**University of California, Berkeley**
**Master of Information and Data Science (MIDS)**

**W205 – Fundamentals of Data Engineering**

**2018 Fall, Sections 1, 2, 3, 4, 5**
**Instructor: Kevin R. Crook**

**Instructor's Supplement to the Syllabus**

**Note:** this supplement to the syllabus is only for Kevin Crook's sections and does not apply to other sections. Information, including rubrics, in this supplement will override the main syllabus.

## Primary Syllabus

Prior to the semester start, the primary syllabus can be found at this link:

https://mids-w205-fund-of-data-eng.github.io/course-content

After the semester starts, the syllabus will can be found in GitHub Classroom at this link (requires that you are logged into GitHub Classroom, otherwise it will not work):

https://github.com/mids-w205-crook/course-content

## Office Hours

- Scheduled
  - Wednesday, 4:00 pm to 6:00 pm Berkeley Time
    - I will stay the whole time, so you can come anytime during this time period.
    - Zoom link will be posted to slack (ISVC doesn't allow for a meeting for more than 1 class)
    - At 6:00 pm I have to take a short break and then reset the Zoom classroom for my 6:30 pm class.
- Before, between, or after any of my classes. If you are not in that class, you will need to slack me so I can give you the Zoom link.

## Checklist Before Your First Synchronous Class Meeting

Here is a quick checklist of things you should complete before our first synchronous class session. More details on these items follow in subsequent sections.

- ✓ Ensure you meet prerequisites
  (or make a plan to be proficient with prerequisites prior to the first synchronous class meeting and ask your advisor about Wyzant tutoring if necessary)

- ✓ Slack account and join relevant channels

- ✓ GitHub account

- ✓ Complete my Google Form with your Slack username and GitHub username
  (at least 48 hours before your first synchronous class meeting so I can setup your virtual machine in the cloud and GitHub classroom permissions, otherwise I won't have time to get you setup and you won't be able to complete in class exercises at the first synchronous class meeting)

- ✓ Google Cloud account

- ✓ Access to the O'Reilly textbooks using the Berkeley library or paid Safari account

## Prerequisites

**For MIDS program entry:**

- A working knowledge of fundamental Computer Science concepts, including:
  - Data Structures
  - Algorithms
  - Analysis of Algorithms
  - Linear Algebra
  - Object-Oriented Programming Skills in Python or C/C++ or Java

- Students should have been given access to bridge courses or other resources to complete deficiencies in these areas with specific instructions of which ones are needed for W205 based on their specific circumstances.

- If you have not been given access to these, or a timetable to complete them, please contact your advisor ASAP and let them know!

**For W205:**

The generally stated prerequisites for W205 are:

- Object-oriented programming with Python
- Linux command line
- GitHub
- Database Management (SQL)

I went through the above prerequisites and made a more detailed list. Please go through this list and make sure you are extremely comfortable with all of these. If you have some things you are not comfortable with, we have resources available, including tutoring from Wyzant.

- **Object-oriented programming with Python**
    - Students should have taken W200 Python Fundamentals for Data Science
      OR
      if exempted from taking W200, students should have been given access to the Python bridge course with instructions to complete it prior to taking W205
    - The first project will require students to demonstrate proficiency in the following:
        - Object-oriented programming skills using Python
        - Jupyter Notebook
        - Professionally quality formatting using Markdown
        - Pandas including professionally formatted tables
        - Mathplotlib including professional quality data visualizations

- **Linux command line**
    - vi Editor
        - Ability to create new files and edit existing files on your own without assistance
    - Logging into cloud based Linux command line using:
        - Windows – PuTTY or another terminal emulator that supports ssh
        - Mac – use of ssh on the command line
    - Determining the user and group
    - Switching users, with or without login shell
    - Using sudo for root privileged operations
    - Creating, deleting, copying, renaming (moving) files and directories

- Links: hard links, symbolic links
- File and directory permissions
  - Getting, setting
  - Understanding octal modes
  - Levels: user, group, world
- File and directory ownership and group
  - Determining the owner and group for a file or directory
  - Changing the owner and group for a file or directory
- Determining mount points for mounted file systems
- Seeing which processes are running, soft kill, hard kills (and why to avoid hard kills)
- Downloading files from the Internet using curl
- Making web API calls using curl
- Knowledge of basic networking using TCP/IP

- **GitHub**
  - Using the git command line utility on Linux (not the GUI!)
    - Clone a repo
    - Update a repo
    - Branches:
      - Create a branch
      - Select a branch
      - Ability to work on a branch leaving the master branch untouched
    - Tracking files, staging files, committing changes, pushing changes to GitHub
  - Using GitHub web interface
    - Creating a pull request with a chosen reviewer
  - GitHub Markdown formatting
    - Headers of various levels
    - Formatting source code with syntax highlighting for SQL, Python, and yml
    - Formatting tables

- **Database Management (SQL)**
  - Reading and understanding Logical Data Models (LDMs) using Entity Relationship Diagrams (ERDs) in Third Normal Form (3NF)
  - Relational algebra: projection, restriction, aggregation, inner joins, outer joins
  - SQL constructs including: SELECT, FROM, WHERE, JOIN, LEFT OUTTER JOIN, GROUP BY, HAVING, ORDER BY, Type 1 Subqueries, Type 2 Subqueries.

## Available Resources to Help You Master Prerequisite Materials

### Official Bridge Courses provided by Berkeley

Bridge courses have been developed for you use for this purpose. These are your best resources because they have been developed specifically for the MIDS program with MIDS courses in mind. Your advisor should have given information on accessing these. If not, please check with them to gain access.

### Some Free Online Resources Provided by 3rd Party (not Berkeley)

These are some free resources, however, they are 3rd party, and not specifically geared towards the MIDS program.

**Linux command line**

http://linuxcommand.org/lc3_learning_the_shell.php

**GitHub**

https://git-scm.com/book/en/v2

**Database Management (SQL)**

https://www.w3schools.com/sql/default.asp

### Private Tutoring (on your own, at your own expense)

Some students have found private tutors on their own at their own expense. We cannot recommend or advise you on a tutor.

## Slack Account and Joining Relevant Channels

- **slack.com**
- Please use slack to communicate as it's the easiest way to reach me. Email will take longer. I try to check slack once a day, except on weekends, but cannot guarantee it.
- Please join the **ucbischool** team using your **@ischool.berkeley.edu** email address, not your @berkeley.edu address, nor any other, as it will not be approved.

- Please join the following channels:
    - **#w205**
        - all sections for all instructors for w205
        - please put all technical questions here so other students may benefit from getting them answered
    - **#w205-crook-fall-2018**
        - all section for Kevin Crook for Fall 2018
        - my primary means of communication with my students about non-technical issues
        - please do not use this channel for technical issues, use #w205
        - please check daily on weekdays if possible

## GitHub Account and GitHub Classroom

- **github.com**
- You will be using GitHub Classroom to submit all of your assignments this semester.
- You will need a GitHub account, which I will add to the permission group for my classroom.
- There is not a separate account for GitHub classroom.
- Assignment link – for security reasons, these links will be placed in slack. You will need to accept these links in a timely manner before they expire.
- While it is not needed for this course, GitHub has an academic discount which essentially allows you to create private repos for free.

## Please Enter Your Slack and GitHub Usernames into my Google Form

I'll post a link in Slack to a Google Form where you can enter your Slack and GitHub usernames. Please be sure and do this at least 48 hours prior to your first synchronous class meeting so I have time to grant access to your virtual machine and GitHub classroom.

For security reasons, I cannot post the link publicly, so that is why I post it in Slack.

## Google Cloud Platform (GCP)

Please sign up for an account with the Google Cloud Platform. It is recommended that you sign up with your @berkeley.edu account in order to receive student discounts. If you have an @ischool.berkeley.edu account, you may need to get an alias to @berkeley.edu in order to receive this discount.

## O'Reilly Textbooks – Berkeley Library or Safari

This course will use part of several O'Reilly Textbooks.

The Berkeley has most of these books available free:

http://www.lib.berkeley.edu/

Safari is a subscription service for online books. You may want to check with your employer to see if they provide a free subscription. They have a free trial period, but it's only a couple of weeks. They have a paid service which runs around $45 a month.

https://www.safaribooksonline.com/

## Setting Student Expectations for the Learning Process in MIDS

**Learning how to learn new skills is the best skill to have**

Think about the Data Science skill set 10 years ago… 5 years ago… even 1 year ago. Obviously, 5 years in Data Science sees as much change as many industries see in 10, 20, or even 30 years.

Whenever you learn something new in the MIDS program, in addition to learning that skill, also think about how you learned that new skill, what worked and what didn't work for you, and use this to develop your ability to learn new skills.

**Make Peace with Theory**

Theory teaches us timeless foundational skills that we can use for a lifetime to learn new skills. If we limit ourselves to only learning the specific skills of the moment, and don't learn theory, we deprive ourselves of the ability to generalize theory to learn new skills as time passes. We essentially date ourselves and become outdated as our skills become no longer useful.

**Make Peace with Confusion and Frustration in Learning**

Confusion and frustration in learning are not necessarily bad things. Learn to embrace them as a good thing.

Being confused when learning something, the frustration involved with struggling with resolving that confusion, and ultimately successfully resolving confusion is actually the best way to learn a new skill!

Also, if you are confused and frustrated when learning something new, it demonstrates that you are pushing yourself out of your comfort zone to achieve your maximum potential.

Think back to a time where you were trying to learn something difficult and were totally confused and frustrated. Eventually when you successfully resolved that confusion, didn't you understand that much better than being "spoon fed" the topic?

If you are limiting yourself to only learning easy skills that can be "watered down" and "spoon fed" to you, and you expect to never be confused nor frustrated with learning, you are cutting yourself and your potential short.

Whenever you experience confusion and frustration in learning, how do you handle it? Is asking for help your first resort, or your last resort? Try making it your goal to embrace confusion and frustration and ask for help as a last resort rather than a first resort.

**You don't have to know everything about a software tool to start using it**

Too often students feel they have to know everything about a software tool to be able to use it. Of course, we want to know as much about tools as we can, especially if we need to frequently use them, however, we don't want to be stuck in a paralysis of not being able to use a new tool until we know everything about it.

**Understand different categories of learning:**

> **Rote Memorization** – the lowest form of learning, however knowing some common facts you frequently use can be very useful.
>
> **Generalization** – being shown how to solve a specific problem and being able to generalize it to a new problem that is very similar
>
> **References** – we have a fact we don't remember or a problem we don't know how to solve, we search for and find the fact or the solution using references.
>
> **Synthesis** – taking the ability to generalize the solutions to several different problems and combine them to solve a new problem you have never seen before. Embrace and develop your ability to synthesize solutions, and don't view it as an evil to be avoided.

**Research** – the highest form of learning. Being presented with a new problem, we attempt to synthesize a solution, but we find parts are missing. We have to invent new generalization to the parts that are missing and then synthesize them to solve to new problem. Embrace and develop your ability to take synthesis to the next level of research, and view research as the highest form of learning instead of viewing it as an impediment to teaching / learning.

**Open ended assignments give you the best opportunity for synthesis and research**

When given an open ended assignment, embrace it as a great chance to find opportunities for synthesis and research. Don't try to turn it into a generalization or reference problem, as you are just hurting yourself by depriving yourself of an opportunity for the highest form of learning.

## Format of the Asynchronous and Synchronous Sessions

In a traditional face-to-face 3 credit hour course, you typically spend 3 hours per week in class. In our format, you spend 1.5 hours watching pre-recorded video lectures (asynchronous) and 1.5 hours in a live class via videoconference (synchronous).

Some students seem to expect the synchronous to be a repeat of the asynchronous. This would not meet accreditation standards which require 3 hours of unique instruction. If we repeated the asynchronous in the synchronous session, we would essential be delivering 1.5 hours of instruction twice.

**Reading Assignments**

- Mandatory
- Must be able to read, study, and understand these on your own

**Asynchronous**

- 1.5 hours of your 3 credit hours each week
- Mandatory
- Videos
- Must be able to watch, study, and understand these on your own
- Will not be directly repeated in Synchronous

**Synchronous**

- 1.5 hours of your 3 credit hours each week
- Mandatory
- Will be complementary to, but not a direct repeat of, the Asynchronous
- Typical activities:
    - **Break Out Sessions** – students will work interactively in groups. In the W205 curriculum this is usually related to the assignments. It is mandatory to participate and contribute to your group.
    - **Instructor Led Activities** – instructor will lead students through an activity. It is mandatory to follow along and keep up. If you have an issue, please let the instructor know before we go on to the next step. Please do not fall several steps behind and then expect the class to stop for several minutes for you to catch up.

## Assignments and Grading Rubrics

**Assignments**

- 12 Assignments
- 10 of 12 will count
    - 2 lowest will be dropped
    - However, to be dropped, the assignment must be turned in with a reasonable attempt
- Each assignment is graded in whole numbers between 0 and 10 points
- General grading rubrics at the assignment level
    - Late Submissions
        - Assignments 1 through 11
            - 1 of 10 points penalty for 1 second late
            - Additional 1 point penalty for each additional 24 hours late
            - After 10 days late, the assignment will be worth zero points
        - Assignment 12
            - Due to end of semester time constraints, late submissions for assignment 12 will have the following late penalties
                - 5 of 10 points for 1 second late
                - Zero for the assignment if 24 hours late
    - Unless otherwise stated in the assignment, GitHub must be used from the git command line utility and follow proper source code control (penalty -1)
    - Submissions in markdown should all be professionally formatted (penalty -1)
    - All data in table format must be formatted in markdown or Pandas (penalty -1)

- For assignments 6 through 12, all steps in class must be repeated and separately annotated, including python and pyspark steps. Instructor will estimate the percentage of steps missing or not separately annotated and assess a penalty of -1 for anything missing and -1 for each additional 10% missing.

| Semester Letter Grade | Rubrics |
|---|---|
| **A +** | • All 12 assignments submitted with a reasonable attempt<br>• The average of the highest 10 of the 12 assignments should be 9.7000 or higher without rounding |
| **A** | • All 12 assignments submitted with a reasonable attempt<br>• The average of the highest 10 of the 12 assignments should be 9.3000 or higher without rounding |
| **A -** | • The average of the highest 10 of the 12 assignments should be 9.0000 or higher without rounding |
| **B +** | • The average of the highest 10 of the 12 assignments should be 8.7000 or higher without rounding |
| **B** | • The average of the highest 10 of the 12 assignments should be 8.3000 or higher without rounding |
| **B -** | • The average of the highest 10 of the 12 assignments should be 8.0000 or higher without rounding |
| **Below B -** | Instructor's discretion, but no harsher than the pattern above. |

**Due Dates for Assignments**

| Assignment Number | Due Date |
|---|---|
| 00 | Does not count. We will work on it in class. Instructor will mock grade 00 after the first class meeting so students can get used to how GitHub classroom works. If you don't complete it during the first class meeting it will not be mock graded. |
| 01 | Monday, September 17th, 6 am Berkeley Time |
| 02 | Monday, September 24th, 6 am Berkeley Time |
| 03 | Monday, October 1st, 6 am Berkeley Time |
| 04 | Monday, October 8th, 6 am Berkeley Time |
| 05 | Monday, October 15th, 6 am Berkeley Time |
| 06 | Monday, October 22nd, 6 am Berkeley Time |
| 07 | Monday, October 29th, 6 am Berkeley Time |
| **Fall Break – Monday, November 5th through Friday, November 9th** | |
| 08 | Monday, November 12th, 6 am Berkeley Time |
| 09 | Monday, November 19th, 6 am Berkeley Time |
| 10 | Monday, November 26th, 6 am Berkeley Time |
| 11 | Monday, December 3rd, 6 am Berkeley Time |
| 12 | Monday, December 10th, 6 am Berkeley Time |