Big Data A	
Project Name:	*Predicting Crime Rates using*
Team Members:	1) Carlos Petricioli (cpa253)
	2) Valerie Angulo (vaa238)
	3) Varsha Muralidharan (vm1370)

Task	Who
Identify data sources	***All***
Plan where data will reside	***All***
	T
Write code to ingest data source 1	Carlos
Write code to profile data source 1	Varsha
Write code to clean/format (ETL) data source 1	Valerie Angulo

NY	
Write code to ingest data source 2	Carlos
Write code to profile data source 2	Varsha
Write code to clean/format (ETL) data source 2	Valerie
We	
Write code to ingest data source 3	Carlos
Write code to profile data source 3	Varsha
Write code to clean/format (ETL) data source 3	Valerie
Design the analytic(s)	***All***
Code the analytic(s)	***All***
Test the analytic(s)	***All***

	Analyze results of analytic(s)	***All***
		All
	Iterate on the analytic	***All***
	Final analytic code due	***All***

nalytics Project Task

g Taxi rides data

Start Date	End Date
Data Planning Stage	
oct/23	oct/24
oct/24	oct/27
axi Rides Processing	
oct/27	oct/31
Nov 2	Nov 9
Nov 2	Nov 11

PD Crimes Processing	
oct/27	oct/31
Nov 2	Nov 9
Nov 2	Nov 11

eather data Processing	
oct/27	oct/31
Nov 2	Nov 9
Nov 2	Nov 11

n	
Oct 23	Nov 2
Nov 11	Nov 21
Nov 21	Nov 25

Nov 25	Dec 3
Nov 25	Dec 3
Dec 3	Dec 10
-	***15-Dec-17***

List

Comments
• *We will initialy have about 250GB of data* • *We will save it on Dumbo*
•In this step, you'll read the data from the source and write it or copy it into HDFS
•This is to characterize the data and the range of values in each column *•You might notice unexpected values in a column - you may decide to normalize the values (e.g. Street vs. St. vs. street) in the ETL stage* *•Find min, max, and averages* *•Find min and max length of text fields*

•In this step, you'll read the data from the source and write it or copy it into HDFS
•This is to characterize the data and the range of values in each column *•You might notice unexpected values in a column - you may decide to normalize the values (e.g. Street vs. St. vs. street) in the ETL stage* *•Find min, max, and averages* *•Find min and max length of text fields*

•In this step, you'll read the data from the source and write it or copy it into HDFS
•This is to characterize the data and the range of values in each column *•You might notice unexpected values in a column - you may decide to normalize the values (e.g. Street vs. St. vs. street) in the ETL stage* *•Find min, max, and averages* *•Find min and max length of text fields*

Based on our model
To try to relate crime, weather and taxi data
To see what patterns we get

•Are the results what you expected? *•Do you need to adjust the analytic(s)?*
•To improve results, and/or to better understand results