Summer 2017 - W205 – Storing and Retrieving Data
Welcome Message
Kevin R. Crook

I would like to give everyone a warm welcome to this course: W205 Storing and Retrieving Data!

I'm very excited to be teaching this course and look forward to working with each of you this semester.

I'm also very excited about Berkeley offering this Masters in Data Science online. It's a great and wonderful opportunity for students to get an elite degree from the world's number one university for Data Science. Each of you should feel very honored and proud of your accomplishment to be a part of this program.

The potential Data Science holds to make our world a much better place is one of the best reasons to be a part of it. Sure, there are seemingly unlimited opportunities to help commercial companies be more successful, and certainly most of us (including myself) need that to pay the bills, but Data Science holds enormous potential in the non-profit sector as well. Giving of your time and Data Science talents can be enormously impactful in stretching resources for those in need.

My teaching philosophy can be summed up as follows:

To help empower students to be extremely successful, by using these techniques:

- Using active learning for knowledge transfer, rather than rote memorization
- Using project based learning to cultivate and sharpen critical thinking and problem solving skills to apply classroom knowledge to real world problems
- Helping students to gain the confidence to attack and attempt to solve any problem, including the so called "it can't be done" problems

As you read through my bullet points above, you will notice that the first two directly line up with the format and curriculum set forth for this course, which makes it a great fit for my teaching style.

I really enjoy working with students on real world projects for project based learning. It's very interesting to me to see the diversity of opinion and approach that students

take. As many students are new to Data Science, it's very interesting to me to see what someone new to Data Science thinks with fresh ideas and ways of approaching problems. Once someone has been in a field for a while, we tend to get set in our thinking of how to solve different types of problems and new and fresh ideas and approaches become less common. So working with students is a very refreshing change for me.

For the last bullet point, I sincerely wish that each of you will leave my course this semester with a boost in confidence to attach any and all problems. Even if you fail, you will learn in the attempt, and make future attempts for yourself and others easier. (At least you will establish what doesn't work.) So many times through the years, I've heard people say "it can't be done" and so many times, sometimes years later, it's being done.

I would like to tell you a little bit about my industry background:

During graduate school, in the late 1980's, I researched in the area of Artificial Intelligence, which heavily uses the same or similar Machine Learning algorithms we use today in Data Science. It was an exciting time as companies were not just thinking about traditional data processing and operational systems, but starting to think about analytical systems.

In the late 1980's, I remember a speaker on campus who was from the Artificial Intelligence Lab at a major US automaker. On the topic of self-driving vehicles, he estimated that we were at least 100 years away from a self-driving vehicle, with significant changes to infrastructure. Less than 30 years later, we have self-driving vehicles without any changes to infrastructure.

In the early 1990's, every company had an Artificial Intelligence (AI) Lab, or was creating one as soon as they could. A new type of computer, designed for AI, called the LISP Machine, was available for commercial purchase, and companies snapped them up as fast as they could be built. The LISP Machines were designed for multi-processing and Machine Learning. I worked in various AI labs up until the late 1990's.

There were a couple of problems: the LISP Machines just weren't powerful enough, and the hype around AI wasn't coming to fruition as fast as executives were wanting. Another problem was that companies had all of their data in silos in operational systems. Data Warehousing was in its infancy and didn't really take off until the early 2000s. So, in many cases, even problems that were solvable didn't have the necessary data available.

The Internet (and the dot com era) took off like a wild fire in the mid-1990s, so AI went to the back burner, with most AI Labs at companies being shut down. After the dot com era hit bust around 2000, the next focus was Data Warehousing and Business Intelligence. Data Warehousing was exciting because companies were able to copy all of their operational data into the Data Warehouse or Data Marts for analytics – something that we sorely lacked during the AI Lab era. Business Intelligence (BI) was the new term for analytics. (I think of BI as "AI light" or "Data Science light") During the 2000's, I worked in Data Warehousing and Business Intelligence.

Eventually, a new term Data Science was coined for a higher order of BI, borrowing all of the knowledge gained from the previous AI Lab era and adding fresh new ideas and approaches. I have been working in this area ever since.

For me, the best part of Data Science is that we finally have the computer hardware and software with the power to do meaningful work. Most companies have built Data Warehouses and Data Marts to hold copies of all of their operational data, so we now have more data available for analytics. We also tons of unstructured data available from the Internet and methods to store and process these data. We also have data brokers with tons of data for purchase and government data free for download to correlate with in house data.

All of these things coming together in the last few years: computer hardware, software, Data Warehouses, Data Marts, unstructured data from the Internet, brokered data, government data, etc. makes us very fortunate to be working in this field of Data Science. I feel we have finally reached a critical mass for Data Science, and over the new few years, we will all be truly amazed at the great advancements that will be made, including surprising areas least expected.

I would like to tell you a little bit about my teaching background:

I started teaching part-time while working full-time in the late 1980's. Through the years I have taught part-time off and on depending on my work schedule.

Up until three years ago, most of my teaching has been programming classes (first C, then C++, then Java, then Python) and assembly language programming for undergraduate Computer Science majors. I have also taught basic computer literacy at the community college for all types of students, including students in vocational programs such as welding, plumbing, auto body, nursing, EMT, etc. Some of these classes were at a halfway house for young, first time offenders recently released from prison. In teaching "non-technical" students from these backgrounds I was able to experience a new side to learning I hadn't seen before.

The last few years, I have also been teaching as professional faculty for graduate students for courses including: Python Programming for Data Science, Big Data Analytics using Spark, Data Management, and Data Warehousing.

Three years ago, I made a bit of a career change, leaving full-time employment with a goal of consulting half-time and teaching half-time. As most of my friends left the technical route years ago for the management route, now I'm now fortunate to have managers and executives friends who allow me to consult part-time for them on an as needed basis.

I've been very impressed with the Berkeley MIDS program: the curriculum, the learning platform, and the format of using live classes via video conferencing. It's a great and effective mix of learning modes, offline and online.

I would like to tell you my thoughts on W205 – Storing and Retrieving Data

As the name implies, you will be studying ways to store and retrieve data for Data Science.

I like to think of it this way: you may have just written the world's greatest Machine Learning algorithm for a problem for a small example data set on your desktop. Now you need to take it and make it work for real data, at huge volume, and using scale out processing. This course is all about how to do that.

One great thing about this course is that we will be using Python, Cloud Computing, Linux, SQL, Hadoop, and Spark, all of which are great choices for Data Science and very popular in the career market.

Python is often thought of a "glue" programming environment. It has libraries for almost everything you can imagine or want to do, including Machine Learning. At least locally, in the Dallas area, most companies I know of are now favoring Python over other languages for Data Science.

We will be using Linux in the Amazon Web Services Cloud. Many companies are switching to cloud based technologies with Linux virtual machines, so this is a very timely skill set to have.

Companies have tons of data in SQL based databases and also want to store the results of analytics in these databases. They need Data Scientists who can make efficient and effective use of SQL.

With Hadoop and Spark, we will study the two most popular architectures for handling huge amounts of data with scale out processing.

I'm truly amazed at the curriculum that has been designed for this course - a great and visionary mix of all of the major ways to acquire data, explore data, cleanse data, store data (architectures with tradeoffs), query data, etc. It even includes the latest on graph oriented storage and retrieval.

This course has many "hands on" labs, exercises, and a project. In the labs, you will acquire basis skills. In the exercises, you will put together the basic skills from the labs to acquire higher level skills. In the project, everything comes together and you will be amazed at your ability to solve a real world problem!

I think you will find you will be using the knowledge and skills acquired in this course on a daily basis for the rest of your Data Science career.