# Map My Fitness

Camille Crumpton, Christopher Daffron,
David Mercer, Josh Willis

# Data Collection and Storage

- MapMyApi: the MapMyFitness API
  - Client Credentials
    - Users
  - Authorization Grant
    - Workouts
- Stored in MongoDB
  - User IDs stored with every document for user uniqueness
- Written to a CSV for analysis
  - Excluded fields with Unicode
    - Ex. workout names and notes

# Data Collection Issues

- Redirection
  - Redirects caused calls to be counted twice
- API call limits
  - Extra API keys
    - Thanks to everyone who sent us keys!
- Poor API documentation
  - Missing information
  - Misplaced information
  - Undocumented features
- MapMyApi server errors
  - The resiliency of our code caused it not to crash, which was actually bad

# Data Collection and Storage Strategy

- Users
  - Start from a single user
  - For every user
    - Get user info from MapMyApi
    - Store the user into a MongoDB collection
    - Get user's friends from MapMyApi
    - Add friends to a list of users
- Workouts
  - For every user
    - Get the user's list of workouts from MapMyApi
    - Store every individual workout into a MongoDB collection

# Proposed Analysis

- Is there a difference between verified and unverified workouts in the Map My Fitness data?
- Verified workouts are recorded with a device (e.g. Garmin watch)
- Unverified are manually input
  - Distance
  - Time
  - Calories

# Data Quality Issues

Large Outliers:

- max max_speed = $2.062*10^{10}$ m/s
  - ~$2.998*10^{8}$ m/s…

- Negative average speeds

- Negative energy use

# Logistic Regression

```
Call:
glm(formula = is_verified ~ distance_total + speed_avg + metabolic_energy_total,
    family = binomial, data = num.data.complete)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8389   0.2171   0.2298   0.2455   0.5230

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             4.031e+00  2.940e-02 137.112  < 2e-16 ***
distance_total         -8.777e-06  2.226e-06  -3.943 8.06e-05 ***
speed_avg              -9.144e-02  1.005e-02  -9.100  < 2e-16 ***
metabolic_energy_total -7.092e-08  9.510e-09  -7.457 8.84e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70304  on 269021  degrees of freedom
Residual deviance: 69554  on 269018  degrees of freedom
AIC: 69562

Number of Fisher Scoring iterations: 6
```

# Logistic Regression

Transform coefficients to odds:

- distance_total = 0.99
- speed_avg = 0.91
- metabolic_energy_total = 0.99

Significance Tests are less than 0.05 due to large samples sizes.

# Linear Model Predicting Energy Burned

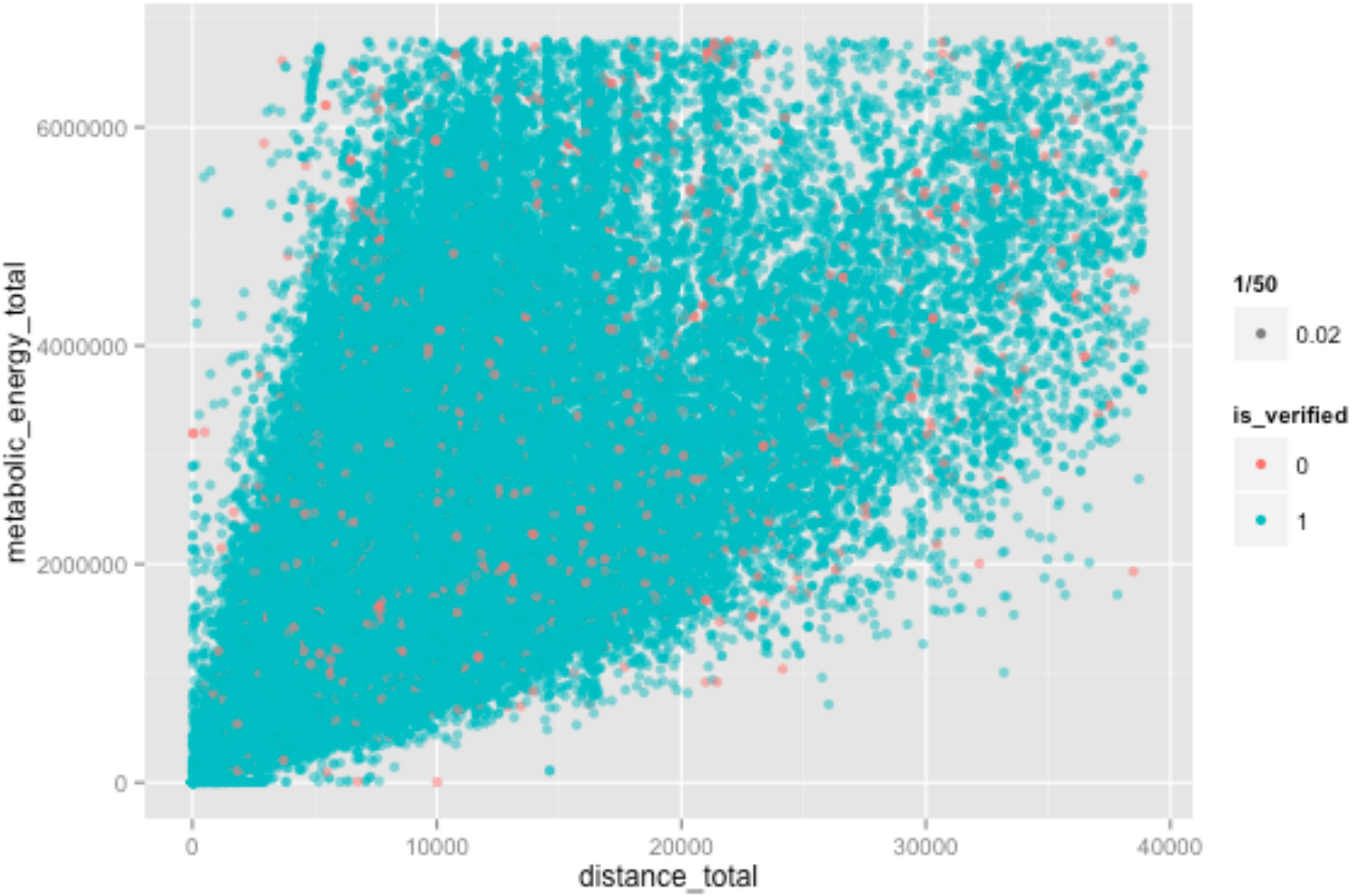
```
Call:
lm(formula = metabolic_energy_total ~ distance_total + speed_avg +
    factor(is_verified), data = num.data.complete)

Residuals:
     Min       1Q   Median       3Q      Max
-5792561  -560141  -139344   436358  5135423

Coefficients:
                          Estimate   Std. Error  t value
(Intercept)           1230717.8039   11786.3145  104.419
distance_total            196.2186       0.4324  453.837
speed_avg              -217582.5683    1863.8315 -116.739
factor(is_verified)1    -38482.4773   11166.7591   -3.446
                                  Pr(>|t|)
(Intercept)          < 0.0000000000000002 ***
distance_total       < 0.0000000000000002 ***
speed_avg            < 0.0000000000000002 ***
factor(is_verified)1              0.000569 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 882500 on 250827 degrees of freedom
Multiple R-squared:  0.5028,    Adjusted R-squared:  0.5027
F-statistic: 8.453e+04 on 3 and 250827 DF,  p-value: < 0.00000000000000022
```

# Visual Analysis



# Questions?