Clue-WATTx Data Hackathon

We are happy you are here with us. Your time and energy is appreciated. We hope you will learn, get to know interesting data people and have fun exploring.

Task

Clue is a mobile app to track the menstrual cycle and to learn about female health. Millions of people tell the app the days of their period and when they experience cycle related symptoms. The challenge for this hackathon is to find patterns in these symptoms and predict them for the upcoming cycle.

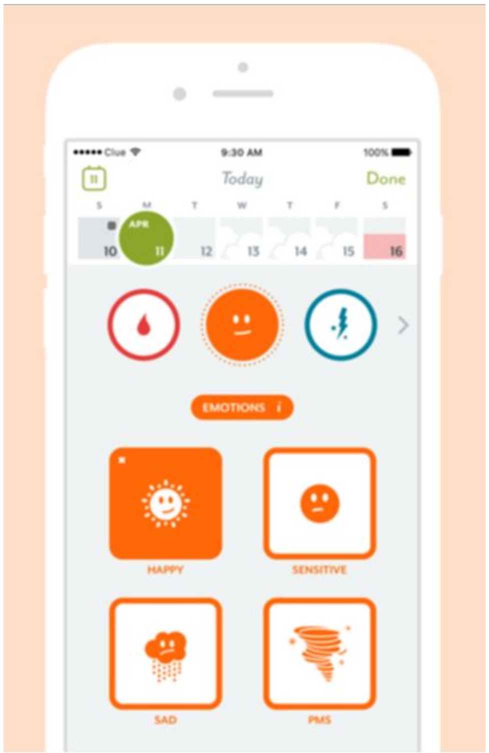
Data

The data comes in 3 files. The **cycles.csv** file contains one row per cycle of a user. By convention a cycle starts with the first day of the period. It contains the following columns:

- **user_id -** String identifying a user
- **cycle_id** - Integer enumerating a given user's cycles. Recent cycle have a lower number
- **cycle_start -** Start date of the cycle
- **cycle_length -** Number of days the cycle contains. Including the first day of the period and including the last day before the next period
- **period_length -** Number of days the period lasted
- **expected_cycle_length -** Cycle length prediction using a median over past cycle lengths

The **tracking.csv** file contains one row per symptom a user enters. On the right you can see how this looks in Clue. The app groups every 4 symptoms together in one category. The file has the following columns:

- **user_id -** String identifying a user
- **cycle_id -** ID of the cycle that contains this symptom
- **day_in_cycle -** The day of cycle the symptoms occurred counted from the start of the cycle starting with 1
- **category -** Group of 4 symptoms
- **symptom**



The tracking.csv contains data for all symptoms tracked in Clue. You will only need to predict symptoms for the 4 categories **Energy, Emotions, Pain and Skin**. We have excluded users who are using hormonal birth control or have extremely long or short cycles.

The **active_days.csv** file contains one row per day in a cycle. Only the days a user has been active (i.e. tracked a symptom) are logged here. The file contains the following columns:

- **user_id -** String identifying a user
- **cycle_id** - Integer enumerating user's cycles
- **day_in_cycle -** The day of cycle the user has been active; counted from the beginning of the cycle, 1-indexed
- **date** - The date that day_in_cycle corresponds to

Evaluation

The data you are provided with excludes the last cycle of each user. Your predictions will be scored on these last cycles. Assume you know the first day of this cycle. You need to predict the symptoms from the second day of the cycle until the expected end of that cycle.

Your predictions will be scored using the **log loss**.

$$log\ loss = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(p_i) + (1-y_i)\log(1-p_i)\right)$$

The y's are the true symptom occurrences and the p's are the assigned probabilities.

The prediction should have the following columns:

- **user_id**
- **day_in_cycle**
- **symptom**
- **predicted probability of symptom**

Setup

The hackathon will use WATTx' new data competition platform to enable you to work with Clue's very personal data while maintaining the privacy of Clue's users.

You get a local copy of a synthetic dataset, that has the same statistical properties than Clue's real dataset, without exposing any real users' data. This dataset will allow you to quickly iterate on your model.
You will also be able to train your model on the real training data by submitting it in a docker container to a server that will run and evaluate it.