Columbia University

Graduate School of Business

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

**Big Data for Finance**

**Business 9334**

**Spring 2019**

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

#### TBA

Instructors:

* Michael Johannes (coordinator)
* Kriste Krstovski (data/computer science portion)
* Finance faculty

Lectures: Fridays, 9:00-12:00.

Room: Warren 311

TA: Pu He and Daheng Yang

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

# Course Description

This course provides an introduction to financial data, data analysis tools and approaches, and analyzing statistical models using output from these datasets.

**Course Requirements**

The course requires basic knowledge of Python and SQL, although much of the data analysis and database tools will be taught. The course will have extensive weekly homework assignments.

**Problem Sets**

1. There are weekly problem sets during the semester, each will be graded on a scale of 0-10 points. The lowest one will be dropped from your grade.
2. Students are allowed to work in groups, but each student must do their own analysis and submit their own original work, which includes computer code. Programming is inherently an individualistic endeavor, and every student is required to do their own coding. It is, however, useful to discuss issues with others so we encourage students to work in groups with each other. Turning in another student’s code is a violation of the honor code.
3. Homework should be turned in at the beginning of class on the day it is due. Late homework is not be accepted under *any* circumstances. Since your lowest homework is dropped, you have a cushion to work with.

**Exams:** There will be a take home final based utilitizing the datasets from the classes and homeworks.

**Course Grade:** Homework: 60%, Final 40%.

# Recommended Text/Reference Books: TBA.

Outline:

1. Accessing, managing and representing data (Kriste, 3 lectures)
   1. Data and file formats (structured, unstructured, etc.)
   2. SQL and databases
   3. Text processing (parsing, tokenizing, stemming, etc.)
   4. Data representation (feature vector representation, etc.)
2. Financial datasets
   1. Lectures (2): Lira Mota: CRSP and Computstat
   2. Simona Abis (2): Mutual fund data
   3. Costis Maglaris (1): high frequency quote data.
   4. Harry Mamasyky (1-2): Thomson-Reuters and NLP
   5. Tomek Piskorski (1): Mortgage loan files
   6. Wei Jiang (1): TBA
   7. Xavier Giroud (1): TBA