# Lecture 1 - Introduction

## DSE 511

Drew Schmidt
2022-08-25

# Introductions

# About Me

- Taught Math/Stats many years ago
- Academic Researcher
  - UTK (2011-2017)
  - ORNL (2017-2021)
- Currently work at private company
- Professional interests:
  - R
  - Computational linear algebra
  - HPC

# Contacting Me

- Email
    - Drew Schmidt mschmid3@utk.edu
    - Please add [DSE511] to the subject line
- Slack - UTKDSE organization (I will invite you)

# Software

- pbdR https://pbdr.org
- fml Project
  - Code https://github.com/fml-fam
  - Blog https://fml-fam.github.io/blog/
- HPCRAN https://hpcran.org/
- GitHub https://github.com/wrathematics

# Publications

- Schmidt, D., 2020, November. A Survey of Singular Value Decomposition Methods for Distributed Tall/Skinny Data. In 2020 IEEE/ACM 11th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems (ScalA) (pp. 27-34). IEEE.
- Hasan, S.S., Schmidt, D., Kannan, R. and Imam, N., 2019, December. A Scalable Graph Analytics Framework for Programming with Big Data in R (pbdR). In 2019 IEEE International Conference on Big Data (Big Data) (pp. 4783-4792). IEEE.
- Schmidt, D., Chen, W.C., Matheson, M.A. and Ostrouchov, G., 2017. Programming with BIG data in R: Scaling analytics from one to thousands of nodes. Big Data Research, 8, pp.1-11.

# Course Structure

# Course Files

- Course content will be posted on Canvas
    - Syllabus
    - Slides
    - Assignments
- Also on GitHub if you can find it
- Assignment 1 is live (it's short)
- Slides will be posted after presentation

# About the Course

- Introduction to Data Science and Computing (511-512)
- 511: version control, scripting languages, relational and non-relational databases, proper use of data structures, introduction to data science work flows, introduction to project management, and applications.
- 512: platforms for scalable computing including Map Reduce, Hadoop, Spark, and HPC, setting up computing in cloud, and modern data science work flows.

# About the Course

- *The computer science of data science*
- 511: introduction to data science tooling and workflows
- 512: profiling and performance optimization (and remote computing)

# Course Content

- Module 1: Version Control
- Module 2: Basic Programming with R and Python
- Module 3: Introduction to the Shell
- Module 4: Databases
- Module 5: Course Wrapup

# Course Structure

## Class Hours

- Lecture component
- Interactive component

## Content

- Some "theory"
- Lots of direct exposure to technologies

## Outside of Class

- ~5 homework assignments
- No projects, exams, etc.

# Assignments

- Daily late penalties
- Some collaboration with other students ok - but there is a line!
- More info in syllabus

> Some limited coordination on the homeworks is fine. You can even credit your friend in the solution if you like - giving credit to people who deserve it is a good thing! However, directly copying code is unacceptable.

# Questions?