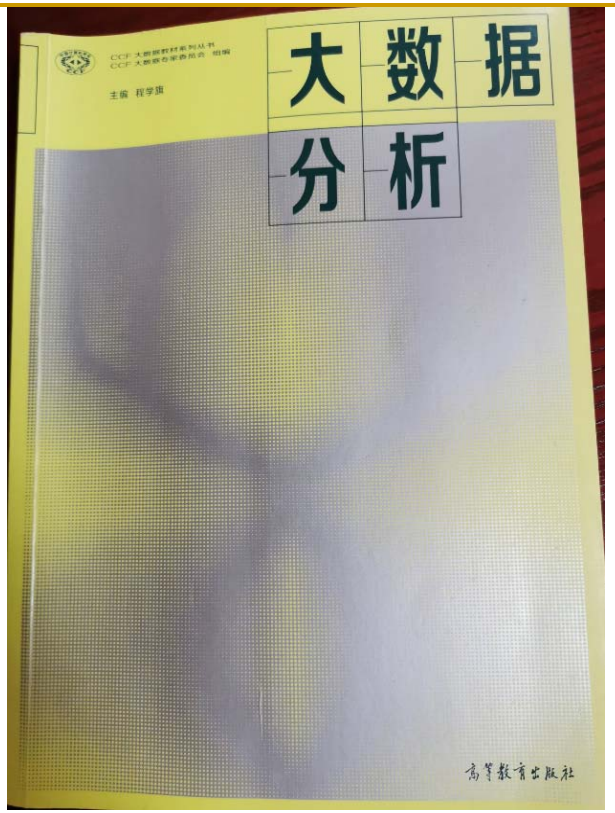
Books



Mining of Massive Datasets, Second Edition

Jure Leskovec,
Anand Rajaraman,
Jeffery David Ullman

Free online:
http://infolab.stanford.edu/~ullman/mmds/book.pdf



《大数据分析》 程学旗 主编
高等教育出版社, 2019

In this class, we will talk about...

chapters	hour (total: 24)
Graph Data: PageRank	3
Graph Data: Other Algorithms	2
MapReduce	3
Recommendation Systems (I)	3
Recommendation Systems (II)	2
Dimensionality Reduction	3
Association Rules (I)	2
Association Rules (II)	2
Clustering (I)	2
Clustering (II)	2

Prerequisites

- Introduction to Big Data
  - Overview for big data store, manage, and analytics
- Algorithms
  - Dynamic programming, basic data structures
- Basic probability
  - typical distributions, Maximum Likelihood Estimate, …
- Programming
  - Python will be very useful
  - e.g. [book]《大数据的Python基础》董付国 著; [video] MOOC: Python语言程序设计; coursera: Python for everybody

- We provide some background, but the class will be fast paced

学生课程成绩综合评价考核办法（选修）

- **总成绩**由以下四项组成：
  - ★ 课堂考勤（10%）
  - ★ 课后作业（10%）
  - ★ 实验课小作业（30%）
  - ★ 实验课大作业（50%）

课程答疑&教学资源获取方式

- 华中科技大学网络教学平台

登录http://hust.fanya.chaoxing.com/portal后加入课程编号为w126044的大数据分析的课堂中;或打开手机的“学习通”软件在首页右上角输入84967919也可加入课堂.

课程答疑&教学资源获取方式

- QQ群：746263866
  1. 扫码关注QQ群：大数据分析(2022年)
  2. 加入群聊，填写学生资料加入，注意提供实名制验证信息：班级_姓名，例如CS2010_张三

群名称：大数据分析(2022年)
群　号：746263866