# 

# The challenge

In this datathon, you will work with historical data obtained from a **fitness application** that provides tailored training plans for runners, based on their fitness levels, goals, and schedules. The data contains detailed information about the workouts. The goal of this challenge is to **analyze** the data and **extract insights** on how users are progressing with their training and what type of workouts they perform.

The data

You will receive all sport activities from a random set of highly active users over the course of 2019. In total, this means over 16 thousand training sessions from 100 users.

**Table *trainings (3 MB, ~16k rows)***

This table contains high-level metadata for each training session of each user. The schema is:

| **Column name** | **Description** | **Unit** | **Data Type** |
| --- | --- | --- | --- |
| user\_id | Internal technical id that identifies a single user | - | str |
| training\_id | Internal technical ID that identifies a single sport activity | - | str |
| start\_date | Time of start of the activity | - | timestamp |
| distance | Total distance covered in the activity | m | integer |
| uphill | Total climb in the activity | m | float |
| downhill | Total descent in the activity | m | float |
| duration | Duration of the activity | s | integer |
| temperature | Ambient temperature recorded during the activity | ºC | float |
| avg\_hr | Average heart rate during the activity | bpm | integer |
| max\_hr | Max heart rate during the activity | bpm | integer |
| power | Average power during the activity | W | integer |
| type | Type of activity (see below) | - | string |

**Table *run\_logs (5.3 GB, ~30M rows)***

This contains detailed logs for every training session in the table above. The schema is:

| **Column name** | **Description** | **Unit** | **Data Type** |
| --- | --- | --- | --- |
| user\_id | Internal technical id that identifies a single user | - | str |
| training\_id | Internal technical ID that identifies a single sport activity | - | str |
| time | The time at which a device recorded GPS and other data | - | timestamp |
| lat | Recorded latitude at point of time | º | float |
| lng | Recorded longitude at point of time | º | float |
| hr | Recorded heart rate at point of time | bpm | float |
| ele | Recorded elevation at point of time | m | float |

**Notes**

* You can join both tables on *user\_id* and *training\_id*.
* Other attributes may not align across both tables.
  + Table *run\_logs* omits any point within a radius of the start or end point, see slides.
* *run\_logs* is over 5GB. Think twice before running slow algorithms on the whole table.
* Attributes can be null and records may contain data quality issues.
* Users might sync activities other than running into the app.
* You may want to consider grade-adjusted pace (GAP) while developing metrics.
* You can complement the data with any publicly available, non-paywalled dataset.

Your Tasks

We will evaluate you on your success at **three tasks**. We have numbered them as 1, 2 and 3 below; this is for clarity, and we suggest tackling them in parallel.

You will be working on **real world, only lightly curated** data. This data will require some purging and transforming; this is your **first task**. You will have to explain what pre-processing you did and why.

In the app, each training is tagged as one of 5 **types**. These are **proposed by the app**, and thus not always reliable – the user may or may not have followed the recommendation. The types are:

* LONG JOG: A long jog is a type of endurance training where the runner maintains a moderate pace for an extended distance or duration.
* INTERVAL: Interval training involves alternating periods of high-intensity exercise with periods of lower-intensity recovery or rest.
* STEADY JOG: A steady jog is a type of run that maintains a consistent, comfortable, medium pace throughout the entire session.
* LOW INTENSITY: Low-intensity runs/walks are slower-paced workouts, focusing on maintaining a comfortable and relaxed pace.
* RACE: competitive running event.

Your **second task** is to create an algorithm that, given the log and metadata of a run, **predicts the type** of training among the five above.   
Out of the ~16k trainings, 250 have column *type* empty. You will have to provide your algorithm’s prediction of the missing value for the 250. For convenience, we provide the table *exam\_dataset*, containing these 250 pairs of *user\_id* and *training\_id* isolated, and a blank column for you to fill in your predicted category.

Finally, the users of the app are interested in **personalized metrics that track their fitness over time**. Your **third task** is to design and prototype such a metric. This should be a time-dependent value, which allows users to compare their current to past performance.

Other than this, you have complete creative control. You are free to enrich the data with any other publicly available data source. We will evaluate your algorithm on the grounds of originality, intuitiveness and sophistication.

At the end of the datathon, your team will present your findings and proposals, and there will be a round of questions. In this 5 min. presentation, you must address the following points:

1. **Data pre-processing (~1 min.)**: Explain what transformations you have applied to the input dataset. You do not have to explain every single one in detail, but we encourage you to present 1-2 interesting problems and how you addressed them. Deeper observations and clever workarounds will be rewarded!
2. **Prediction (~1 min.)**: Present your solution for the labeling of sport sessions.
3. **Design (~3 min.)**: Present your designed fitness metric, and some results on the data.

# Submission

Please submit your code and prediction by Sunday 7th of May 12:00h. This should comprise:

* A compressed version of your code (for all three tasks)
* A .csv file, prefixed by your team name and ‘\_submission.csv’ ; i.e ‘**team1\_submission.csv**’. It should be in the same format as *exam\_dataset.csv*, with column *type* filled in for each of the 250 rows with one of “LONG JOG”, “INTERVAL”, “STEADY JOG”, “LOW INTENSITY” or “RACE” (without the quotation marks).   
  If the submission does not match exam\_dataset.csv exactly on non-missing values, your score will be 0 for task 2. The same goes for any other attempt of “cheating” the scoring function. Order of rows after the header is not important.

Please submit your slides by Sunday 7th of May 14:00h.

# How to access the data

The link to the challenge’s Google Drive is on the Notion page. If you are having problems accessing the data with your provided email, please share with us your @gmail.com email account over Slack.

We offer you two options to access the datasets. The first option (Google Drive) is best suited if you will develop and run on your own machines. The second option (Google Colab) is a better option if you would rather develop and run on the cloud.

### Option 1: manually through Google Drive

In the Datathon Google Drive, you can find the three datasets (*trainings, run\_logs* and *exam\_dataset)* as .csv files, which you can download directly.

For your convenience, we provide you with a folder in Google Drive called *user\_run\_logs*. This contains a copy of dataset *run\_logs*, partitioned by *user\_id* into 100 files. That is, logs from *run\_logs* with *user\_id=a1b2c3* can also be found in *user\_run\_logs/a1b2c3\_run\_logs.csv*. This is for easier handling and provided as-is; in case of inconsistencies *run\_logs* is the correct file.

### Option 2: directly query on Google Colab

To work with Google Colab (a free Python notebook environment from Google), you can simply log in with your Google account to <https://colab.research.google.com/>.

You don’t need to manually upload the datasets to Colab; they can be queried directly from Google’s BigQuery where they are hosted. To access the data in a Colab notebook, please follow these steps:

1. Download files *service-account-key.json* and Jupyter notebook *colab\_load\_data.ipynb* from Google Drive.
2. Go to Colab (link above) and log in with your Google account
3. Upload notebook *colab\_load\_data.ipynb*
4. On the left tab for files, upload *service-account-key.json*.
5. Run the cells in the notebook as provided.

Please keep in mind that the fourth cell in the notebook queries and downloads a 5GB dataset and will take at least a minute – we ask you to run this responsibly and as little as possible, for everyone’s benefit.

# Evaluation

For your team to be evaluated fully, you will need to address the following points with their corresponding weight:

* Data pre-processing (task 1): 10 points.
* Prediction of training session type (task 2): 50 points.
* Design of a fitness metric (task 3): 40 points.

Tasks 1 and 3 will be graded based on your presentation, and if possible, inspection of your code.

Task 2 will be scored on percent of correctly labeled training sessions with respect to the reference dataset, and graded on a curve. In task 2, a bonus of up to 20 points may be directly added to your score for exceptional ideas; total may not exceed 50 points.

For the three tasks, more intuitive, explainable, business driven solutions are preferred over brute force computational models.

Good luck!