OK Google, Tell Me About Myself

Lisa Chang

About Me

- ~~Process Contact Engineer~~
- ~~Process Development Engineer~~
- ~~Applications Engineer~~
- ~~Scientist~~
- Software Engineer
- Data Scientist

Equifax data breach may affect nearly half the US population

CNET (September 7, 2017)

names
social security numbers
birthdates
addresses

Deloitte Gets Hacked: What We Know So Far

Fortune (September 25, 2017)

confidential info
usernames
passwords
IP addresses

Uber Hid 2016 Breach, Paying Hackers to Delete Stolen Data

The New York Times (November 21, 2017)

names
e-mail addresses
telephone numbers

What Amazon Wants From Whole Foods: Data on Shopping Habits

Voice of America (June 29, 2017)

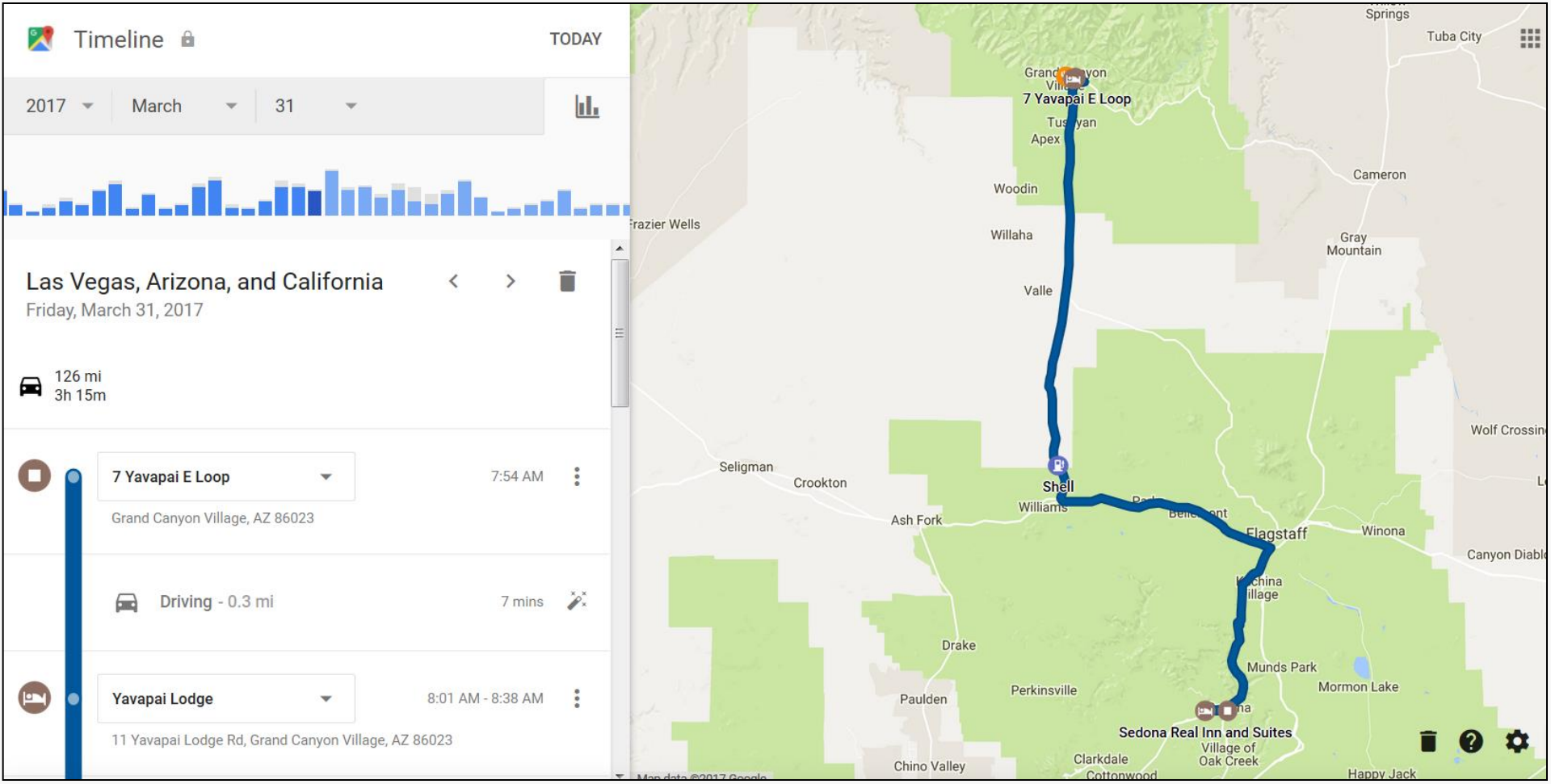
Your Roomba May Be Mapping Your Home, Collecting Data That Could Be Shared

The New York Times (July 25, 2017)

GOOGLE MAPS SUPERCHARGES LOCATION SHARING, BEGINS DROOLING OVER YOUR DATA

Wired (March 22, 2017)

Google Timelines



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.

EXPLORE YOUR TIMELINE

8 cities visited this month

28 places visited this month (5 new)

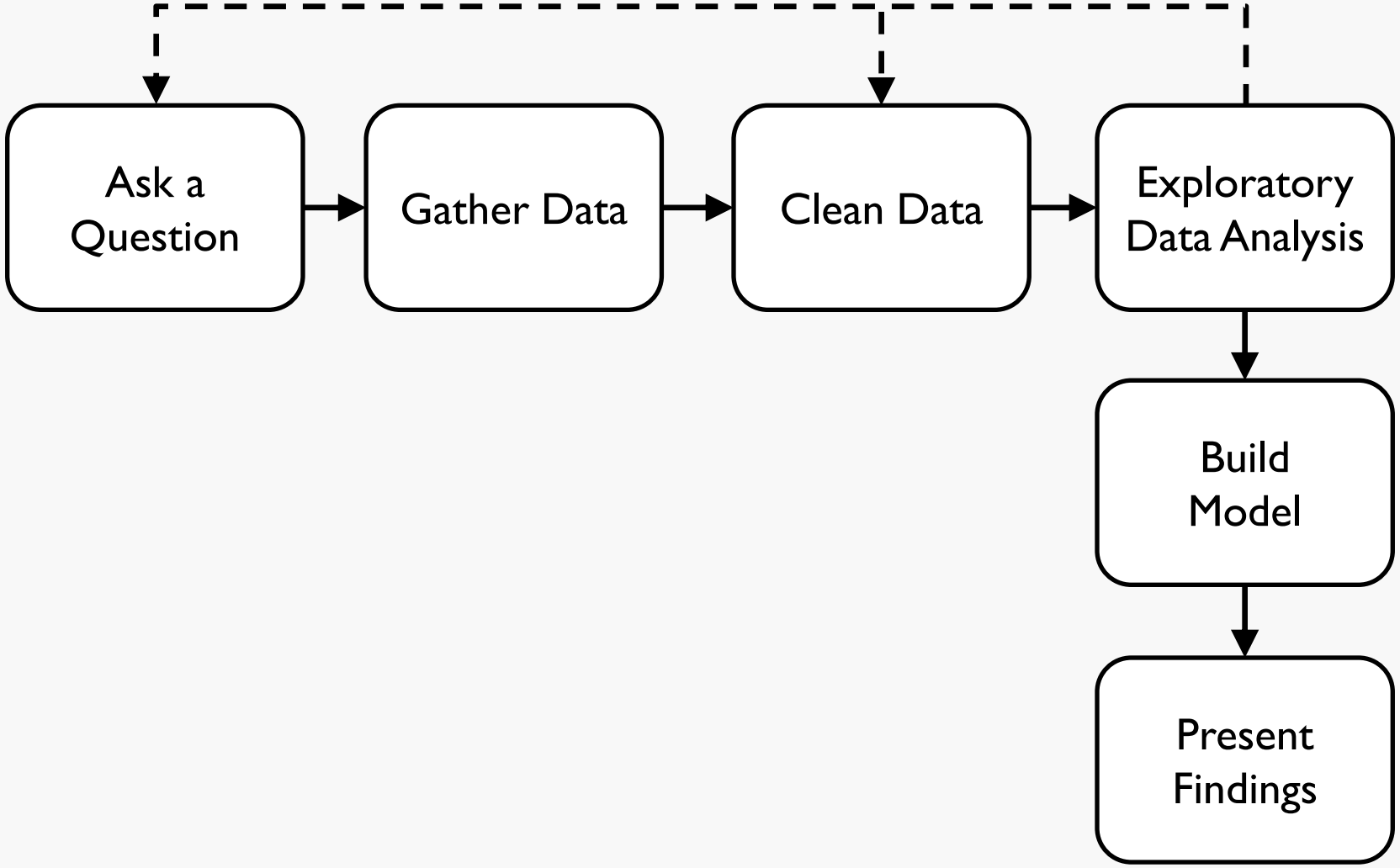
Your activity in timeline

6 mi (11 km)
walked this month

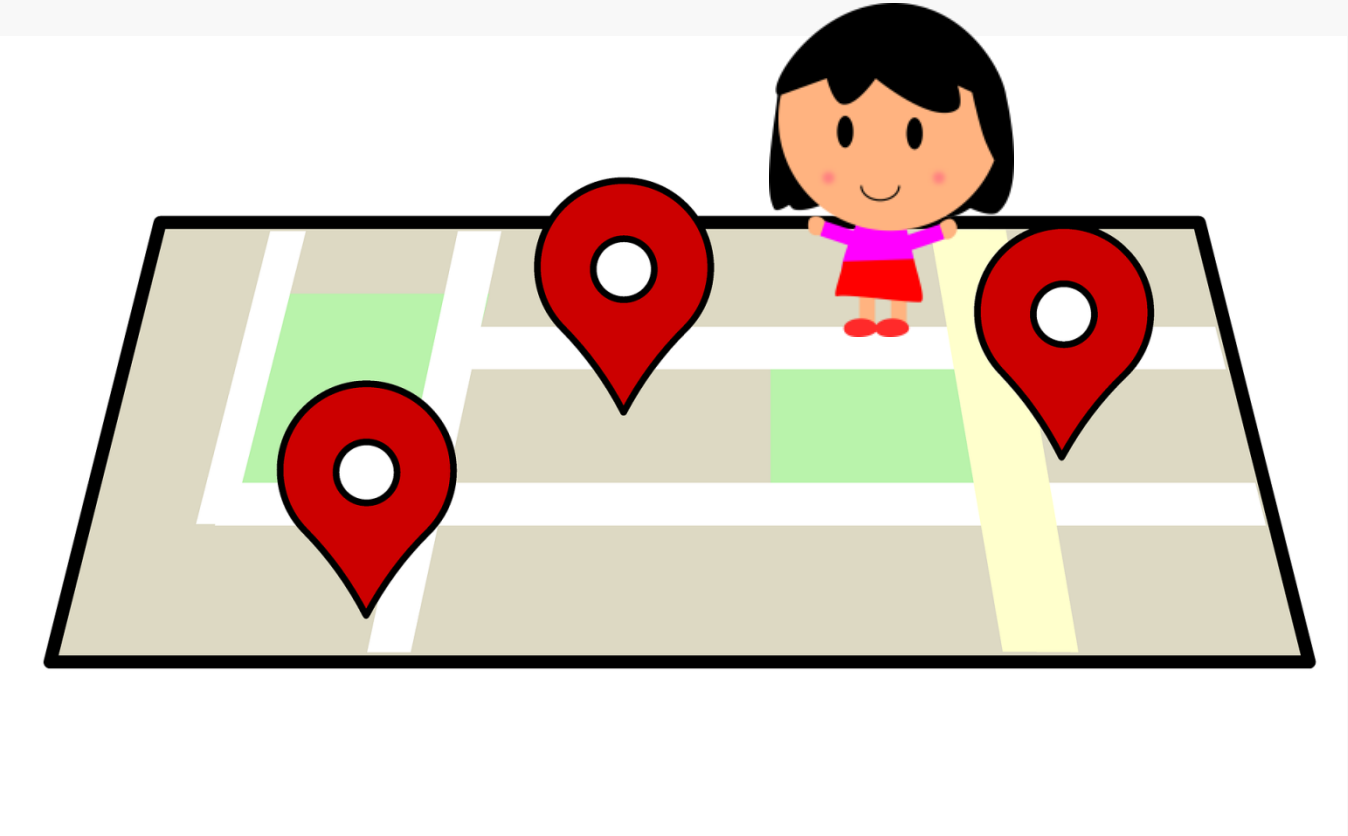
22 mi (37 km)
run this month

28 hours spent in a vehicle this month

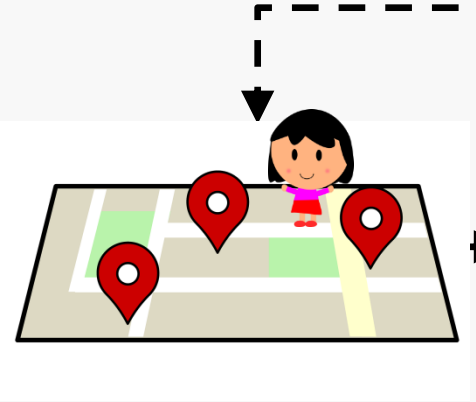
Data Science Process



Can I infer details and create a model of my life from location data?



Data Science Process



Gather Data

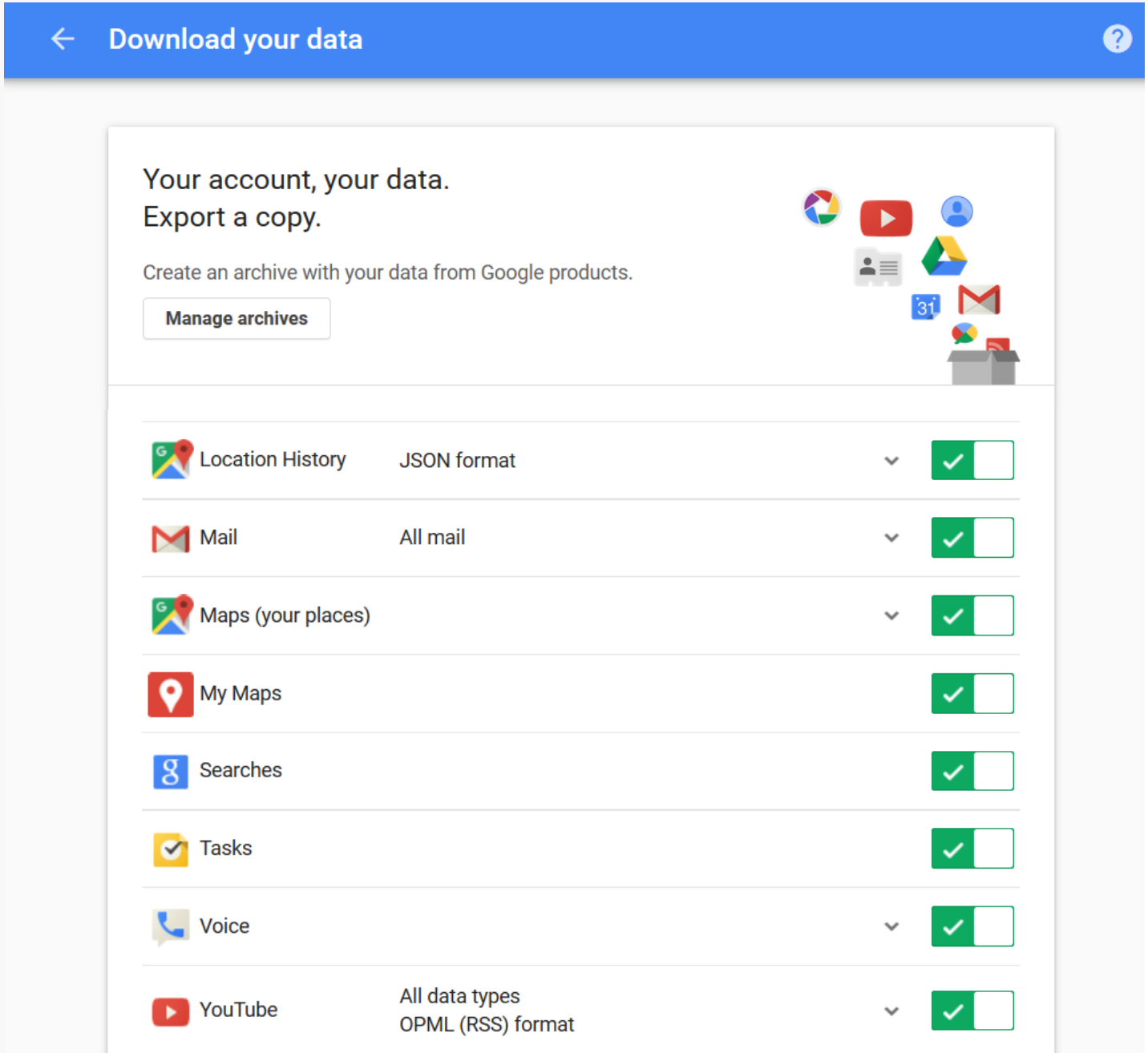
Clean Data

Exploratory Data Analysis

Build Model

Present Findings

https://takeout.google.com



KML format example

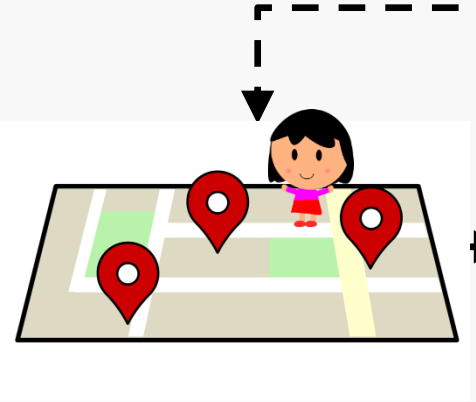
```
<when>2017-03-30T22:16:05Z</when>
<gx:coord>-112.1206089 36.0538447 2110</gx:coord>

<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

Not always available

```
"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",
```

Data Science Process



Gather Data

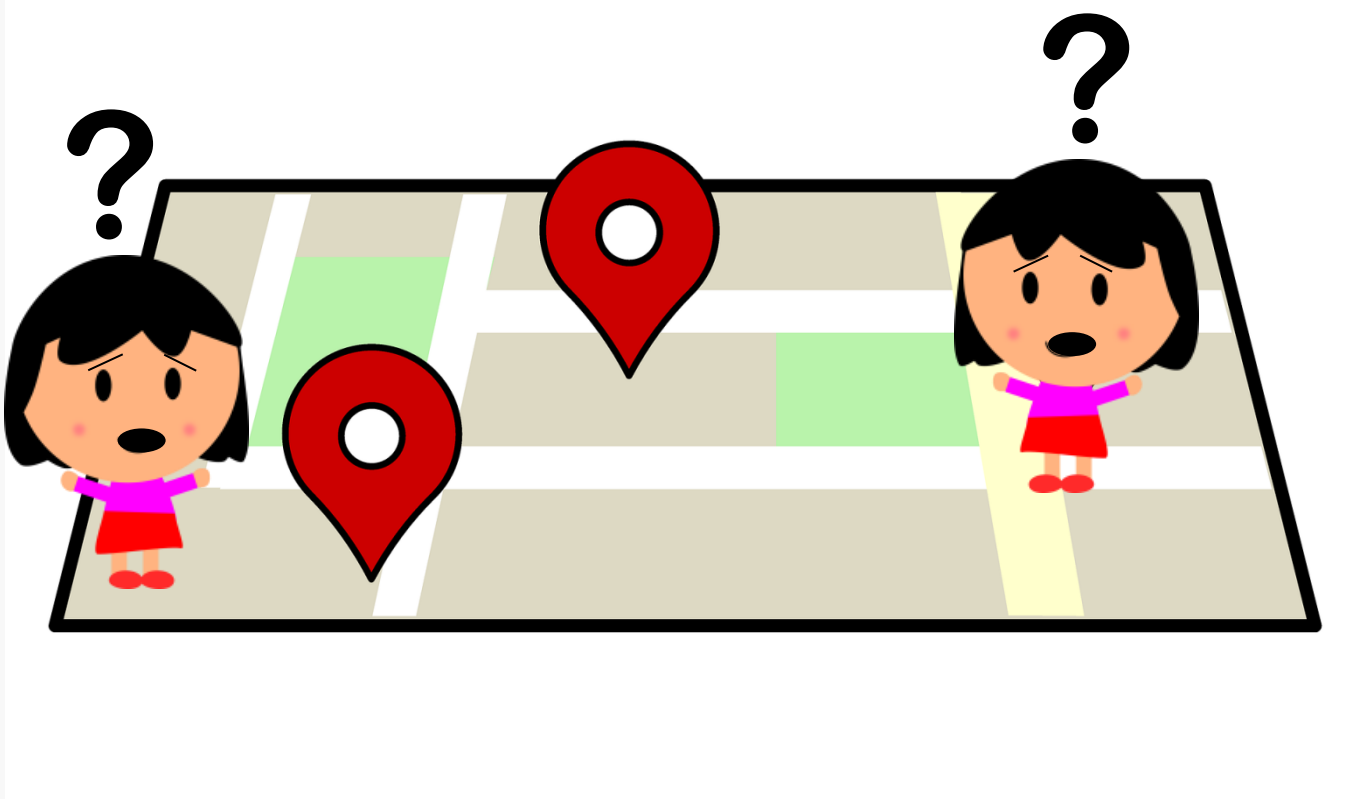
Clean Data

Exploratory Data Analysis

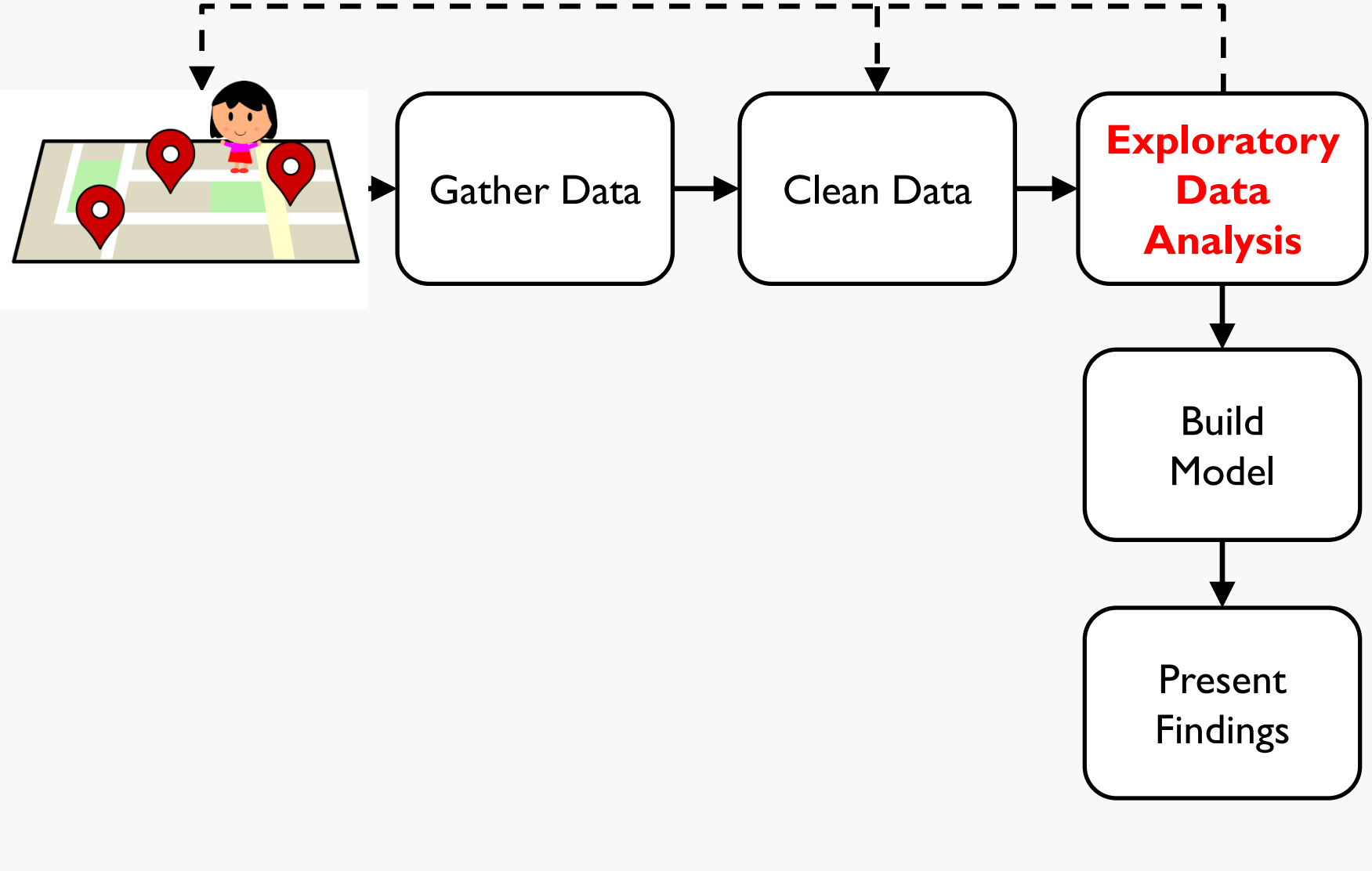
Build Model

Present Findings

Traveling at the speed of light



Data Science Process

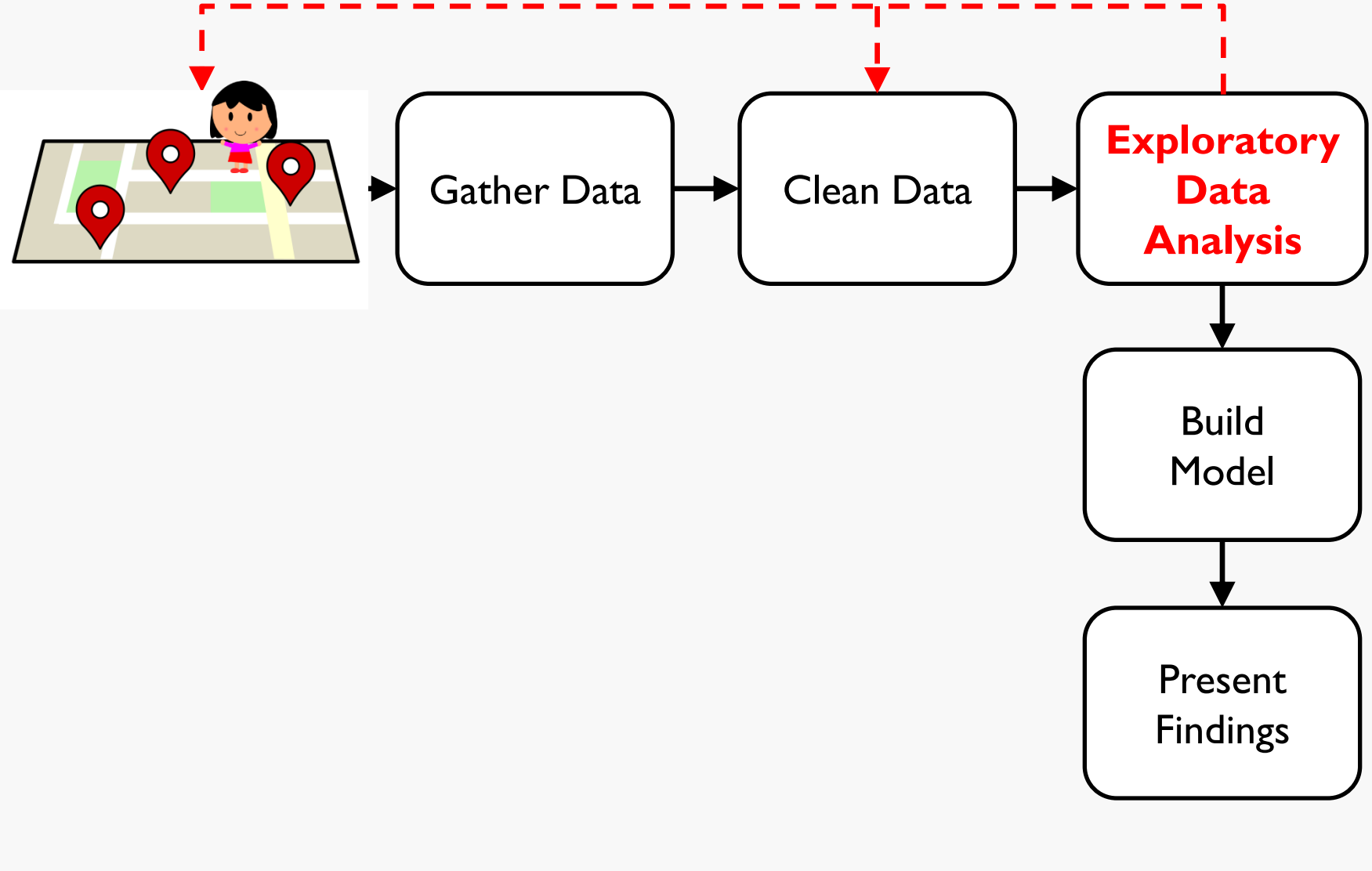


What is EDA?

- Define characteristics
  - Trends
  - Biases
  - Variability
  - Breadth

- Test Assumptions
- Visualize

Data Science Process



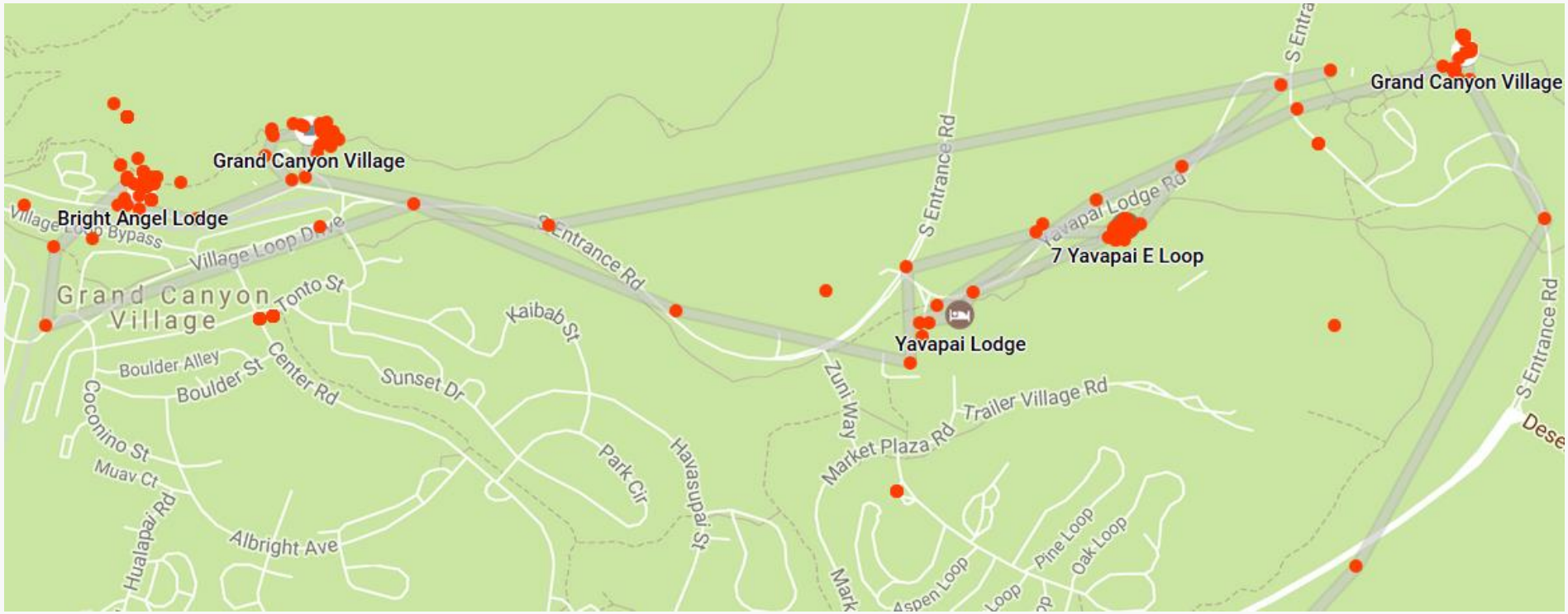
What's in the data?

```
<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>
```

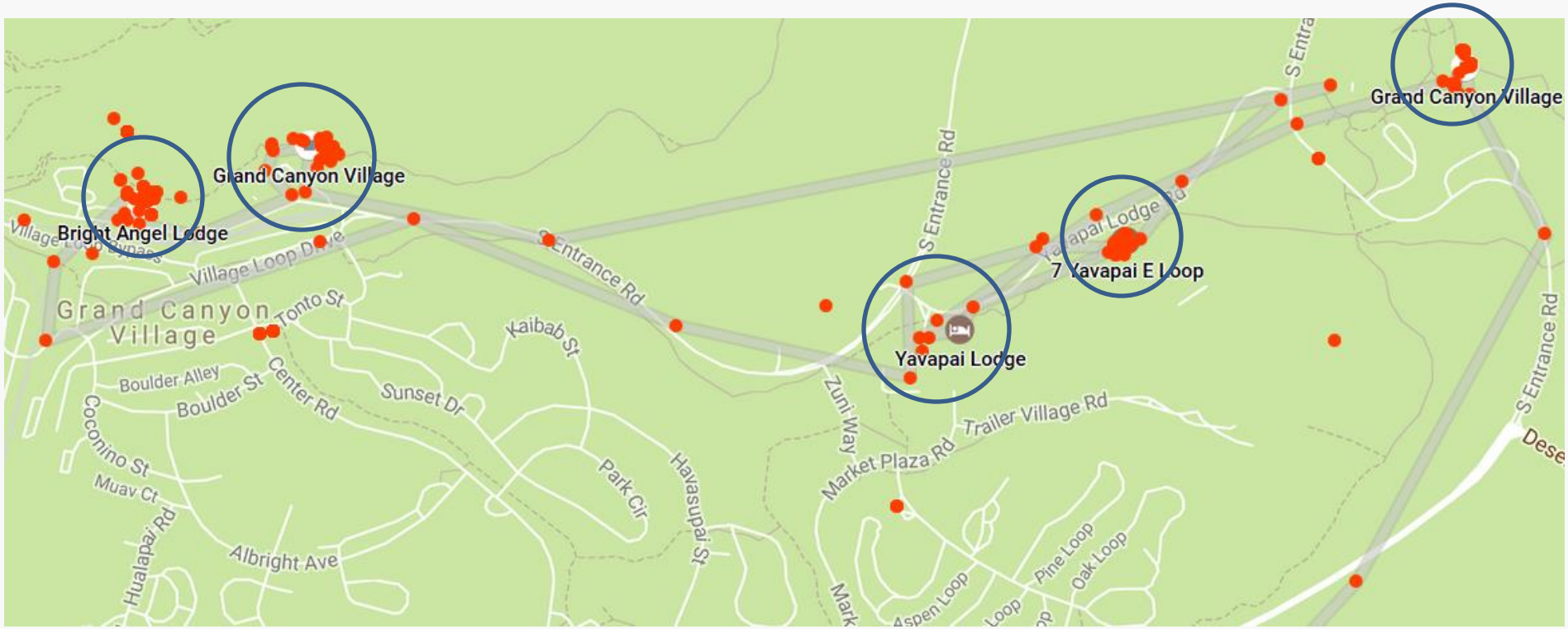
Location

```
<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

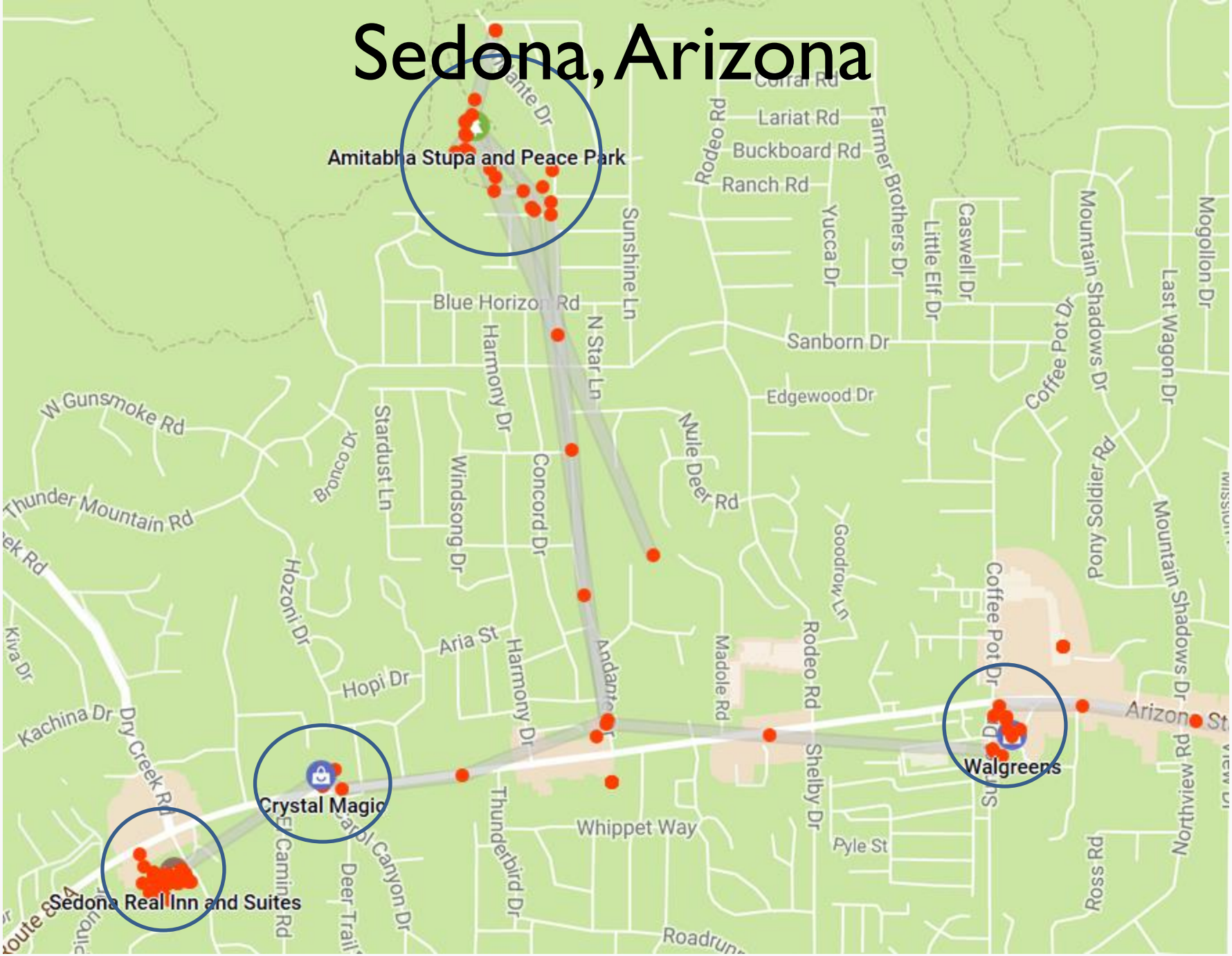


Grand Canyon

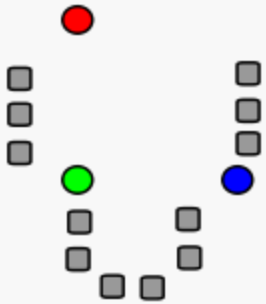


Sedona, Arizona

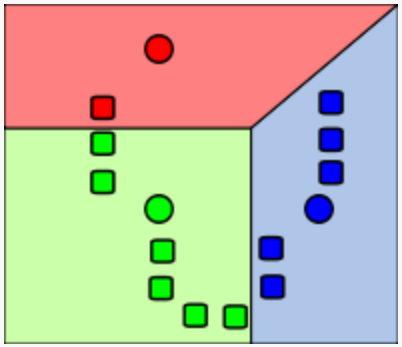
Sedona, Arizona



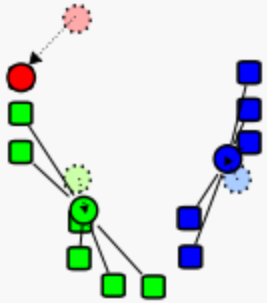
K-Means



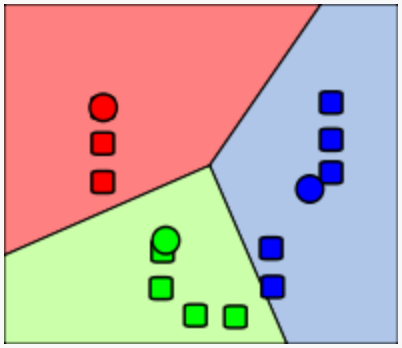
Randomly pick K = 3 points (initial centroids)



Assign each point to its closest centroid



Using points in clusters, calculate new centroids

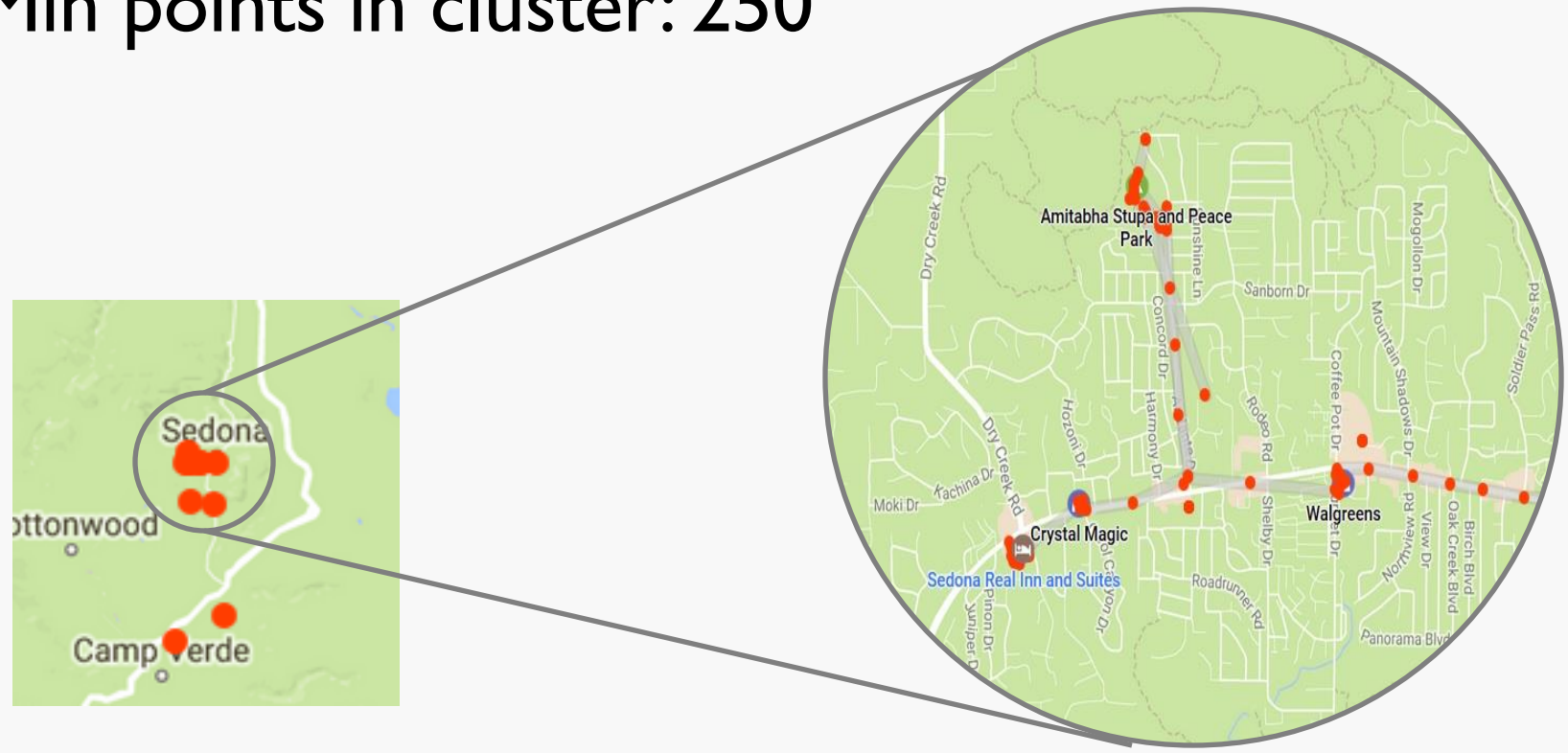


Assign each point to its closest centroid

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

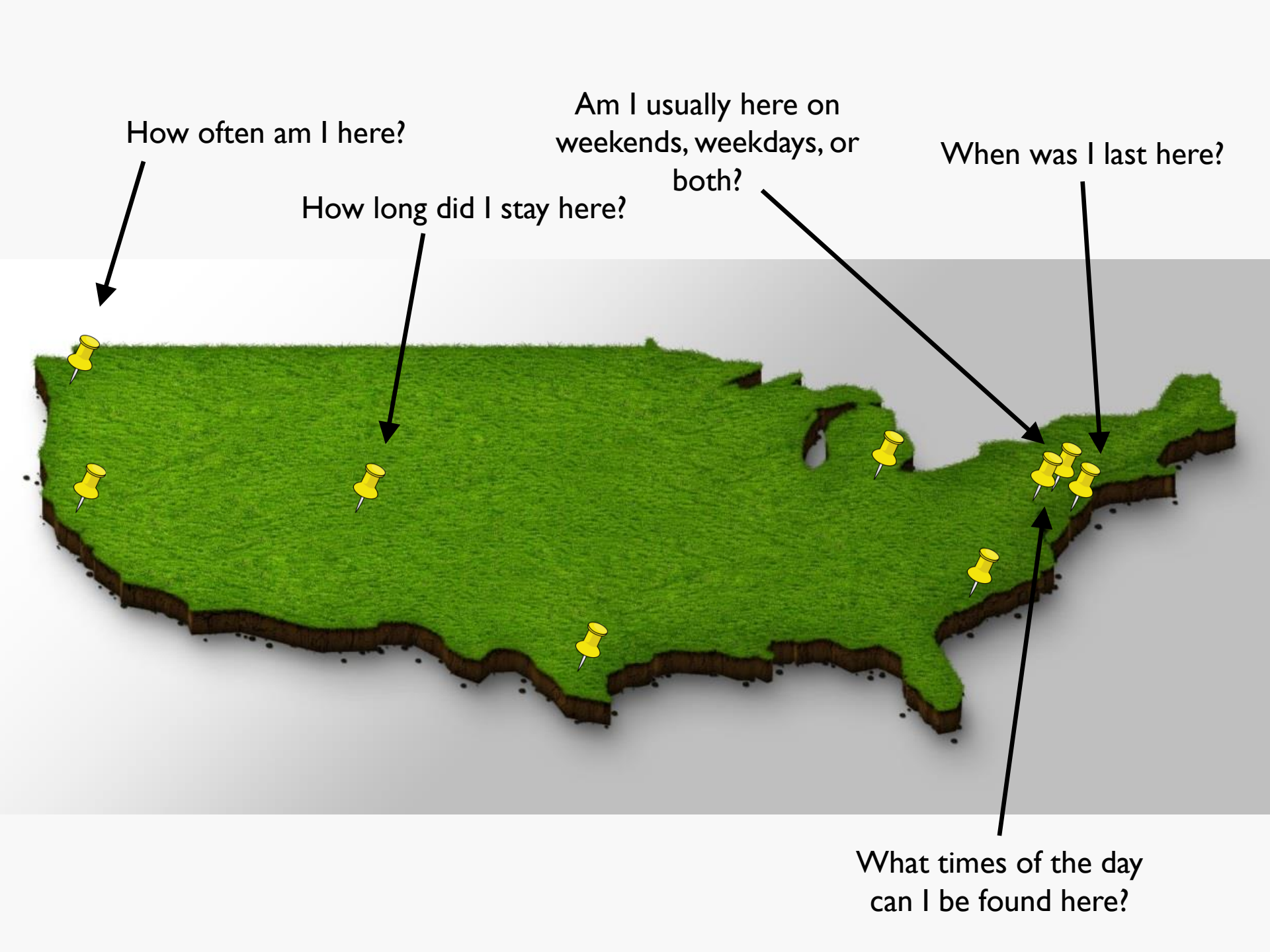
- Max radius: 0.1 miles
- Min points in cluster: 250



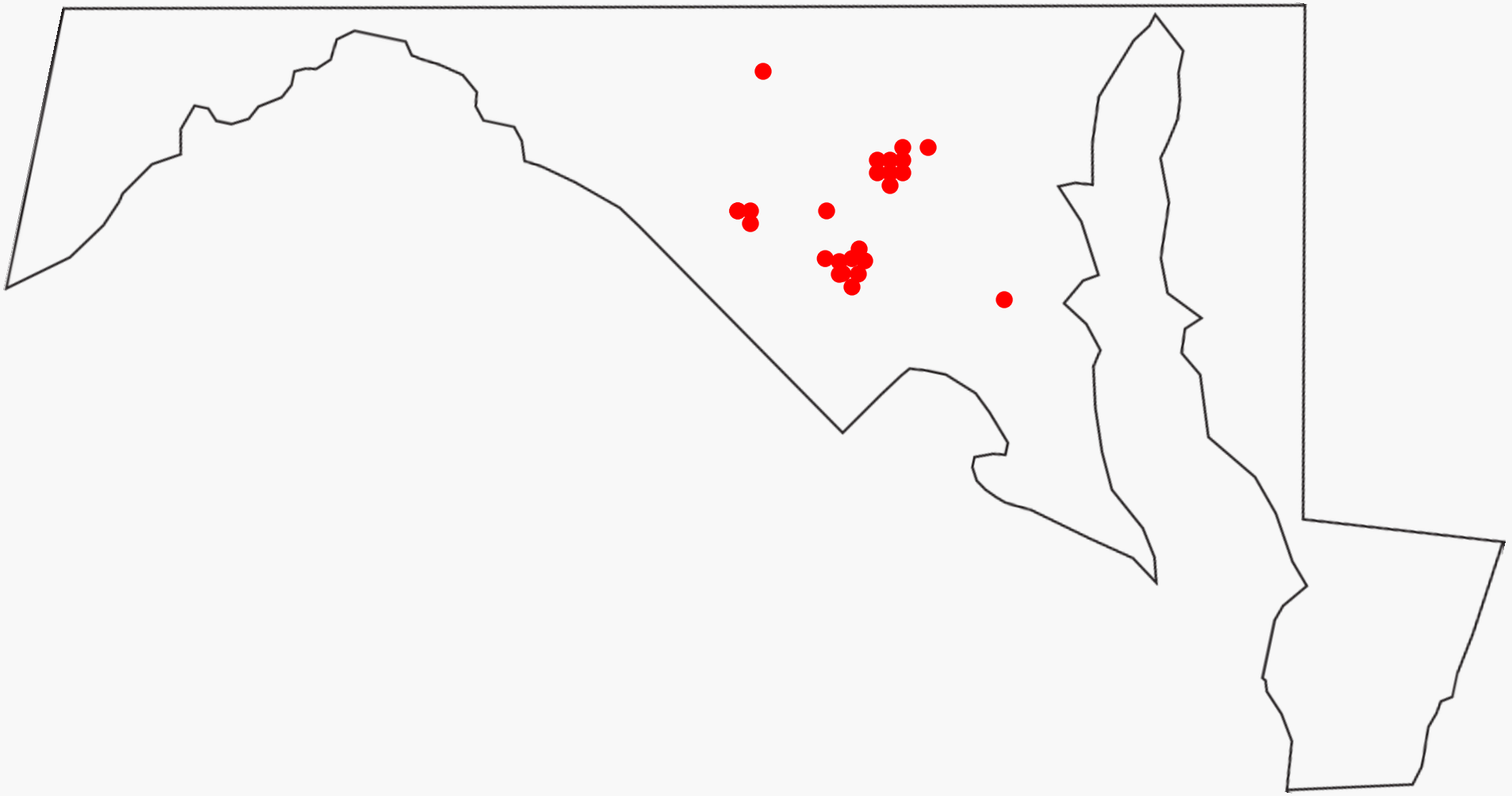
Now what?

Time

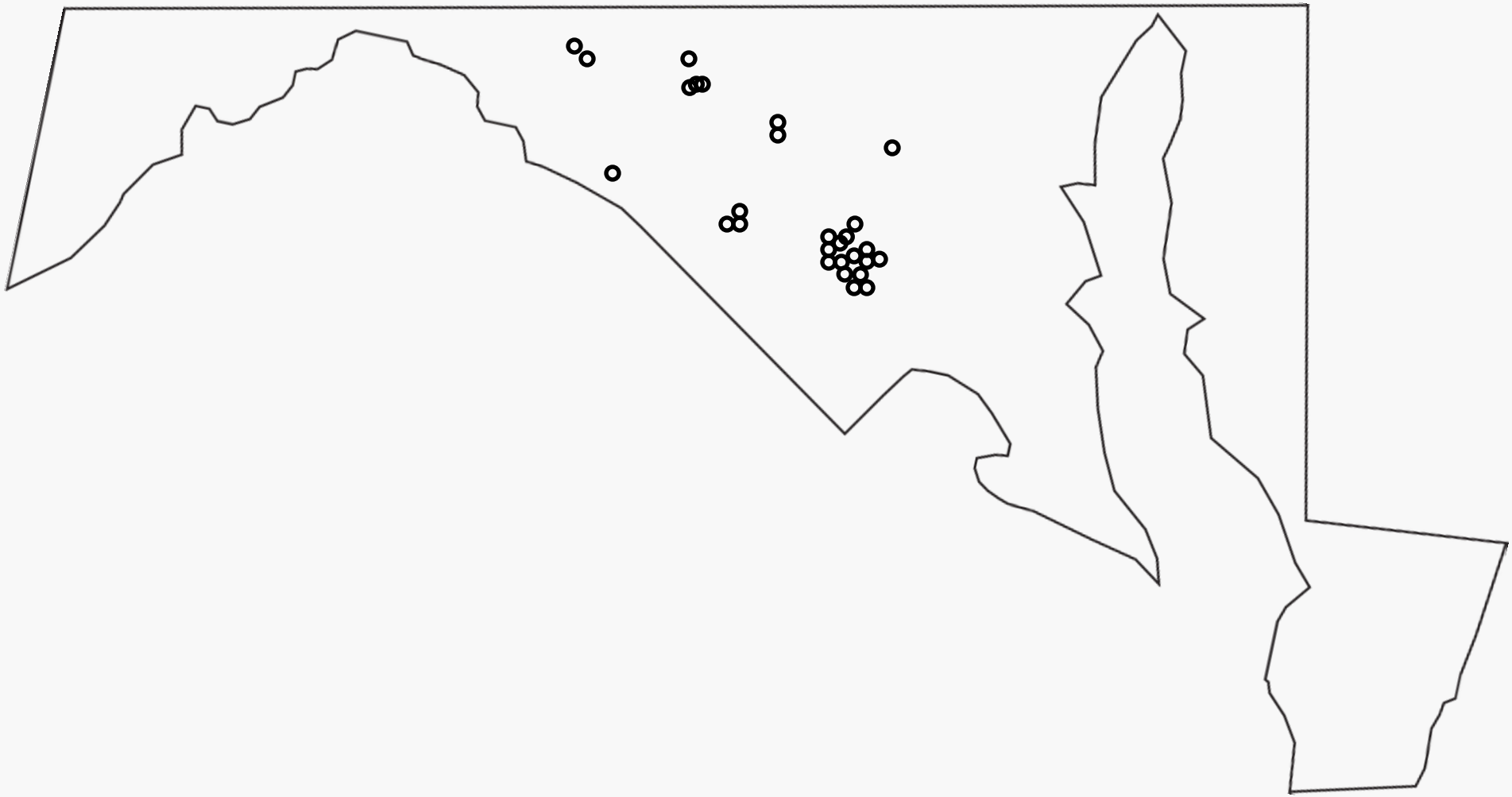
```
<when>2017-03-30T22:16:05Z</when>
<gx:coord>-112.1206089 36.0538447 2110</gx:coord>

<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 am I here?
How long did I stay here?
Am I usually here on weekends, weekdays, or both?
When was I last here?
What times of the day can I be found here?

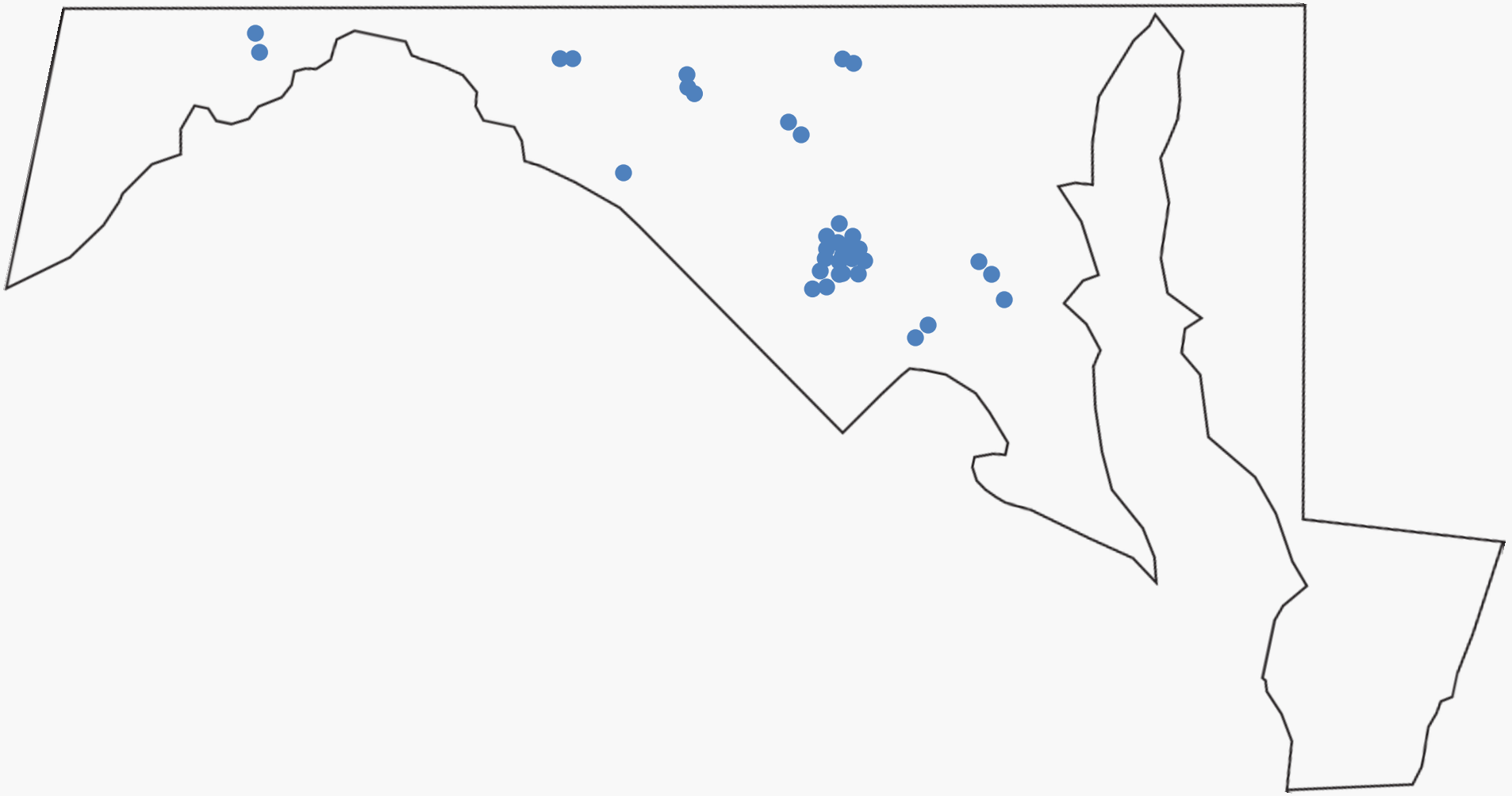
Weekday Day Points



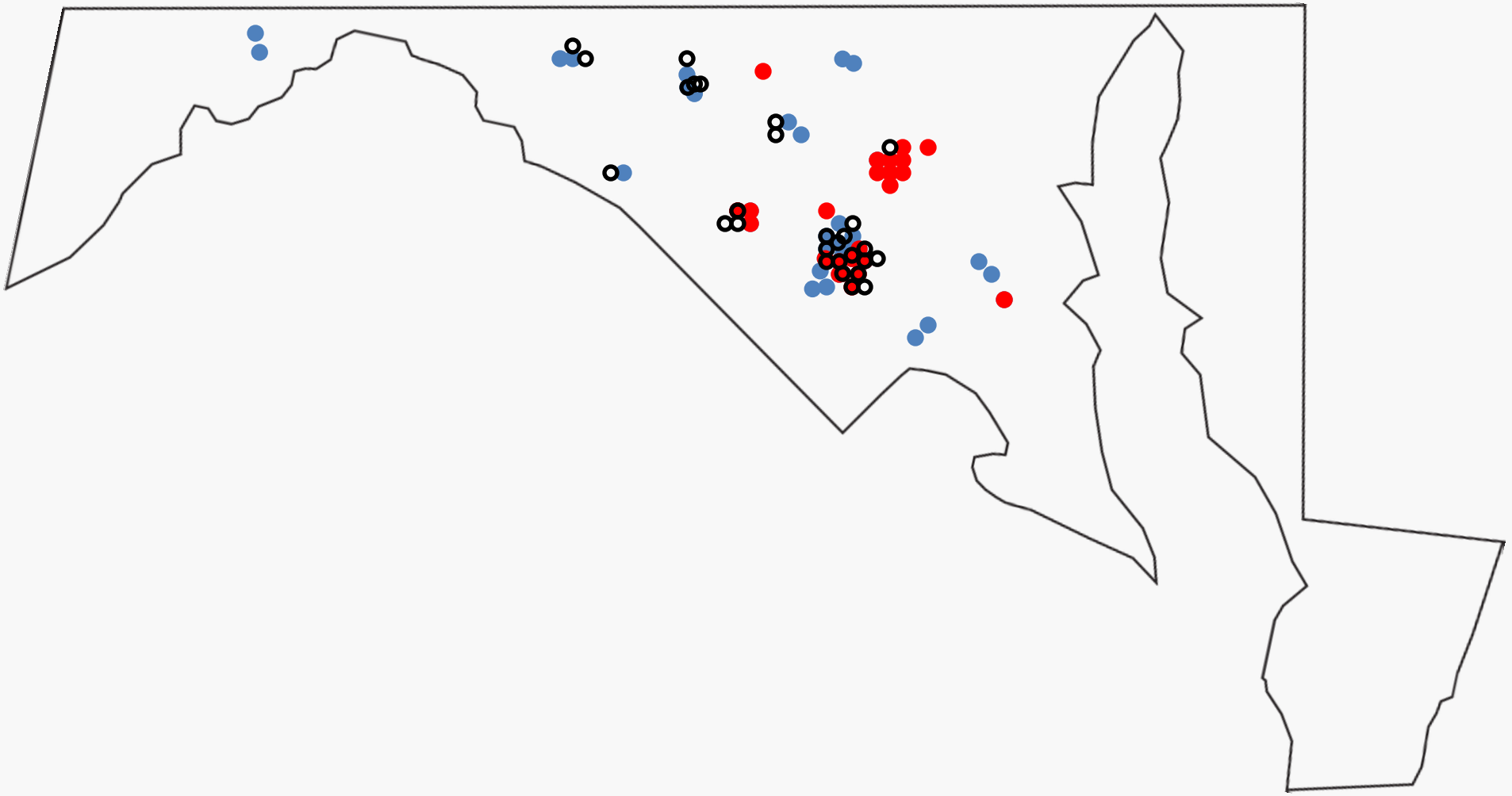
Weekday Evening Points



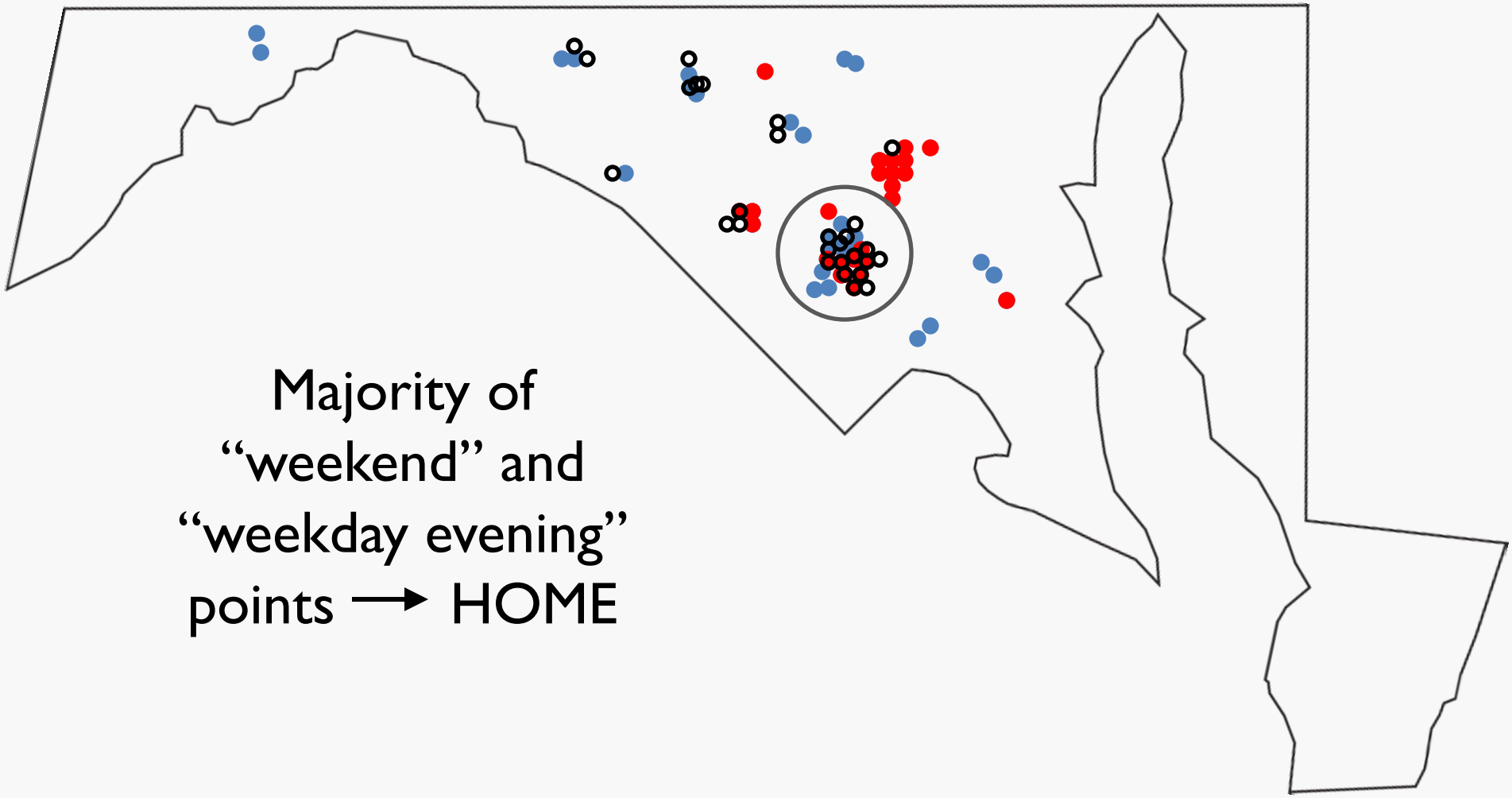
Weekend Points



All Points



All Points



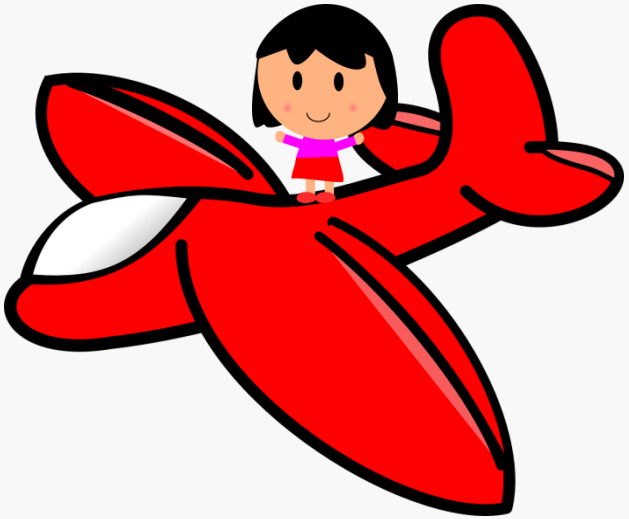
Majority of
"weekend" and
"weekday evening"
points → HOME

If I Know Home, Then I Know…

- When I'm likely to be
  - home
  - away from home
- Local travel radius
- Distance from home at any given point

Significant Location Types

HOME



AWAY FROM HOME

LOCAL

Non-Home Locations

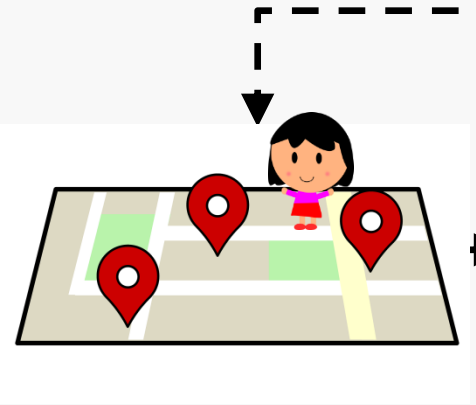
- Away From Home
  - Visited on days I was never home
    - Vacation hotels
    - Conference center
    - Parents' home
- Local
  - *Work*: If there on weekdays for at least 5 hours
  - *Weekend*: If I'm only there on Sat / Sun
    - Concert Venue
  - *Same Day*: Only there on a specific day of week
    - Farmer's Market

Psst…

Google already asks for your Home and Work addresses.

To help you with your commute, or … ??

Data Science Process



Gather Data

Clean Data

Exploratory Data Analysis

Build Model

Present Findings

Significant Location Details

- Lat / Long Boundaries
- Type of location
  - *home, work, weekend, sameDay*

Data Point Details

Original

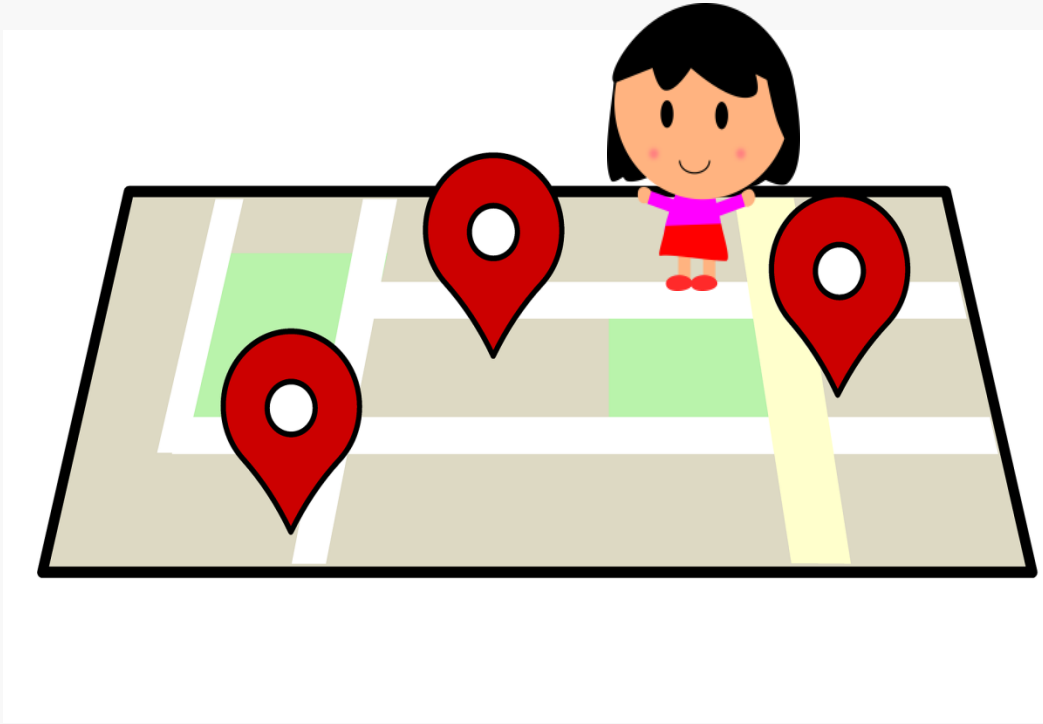
```
<when>2017-03-30T22:16:05Z</when>
<gx:coord>-112.1206089 36.0538447 2110</gx:coord>
```

New

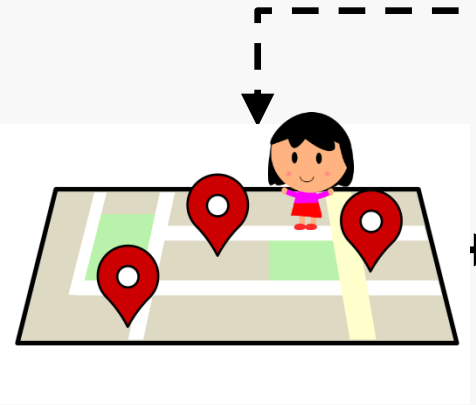
```
<distanceFromHome>1949.46</distanceFromHome>
<locationLabel>cluster3</locationLabel>
<description>awayFromHome</description>
```

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.
- Will I 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

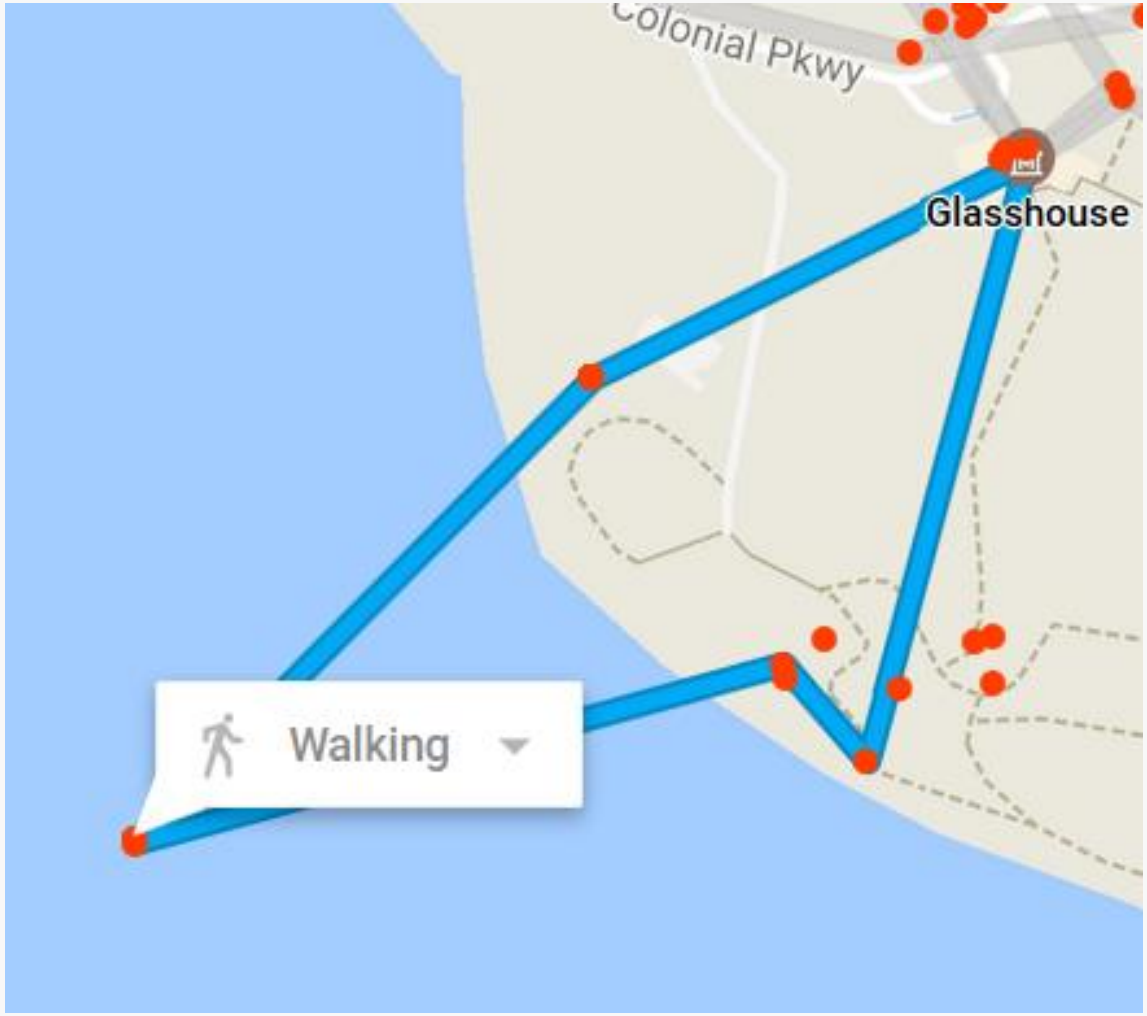
- Regular schedule

- 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)

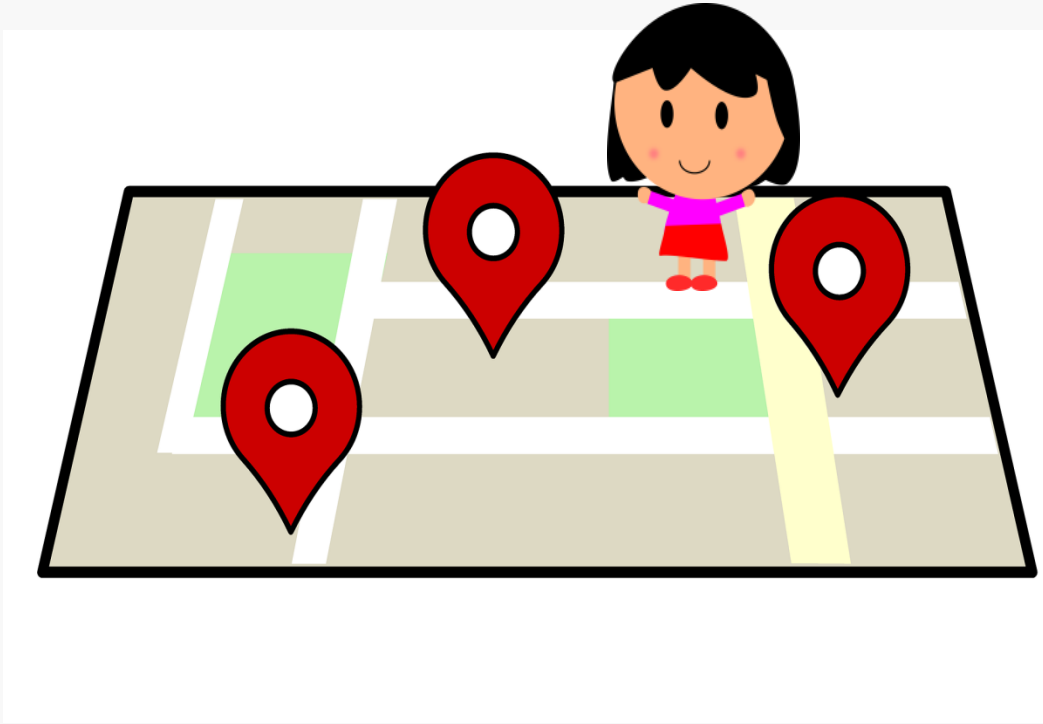
Don’t let Google Ruin Your Life!

My girlfriend likes to track me on google maps and i encourage that but the fact that the maps are showing me in places and addresses i am not nor have ever been is really creating trust issues she now thinks im a liar and a cheat and trying to hide things from her. THANK YOU GOOGLE FOR THE PROMBLEMS YOU ARE CREATING IN MY LIFE i am not s liar or a cheater but your maps say different i hope you google bastards have the same issues with the people you care about so you know how it feel

Categories: Chrome Map errors... 2 posts 16 views 1 expert reply

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

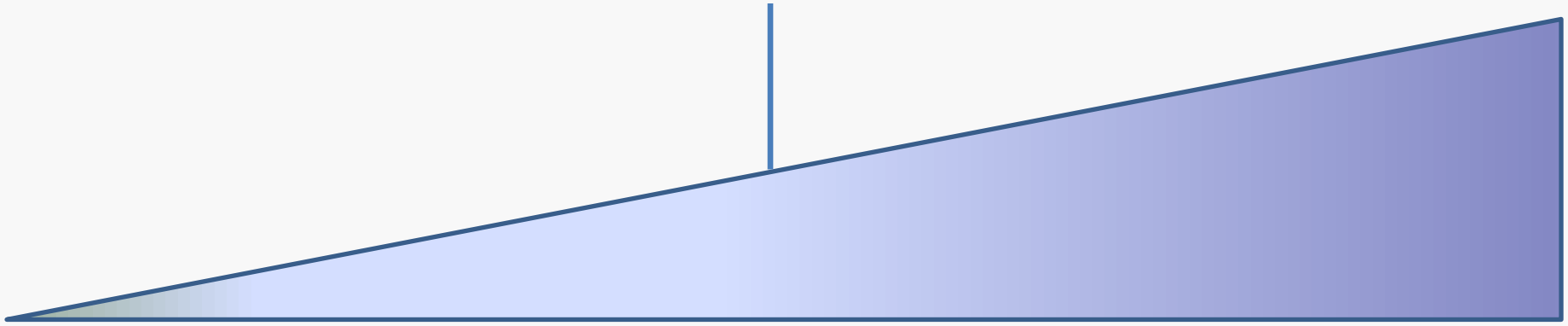
- Products and apps
- Companies that access data
- Companies that buy / share data

Implications

Benign

Implications

Your insurance claim was denied due to…



Benign

Worrisome

Big Brother on wheels: Why your car company may know more about you than your spouse.

The Washington Post (January 15, 2018)

"… an automaker can vacuum up a massive amount of personal information about someone … where he shops, the weather on his street, how often he wears his seat belt, what he was doing moments before a wreck — even where he likes to eat and how much he weighs."

Companies race to gather a newly prized currency: Our body measurements

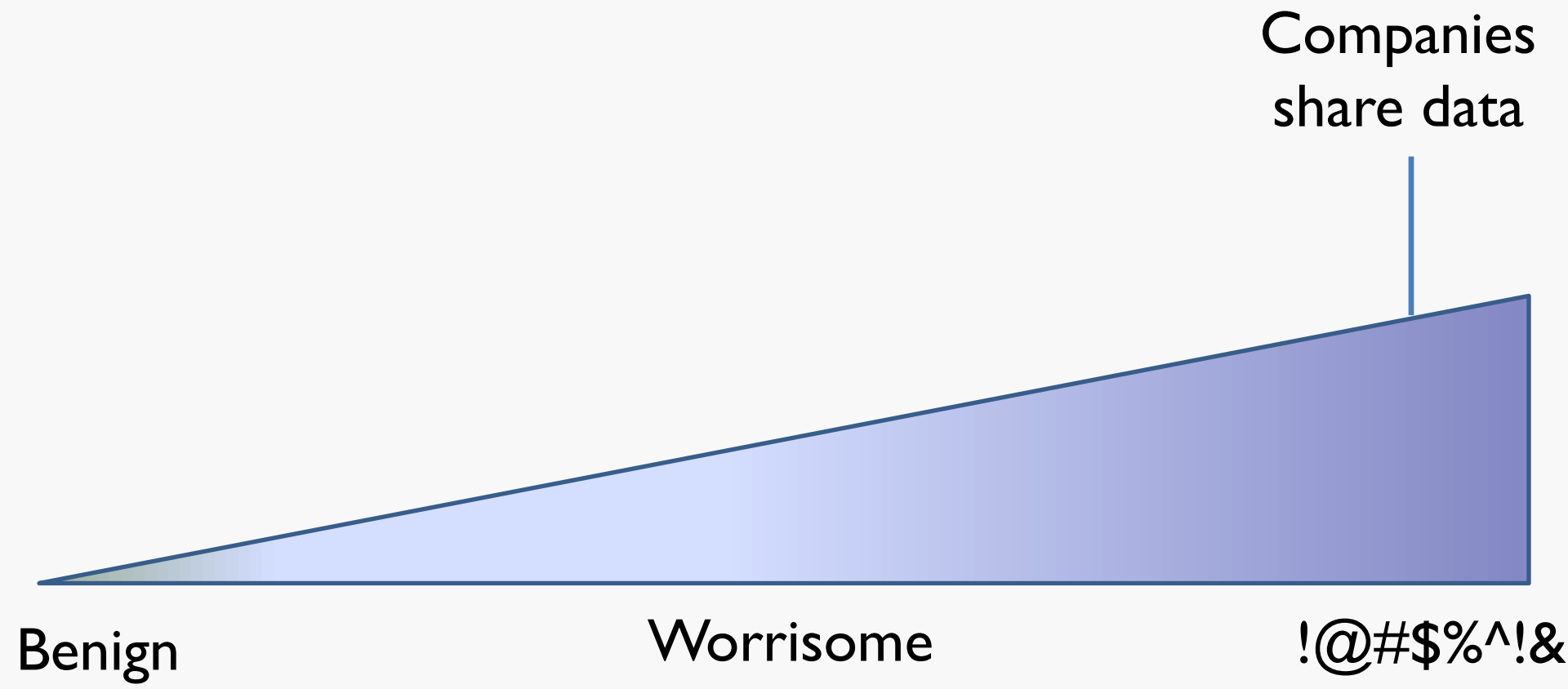
The Washington Post (January 16, 2018)

"Clothing companies now see body measurements as one of their most prized currencies, and millions of Americans are increasingly offering up their innermost personal data in search of customized pieces or a better fit."

"These body measurements look a lot like medical records…"

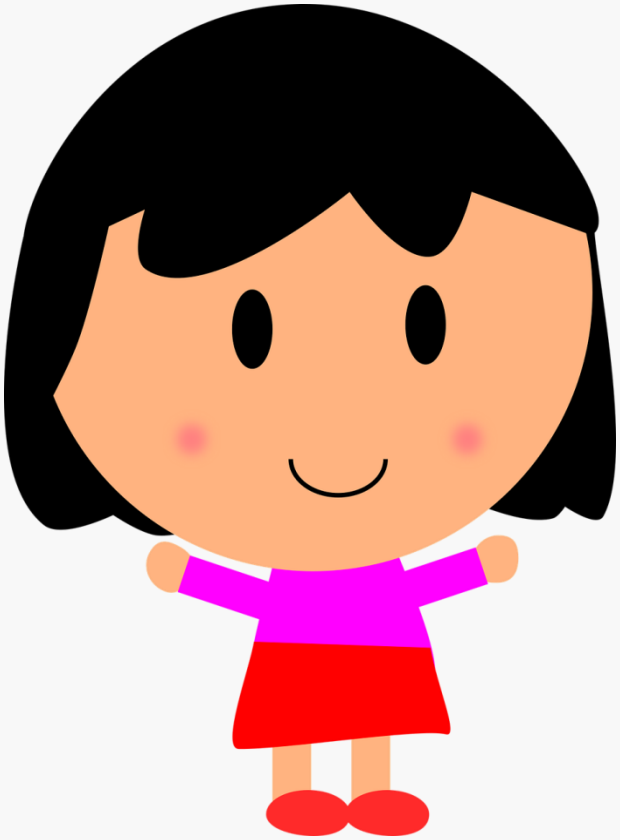
"… privacy experts worry that the retailers eventually will be tempted to sell the data … Or the information could become the target of hackers. Passwords can be changed; body sizes can't."

Implications



Your Data, Your Choice

Further Information

- Code
  - Python: Jupyter Notebook
  - https://github.com/laconicllama
- https://laconicllama.blogspot.com/
- laconicllama@hotmail.com

Questions?