CIS 9 - End of quarter discussion

Congratulations on finishing up the CIS 9 class! I hope you've gained a few insights into the vast field of Data Science. The following are some next steps that could be helpful if you'd like to investigate data analytics or ML further.

Lower division classes to take to prepare for a Data Science major

[Math](https://www.deanza.edu/math/courses.html): Calculus (Math 1A - 1D, click on the Math link for the class description, at De Anza)

Linear Algebra (Math 2B)

Statistics (Math 10)

[CS](https://www.deanza.edu/cis/schedule.html): Data Structures (CIS 22C, click on the CS link for the class description, at De Anza)

Python (CIS 41A/B)

R (CIS 64H)

Database (CIS 64A/B)

Online classes or MOOCs to take on your own, before formally doing undergrad or graduate work in Data Science

Among many courses on Udemy, Coursera, Edx, etc. here are 2 courses associated with a university:

[CS 109](http://cs109.github.io/2015/index.html) from Harvard: Intro to Data Science for upper division undegrad or first year grad students. Prerequisites are Python and a college level statistics class. [Here](https://github.com/cs109/2015lab4/blob/master/Lab4-stats.ipynb)'s an example lab from the class from a few years ago, some of the material will look familiar to you. The lectures, class notes, labs are all available on github that you can download and go through at your own pace.

[Machine Learning](https://www.coursera.org/learn/machine-learning) by Andrew Ng (Stanford), offered through Coursera. This is "the" classic ML course that is very well known in the field. Prerequisites are Linear Algebra, MATLAB, and a college level statistics class. For a small fee (less than $100) you have access to the videos, lecture notes, assignments, TAs, and have your assignments graded.

Data Science as a major

Up until a few years ago, data science was a specialty in either the computer science department or statistics department at most schools. Recently data science has become a course of study on its own, where students can major or minor in it. Most universities currently offer data science graduate programs (Master's or PhD), but there is a growing number of universities that begin to offer undergraduate (Bachelor's) program in data science.

As of this writing there are 2 CA universities that offer an undergraduate, lower division data science class similar to CIS 9: UCB [Data 8](https://data.berkeley.edu/education/courses/data-8) and UCSD [DS 10](https://sites.google.com/view/dsc10winter2021/syllabus). If you choose to enroll at these schools, the CIS 9 credit can be used for these classes.

For upper division, a well rounded program will include classes in probability and statistics, modeling and machine learning (including AI and deep learning), computer systems, and ethics or social science classes

Since data science is at its best when it's the intersection of computer science / statistics and a specialized domain (a fields such as health care, agriculture, education, physics, business, linguistics...), there are many students with undergraduate degrees in various fields who do graduate work in data science. Having the training in data science and having a background in a specialized domain make these data scientists valuable in their field.

Roles in a data science team

There are specialized roles in a data science team, these roles vary slightly in title and job description from one company to another. The following are a general description of these roles.

* A *data scientist* builds the mathematical ML models. Most data scientists have graduate level training in computer science, mathematics, or statistics, and they have a good understanding of a company’s business objectives in order to build effective models.
* A *data engineer* creates easily accessible data pipelines for the ML models. This means gathering data from different sources and pre-process them (transforming, grouping, filtering) so that the data can be used as input for the model. For large projects, a data engineer needs to work with big data and parallel processing.
* An *ML engineer* uses the pre-processed data with the ML model and to train the model and gain insights into the data. An ML engineer is also the bridge between the data science team and the software team and the business strategy team.
* A *data architect* designs and manages the set of policies and standards that specify how data are collected, used, and stored.
* A *business analyst* has some training in data science but is more knowledgeable in business processes. Business analysts are skillful at presenting data insights and results to the rest of the organization and customers.
* A *software engineer* is responsible for integrating tools, automating tasks, implementing software version control, and developing APIs so the team can do their job efficiently.

As the data science field continues to grow and change, new roles will develop and existing roles can change to adapt with new technologies and methodologies.

This document will need to be updated on a regular basis!