University of California, Berkeley
Master of Information and Data Science (MIDS)
W205 – Fundamentals of Data Engineering

# Week 1 – Introduction to Data Engineering

# Agenda for Today's Class

- Welcome
- Introductions
- Overview of the Course
- Attendance
- Asynch High Level Review in a Nutshell
- Breakouts
- Summary

# Welcome !

- For students new to UC Berkeley:
  - Welcome to UC Berkeley !
  - Welcome to the iSchool !
- For all students:
  - Welcome to w205 !

Our goal is for everyone to be extremely successful in w205 !

# Introductions

- Instructors introduce themselves
- Students introduce themselves

# Overview of w205

GitHub => ucb_mids_w205_repo => README.md

- Spend about 5 minutes briefly scrolling through and pointing out the major sections
  - Slack
  - Schedule and Due Dates
  - Office Hours
  - Grading
  - Grading Scale
  - Attendance and Participation
  - Asynch Assessments
  - Projects
  - Readings
  - Cloud

# Attendance and Participation

Please record your attendance and participation for today's class:

GitHub => ucb_mids_w205_repo => README.md => Attendance and Participation

# Asynch High Level Review in a Nutshell

Each week we will spend about 15 minutes reviewing the most important high level concepts from the asynch

# Prerequisite Knowledge for w205

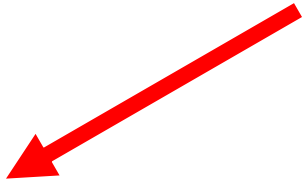
- Python programming
  - Taken or gone through the Python course
  - Week 2 asynch has optional review of data visualization
- SQL
  - Week 2 asynch has a SQL refresher
- Linux CLI
  - Week 3 asynch has a Linux CLI refresher
- GitHub Git CLI
  - Week 3 asynch has a GitHub Git CLI refresher

## Data Engineers

Design, create, and deploy data pipelines

## Data Scientists

Process the data using sophisticated mathematical/statistical models, artificial intelligence, machine learning, deep learning, etc.



Data engineers design, create, and deploy data serving layers

# Full Stack Data Scientists

Data scientists who can do their own data engineering

# Degree of Separation of Roles

- Data engineers separate from data scientists
  - Data scientists need to work effectively with data engineers
  - Breakout exercise later today

- Full stack data scientists
  - Typically smaller companies and startups
  - Although becoming popular at larger companies

# Learning How to Learn is the Best IT Skill to Have

- Technology and products come and go at fast pace
- Must constantly learn new skills
  - Videos, tutorials, books, etc.
- Bleeding edge technologies are more challenging to learn than established technologies
- Data science often requires bleeding edge
- Frustration and confusion in learning are not always a bad thing
- Breakout exercise later today

# Debugging Skills

- Work through problems
  - Calmly, systematically, logically, orderly
- Optimized debugging
  - Each step cuts down on search space as much as possible
  - Reach successful resolution as soon as possible
- Working Independently
  - Learn to do as much as we can before involving others
  - Solve most issues without involving others

# Cloud

Brief overview (we will spend more time in coming weeks)

- VM: emulation of physical hardware computer
- Containers: emulation of OS
- Object store: scale up storage, AWS S3, GCP Store, Swift, etc.
- Edge servers: servers all over the world
- Elastic computing: increase / decrease computing resources on an as needed basis
- Managed services: ready to use software, everything handled for us, pay as you go

# Breakouts

GitHub => ucb_mids_w205_repo => breakouts

(time permitting, we may not get to all of them)

# Summary

Instructor will give a brief (about 2 minute) summary of today’s class.