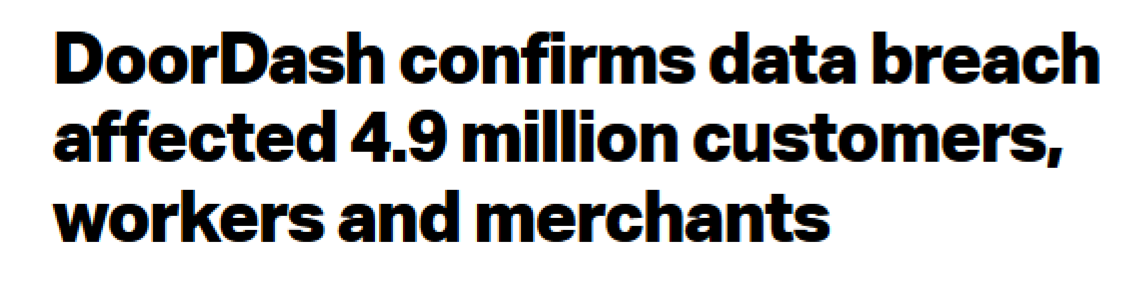
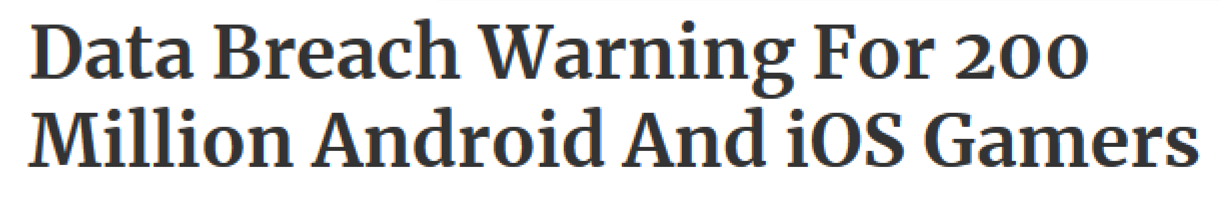
OK Google, Tell Me About Myself

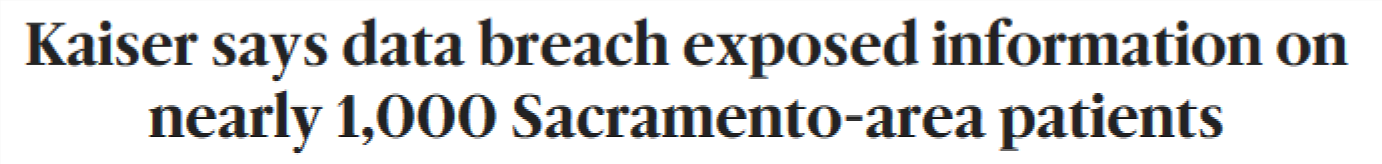
Lisa Chang

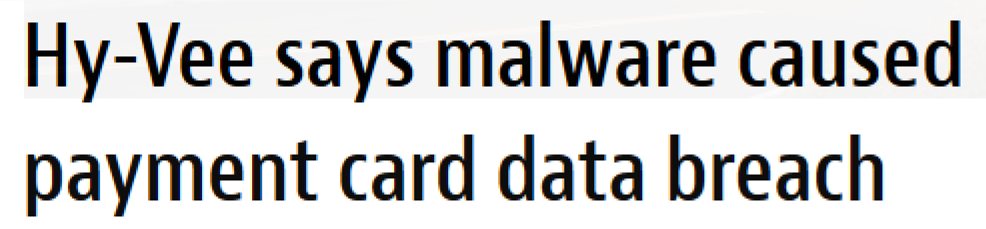
Data Scientist, Praxis Engineering

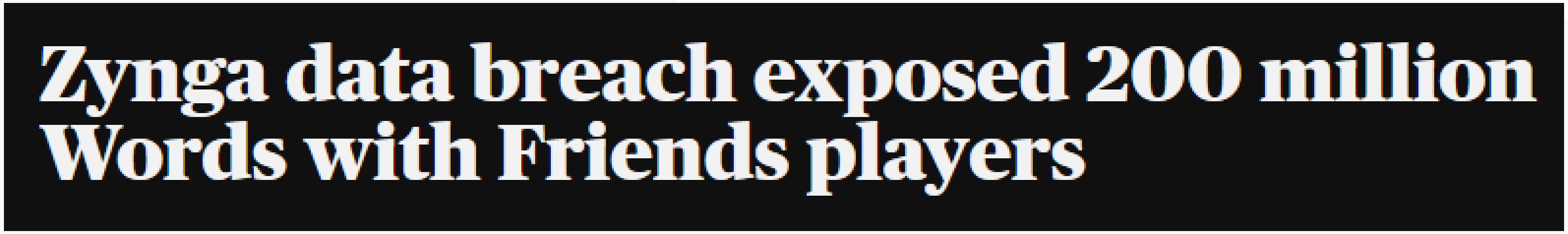
Last Week's Data Headlines


DoorDash confirms data breach affected 4.9 million customers, workers and merchants

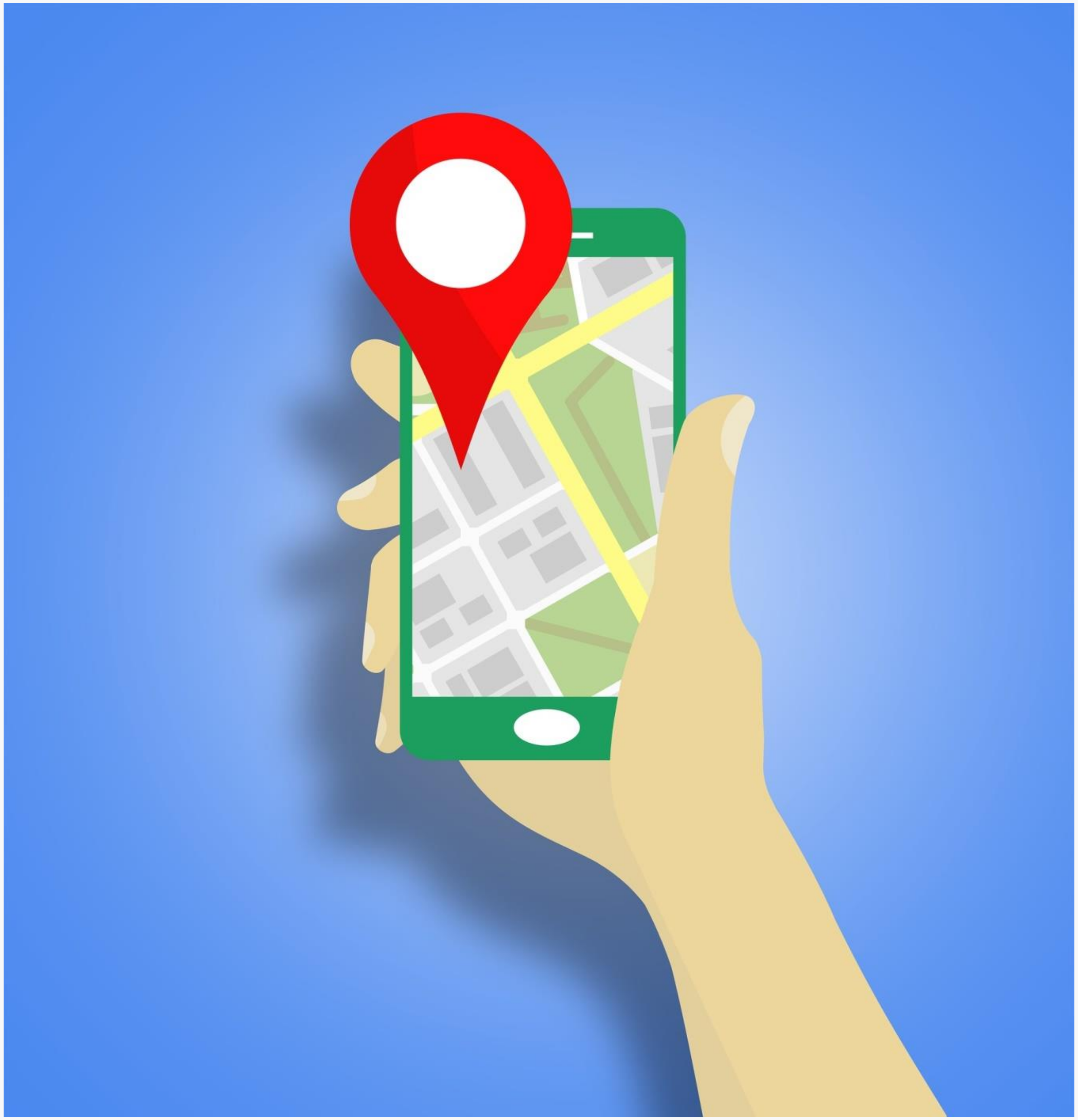

Data Breach Warning For 200 Million Android And iOS Gamers


Kaiser says data breach exposed information on nearly 1,000 Sacramento-area patients

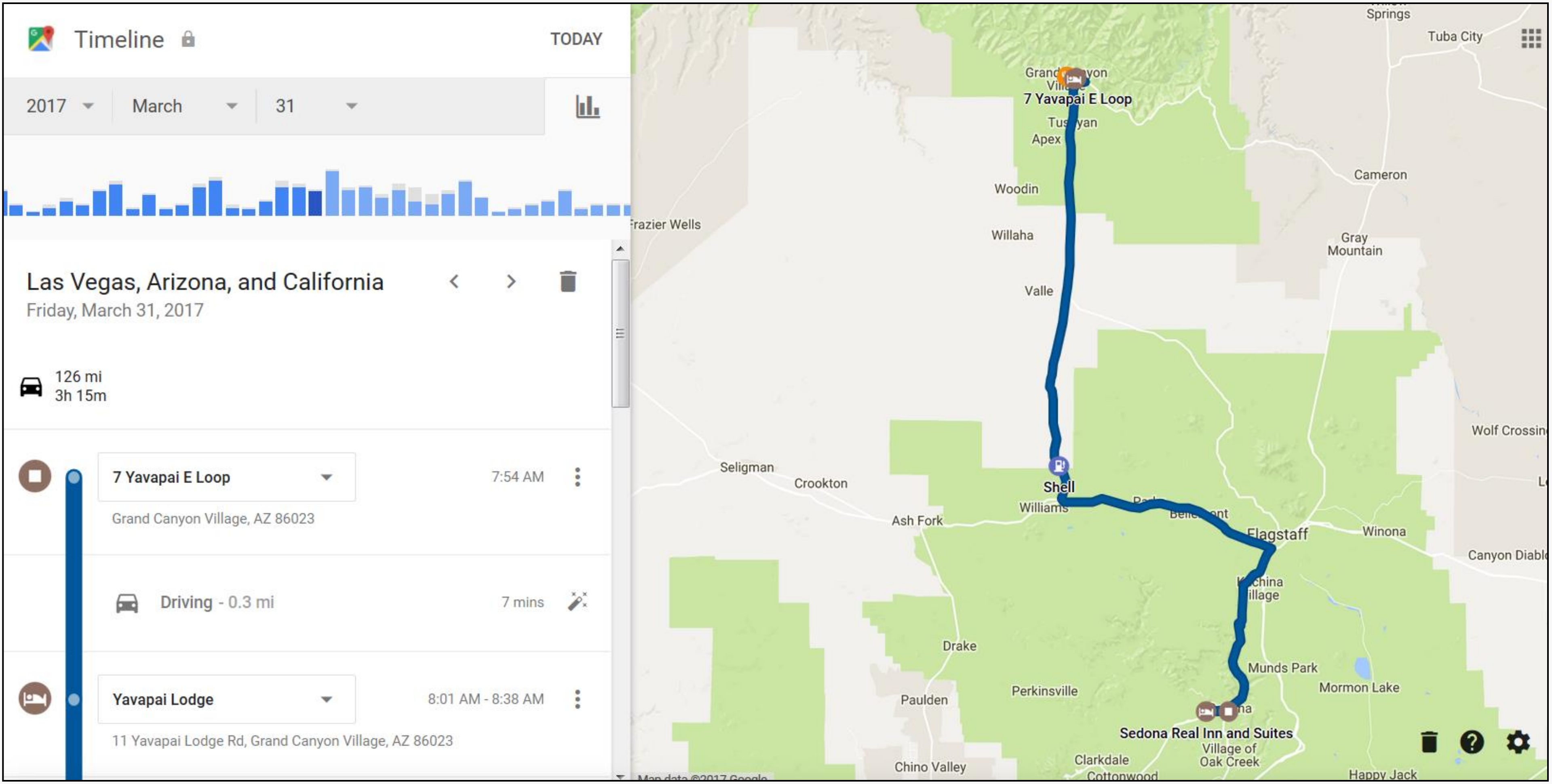

Hy-Vee says malware caused payment card data breach


Zynga data breach exposed 200 million Words with Friends players

Google™



Google Timelines



Timeline

TODAY

2017 ▾ March ▾ 31 ▾

Las Vegas, Arizona, and California

Friday, March 31, 2017

126 mi
3h 15m

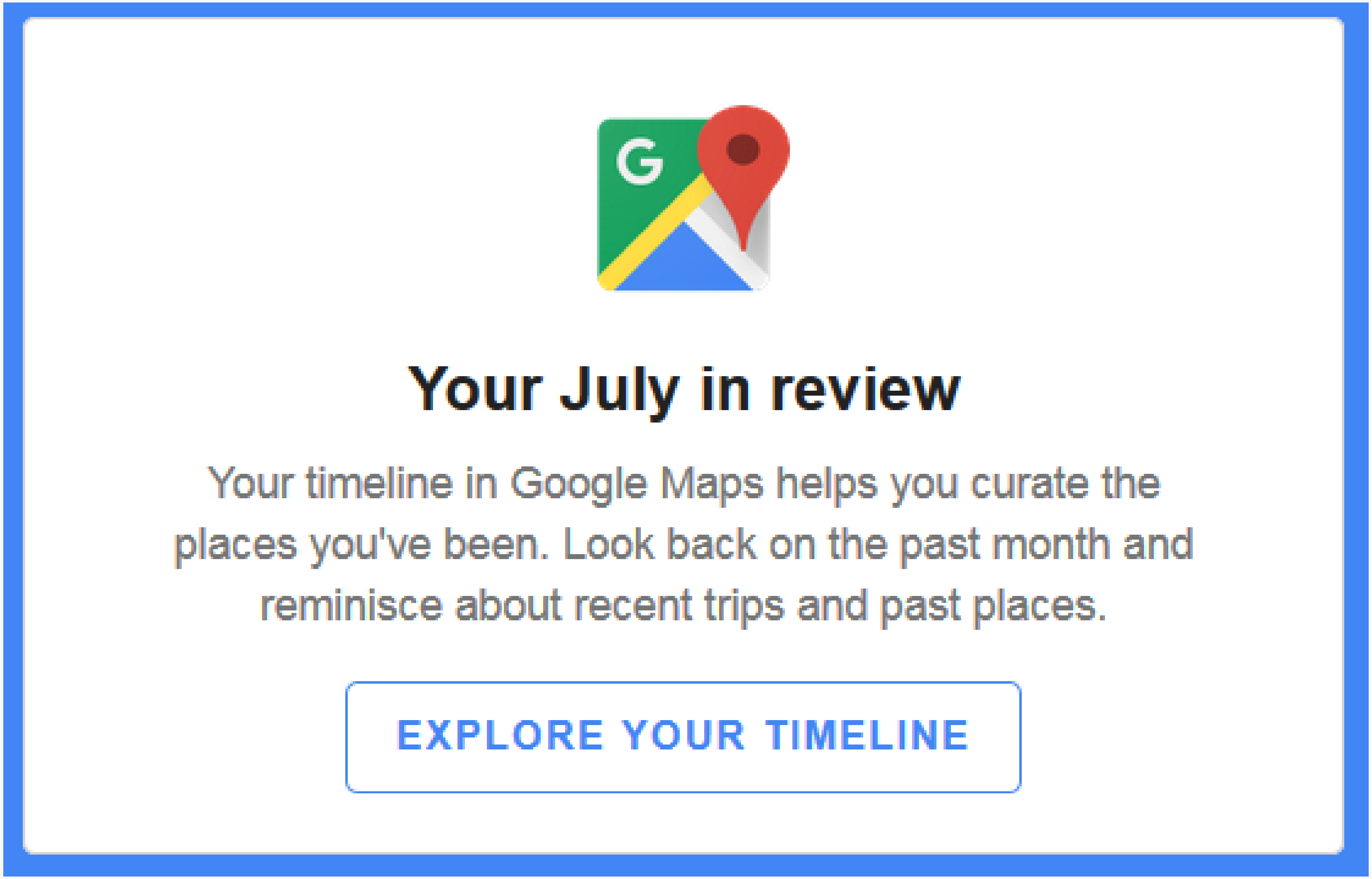
7 Yavapai E Loop ▾ 7:54 AM

Grand Canyon Village, AZ 86023

Driving - 0.3 mi 7 mins

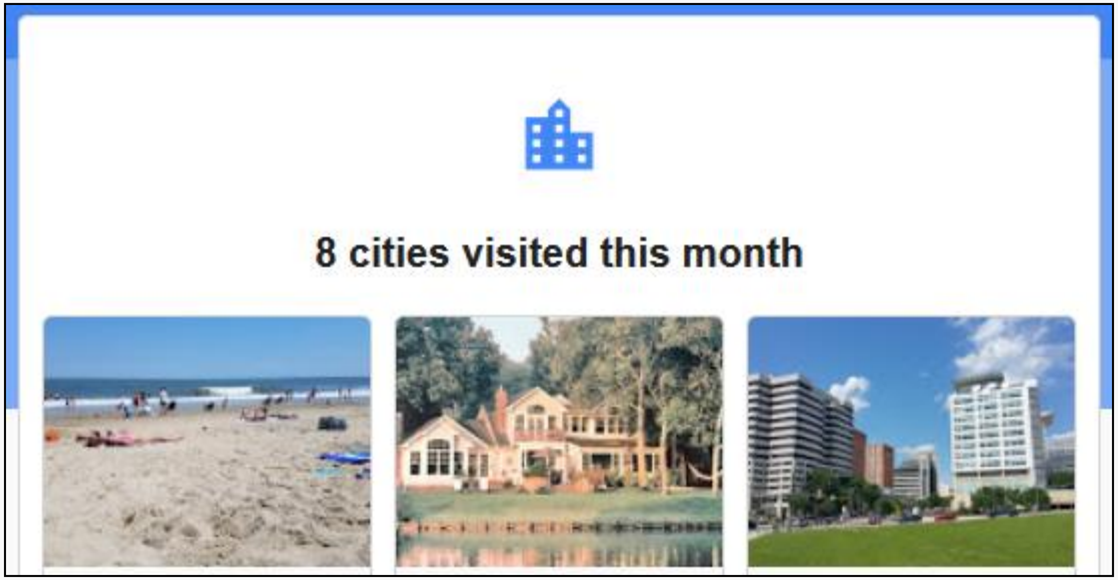
Yavapai Lodge ▾ 8:01 AM - 8:38 AM

11 Yavapai Lodge Rd, Grand Canyon Village, AZ 86023

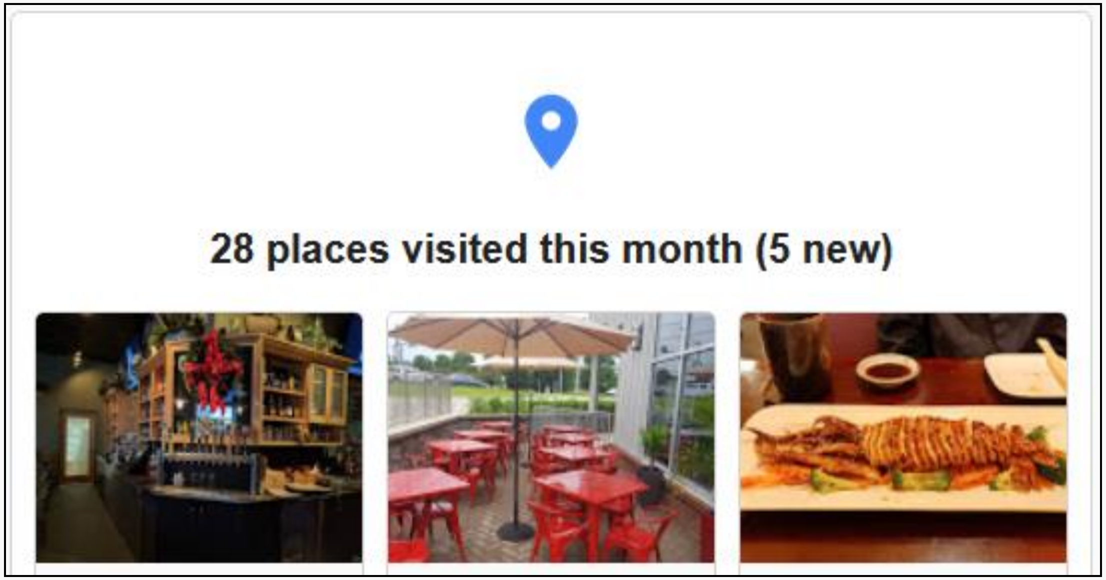


Your July in review

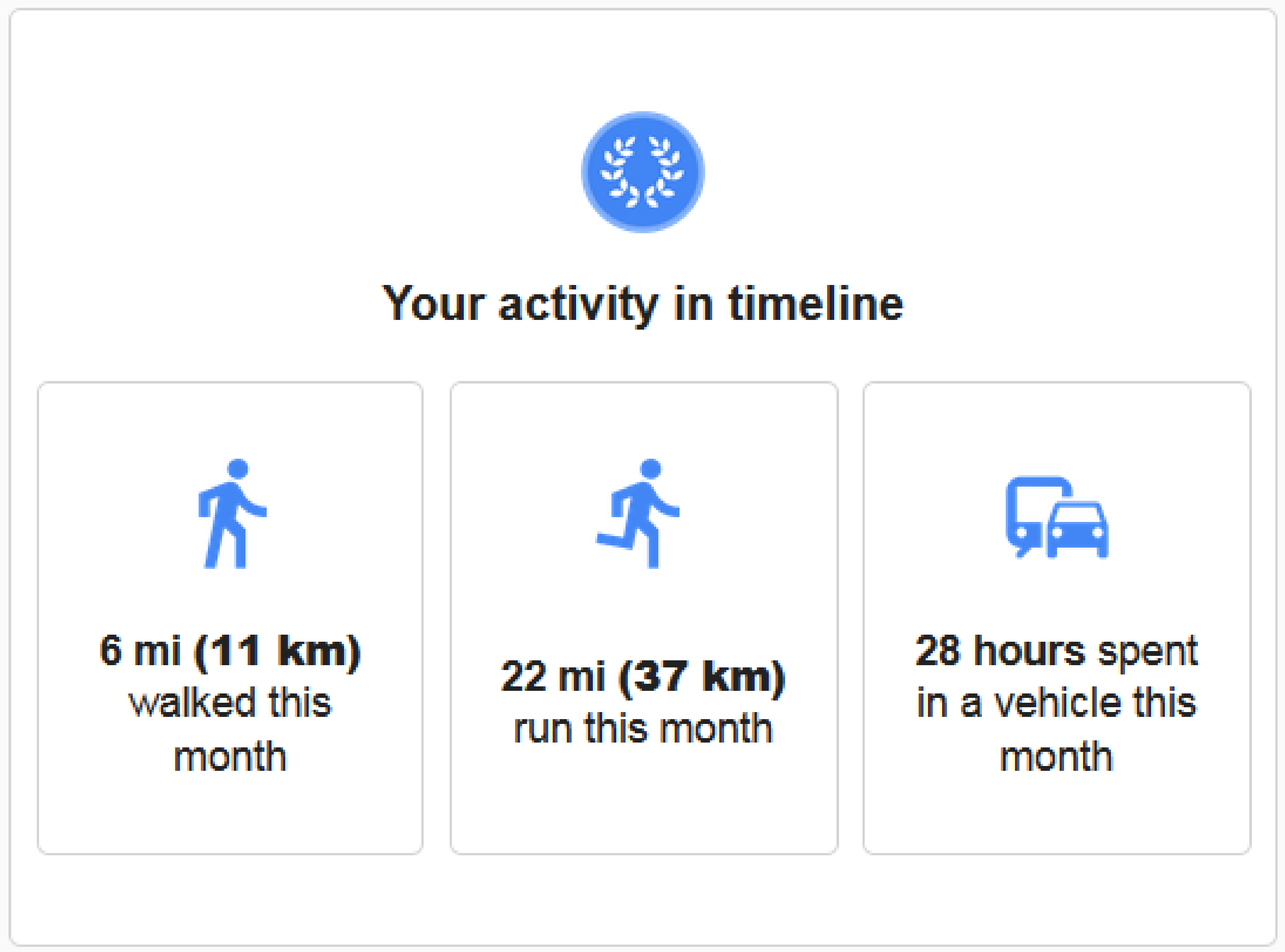
Your timeline in Google Maps helps you curate the places you've been. Look back on the past month and reminisce about recent trips and past places.



8 cities visited this month



28 places visited this month (5 new)



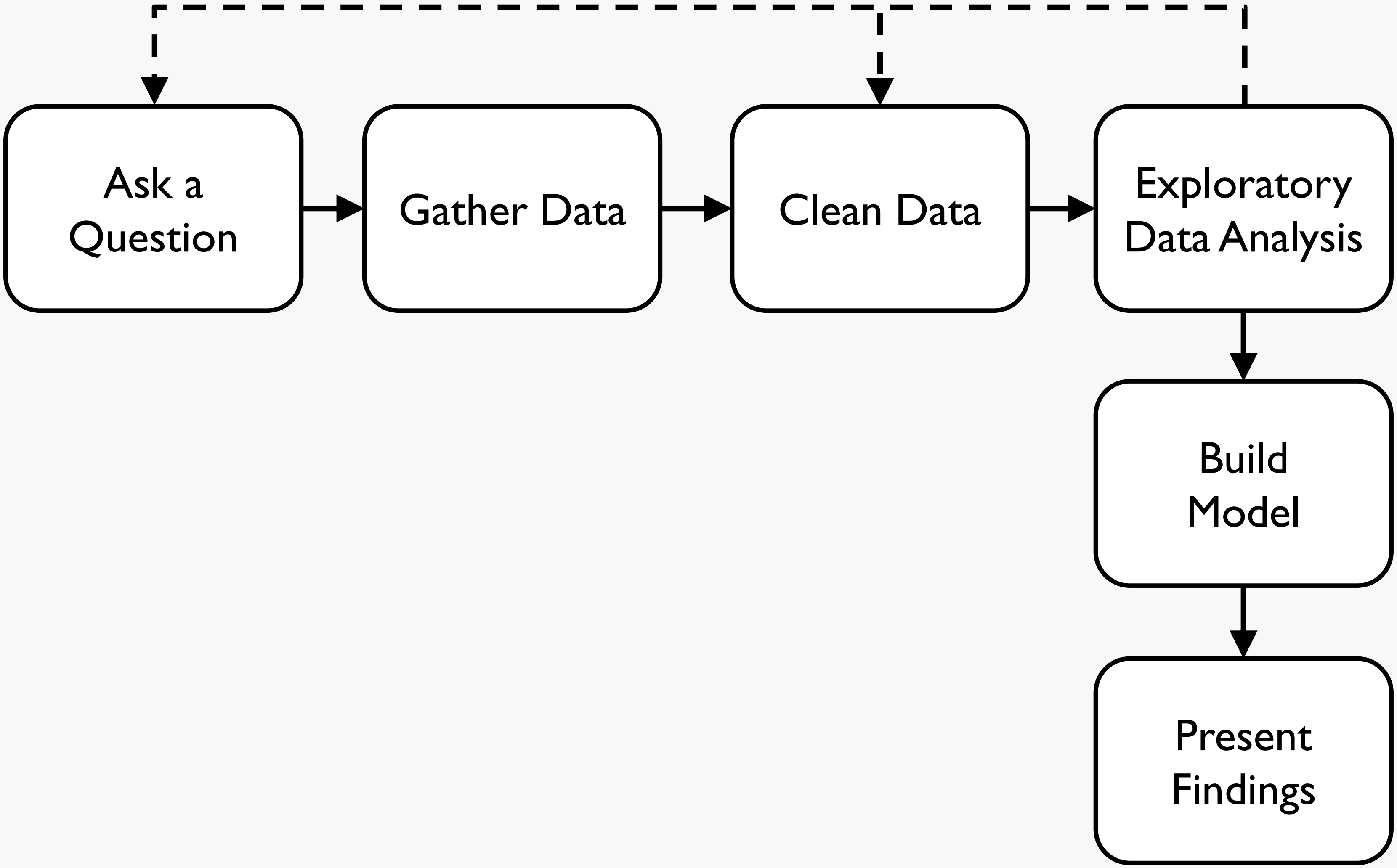
Your activity in timeline

6 mi (11 km) walked this month

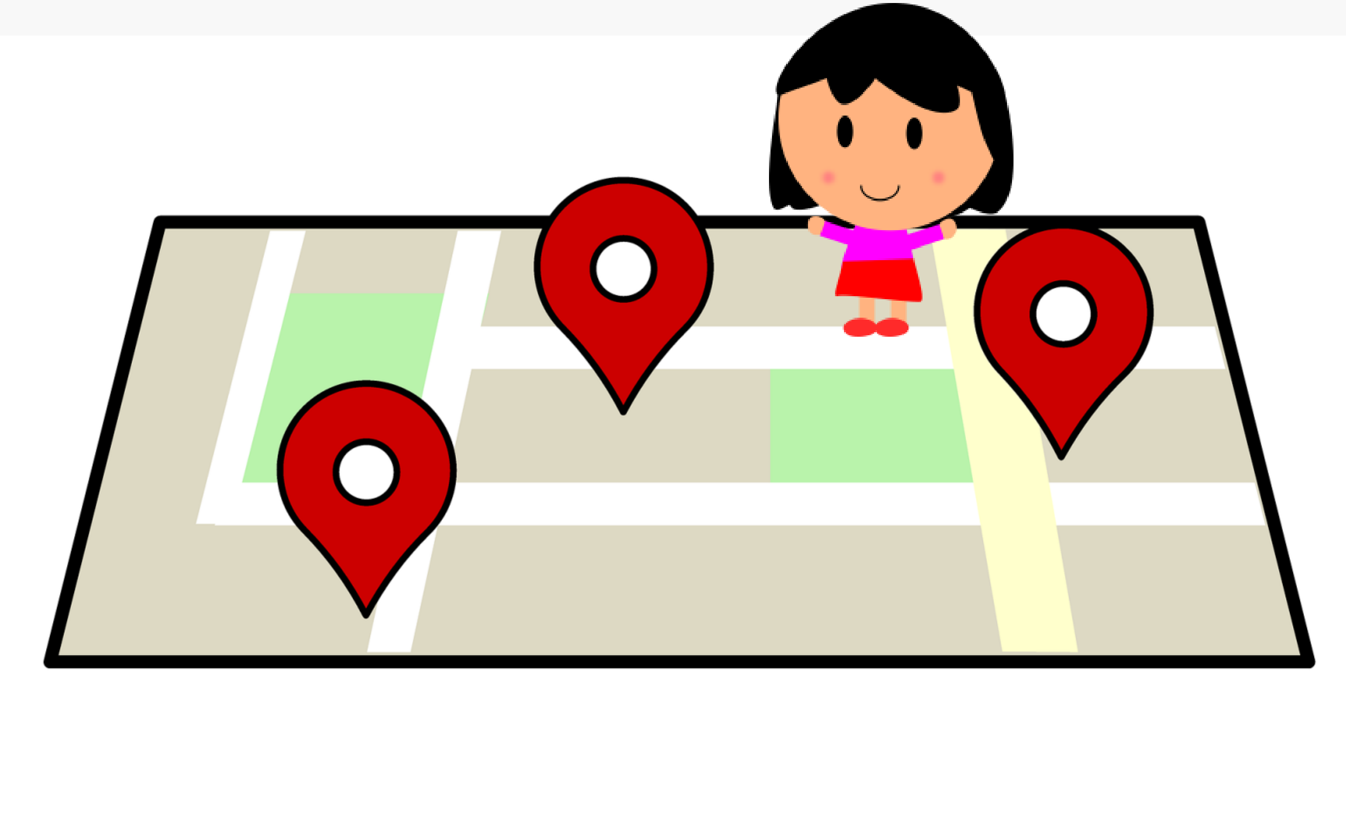
22 mi (37 km) run this month

28 hours spent in a vehicle this month

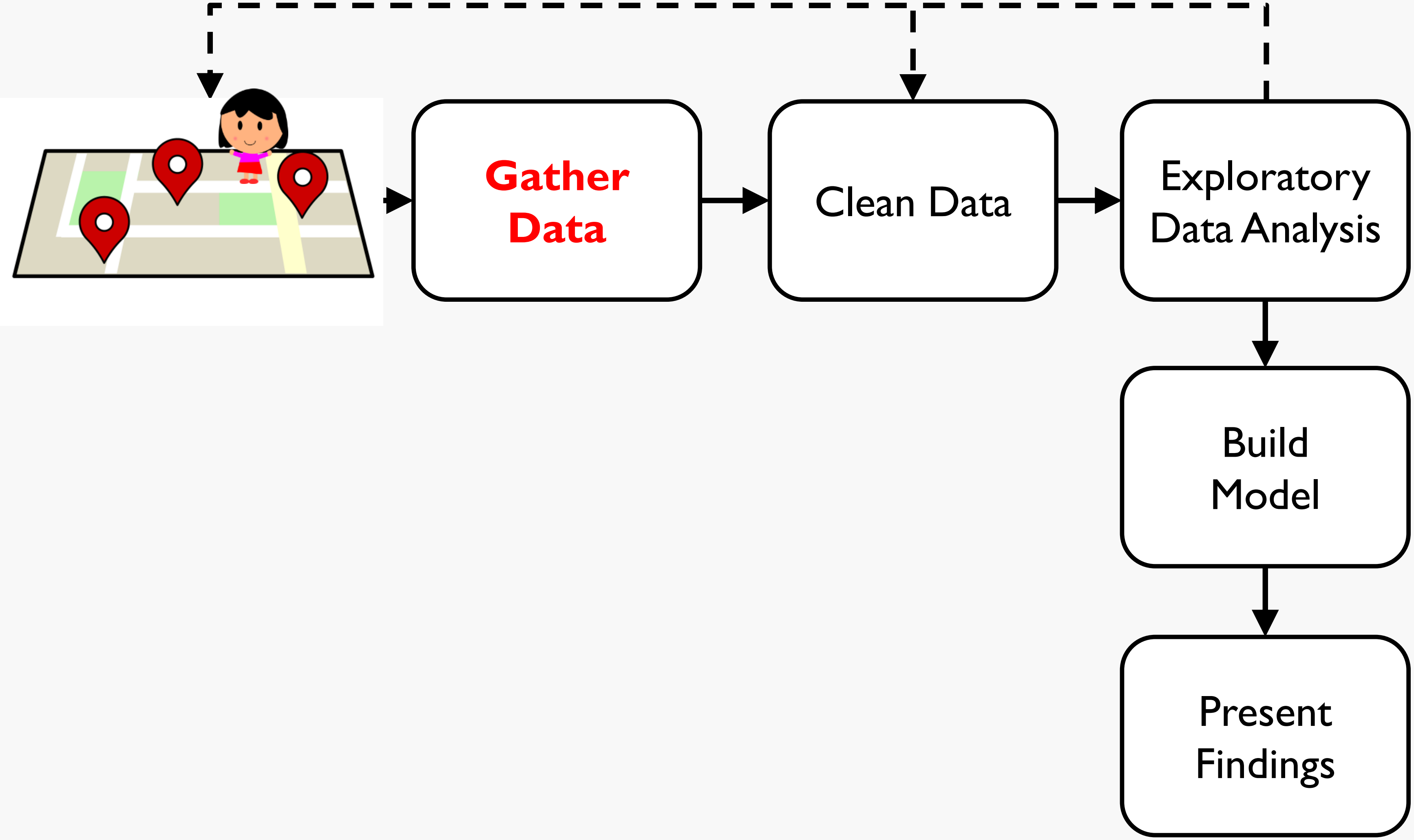
Data Science Process



Can I create a model of my life from stored location data?



Data Science Process



Gather Data

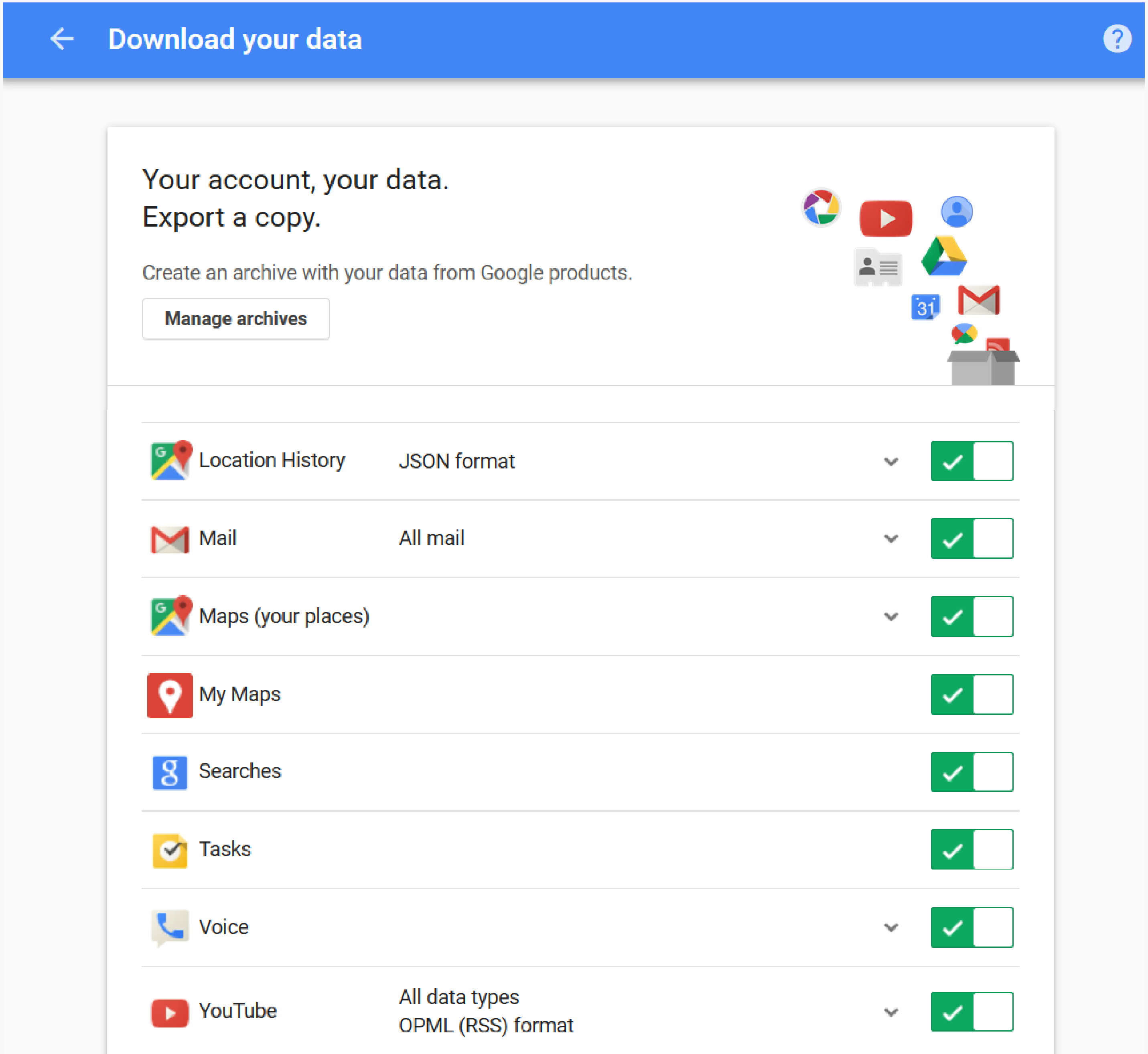
Clean Data

Exploratory Data Analysis

Build Model

Present Findings

https://takeout.google.com

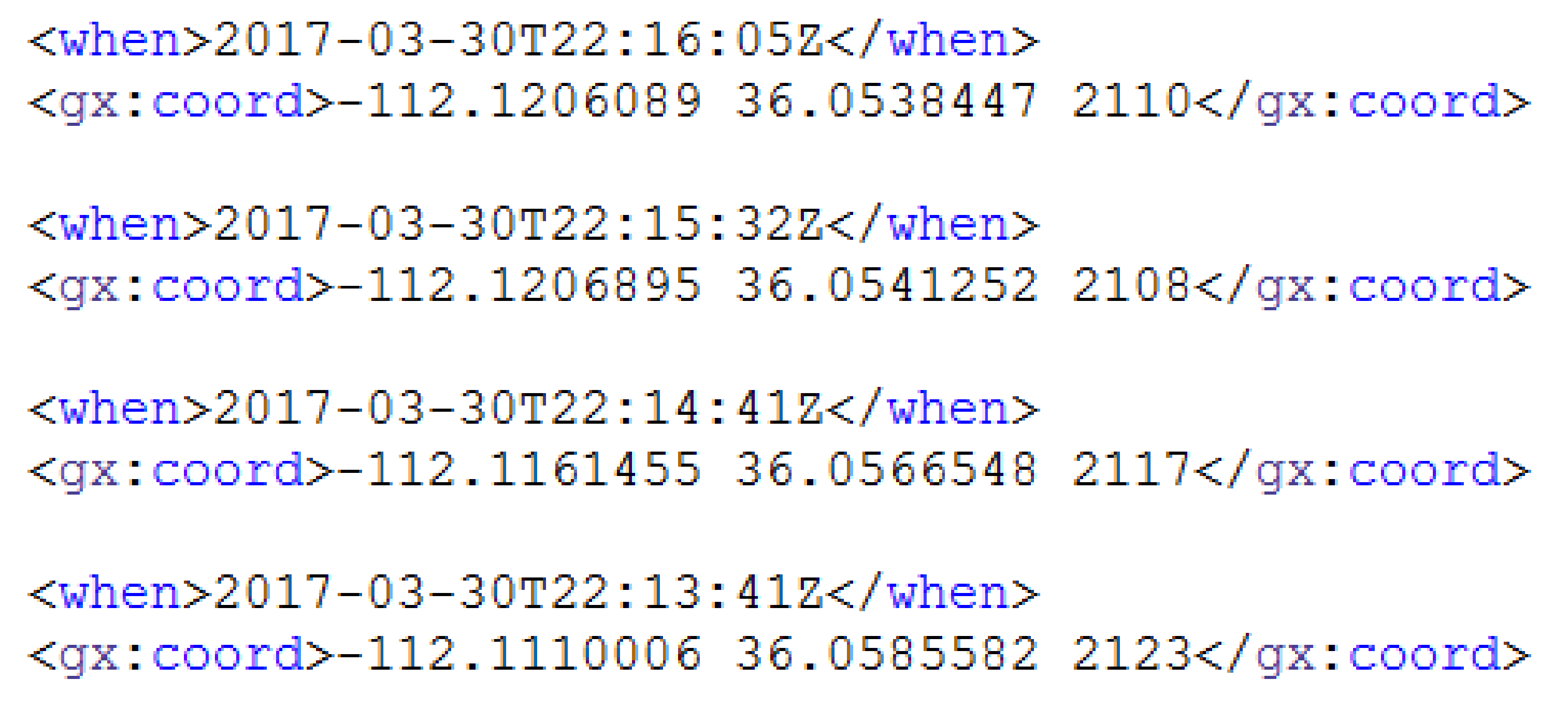


Your account, your data.
Export a copy.

Create an archive with your data from Google products.

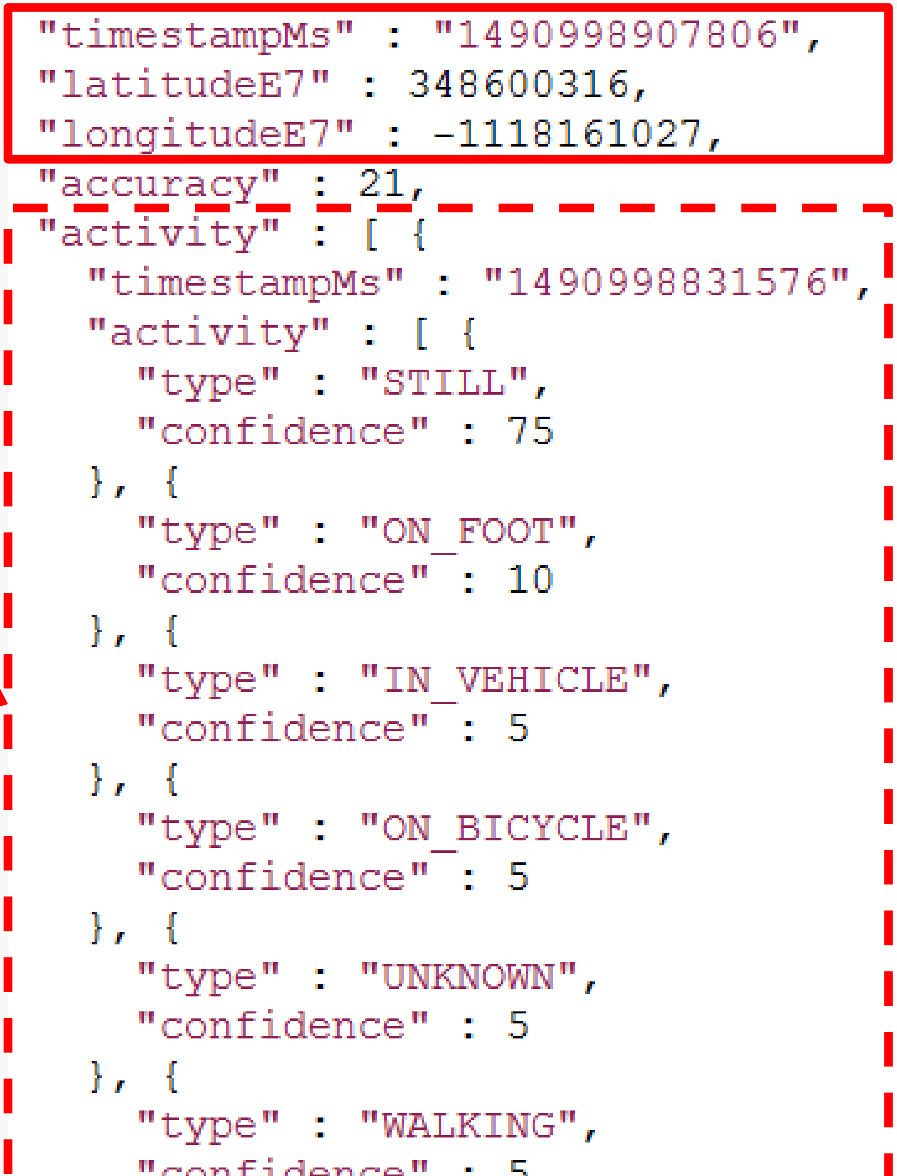
Location History	JSON format	☑
Mail	All mail	☑
Maps (your places)		☑
My Maps		☑
Searches		☑
Tasks		☑
Voice		☑
YouTube	All data types OPML (RSS) format	☑

KML format example



```
<when>2017-03-30T22:16:05Z</when>
<gx:coord>-112.1206089 36.0538447 2110</gx:coord>

<when>2017-03-30T22:15:32Z</when>
<gx:coord>-112.1206895 36.0541252 2108</gx:coord>

<when>2017-03-30T22:14:41Z</when>
<gx:coord>-112.1161455 36.0566548 2117</gx:coord>

<when>2017-03-30T22:13:41Z</when>
<gx:coord>-112.1110006 36.0585582 2123</gx:coord>
```

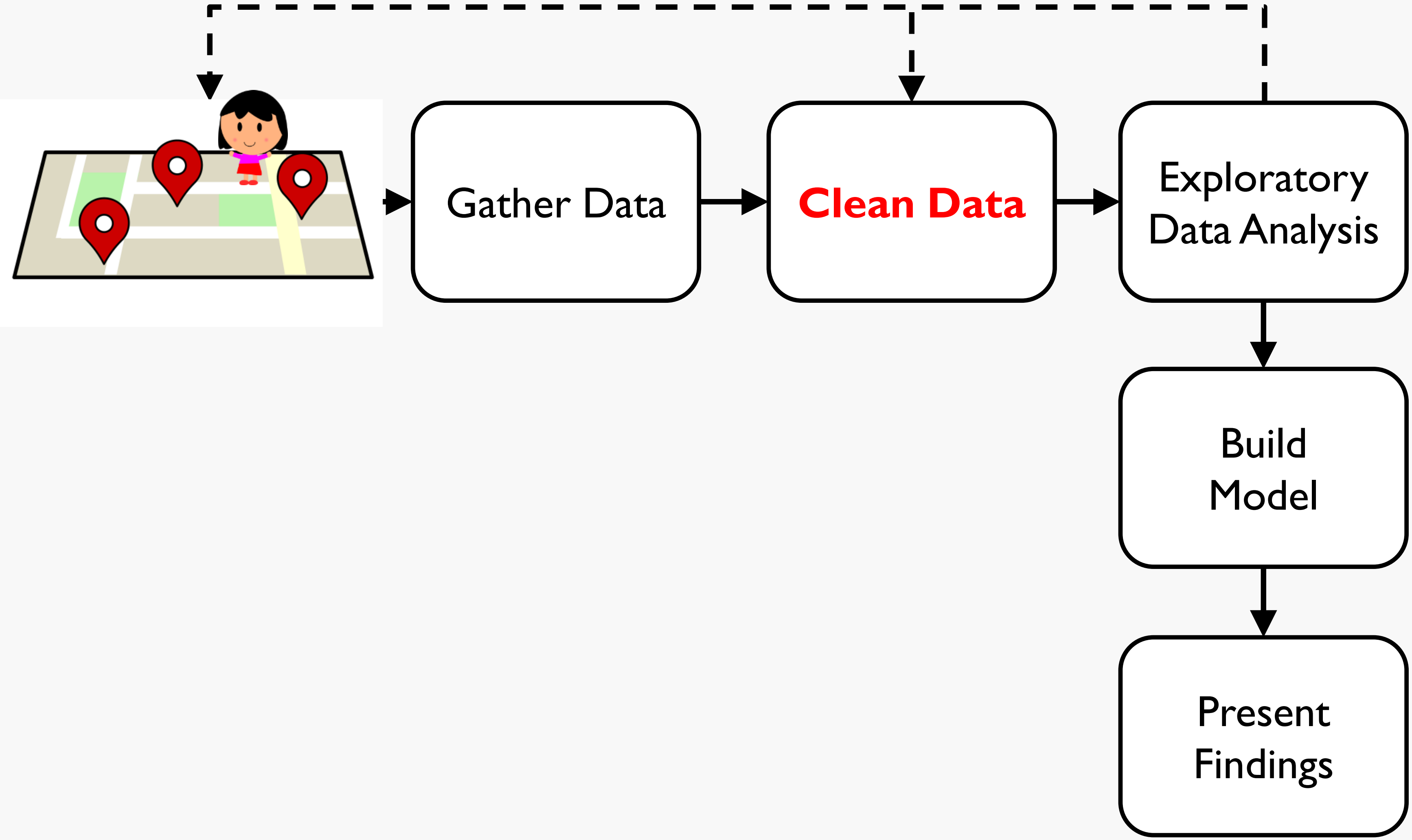
JSON format example



```
"timestampMs" : "1490998907806",
"latitudeE7" : 348600316,
"longitudeE7" : -1118161027,
"accuracy" : 21,
"activity" : [ {
  "timestampMs" : "1490998831576",
  "activity" : [ {
    "type" : "STILL",
    "confidence" : 75
  }, {
    "type" : "ON_FOOT",
    "confidence" : 10
  }, {
    "type" : "IN_VEHICLE",
    "confidence" : 5
  }, {
    "type" : "ON_BICYCLE",
    "confidence" : 5
  }, {
    "type" : "UNKNOWN",
    "confidence" : 5
  }, {
    "type" : "WALKING",
    "confidence" : 5
```

Not always available

Data Science Process



Gather Data

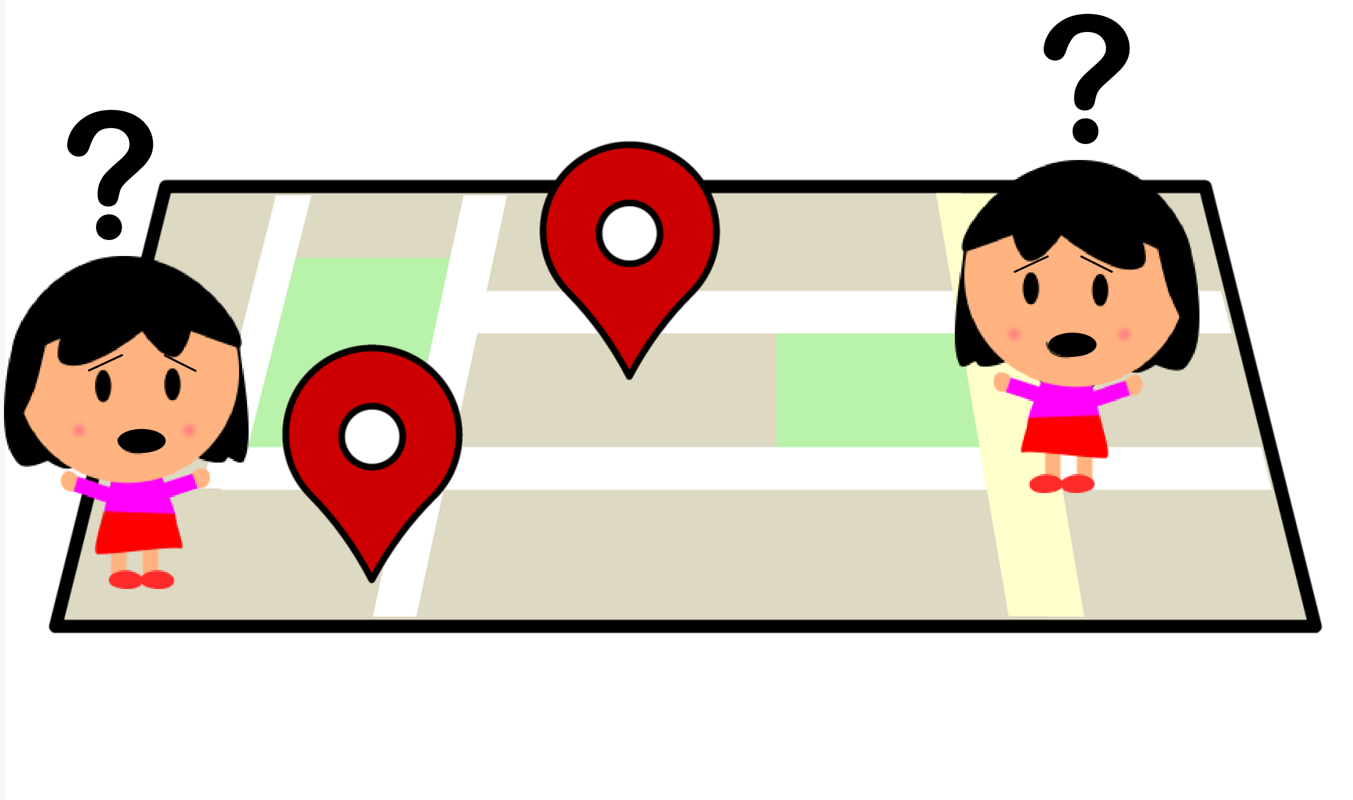
Clean Data

Exploratory Data Analysis

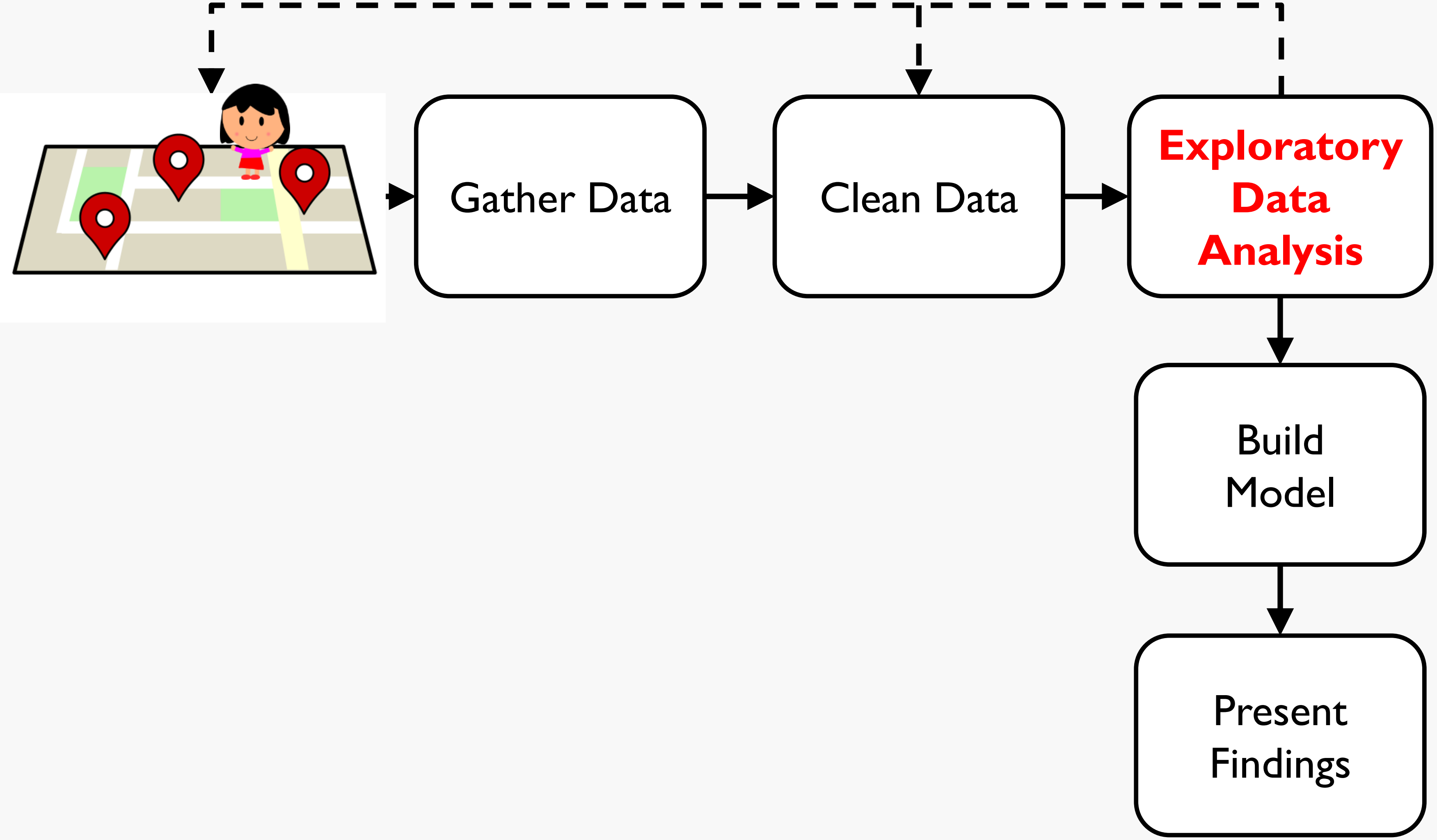
Build Model

Present Findings

Traveling at the speed of light



Data Science Process



Gather Data

Clean Data

Exploratory Data Analysis

Build Model

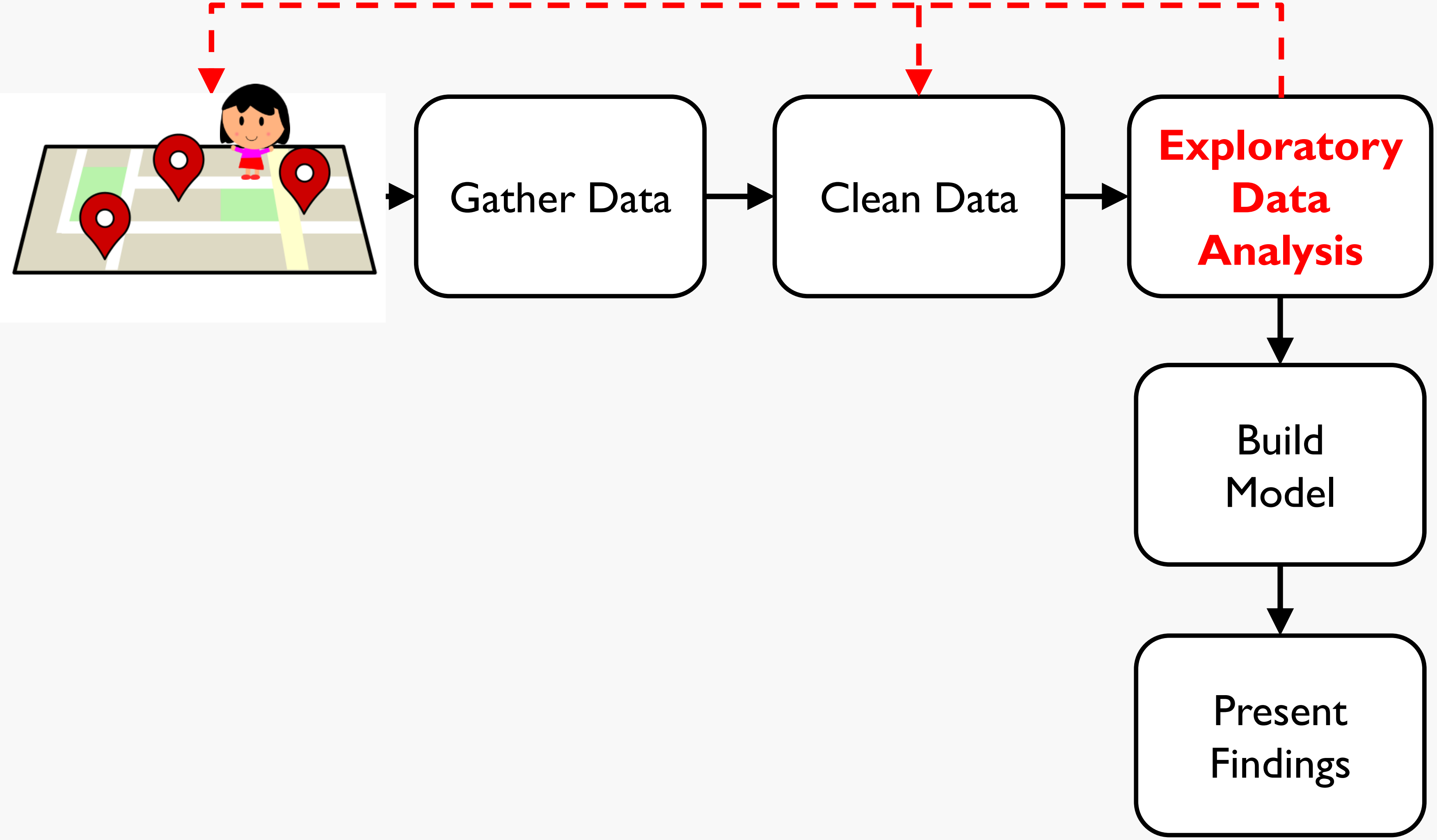
Present Findings

What is EDA?

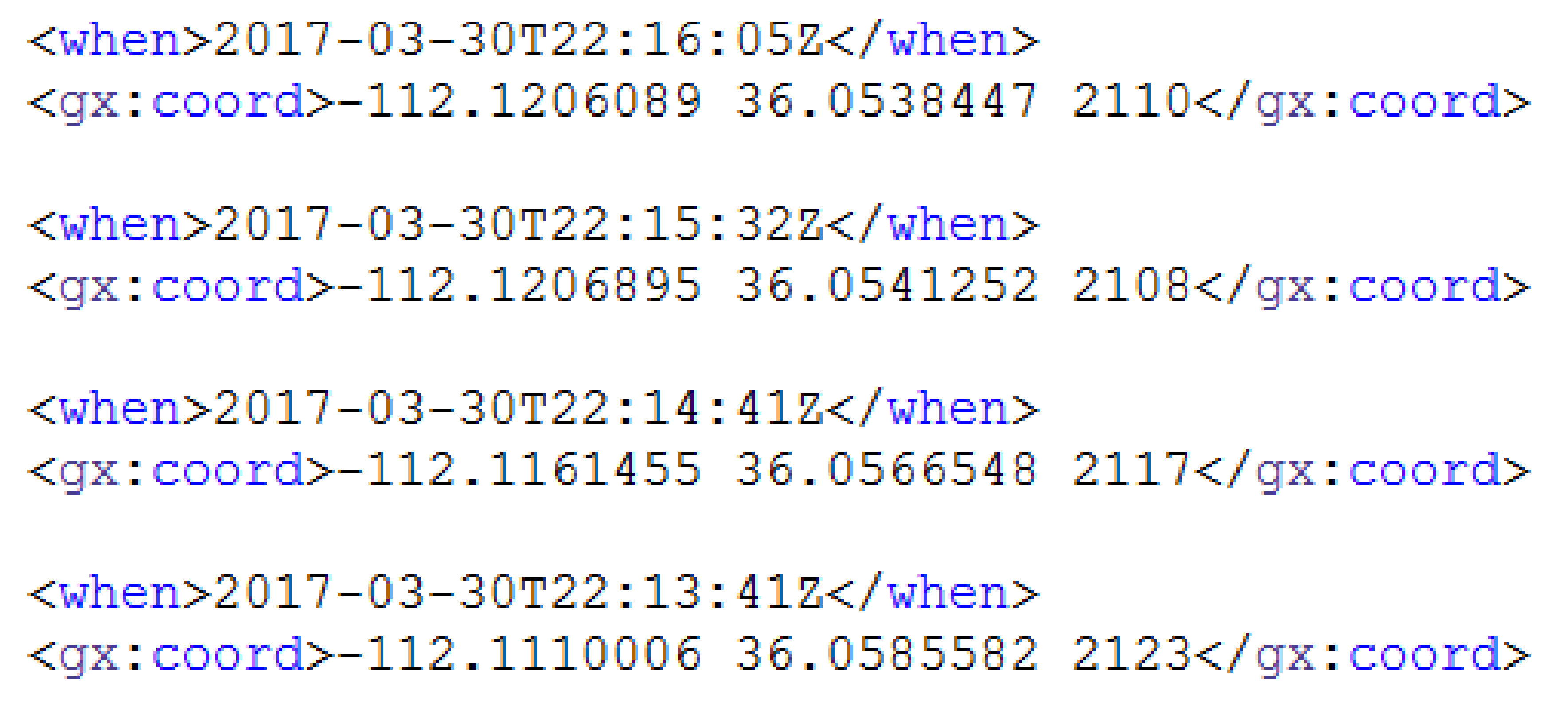
- Define characteristics
  - Trends
  - Biases
  - Variability
  - Breadth

- Test Assumptions
- Visualize

Data Science Process

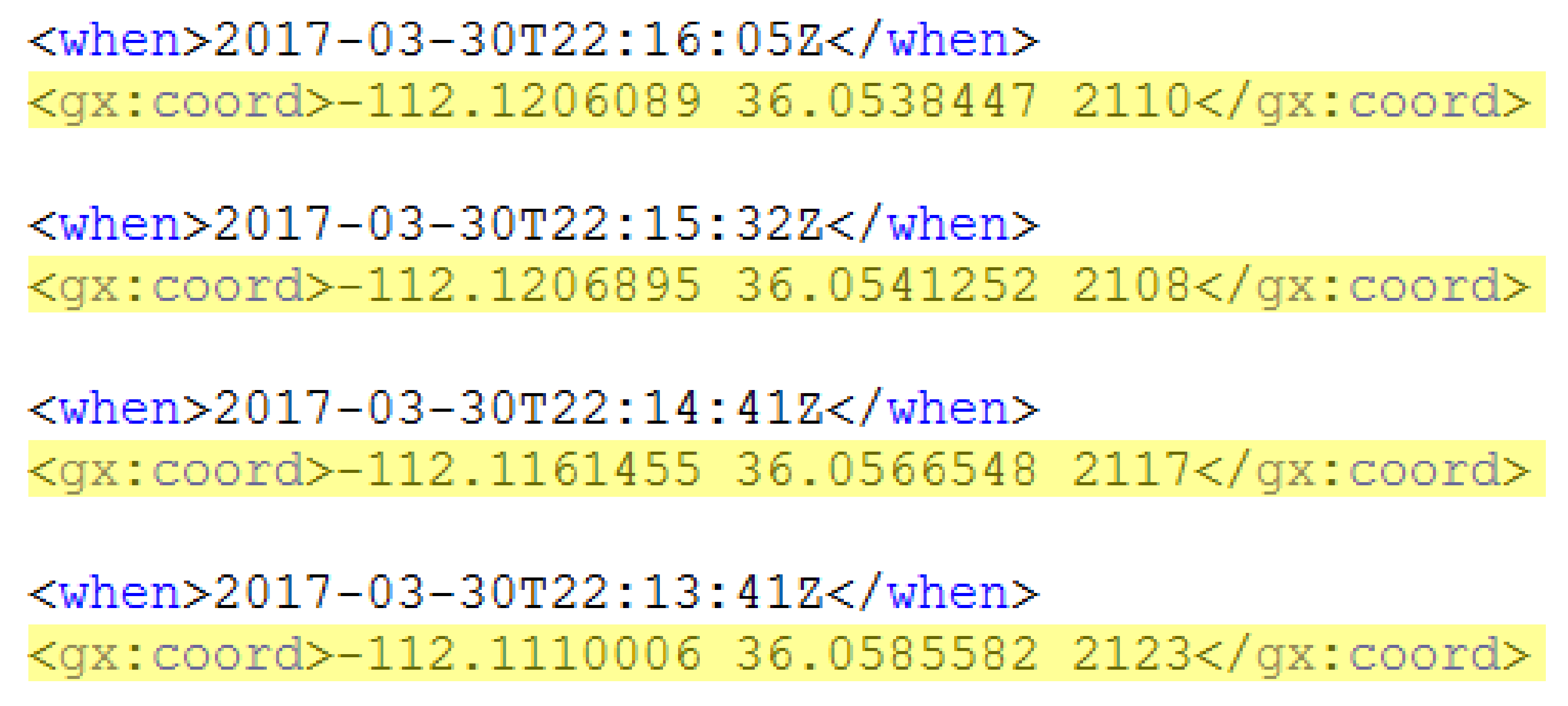


What's in the data?



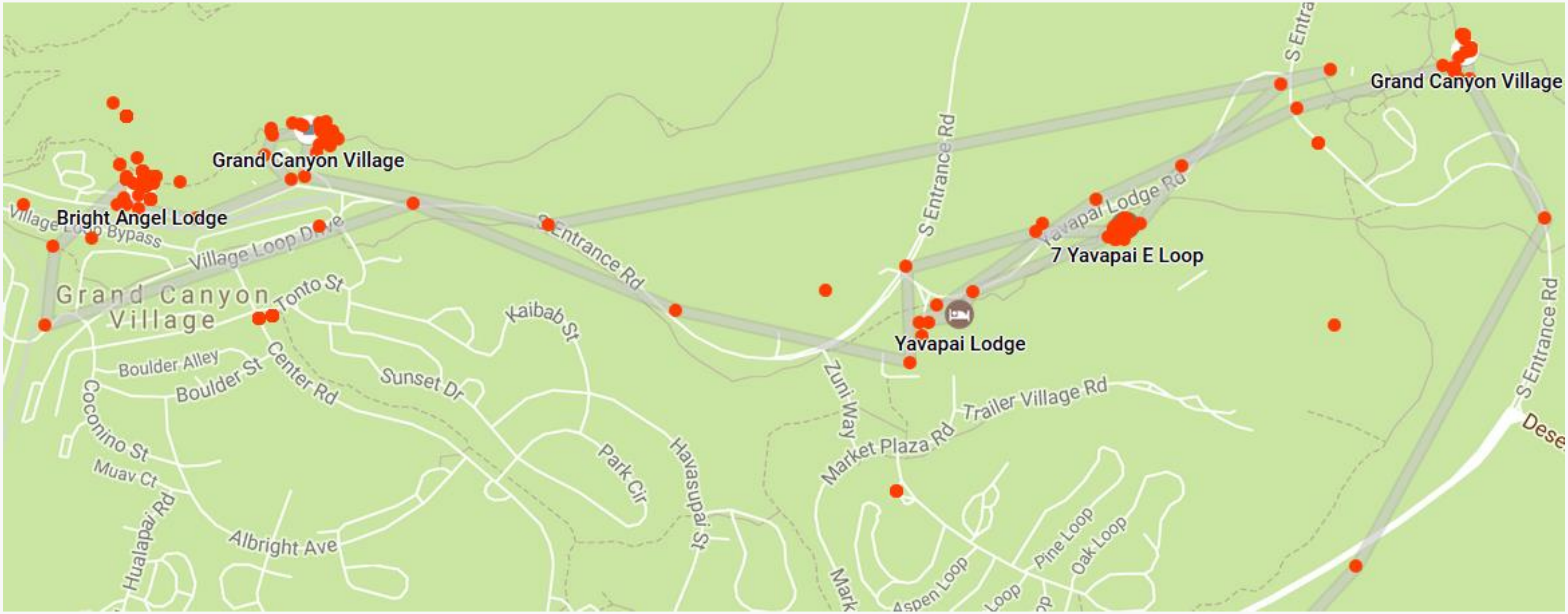
```
<when>2017-03-30T22:16:05Z</when>
<gx:coord>-112.1206089 36.0538447 2110</gx:coord>

<when>2017-03-30T22:15:32Z</when>
<gx:coord>-112.1206895 36.0541252 2108</gx:coord>

<when>2017-03-30T22:14:41Z</when>
<gx:coord>-112.1161455 36.0566548 2117</gx:coord>

<when>2017-03-30T22:13:41Z</when>
<gx:coord>-112.1110006 36.0585582 2123</gx:coord>
```

Location

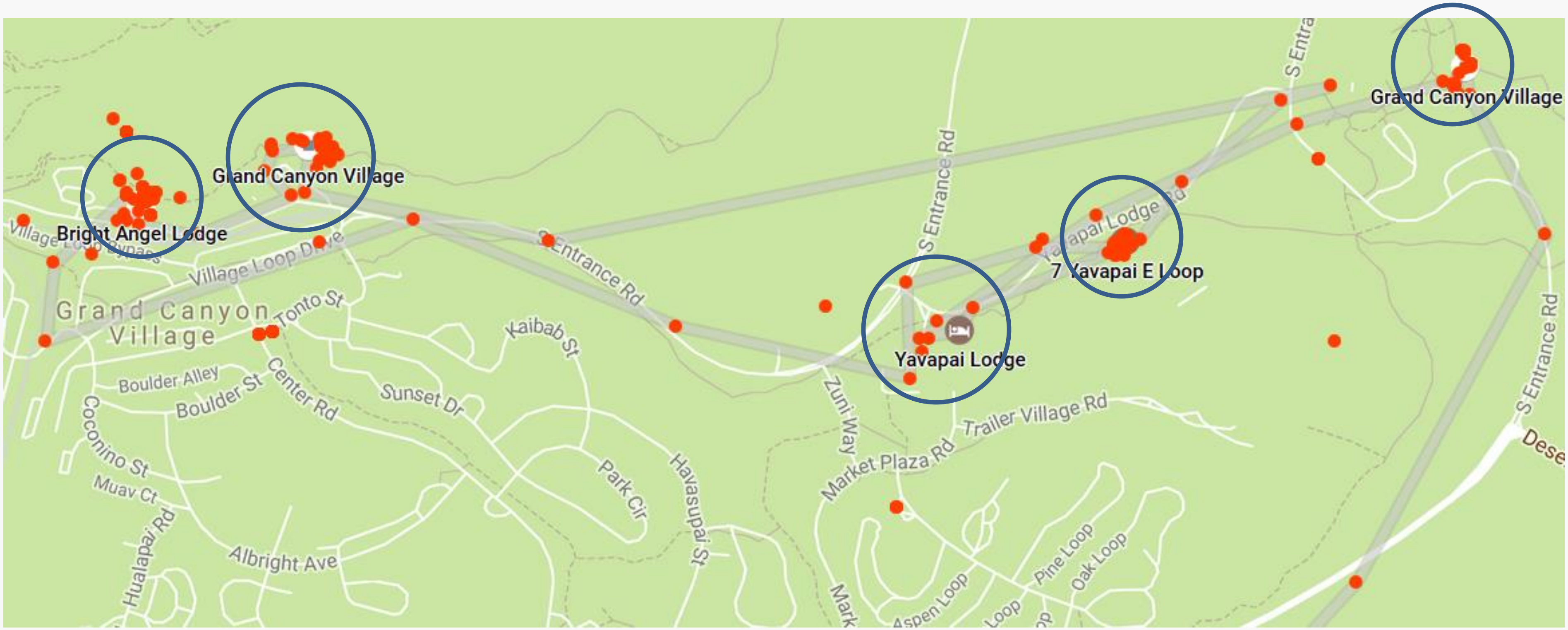


```
<when>2017-03-30T22:16:05Z</when>
<gx:coord>-112.1206089 36.0538447 2110</gx:coord>

<when>2017-03-30T22:15:32Z</when>
<gx:coord>-112.1206895 36.0541252 2108</gx:coord>

<when>2017-03-30T22:14:41Z</when>
<gx:coord>-112.1161455 36.0566548 2117</gx:coord>

<when>2017-03-30T22:13:41Z</when>
<gx:coord>-112.1110006 36.0585582 2123</gx:coord>
```

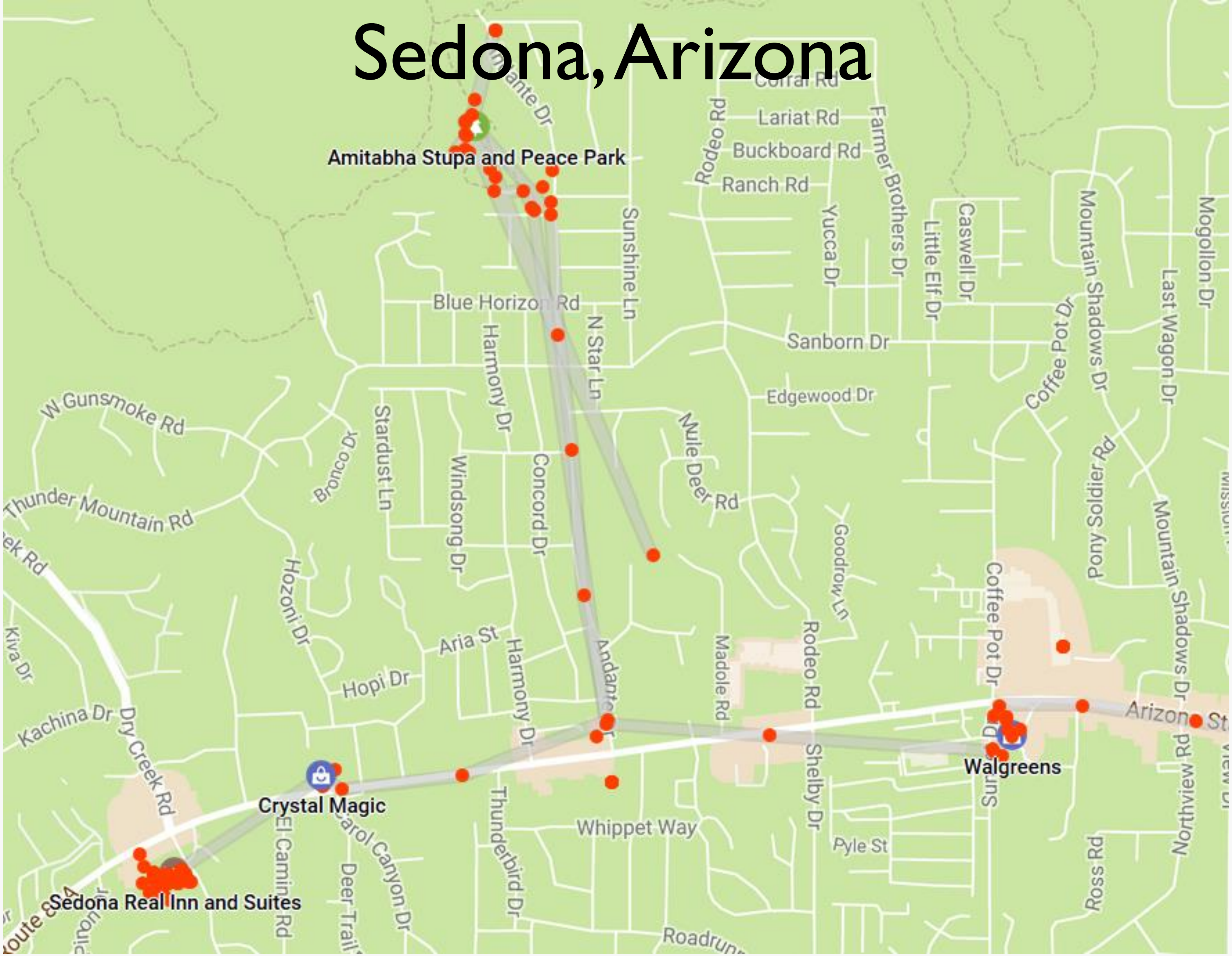
Grand Canyon



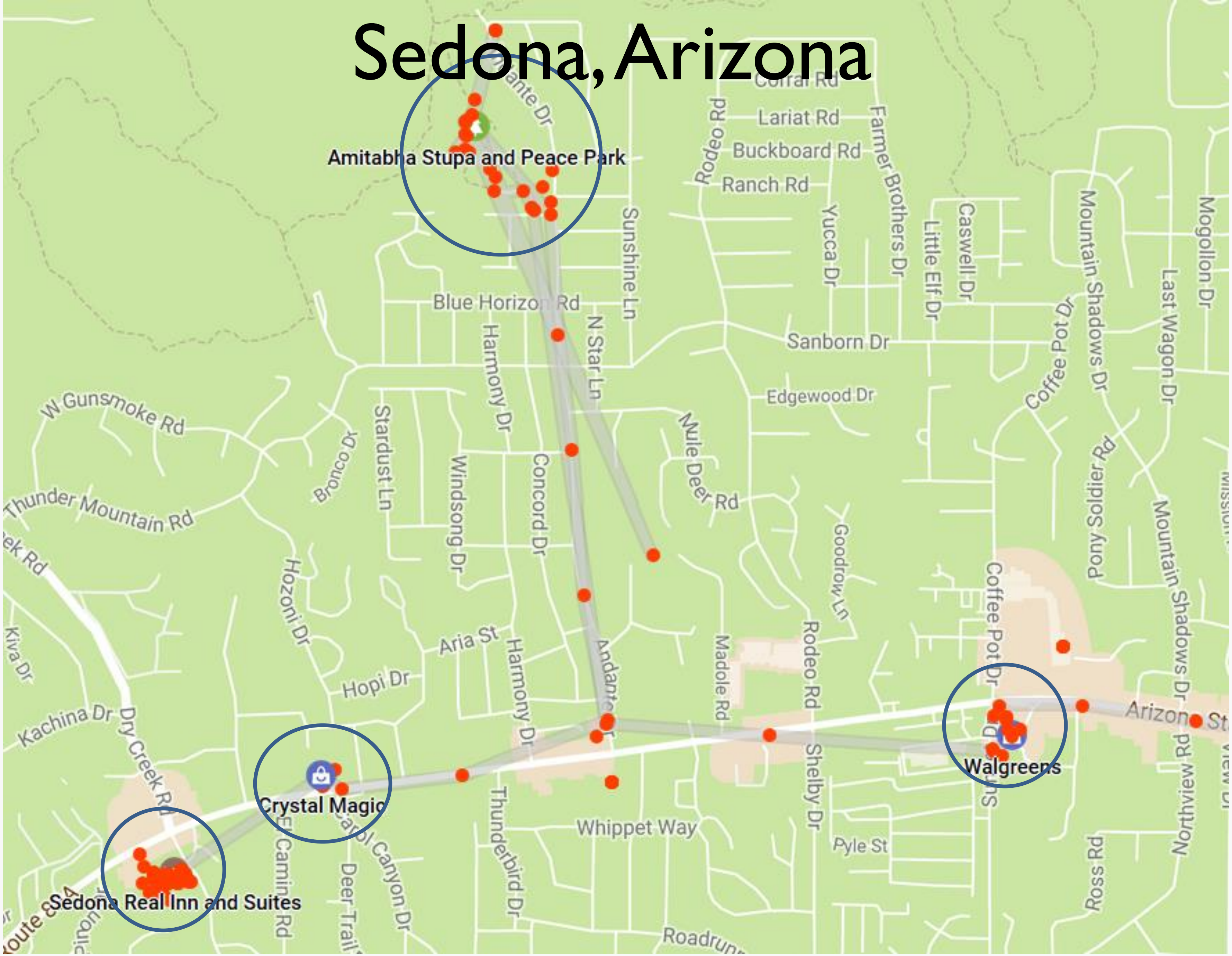
Grand Canyon



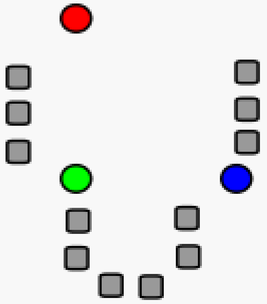
Sedona, Arizona



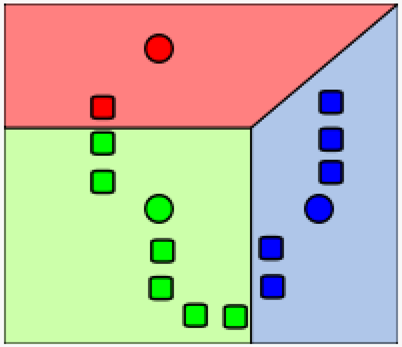
Sedona, Arizona



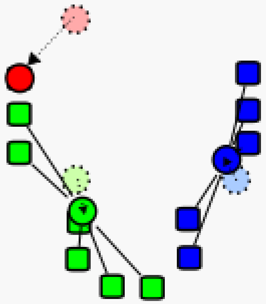
K-Means



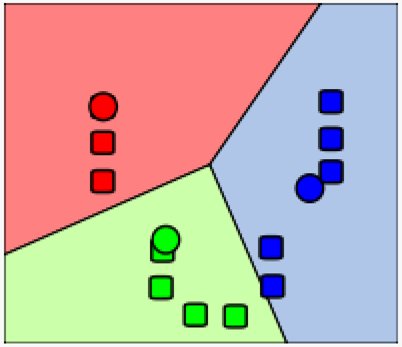
Randomly pick K = 3 points (initial centroids)



Assign each point to its closest centroid



Using points in clusters, calculate new centroids



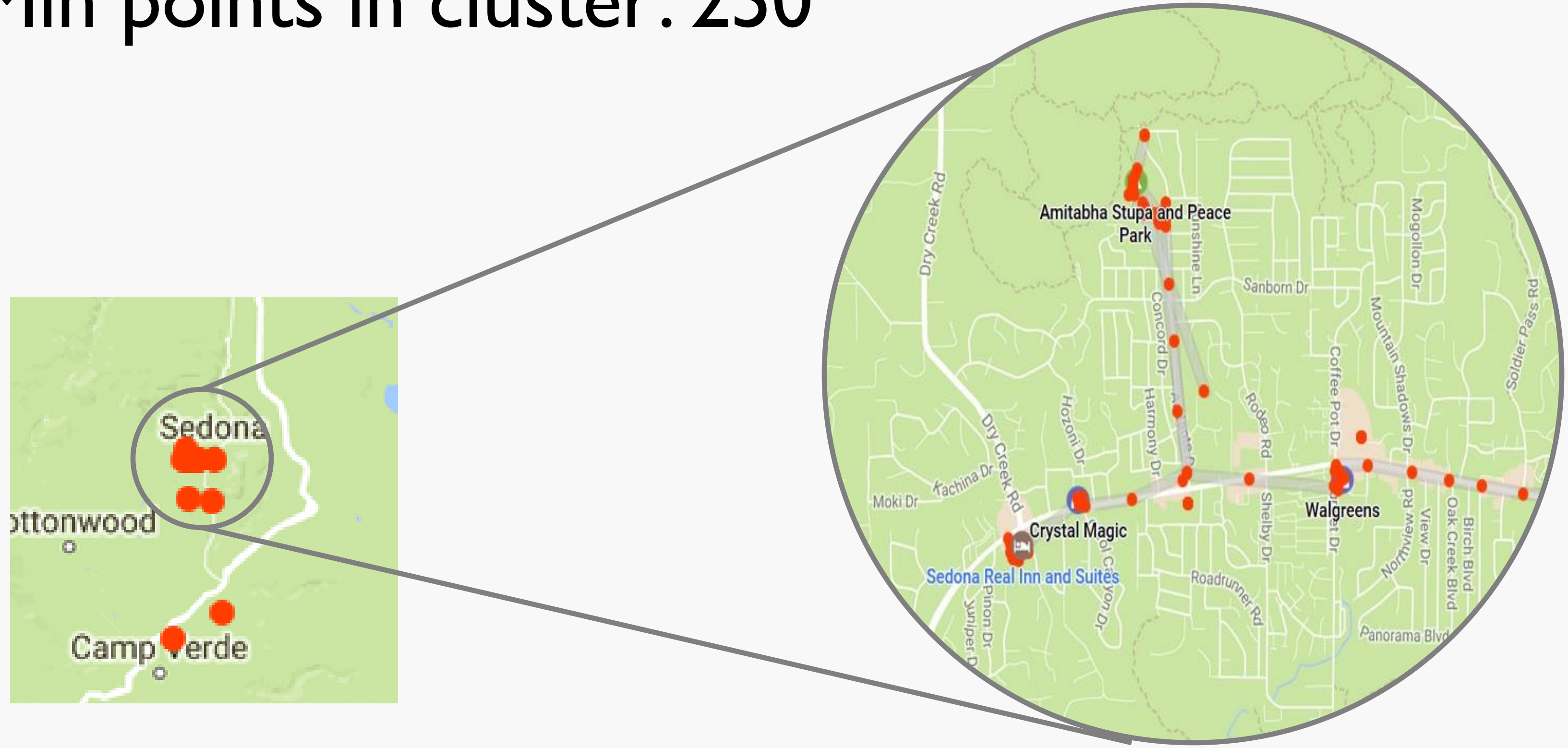
Assign each point to its closest centroid

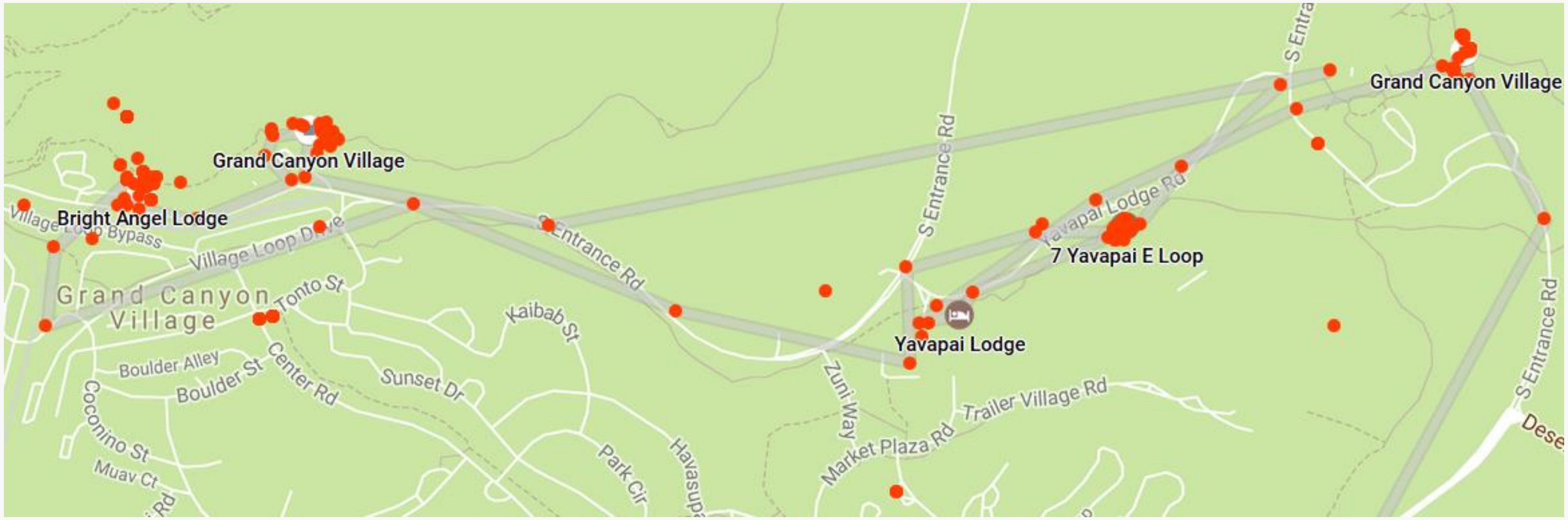
K_Means_Example_Step_1.svg,
K_Means_Example_Step_2.svg,
K_Means_Example_Step_3.svg,
K_Means_Example_Step_4.svg

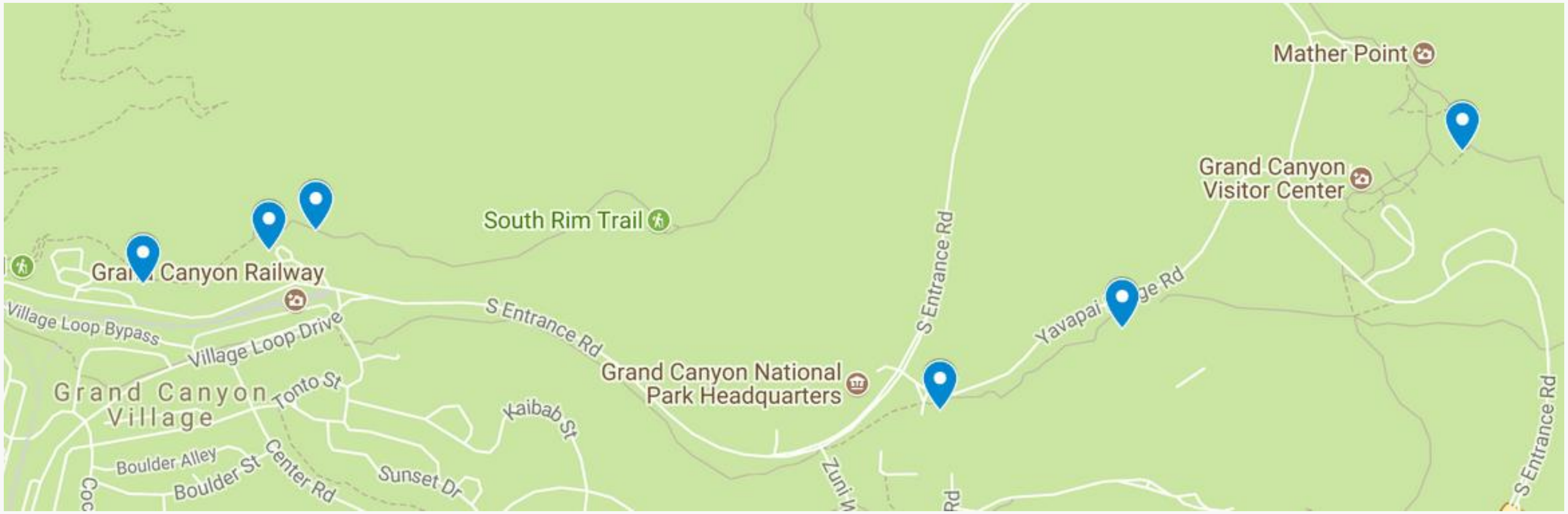
by Weston.pace are licensed under CC BY-SA 3.0

Recursive K-Means

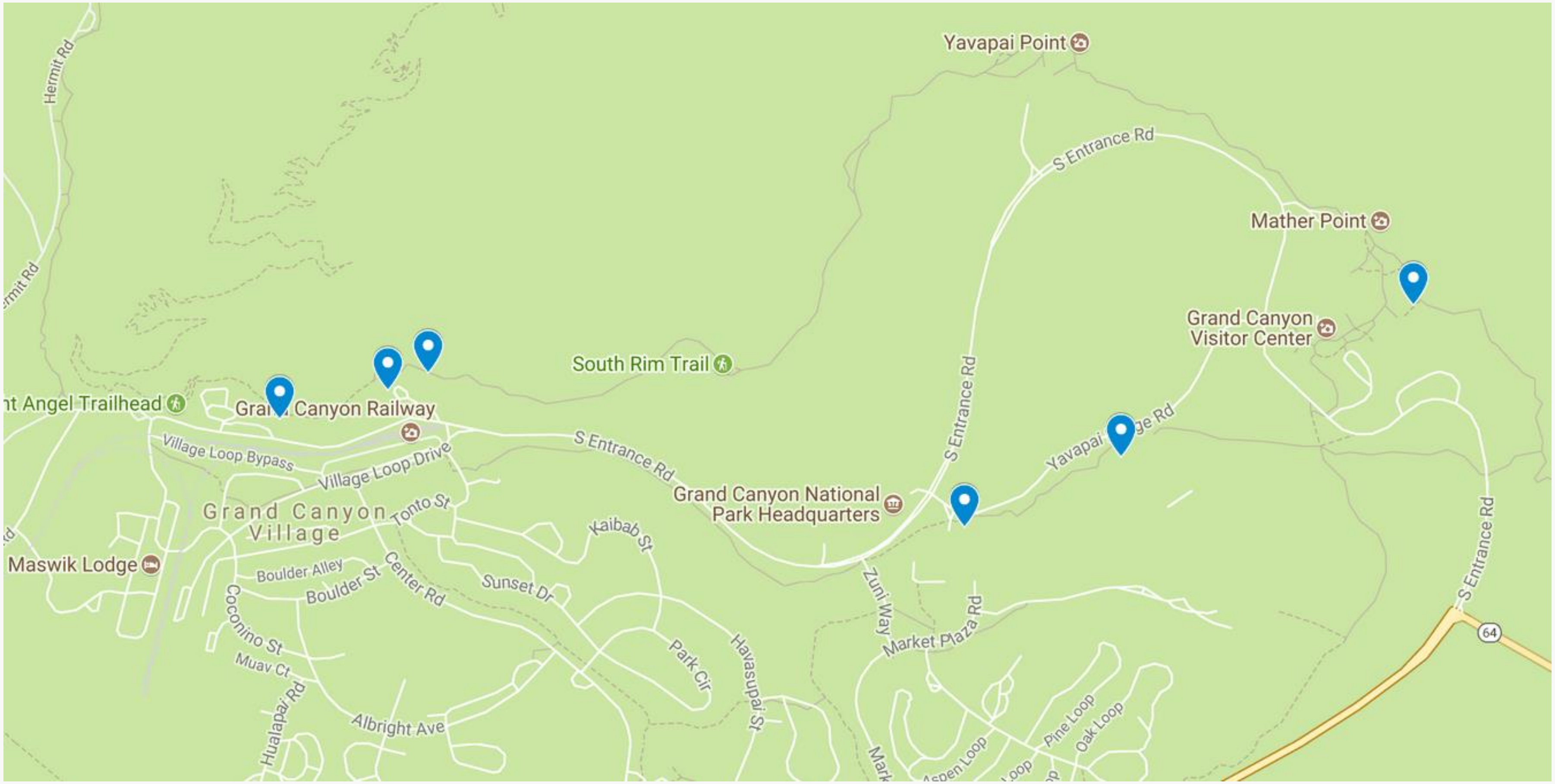
- Max radius: 0.1 miles
- Min points in cluster: 250



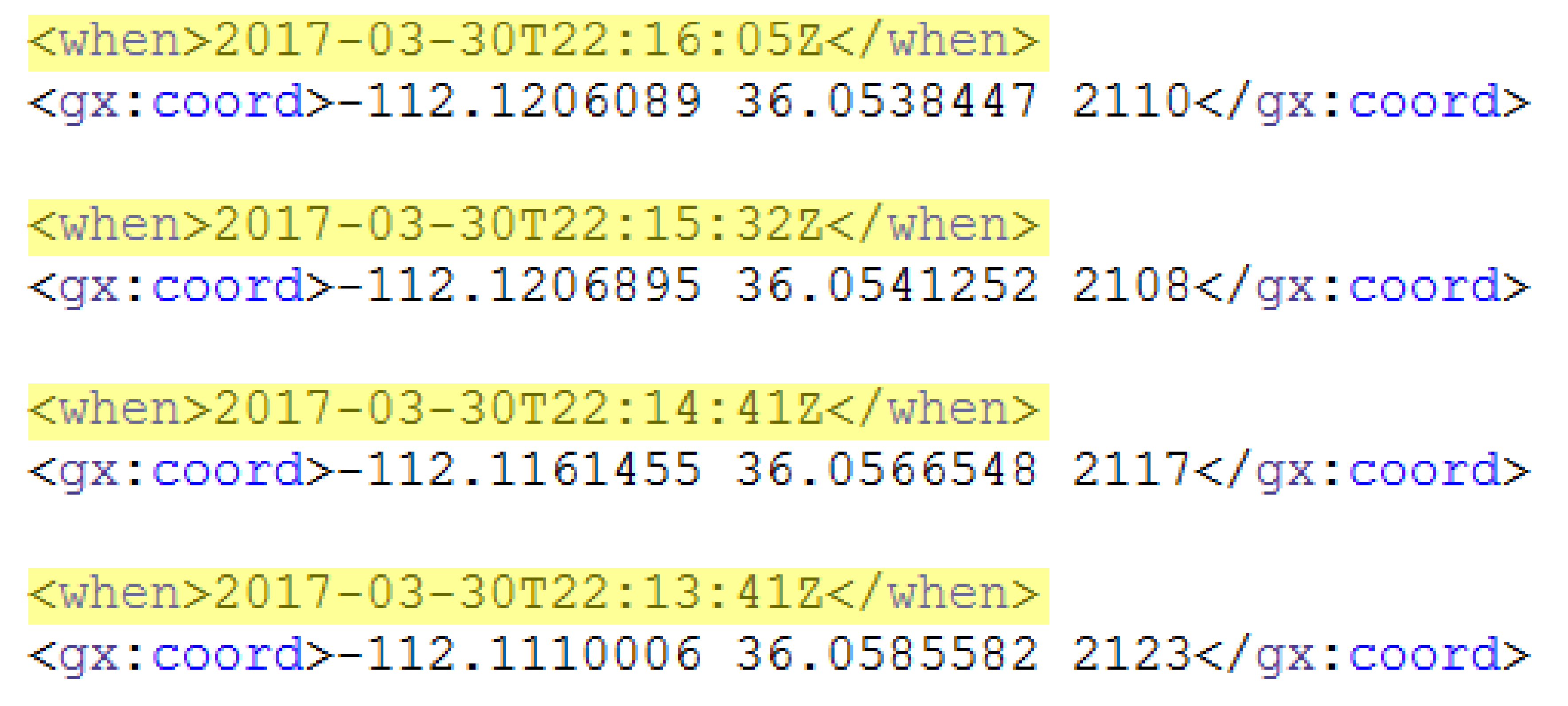

Grand Canyon Village
Bright Angel Lodge
Grand Canyon Village
Grand Canyon Village
7 Yavapai E Loop
Yavapai Lodge
Village Loop Bypass
Village Loop Drive
S Entrance Rd
Yavapai Lodge Rd
Tonto St
Kaibab St
Boulder Alley
Boulder St
Center Rd
Sunset Dr
Coconino St
Muav Ct
Park Cir
Havasupai
Zuni Way
Market Plaza Rd
Trailer Village Rd
S Entrance Rd
S Entrance Rd


Mather Point
Grand Canyon Visitor Center
South Rim Trail
Grand Canyon Railway
Village Loop Bypass
Village Loop Drive
S Entrance Rd
Grand Canyon National Park Headquarters
S Entrance Rd
Yavapai Lodge Rd
Grand Canyon Village
Tonto St
Kaibab St
Boulder Alley
Boulder St
Center Rd
Sunset Dr
Zuni Way
S Entrance Rd

Now what?


Yavapai Point
S Entrance Rd
Mather Point
Grand Canyon Visitor Center
South Rim Trail
nt Angel Trailhead
Gra Canyon Railway
Village Loop Bypass
Village Loop Drive
Grand Canyon Village
Tonto St
Grand Canyon National Park Headquarters
Yavapai ge Rd
Maswik Lodge
Boulder Alley
Boulder St
Center Rd
Kaibab St
Sunset Dr
Coconino St
Muav Ct
Hualapai Rd
Albright Ave
Park Cir
Havasupai St
Zuni Way
Market Plaza Rd
Aspen Loop
Pine Loop
Oak Loop
Hermit Rd
64

Time



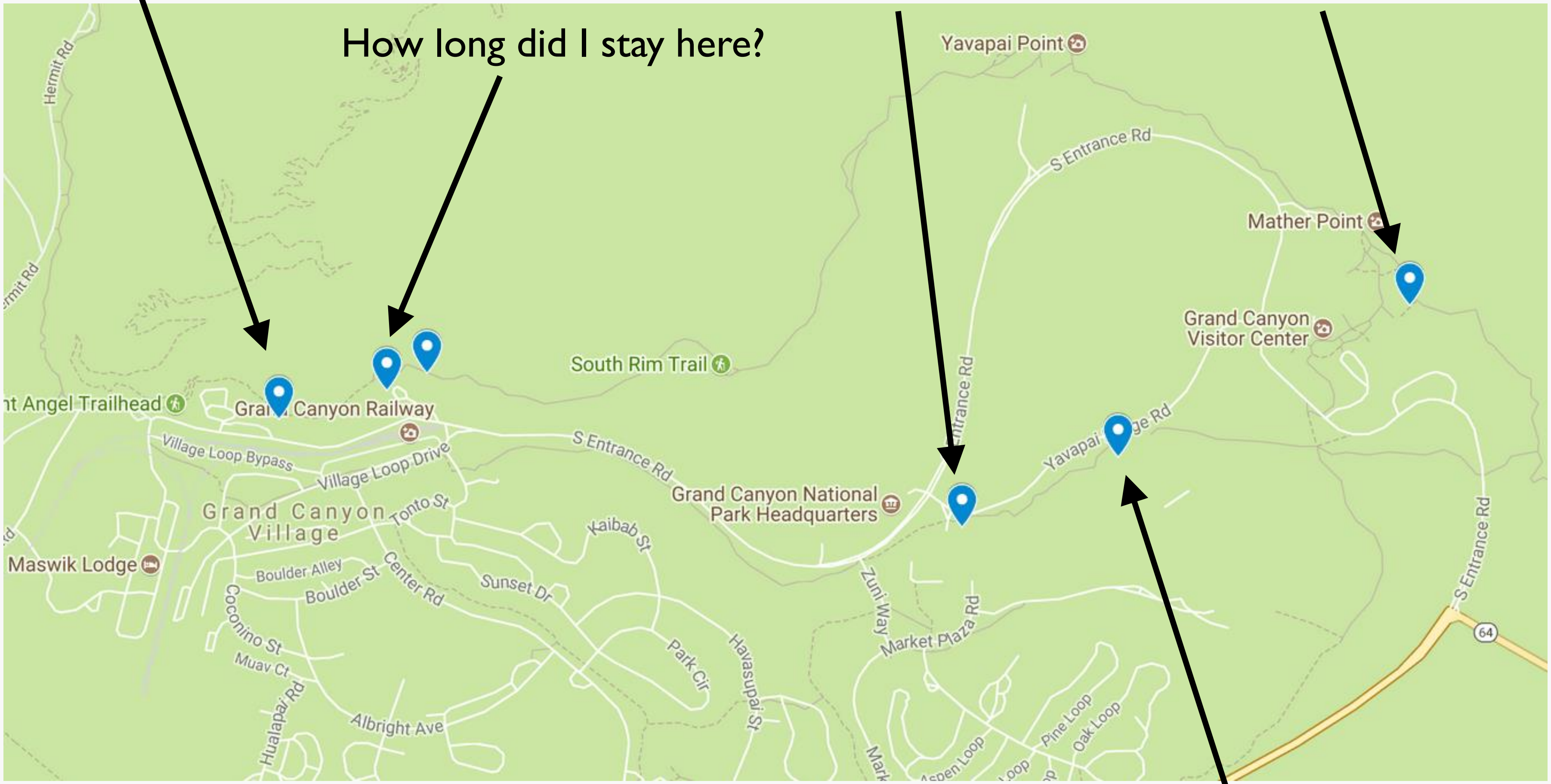
```
<when>2017-03-30T22:16:05Z</when>
<gx:coord>-112.1206089 36.0538447 2110</gx:coord>

<when>2017-03-30T22:15:32Z</when>
<gx:coord>-112.1206895 36.0541252 2108</gx:coord>

<when>2017-03-30T22:14:41Z</when>
<gx:coord>-112.1161455 36.0566548 2117</gx:coord>

<when>2017-03-30T22:13:41Z</when>
<gx:coord>-112.1110006 36.0585582 2123</gx:coord>
```

How often do I visit this location?

Am I usually here on weekends, weekdays, or both?

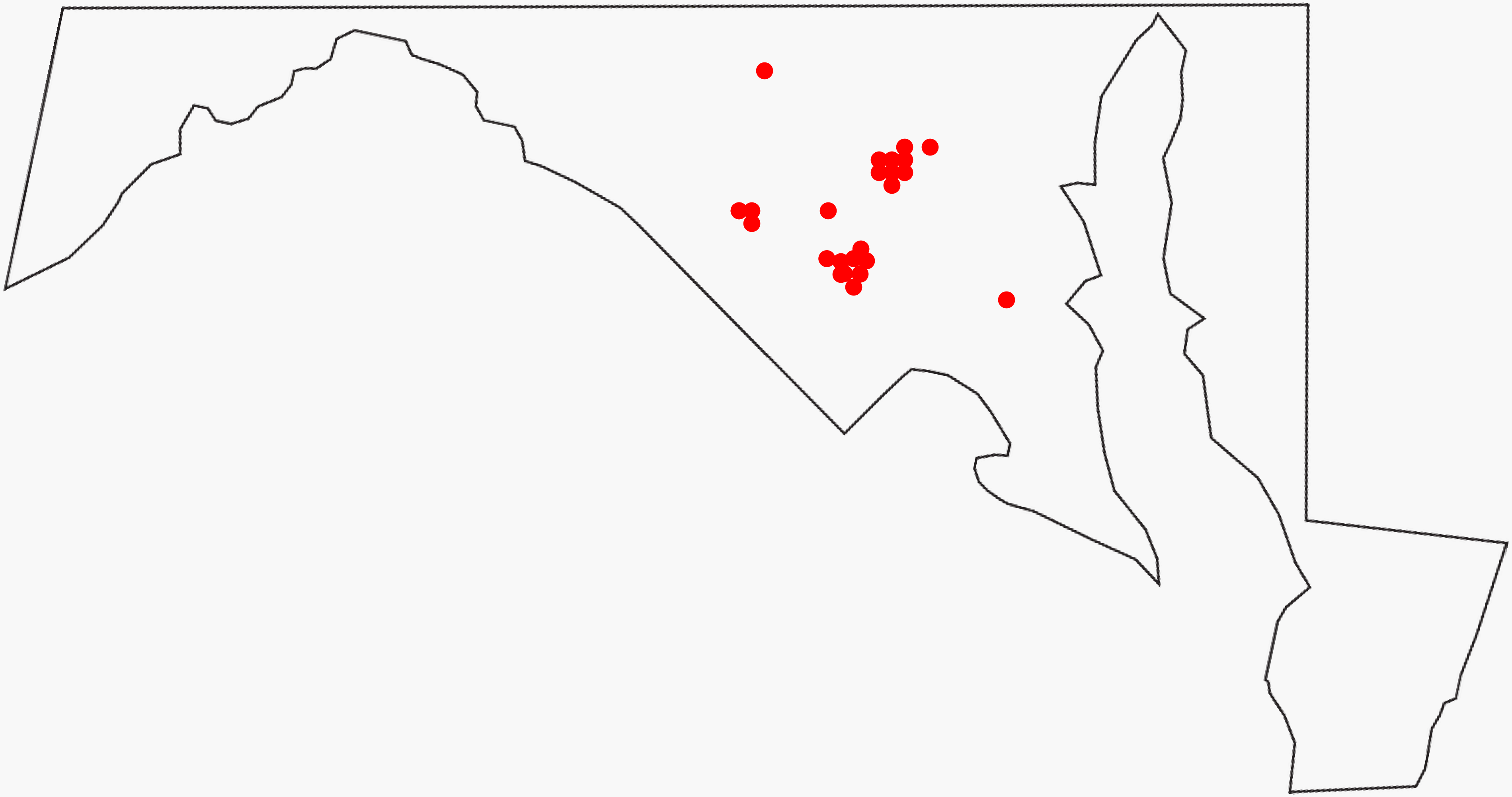
When was I last here?

How long did I stay here?

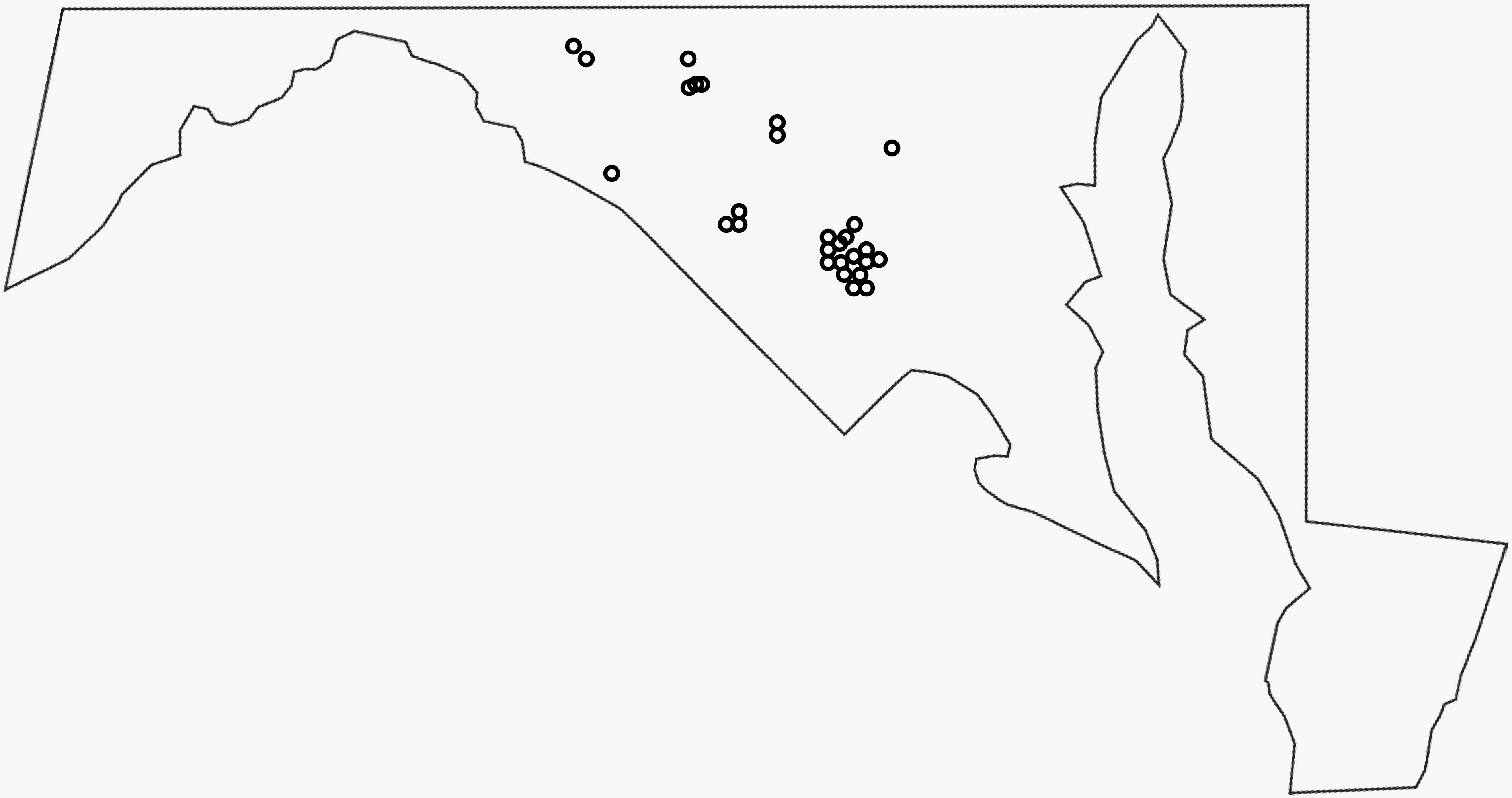


What times of the day can I be found here?

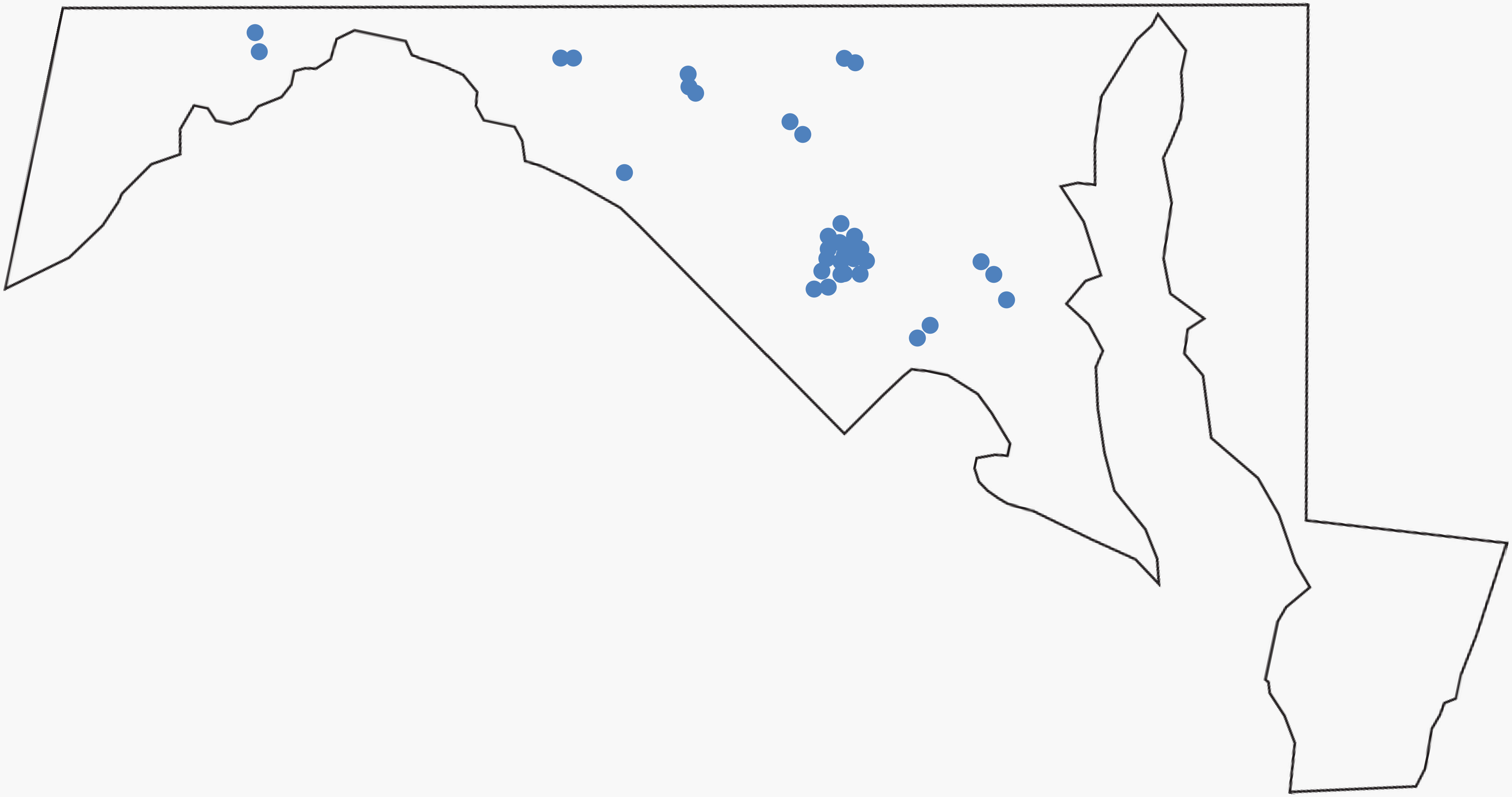
Weekday Day Points



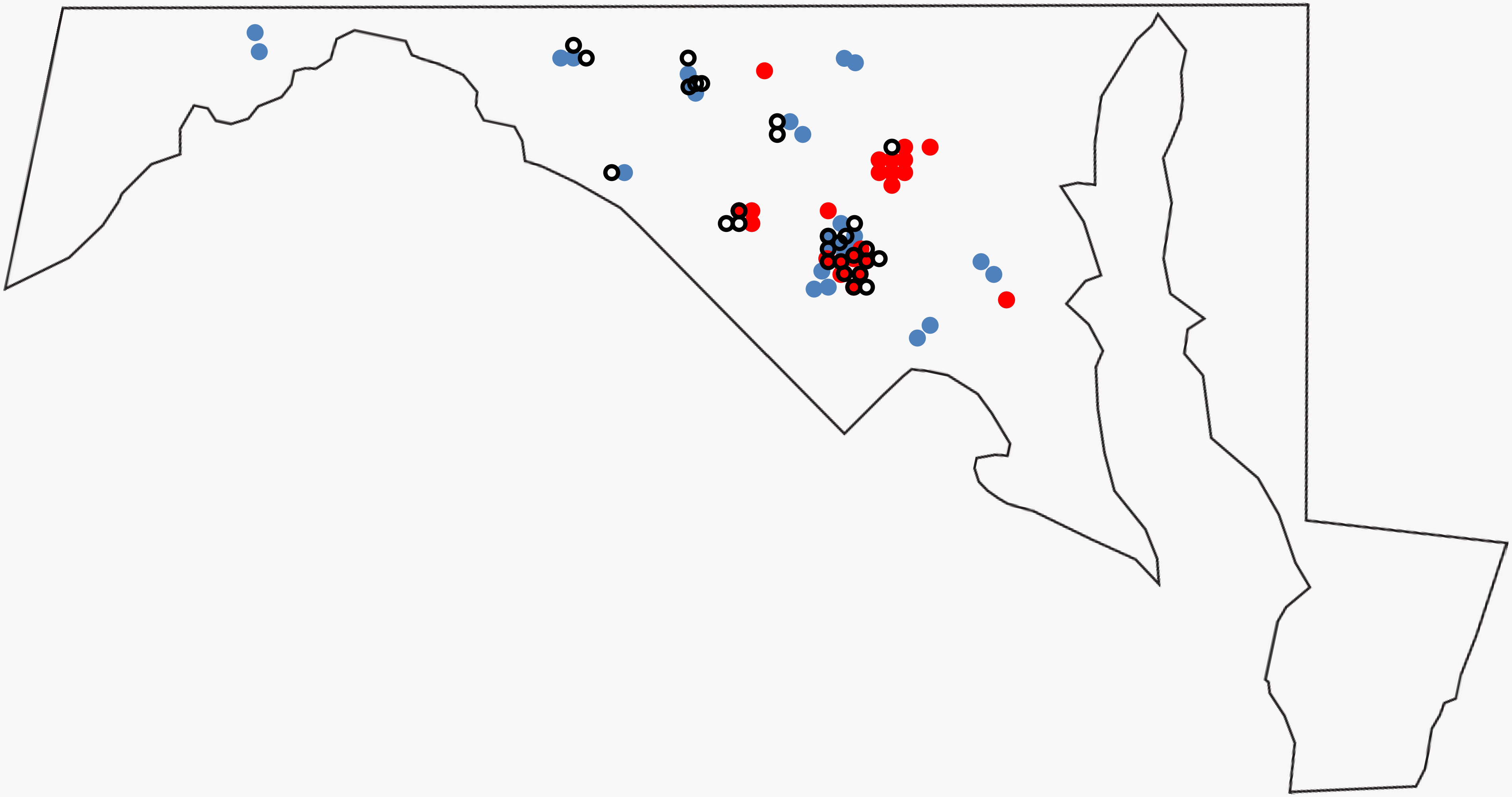
Weekday Evening Points



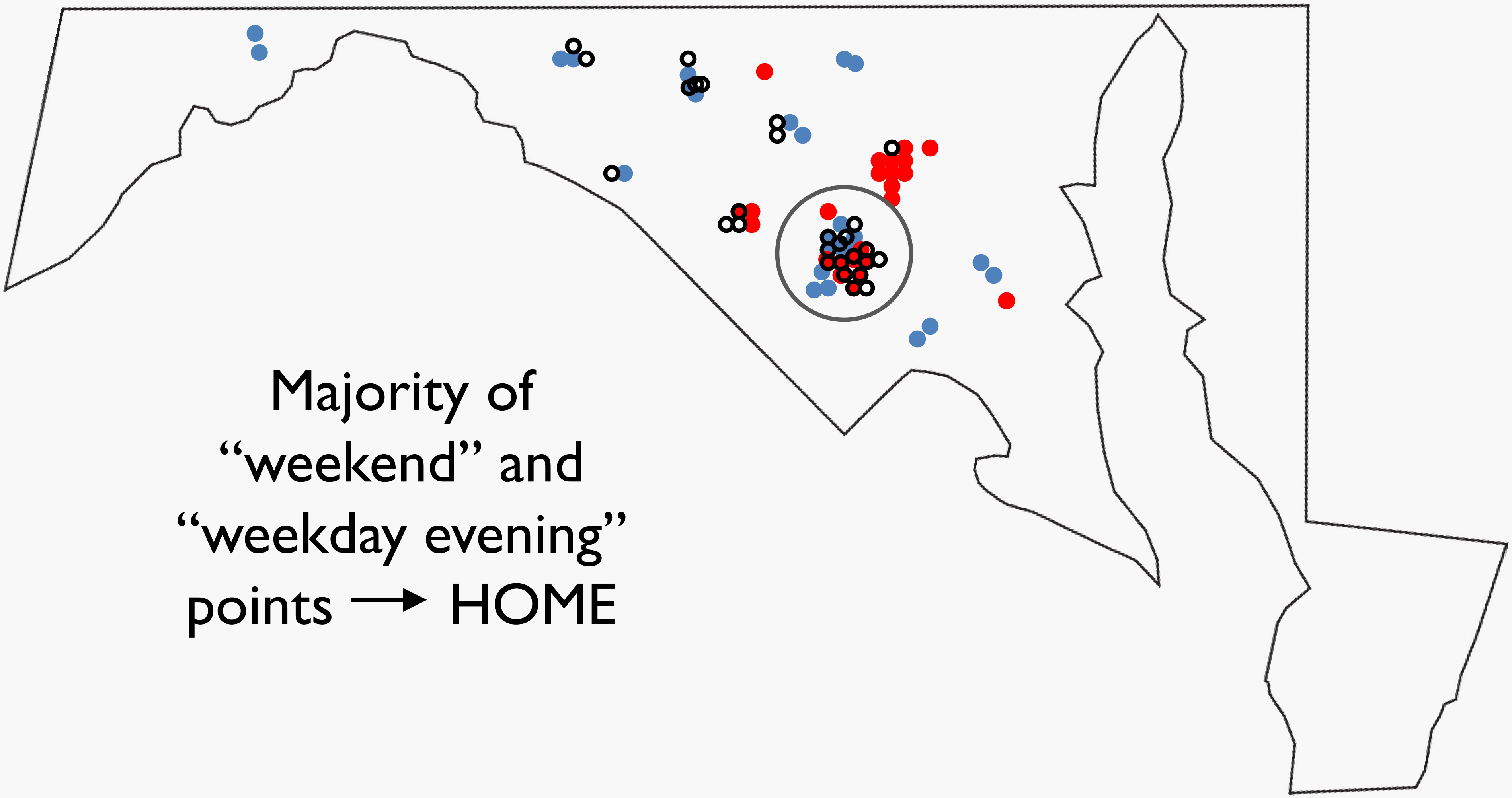
Weekend Points



All Points



All Points



Majority of “weekend” and “weekday evening” points → HOME

If I Know Where "Home" Is, I Can Calculate…

- At a specific time / day of week, probability of being:
  - home
  - away from home

If I Know Where "Home" Is, I Can Calculate…

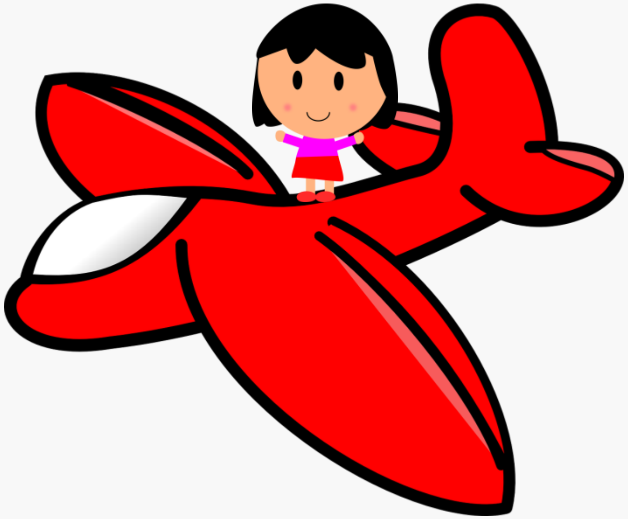
- Average local travel radius
  - Use points when I was home at **both** the beginning/end of the day
- Max local distance I've driven

If I Know Where "Home" Is, I Can Calculate…

- Distance from home for any recorded data point
  - Assume traveling if:
    - Not home that day
    - Home only at beginning/end of day

Significant Location Types Identified

HOME

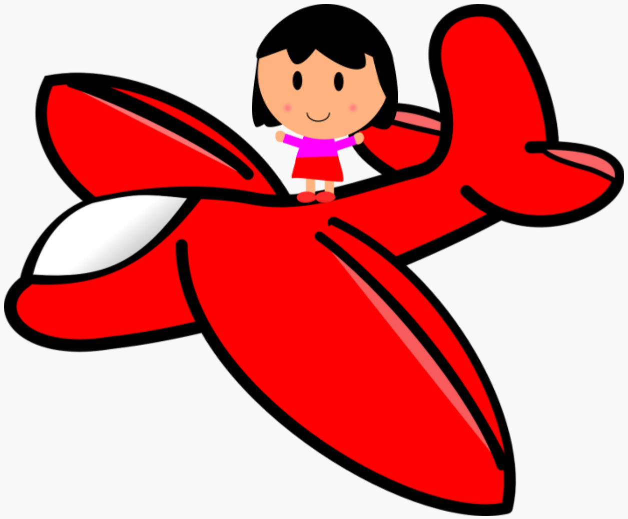


AWAY
FROM
HOME

LOCAL

Away From Home Locations

- Hotels (Vacation, Conferences)
- Tourist Venues



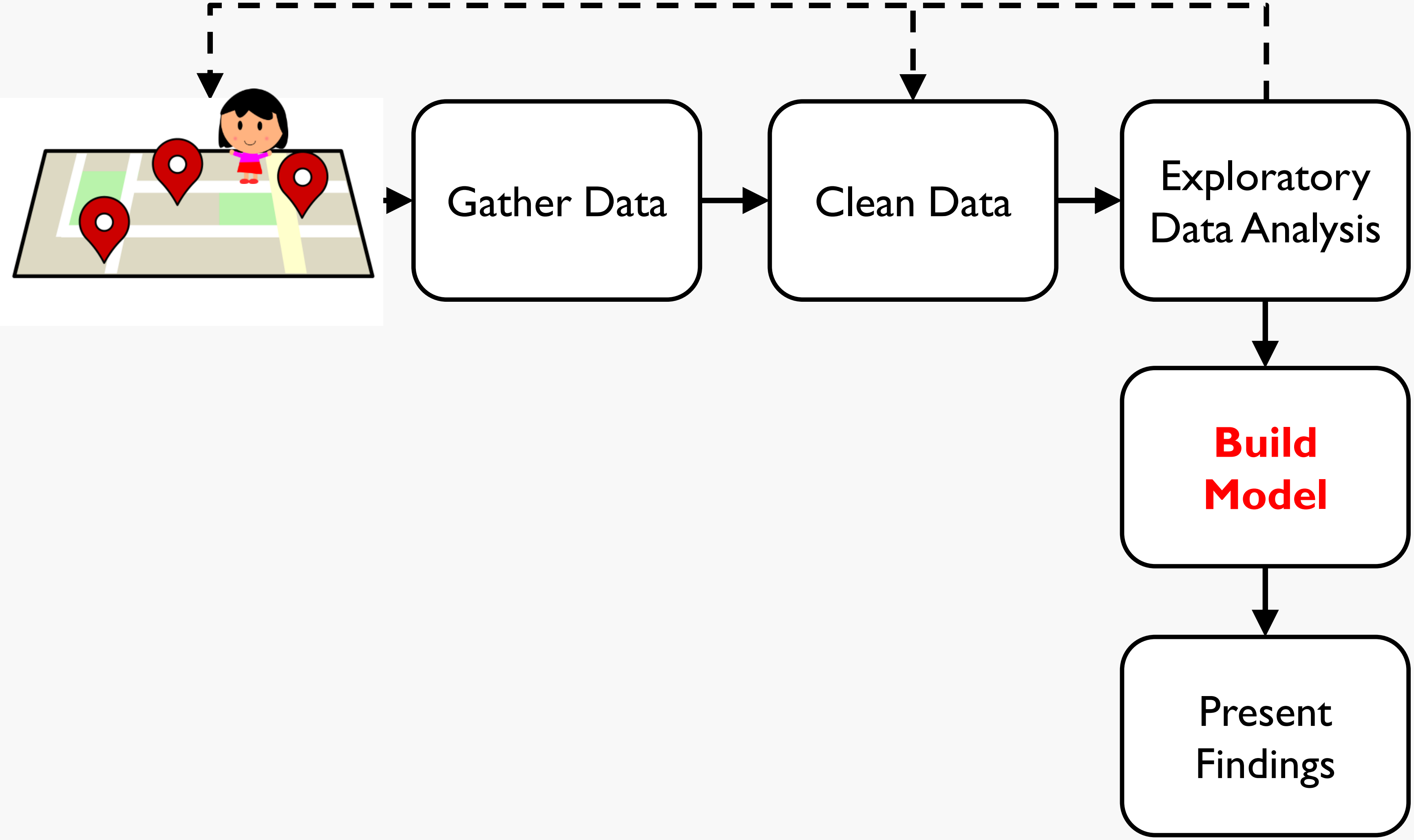
Local Locations

- Favorite Breakfast / Lunch / Dinner Spots
- Grocery Store
- Running Trails

Specific Local Locations

- *Work*: If I'm there on weekdays >= 5 hours
- *Weekend*: If I'm only there on Sat / Sun
  - Concert Venue
- *Same Day*: If I'm only there on a specific day of the week
  - Farmer's Market
  - Trivia Night

Data Science Process



Gather Data

Clean Data

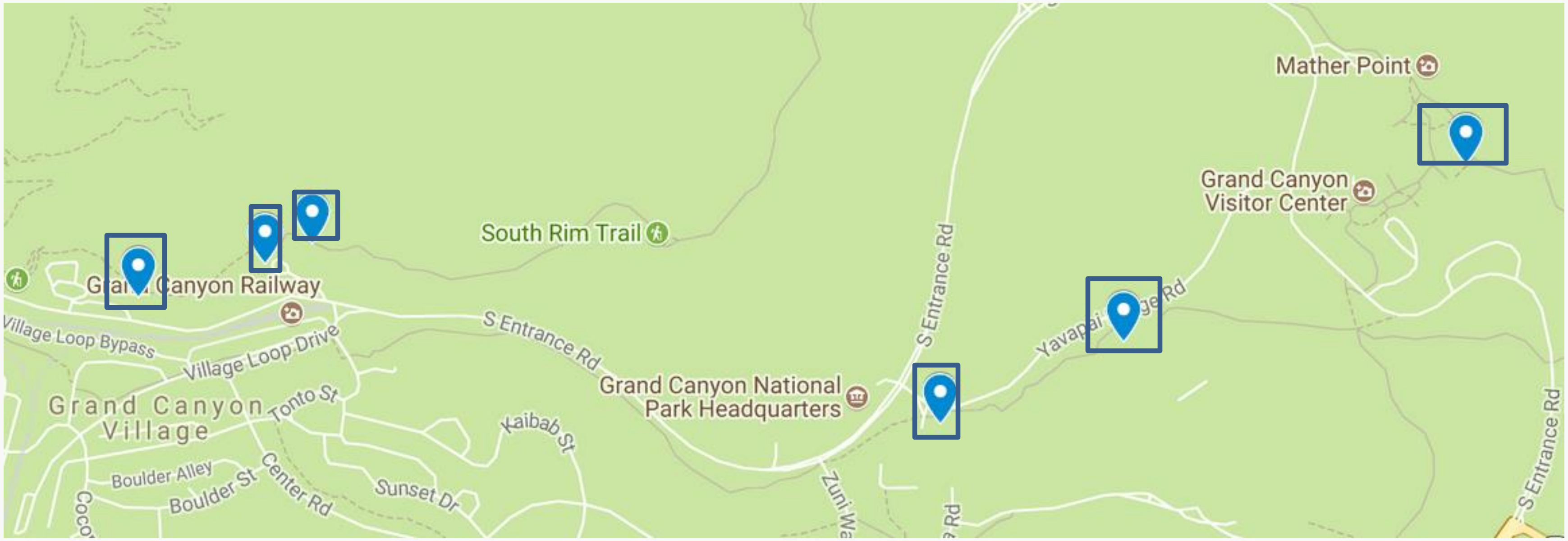
Exploratory Data Analysis

Build Model

Present Findings

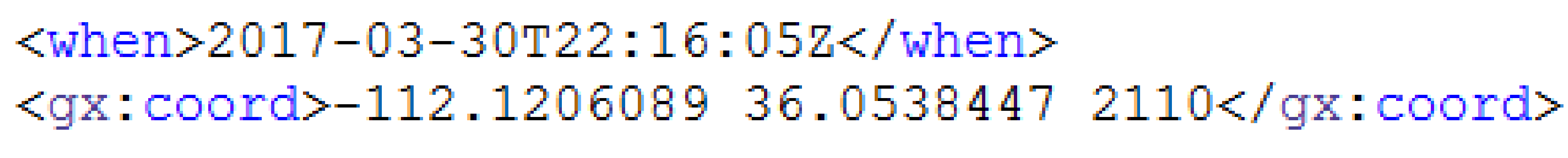
Significant Location Details

- Lat / Long Boundaries
- Location Type
  - *AwayFromHome*
  - *Local (home, work, weekend, sameDay)*



Data Point Details

Original



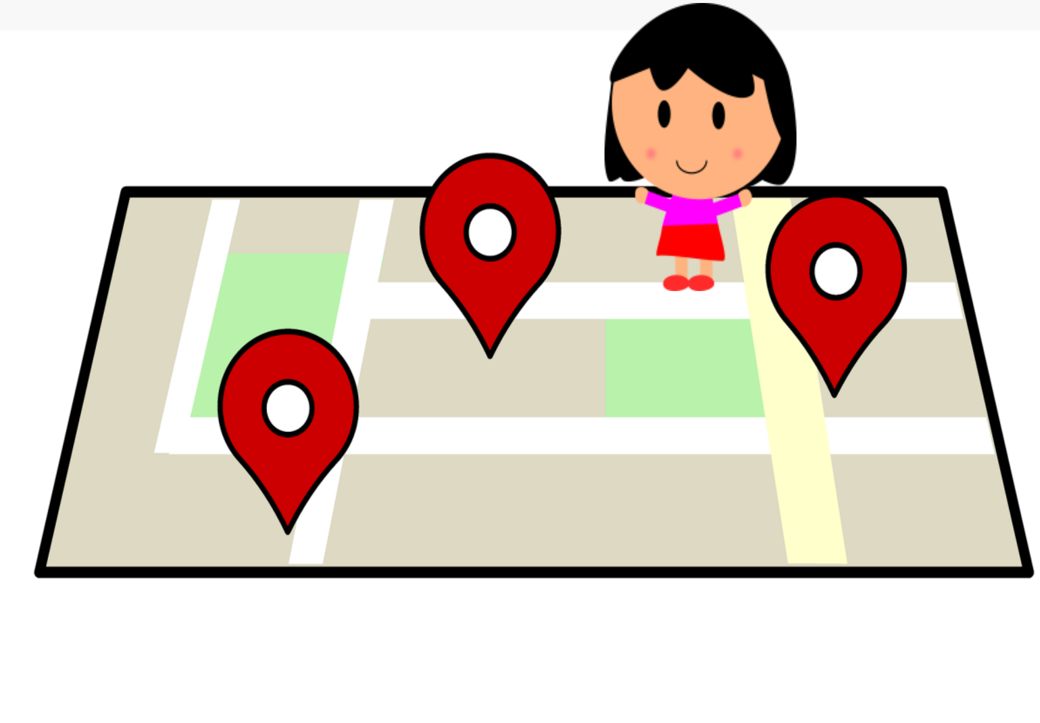
```
<when>2017-03-30T22:16:05Z</when>
<gx:coord>-112.1206089 36.0538447 2110</gx:coord>
```

New

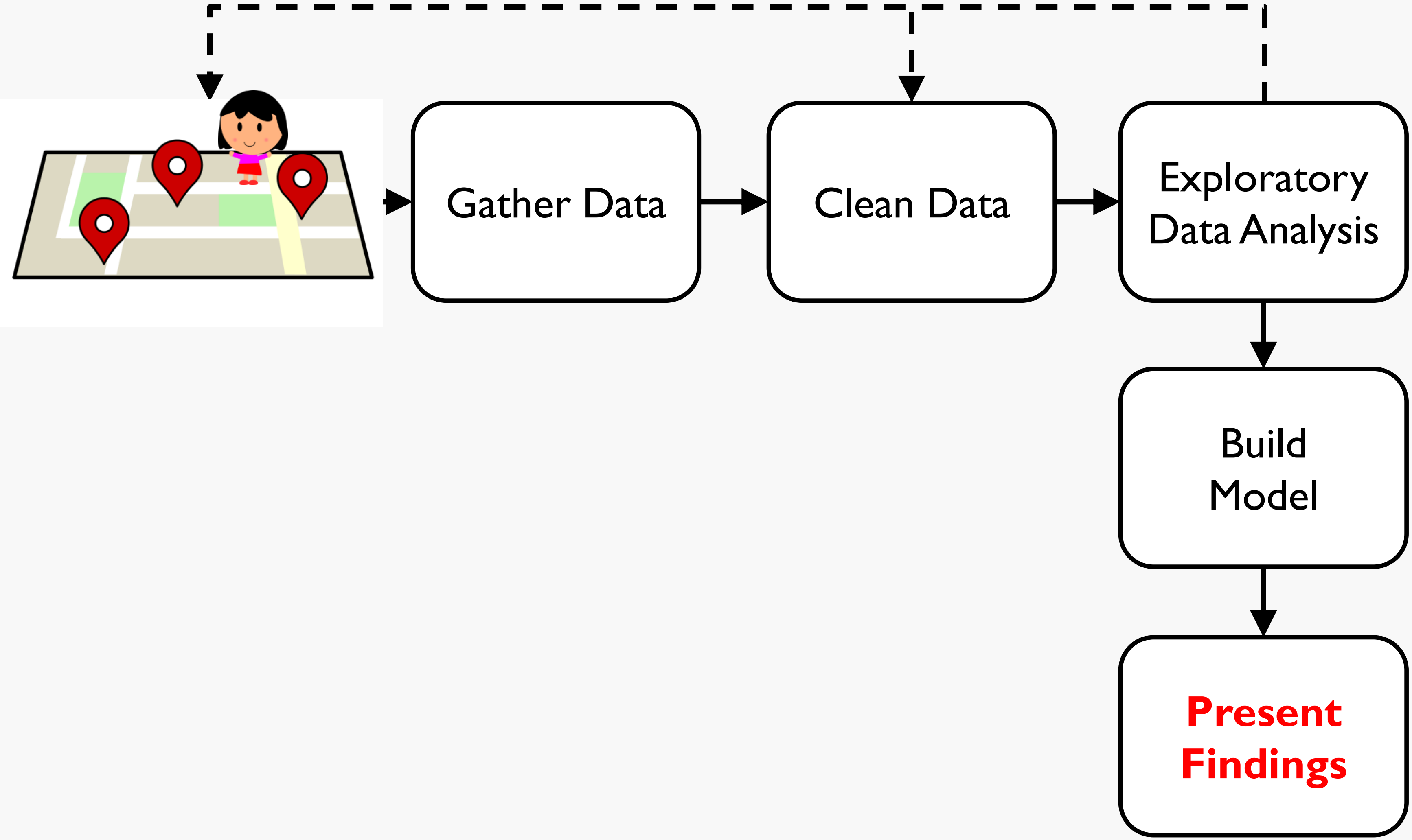
```
<distanceFromHome>1949.46</distanceFromHome>
<locationLabel>cluster3</locationLabel>
<description>awayFromHome</description>
```

Model of Me

- Dates away from home
- Local travel radius
- Likelihood of being at a location by day / time
- Significant locations



Data Science Process



Gather Data

Clean Data

Exploratory Data Analysis

Build Model

Present Findings

DEMO

Questions I can ask the data

- Where was I on August 9, 2017 at 2:18PM ?
- Predict where I will be on Monday at 8:45AM.
- Predict when I am likely to be away on Saturday.
- Predict whether I'll be home on Sunday at 10PM.

Expanded Questions

- How many days was I out of town in July?
- When was I at work on a weekend?
- How many times did I visit the grocery store last month?
- How long does it usually take to drive to work?
- When was I last at the Grand Canyon?

DISCUSSION

Assumptions

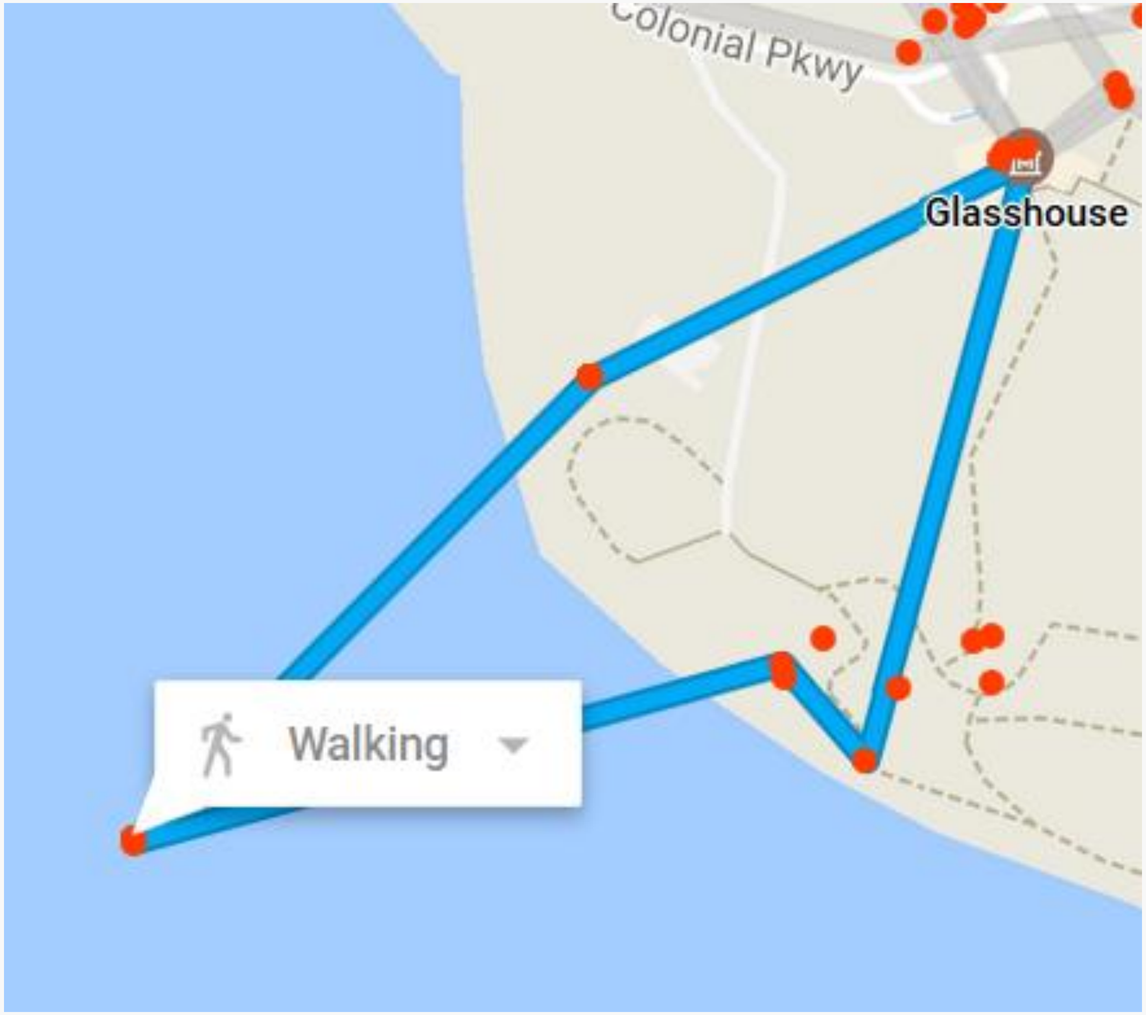
- Regular schedule
- “Normal” work habits

- Home
  - More often than anywhere else
  - More often on weekday evenings & weekends

When This Doesn’t Work

- Irregular schedule / lots of travel
- Not enough points
- Bad technology
  - signal
  - hardware

Bad Technology



Cautions

- Analysis is a general pattern of behavior

- Locations may be inaccurate (Google itself asks for corrections)

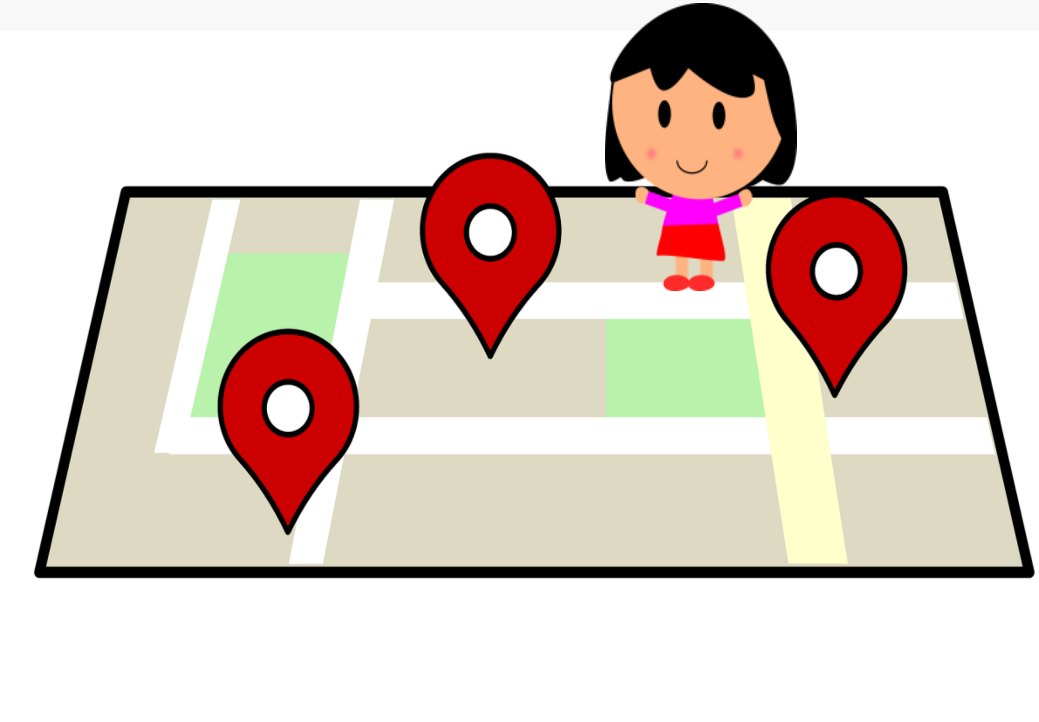
- A long traffic light can be a "location"

Google already asks for your Home and Work addresses…

…which means that they already know your significant locations!

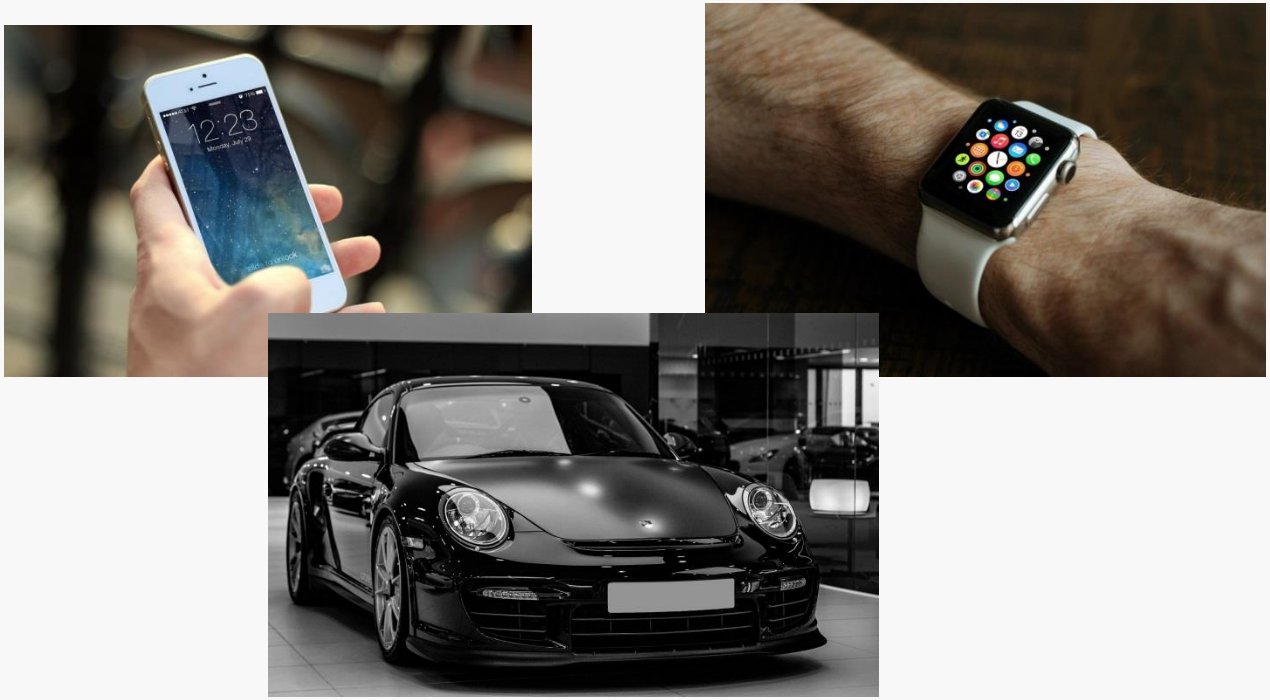
Would you share this info?

- Dates away from home
- Local travel radius
- Likelihood of being at a location by day / time
- Significant locations

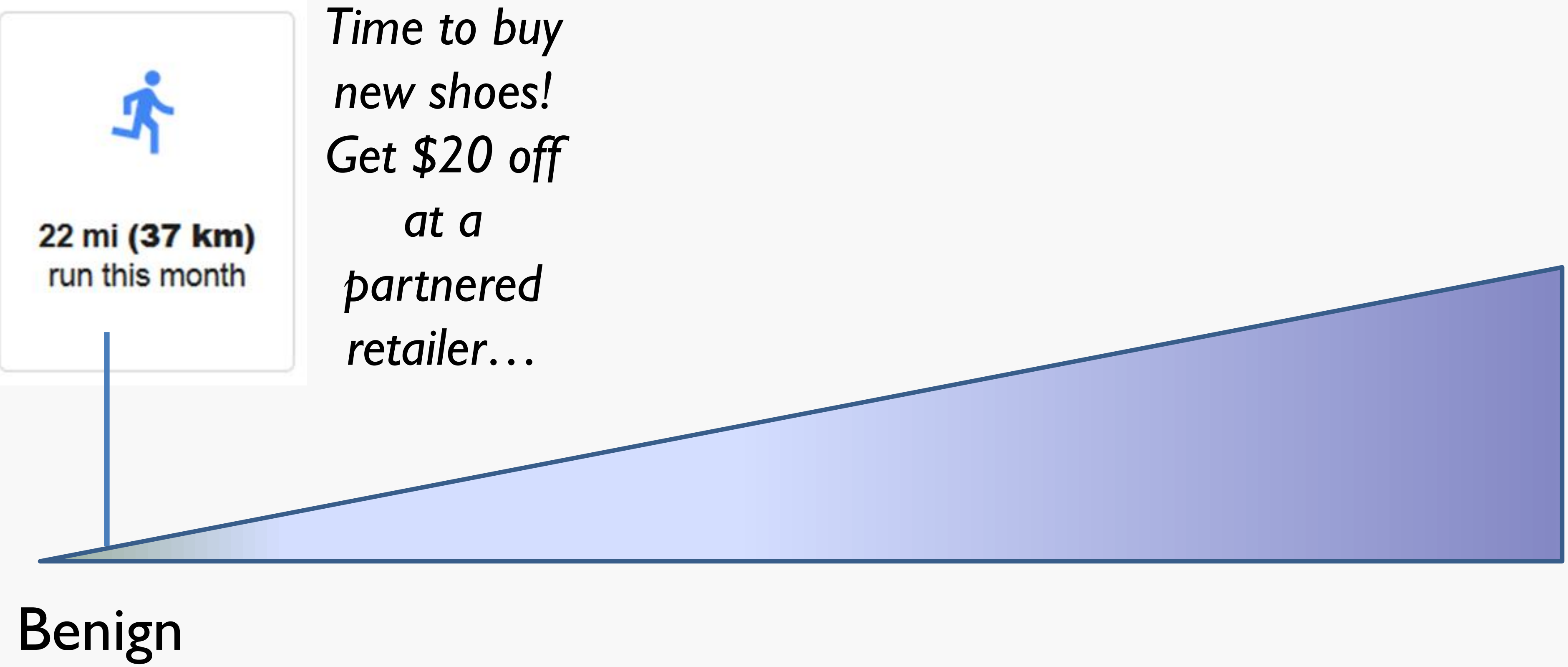


Who could have it

- Products and apps
- Companies that access data
- Companies that buy / share data



Implications

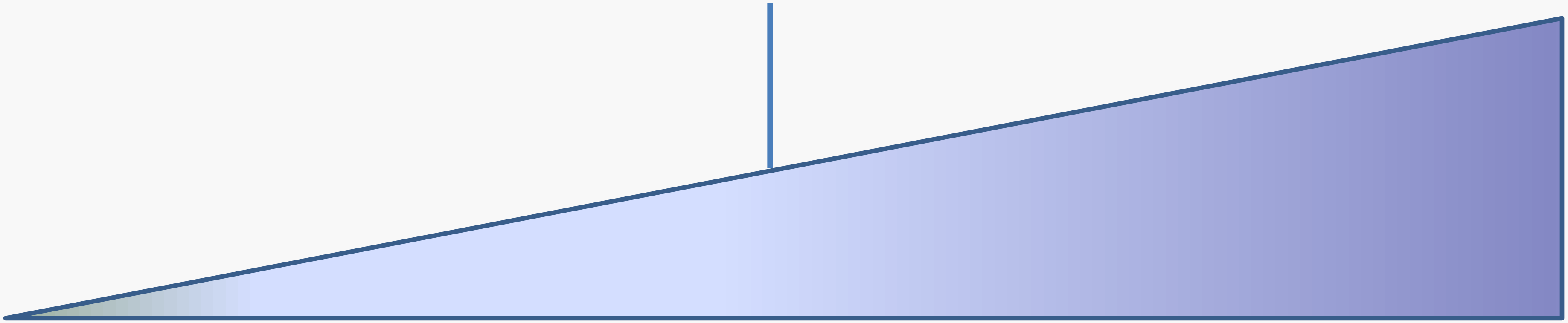


22 mi (37 km)
run this month

Time to buy new shoes! Get $20 off at a partnered retailer…

Implications

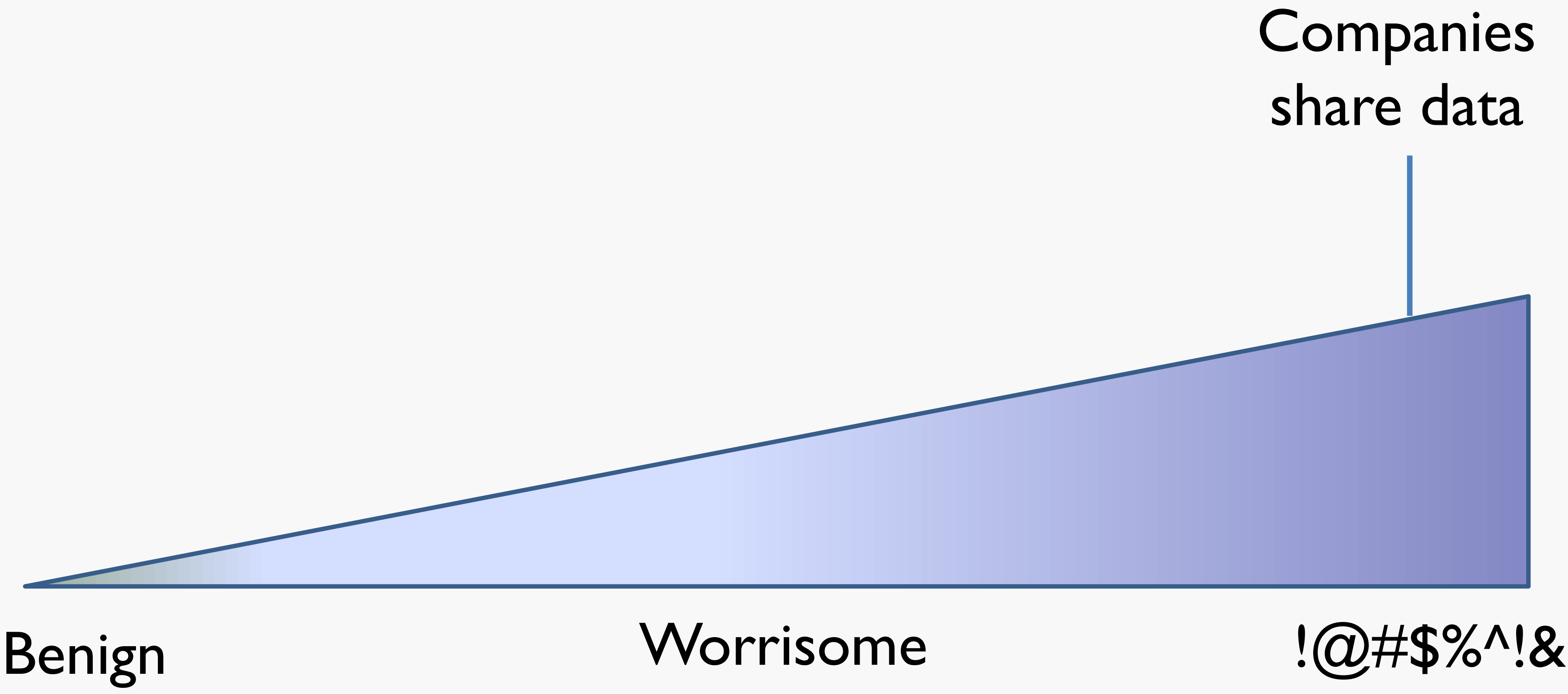
Your insurance claim was denied due to…



Benign

Worrisome

Implications

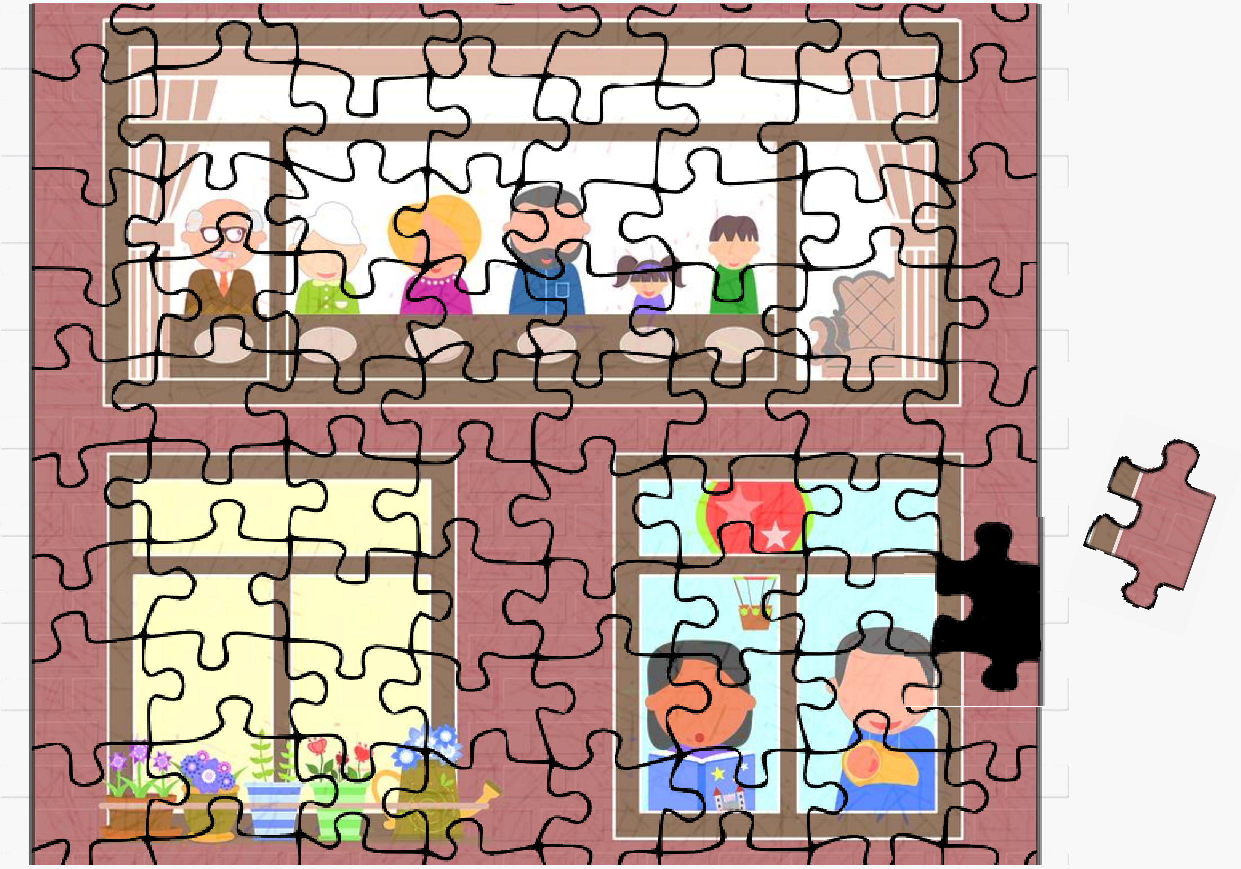


Companies share data

Benign

Worrisome

!@#$%^!&



Your Data, Your Choice

Further Information

- Code (Jupyter Notebook)

  https://github.com/laconicllama

- Contact

  laconicllama@hotmail.com

Questions?