# Third Homework Assignment (20% of grade)

**Submission Guidelines:**

Please follow carefully the instructions posted on the course’s github page when submitting your solutions (<https://github.com/MSIA/bigdatacourse/blob/master/README.md>). Failure to follow the instructions will result in lost points.

**Deliverables:**

1. Your spark source code file: name the file lastname\_x.py (scala,java)
2. Output files: name them lastname\_x.txt
3. Short write-up of findings as instructed below: name them lastname\_findings\_x.txt

Here ‘x’ is the number of the task.

You have to use spark. You can use scala, python, or java and you can use all libraries available in spark. You are not allowed to grab other code form the internet that is based on spark. (It is allowed to use python specific libraries such as nltk, scikit-learn, etc.)

**Note:**

All input files are available in /home/public/crime. Please copy them directly from /home/public/crime to HDFS and not to your home directory on wolf.

Due on Wednesday, May 24 at 5 pm.

## Crime in Chicago

Yes, Chicago has crime, and 6 million events since 2001. If we live in a wonderland, there would be no Spark homework assignment. But we don’t.

The Chicago crime data is available in /home/public/crime. The file has the header that explains many fields. Less obvious fields: block = the first 5 characters correspond to the block code and the rest specify the street location; IUCR = Illinois Uniform Crime Reporting code; X/Y coordinates = to visualize the data on a map, not needed in the assignment; District, Beat = police jurisdiction geographical partition; the region is partitioned in several districts; each district is partitioned in several beats; <http://gis.chicagopolice.org/pdfs/district_beat.pdf>; community areas and wards: <https://www.chicago.gov/city/en/depts/dgs/supp_info/citywide_maps.html>

Perform the following tasks.

1. By using SparkSQL, generate a histogram of average crime events by month. Find an explanation of results. (10 pts)
2. By using plain Spark (RDDs): (1) find the top 10 blocks in crime events in the last 3 years; (2) find the two beats that are adjacent with the highest correlation in the number of crime events (this will require you looking at the map to determine if the correlated beats are adjacent to each other) over the last 5 years (3) establish if the number of crime events is different between Mayors Daly and Emanuel at a granularity of your choice (not only at the city level). Find an explanation of results. (20 pts)
3. Predict the number of crime events in the next week at the beat level. Violent crime events represent a greater threat to the public and thus it is desirable that they are forecasted more accurately (IUCR codes available here: <https://data.cityofchicago.org/widgets/c7ck-438e>). (45 pts) You are encouraged to bring in additional data sets. (extra 10 pts if you mix the existing data with an exogenous data set) Report the performance of your models (accuracy, F1, MAPE, etc). You must use Spark dataframes and ML pipelines.
4. Find patterns of crimes with arrest with respect to time of the day, day of the week, and month. Use whatever method in spark you would like. (25 pts)

## Bias/Fairness Analysis

Watch the video demonstrating the use of chatGPT to identify protected attributes and to perform bias/fairness analysis. (Insert link for video here)

Using chatGPT apply a similar methodology to the Chicago crime dataset.

You must add at least 2 exogenous features to this dataset that will enhance the bias/fairness analysis.

**Data**

* + Did chatGPT capture the required attributes? If not, which attributes did it miss? Were any attributes incorrectly identified?
  + Add/remove protected (including proxy) attributes based on your evaluation of the outputs from chatGPT
  + Analyze the data. Do you detect bias in the data?
  + Present the results visually to show salient insights wrt to bias.
  + Based on the EDA and your project objective (predicting crime at the beat level), develop a hypothesis about where b/f issues could arise in the modeling.

**Modeling**

* + Based on your hypothesis, assess the fairness of your model/analysis by applying the fairness-related metrics that are available in any of the R and Python packages for fairness analysis.
  + Explain the reasoning for the groups that you selected for the fairness metrics.
  + Compare the fairness metrics for the different groups.
  + If you developed multiple models compare the fairness metrics for the models.
  + Comment on the results.
  + Suggest how the bias/fairness issues could be mitigated.
  + Present the results visually to show salient insights.

**chatGPT**

* + Document how you used chatGPT for your work.