Lecture 1: Syllabus and Motivation

COSC 526: Introduction to Data Mining

Spring 2020

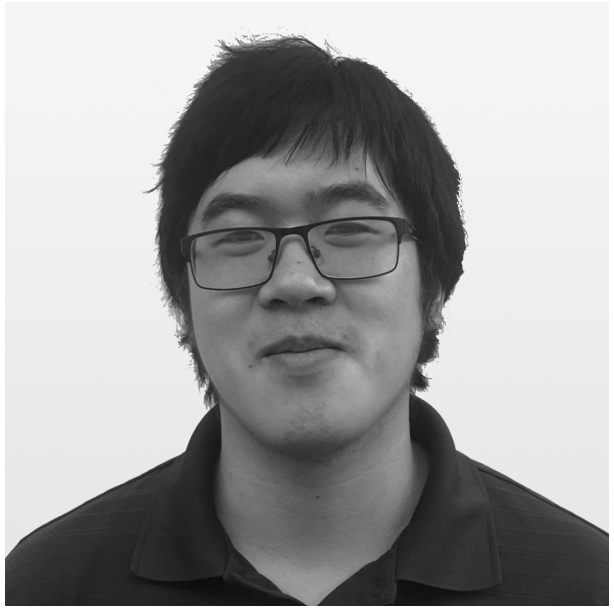
THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Instructors:

Michela Taufer

Danny Rorabaugh

GRA:



Nigel Tan

Course goals

- Build and use environments in which research on data can be designed, performed, and shared
  - GitHub, Jupyter Notebook, XSEDE Jetstream cloud, HPC systems
- Use distributed programming models and associated framework to analyze the data
  - MapReduce and Spark
  - MapReduce Benchmarking
- Design MapReduce-based algorithms and run them in the Cloud
  - Well-known algorithms: Clustering and data processing
- Challenge yourself in the 4-week Hackathon

Lecture structure

- Short lecture (~60 minutes) to introduce a topic and define one or multiple practical problems related to that topic
- Work on the practical problems in team or by yourself
- Group discussion and assessment of achievements
- Push of results (e.g., solutions and comments) in your private GitHub
- What if you need some more time to solve your problems?
  - Complete the unfinished work during the week and submit before the next lecture on Friday at 8AM (hard deadline)

Assignments

- Complete unfinished work during the week and submit before the next lecture on Friday at 8AM ET (hard deadline)

Course requirements

- Students have to bring their own laptop to the lecture
- No text is required
- Python programming skills requested
  - If you feel Python is not your forte, you are welcome to stay in the course but you will need to catch up with the programming skills in the next three weeks by yourself
- Weekly submissions are mandatory
  - But if you complete your work in class, you have only to read the paper and submit the summary (mandatory format)

Grades

- Participation and submission of practical problems: 50%
- Project with poster and 2-page paper: 50%

Office hours

- Instructor: Friday 2:00PM – 3:15PM or by appointment (sent email to taufer@utk.edu) – room: Min Kao 620
- GRA: Wed 3-4pm and Thur 10-11am, or by appointment (with 24 hours notice) – room: RA office

Outline of today's lecture

- VIP talk:
  - Genevieve Bell (Intel) on the origin of data analytics
- Establish a collaborative environment
  - Install git, GitHub, Jupyter, and learn how to use the tools
- Establish familiarly with text parsing
  - Handling files in different formats and different text formats
  - Code developed today will be used to get familiar with GitHub next week
- Learn to share solutions and discuss ideas

Building our motivation

Building our motivation

- Intel's Genevieve Bell shows that we have been dealing with big data for millennia, and that approaching big data problems with the right frame of reference is the key addressing many of the problems we face today from the keynote of Supercomputing 2013:

  https://youtu.be/CNoi-XqwJnA

- Your task:
  - List three key concepts you learned by watching the video