

# CLUSTERING

---

~Abhishek Kumar

# Scope

---

- What is Clustering?
- Types of Clustering Algorithms
- Pros & Cons
- Areas of Applications
- Python code implementation
- Discussion





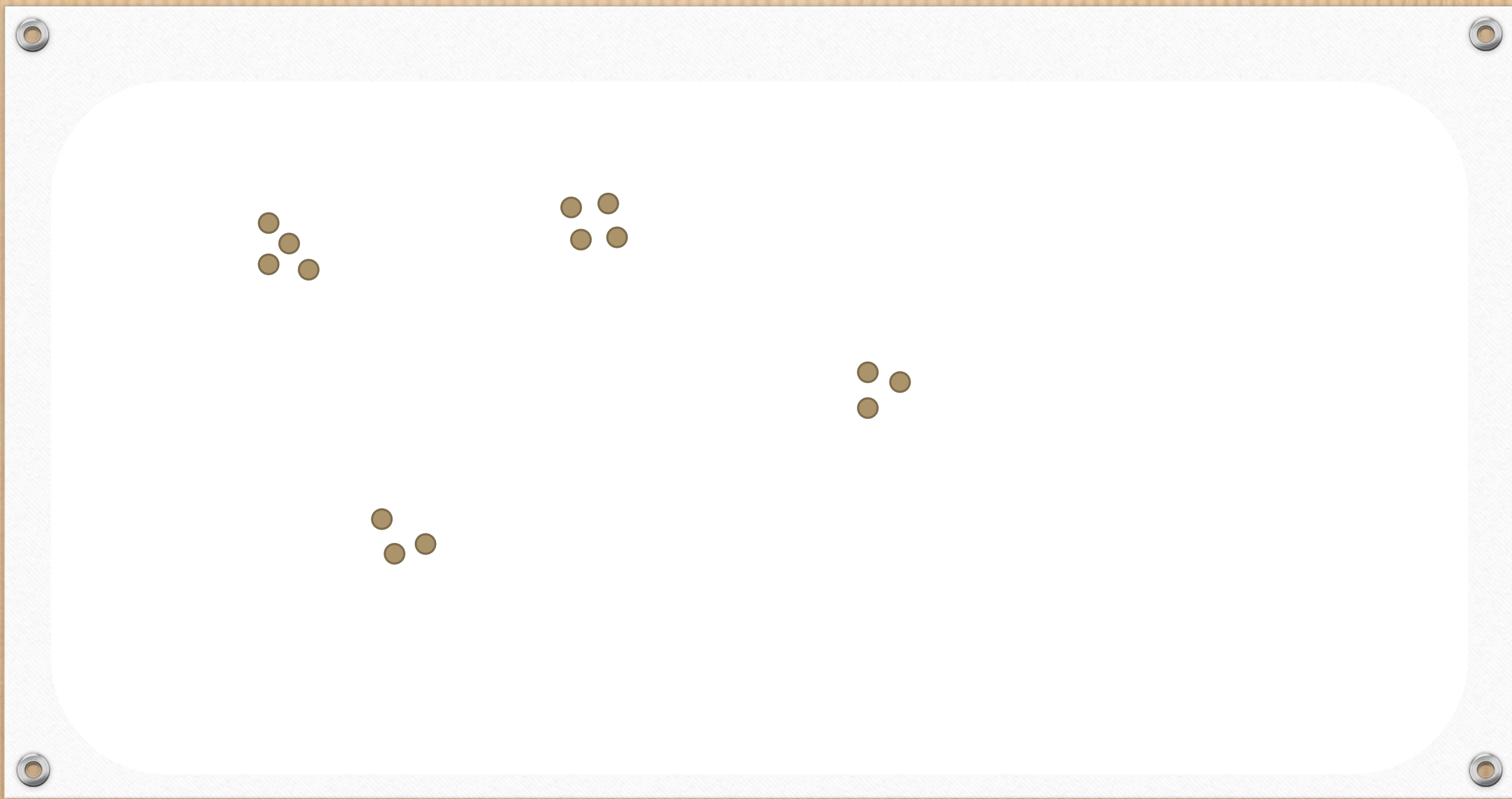
---

Unsupervised learning ?

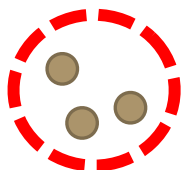
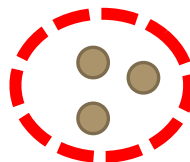
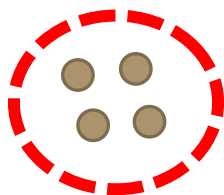
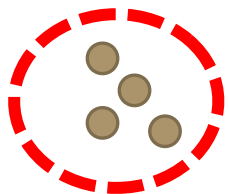
# Clustering

---

- Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group called a cluster are more similar to each other than to those in other groups.
- Sometimes called
  - sorting by psychologists
  - segmentation by people in marketing







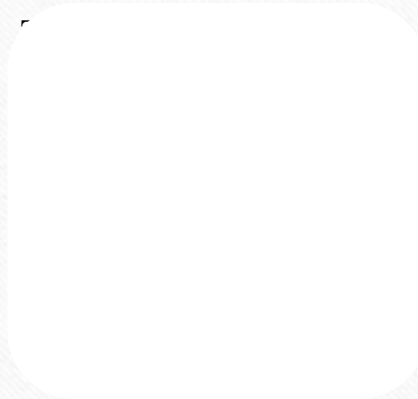
# What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.

---



Similarity is hard  
to define, but...  
*"We know it when  
we see it"*





# Distance measure

---

- Euclidean distance

$$d(g_1, g_2) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- Manhattan distance











$$d(g_1, g_2) = \sum_{i=1}^n |x_i - y_i|$$

- Minkowski distance

$$d(g_1, g_2) = \sqrt[m]{\sum_{i=1}^n (x_i - y_i)^m}$$



# Customer Segmentation

									
SAM	DAN	ROBBIE	SHAUN	DANNA	ROBIN	SANDY	DANE	RONNIE	ROB
AGE : 18	49	37	20	51	40	23	50	42	40
ENG: 30	10	70	40	10	60	20	20	70	80
(times/month)									

Hint: Similar Names!!

engagement  
(times/month)

70

60

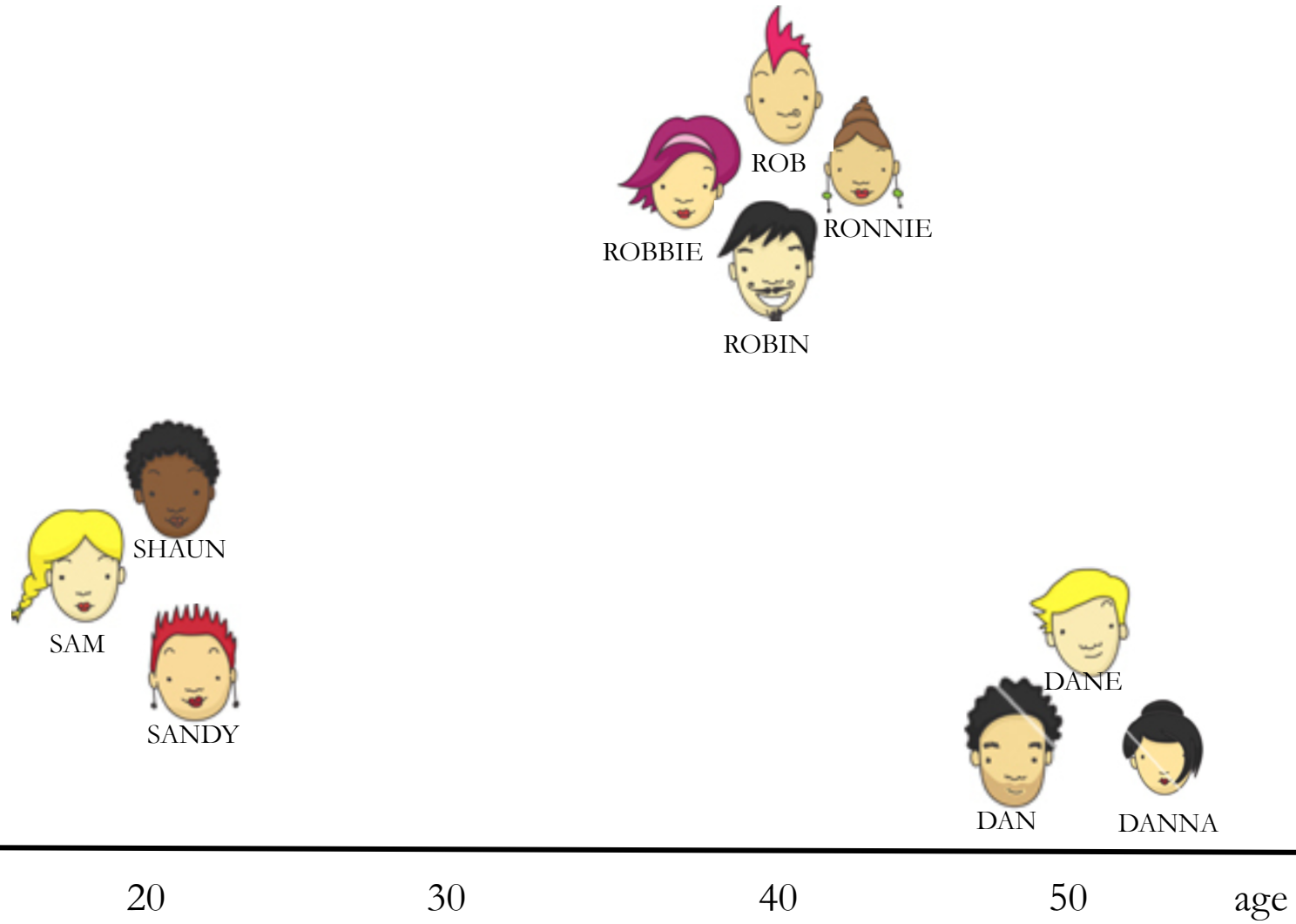
50

40

30

20

10



age



# What Is A Good Clustering?

---

- Organizing data into classes such that there is
  - high intra-class similarity
  - low inter-class similarity

# Types of Clustering?

---

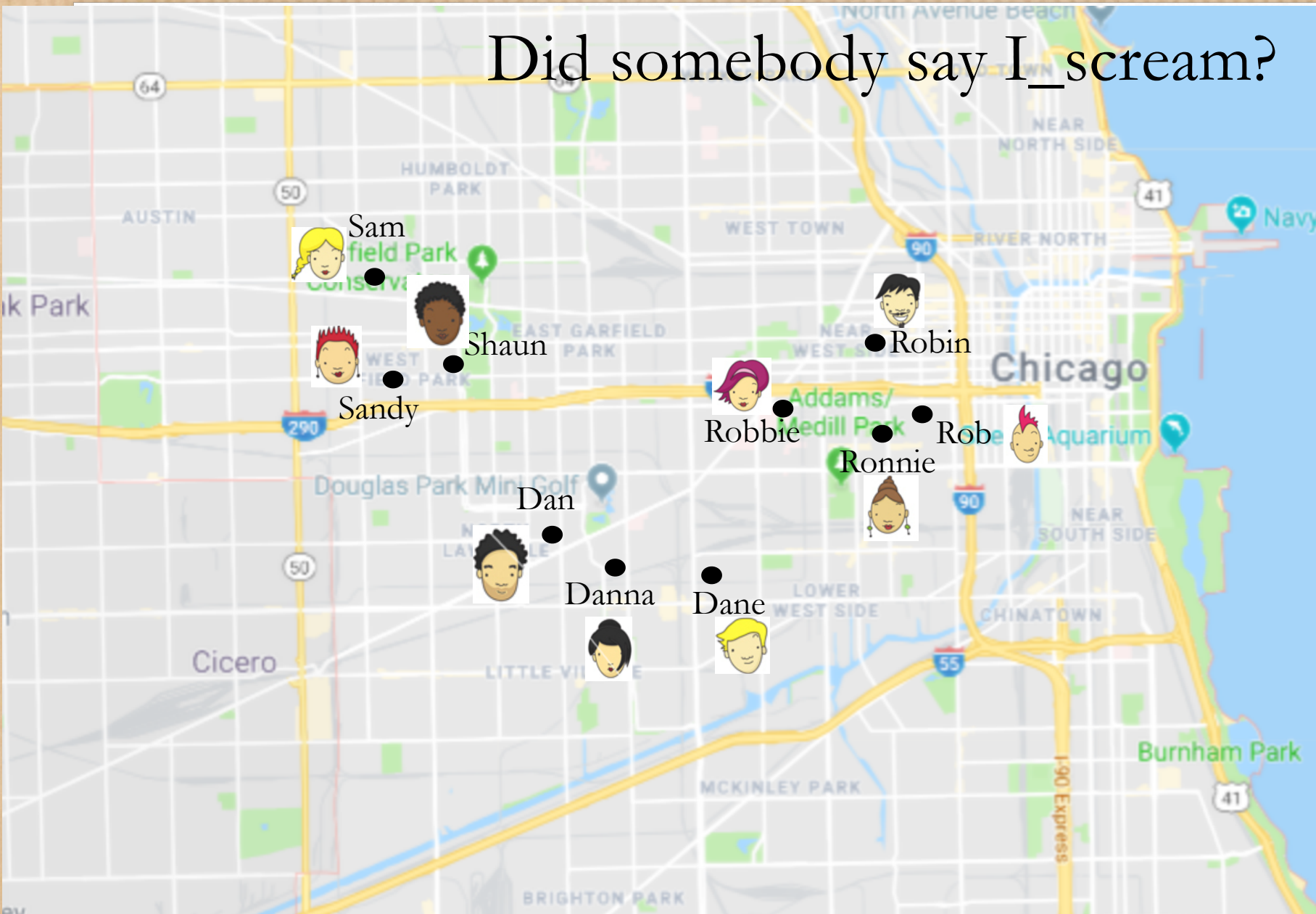
- K means clustering
- Hierarchical clustering



---

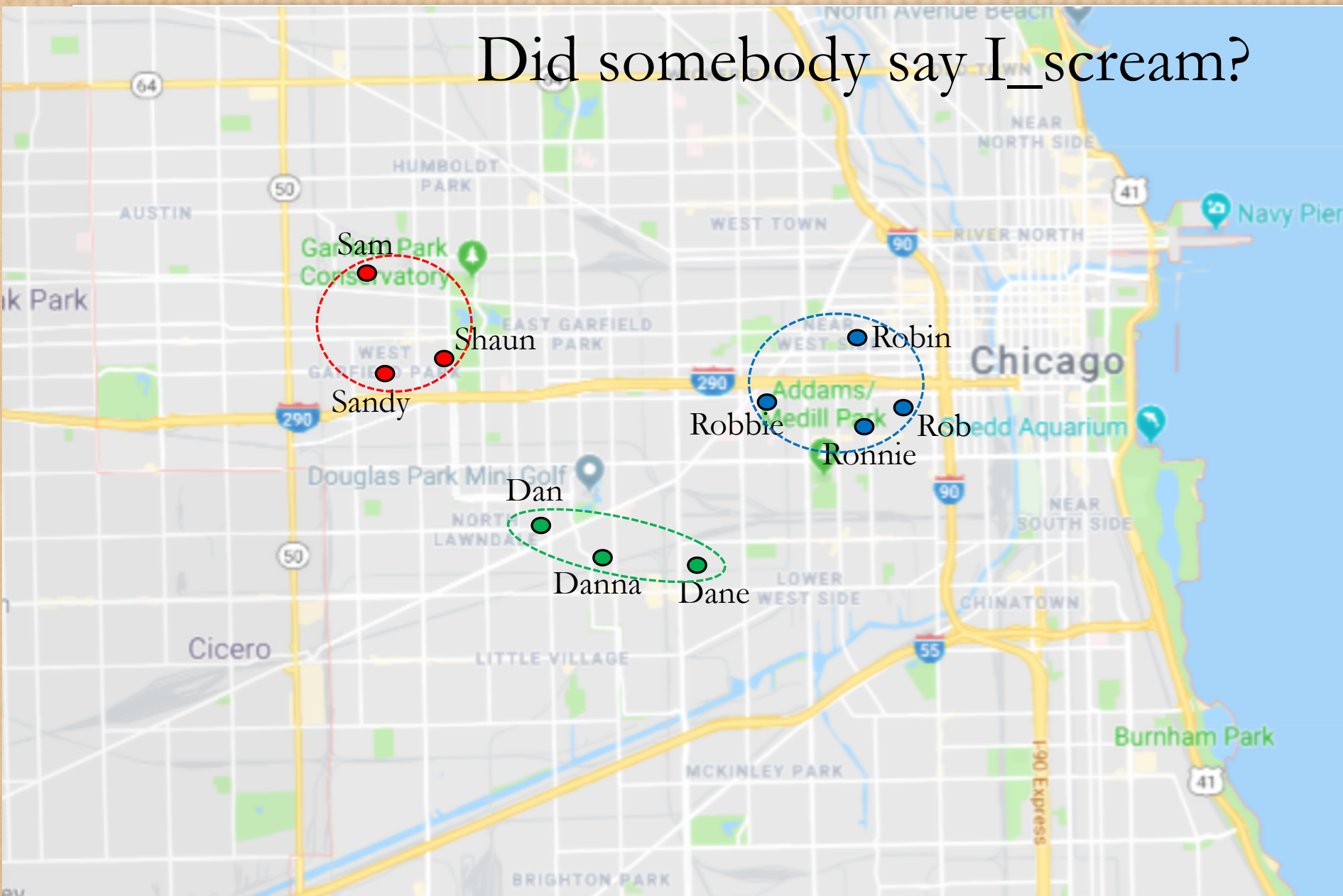
# K-means Clustering

# Did somebody say I\_scream?

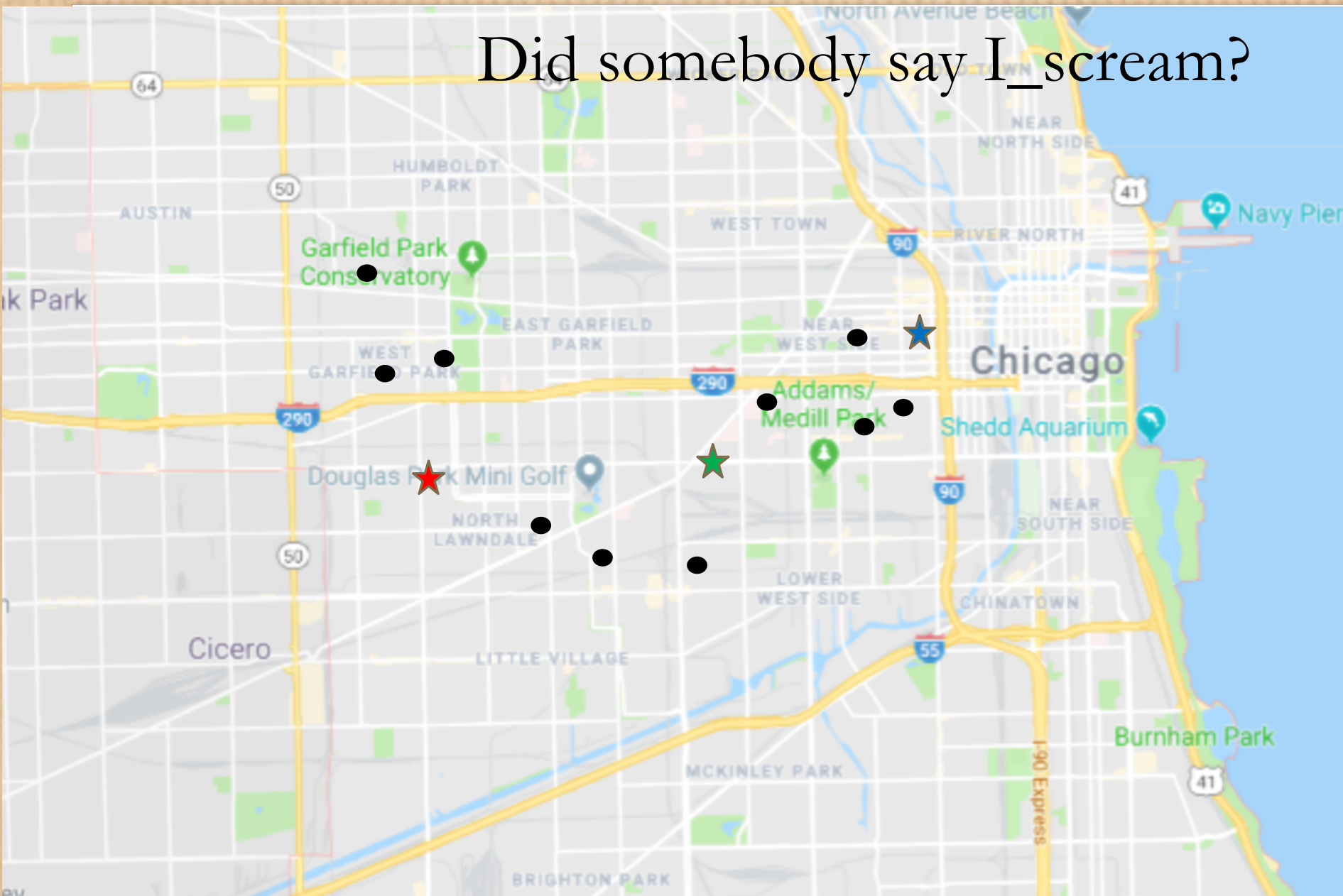




Did somebody say I\_scream?

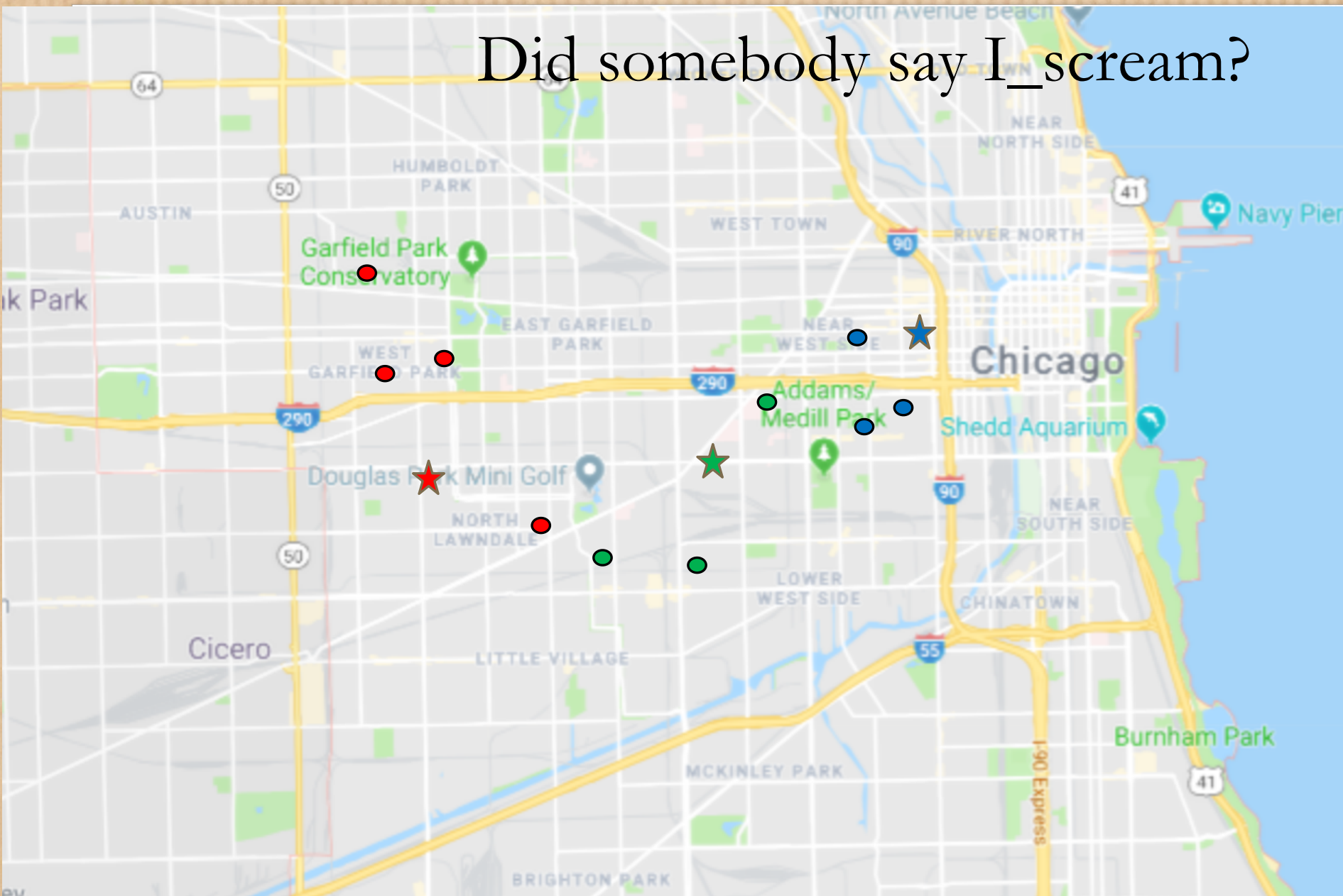


Did somebody say I\_scream?

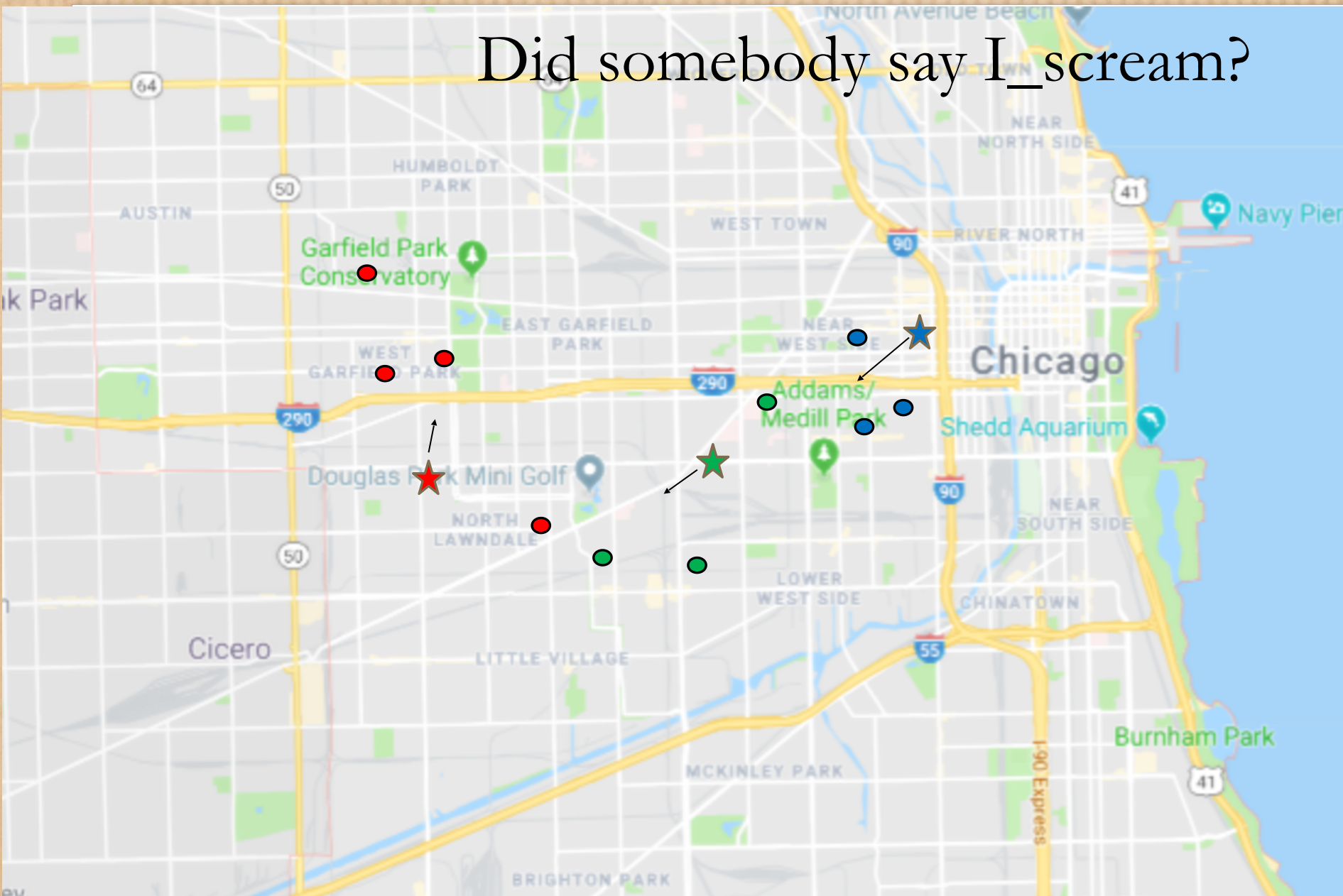




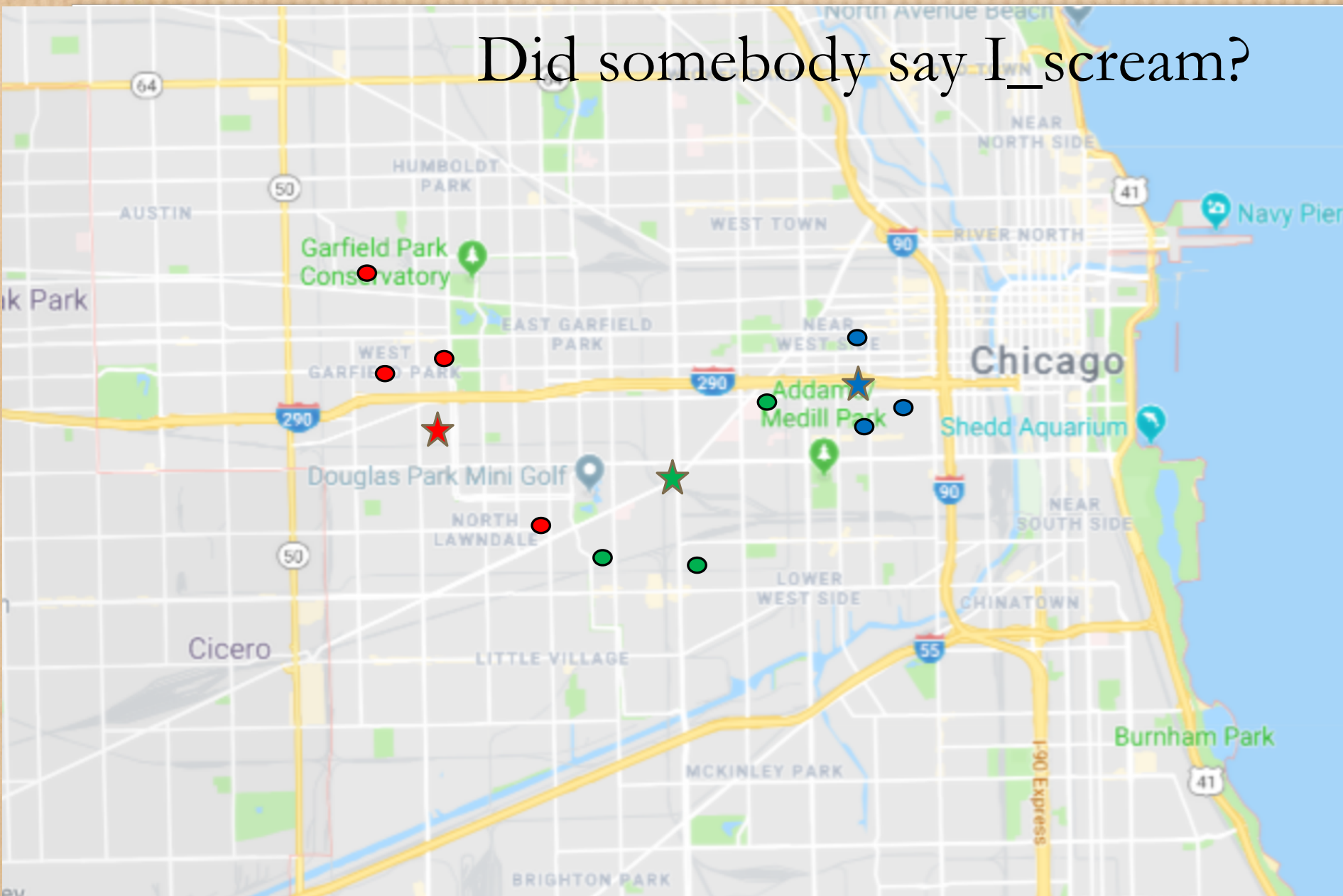
Did somebody say I\_scream?



Did somebody say I\_scream?

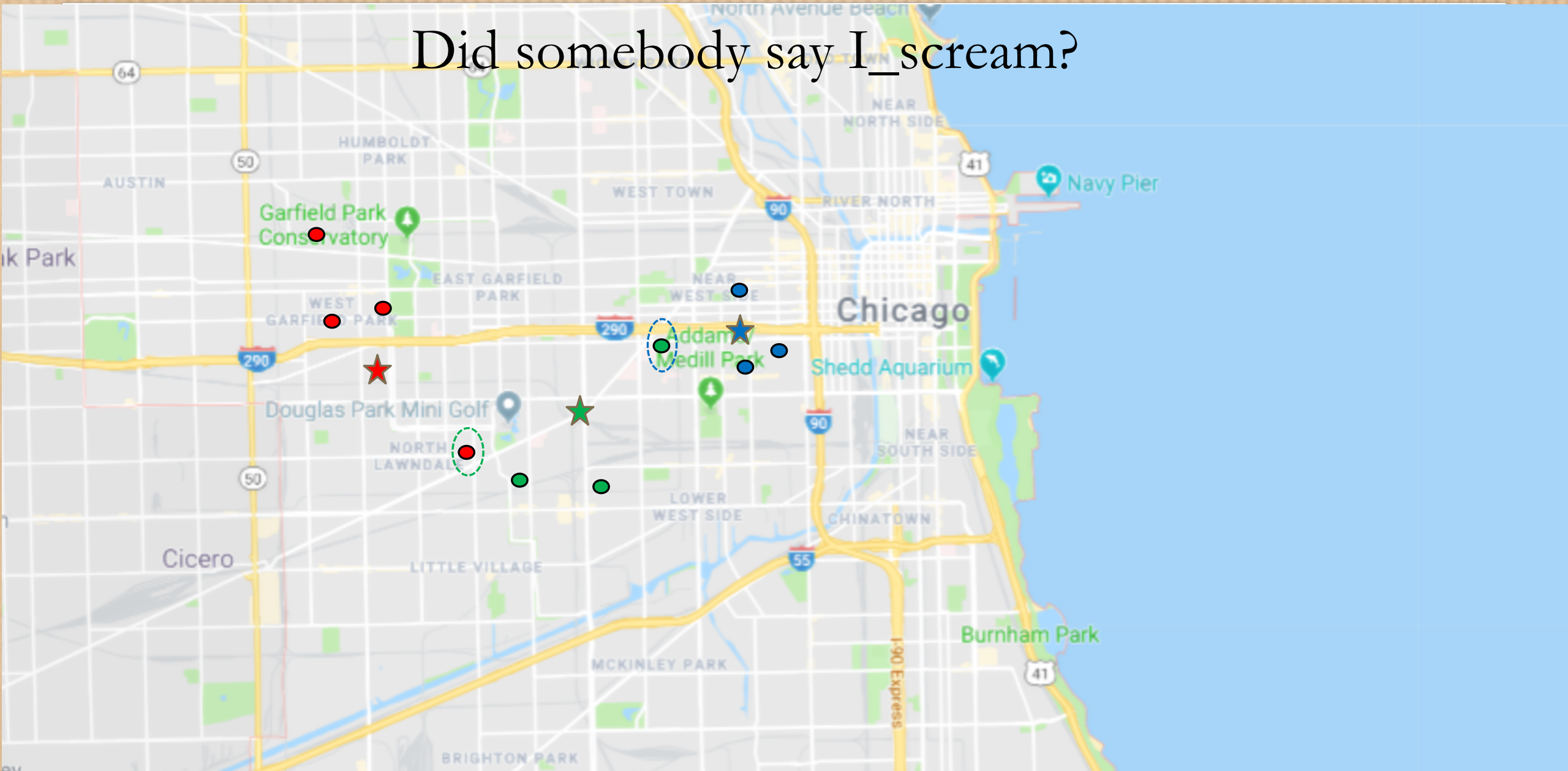


Did somebody say I\_scream?





# Did somebody say I\_scream?



A map of Chicago, Illinois, showing various neighborhoods and landmarks. The map is overlaid with several colored markers: red dots, green dots, blue dots, and stars. The markers are distributed across the city, with a concentration in the central and western parts. The text "Did somebody say I\_scream?" is displayed at the top of the map.

Neighborhoods and landmarks visible on the map include:

- AUSTIN
- HUMBOLDT PARK
- Garfield Park Conservatory
- WEST GARFIELD PARK
- EAST GARFIELD PARK
- NEAR WEST SIDE
- CHICAGO
- RIVER NORTH
- NEAR NORTH BEACH
- NEAR SOUTH SIDE
- CHINATOWN
- Burnham Park
- MCKINLEY PARK
- BRIGHTON PARK
- LITTLE VILLAGE
- DOUGLAS PARK MINI GOLF
- NORTH LAWNDALE
- CICERO
- SHedd AQUARIUM
- NAVY PIER

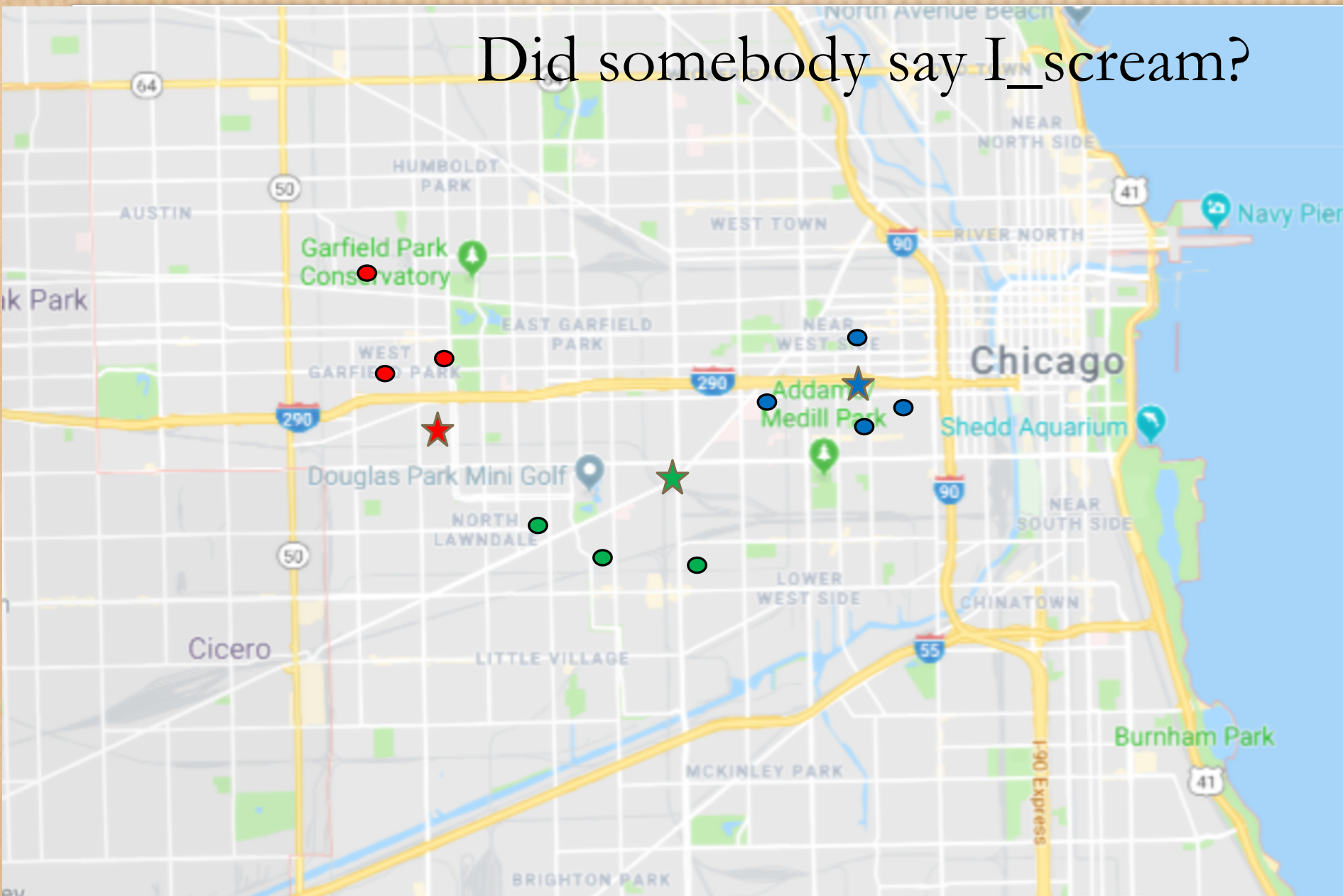
Highways shown include:

- 64
- 50
- 290
- 90
- 55
- 41

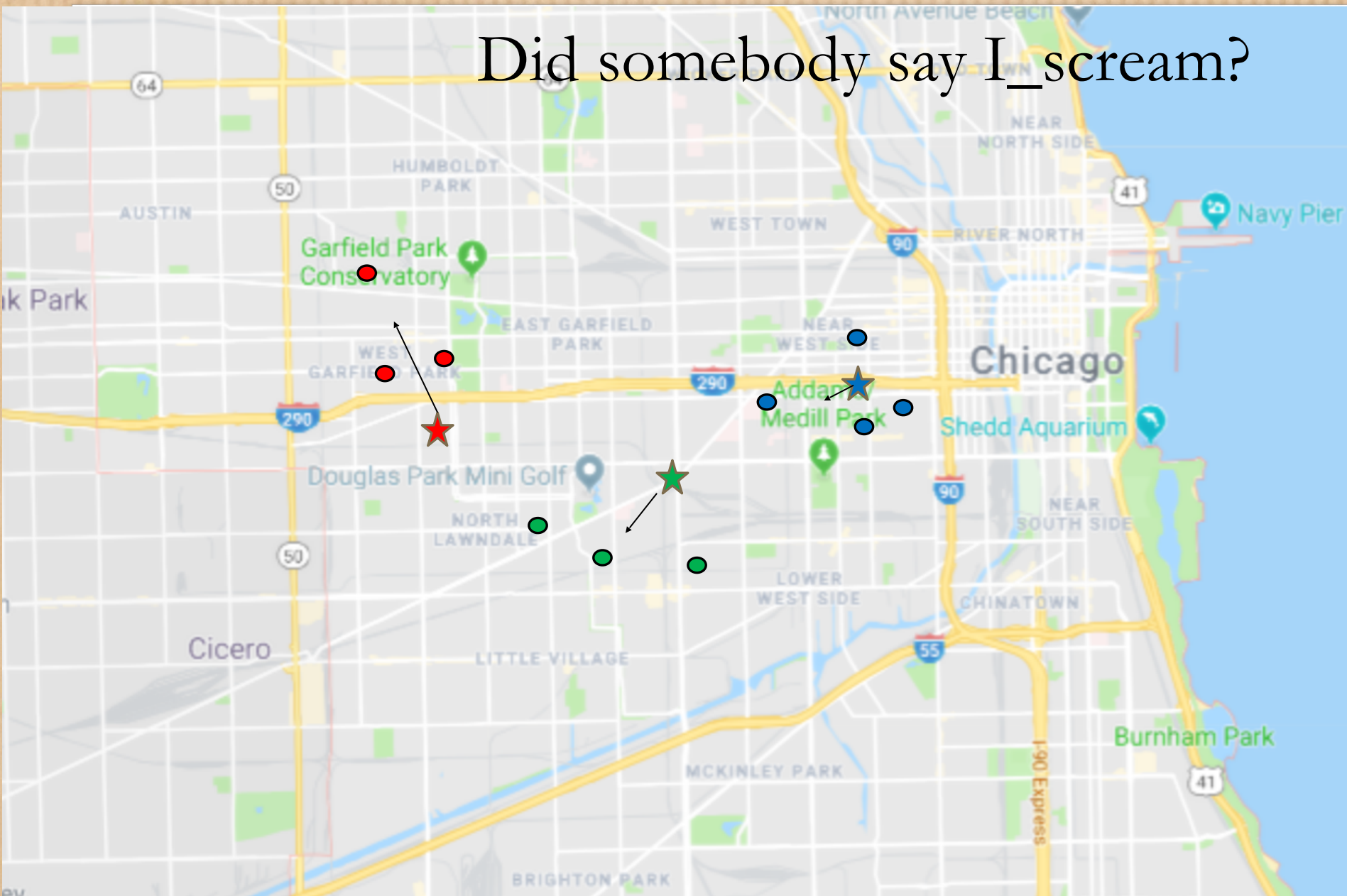
Other markers include:

- Red dots: Located in Garfield Park, West Garfield Park, and North Lawndale.
- Green dots: Located in Garfield Park, West Garfield Park, and North Lawndale.
- Blue dots: Located in Near West Side and Near South Side.
- Stars: A red star in West Garfield Park and a green star in North Lawndale.

Did somebody say I\_scream?

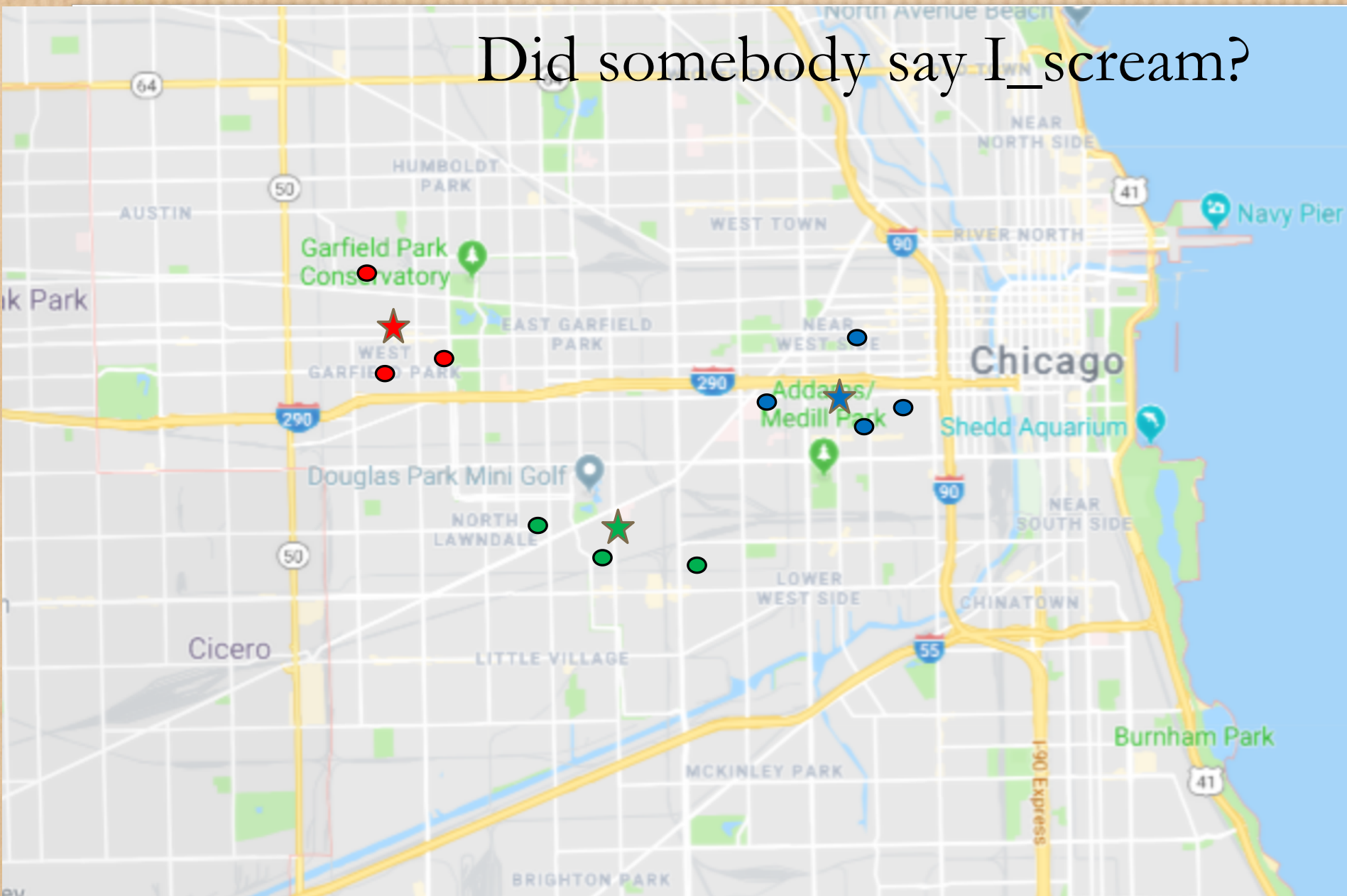


Did somebody say I\_scream?





Did somebody say I\_scream?



# Did somebody say I\_scream?

The map displays the following locations and markers:

- Sam**: Red dot near Garfield Park Conservatory.
- Shaun**: Red dot near West Garfield Park.
- Sandy**: Red dot near West Garfield Park.
- Robin**: Blue dot near Addams Park.
- Robbie**: Blue dot near Addams Park.
- Ronnie**: Blue dot near Addams Park.
- Dan**: Green dot near Douglas Park Mini Golf.
- Danna**: Green dot near Douglas Park Mini Golf.
- Dane**: Green dot near Douglas Park Mini Golf.

Other landmarks and parks visible on the map include Humboldt Park, Garfield Park Conservatory, East Garfield Park, West Garfield Park, Douglas Park Mini Golf, North Lawndale, Little Village, Mckinley Park, Brighton Park, Burnham Park, Navy Pier, and the Chicago River.

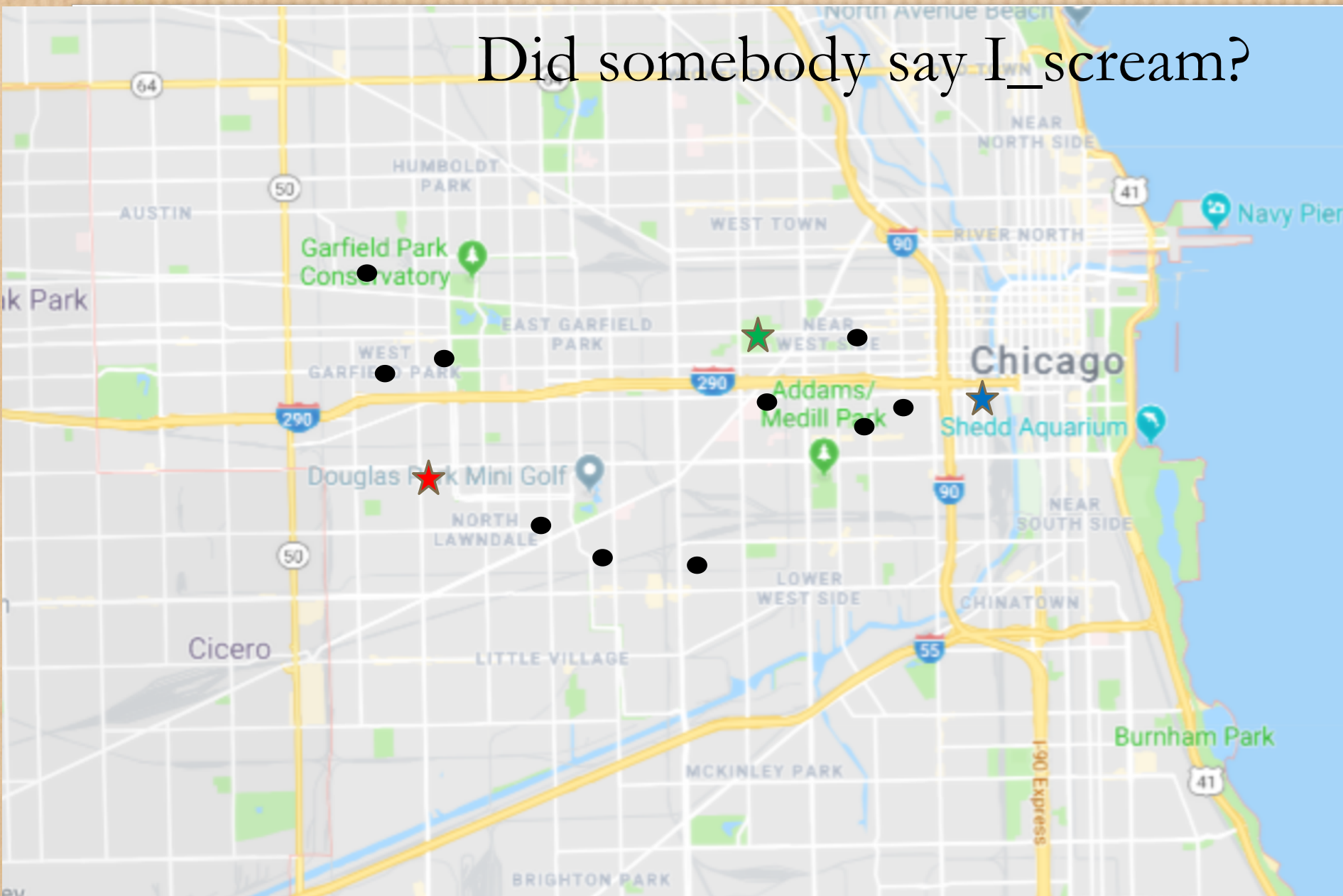
# Pros & Cons

---

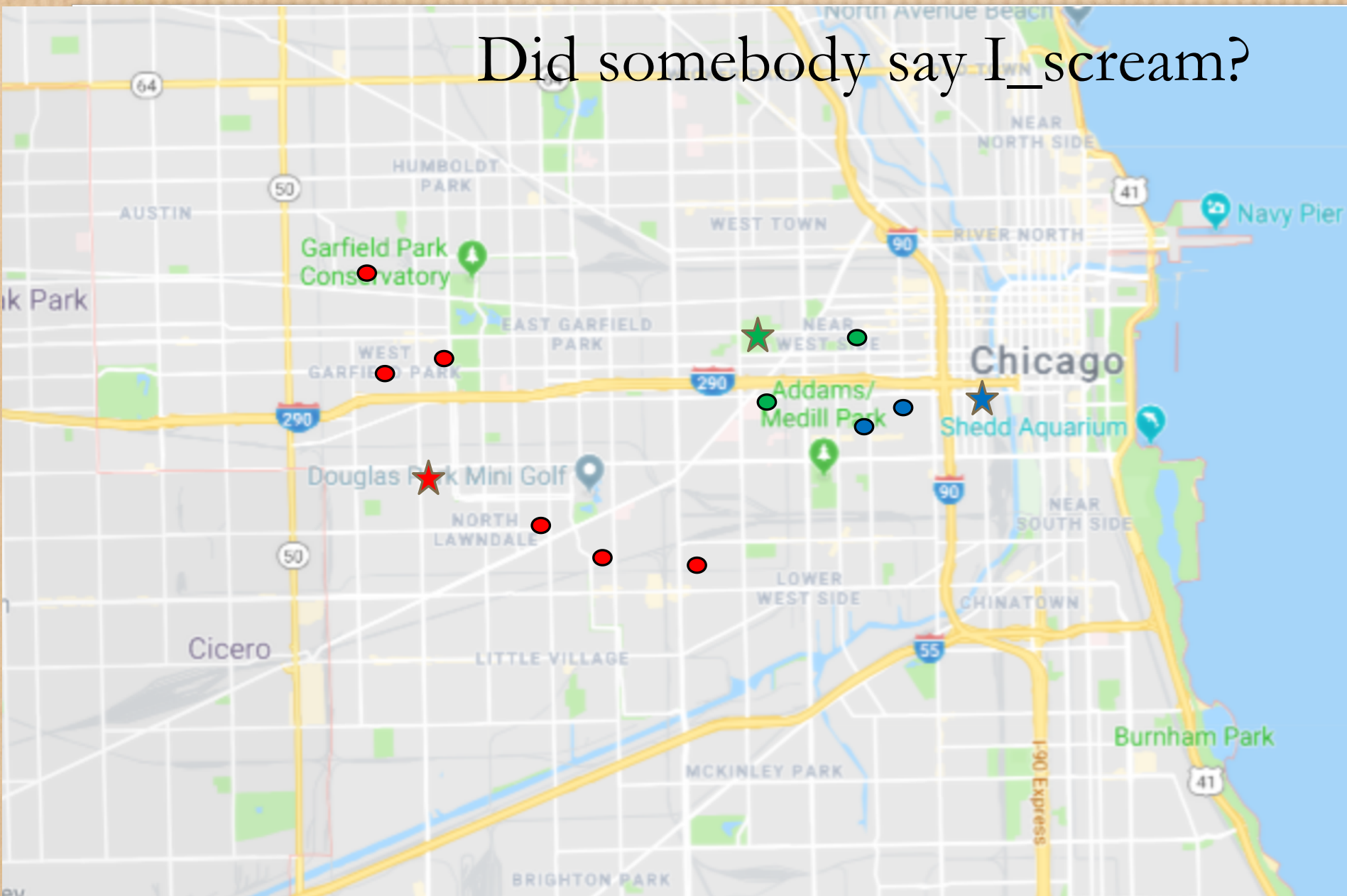
- Low complexity *complexity is  $O(nkt)$ , where  $t = \text{\#iterations}$*
- Clusters are sensitive to initial assignment of centroids
- Necessity of specifying  $k$
- Sensitive to noise and outlier data points



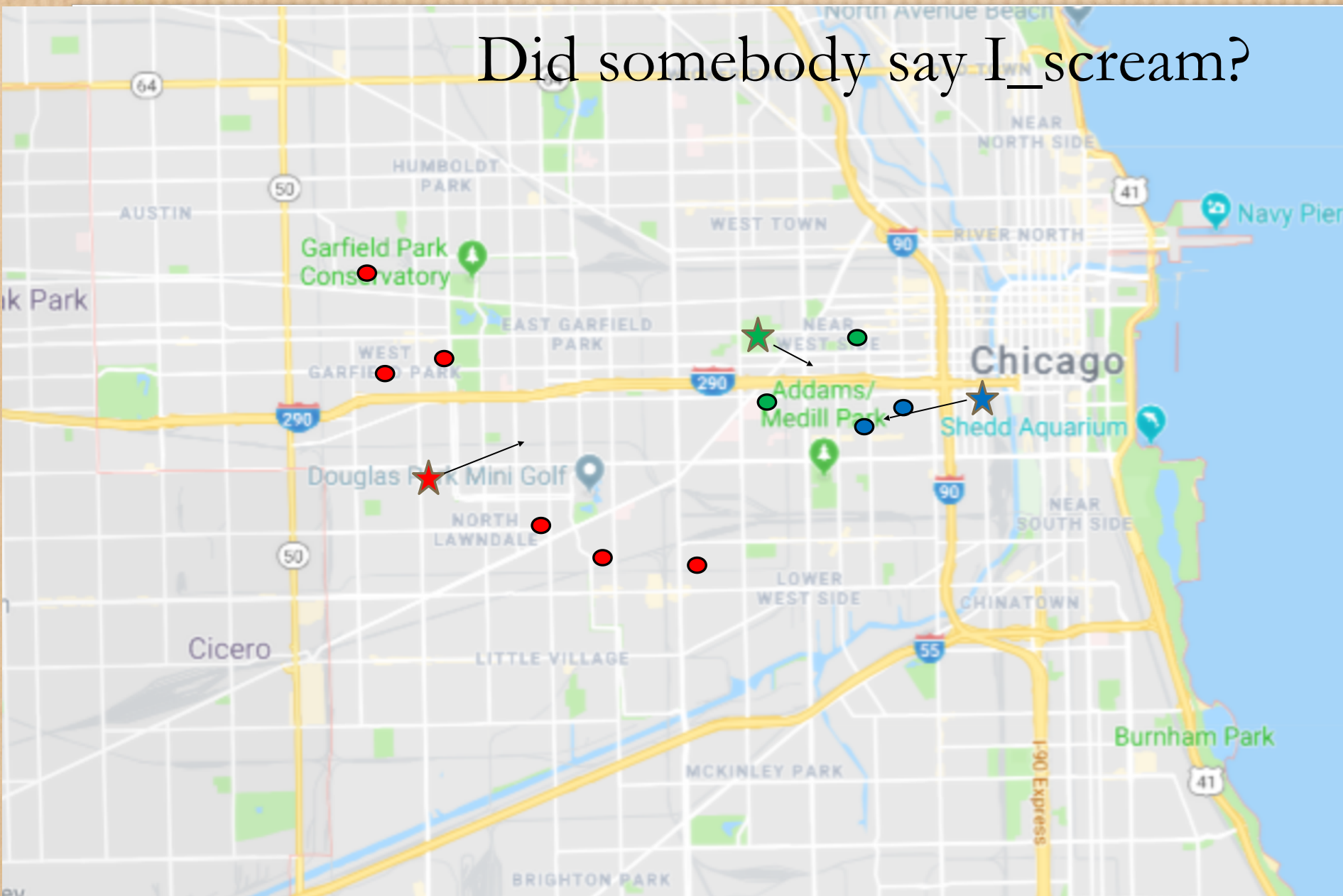
Did somebody say I\_scream?



Did somebody say I\_scream?

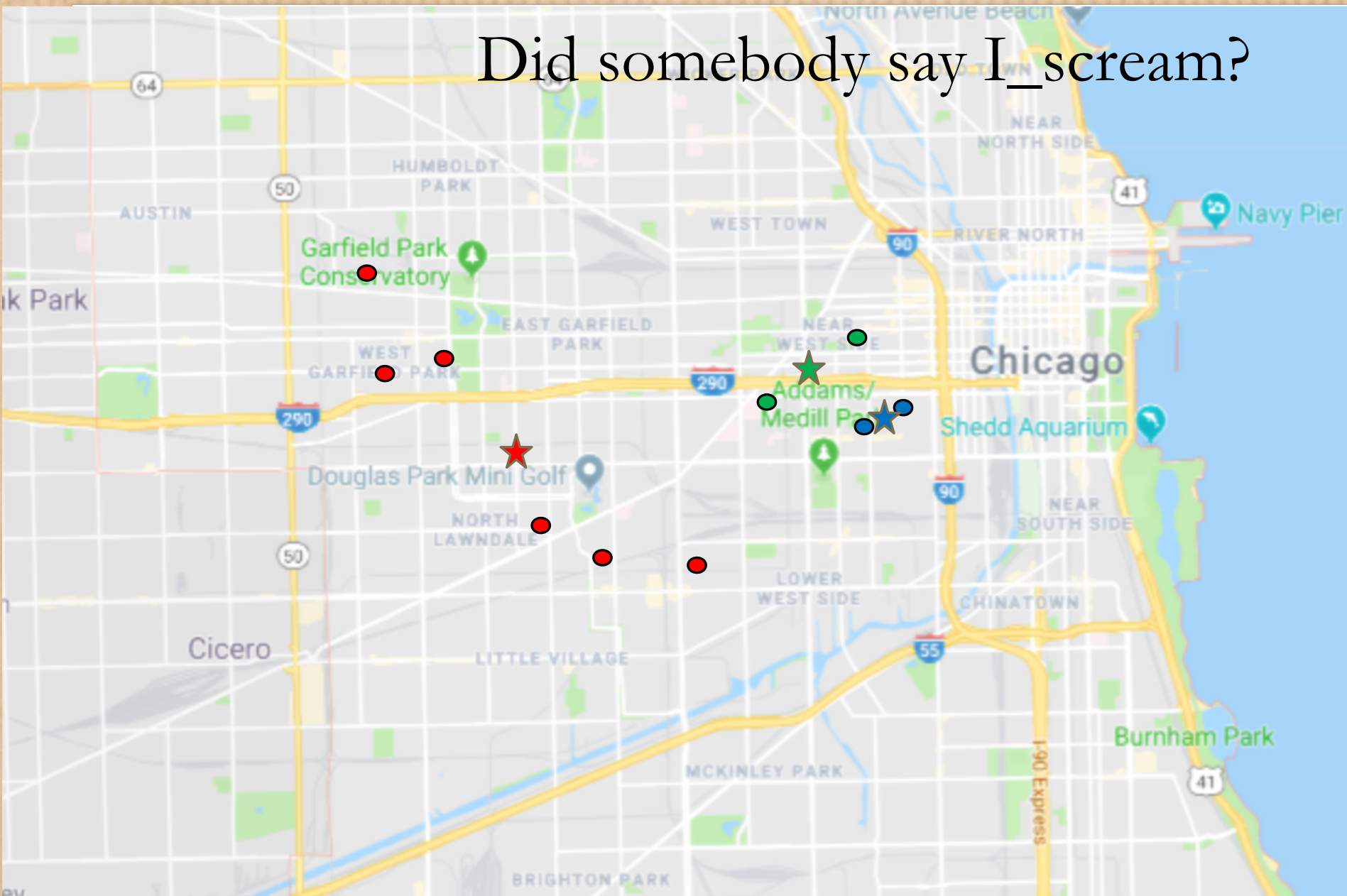


Did somebody say I\_scream?

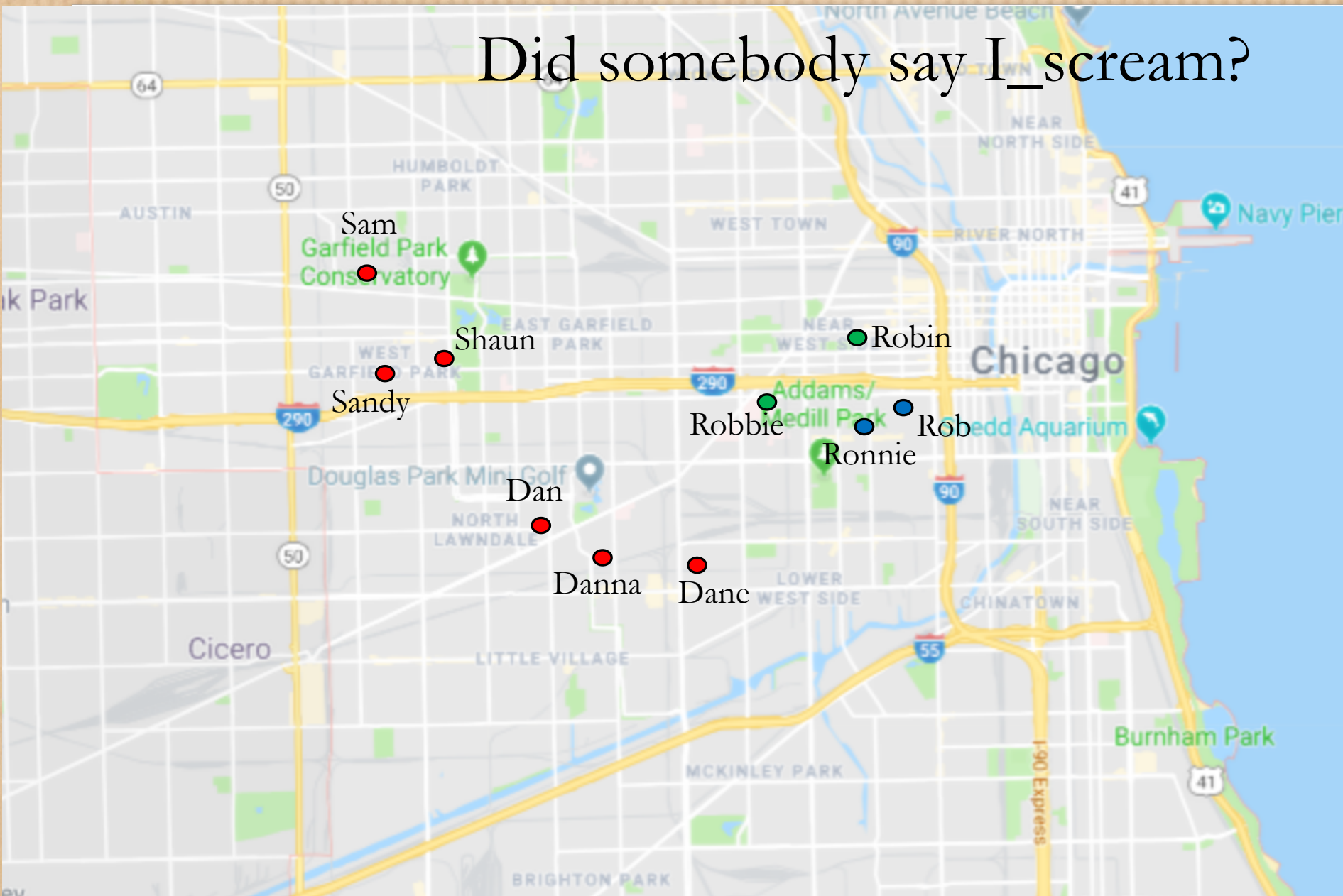




Did somebody say I\_scream?



Did somebody say I\_scream?



---

K means is a non-deterministic algorithm



# ELBOW method

## (How Many Clusters?)

---



1 cluster



2 clusters



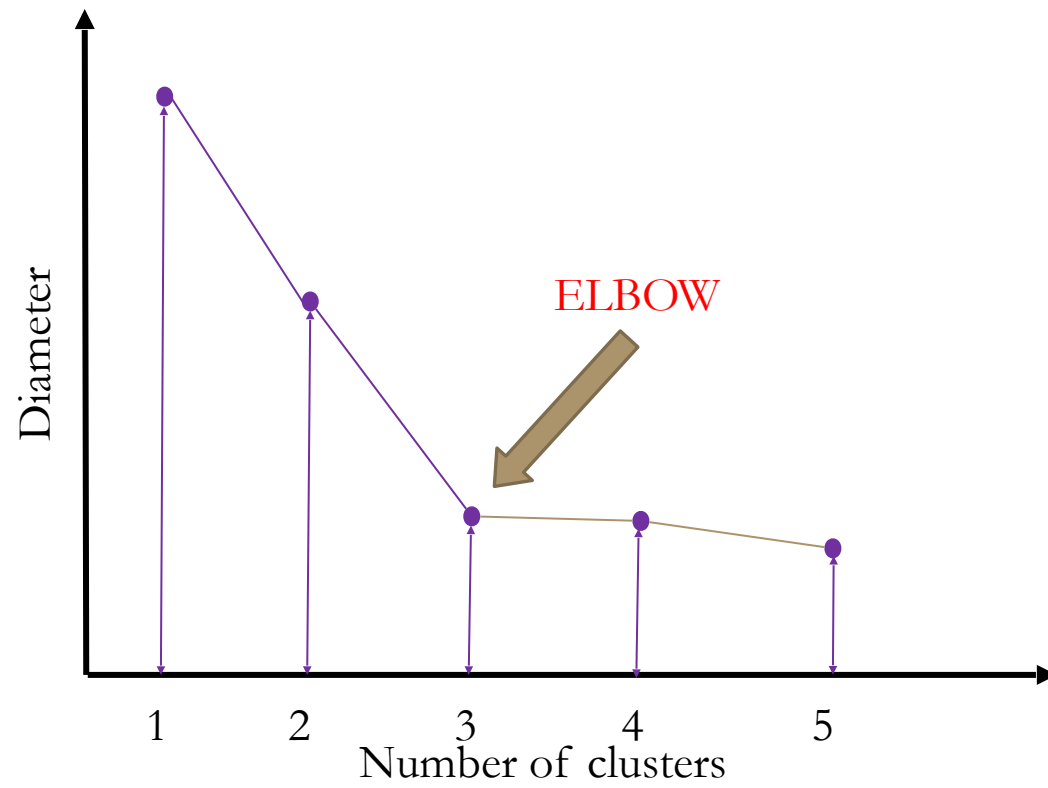
3 clusters



4 clusters



5 clusters

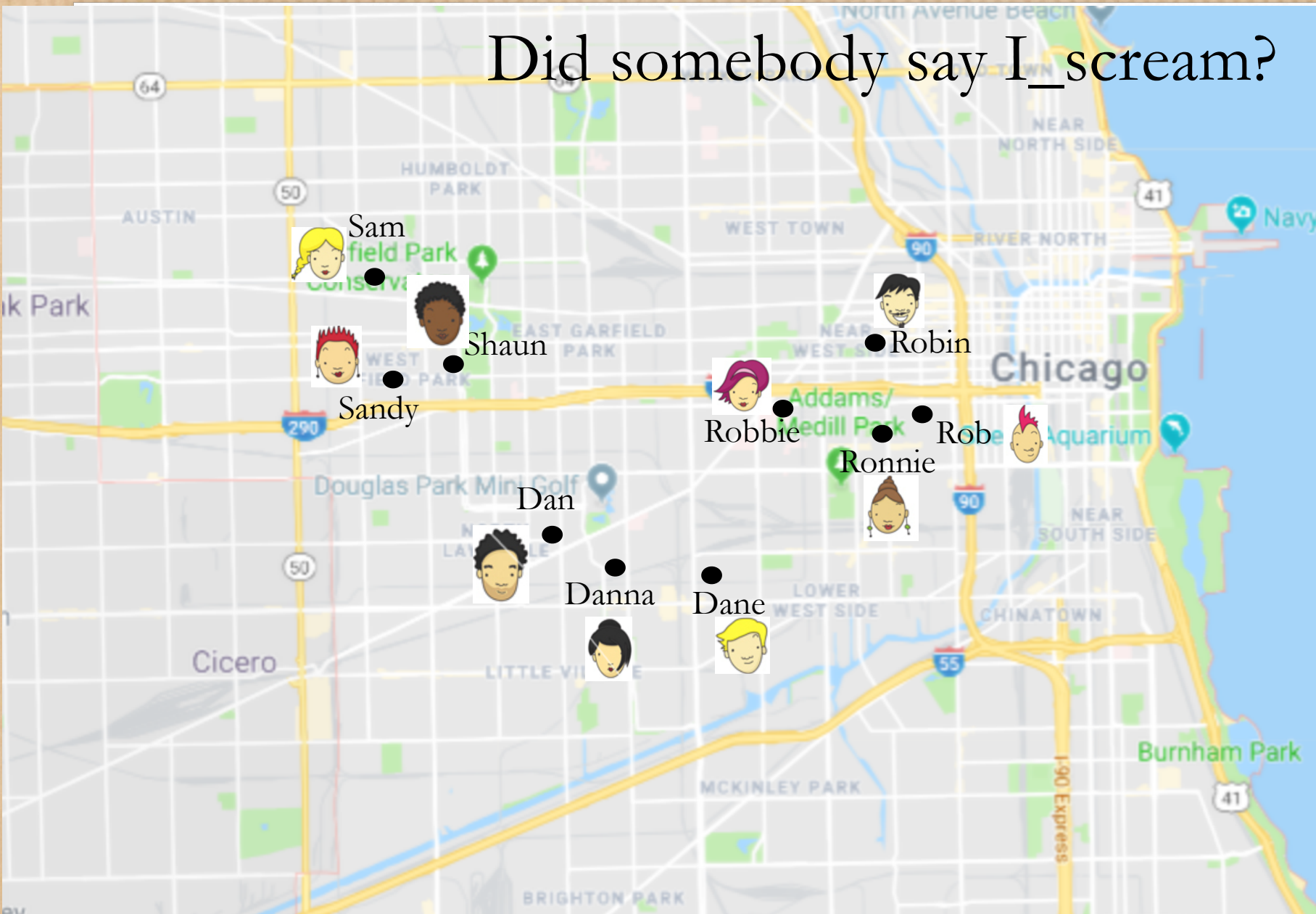


---

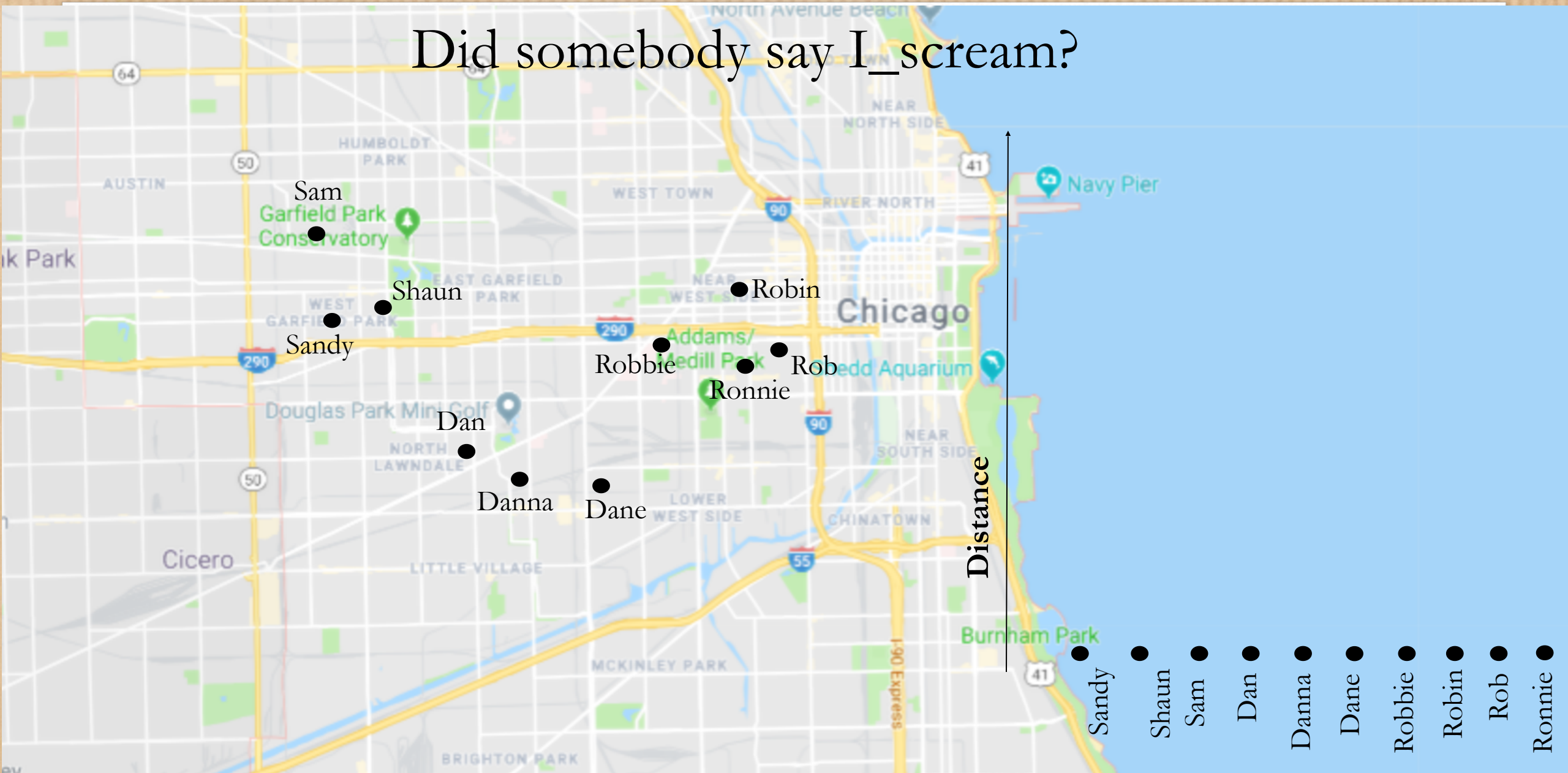
# Hierarchical Clustering



# Did somebody say I\_scream?

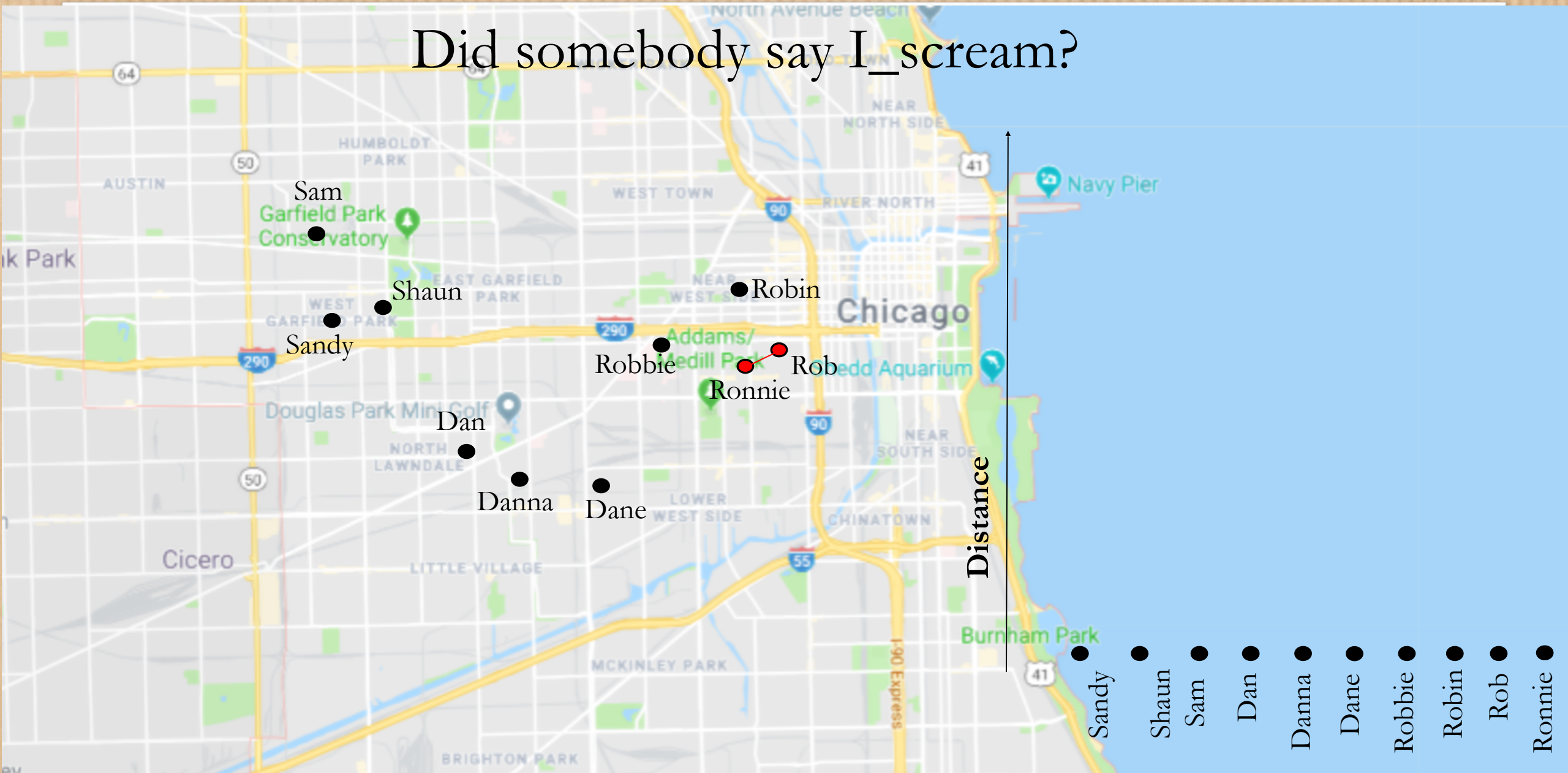


# Did somebody say I\_scream?



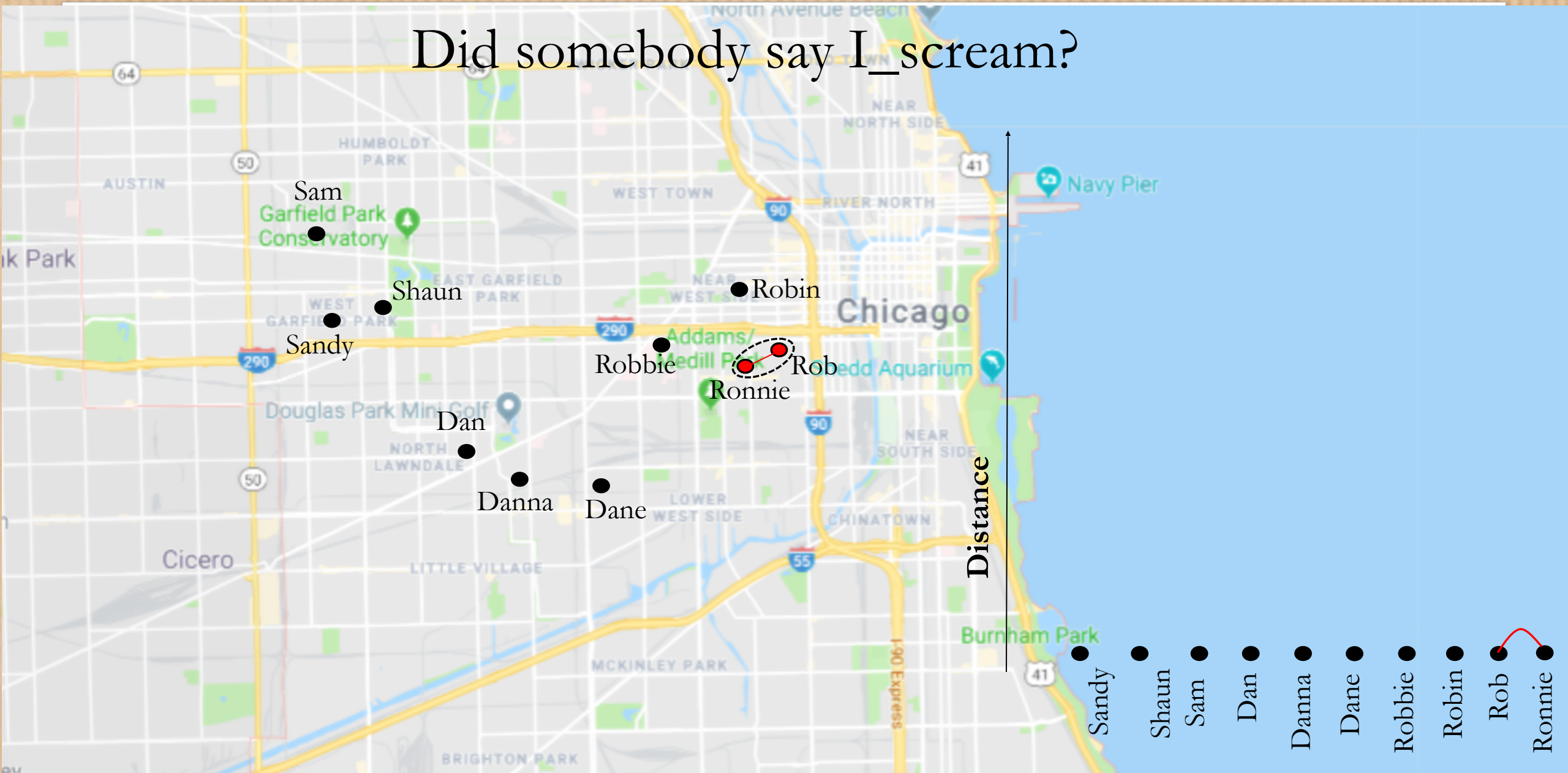


# Did somebody say I\_scream?

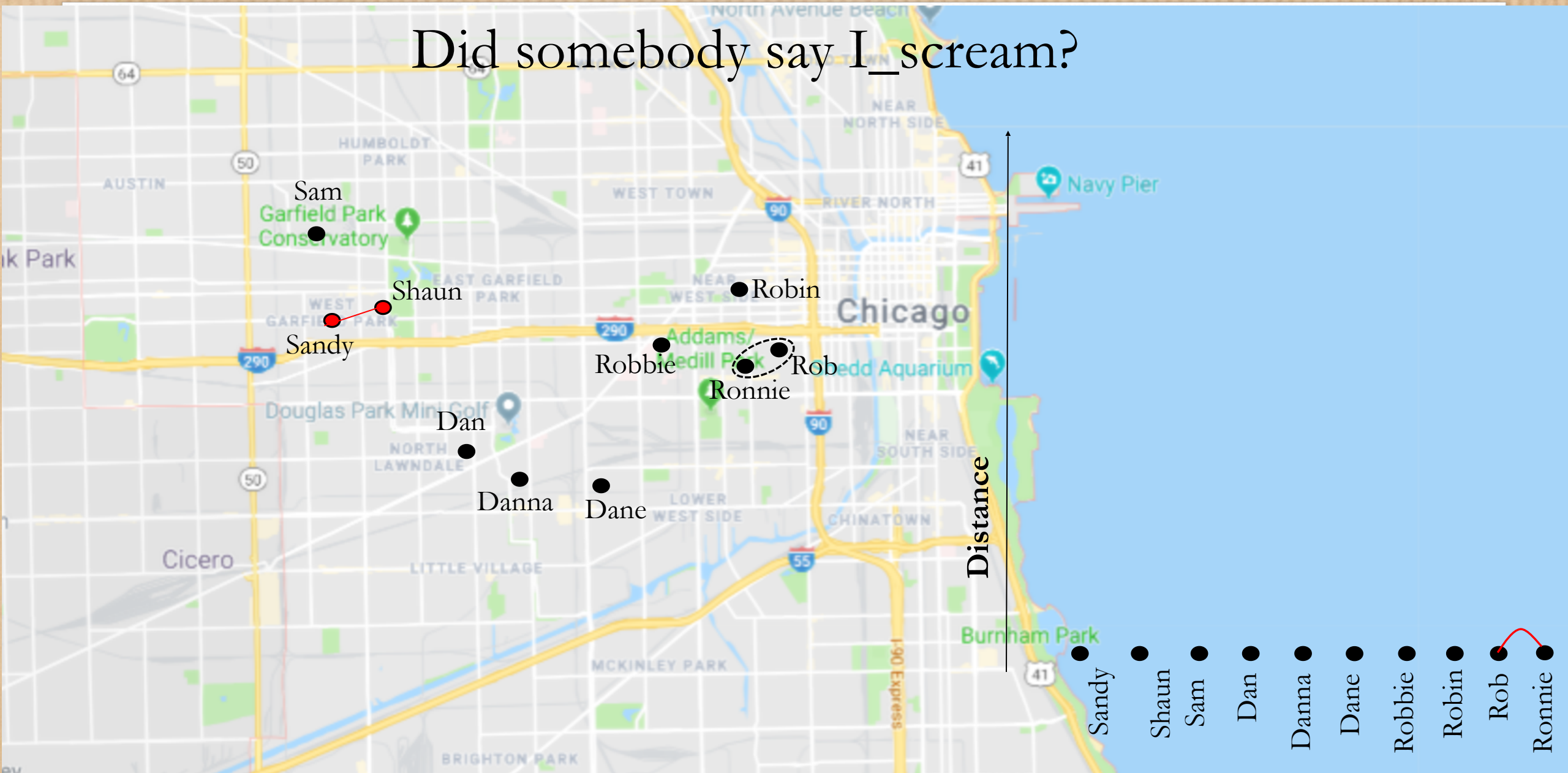




# Did somebody say I\_scream?

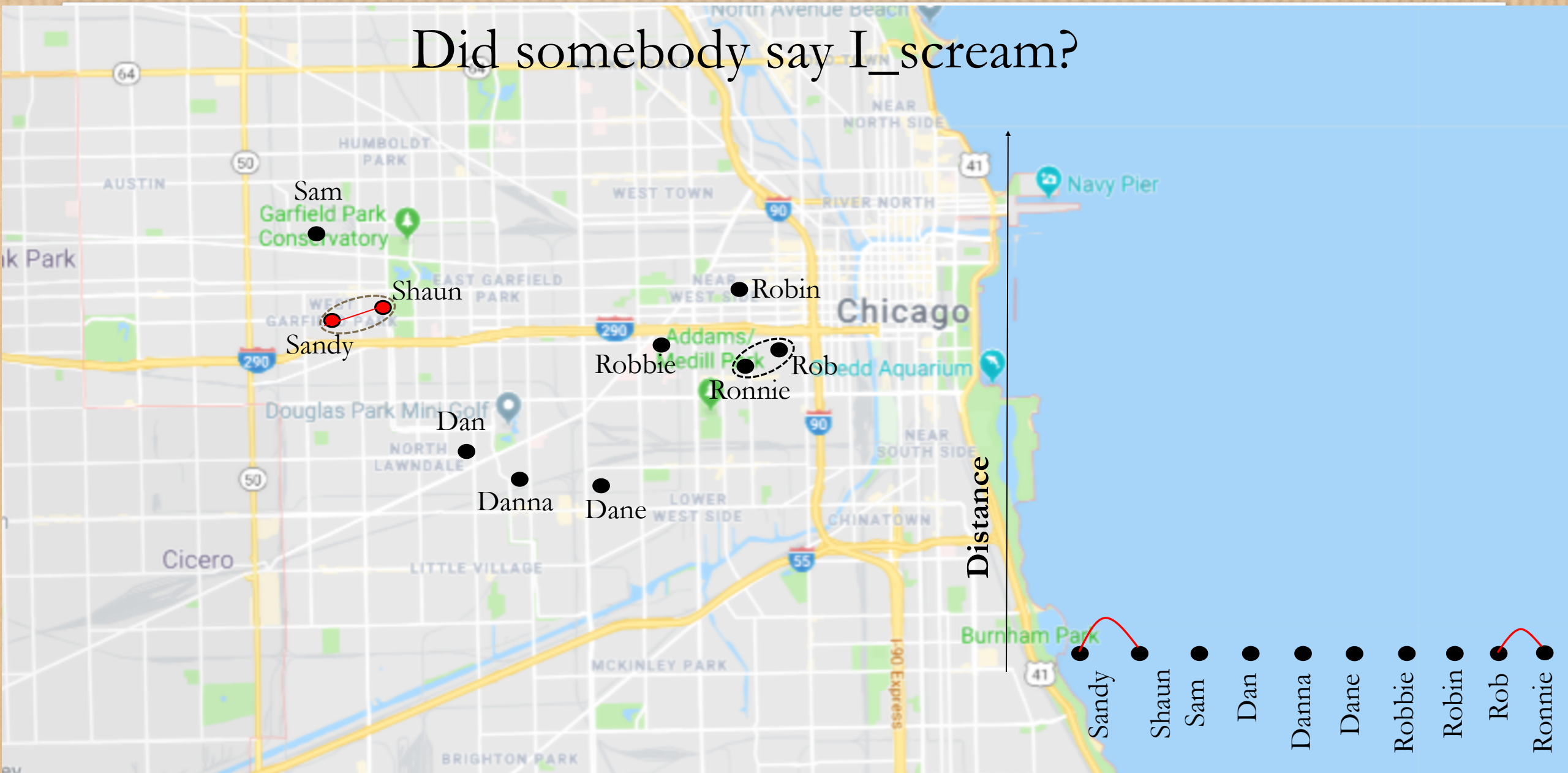


# Did somebody say I\_scream?



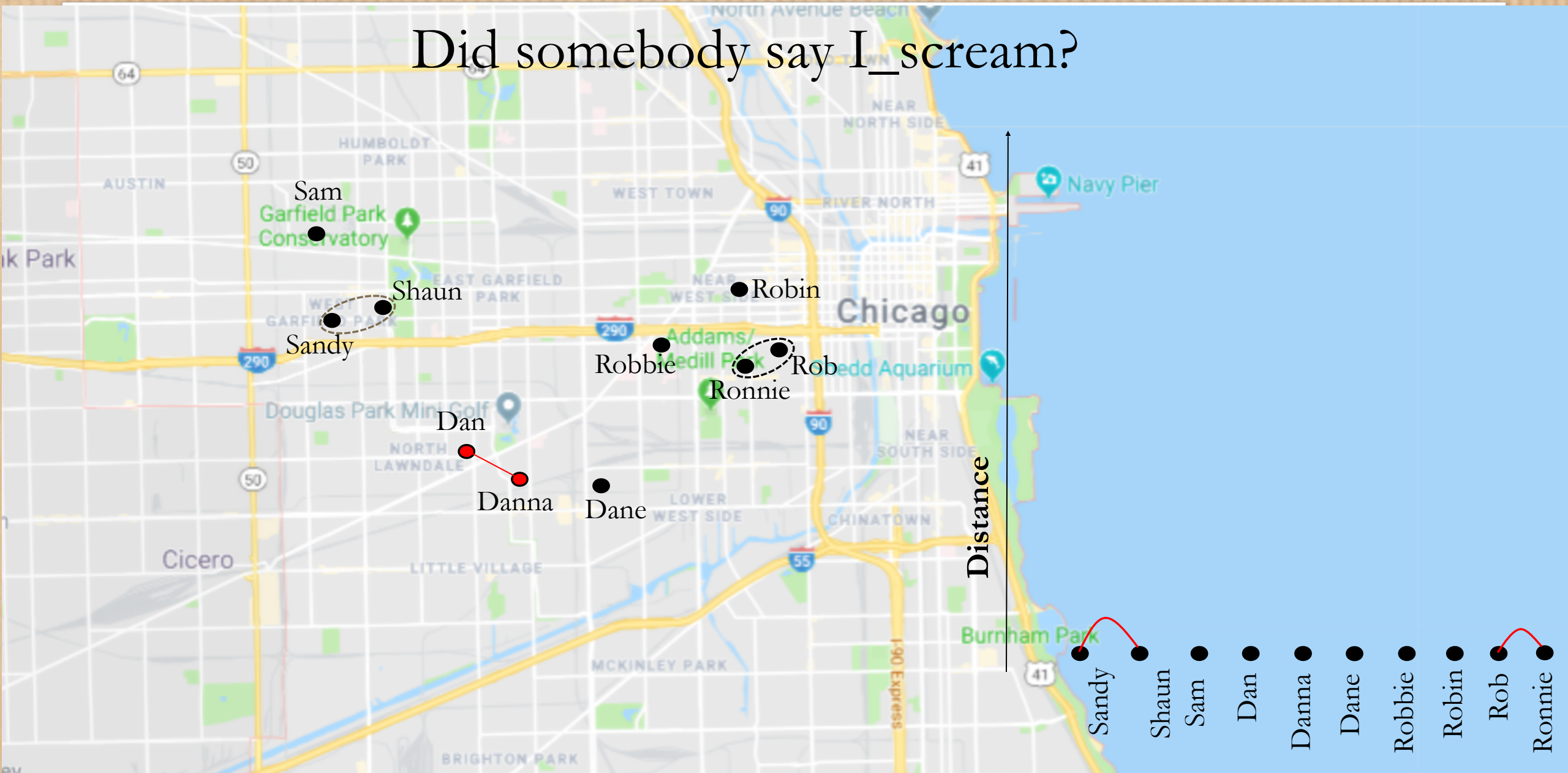


# Did somebody say I\_scream?

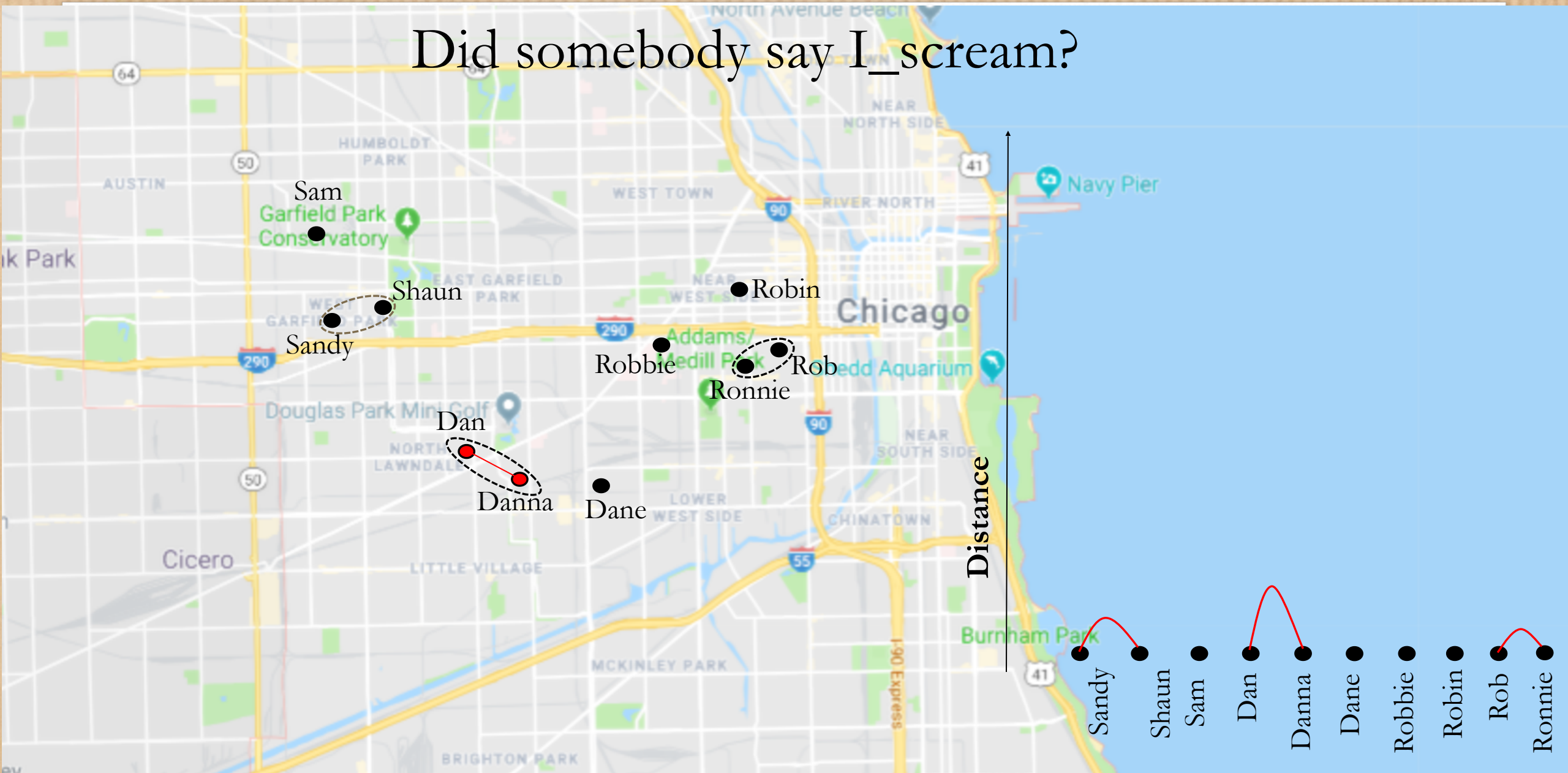




# Did somebody say I\_scream?

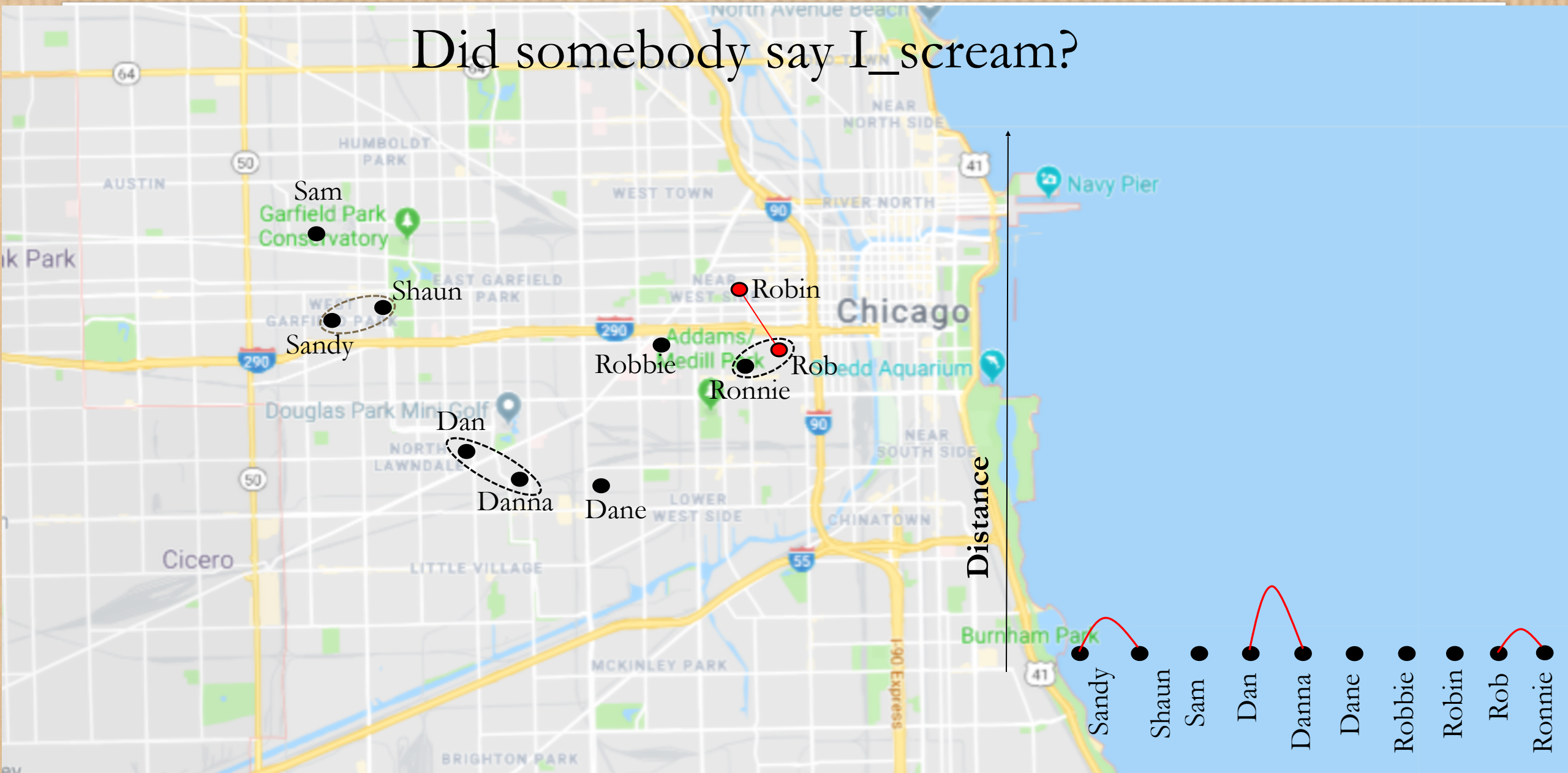


# Did somebody say I\_scream?



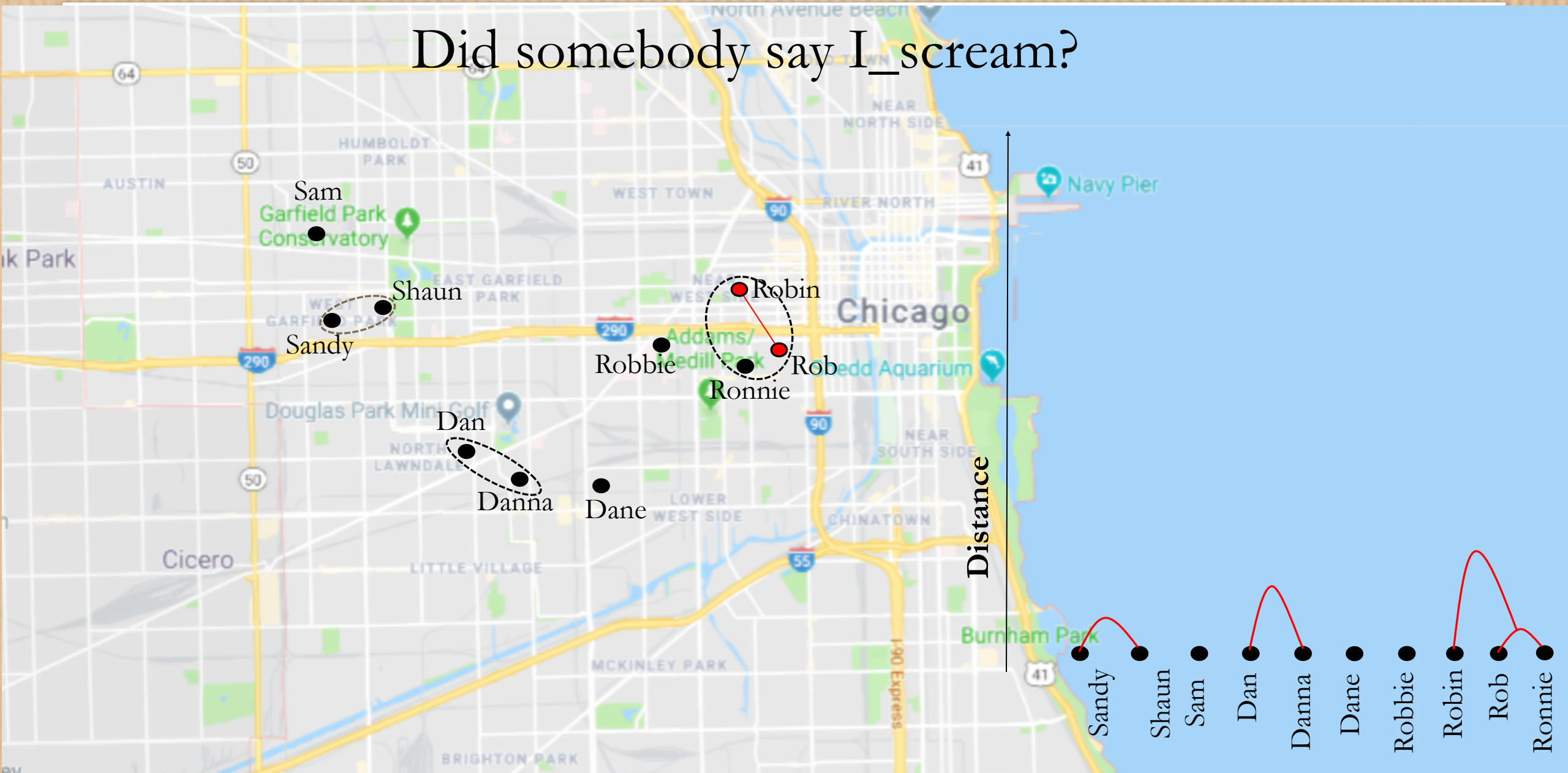


# Did somebody say I\_scream?

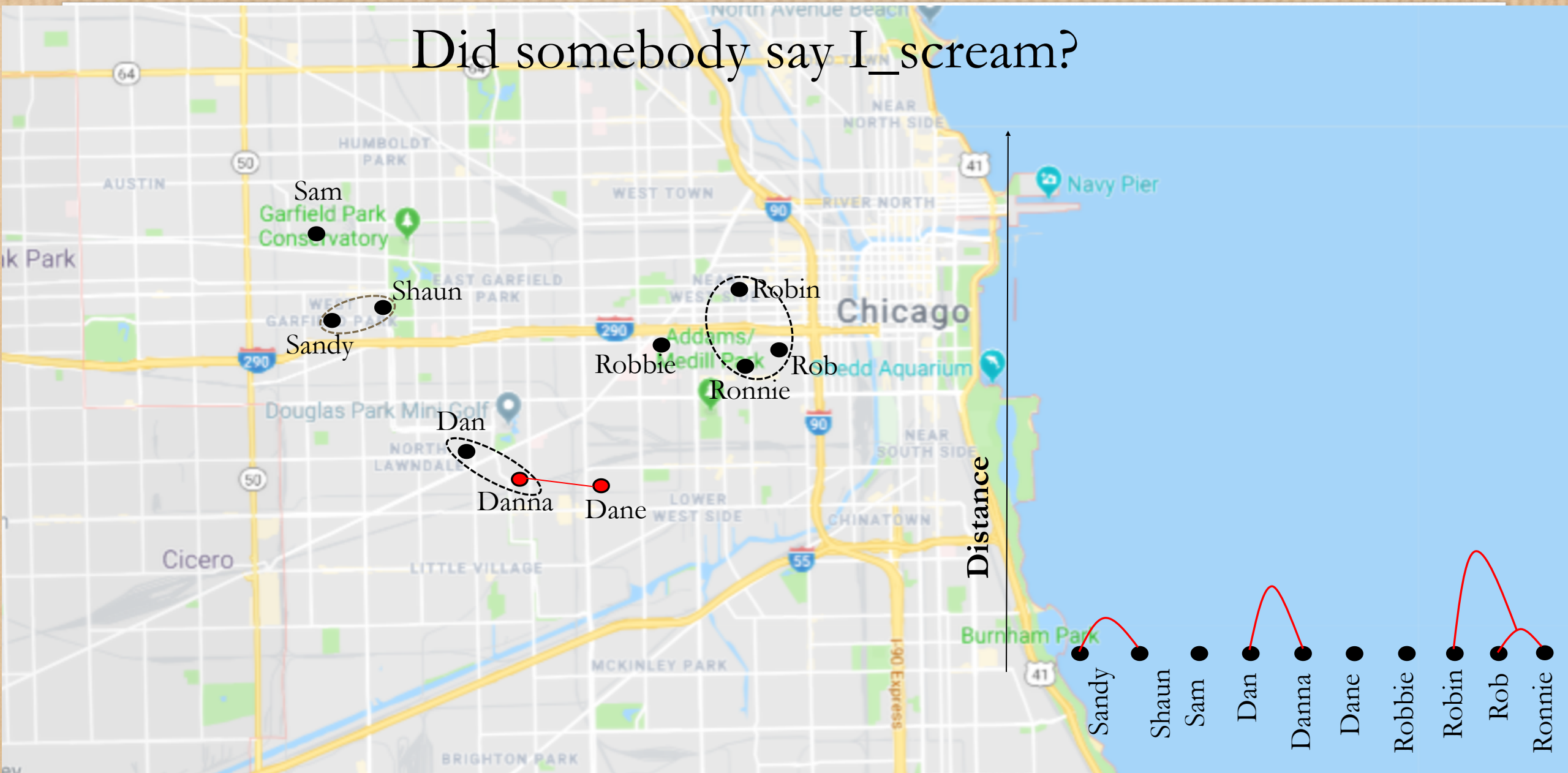




# Did somebody say I\_scream?

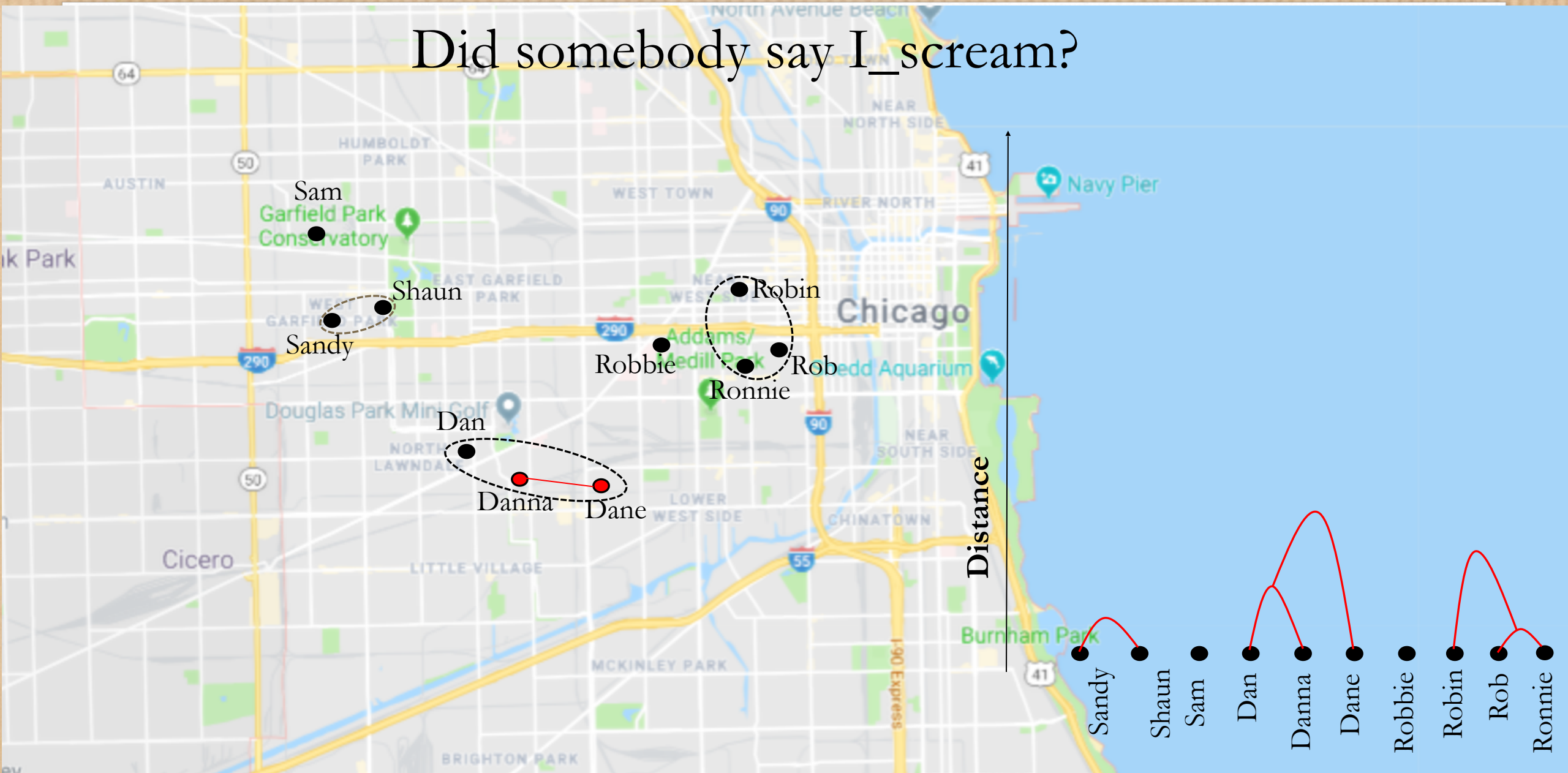


# Did somebody say I\_scream?



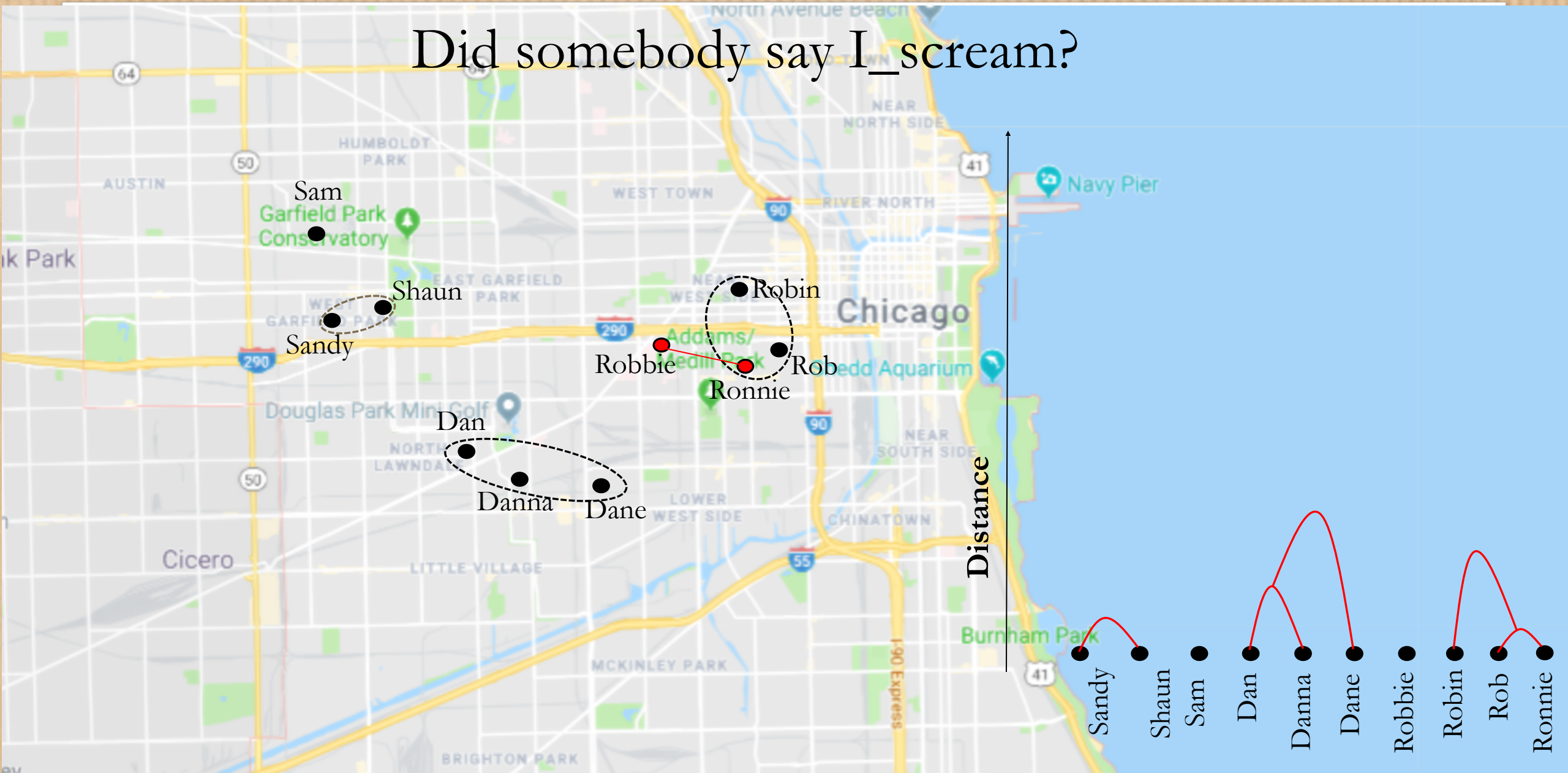


# Did somebody say I\_scream?

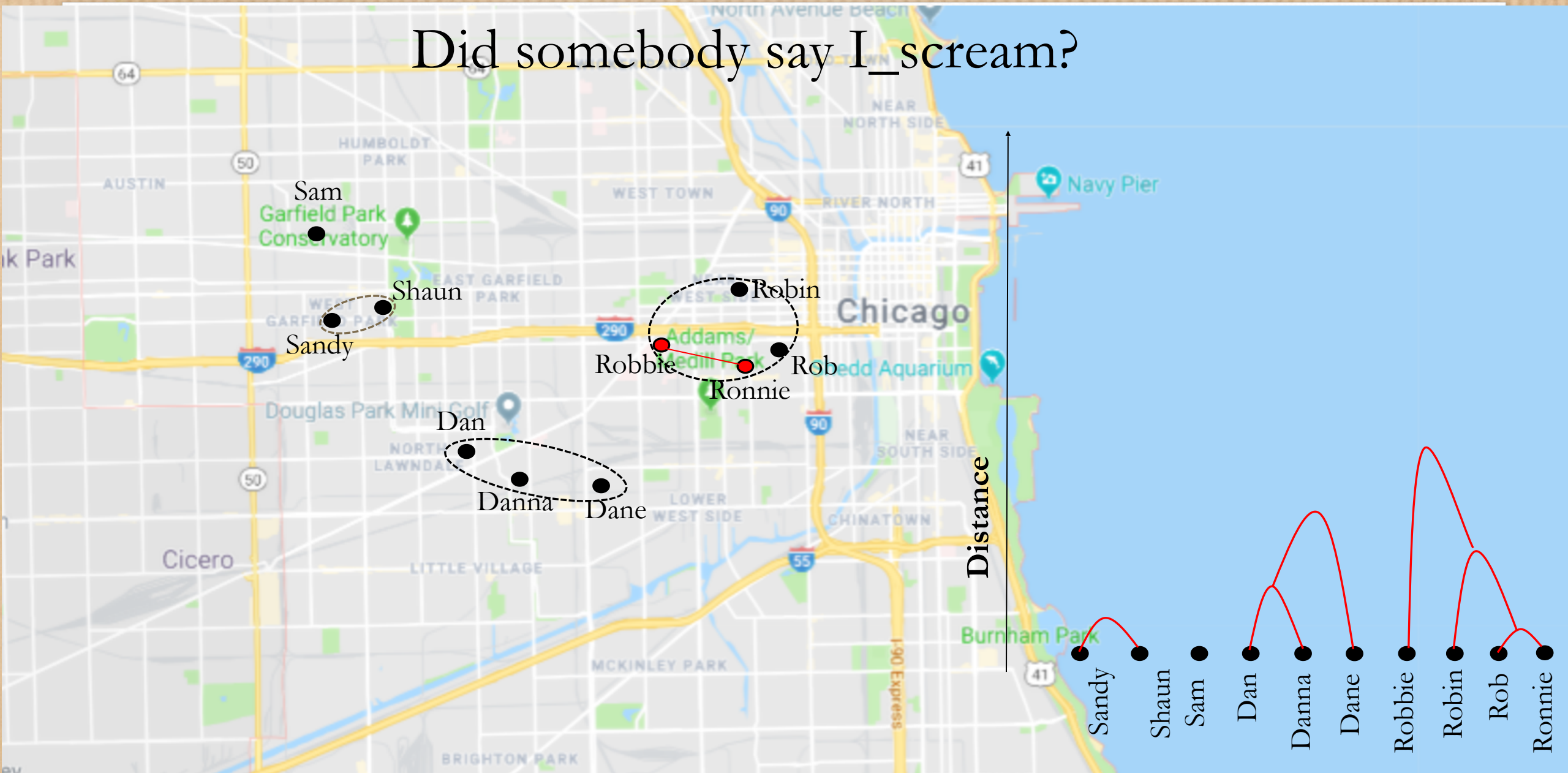




# Did somebody say I\_scream?

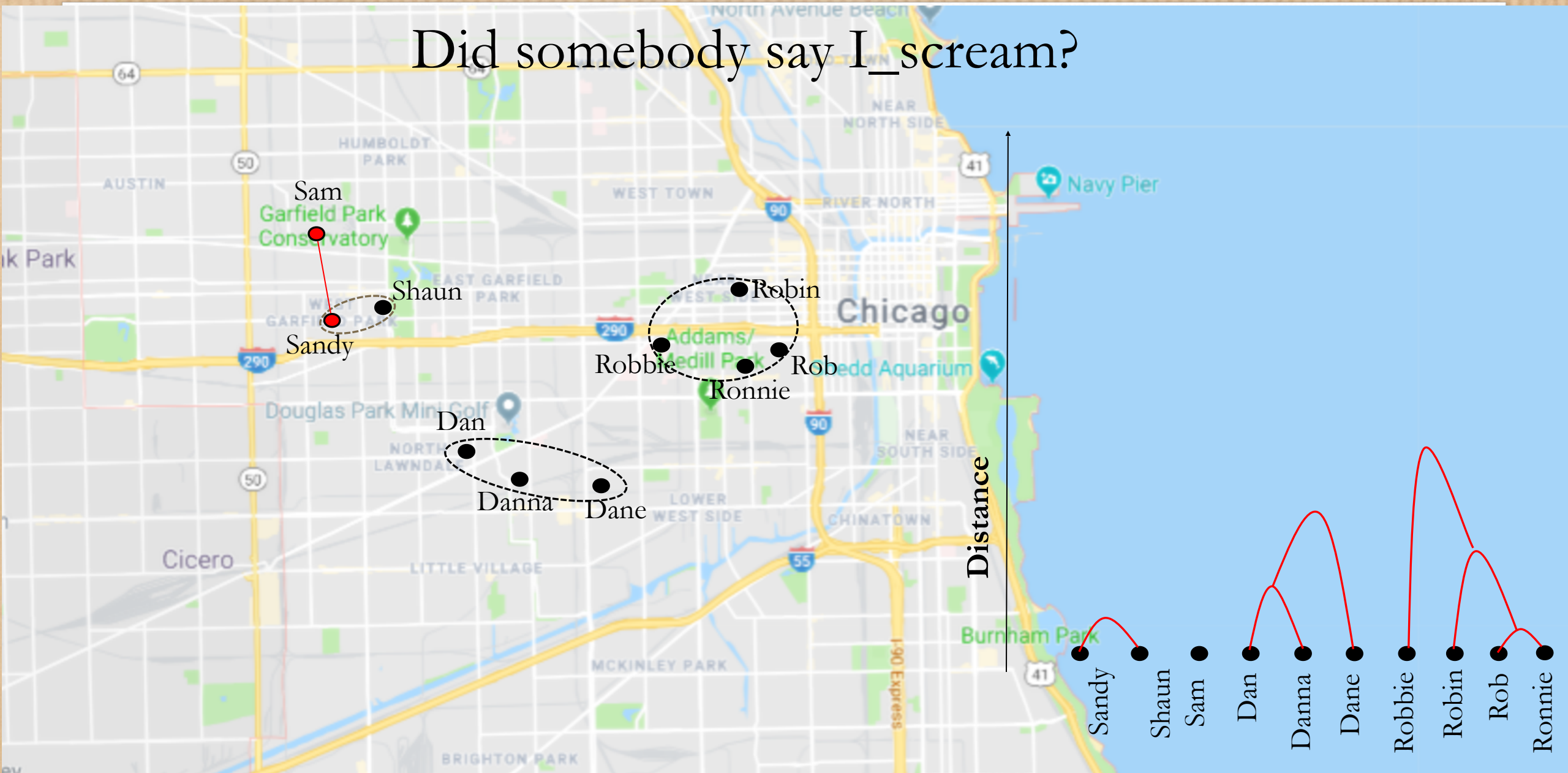


# Did somebody say I\_scream?



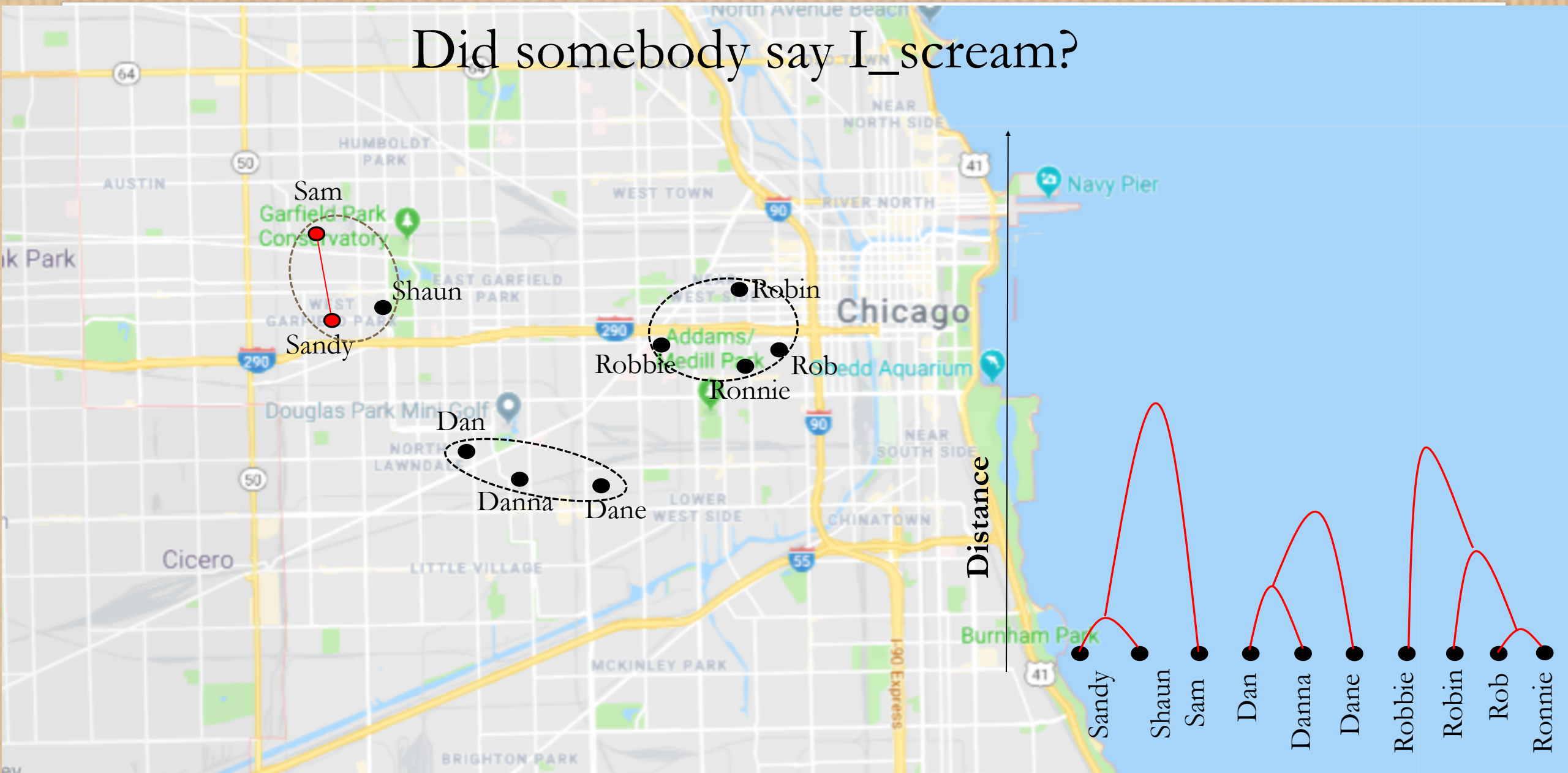


# Did somebody say I\_scream?

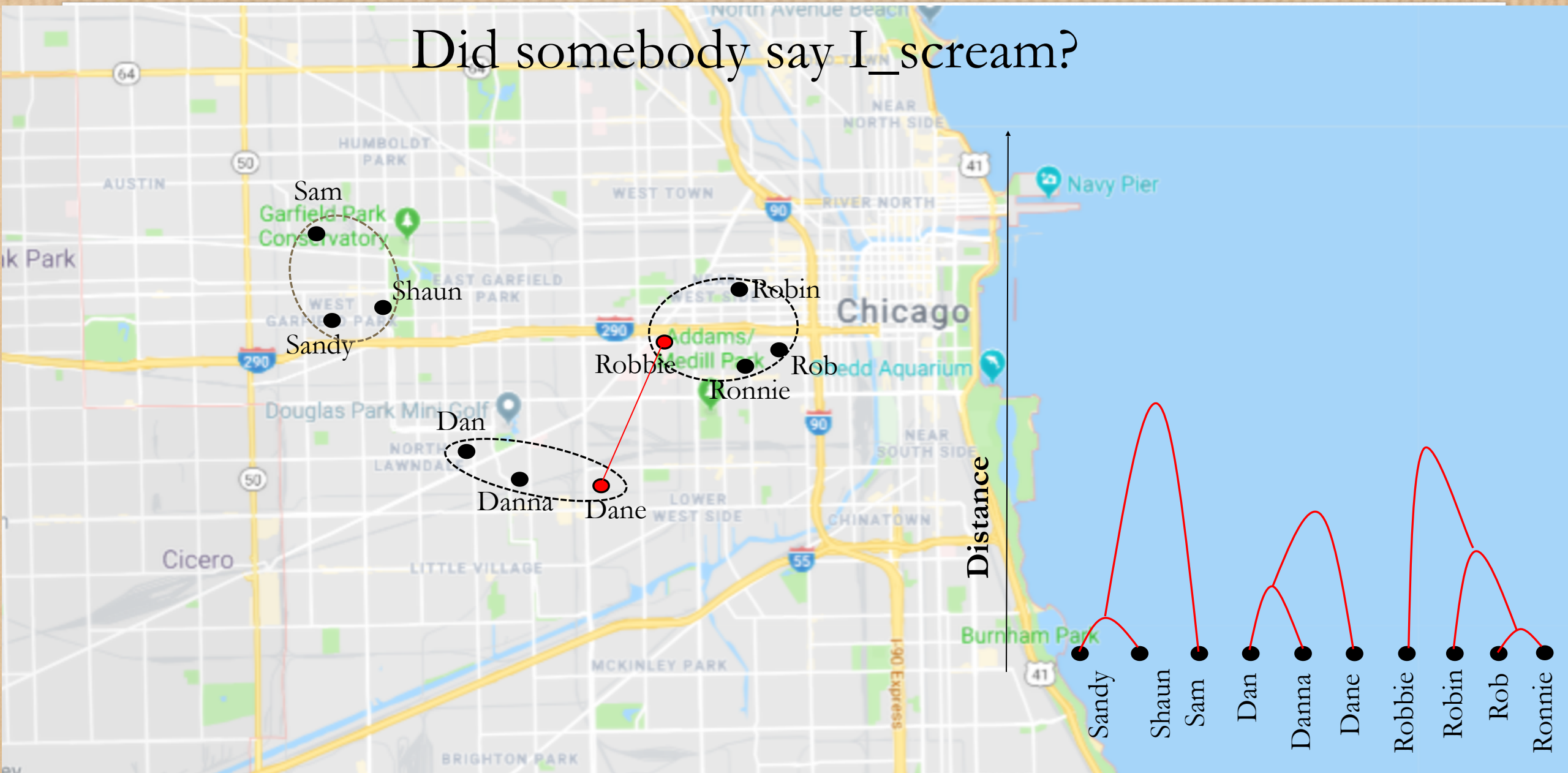




# Did somebody say I\_scream?

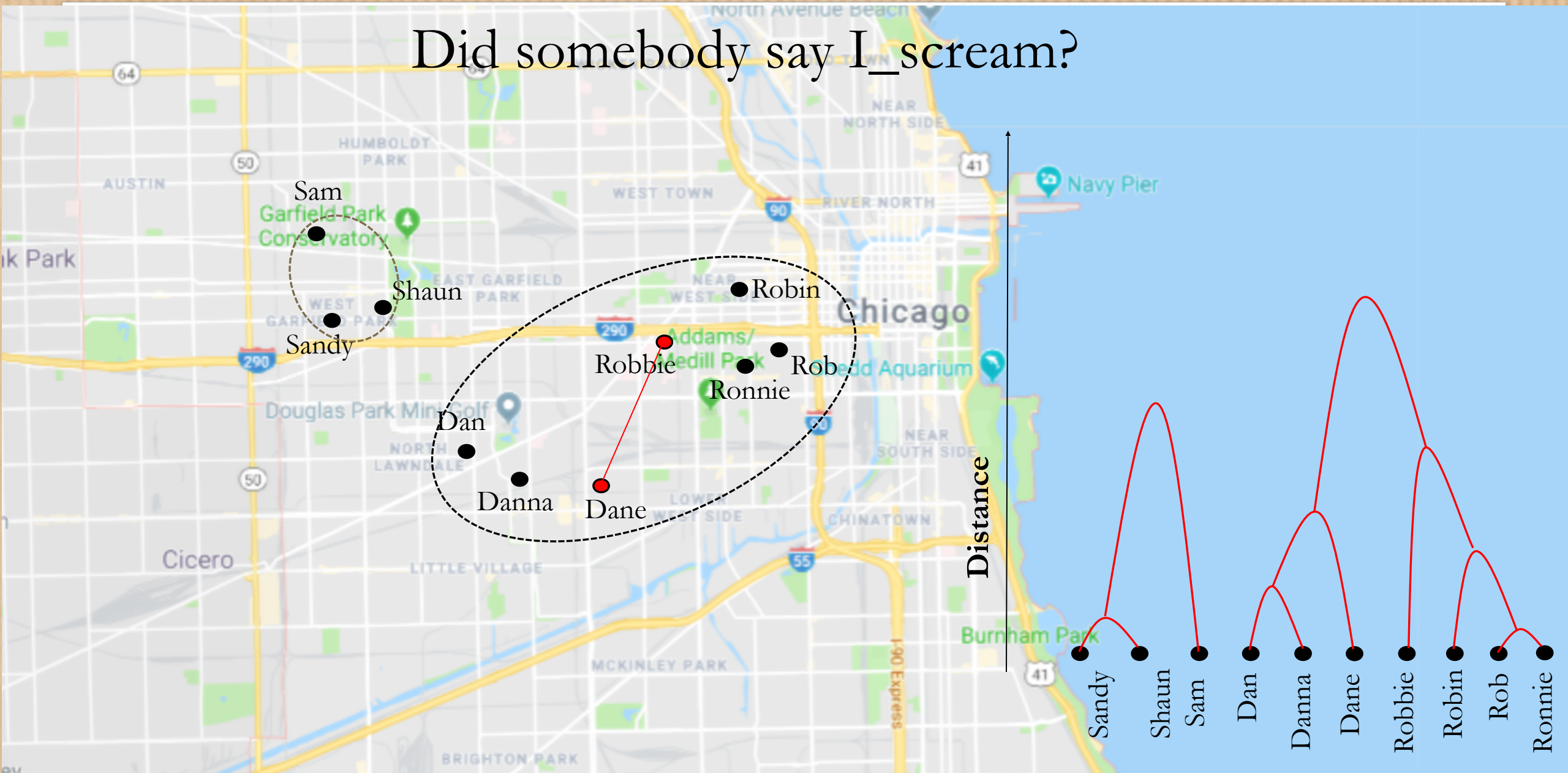


# Did somebody say I\_scream?



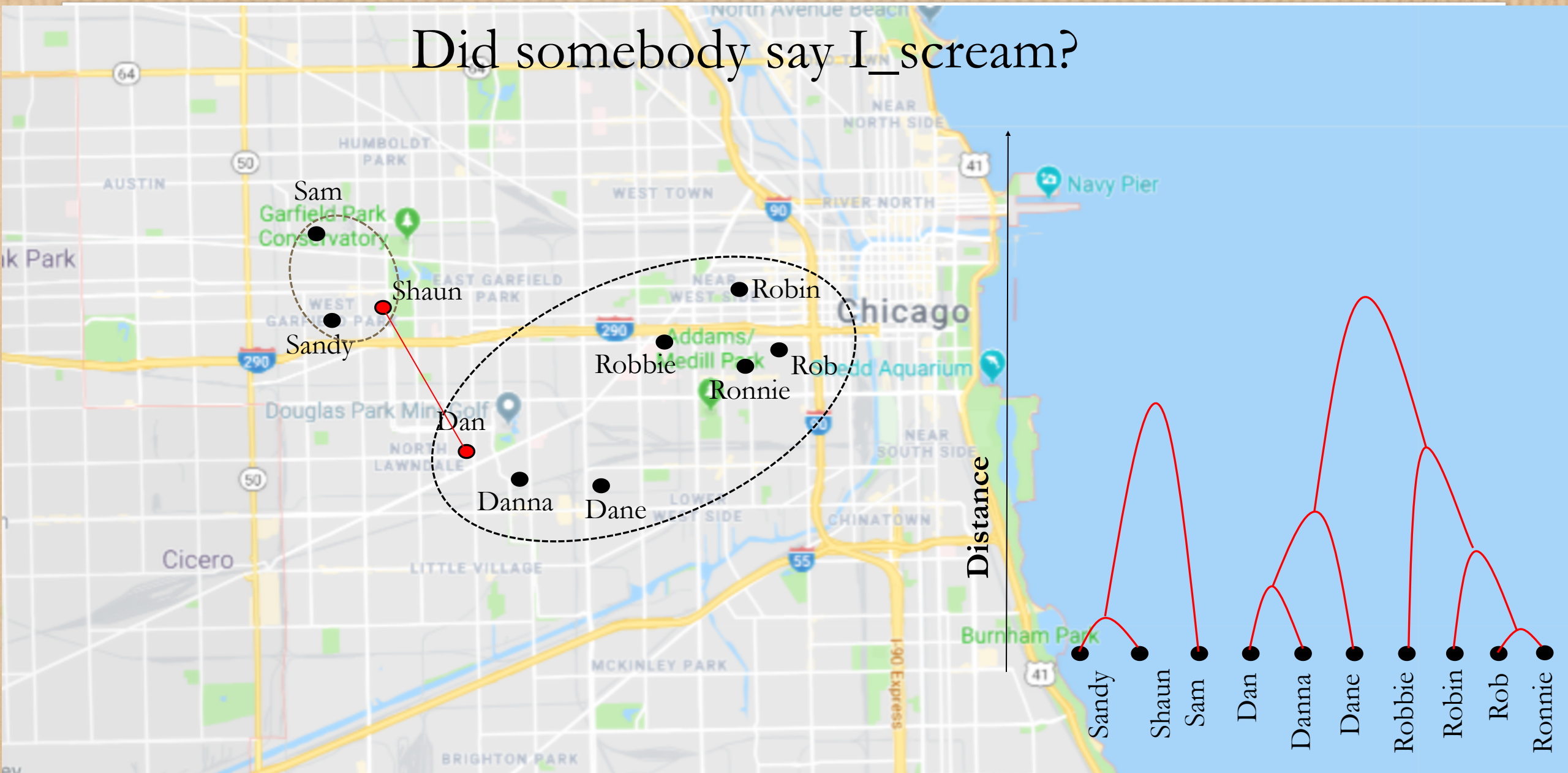


# Did somebody say I\_scream?

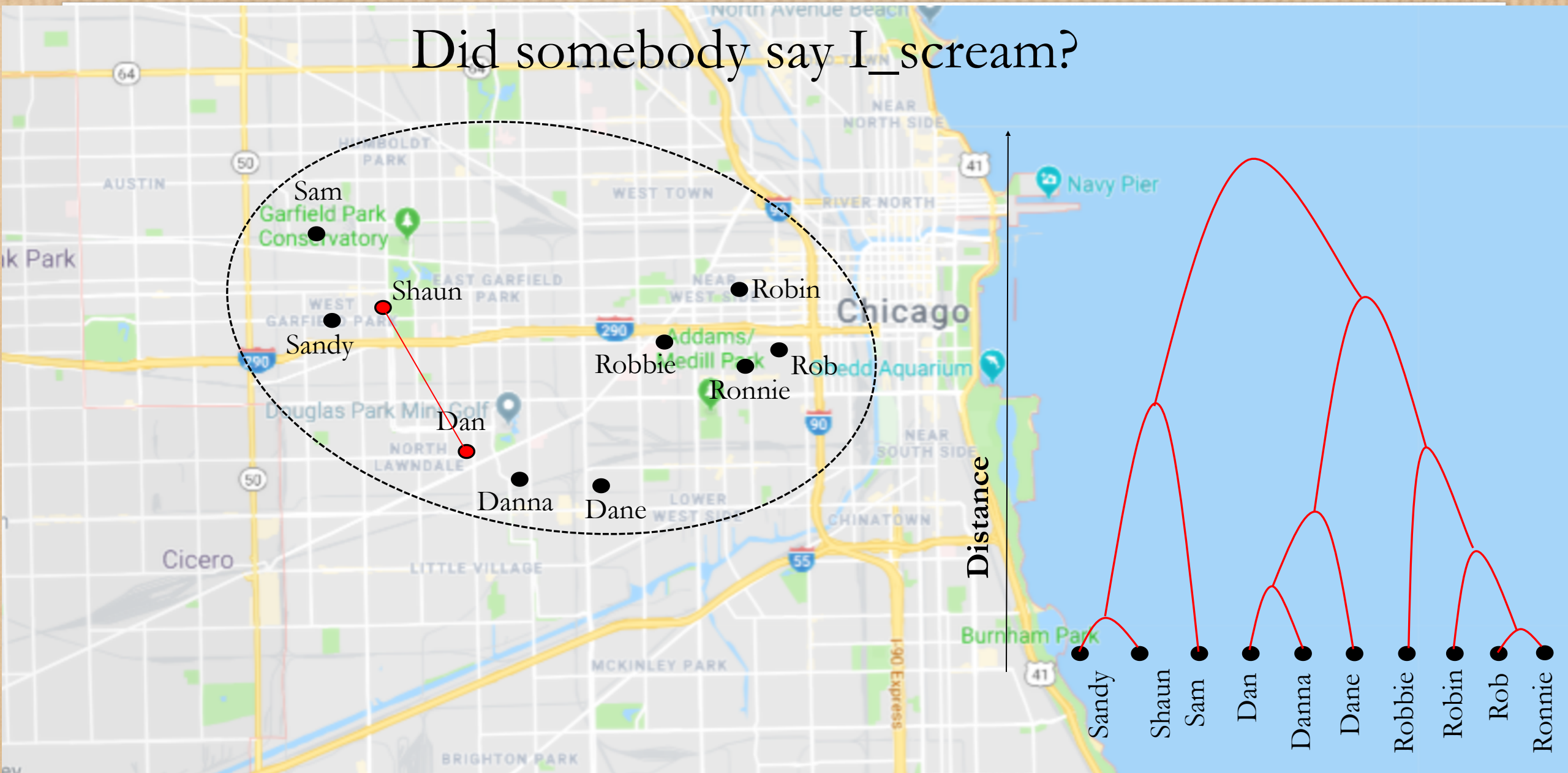




# Did somebody say I\_scream?



# Did somebody say I\_scream?



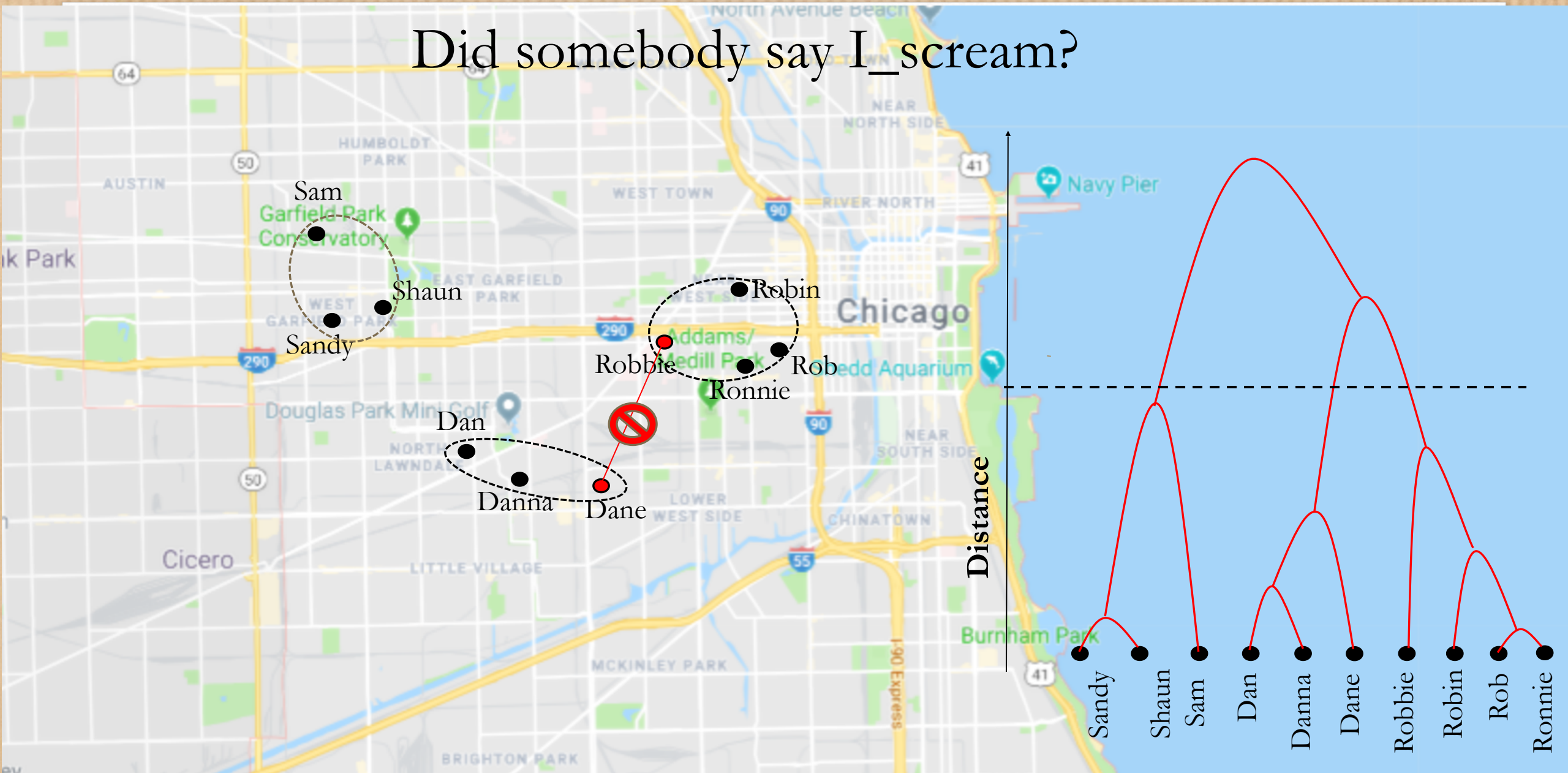


# Did somebody say I\_scream?

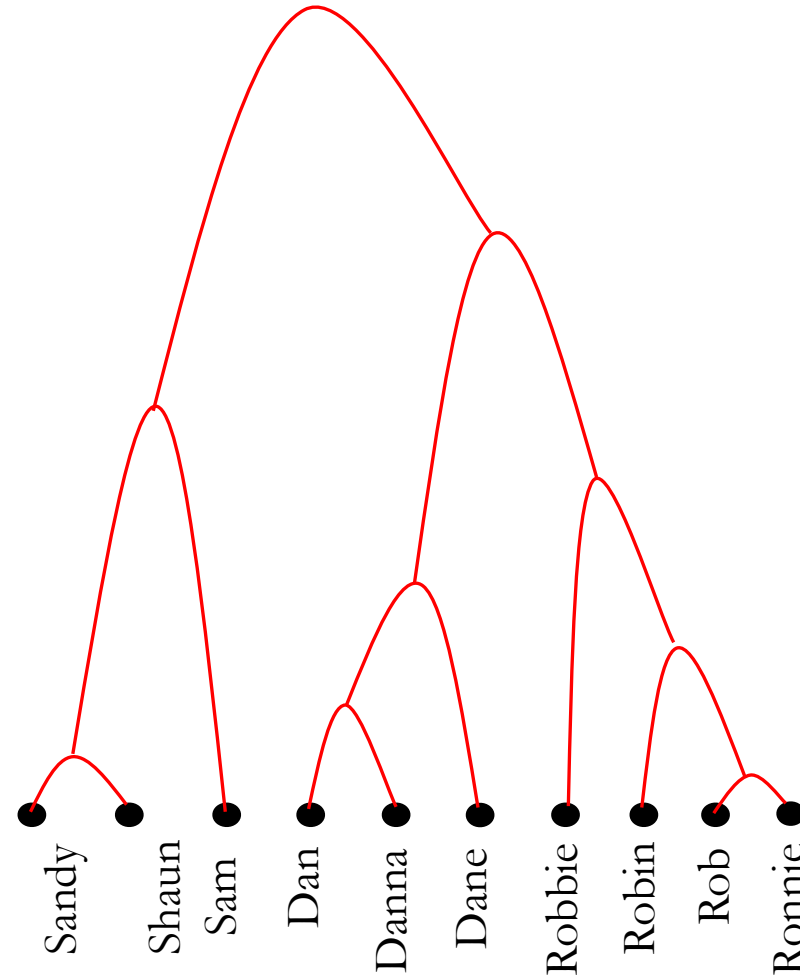




# Did somebody say I\_scream?



# Agglomerative



## Divisive

**use** of a **dendrogram** is to work out the best way to allocate objects to clusters.

# Pros & Cons

---

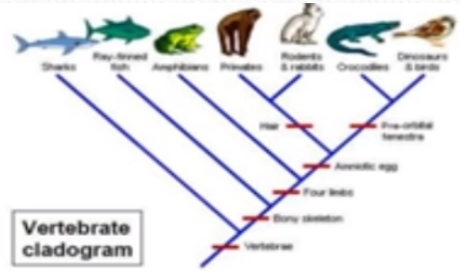
- Dendograms are great for visualization (provides hierarchical relations between clusters)
- Good in outlier detection
- Computationally intensive  $O(n)^3$  (Useful when there a small number of observations)



# Applications in real world



Genetics



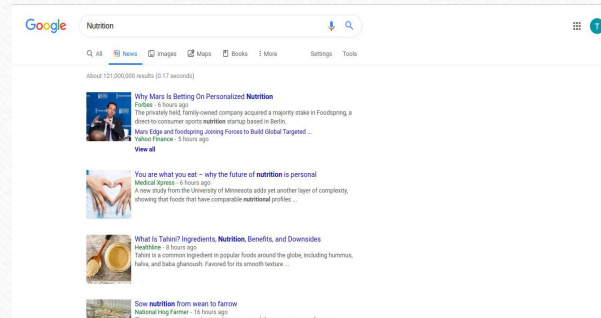
Evolutionary Biology



Social Network



Recommender System



Google News

# Python code implementation

---

- Jupyter Notebook





# Discussion

---





Thank you!!!!!!

---

