CLUSTERING

~Abhishek Kumar

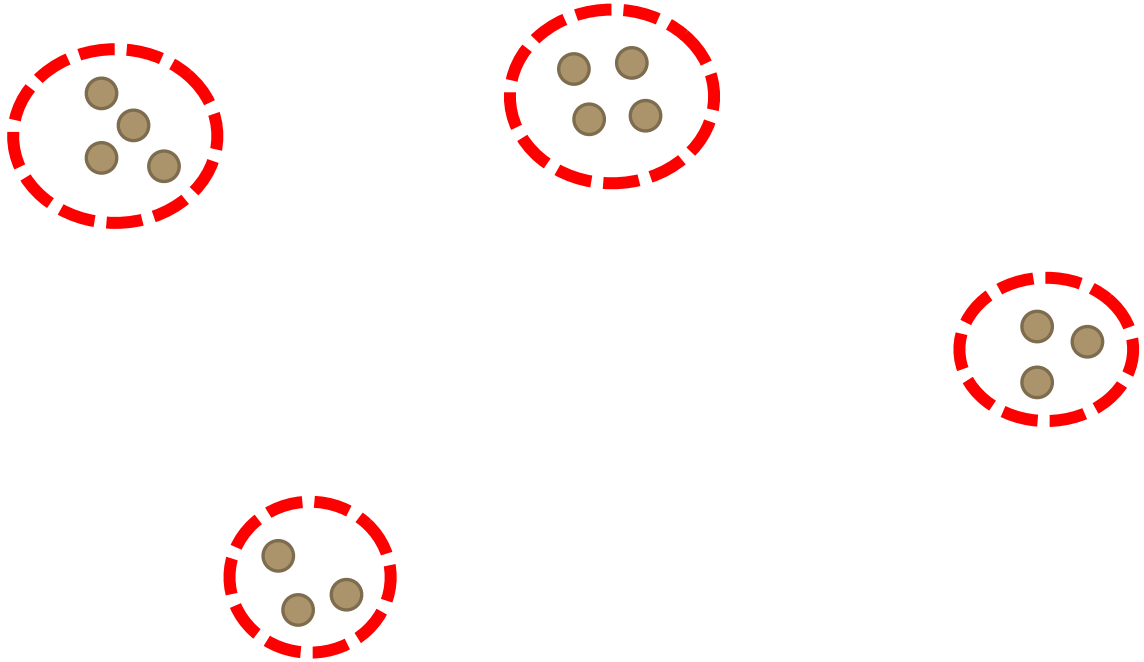
Scope

- 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

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

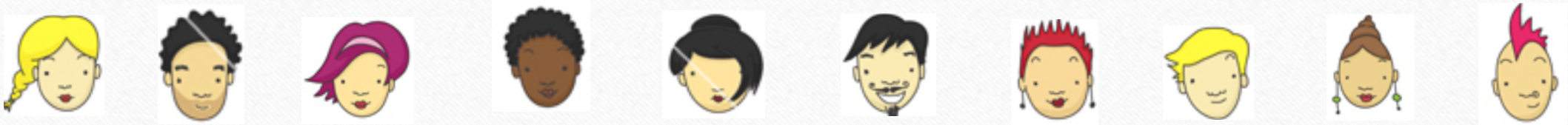
- Manhattan distance

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

- Minkowski distance

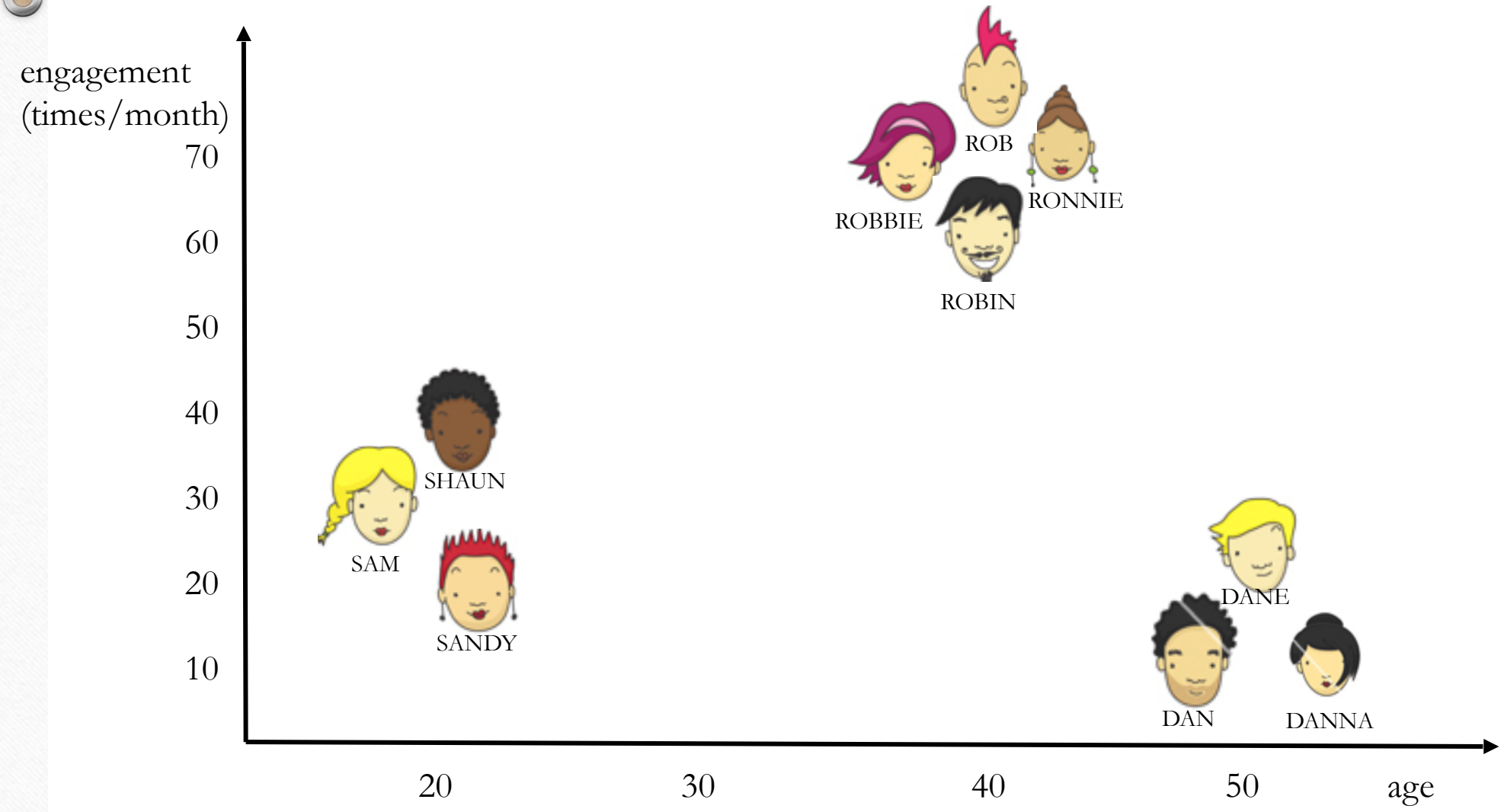
$$d(g_1, g_2) = \sqrt[m]{\sum_{i=1}^{n} (x_i - y_i)^m}$$

Customer Segmentation



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

Hint: Similar Names!!


engagement
(times/month)
70
60
50
40
30
20
10
ROB
ROBBIE
RONNIE
ROBIN
SHAUN
SAM
SANDY
DANE
DAN
DANNA
20
30
40
50
age

What Is A Good Clustering?

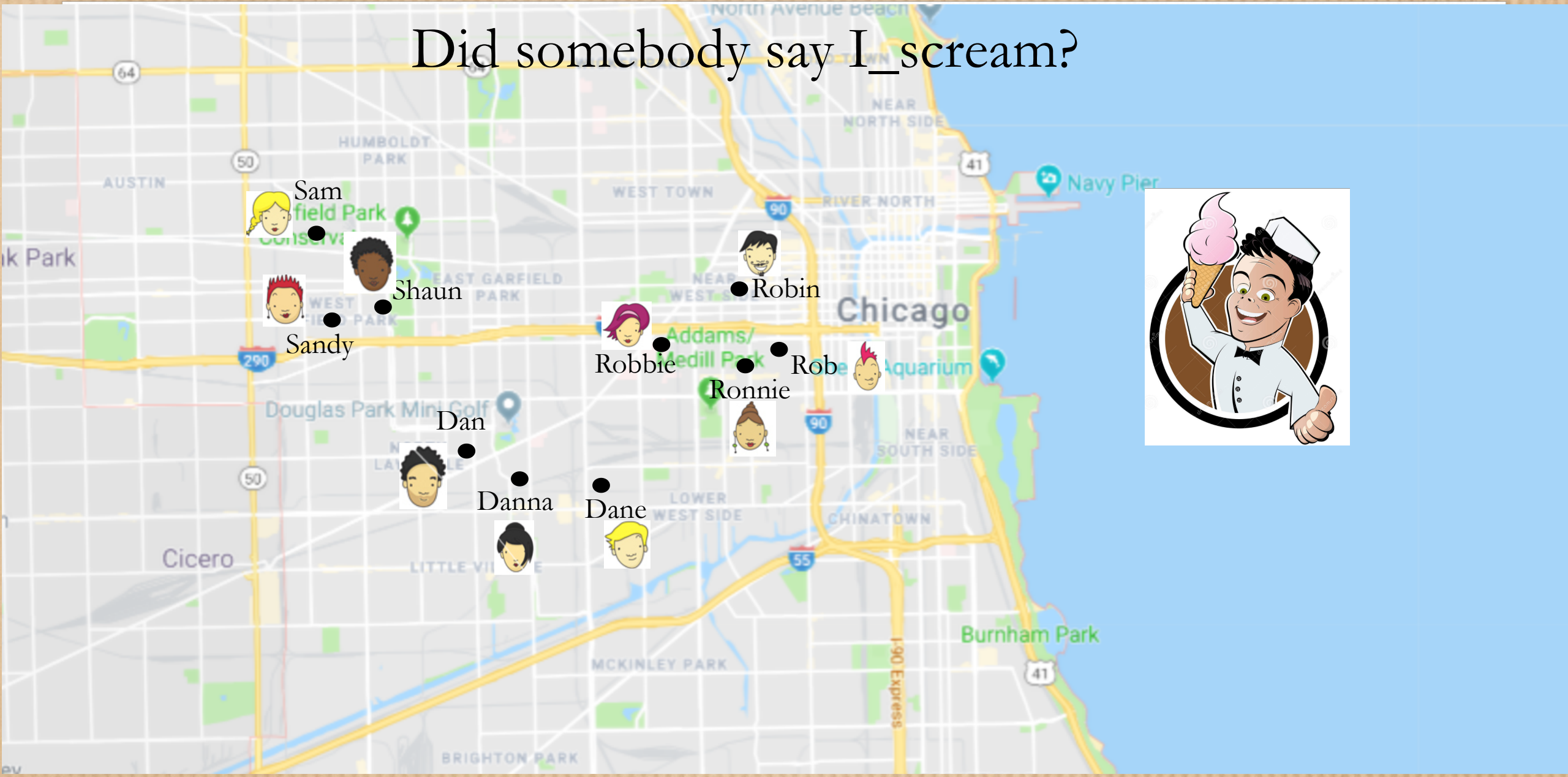
- Organizing data into classes such that there is
  - high intra-class similarity
  - low inter-class similarity

Types of Clustering?

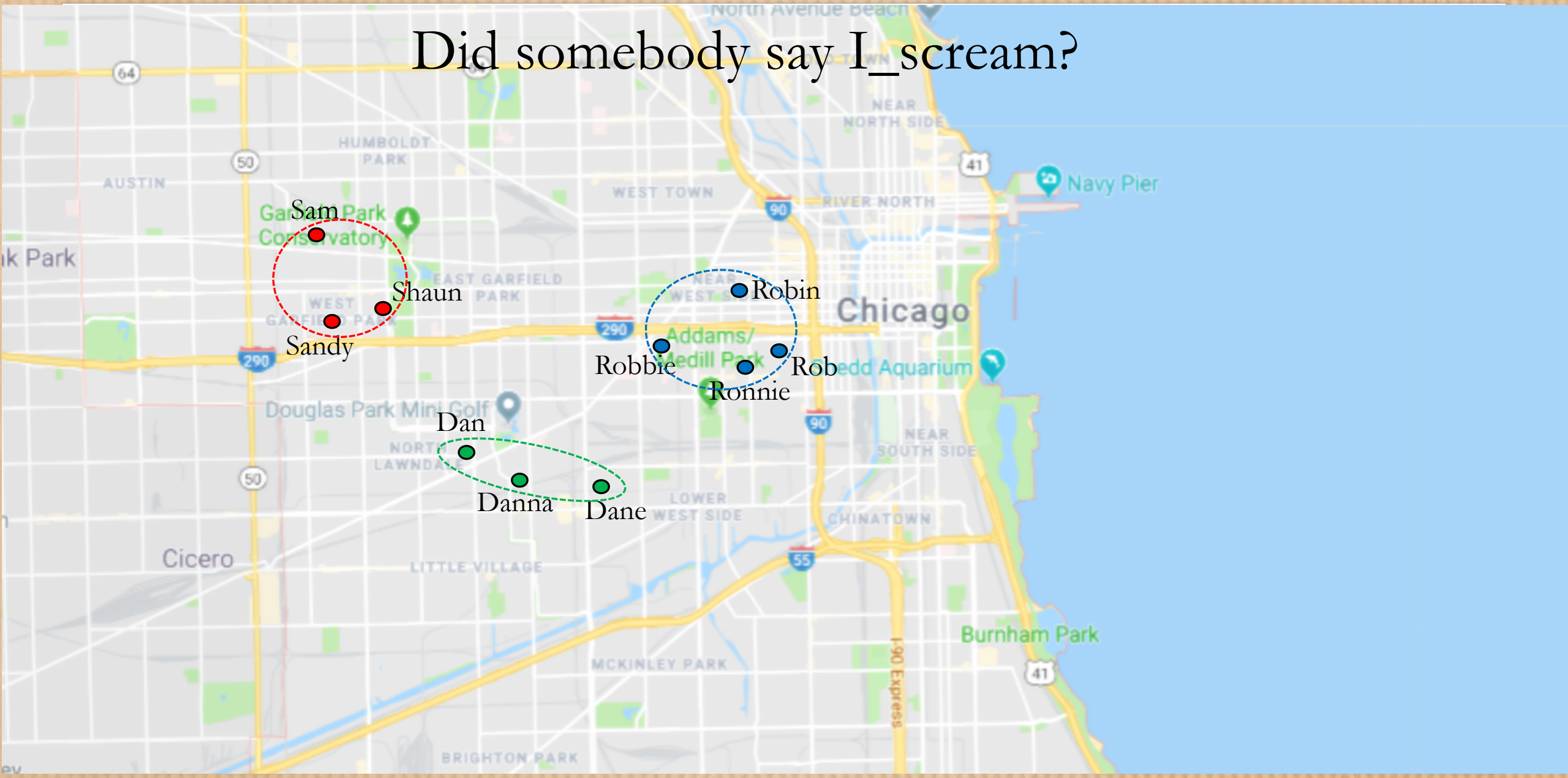
- K means clustering
- Hierarchical clustering

K-means Clustering

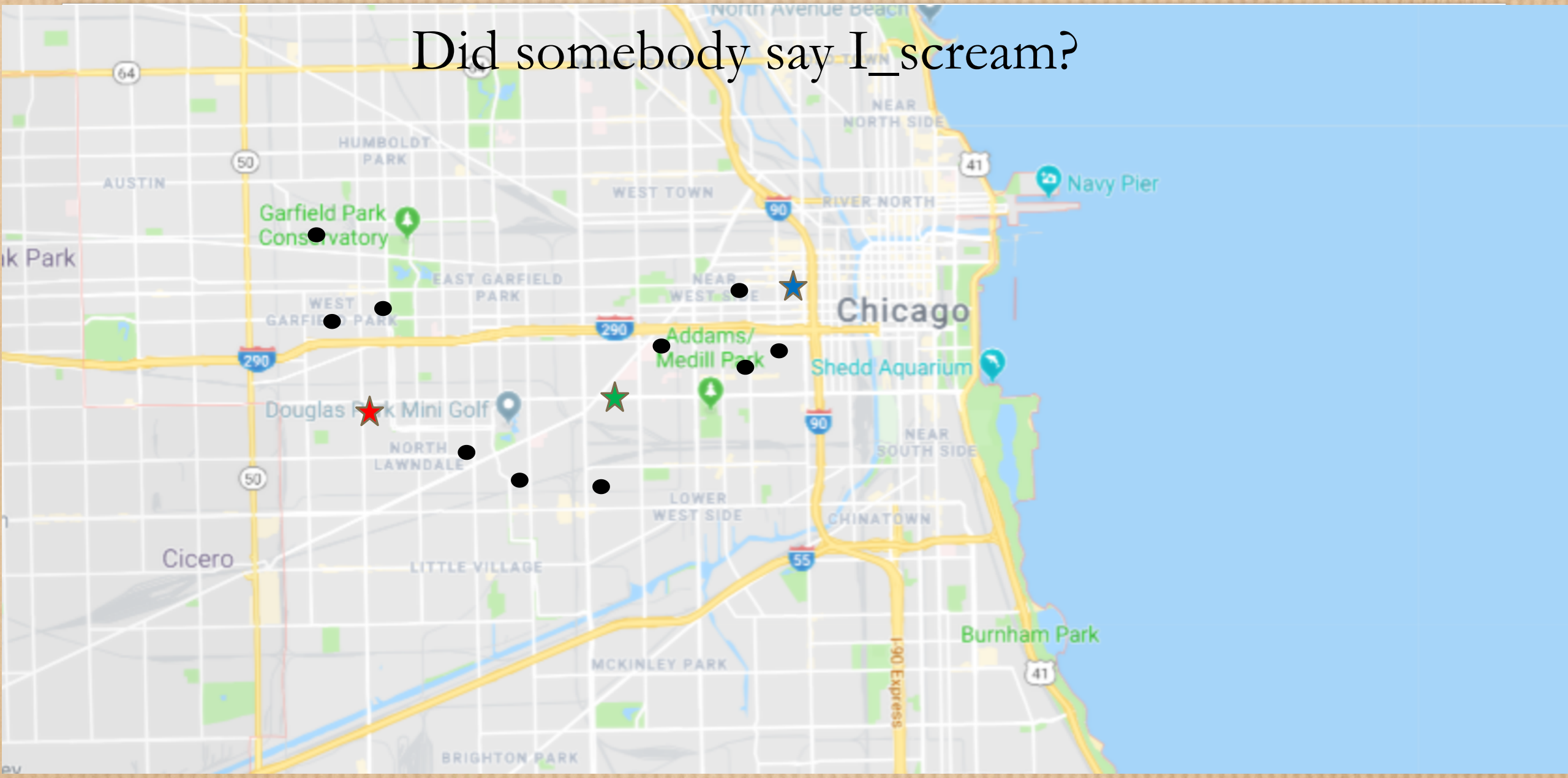
Did somebody say I_scream?



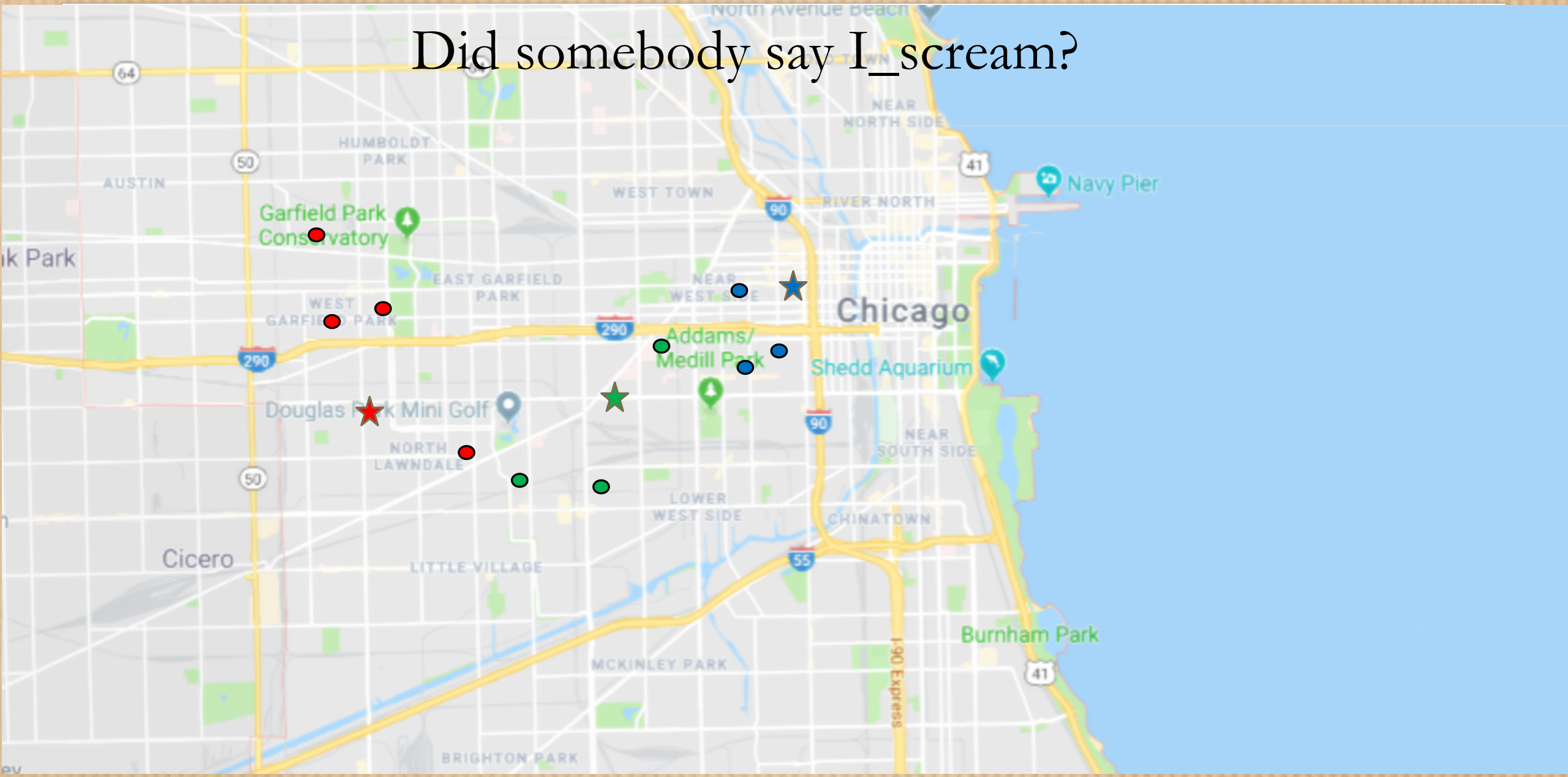
Did somebody say I_scream?



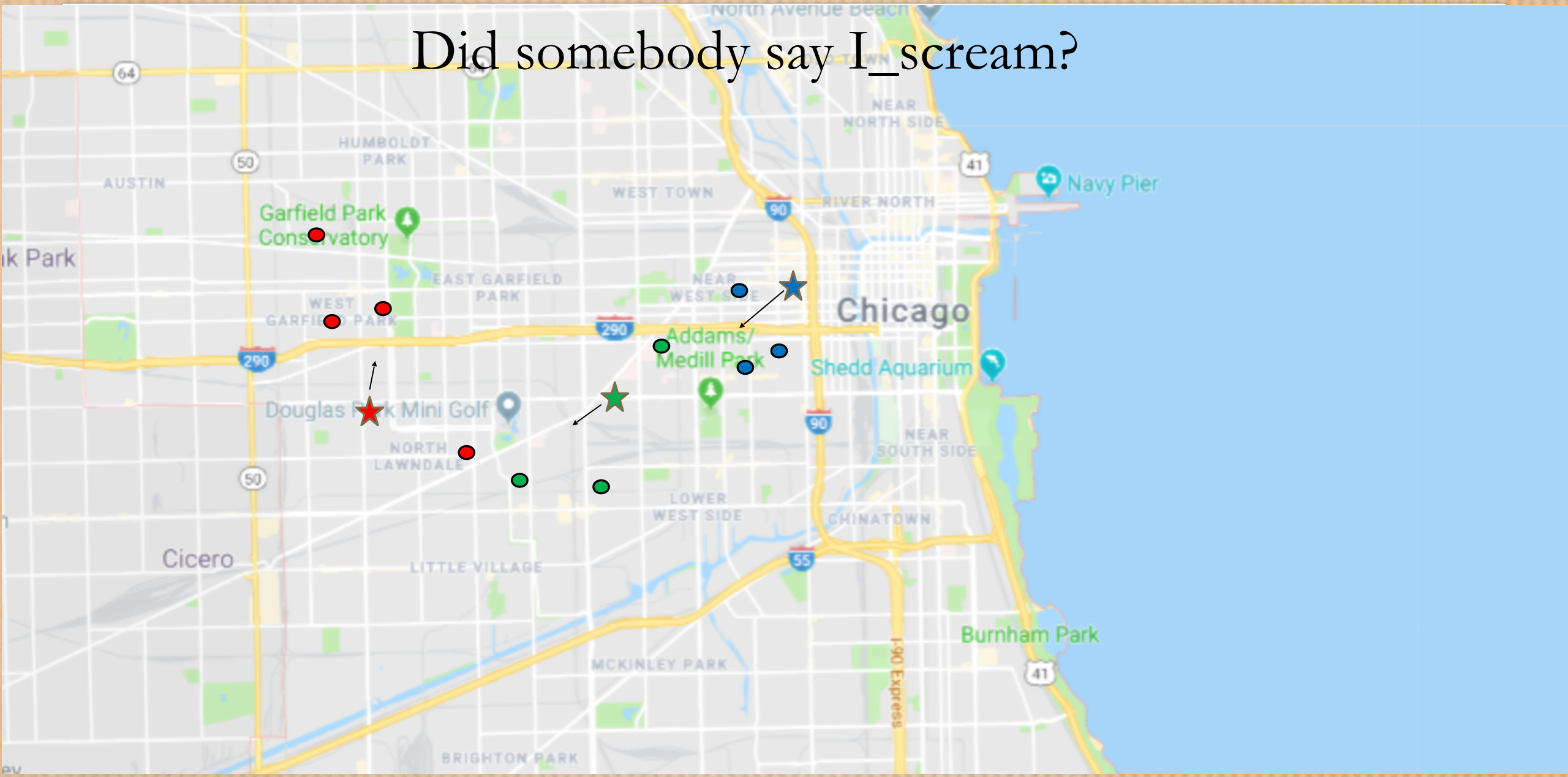
Did somebody say I_scream?



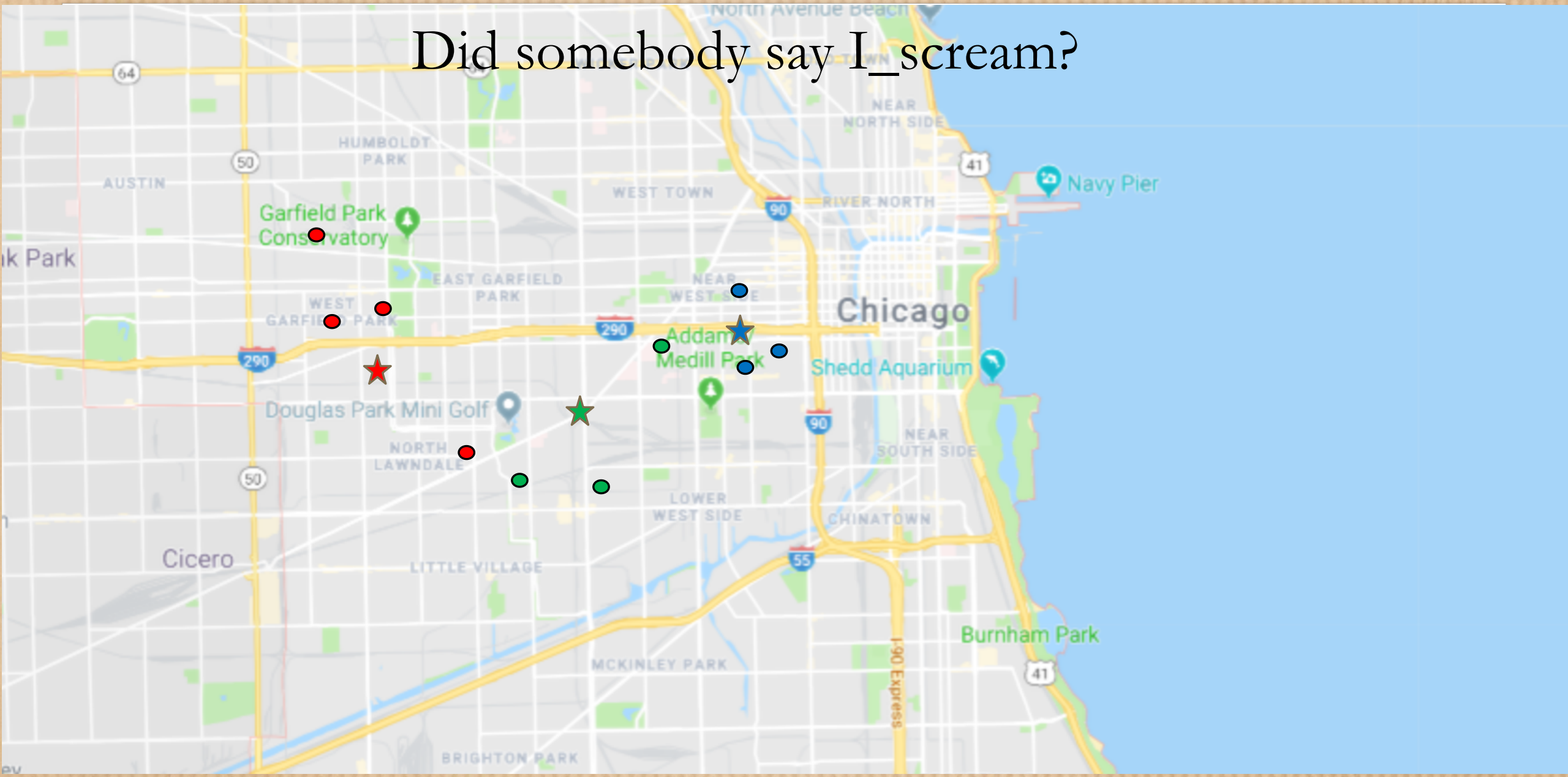
Did somebody say I_scream?



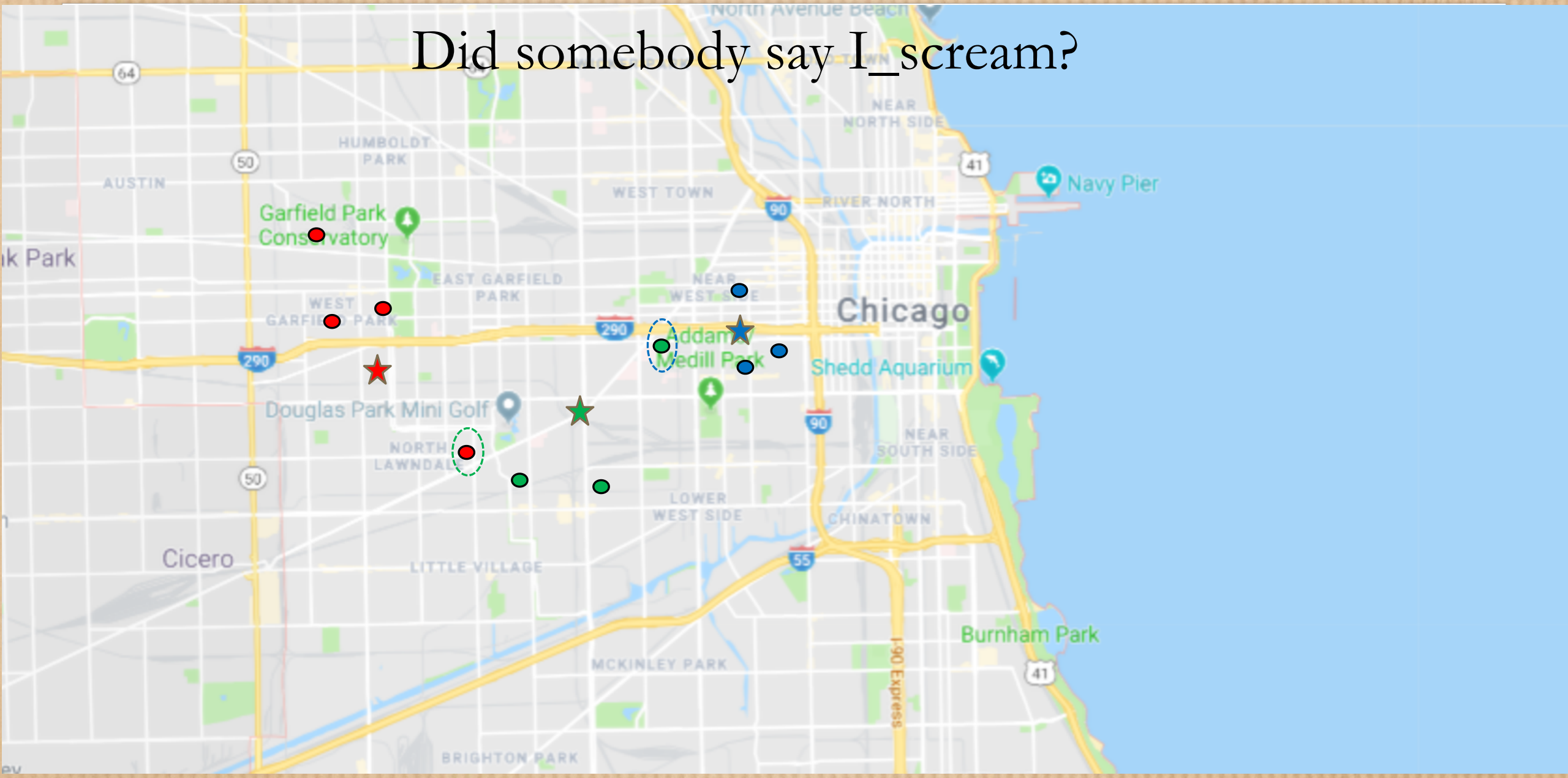
Did somebody say I_scream?



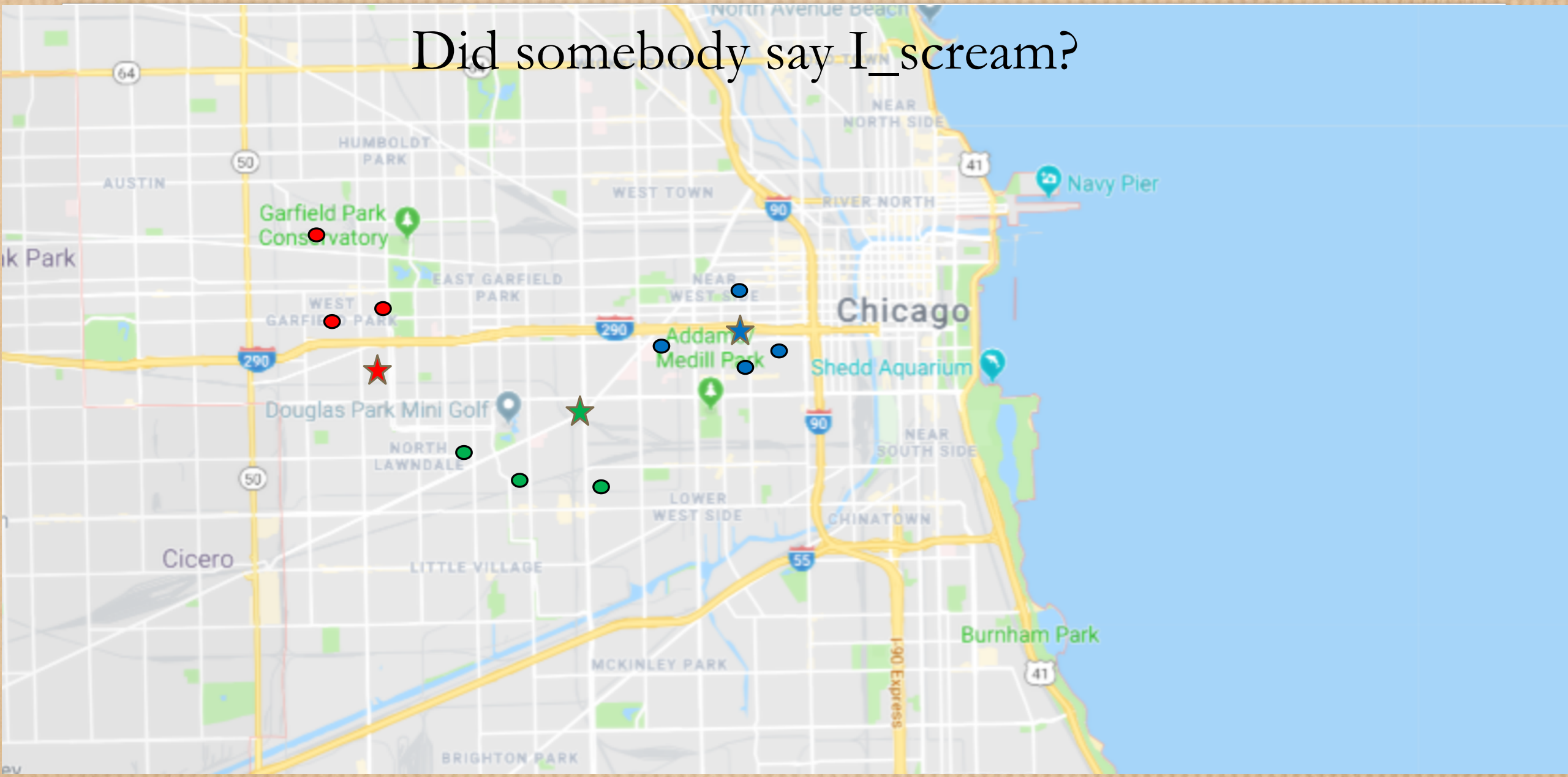
Did somebody say I_scream?



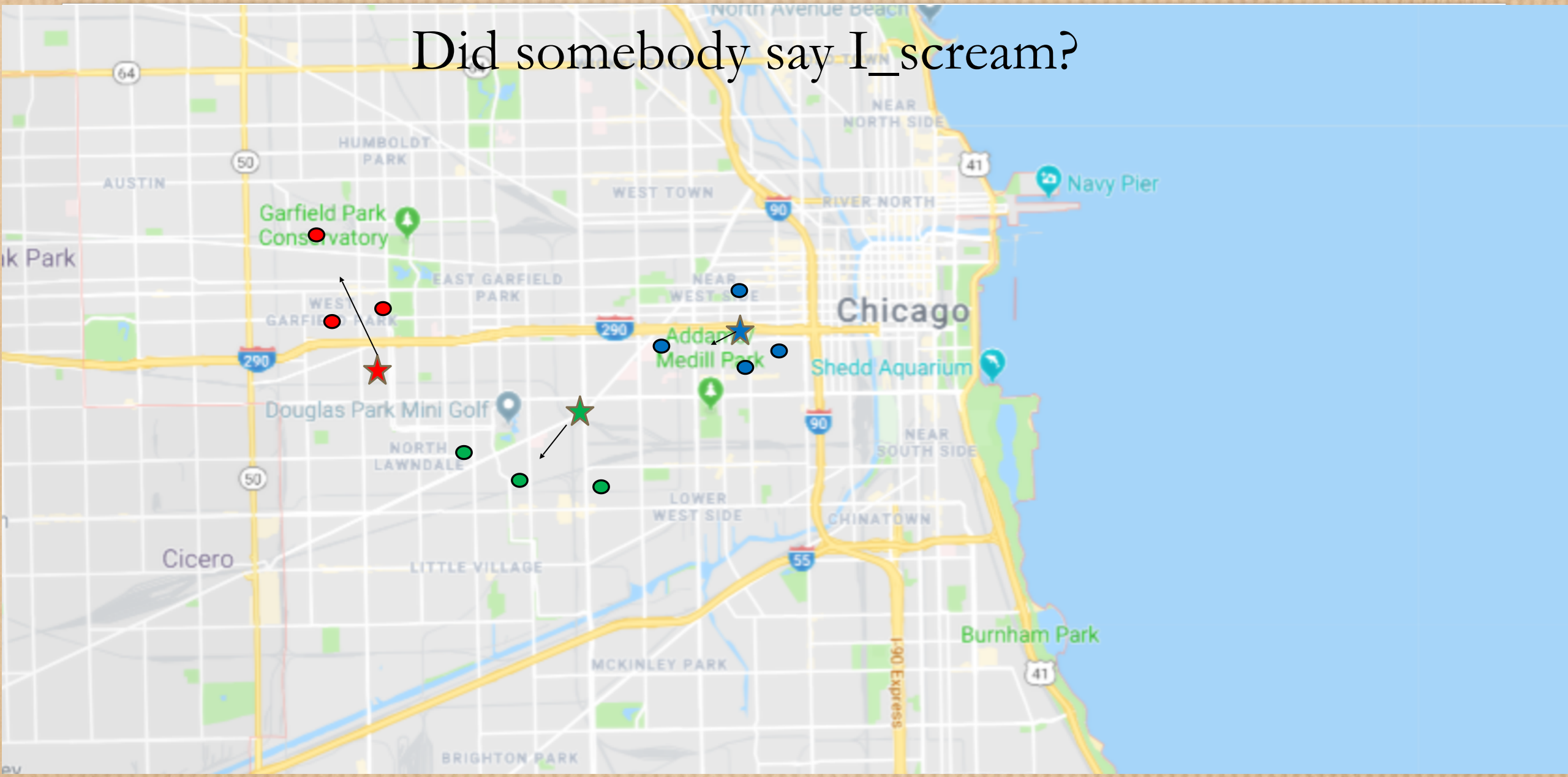
Did somebody say I_scream?



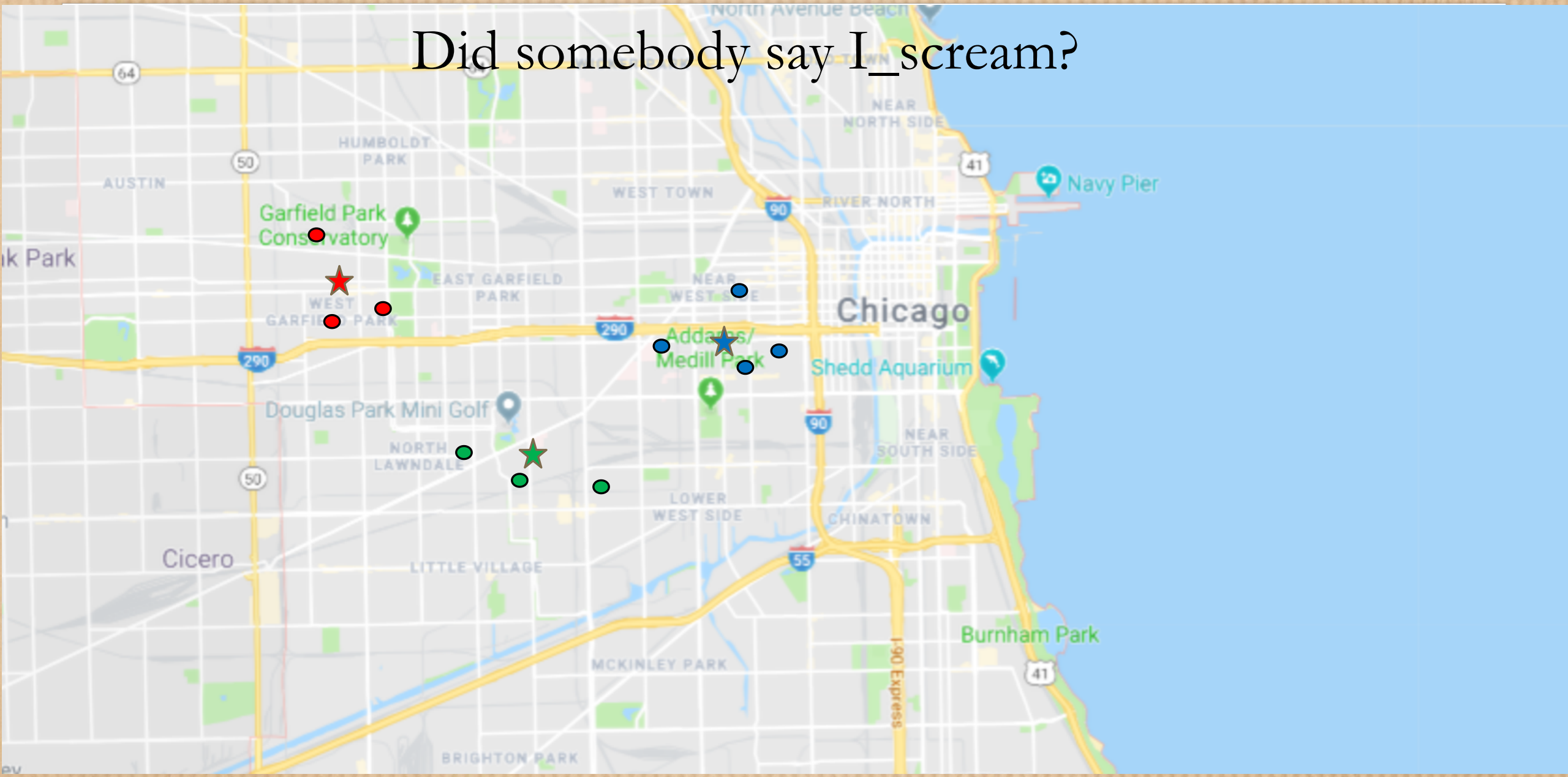
Did somebody say I_scream?



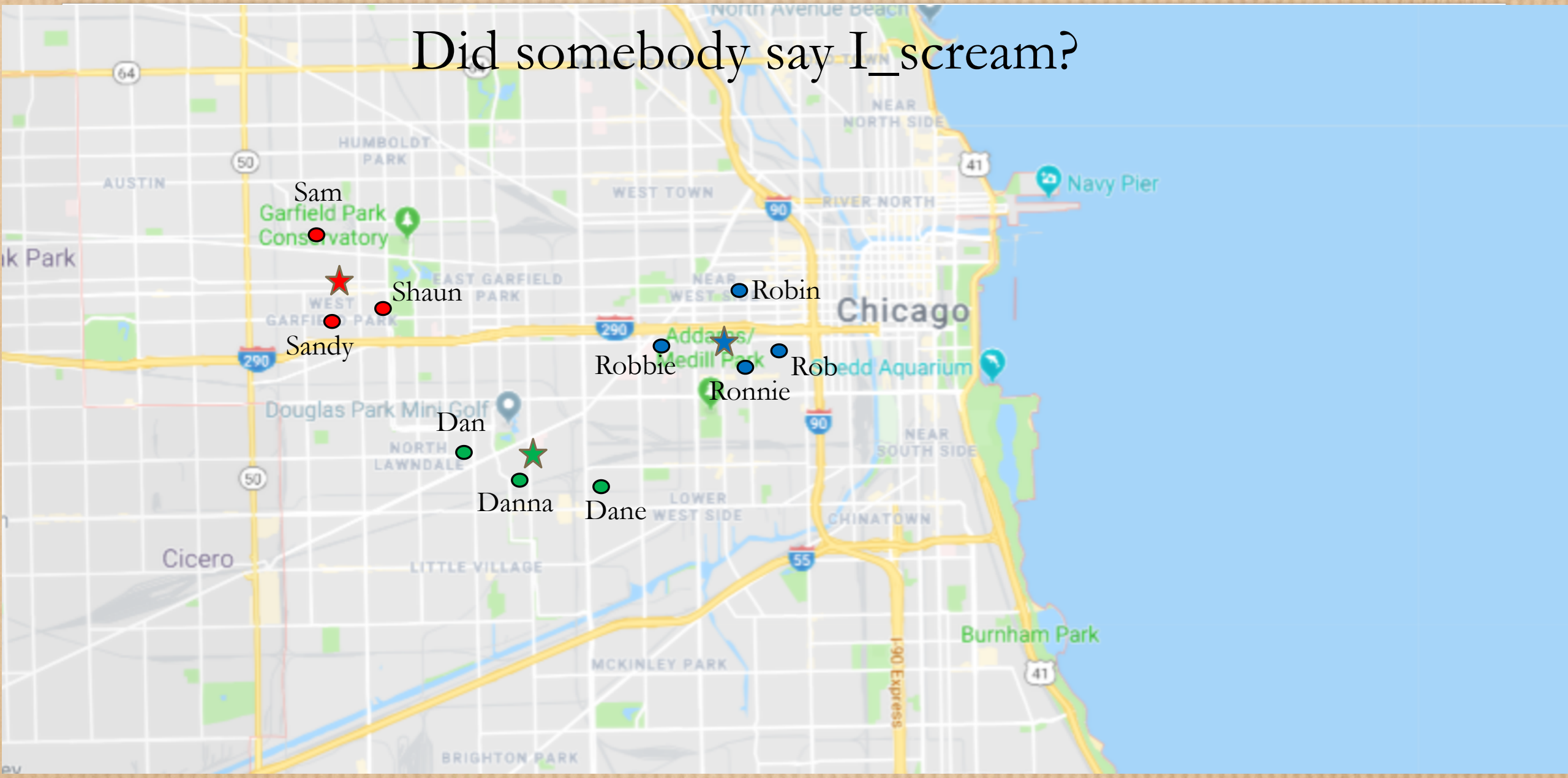
Did somebody say I_scream?



Did somebody say I_scream?



Did somebody say I_scream?

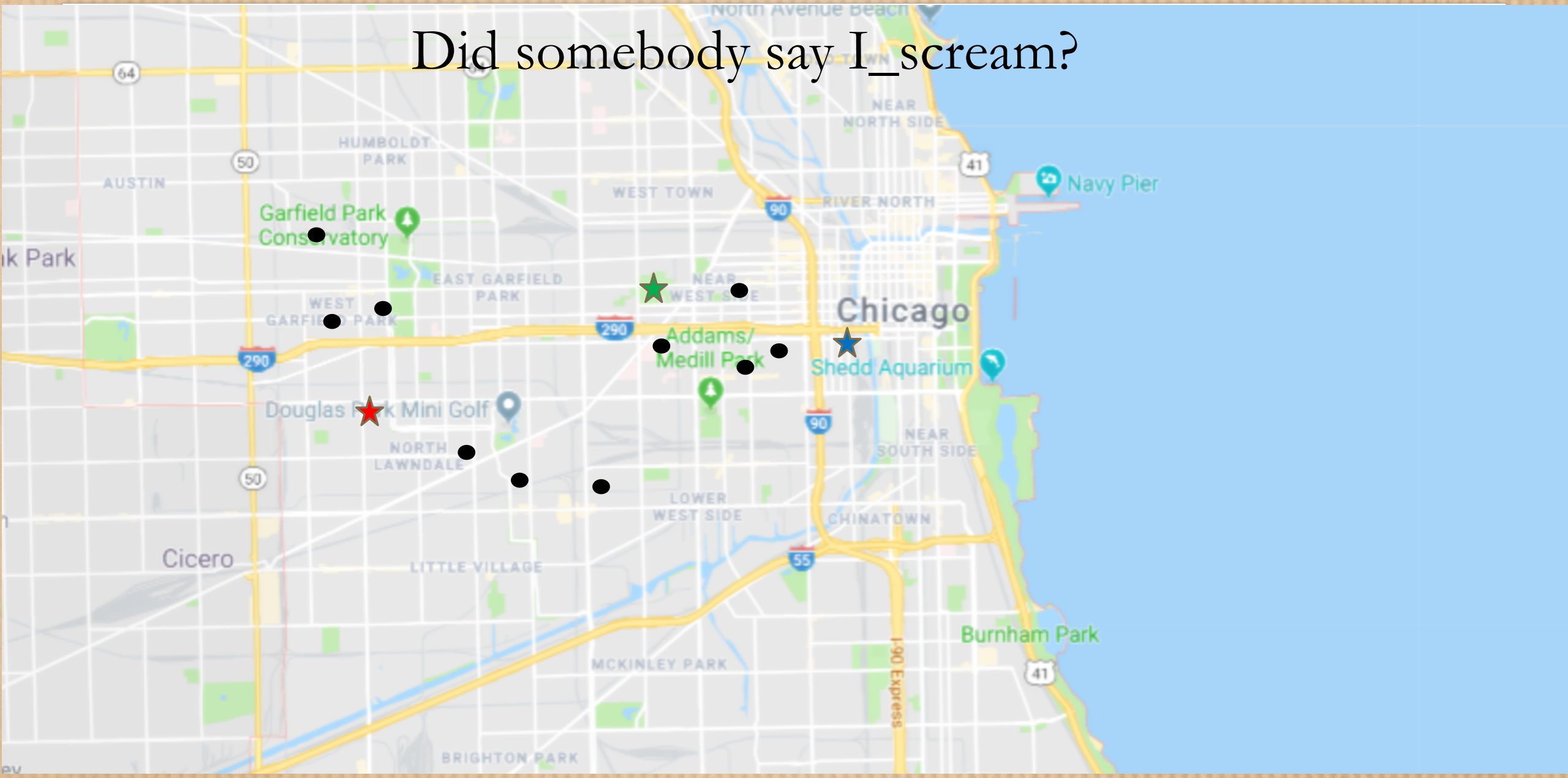


Pros & Cons

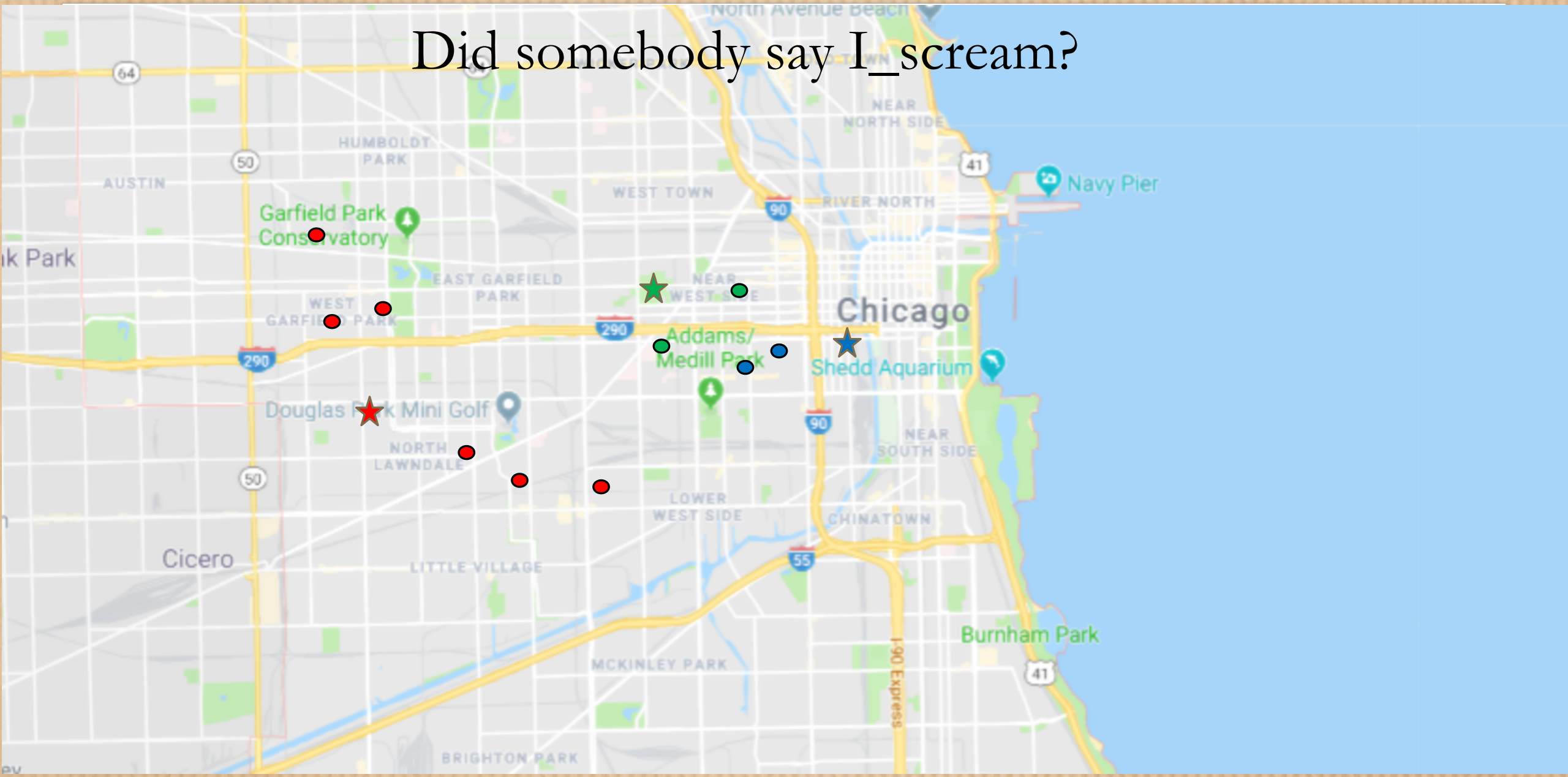
- Low complexity *complexity is O(nkt), where t = #iterations*

- Clusters are sensitive to initial assignment of centroids
- Necessity of specifying k
- Sensitive to noise and outlier data points

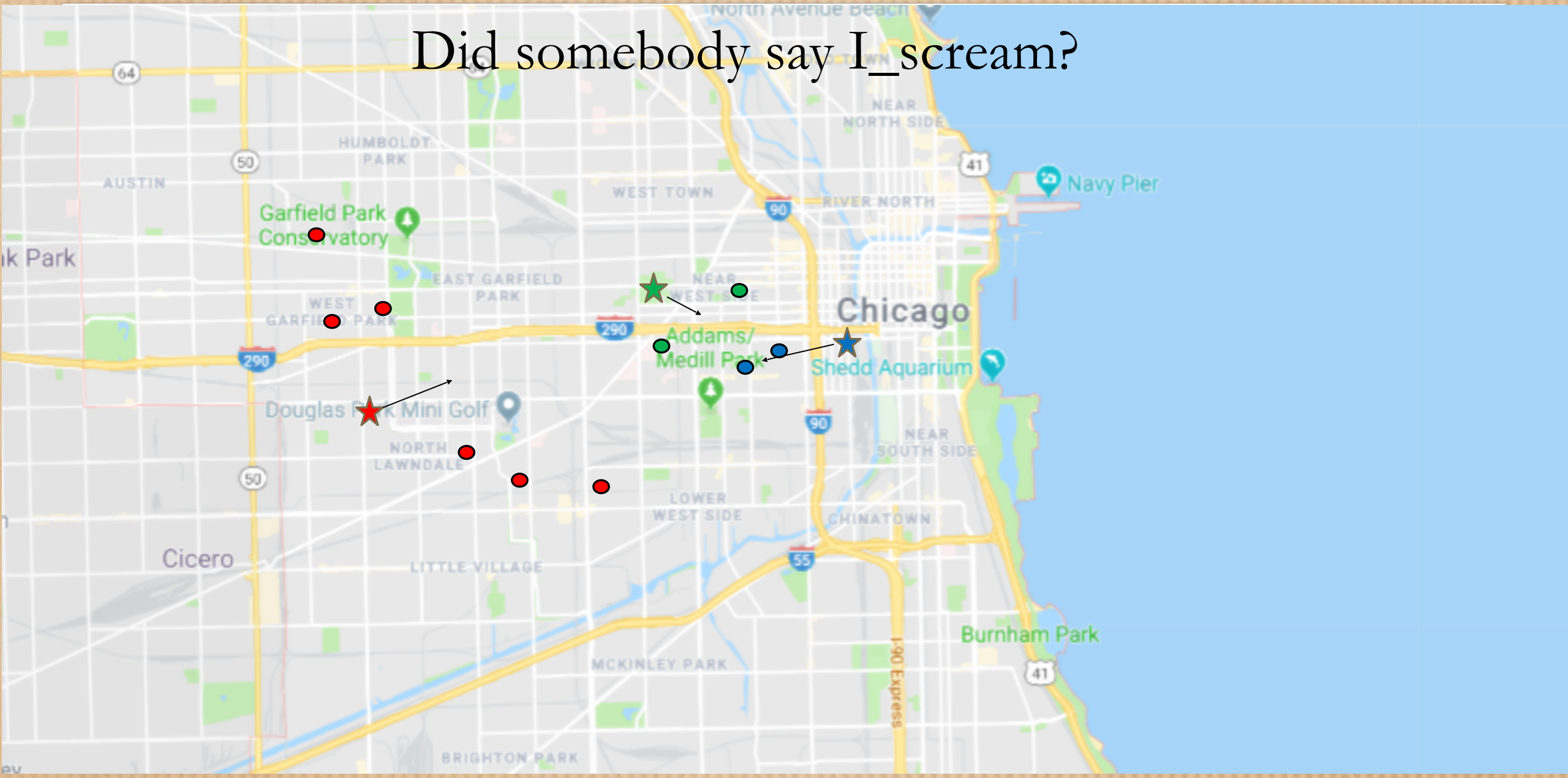
Did somebody say I_scream?



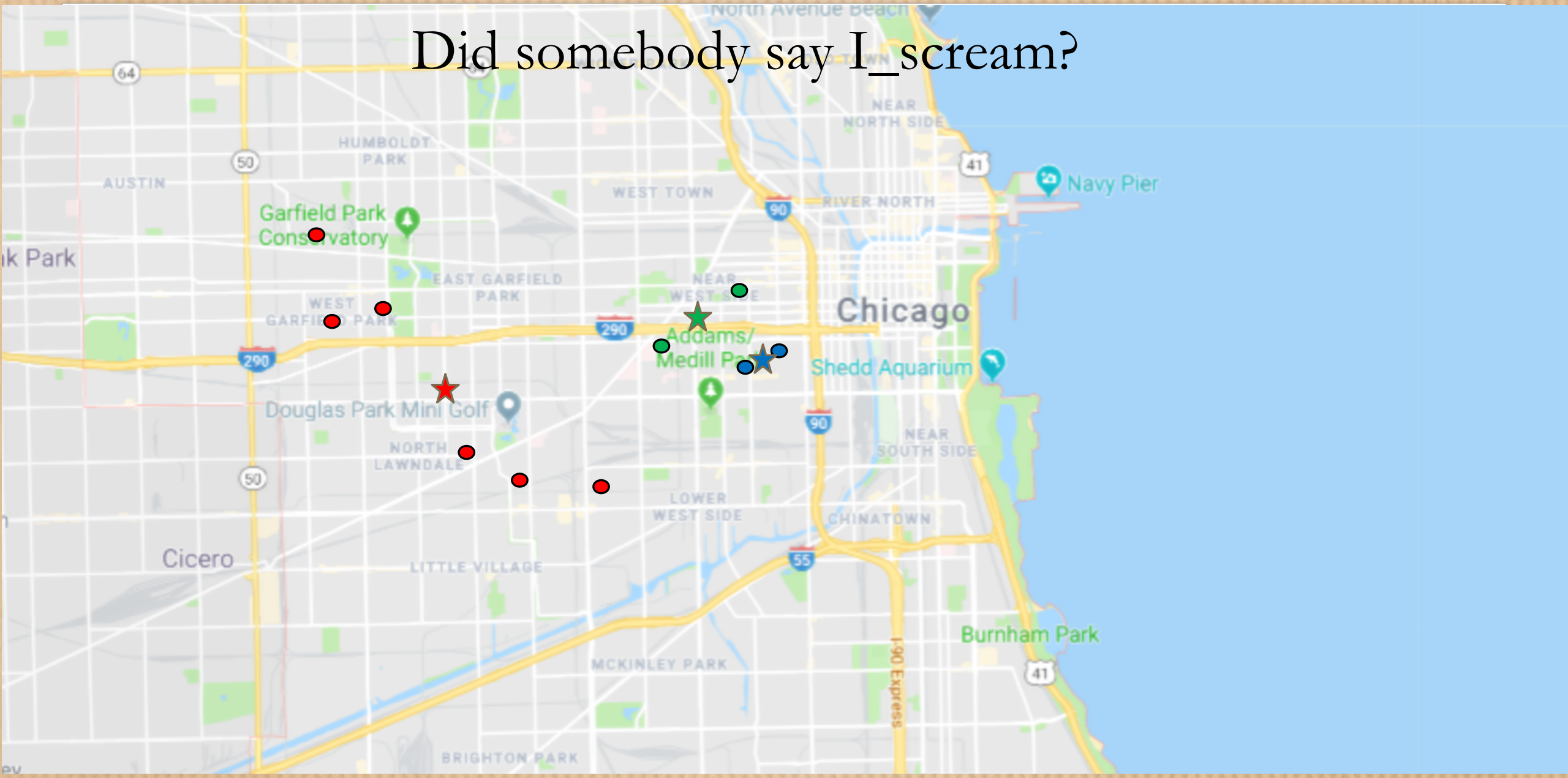
Did somebody say I_scream?



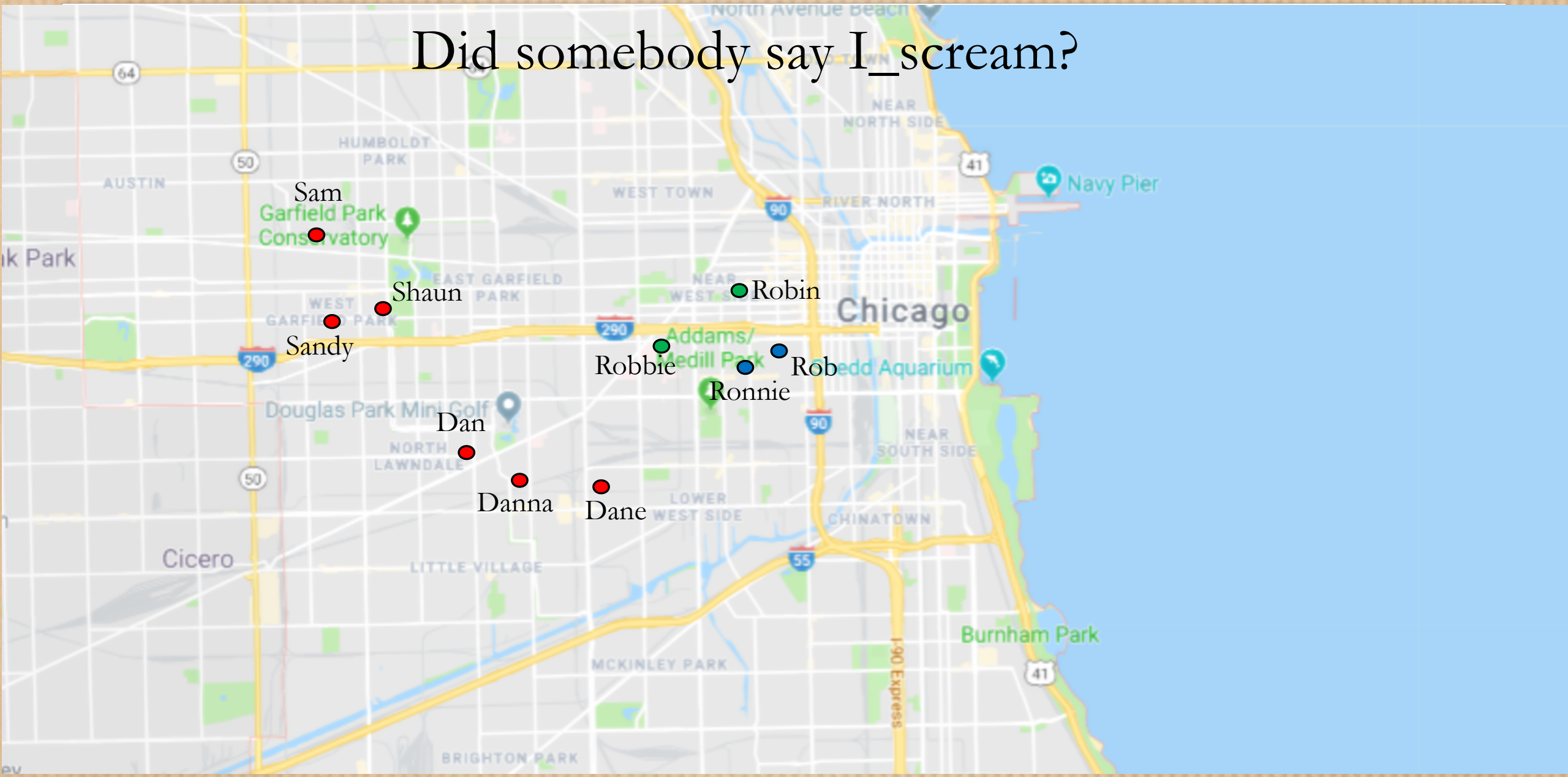
Did somebody say I_scream?



Did somebody say I_scream?

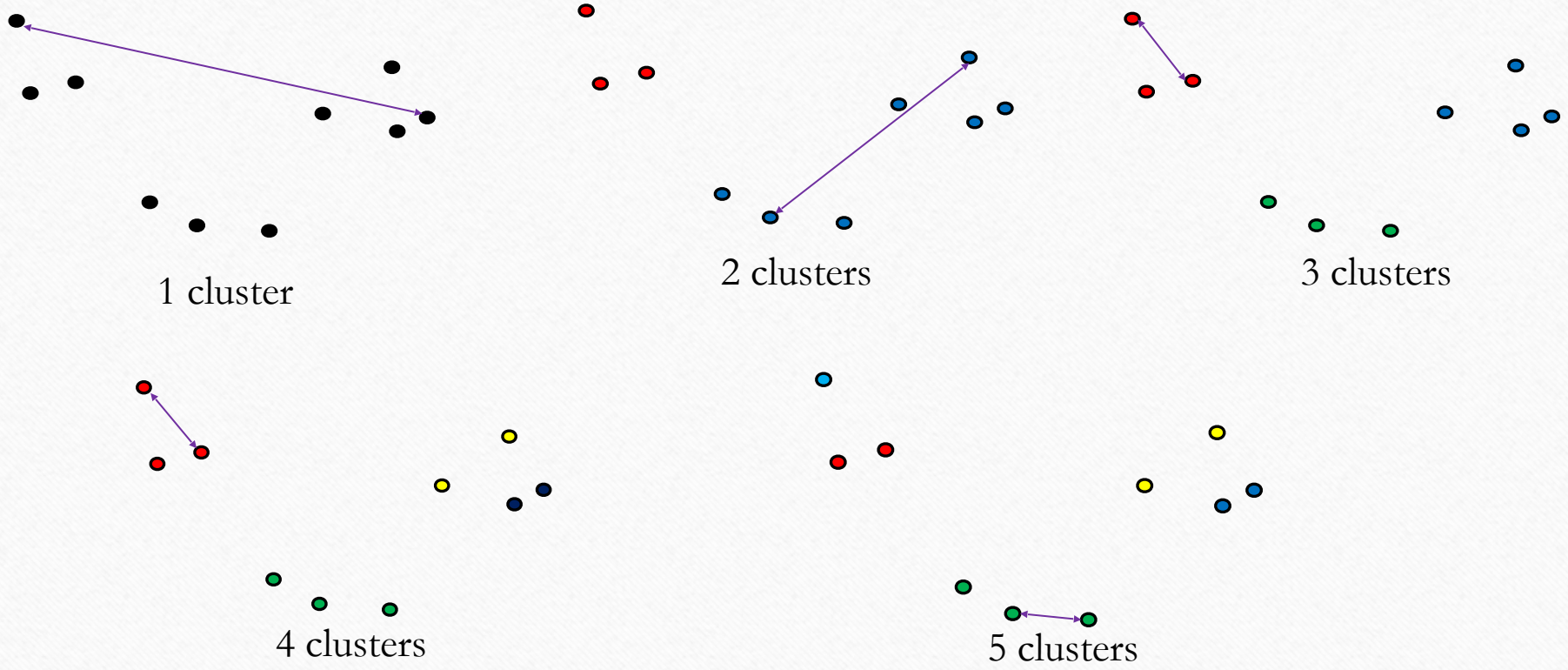


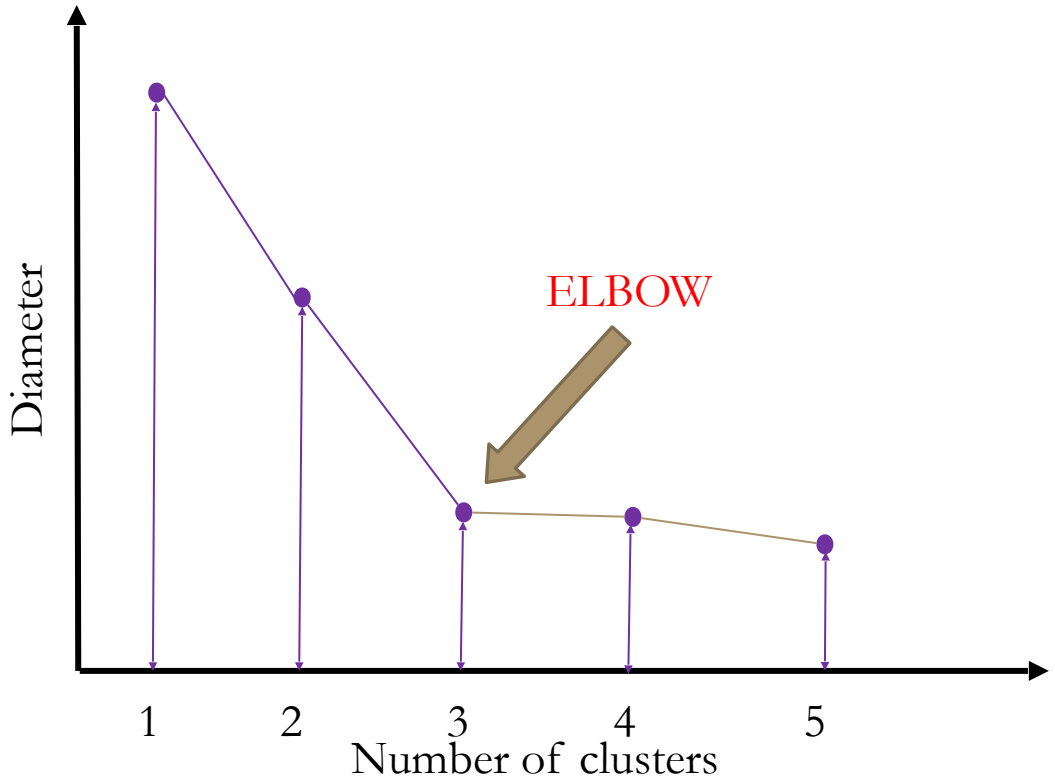
Did somebody say I_scream?



K means is a non-deterministic algorithm

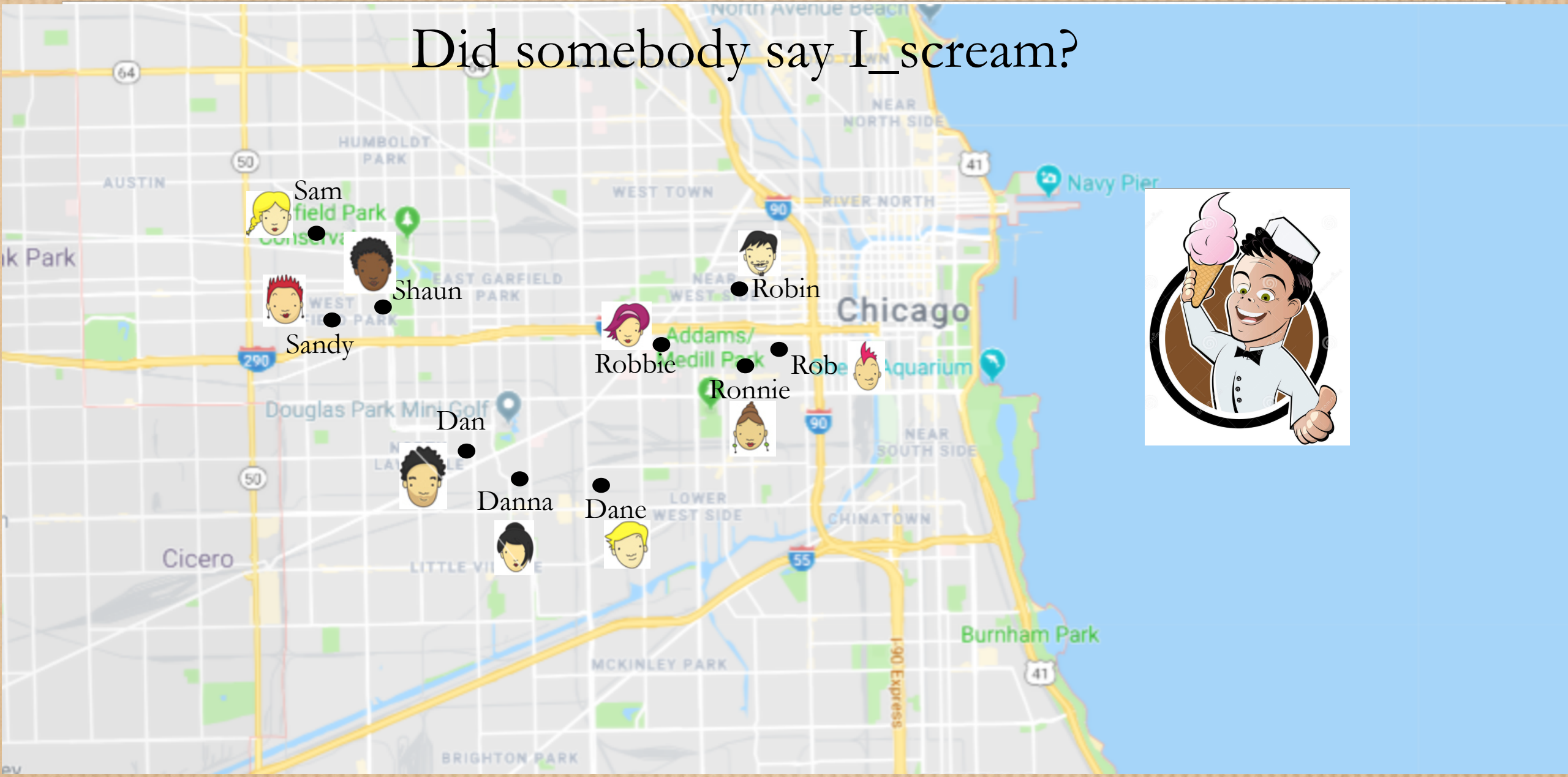
ELBOW method
(How Many Clusters?)



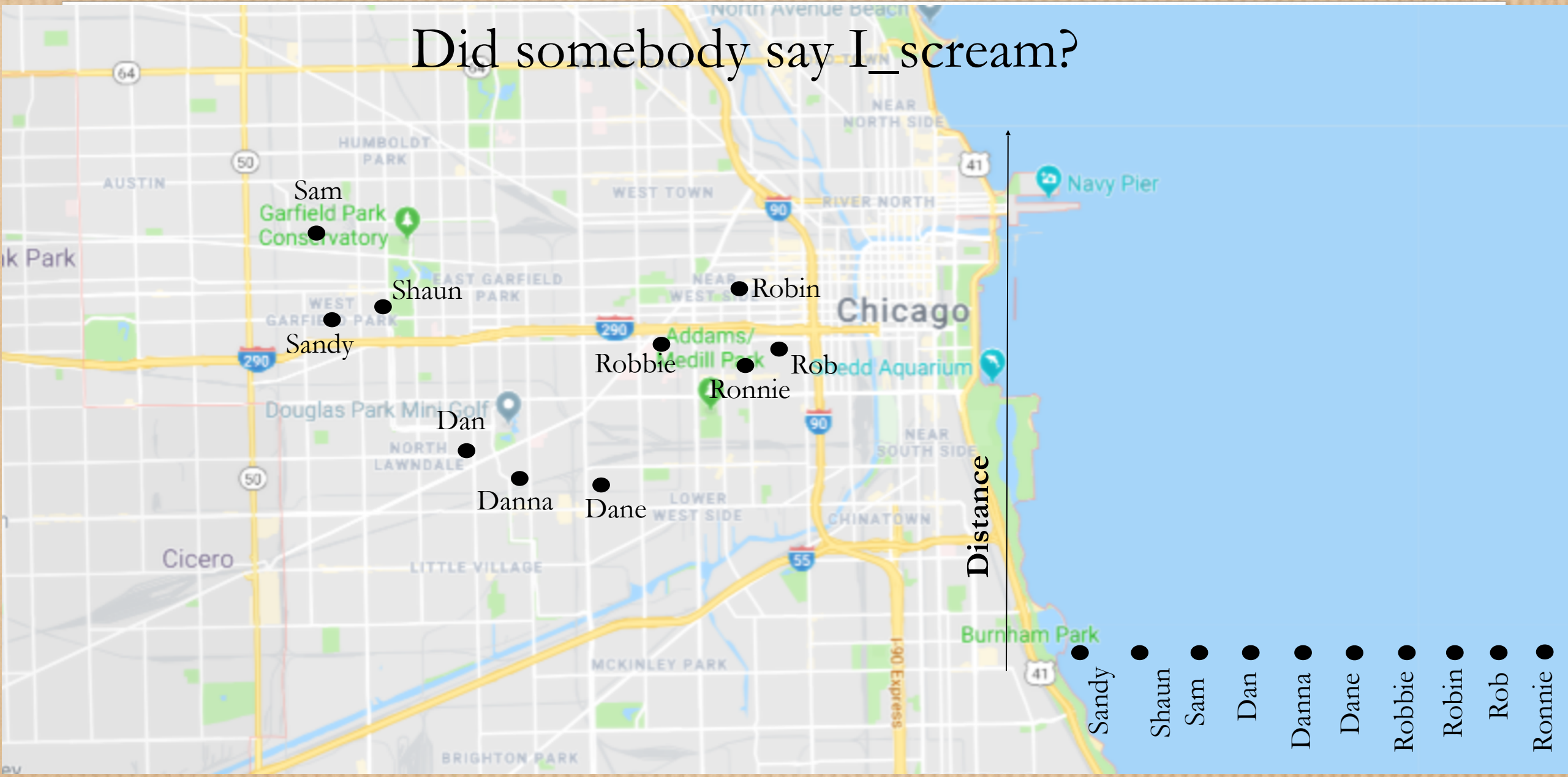

ELBOW
Diameter
1
2
3
4
5
Number of clusters

Hierarchical Clustering

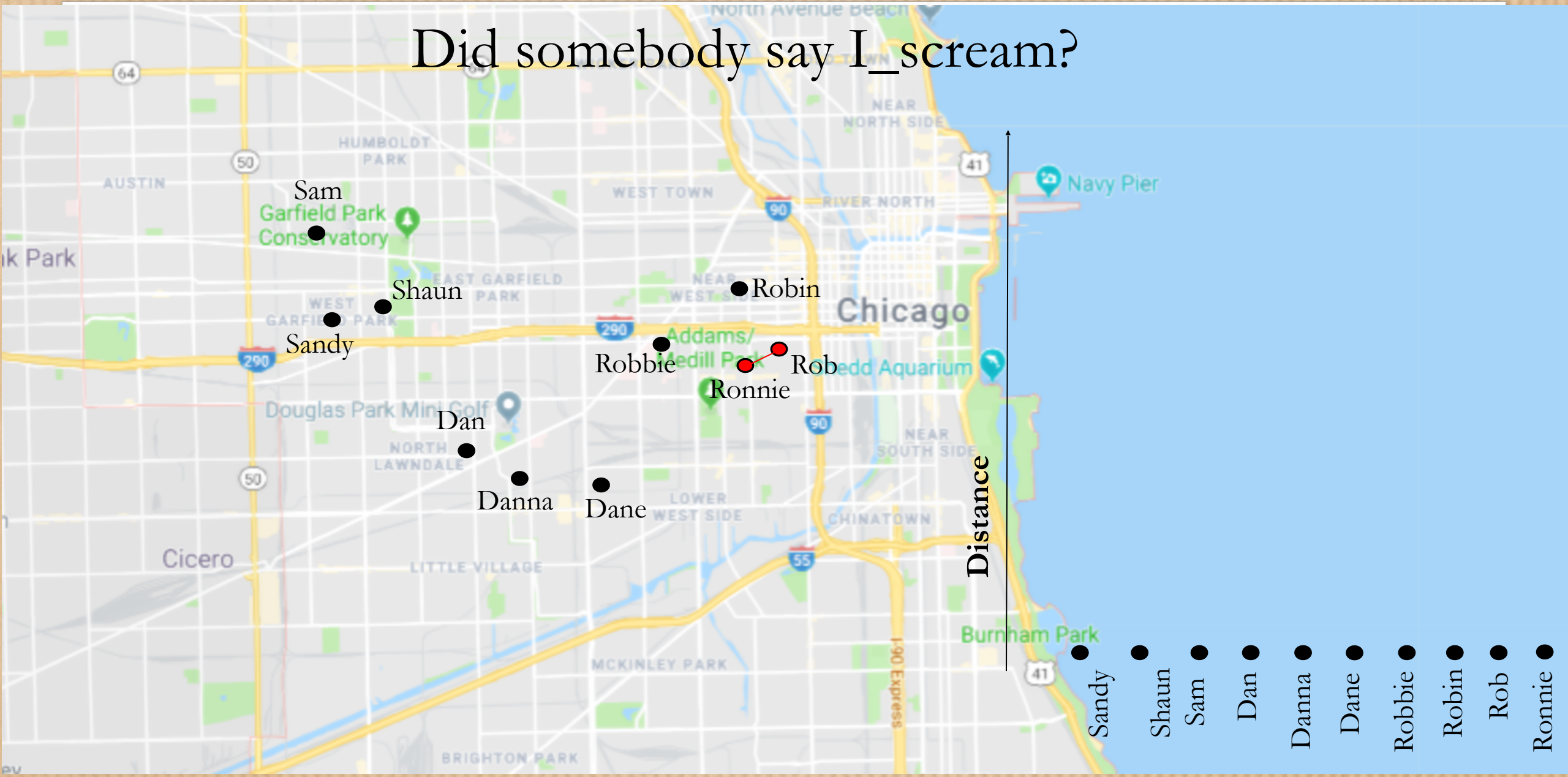
Did somebody say I_scream?



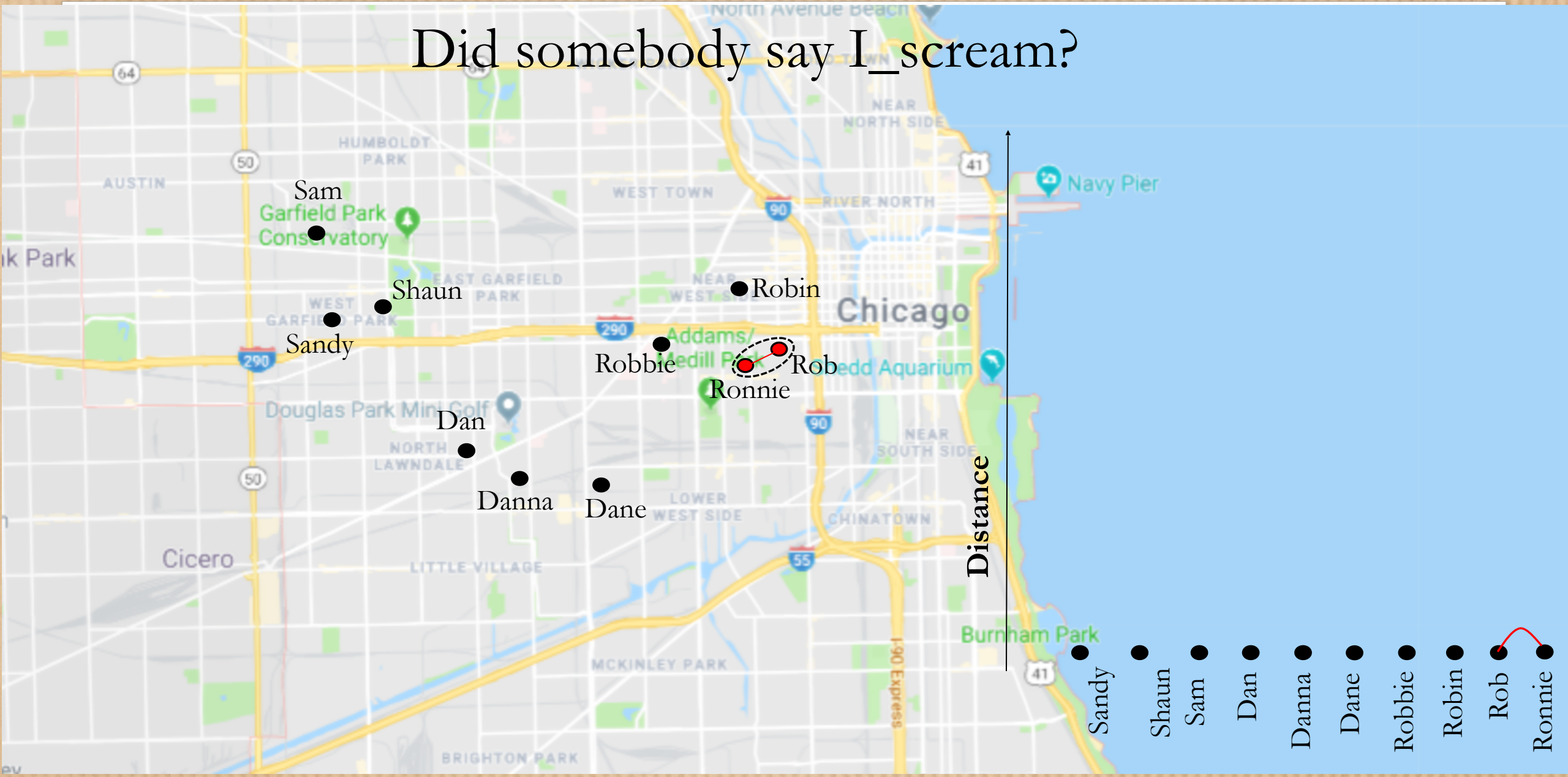
Did somebody say I_scream?



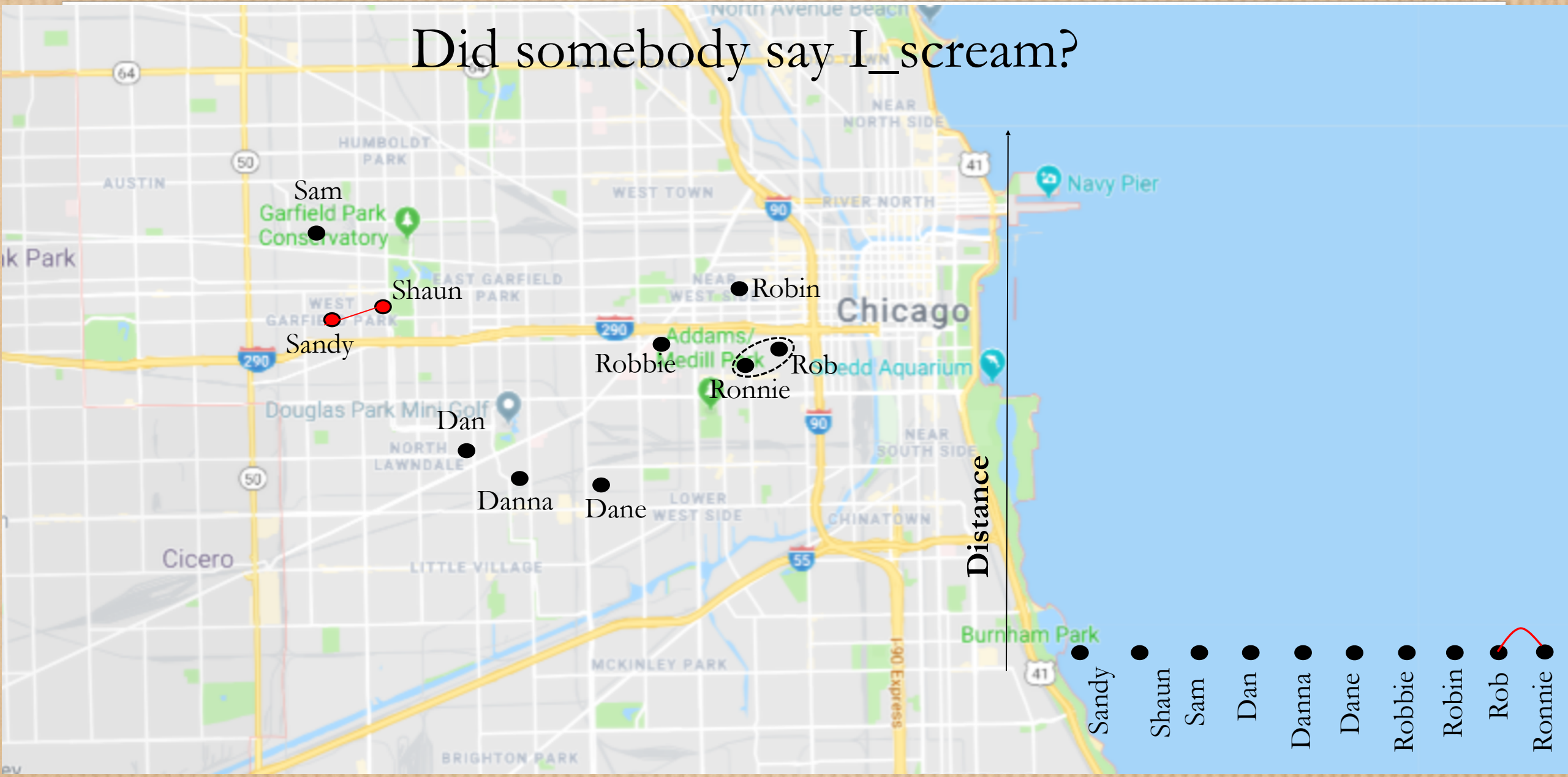
Did somebody say I_scream?



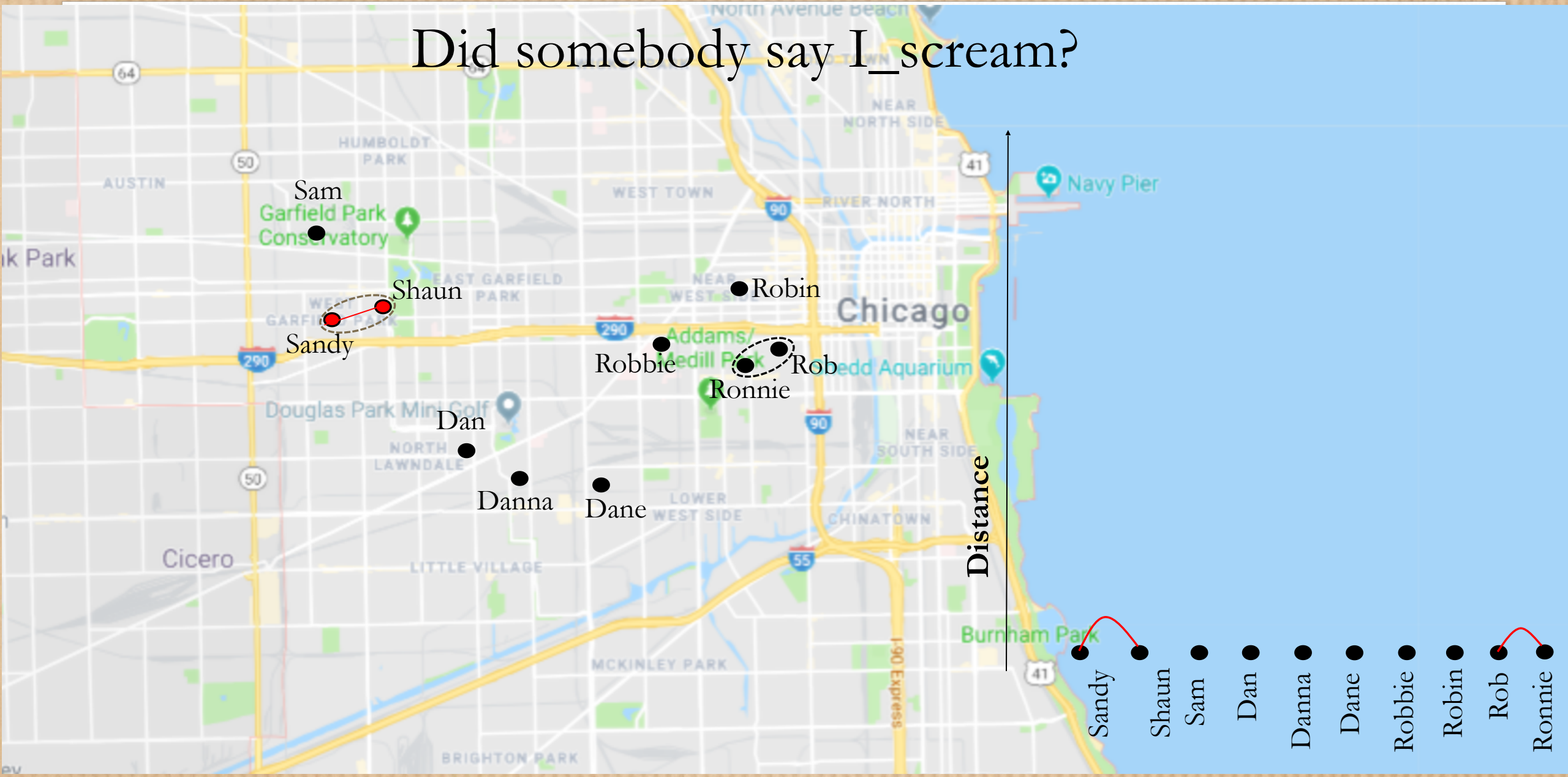
Did somebody say I_scream?



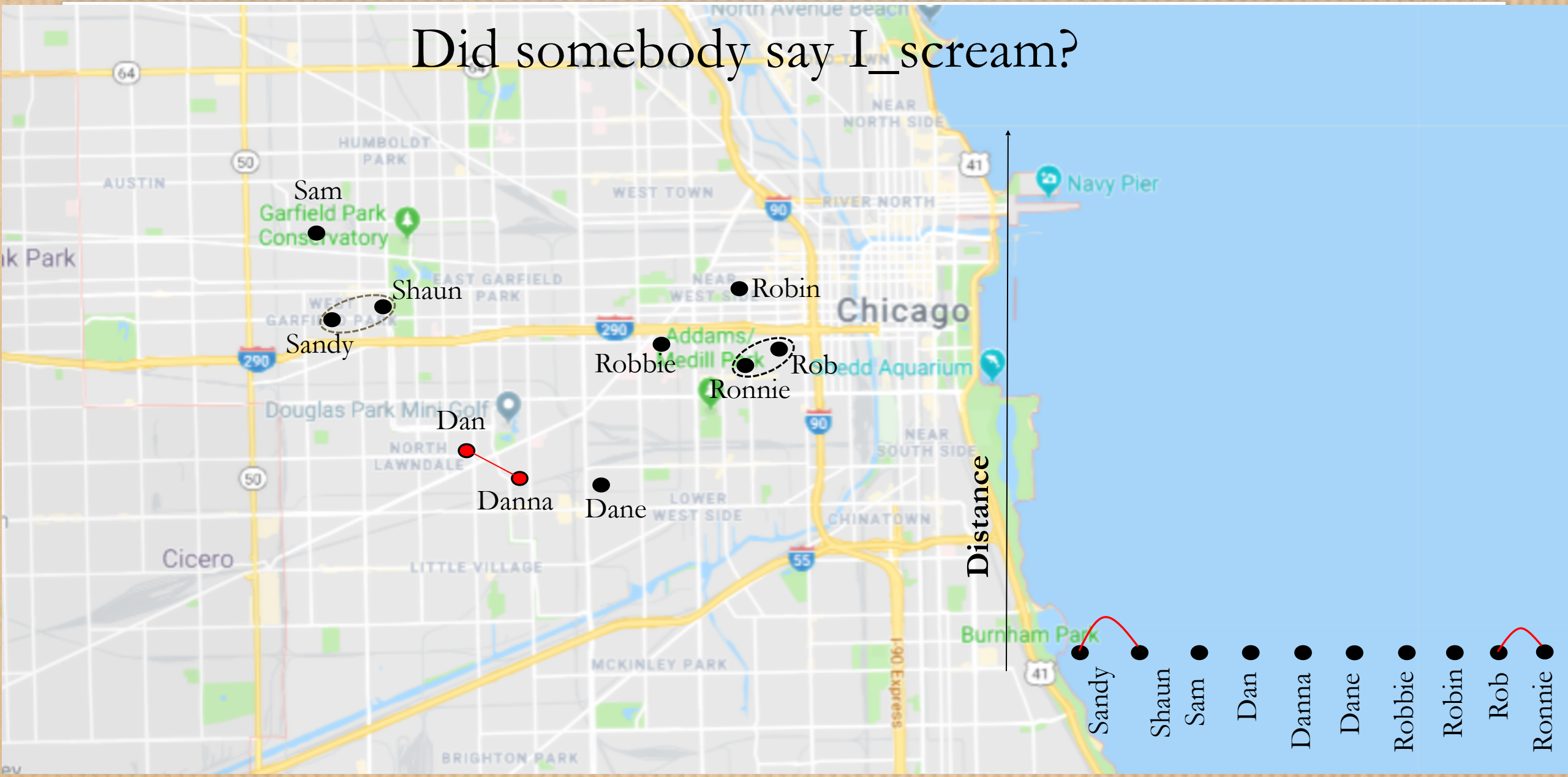
Did somebody say I_scream?



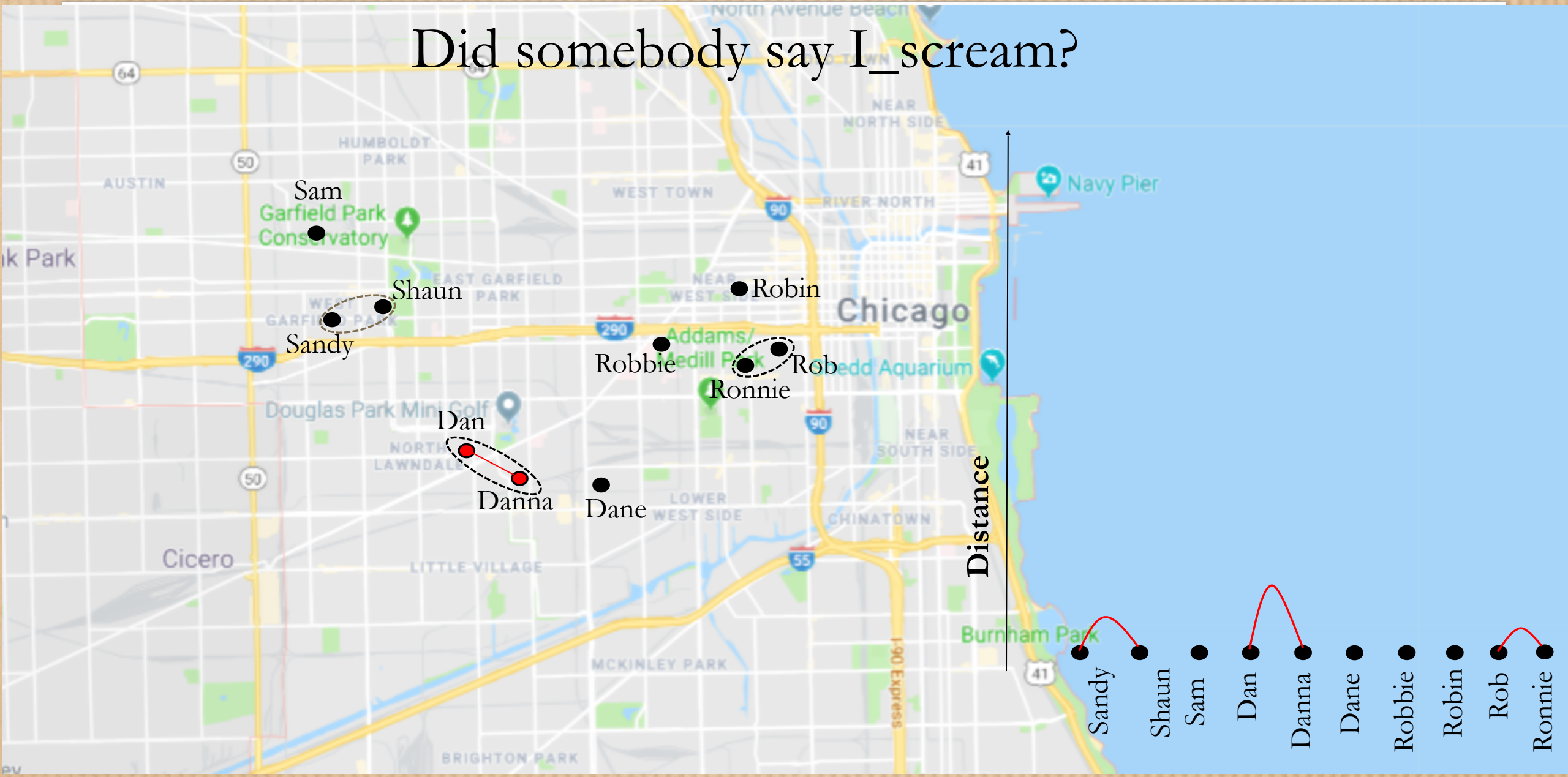
Did somebody say I_scream?



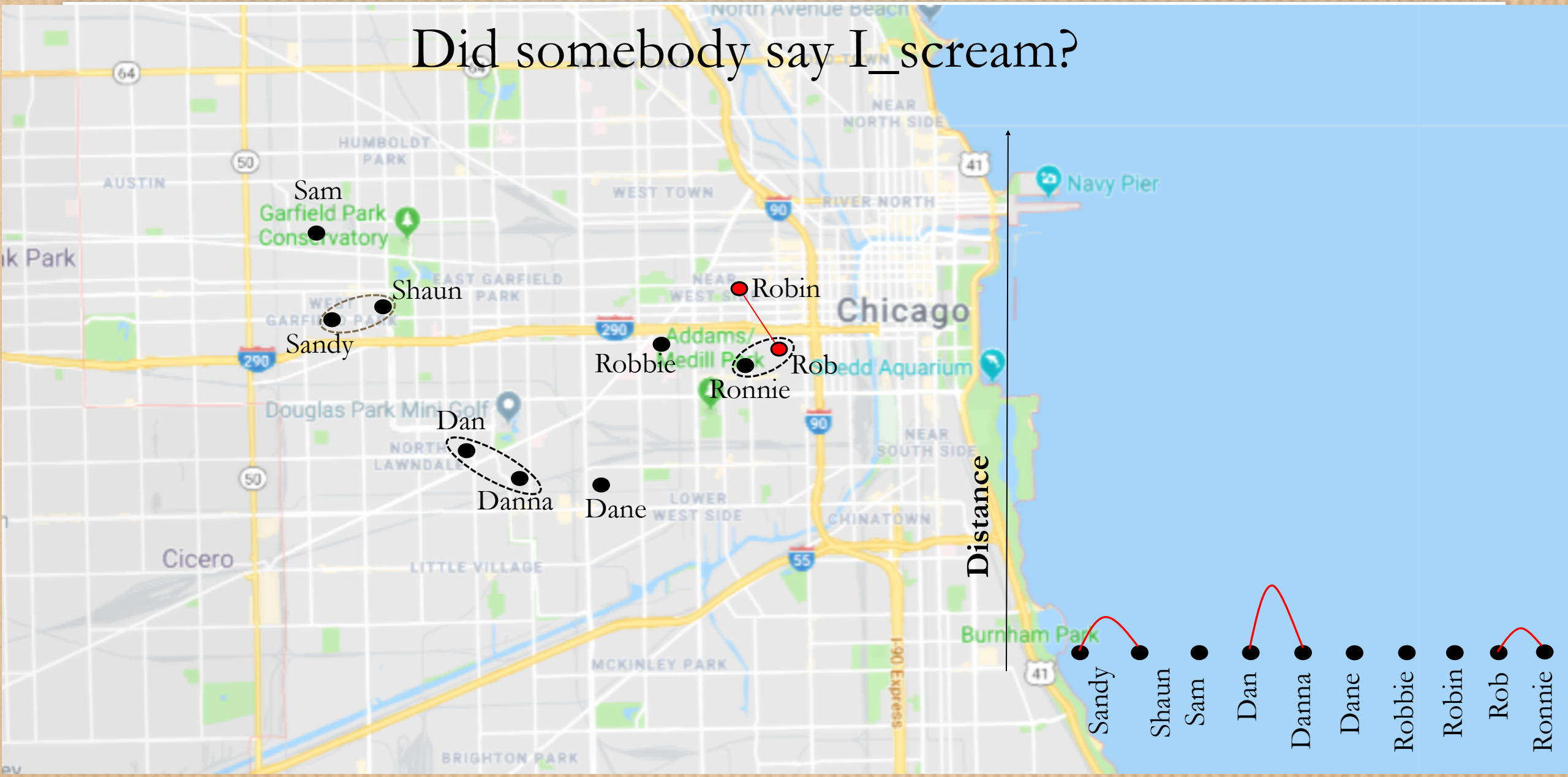
Did somebody say I_scream?



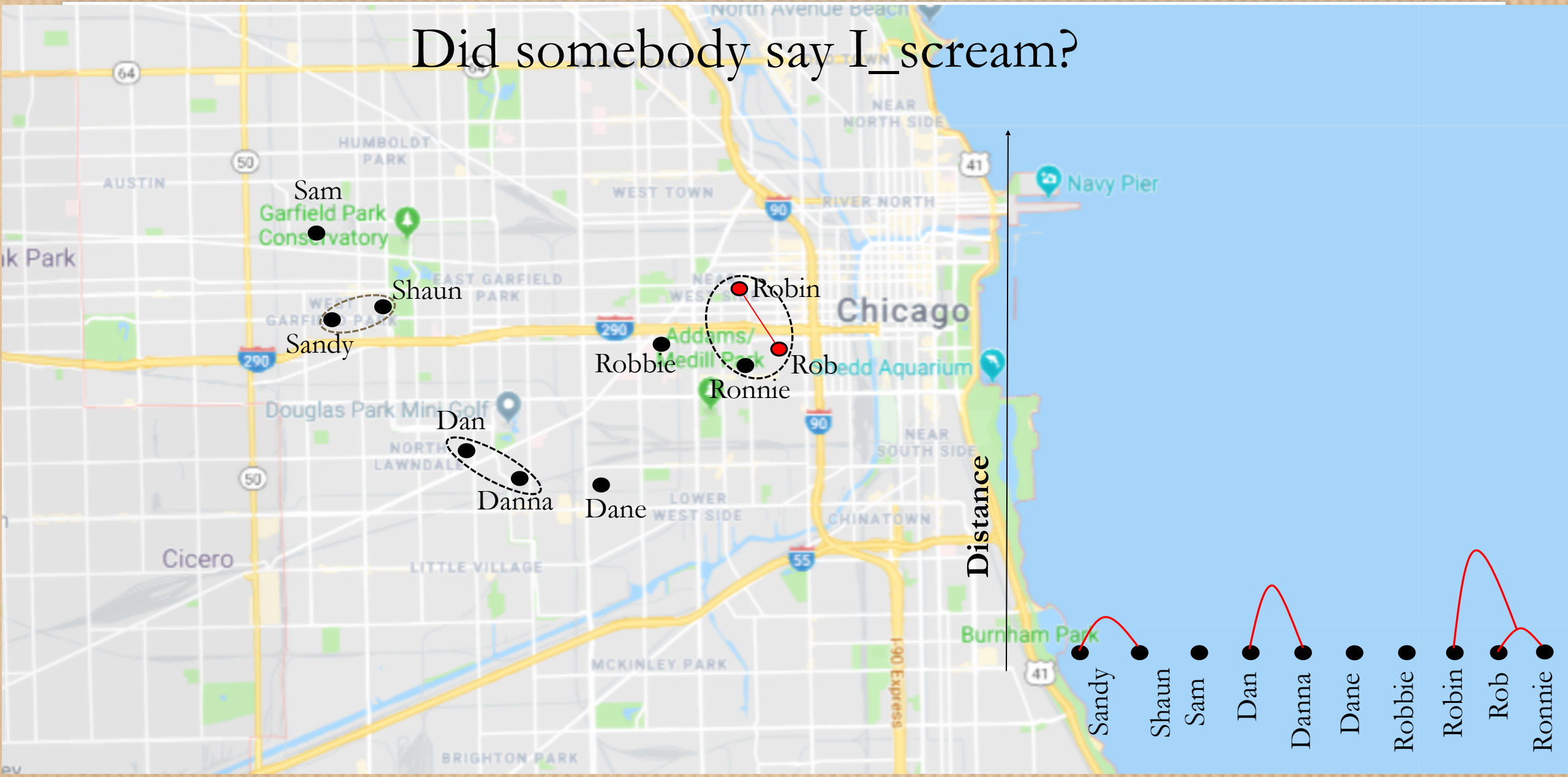
Did somebody say I_scream?



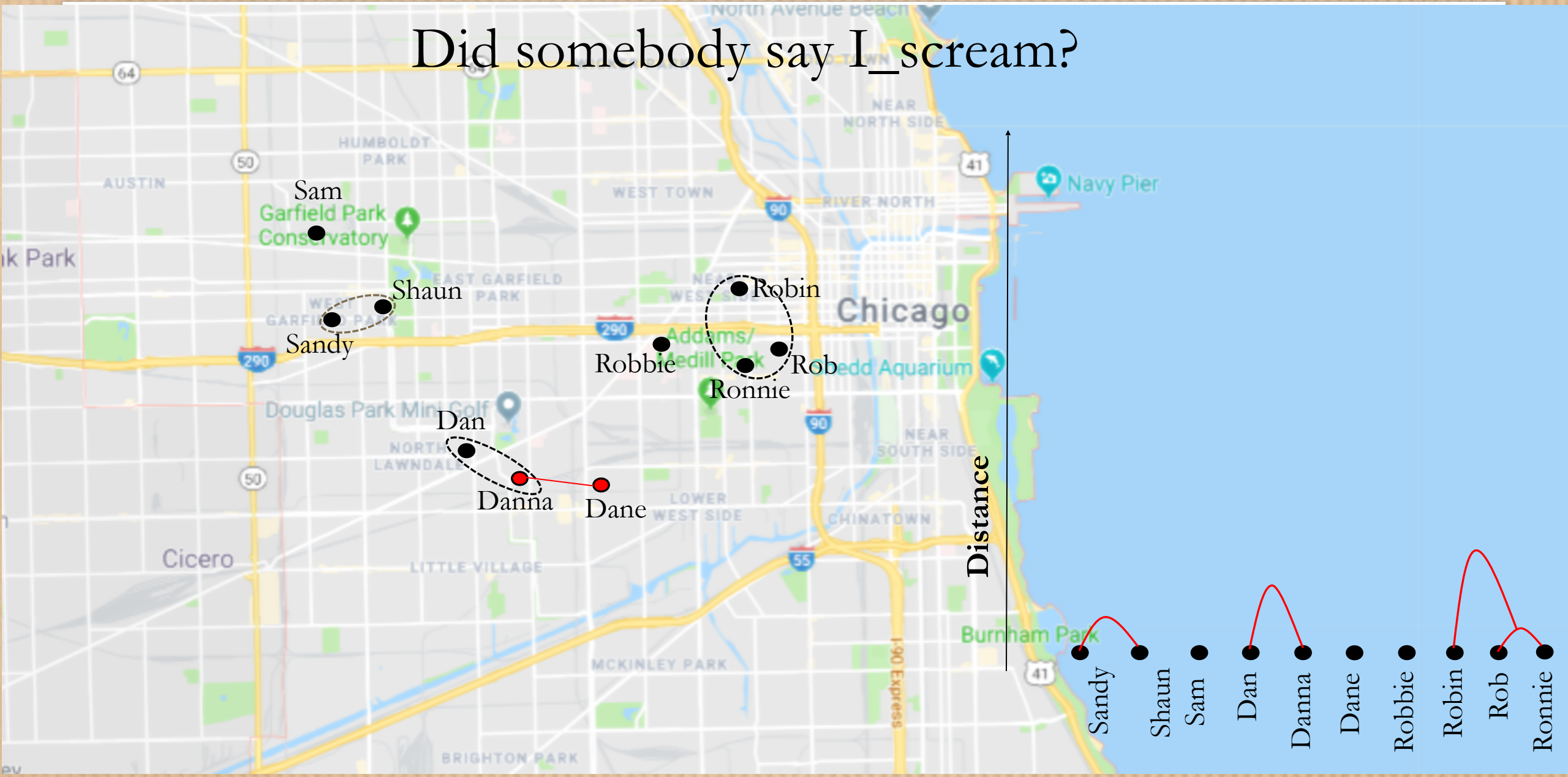
Did somebody say I_scream?



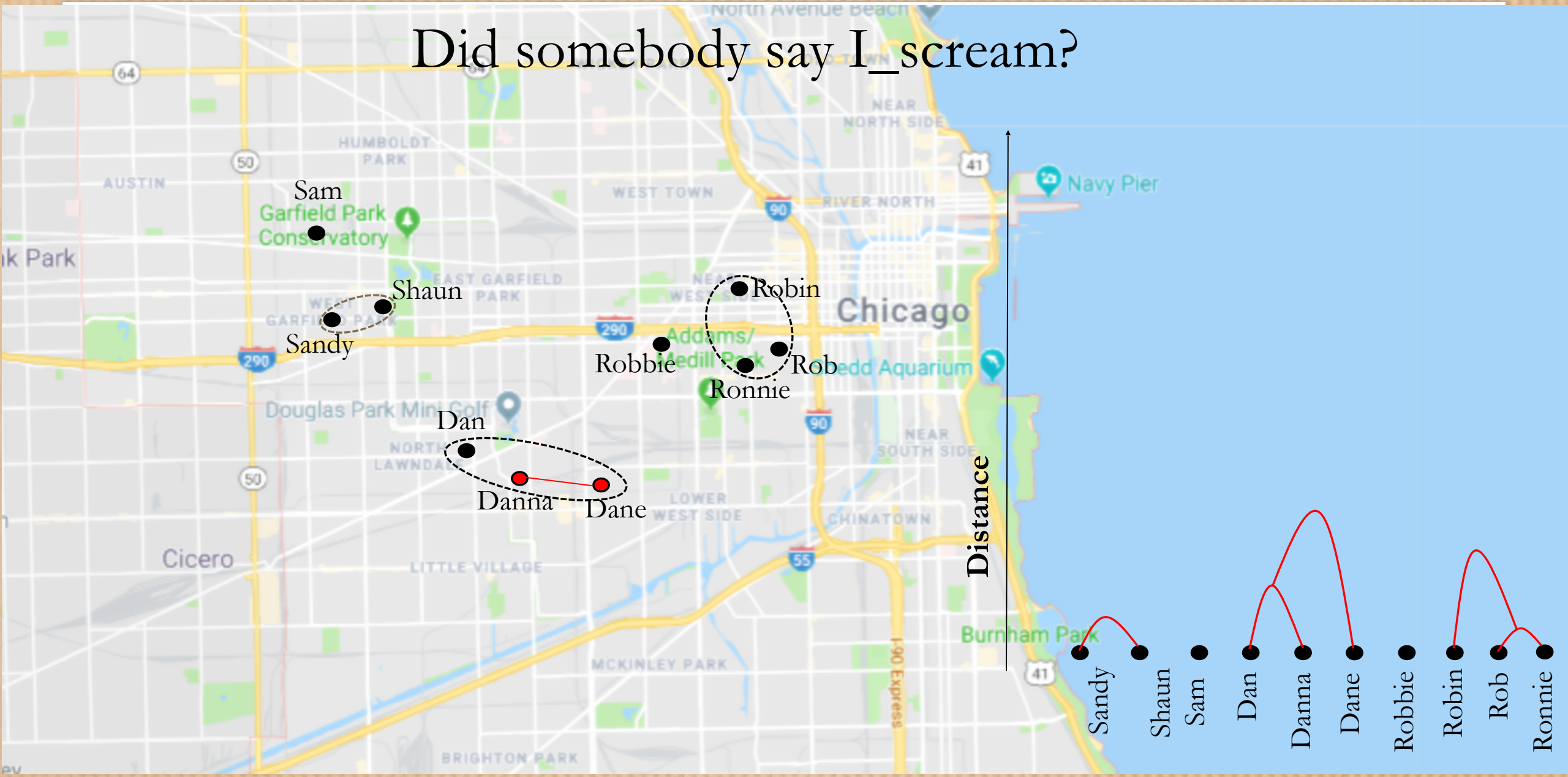
Did somebody say I_scream?



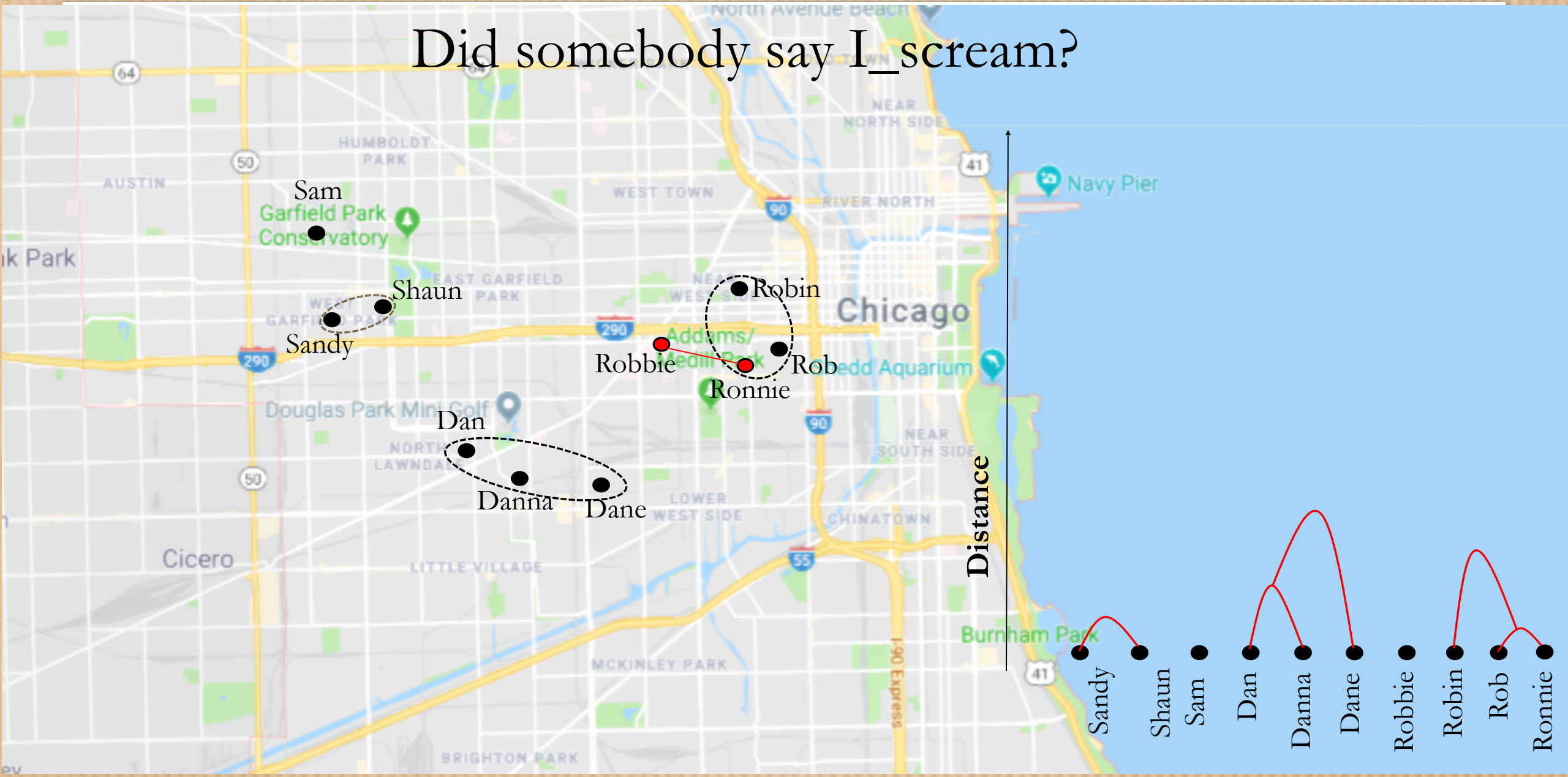
Did somebody say I_scream?



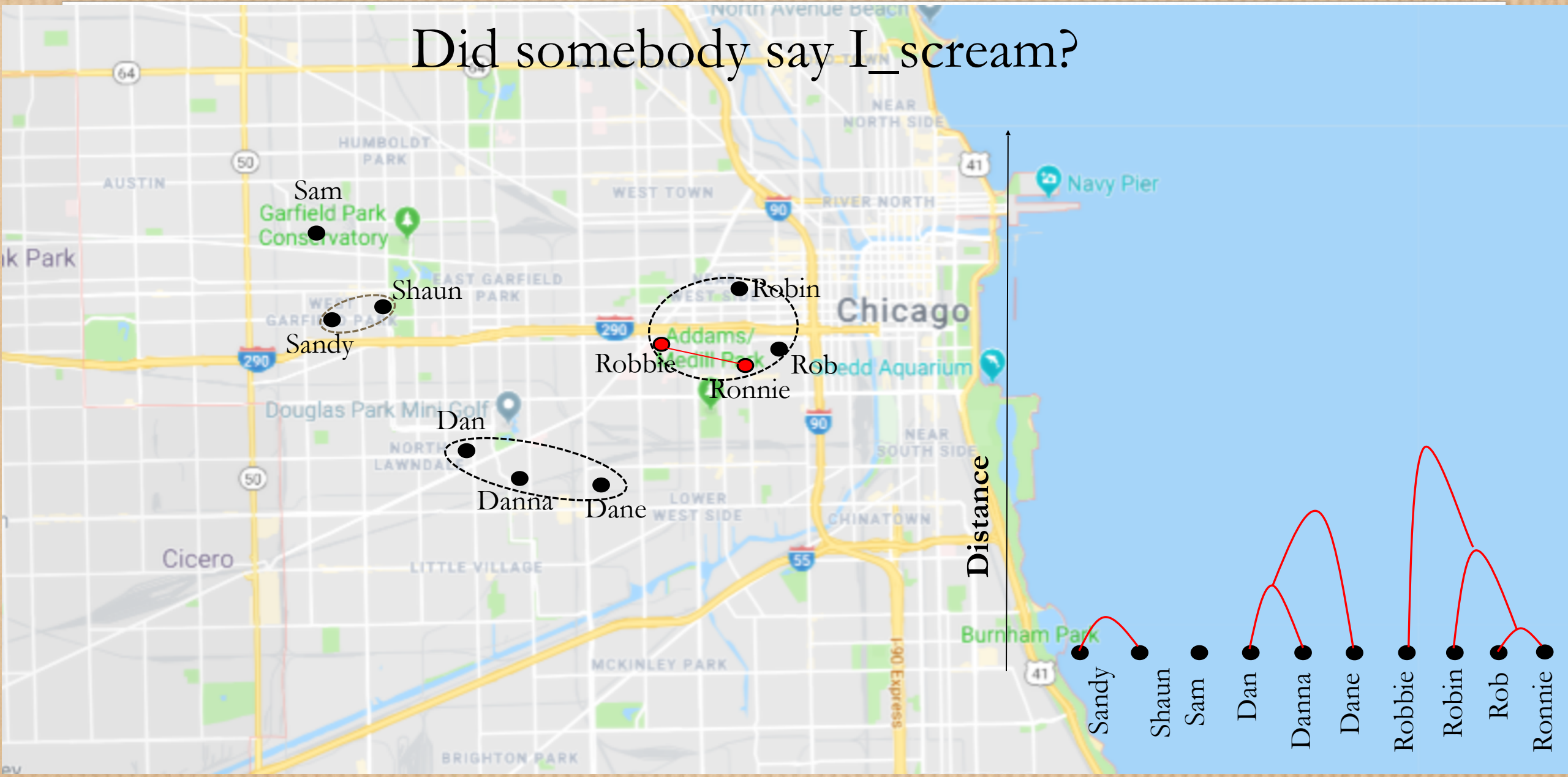
Did somebody say I_scream?



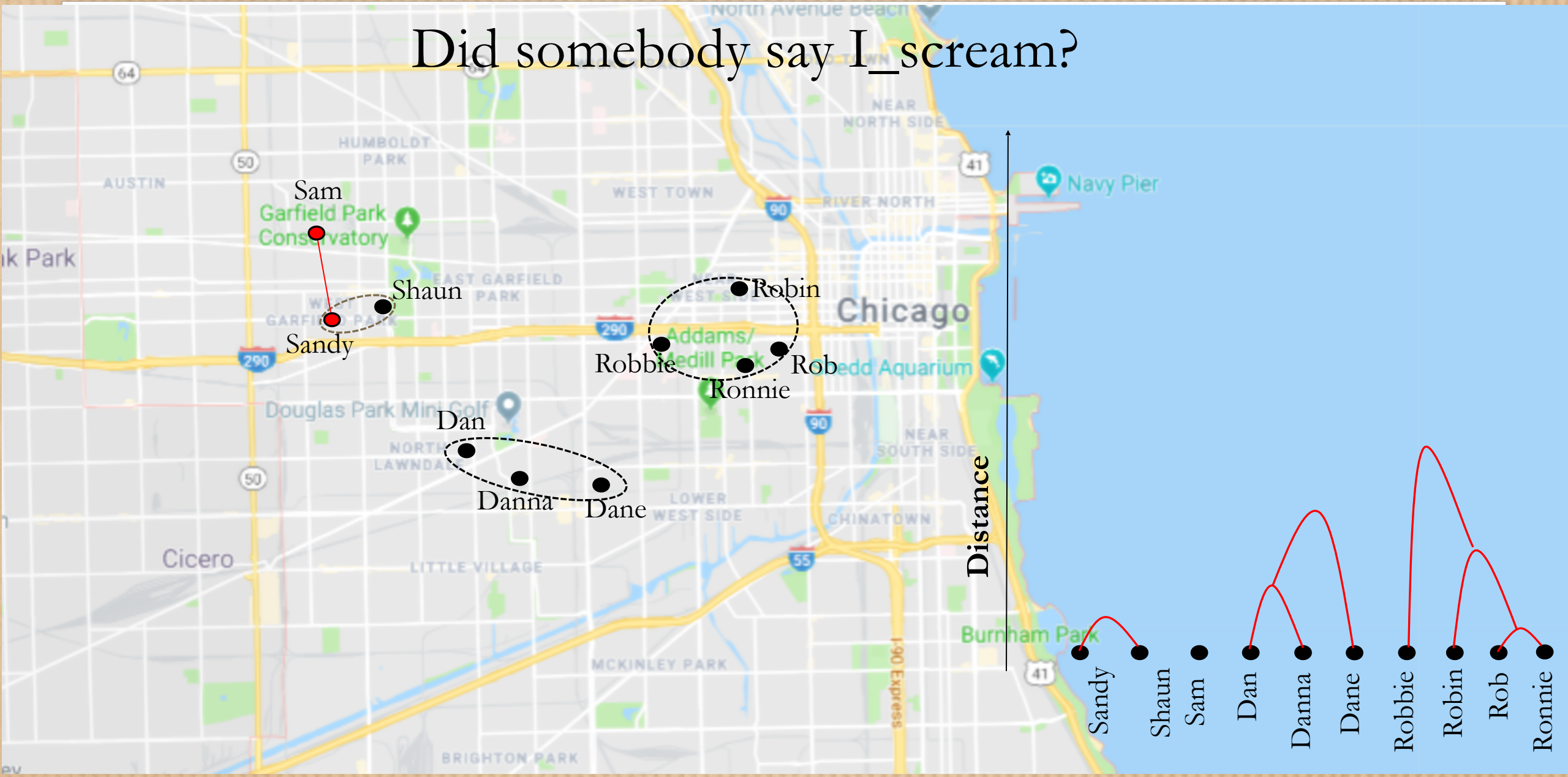
Did somebody say I_scream?



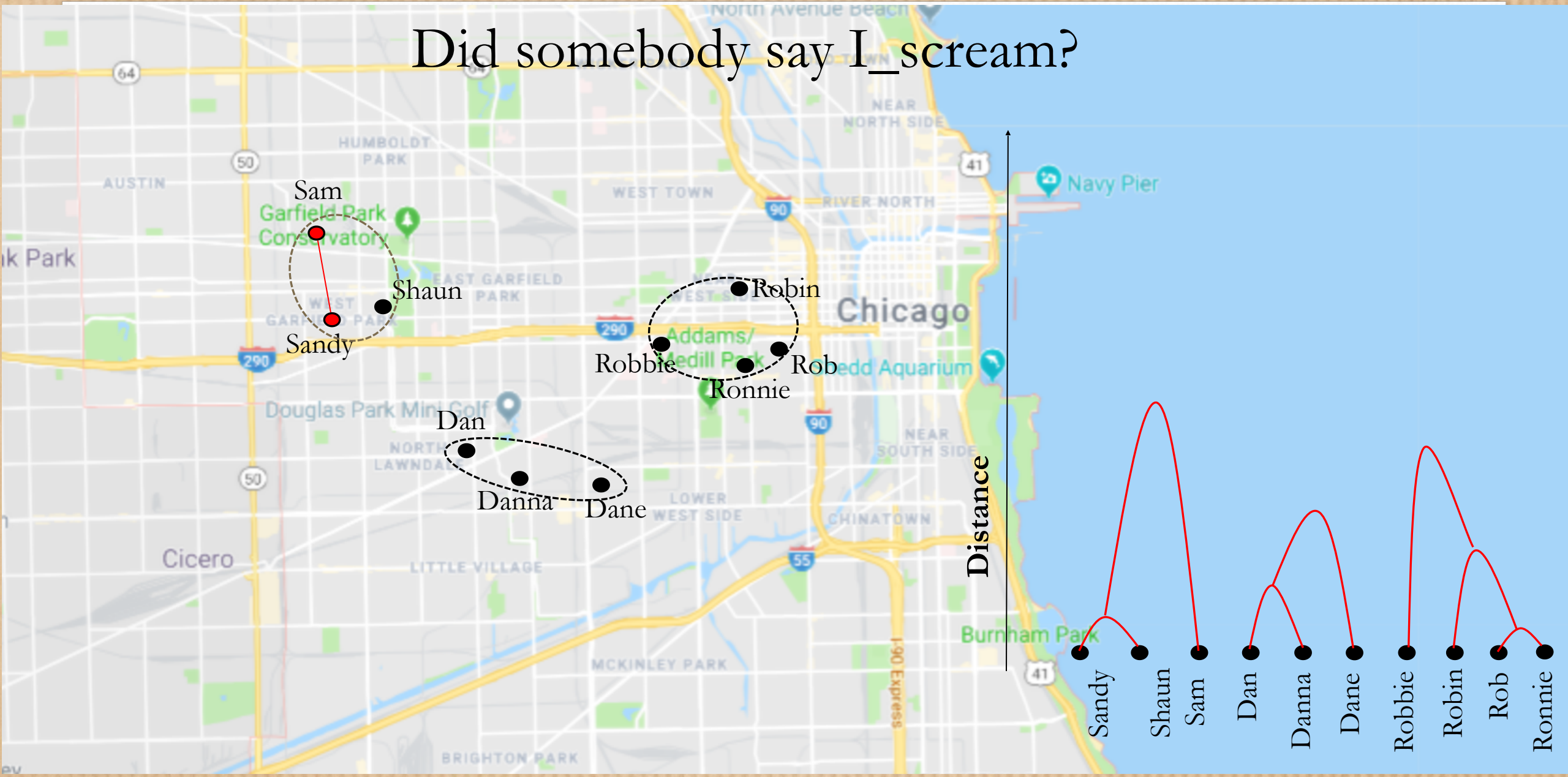
Did somebody say I_scream?



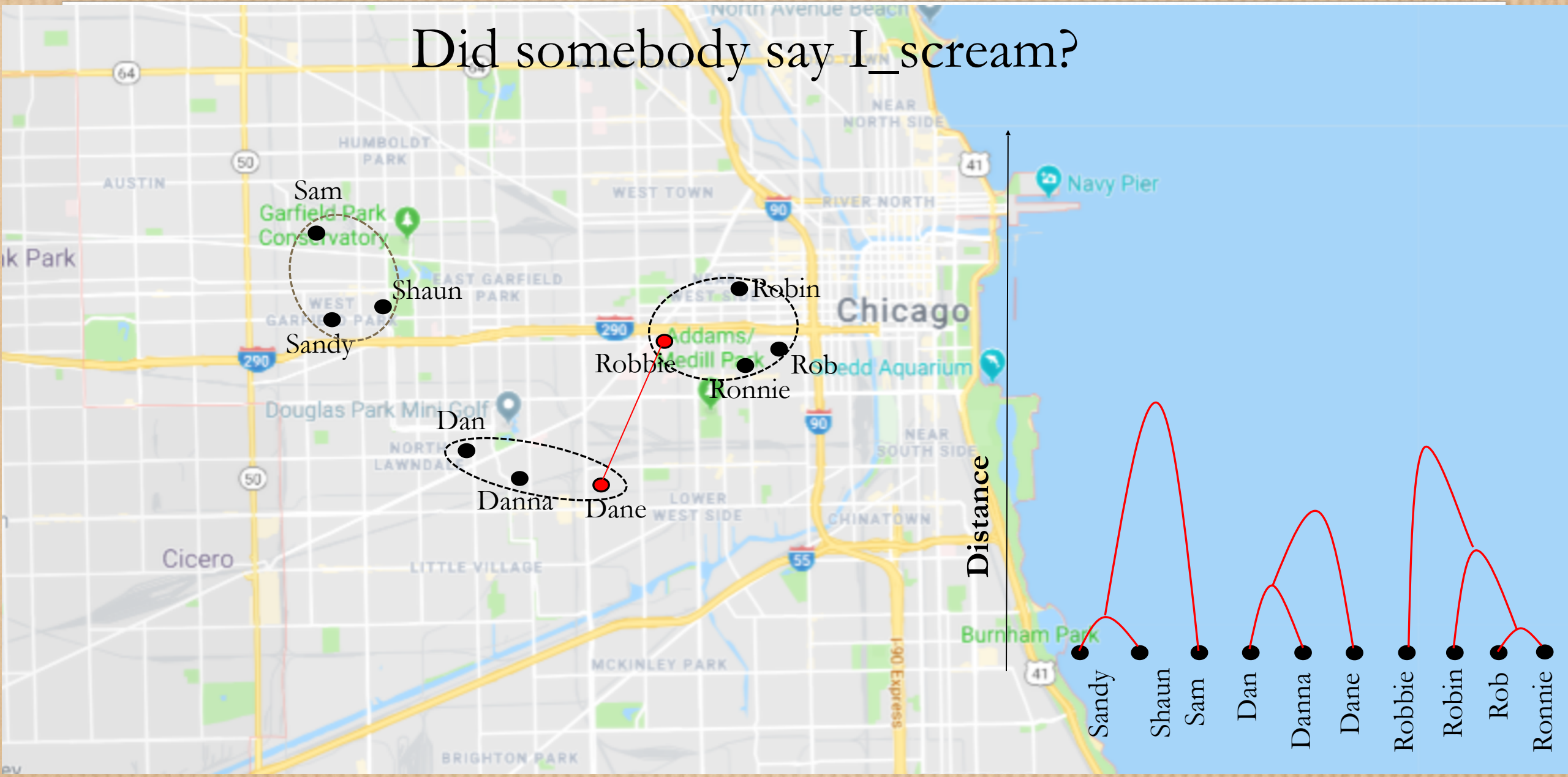
Did somebody say I_scream?



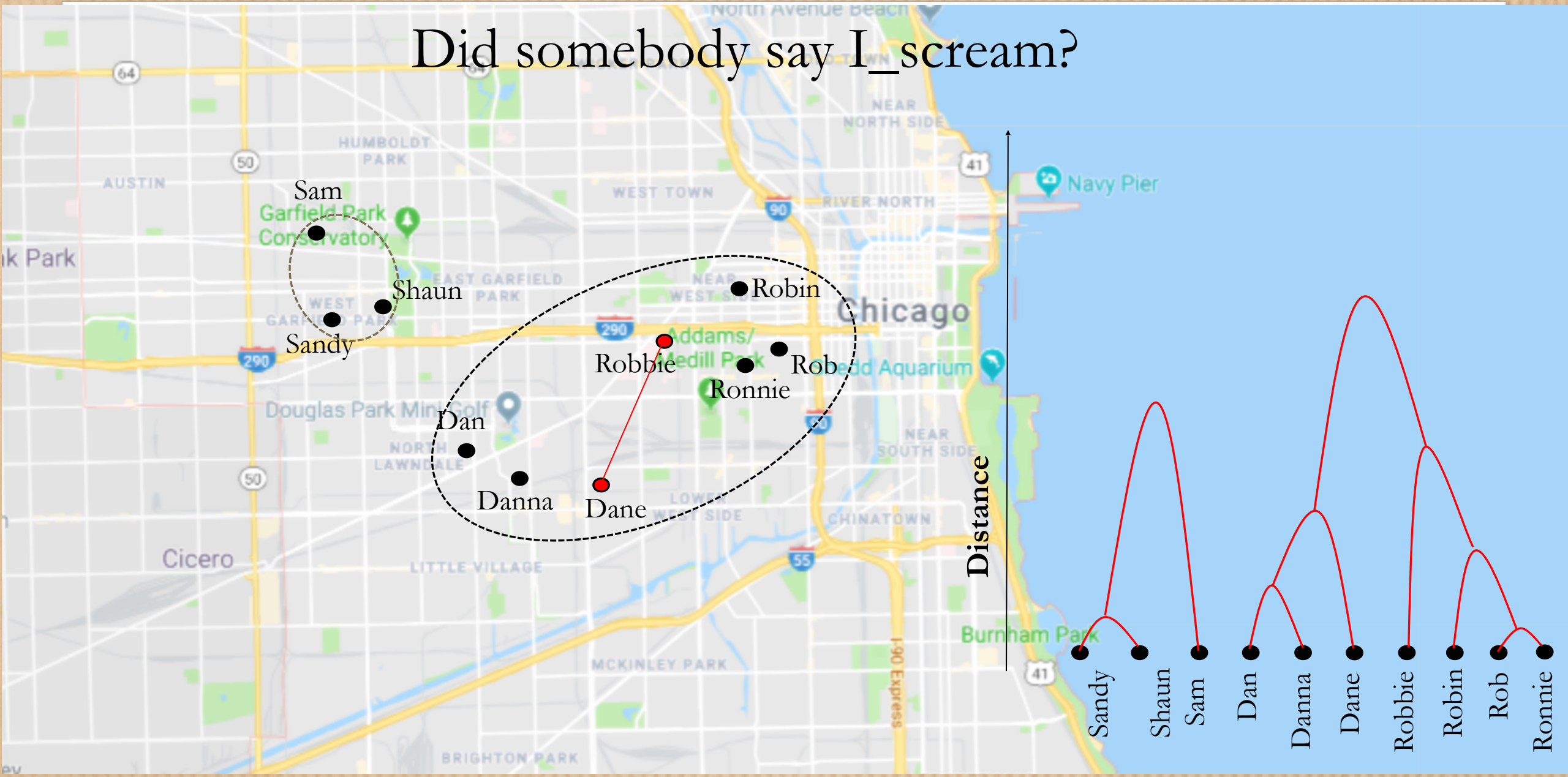
Did somebody say I_scream?



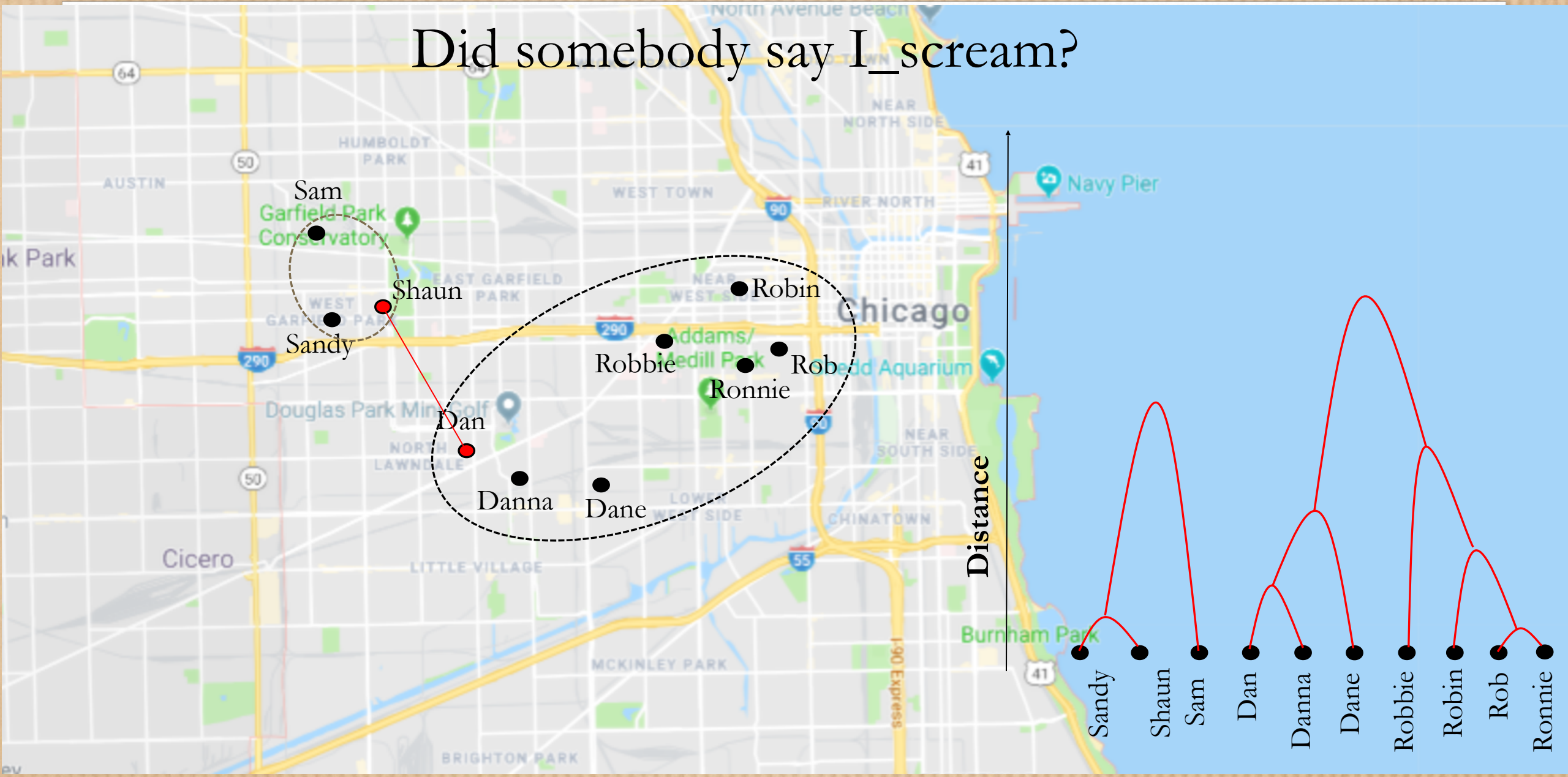
Did somebody say I_scream?



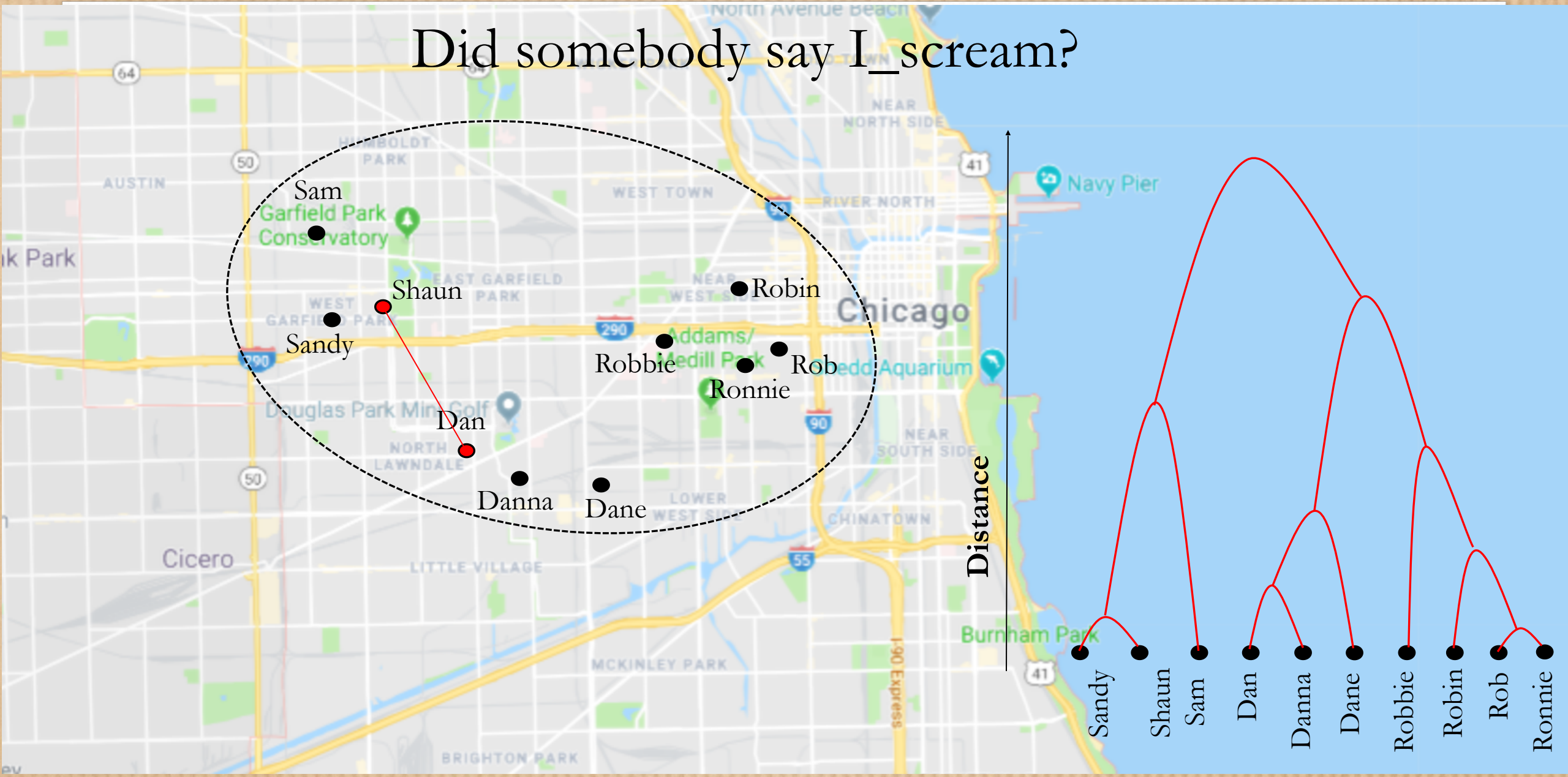
Did somebody say I_scream?



Did somebody say I_scream?

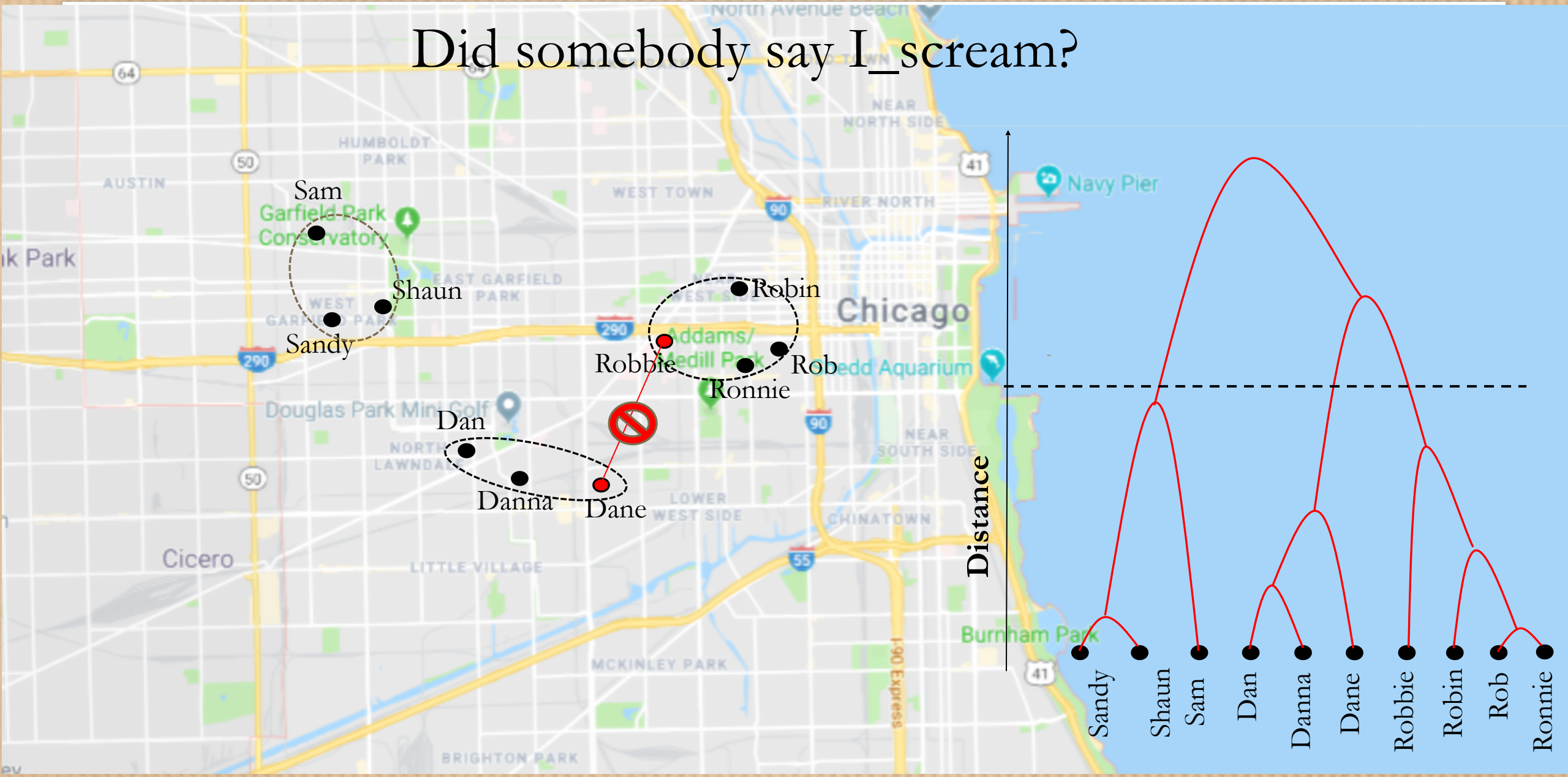


Did somebody say I_scream?

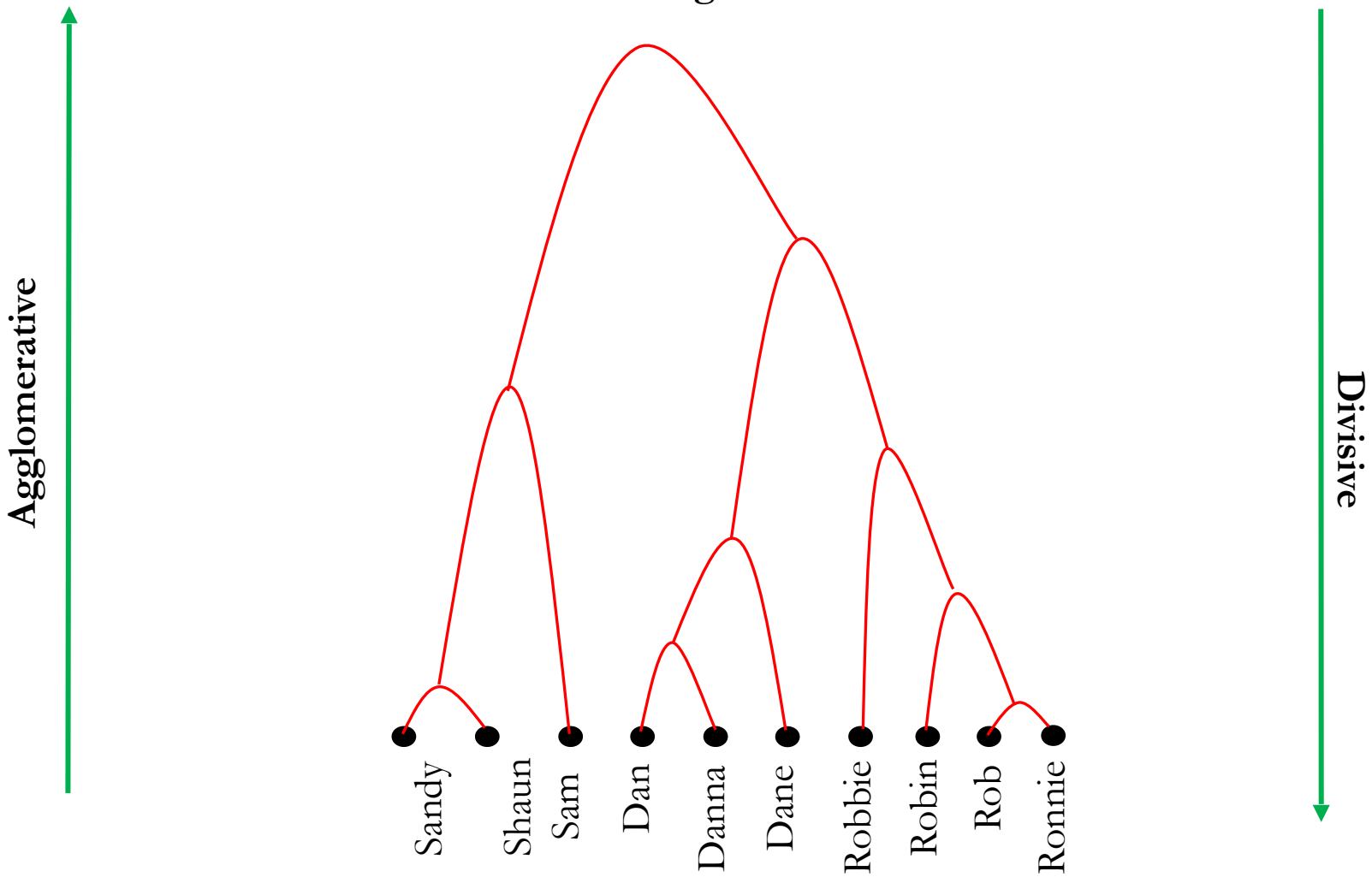


Did somebody say I_scream?

Did somebody say I_scream?



Dendrogram



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

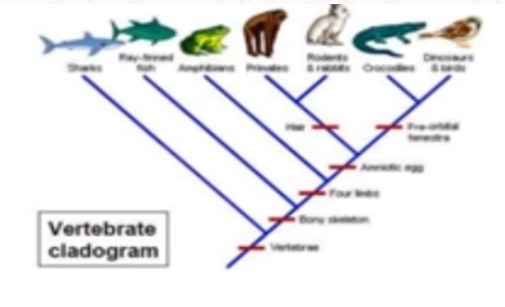
Pros & Cons

- 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

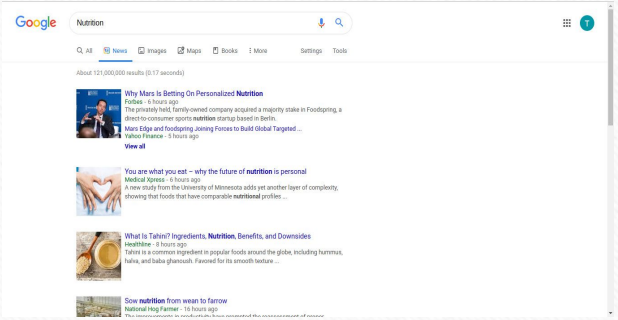
Genetics



Evolutionary Biology

Social Network

Recommender System



Google News

Python code implementation

- Jupyter Notebook

Discussion

Thank you!!!!!!