**PROBLEM STATEMENT**

**Overview:**

In the context of telecommunications industry, one of the most important issues that industry faces is network congestion. It has been shown that congestion, even if for smaller durations, has a negative impact on customer loyalty, especially in price sensitive markets. To solve this problem effectively, it becomes imperative for firms to be able to predict congestion in advance and take proactive actions.

In this competition, you are required to train machine learning models that use cell tower statistics such as usage, customer count, etc, to predict the type of congestion that might occur.

We are providing a subset of original dataset, while also randomizing/masking some values to avoid leakage of proprietary information. Hence, this dataset only has sample data for December, 2018 transactions.

Some fields in the current dataset are anonymised to avoid data leakage; while the usage data has also been anonymously scaled/randomized by a single/constant factor.

**Data Collection Methodology overview:**

**ESR Records**

1. User level activity

**Incidents Table**

1. Tower level activity
2. Aggregated over 5 mins period
3. Inner joined

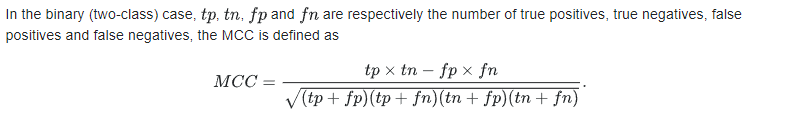
**Data Dictionary:**

You are provided data for cell towers with the following fields in training dataset:

1. cell\_name : Cell tower number/name – Masked name for cell towers
2. 4G\_rat: Tower supports 3G/4G indicator
3. Par\_year – Year under consideration (2018)
4. par\_month – Month under consideration (December)
5. par\_day – Day under consideration
6. par\_hour – Hour under consideration
7. par\_min – Minute bucket under consideration
   1. Buckets of 5 min interval. Eg: value of 15 implies statistics are compiled/aggregated over a time period from 10-15 mins
8. subscriber\_count: Count of total subscribers for the cell in the specified time period
9. Usage data: Data usage by activity type; includes both upload and download bytes
   1. web\_browsing\_total\_bytes
   2. video\_total\_bytes
   3. social\_ntwrking\_bytes
   4. cloud\_computing\_total\_bytes
   5. web\_security\_total\_bytes
   6. gaming\_total\_bytes
   7. health\_total\_bytes
   8. communication\_total\_bytes
   9. file\_sharing\_total\_bytes
   10. remote\_access\_total\_bytes
   11. photo\_sharing\_total\_bytes
   12. software\_dwnld\_total\_bytes
   13. marketplace\_total\_bytes
   14. storage\_services\_total\_bytes
   15. audio\_total\_bytes
   16. location\_services\_total\_bytes
   17. presence\_total\_bytes
   18. advertisement\_total\_bytes
   19. system\_total\_bytes
   20. voip\_total\_bytes
   21. speedtest\_total\_bytes
   22. email\_total\_bytes
   23. weather\_total\_bytes
   24. media\_total\_bytes
   25. mms\_total\_bytes
   26. others\_total\_bytes
10. beam\_direction: Tower beam direction
11. cell\_range: Cell tower range
12. tilt: Cell tower tilt
13. ran\_vendor: Service Vendor
14. **Congestion\_Type: Type of congestion observed (Target Variable)**

**Evaluation Criteria:**

Evaluation of outputs is based on Matthews correlation coefficient (<https://en.wikipedia.org/wiki/Matthews_correlation_coefficient>)



Scikit Learn: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html>

**Rules:**

1. Use of external data is not allowed
2. Using Cell name as a predictor variable is not allowed
3. Students are encouraged to use Python 3+ modules for all their work

**Useful Links:**

1. Backhaul: <https://en.wikipedia.org/wiki/Backhaul_(telecommunications)>
2. RAN: <https://en.wikipedia.org/wiki/Radio_access_network>