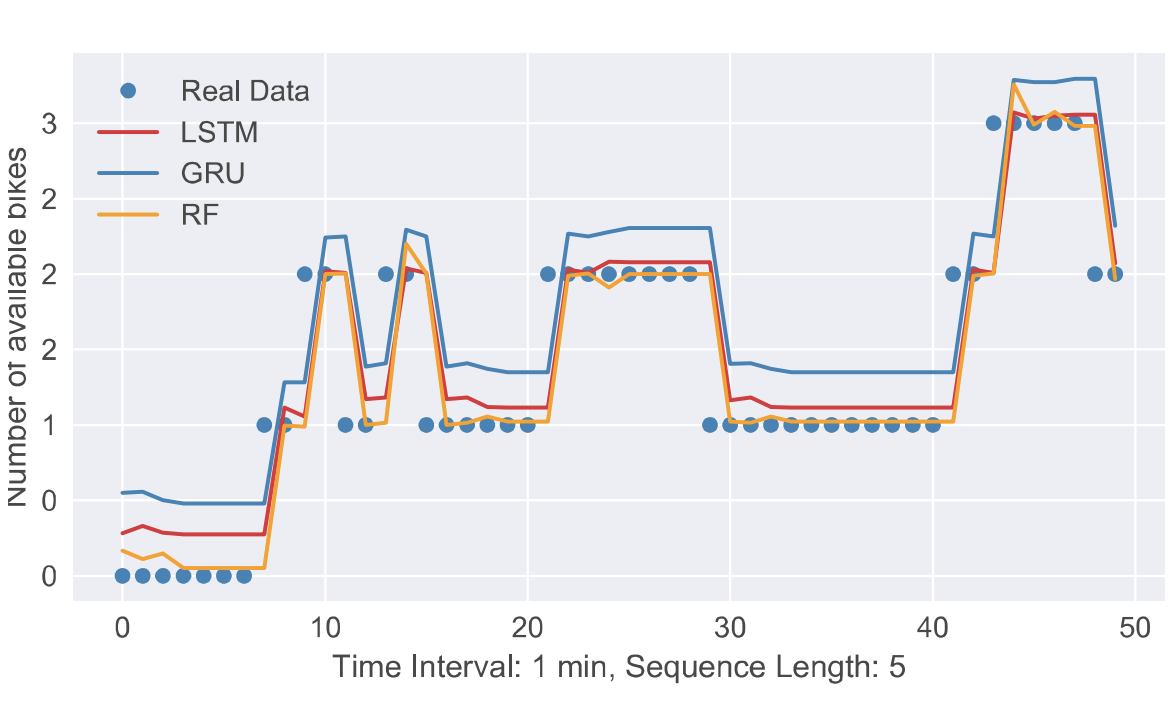
What our dataset are gonna be like:

| 30 minutes Bike activity change | Is\_not\_weekday | Time label (rush hour  or not) | Cluster label | Precipitation | Temperature | Availability |
| --- | --- | --- | --- | --- | --- | --- |
| 5/10 | Yes | 1 (morning rush hour) | 1 | 0.1 | 72 | Yes |
| 12/10 | No | 2 (night rush hour) | 3 | 0 | 50 | No |
| 2/1 | No | 0 (no rush hour) | 2 | 0 | 68 | Yes |

30 minutes

Bike activity change: we can figure out a way to quantify the change, maybe by (come in times/ come out times)



* This dataset will be used for both RF and Time Series Models.
* Split the dataset into 95% percent training and 5% testing
* Should we use available bikes or availability as the response variable?

Second dataset:

| Date  (not gonna use for model) | Time interval  (not gonna use for model) | Hour\_of\_day  (5) | 10 minutes Bike activity change | weekend | reshuffle（proportion） | Precipitation | Temperature | Availability |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2019/01/01 | 00:00AM | 0 | 5/10 | Yes | 1 | 0.1 | 72 | Yes |
| 2019/01/01 | 00:30AM | 0.5 | 12/10 | No | 0 | 0 | 50 | No |
| … | … |  |  |  |  |  |  |  |
| 2019/06/30 | 13:00PM | 13 | 2/1 | No | 1 | 0 | 68 | Yes |
| … | … |  |  |  |  |  |  |  |