Modeling, Prediction, Recommendation from Large-Scale Fitness Data

Project 4

David Doerner
Jason Gilberg
Patrick Mulrooney
Masa Omori

The Team

David Doerner

Predictive Modeling

Feature Engineering

Jason Gilberg

Predictive Modeling

Presentations

Patrick Mulrooney

Data Management

Data Visualization

Masashi Omori

Parallel Computing

Treasurer

Advisor

Prof. Julian McAuley

The Challenge

- Create a predictive model for speed or heart rate in a particular workout
- Generate a route or pace recommendation system to increase efficiency of users' end goals
- Visualize route and historical data to challenge users and encourage growth

Question Formulation

Data Understanding

What impact does altitude have on speed and heart rate?

What are experienced users better at than beginners?

Fitness Performance

Are there ways to increase fitness levels more efficiently (general activity, casual users)?

Can there be more specific training to reach performance goals (times, races)?

Implementation

Can we provide recommendations to enable the users to reach their goals faster?

How can we effectively visualize the information to encourage the users?

Datasets

Selection

- Endomondo Fitness Application
  - User data
  - HR + GPS data
  - GPS-only data

Collection

- Investigating the inclusion of Weather and Air Quality Data
- Map Data from Google Maps for terrain/elevation data

Management

- InfluxDB
- S3
- Scalability
  - Spark

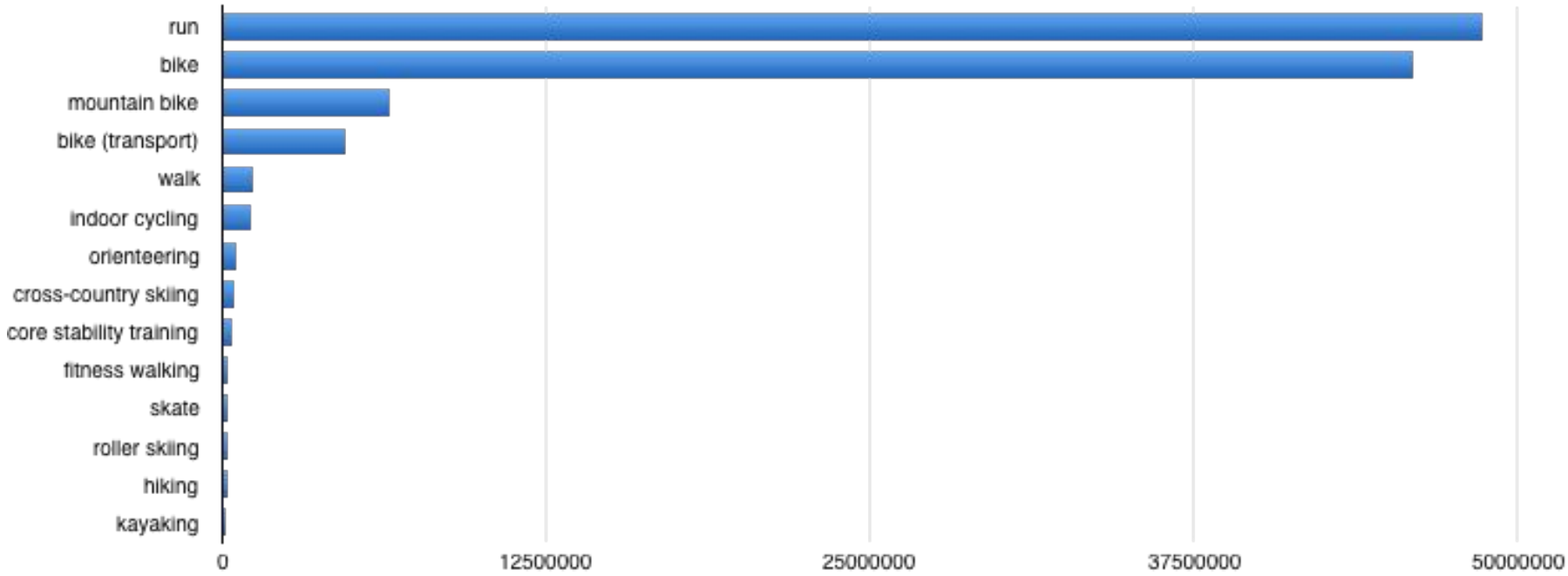
Data Preparation

- Data Parsing
  - Invalid Json
  - Used Perl/Python to format properly
- Data Inaccuracies
  - User input data
  - Inaccuracies in collection from the hardware

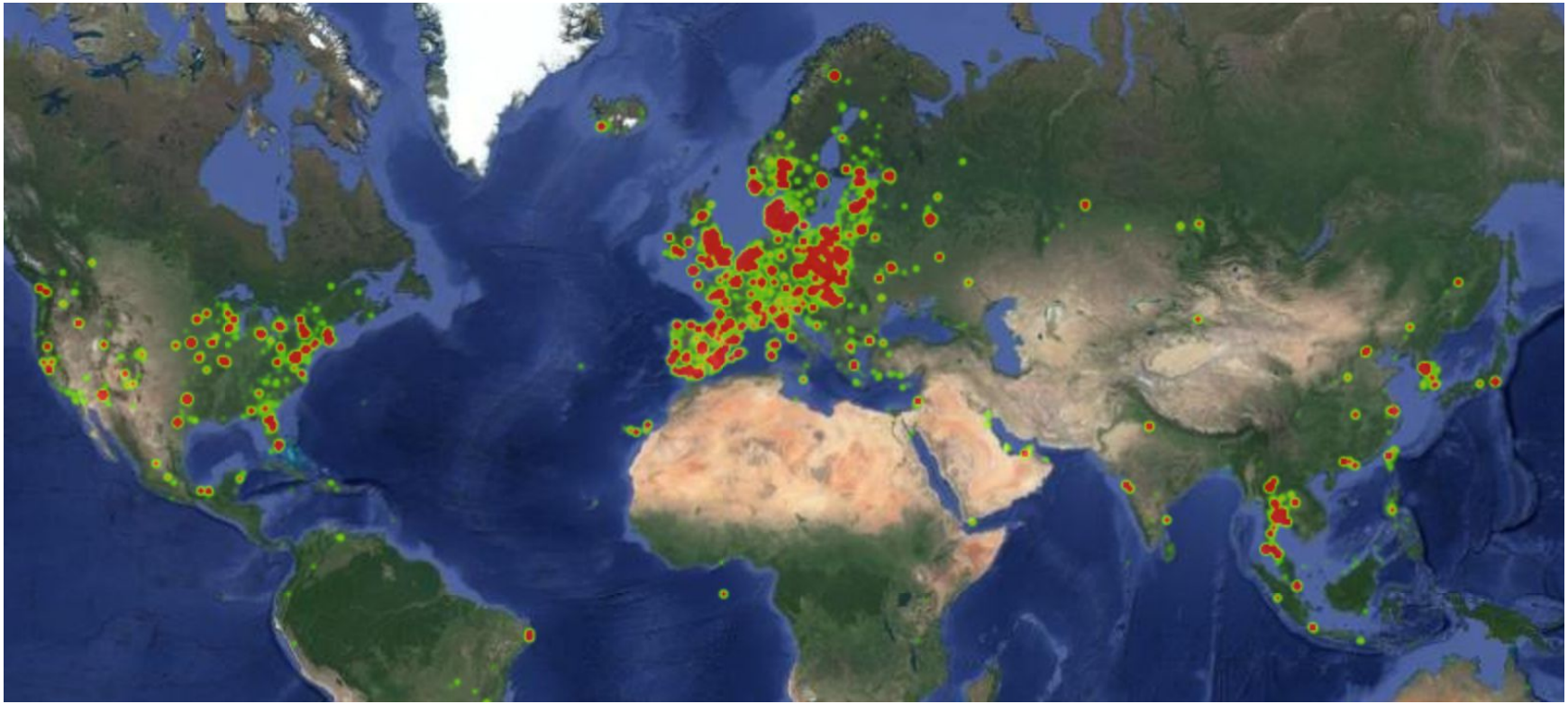
Why InfluxDB?

- Time Series Oriented Data
- SQL-like Querying Language
  - Regex Support
- Built-in Libraries for Python
  - Communicates using HTTP
- Visualization Tools
- Ease of Use
  - 100M+ Records on Laptop
  - Count of 50M records in 26s
- Measurements, tags, & values
- FREE!
  - Unless Clustering Needed

Exploratory Analysis: Time series data point count

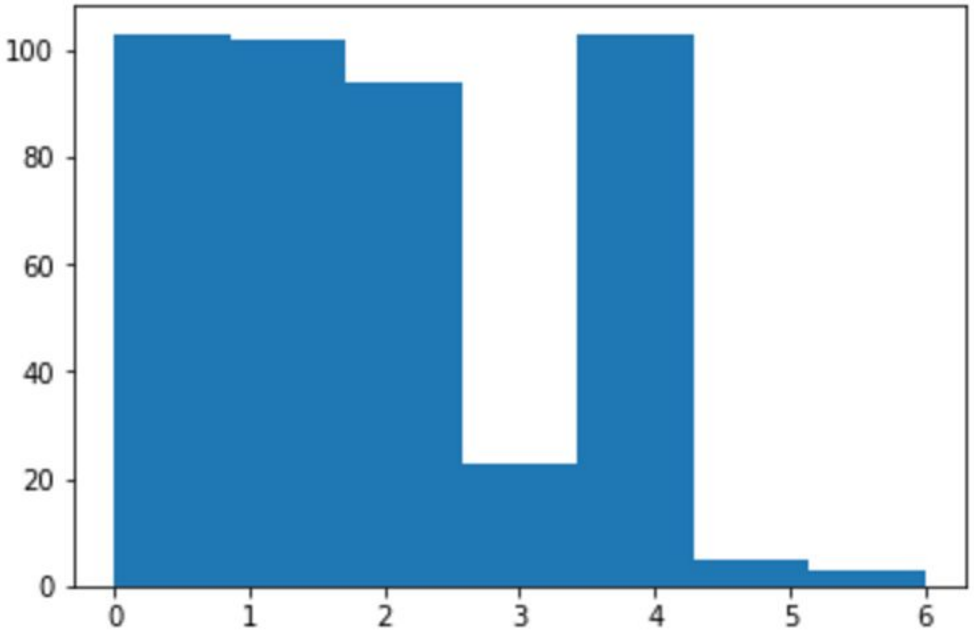


Exploratory Analysis: All workout starting points

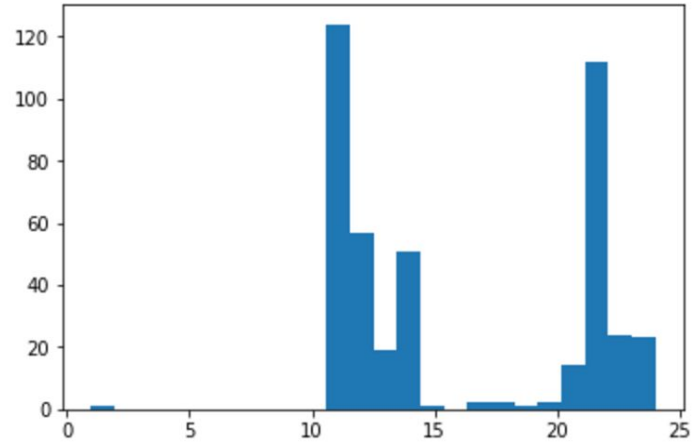


Example: Antwerp Biker

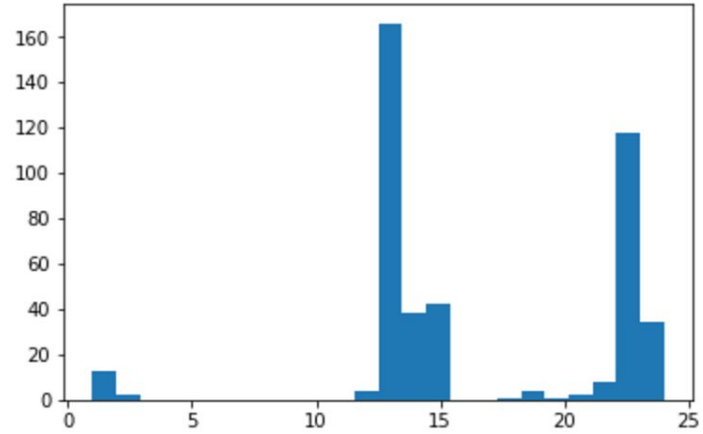
Day of week



Start hour



End hour



ll days combined

Sunday

nday

Tuesday

ednesday

Thursday

Progress To Date

Data Storage	• InfluxDB • Databricks and S3
Data Cleansing	• Garbage Data • Obvious Outliers
Github	• Repo Created
Exploratory Analysis	• Most Common Exercises • Regular Routes

Lessons Learned

Understand and Clean the Data

- Heart Rate Time Series Data vs. non-HR Time Series Data
- Running & Biking are most dominantly logged activities
- Garbage Data
  - HR below 40 bpm (unlikely)
  - HR above 250 bpm (death)
  - 4000 ft delta altitude in 5s
- Data Inconsistency
  - Time Series capped at 500 points regardless of exercise duration -- timestamp delta not uniform

Timeline & Next Steps