Modeling, Prediction, Recommendation from Large-Scale Fitness Data

Project 4

Exercise Freak Consulting, LLC

The Exercise Freak Team

David Doerner

Chief Analytics Officer

Jason Gilberg

Chief Business Development Officer

Patrick Mulrooney

Chief Data Architect

Masashi Omori

Chief Financial Officer

Advisor

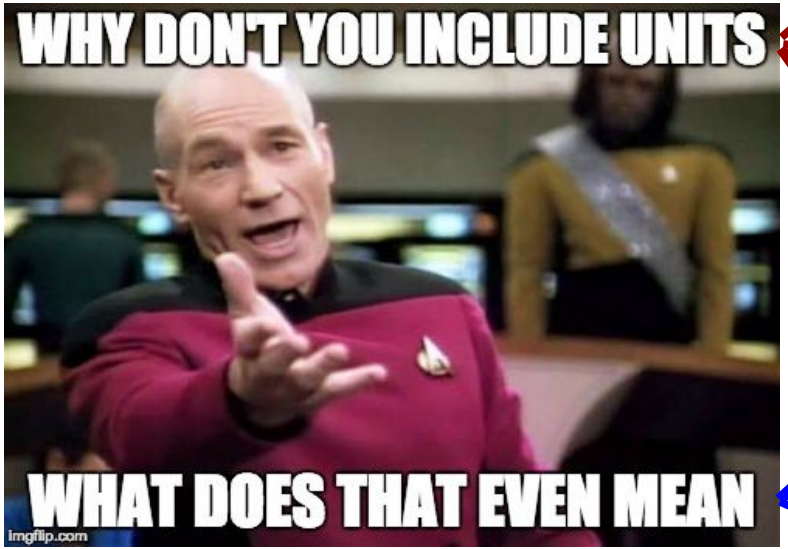
Prof. Julian McAuley

The Challenge

Given a user with certain features, goals, and workout history, recommend some workouts with predictions for their performance.

- Can these recommendations be enhanced by utilizing other users' data?
- Can we update our recommendations and predictions as users perform more workouts?

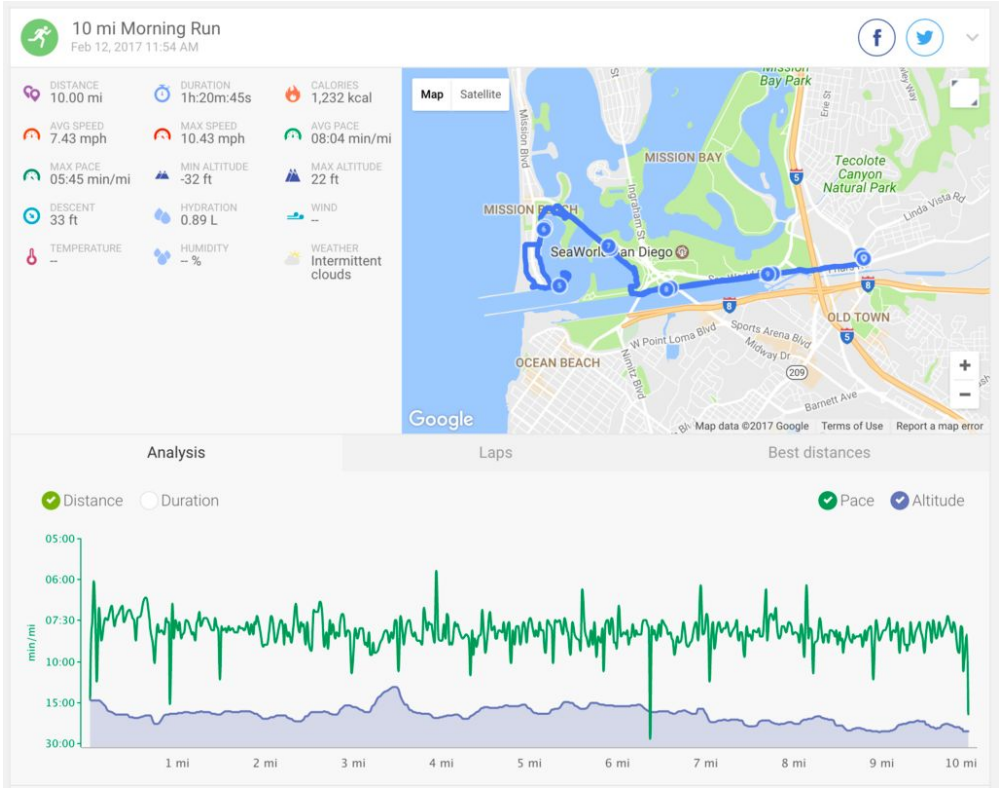
Endomondo: The Data




```
"altitude_max": 92.0,
"altitude_min": -3.0,
"ascent": 353.0,
"author": {
    "expand": "abs",
    "first_name": "Jussi",
    "gender": 0,
    "id": 10921915,
    "is_premium": false,
    "last_name": "Lattu",
    "middle_name": "",
    "name": "Jussi Lattu",
    "picture": { "url": "https://www.endomondo.com/..
},
"calories": 978.0,
"can_copy": true,
"descent": 348.0,
"distance": 25.035982131958008,
"duration": 5359.0,
"expand": "full",
"feed_id": 281475496560571,
"hydration": 0.810823,
"id": 546556161,
"local_start_time": "2015-06-21T13:24:58.000+03:00",
"pb_count": 0,
"points": {
    "expand": "full",
    "id": 2199569811713,
    "points": [
        {
            "distance": 0.0,
            "duration": 0.0,
            "instruction": 2,
            "latitude": 60.19968,
            "longitude": 24.653068,
            "sensor_data": {},
            "time": "2015-06-21T10:24:58.000Z"
        },
    ]
},
"speed_avg": 16.818349631470205,
"speed_max": 45.5241,
"sport": 2,
"start_time": "2015-06-21T10:24:58.000Z",
"weather": { "type": 3 }
```

Endomondo: The App

- The JSON data in a readable format.
- What is Hydration? There's missing weather data. Still no units.
- **Go For a Run!**
- We have our units by mapping the output with units to the raw JSON data.

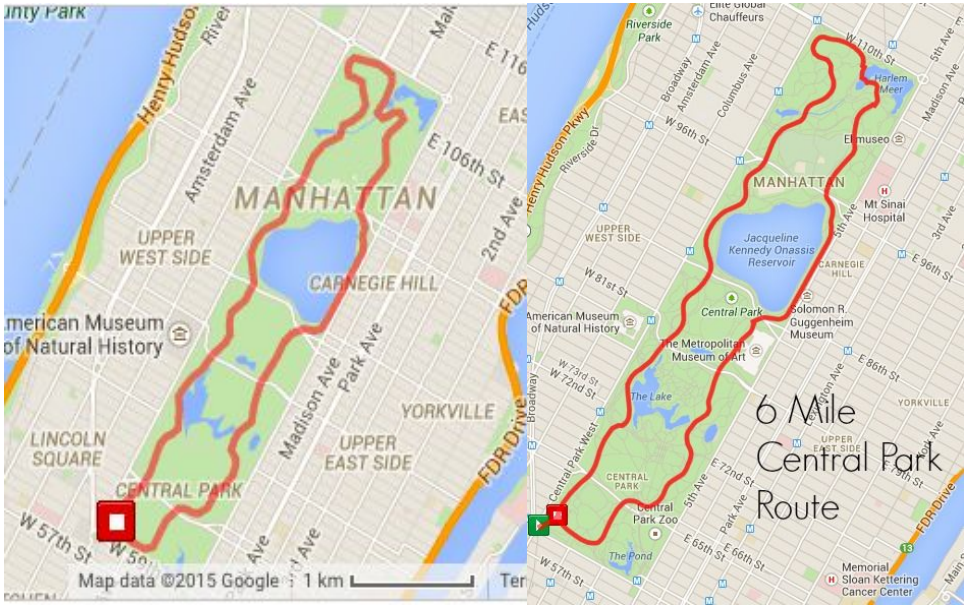


AWS/Postgres

- Configured Postgres on EC2 instance
- Edited config to allow PySpark/PGAdmin to connect to EC2
- Reviewing PySpark's machine learning API
- Short term goal: Create a sample clustering with a subset of the data.

Clustering Techniques

Clustered Routes



Cluster the Routes based on Features:

- Total Distance
- Path characteristics
- Changes in Altitude
  - Net Change
  - Gross Change

Cluster the Workouts on Features:

- Pace
- Route Features
- Metadata
  - Mostly temporal

Why: To group users with similar routes and performance for prediction

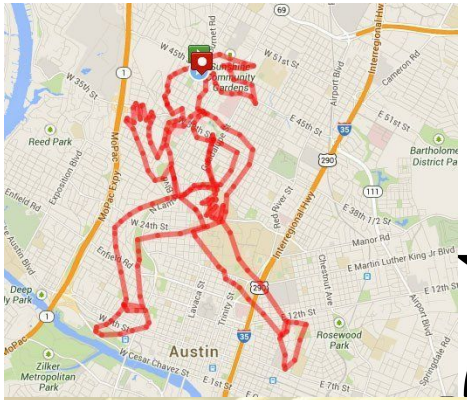
Regression

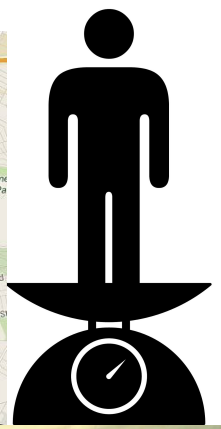
Factors:

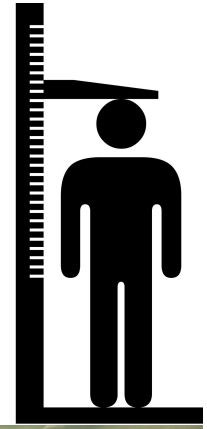
- Route Information
  - Elevation, Distance, Weather
- User Data
  - Biomeasurements
  - Historical Totals

Targets:

- Performance Data
  - Average Pace
  - Instantaneous Pace







Why: To predict performance better than using simple or moving averages

Improvements with Dimensionality Reduction

What does it accomplish:

- Reduce granularity while preserving information
- Reduces calculation complexity for Clustering and Regression

Why is it important:

- To increase performance of prediction and recommendation system
  - Allow updates to recommendations and predictions
  - Dynamic system more relevant than static system

Summary of ML Steps

1. Cluster Routes for recommending similar routes
2. Cluster Workouts for prediction
    a. Frame a regression problem (improves predictions)
    b. Define a user's probability of different types of workouts
3. Predict performance within workout clusters using regression
4. Aggregate user's performance predictions based on probability

Deliverable

Create a set of 10 recommended routes per user

Provide performance predictions for the 10 recommended routes

Update performance predictions and route recommendations given new data

Live performance visualization against prediction

Additional Utility

Applying this analysis platform to other problems to create value

Busting neon collar criminals

FITBILK —

Marathon runner's tracked data exposes phony time, cover-up attempt

A cut corner, a retraced route on a bike, and the Garmin tracker that exposed the lies.

Previous run

Winning run

- In mid Feb a woman in Florida was caught in an elaborate lie thanks to fitness tracking data.
- After she placed second in a half marathon race detectives (it is a thing) uncovered her elaborate cover up.
- The significant difference in previous workout data versus her winning run was the primary source of suspicion.
- What else could we track?

Workplace Initiatives

- Are they moving? What path are they taking? Are they straining?
  - Warehouse employees
  - Construction workers
  - Nurses
- Workplace wellness.
  - Get moving in the workplace
  - Get moving out of the workplace
  - Healthier (and happier?) employees
  - Lower premiums
  - Fewer sick days