EY NEXTWAVE DATA SCIENCE CHALLENGE 2019 - MANUAL

To make sure you are **eligible to participate**, please check the Terms and Conditions since country/region conditions may apply. Also, you will find there all dates and details of the different phases of this challenge.

1. Context of the challenge

The EY NextWave Data Science Challenge 2019 focuses on how data can help the next smart city thrive, and boost the mobility of the future. Global urbanization is on the rise, with more than 50% of the world's population living in cities; according to the UN, that number will reach 60% by 2030 – that's nearly 1.5 billion more than in 2010.

While this trend creates great opportunities for cities, it also presents challenges to governments on how to upgrade infrastructure, alleviate congestion and address pollution. Electric and autonomous vehicles, along with the explosion of the ride sharing economy, are helping to address these challenges which also disrupt mobility and demand innovative solutions.

In parallel, public authorities have more information than ever on how citizens move around in the city. However, a gap exists between having this data and using it to improve the user travel experience for citizens. Forward-looking authorities have a chance to innovate infrastructure to make their city a better place to live in a better working world.

Here's your chance to narrow that gap. As a challenge participant, you will be able to download a dataset with a vast number of anonymous geolocation records from the US city of Atlanta (Georgia), during October 2018. Your task is to produce a model that helps authorities to understand the journeys of citizens while they move in the city throughout the day. If you dig deep enough, your work could inspire solutions that help city authorities anticipate disruptions, make real-time decisions, design new services, and reshape infrastructures in order that cities as smart as their citizens.

2. Teaming Up

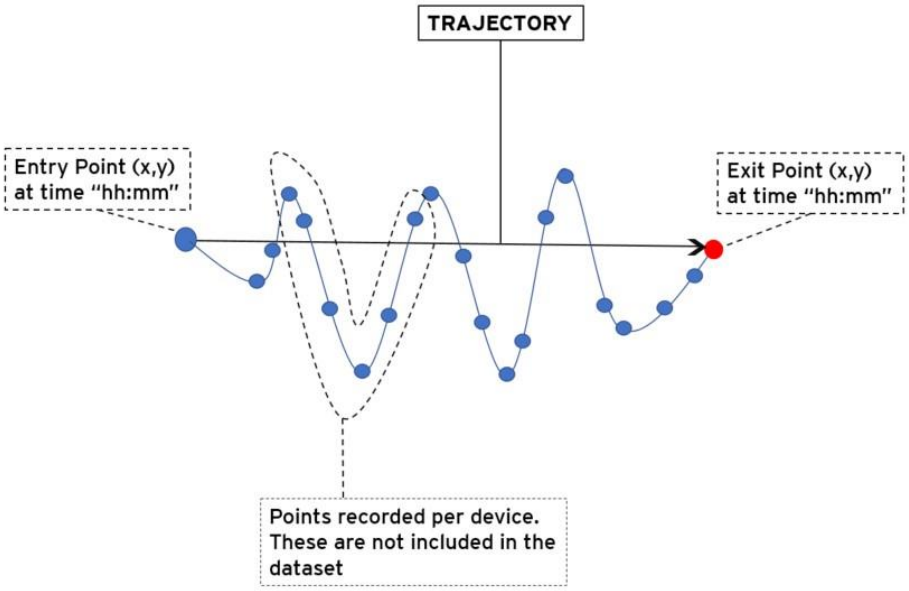
Once you have made the first submission, you will be able to see yourself in the ranking. When accessing the ranking, you will see all the participants that submitted any piece of work in the challenge. Next to them you will see a yellow button saying: "Request Team Up". Click on that button and you will be taken to a window where you will be able to request to that user the option to team up. The other member will then see the team request in the "Profile" tab on the left banner under "Request", where he/she will have to accept. Once the request is accepted you will become a team and both of your names will appear in the ranking.

3. Dataset Description

Students will have access to a data file data_train.csv that contains the anonymized geolocation data of multiple mobile devices in the City of Atlanta (US) for 11 working days in October 2018.

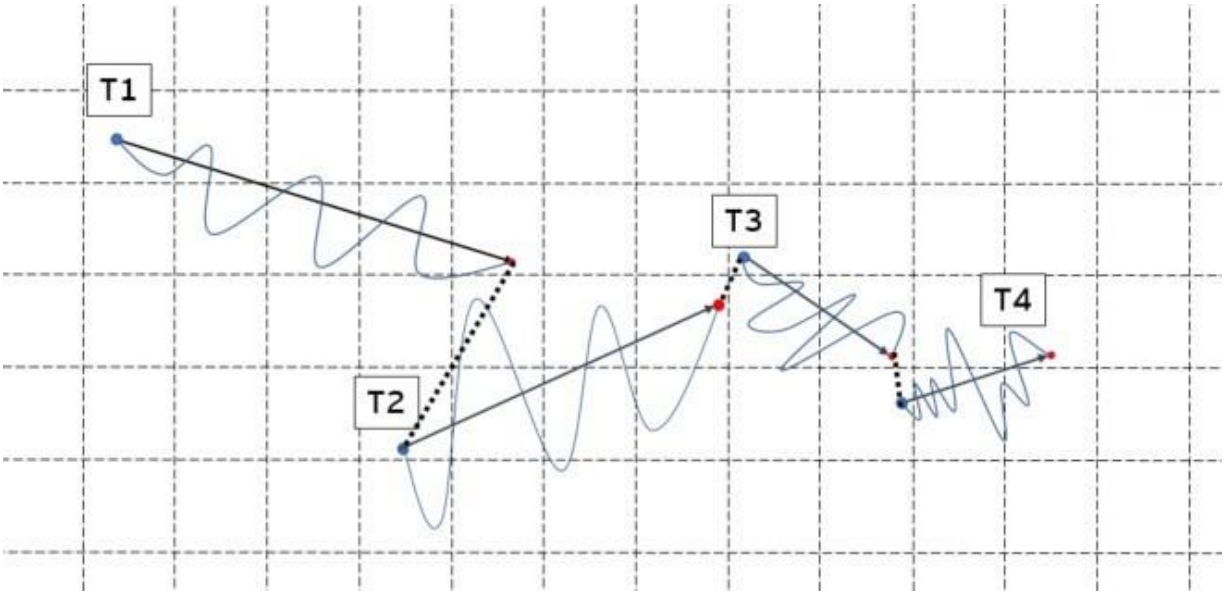
The devices' ID resets every 24 hours; therefore, you will not be able to trace the same device across different days. Therefore, every device ID represents a 1-day journey.

Each journey is formed by several trajectories. A trajectory is defined as the route of a moving person in a straight line with an entry and an exit point. See an example below of one trajectory from one of the devices:



As you can see, trajectories are a simplification of the real path of a person.
A trajectory ends when a person stops moving and stays in the same place for a while and when the device stops recording for some time.

For each device you will get multiple trajectories. The set of all trajectories of a device represents a simplification of the journey of one person for 24 hours. The graphic below shows a full journey of a device.



It is important to note that each device has a different number of trajectories

Trajectories are separated. In the graph, this separation is shown as a dotted line between the exit point of a trajectory and the entry point of the next one. These dotted lines represent blind parts of the journey where the device did not record the location.

4. Dataset Details

There are approximately 210,000 devices and 11 columns in the database. You will receive these records separated into two datasets to download:

- A train dataset (data_train.csv)
- A test dataset (data_test.csv)

The train dataset contains 80% of the records, while the test dataset contains 20%. The test dataset will then be split into public and private datasets.

The variables in the dataset are as follows:

Variable name	Type	Description
hash	String	Represents the unique identifier of a device
trajectory_id	String	Represents the unique identifier of a trajectory associated to a device
time_entry*	Date	Indicates the local time for the starting point of the trajectory (HH:mm:ss)
time_exit*	Date	Indicates the local time for the ending point of the trajectory (HH:mm:ss)
Vmax	Integer	Represents the maximum velocity registered in the course of a trajectory.
Vmin	Integer	Represents the minimum velocity registered in the course of a trajectory.
Vmean	Integer	Represents the average velocity registered in the course of a trajectory.
x_entry	Double	Entry x coordinate (cartesian projected position)
y_entry	Double	Entry y coordinate (cartesian projected position)
x_exit	Double	Exit x coordinate (cartesian projected position)
y_exit	Double	Exit y coordinate (cartesian projected position)

*All the data related to time is shown in Atlanta's local time (Eastern Time).

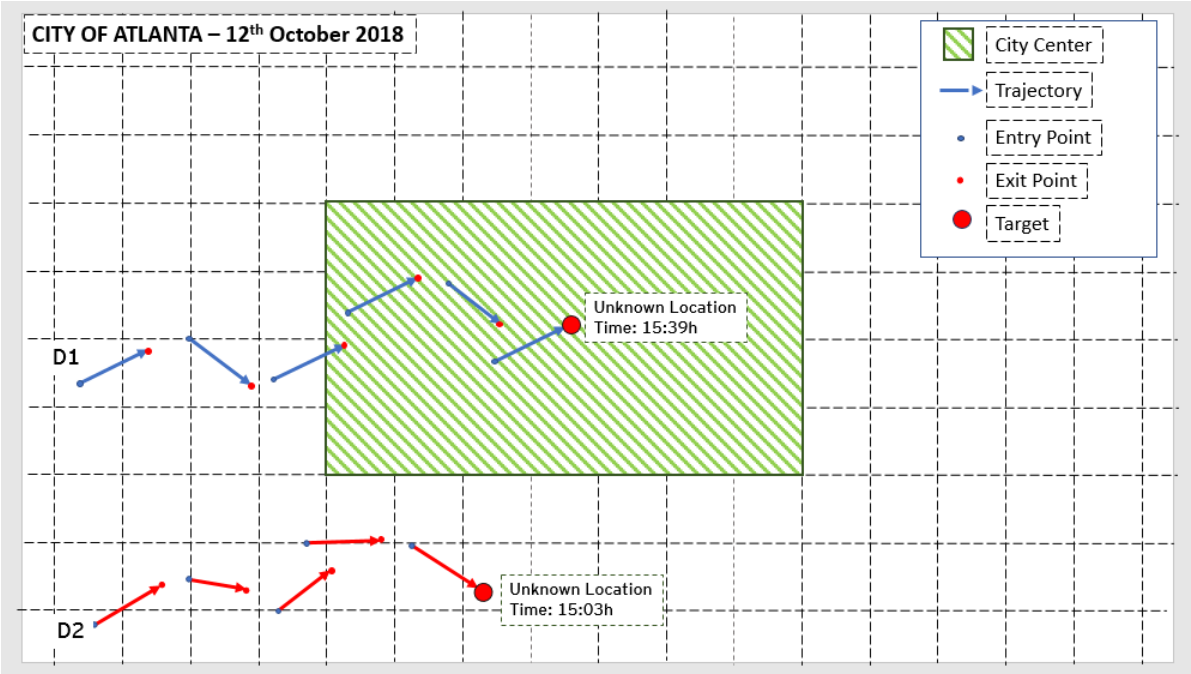
5. The challenge

You must predict how many people are in the city center between 15:00 and 16:00.

The test dataset contains a number of devices where the trajectories after 15:00 have been removed. All but one: After 15:00, you will find one last trajectory, with (1) entry location, (2) entry time and an exit time that is between 15:00 and 16:00. But the exit point has been removed.

Your task is to predict the location of this last exit point and whether this device is within the city center or not. The target variable is the latter.

See the graphic example below.



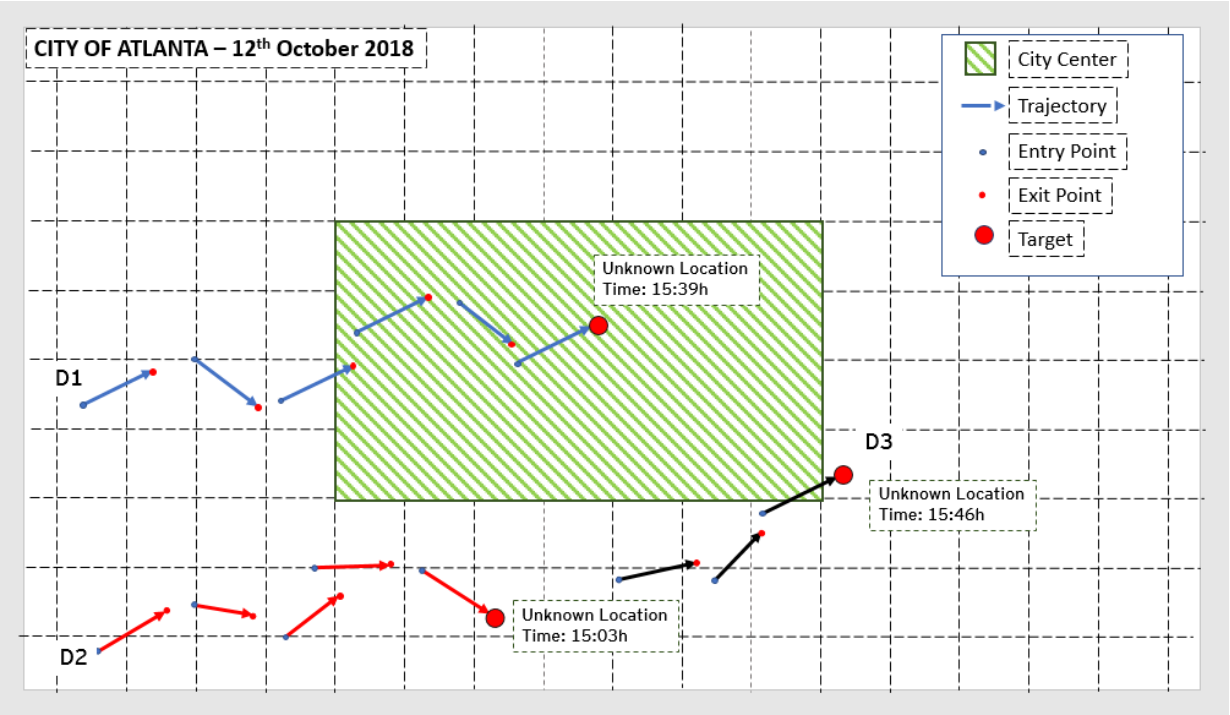
After you estimate the position of each target, you will have to classify that point based on whether it is located inside the city center or not. To do so, you will have to implement a rule that outlies the limits of the city center of Atlanta (decimal point "."):

$$3750901.5068 \leq x \leq 3770901.5068$$

$$-19268905.6133 \leq y \leq -19208905.6133$$

You will need to classify each of the exit points whether they are within (1) or outside (0) the limits of the city center. For example, Device 1 in the graph is in the city center between 15:00 and 16:00, therefore Device 1 will be classified with a "1" while Device 2, which is not in the city center in that timestamp will be classified with a "0".

Some trajectories may "cross" the city center, but their exit point will be outside the city center. See the example of Device 3 (D3) in the graph below. For the sake of simplicity, these trajectories are considered outside the center, given that we only consider if the exit point is within boundaries or not.



6. Submission

After classifying each of the targets, you will have to submit your results in the following format: trajectory_id ; city_center

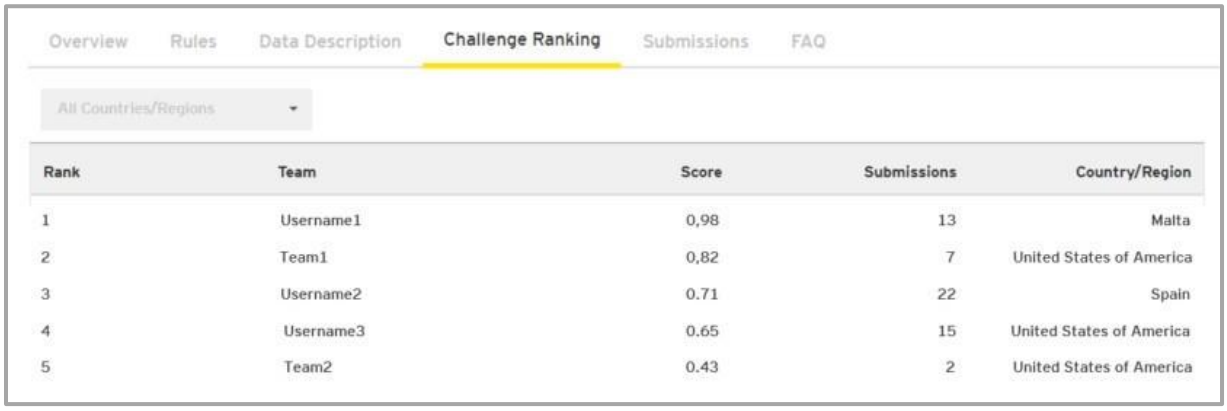
The trajectory id identifies the last trajectory of a device and the city center identifies the location of that point. Here's an example of how a submission would look:

trajectory_id	city_center
123df5	1
345rgf	0
678lsp	0
910dcw	1

Trajectory "123df5" ends in the center, while trajectory 345rgf does not.

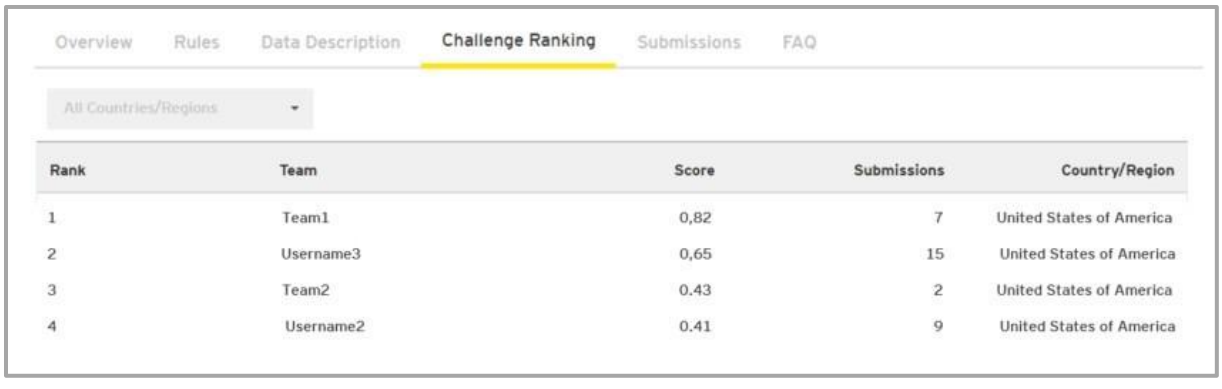
Submissions are evaluated using the F1-score between the predicted and the observed target. Your results will be compared to the real data both using public and private datasets. In the challenge ranking you will be able to see the score based on the public dataset only. Your score will be between 0 and 1, 1 meaning you got all the values correct and 0 meaning you didn't get any.

This an example of how the global (All countries) ranking would look:



Rank	Team	Score	Submissions	Country/Region
1	Username1	0,98	13	Malta
2	Team1	0,82	7	United States of America
3	Username2	0.71	22	Spain
4	Username3	0.65	15	United States of America
5	Team2	0.43	2	United States of America

And this is an example of how the "United States of America" ranking would look:



Rank	Team	Score	Submissions	Country/Region
1	Team1	0,82	7	United States of America
2	Username3	0,65	15	United States of America
3	Team2	0.43	2	United States of America
4	Username2	0.41	9	United States of America

7. Suggestions

- Remember you can participate as a team of up to two persons.
- Be mindful that there is a time component in the dataset. When proceeding with the analysis, data may need to be aggregated or grouped to be successfully processed.
- Be rigorous during the data cleaning process.
- Do not forget to consider the presence of outliers.
- It would be advisable to work on cloud solutions for better performance.
- External datasets can be used to complement the analysis. Read the terms and conditions for more clarification on this.

8. Important competition dates

- Competition starts on 1 April 00:00 (CET) and ends on 10 May 2019, 23:59 (CET).
- The global and local rankings will be available on the 11 May onwards.
- Before 14 May, EY will announce the country / region winners.
- The global award ceremony is 14 June 2019 in New York City, New York.

Please read in detail Terms & Conditions to understand all dates and details of the different phases of the competition.

9. Country / region finals

If you are among the top performers in your country / region on 10 May, you will be invited by the local EY firm to take part in the country / region final. Check Terms & Conditions for more details.

Finalists will present their findings to a group of judges to compete to become the country / region winner.

Before participating in the Finals, please take the following points into consideration to ensure you are prepared:

1) Get prepared before the competition ends

Check the national (your country's / region's) ranking periodically. If you're at the top or among the leaders who might be invited to the finals, get ready for the next stage. Do not wait until the last minute or the final confirmation to make sense of your results. During the competition take time to identify non-obvious patterns, consider alternative approaches, and think of potential opportunities of how cities could use geolocation information.

2) Be prepared to demonstrate your eligibility and support your findings

If you are shortlisted for the country / regional finals, you will be required to send information that demonstrates your eligibility to participate in the challenge.

Also, you will need to send information supporting your work. See Terms and Conditions for full details.

3) What to expect during country / regional finals?

Data science goes beyond deciding what model or calculation to use. Good data scientists make sense of their findings, think of real-world applications, and articulate their ideas to colleagues. For that reason, all country / regional finalists are required to present their analysis to local EY leadership.

The format of the country / region final will consist of selected participants presenting their work to a panel of EY judges. Requirements for the presentations will be issued to finalists in advance.

Each judge will use a standard scorecard to assess a finalist's performance separately. The assessment will include: methodologies and algorithms used, depth of insights provided, use of external information, ability to communicate (including level of English), and quality of the presentation.

See Terms and Conditions for full details.

10. Requesting help

In case you need help using the platform, please visit the help section here: https://datascience.ey.com/help

In case you have questions regarding the challenge, we suggest you take a look at the FAQ section. There we will be uploading most common questions from participants.

For further help, please write us at eydatasciencechallenge@ey.com .