DSC 291: Data Science for scientists and engineers
====================================================================

The last decade has seen an explosion in the size and complexity of datasets. Many popular data analysis software packages: R, Matlab, SPSS, etc. are not adequate when the data size is many times the memory size. New packages such as Hadoop and Spark have been developed that process massive data using commodity computers in the cloud.

In this class you will learn the foundations of big data Analytics, parallel computation, and cloud computing. See detailed Curriculum below.

Data science projects require a close collaboration of two types of people: domain experts and methods experts. This course will consist of teams of two or three students. One student will be a domain expert, the other 1-2 will be methods experts.

What makes a good methods expert? A methods expert is likely to have a BSc in computer science, electrical engineering, mathematics or other computational fields. The methods expert should be expert programmer in the unix / python / jupyter notebooks ecosystem. S/he should be familiar with computer organization and with computing in the cloud. S/he should also have a strong math background, especially Linear Algebra, Calculus of several variables, probability and statistics.

The ideal domain expert is a Phd student that is working in a Laboratory or some other organization in the UCSD campus whose focus on collecting and analyzing data and publishing the results found. S/he should be familiar with statistical concepts such as p-value and percentage of variance explained. The goal of taking the course is to understand big data analysis and go back to his/her organization to implement such methods. Each domain student needs a recommendation letter from the head of his lab certifying that the student is intimately involved in data collection and data analysis in the lab.

In order for the teams to be successful, good communication skills are critical. Students need to be able to explain concepts that they are familiar with to students of the other type without getting bogged down in jargon.

To be accepted to this course, each student would need to fill in a google form. The instructor will accept/reject applicants based on these forms and an interview.

Methods Students:  https://forms.gle/GknVWR8y4xqs6oth8
Domain Students:  https://forms.gle/vZ85QyvifSnrnRPb7
Curriculum:

- I/O limited computation and the memory bottleneck.
- Data parallel computation in the cloud.
- Map-reduce
- Spark
- Unsupervised learning: PCA
- Unsupervised learning: Manifolds, intrinsic dimension, and the graph Laplacian.
- Supervised learning: random forests and boosted trees
- Train/test pitfalls: working with heterogeneous datasets.
- Bootstrap, margins and Robust prediction.
- Neural Networks and TensorFlow.