Lecture 11
Unsupervised learning

GEOL 4397: Data analytics and machine learning for geoscientists

Jiajia Sun, Ph.D.

March. 26th, 2019

UNIVERSITY of HOUSTON
YOU ARE THE PRIDE
EARTH AND ATMOSPHERIC SCIENCES

Week	Date	Topics	Comments
1	01/15 Tues	Overview of syllabus Lecture: Introduction to Machine learning: applications	
	01/17 Thur	Lecture: Review of linear algebra	
2	01/22 Tues	Lab: Linear algebra in Python	Not graded
	01/24 Thur	Lecture: Introduction to optimization	
3	01/29 Tues	Lab: Gradient descent + Linear regression	Report due on 02/05 at 5:30 pm
	01/31 Thur	Lecture: Introduction to machine learning: concepts	
4	02/05 Tues	Lecture: Logistic regression	
	02/07 Thur	Lab: Logistic regression	Report due on 02/14 at 5:30 pm
5	02/12 Tues	Lecture: Support vector machine	
	02/14 Thur	Lab: Support vector machine	Report due on 02/21 at 5:30 pm
6	02/19 Tues	Lecture: Decision trees	
	02/21 Thur	Lab: Decision trees	Report due on 02/28 at 5:30 pm
7	02/26 Tues	Lecture: Random Forest	
	02/28 Thur	Lab: Random forest	Report due on 03/07 at 5:30 pm
8	03/05 Tues	Lecture: Ensemble learning	
	03/07 Thur	Lab: Ensemble learning	Reprot due on 03/19 at 5:30 pm
9	03/12 Tues	No class due to spring break	
	03/14 Thur	No class due to spring break	
10	03/19 Tues	Review & Recap	
	03/21 Thur	Exam	
11	03/26 Tues	Lecture: Clustering	
	03/28 Thur	Lab: Clustering	Report due on 04/04 at 5:30 pm
12	04/02 Tues	Lecture: Introduction to TensorFlow	
	04/04 Thur	Lab: TensorFlow	Not graded
13	04/09 Tues	Lecture: Introduction to neural networks 1	
	04/11 Thur	Lecture: Introduction to neural networks 2	
14	04/16 Tues	Lab: Deep learning	Report due on 04/23 at 5:30pm
	04/18 Thur	Lecture: Convolutional neural networks 1	
15	04/23 Tues	Guest lecture: Convolutional neural networks 2	
	04/25 Thur	Lab: CNN (optional)	Report due on 05/02 at 5:30 pm
16	04/30 Tues	final presentation??	
	05/02 Thur	final presentation??	
Note	28 class meetings		04/29 last day of class

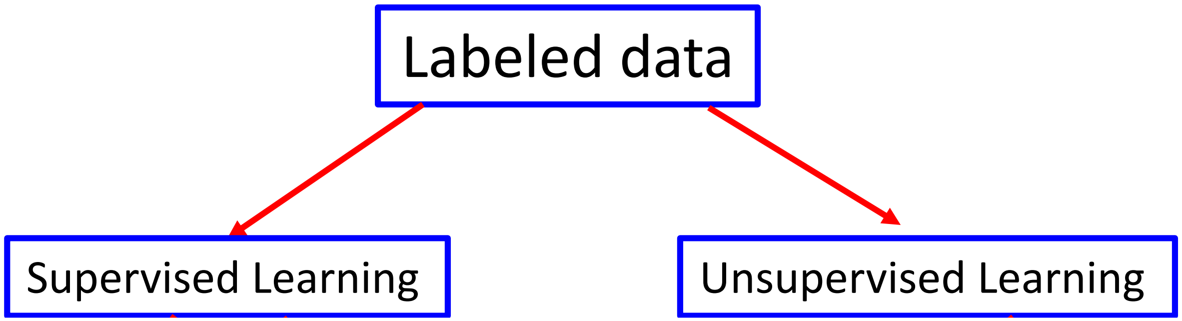
Outline

- Dimensionality reduction
- K-means Clustering
- Implementation in Scikit-Learn

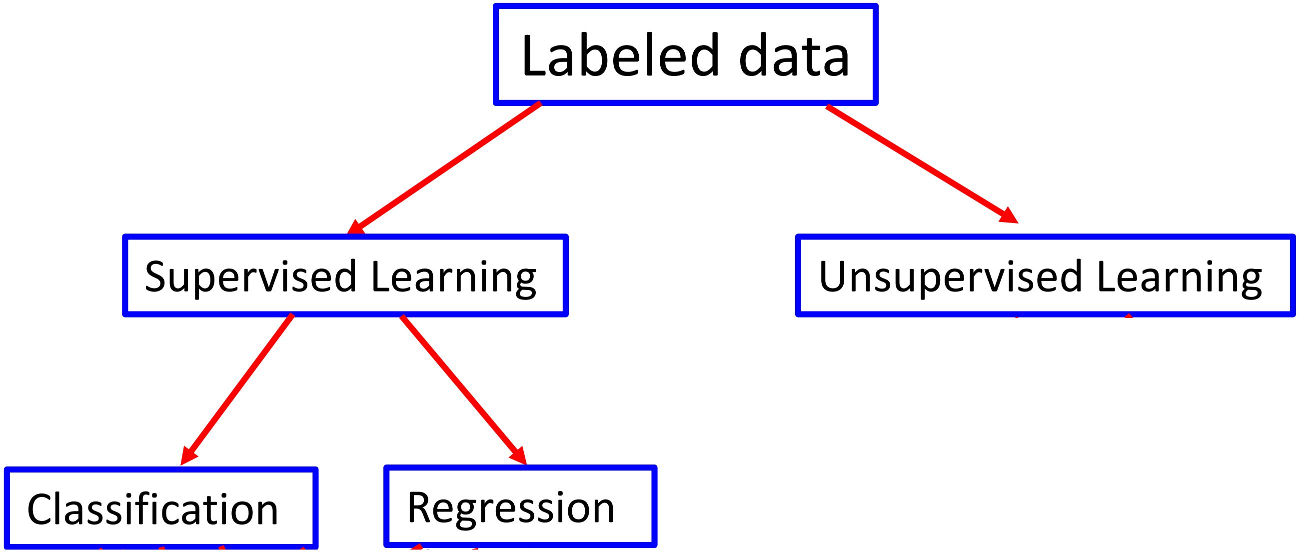
Acknowledgments

- Youtube video by Joshua Starmer: https://goo.gl/RDMb4P
- Youtube video by Luis Serrano: https://goo.gl/wuSYXK

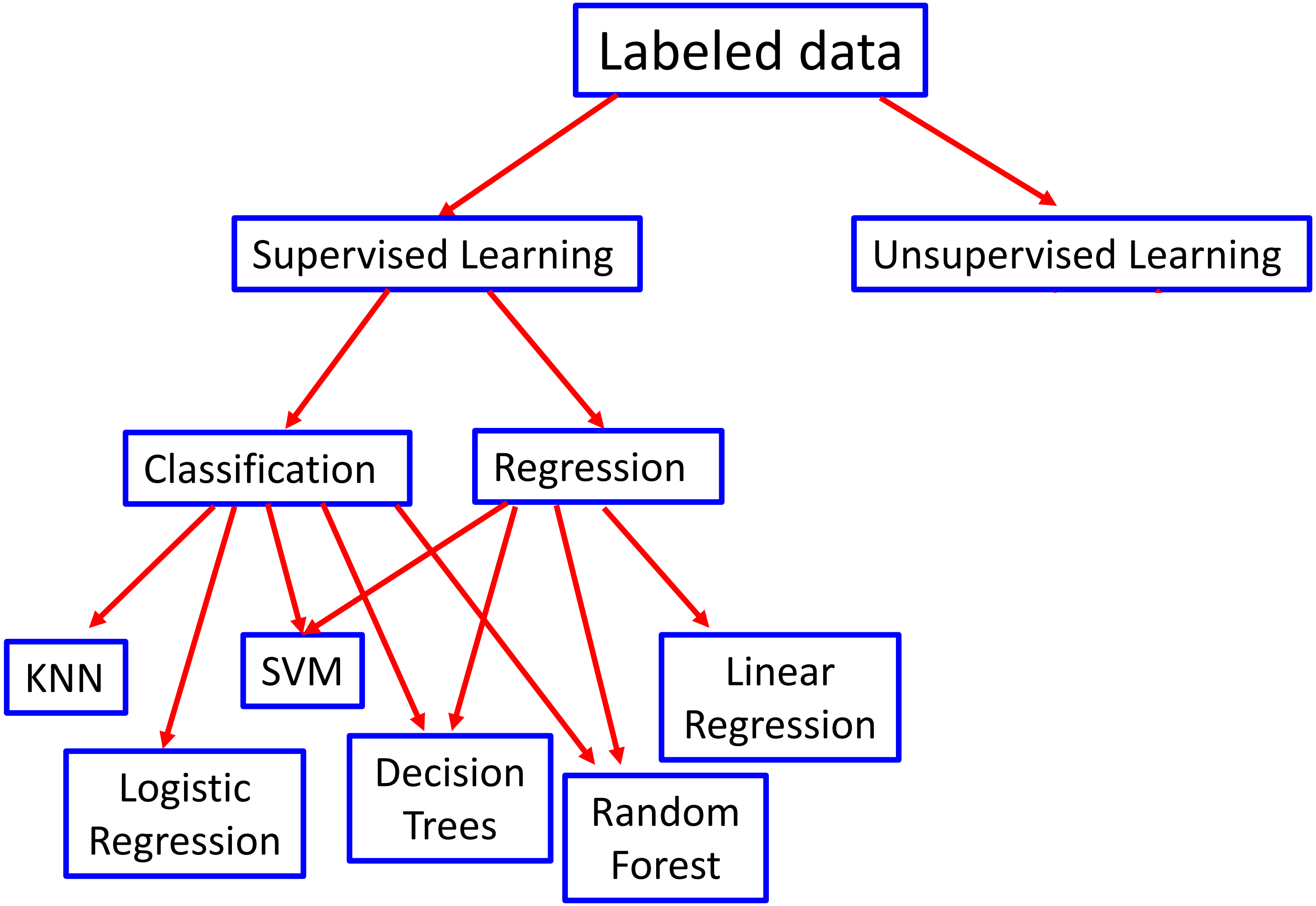
Machine learning algorithms



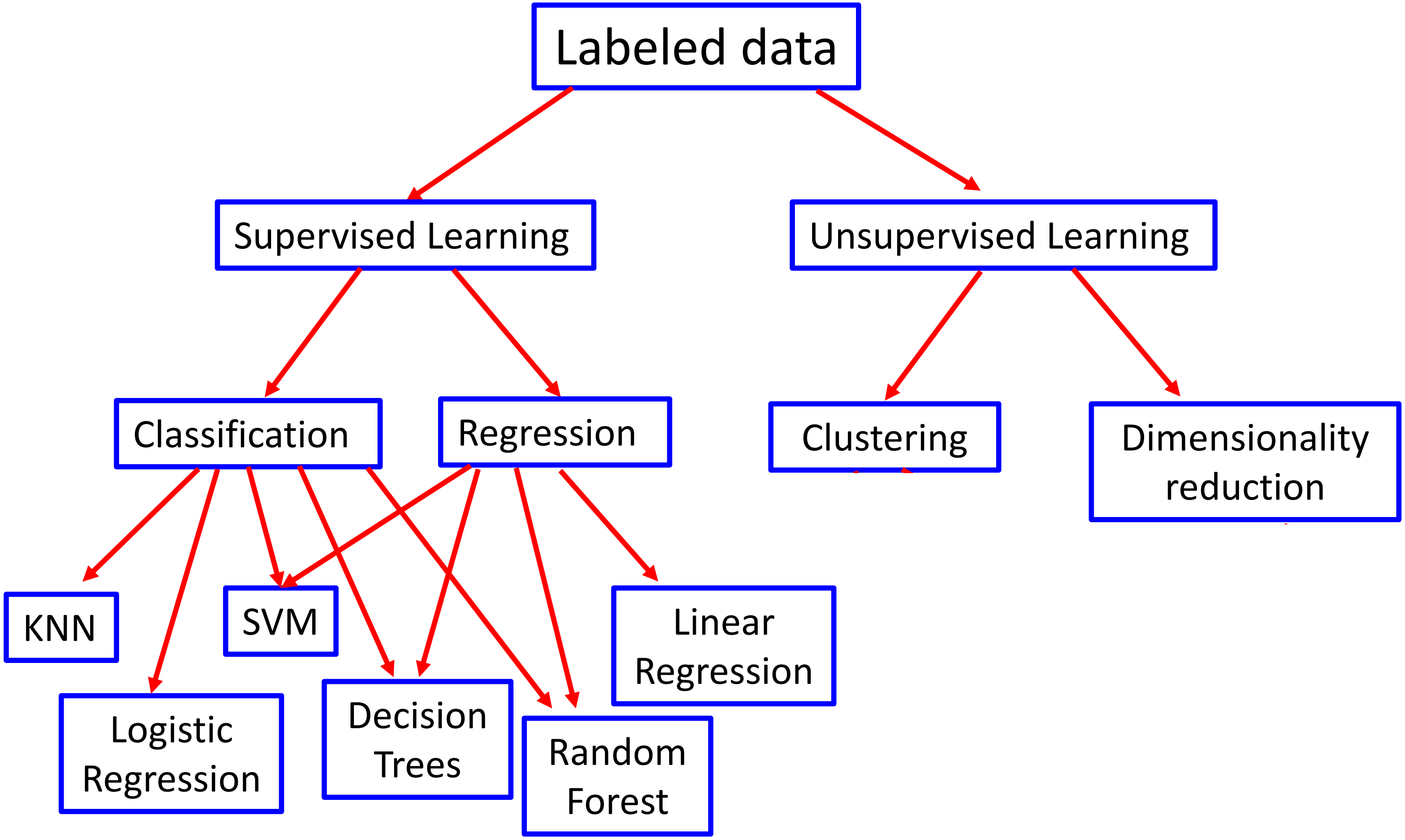
Machine learning algorithms



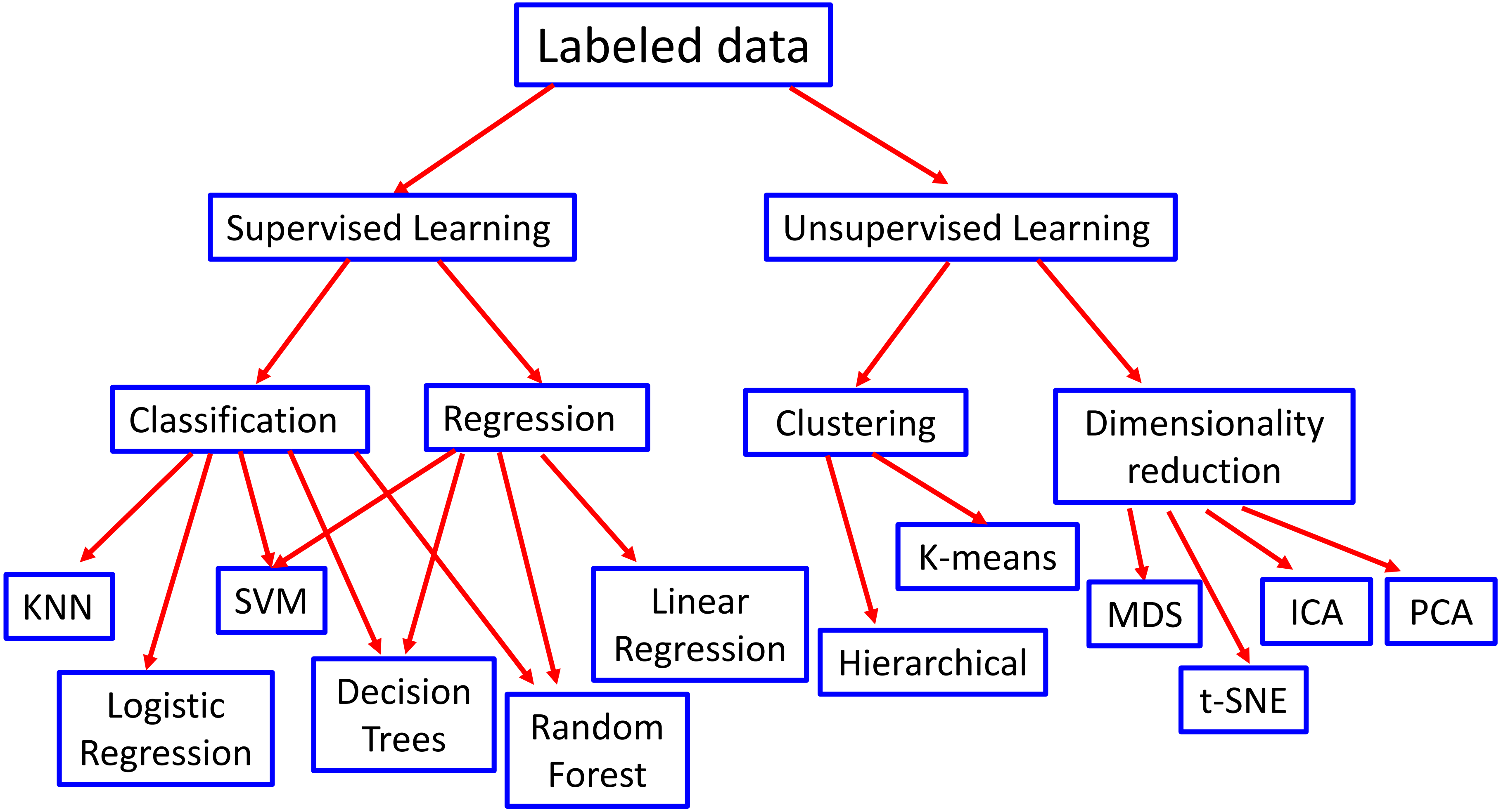
Machine learning algorithms



Machine learning algorithms



Machine learning algorithms



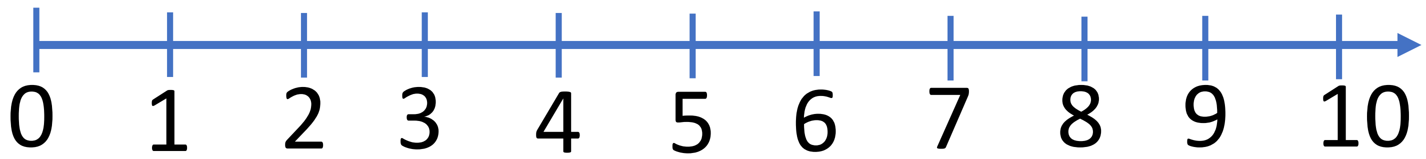
Dimensionality reduction

- Reduces high-dimensional data into 2D (or 3D) space for better visualization and analysis

Introduction to dimensions

- Fundamental yet important concepts

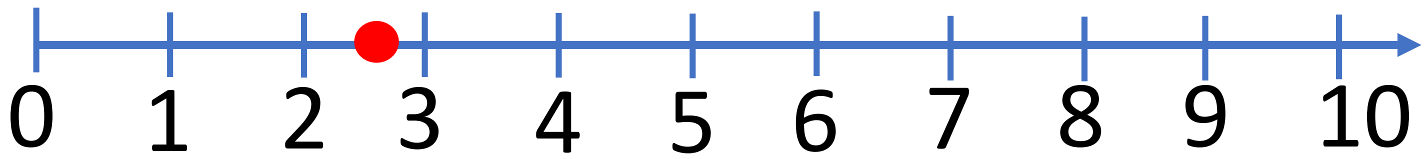
1-Dimension (1D)



1-Dimension (1D) = a number line

1-Dimension (1D)

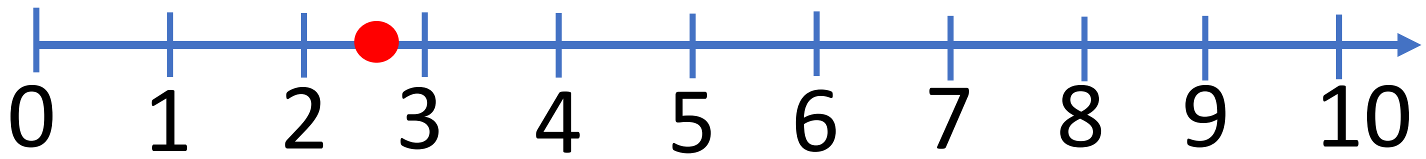
1-Dimension (1D) = a number line



Suppose I measure the average density of some crustal rocks:
2.67 g/cc

1-Dimension (1D)

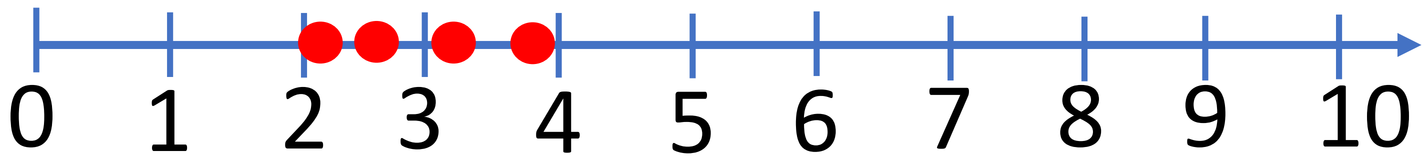
1-Dimension (1D) = a number line



Suppose I measured densities on several different rocks:
2.14 g/cc, 2.67 g/cc, 3.25 g/cc, 3.86 g/cc

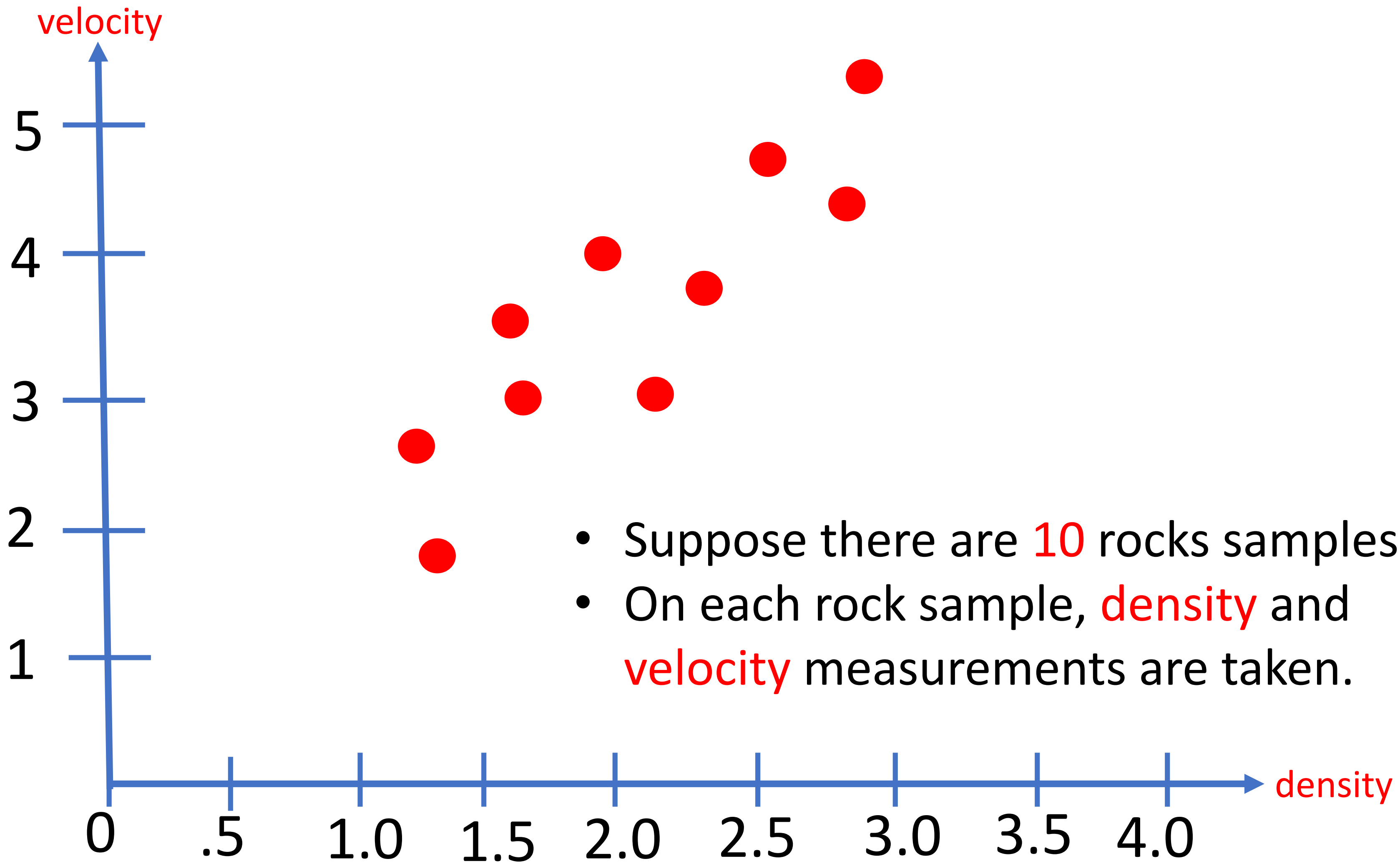
1-Dimension (1D)

1-Dimension (1D) = a number line



Suppose I measured densities on several different rocks:
2.14 g/cc, 2.67 g/cc, 3.25 g/cc, 3.86 g/cc

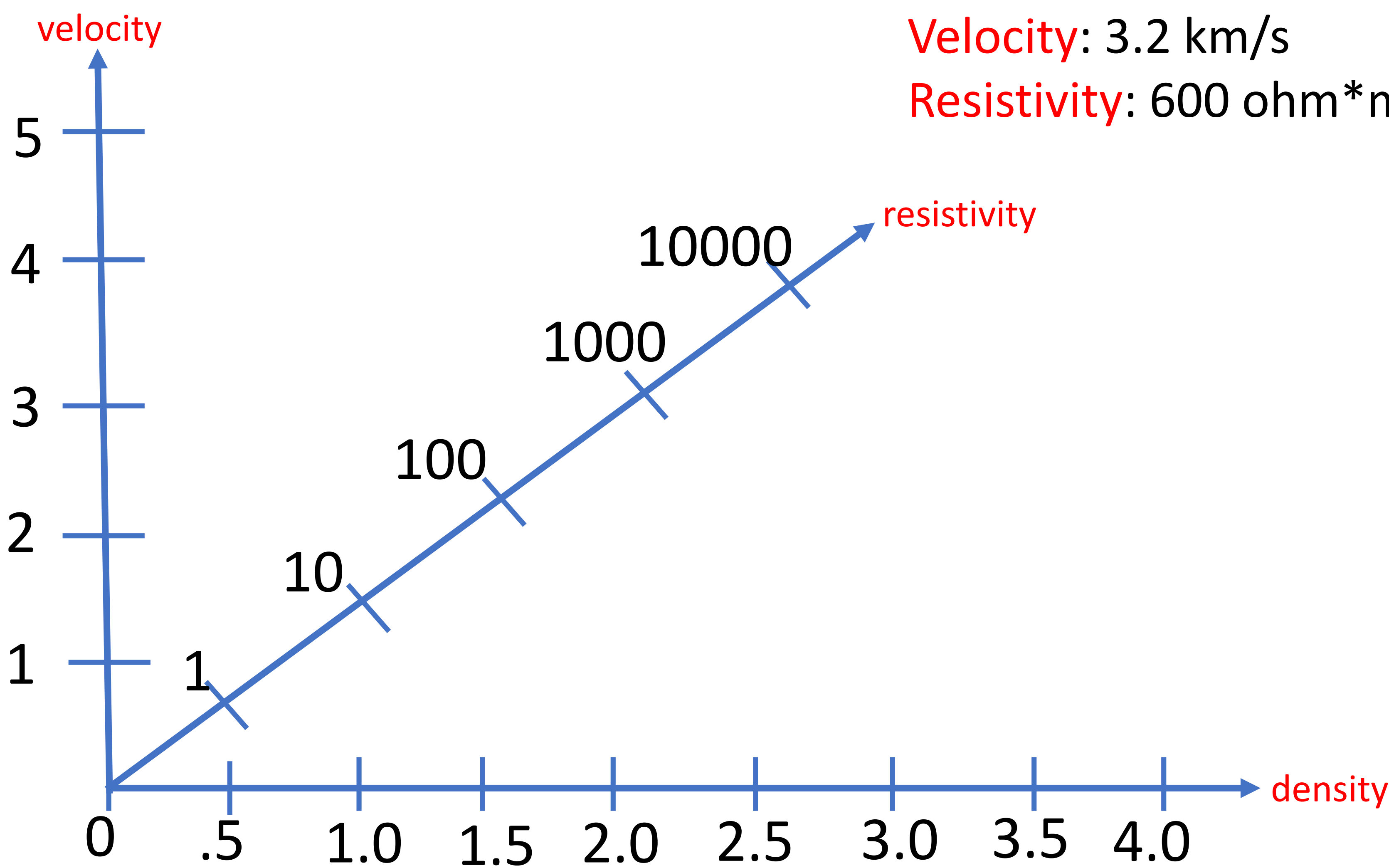
2-Dimension (2D)



- Suppose there are 10 rocks samples.
- On each rock sample, density and velocity measurements are taken.

3-Dimension (3D)

Density: 2.65 g/cc
Velocity: 3.2 km/s
Resistivity: 600 ohm*m



Dimensions so far …

- 1 measurement = 1D graph
- 2 measurements = 2D graph
- 3 measurements = 3D graph

Dimensions so far …

- 1 measurement = 1D graph
- 2 measurements = 2D graph
- 3 measurements = 3D graph
- 4 measurements = 4D graph (you cannot draw it)

Dimensions so far …

- 1 measurement = 1D graph
- 2 measurements = 2D graph
- 3 measurements = 3D graph
- 4 measurements = 4D graph (you cannot draw it)
- 200 measurements = 200D graph

Dimensions so far …

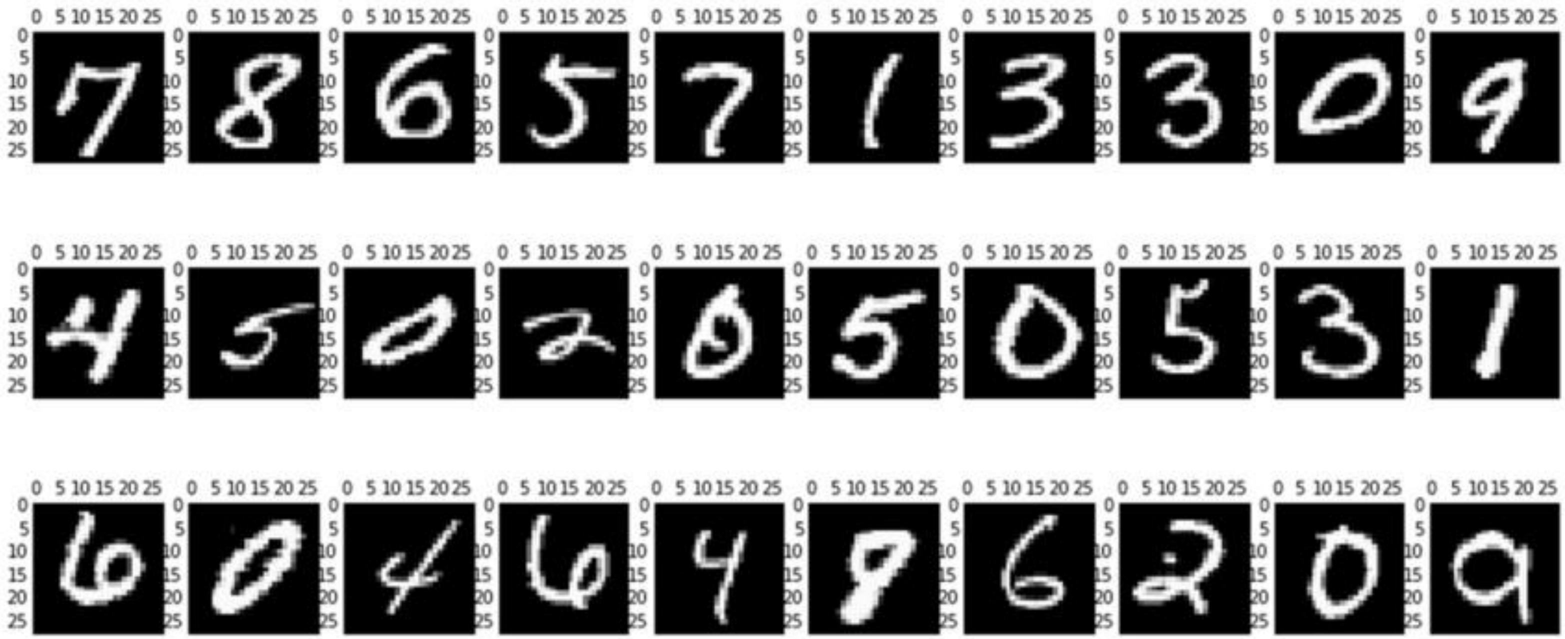
- 1 measurement = 1D graph
- 2 measurements = 2D graph
- 3 measurements = 3D graph
- 4 measurements = 4D graph (you cannot draw it)
- 200 measurements = 200D graph

Each more physical property measurement we make on one rock sample adds one more dimension.

In-class quiz

- Geochemical facies analysis
- Data consists of XRF measurements of cuttings from the lateral section of an unconventional well
- Measurements made at approximately 10 m intervals
- For each cutting sample, there were 22 measurements.
- A total of 269 cutting samples
- Question: If we plot up the measurements, what is the dimension of the space? How many points are there in this high dimensional space?

In-class quiz 2



Some examples of MNIST handwritten digits
Each image is a 28 X 28 pixel images
A total of 70,000 images
Question: Suppose we want to plot up all 70,000 images in a high dimensional space, what should the dimension of this space be?

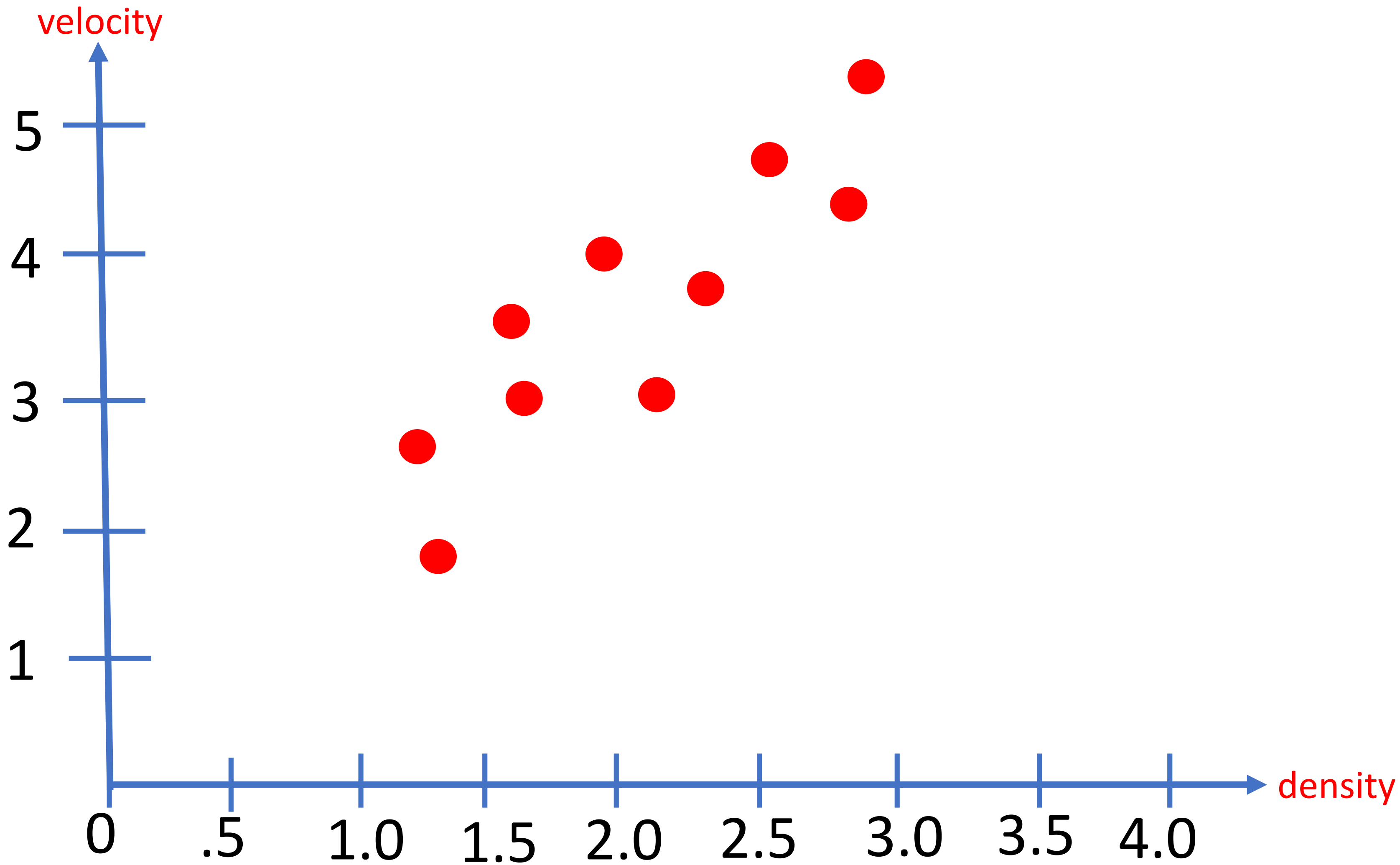
Questions

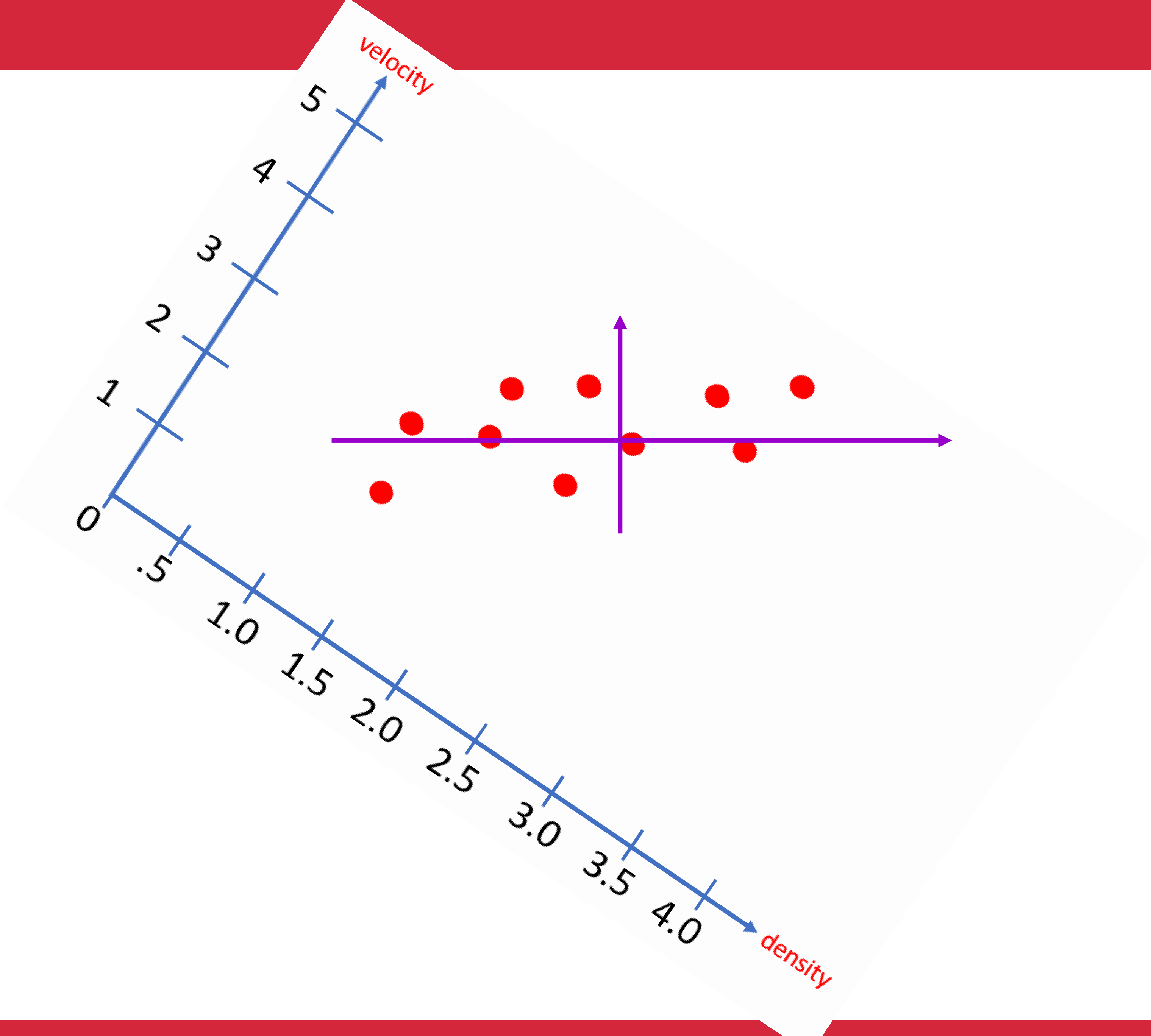
- How to visualize high-dimensional data?
- Are all those dimensional equally important? Or some more important than others?

PCA

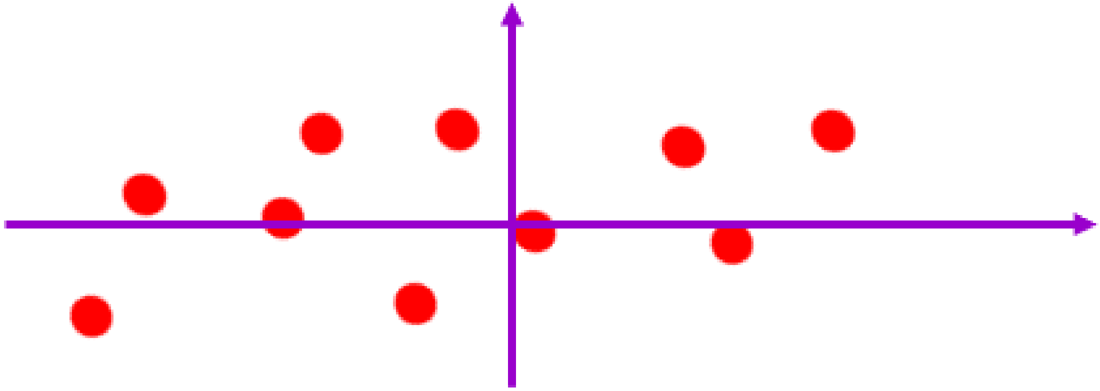
- Reduces a high-dimensional dataset (e.g., 22-D dimensional) to 2 or 3 dimensions for better visualization

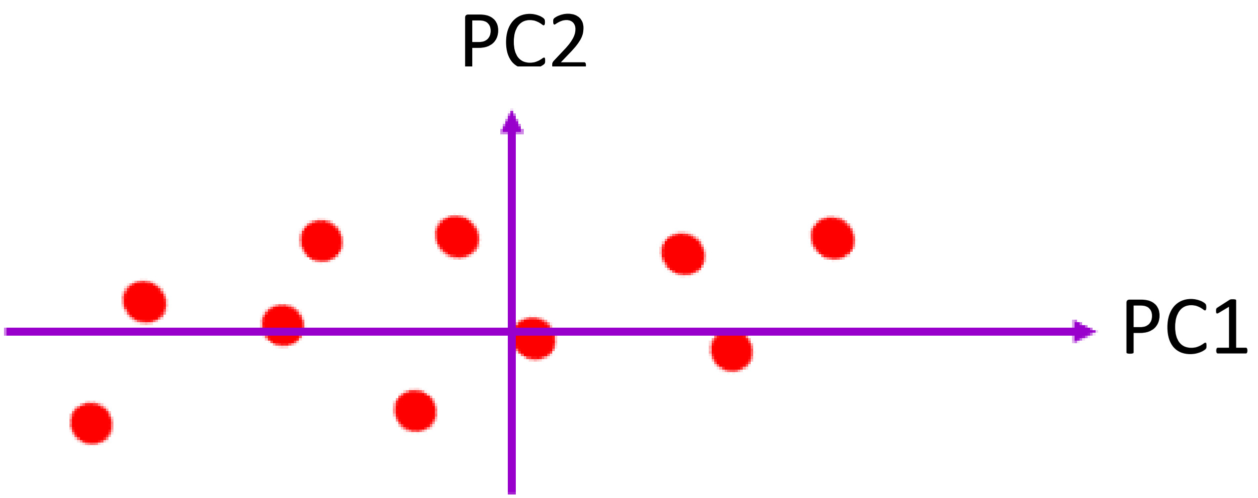
2-Dimension (2D)





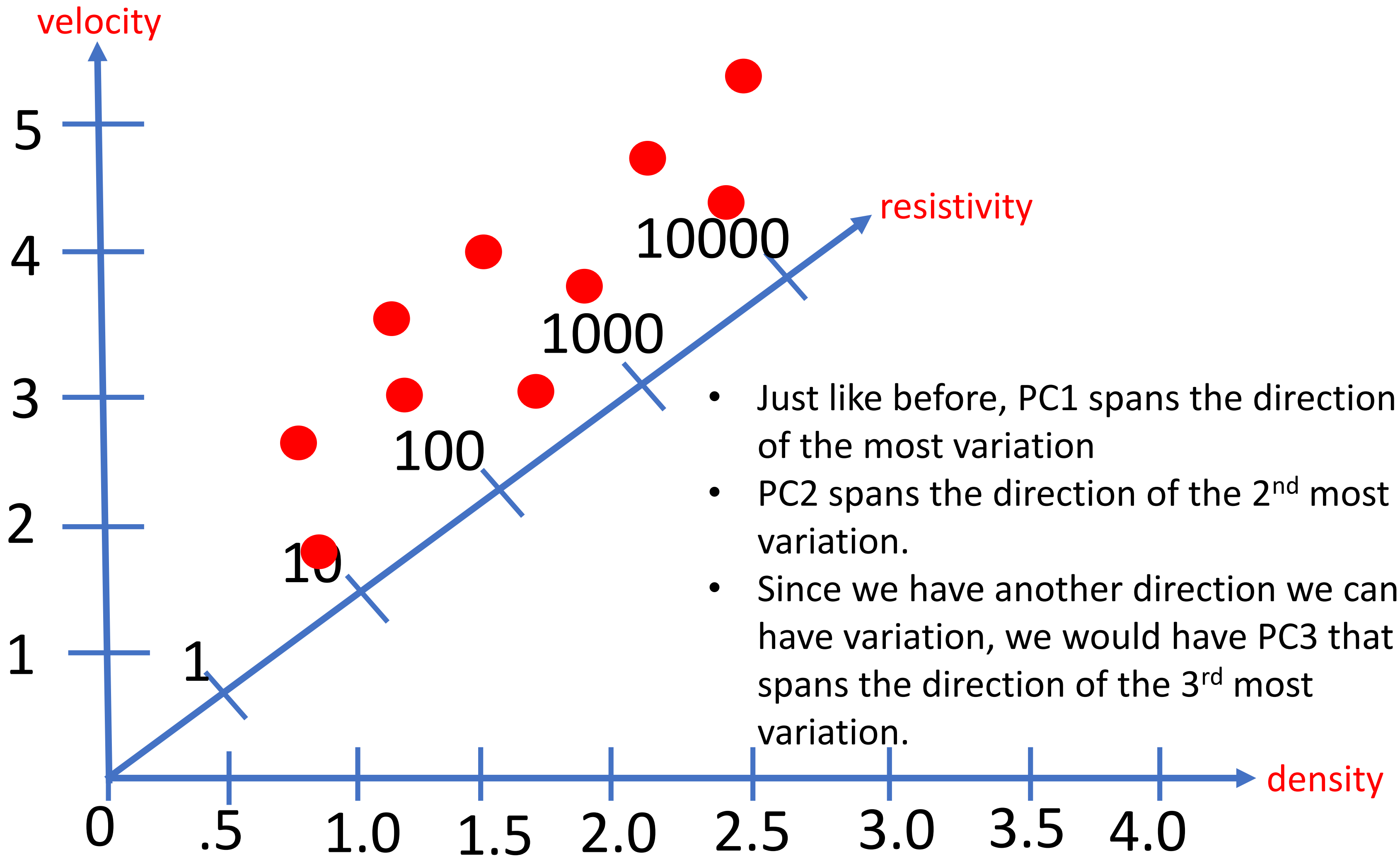
velocity
5
4
3
2
1
0
.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
density





- PC1 (the first component) is the axis that spans the most variation
- PC2 (the second component) is the axis that spans the second most variation

What if we 3-Dimension (3D) data?



- Just like before, PC1 spans the direction of the most variation
- PC2 spans the direction of the 2nd most variation.
- Since we have another direction we can have variation, we would have PC3 that spans the direction of the 3rd most variation.

What if we had 4-D data?

- PC1 would span the direction of the most variation.
- PC2 would span the direction of the 2^{nd} most variation.
- PC3 would span the direction of the 3^{rd} most variation.
- PC4 would span the direction of the 4^{th} most variation.

What if we had 4-D data?

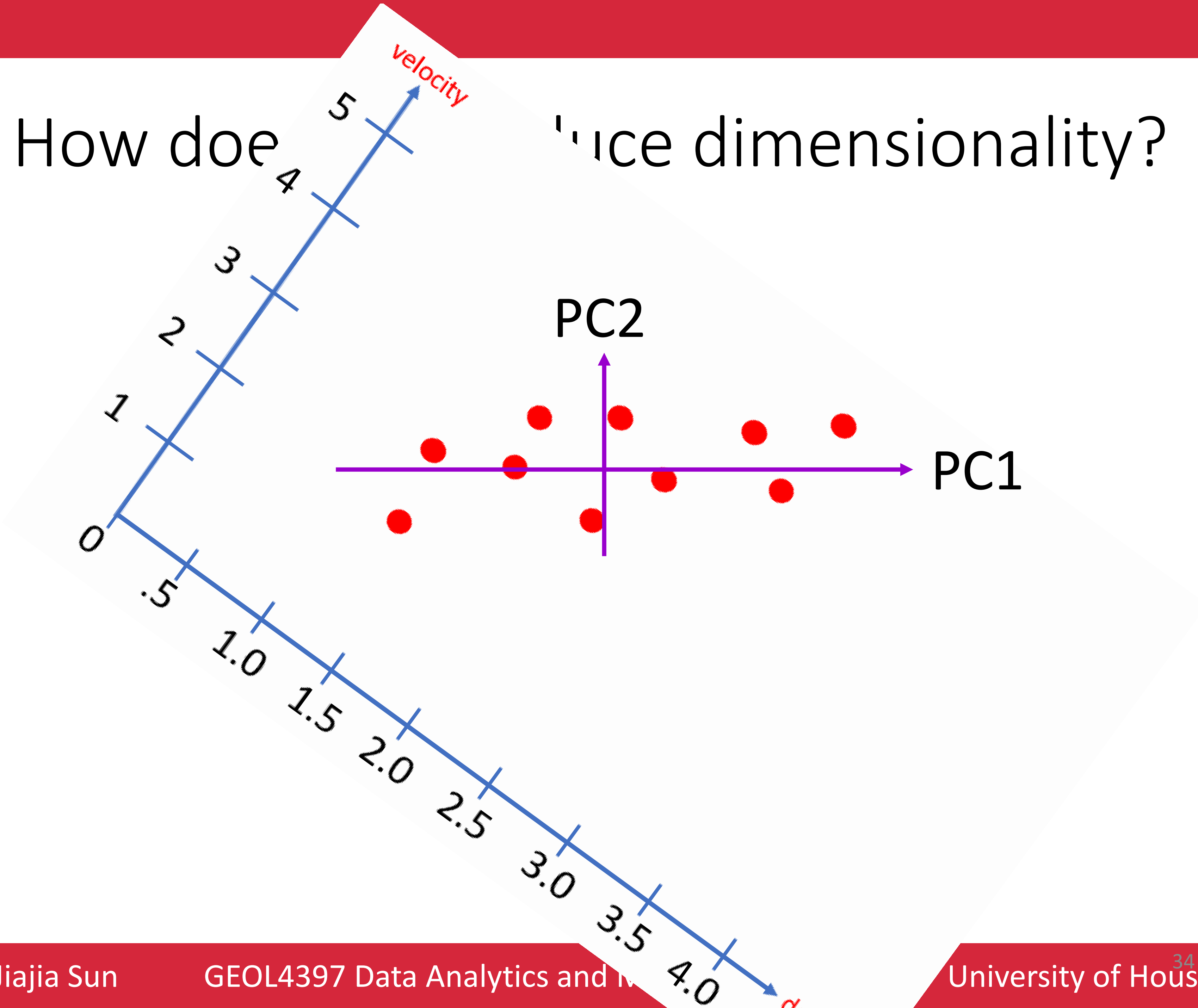
- PC1 would span the direction of the most variation.
- PC2 would span the direction of the 2^{nd} most variation.
- PC3 would span the direction of the 3^{rd} most variation.
- PC4 would span the direction of the 4^{th} most variation.

In general,

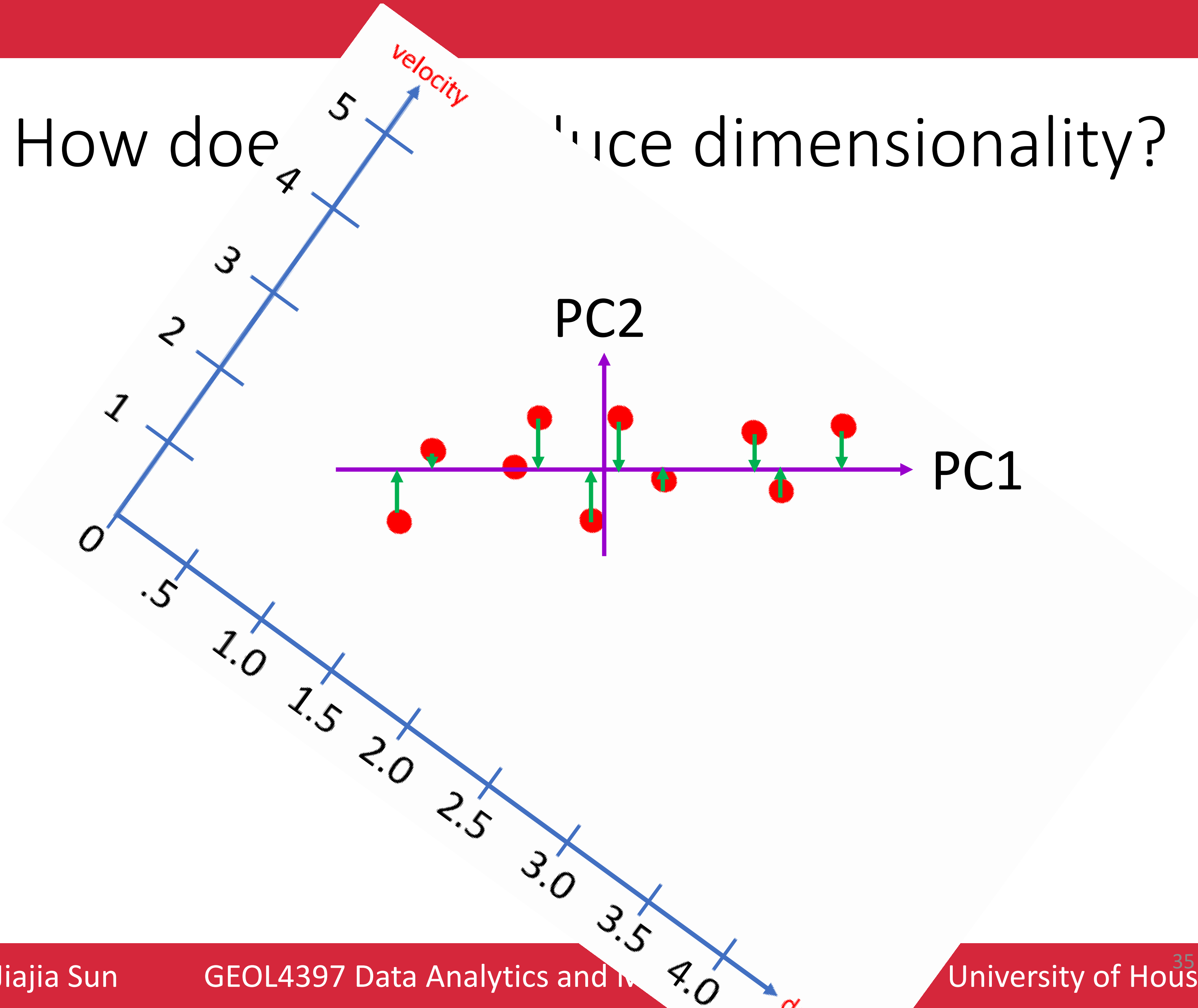
- There is a PC for each physical property measurement.
- If we had 22 measurements on each cutting sample, we would have 22 PCs.
- PC22 would span the direction of the 22^{nd} most variation.

How does PCA reduce dimensionality?

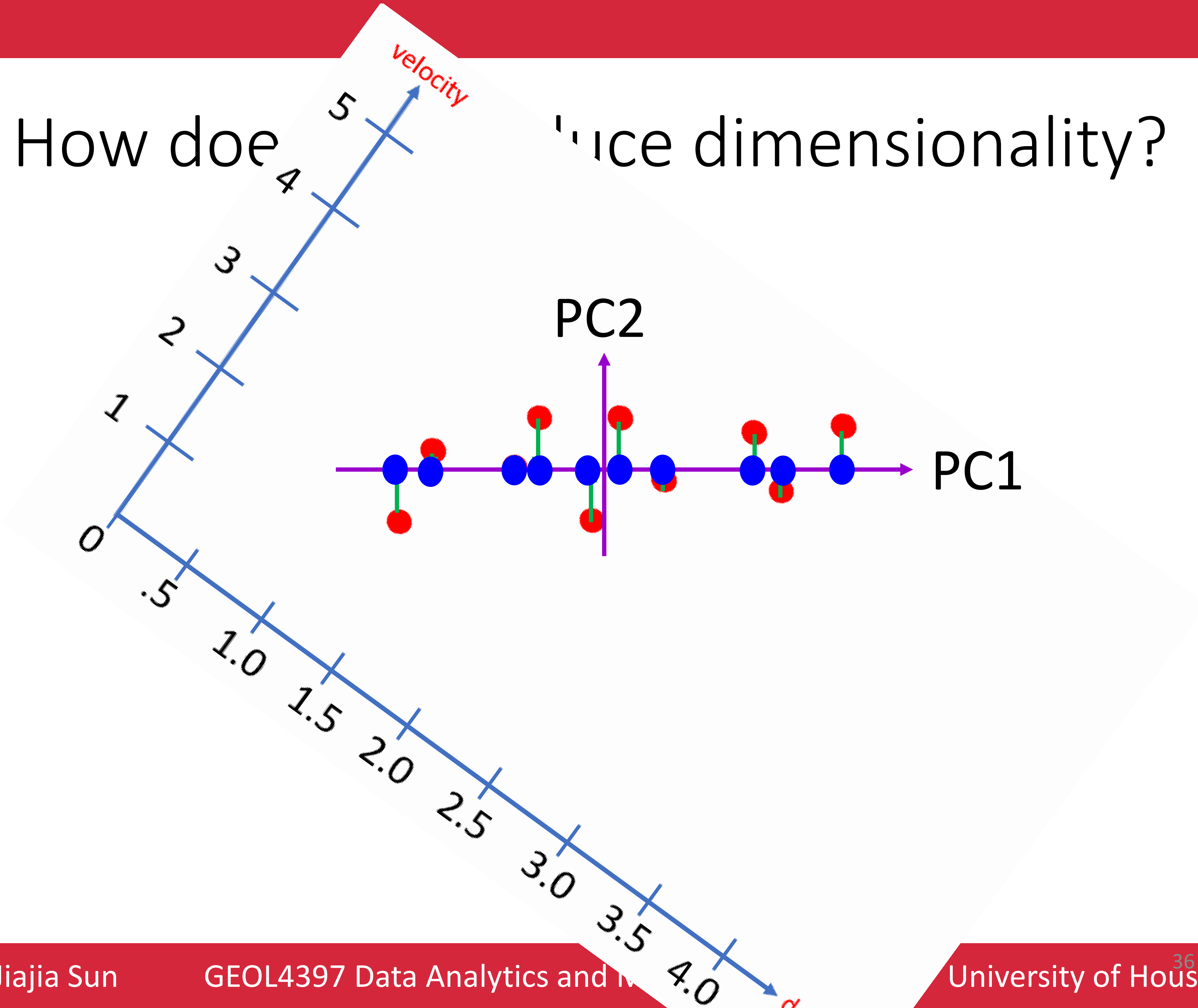
How doe… …uce dimensionality?



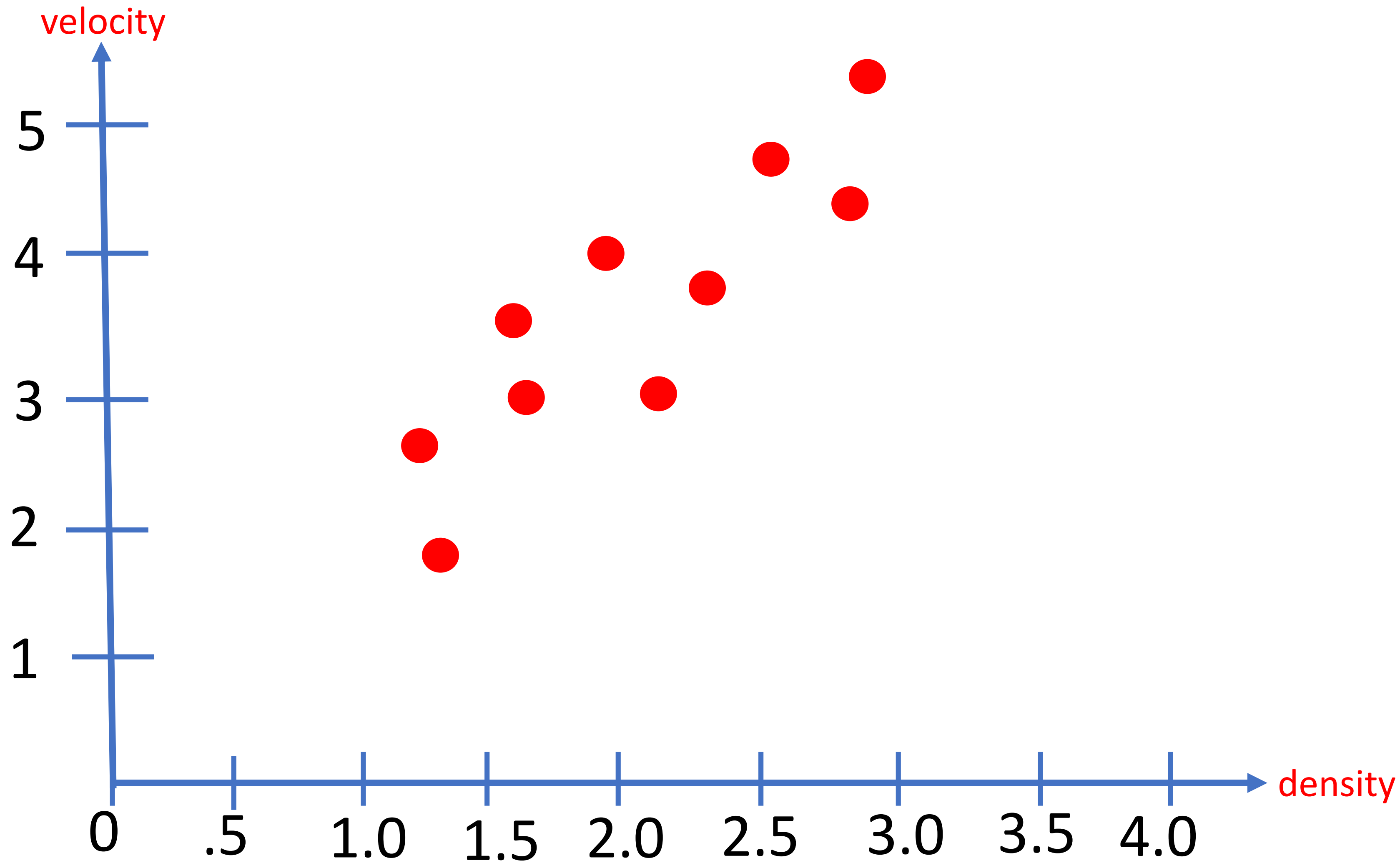
How doe … uce dimensionality?



How doe ... 'uce dimensionality?



Dimensionality reduction



MNIST dataset: PCA

- Original data in 784 dimensional space

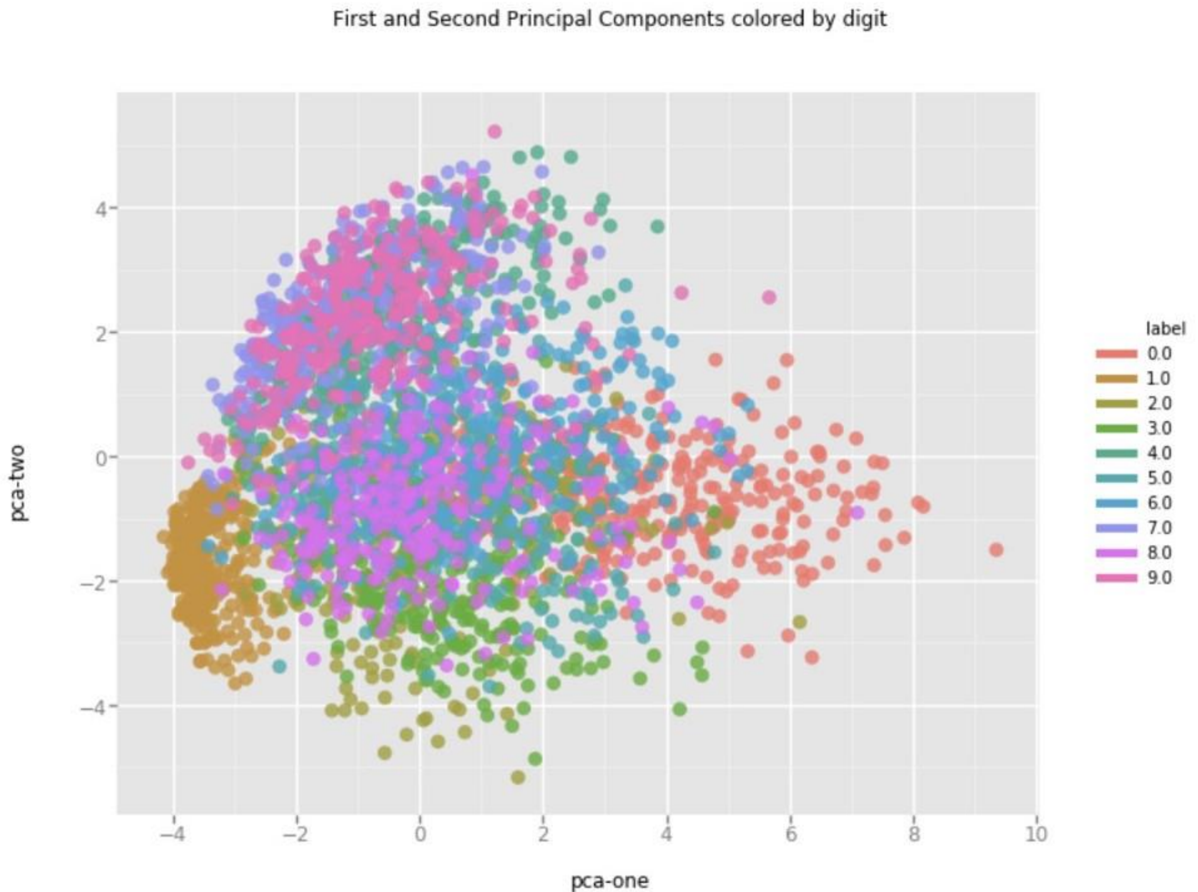


Image source: https://goo.gl/Dj97T7

MNIST dataset: t-SNE

- Original data in 784 dimensional space

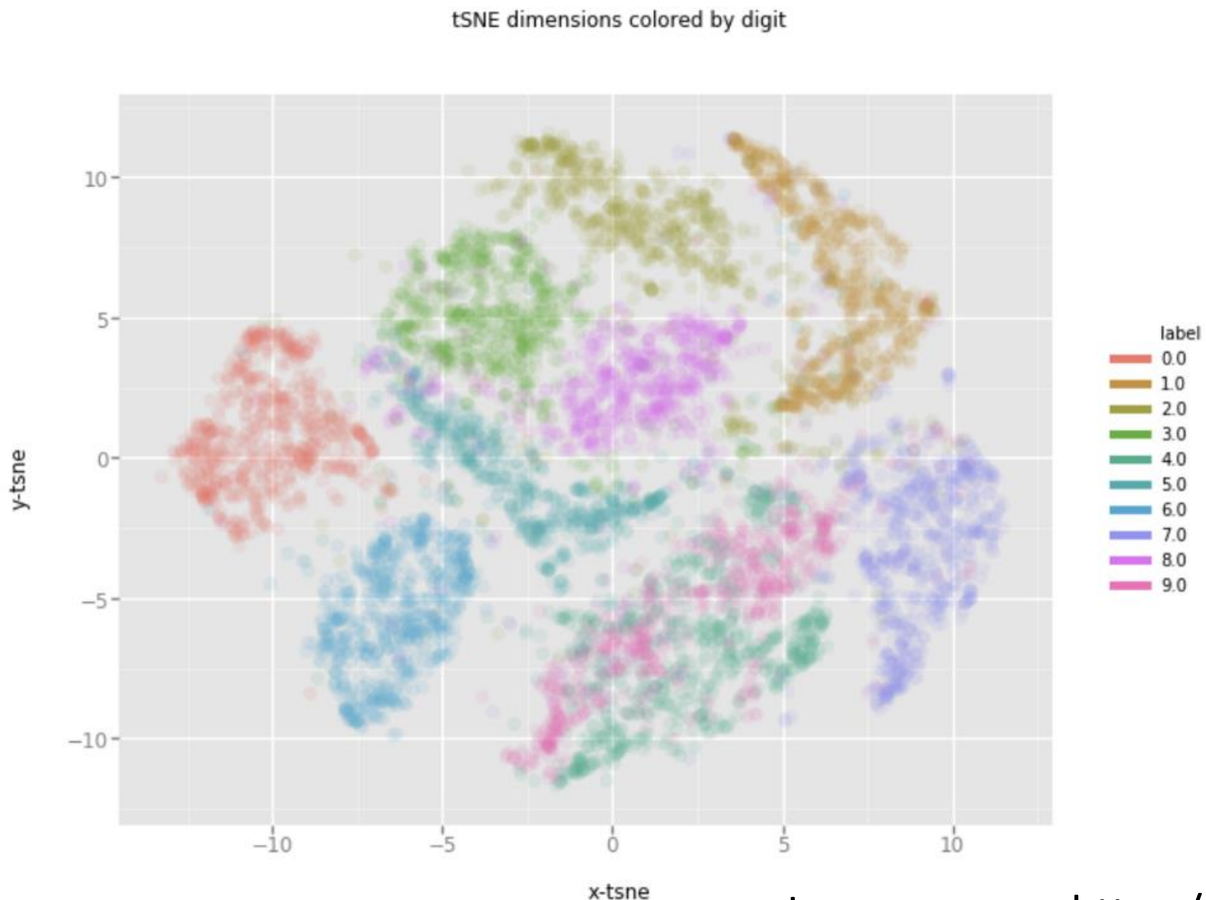
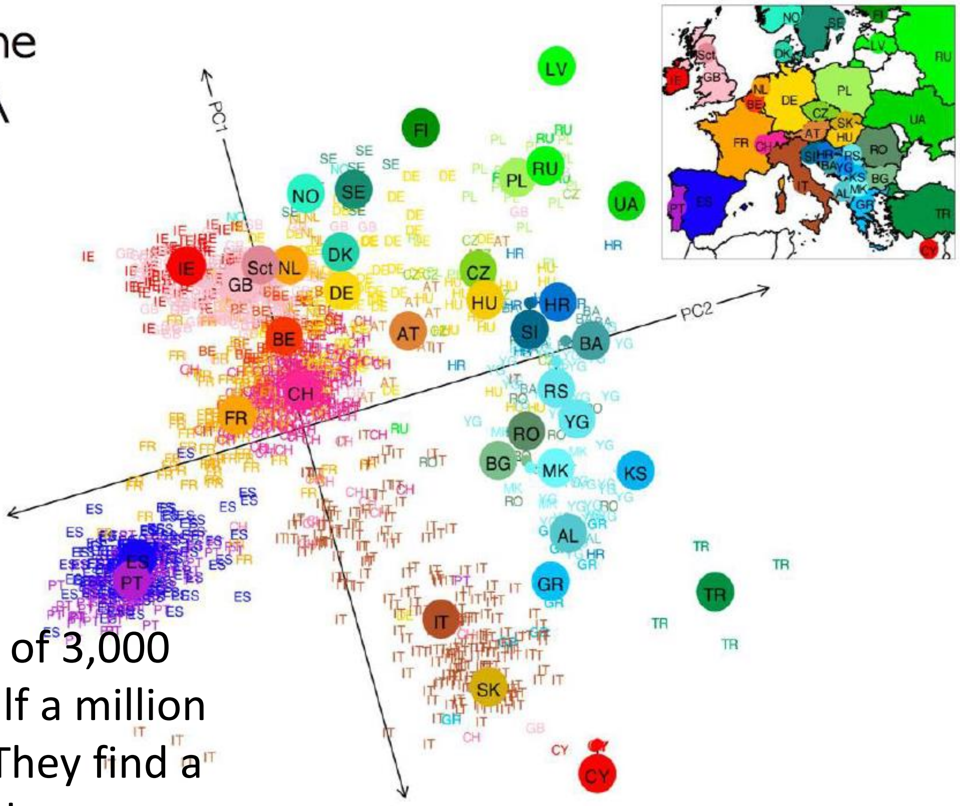


Image source: https://goo.gl/Dj97T7

Example: PCA for high-dimensional data

500,000 DNA sites in human genome projected to 2 dimensions with PCA

Principal components correspond to geography → ancestry



Novembre et al. (2008), Nature

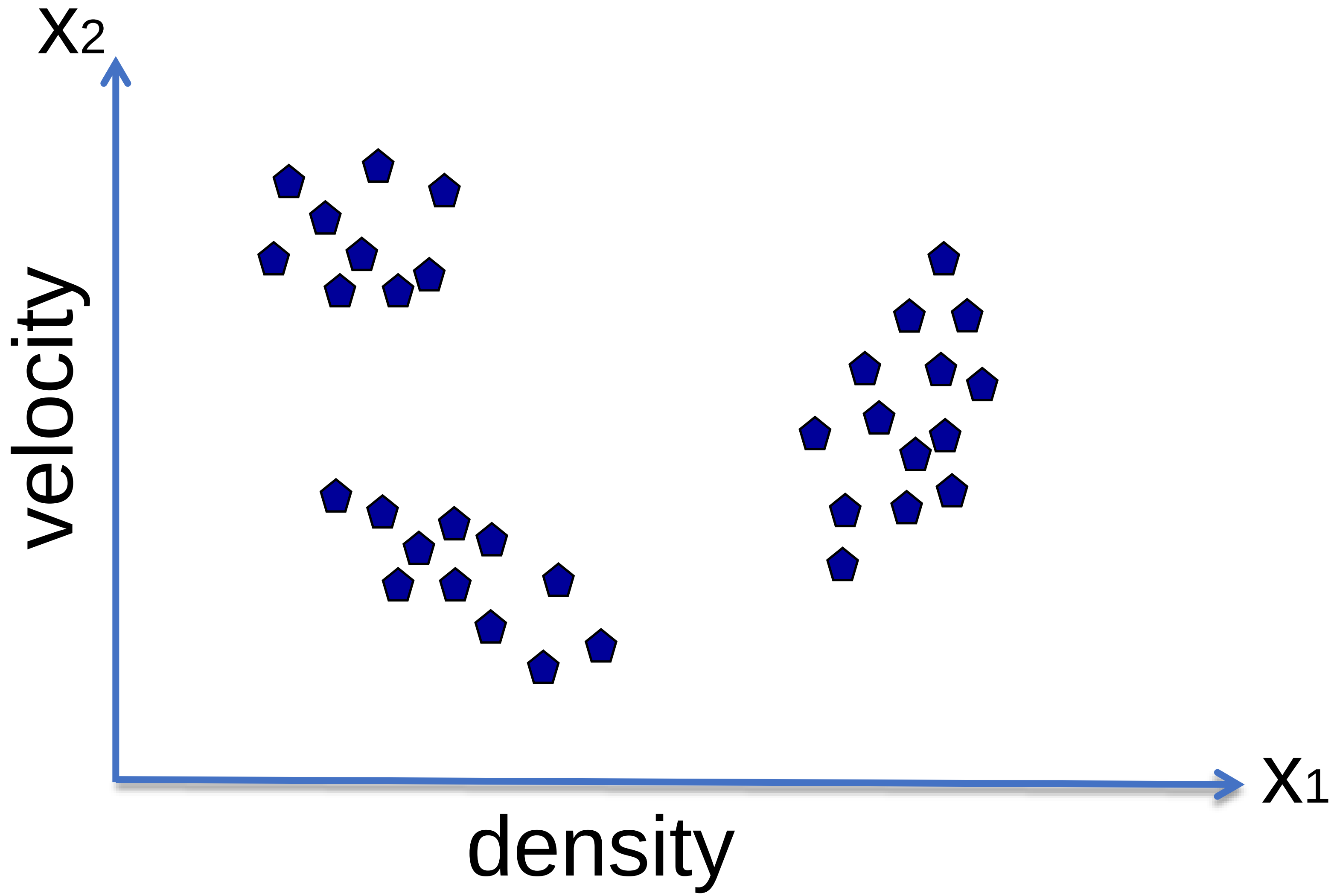
Characterize genetic variations in a sample of 3,000 European individuals genotyped at over half a million variable DNA sites in the human genome. They find a close correspondence between genetic and geographic distances. → genetic ancestry testing an individual's DNA can be used to infer their geographic origin with surprising accuracy—often to within a few hundred kilometers.

Thanks to Karianne Bergen

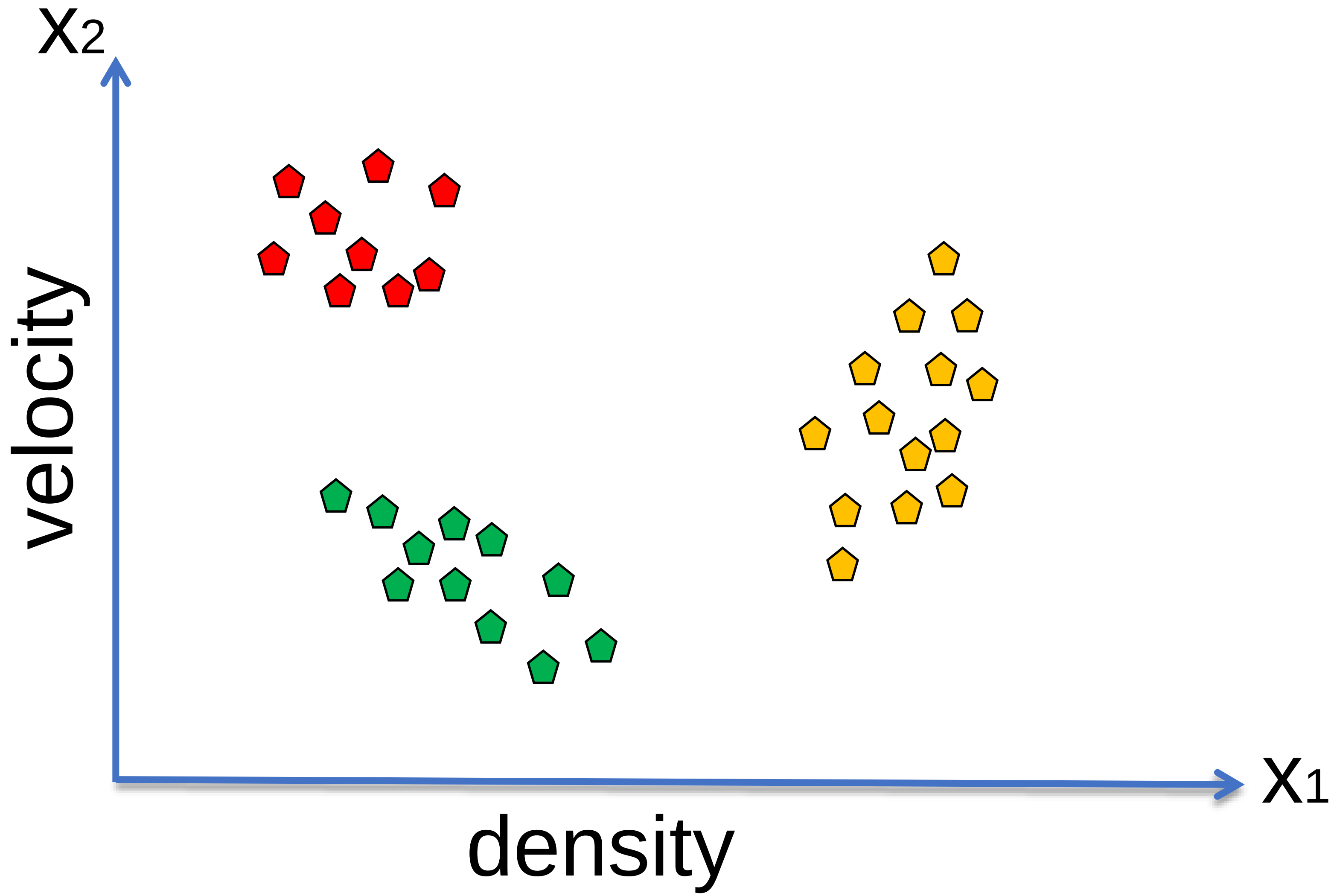
Outline

- Dimensionality reduction
- K-means Clustering
- Implementation in Scikit-Learn

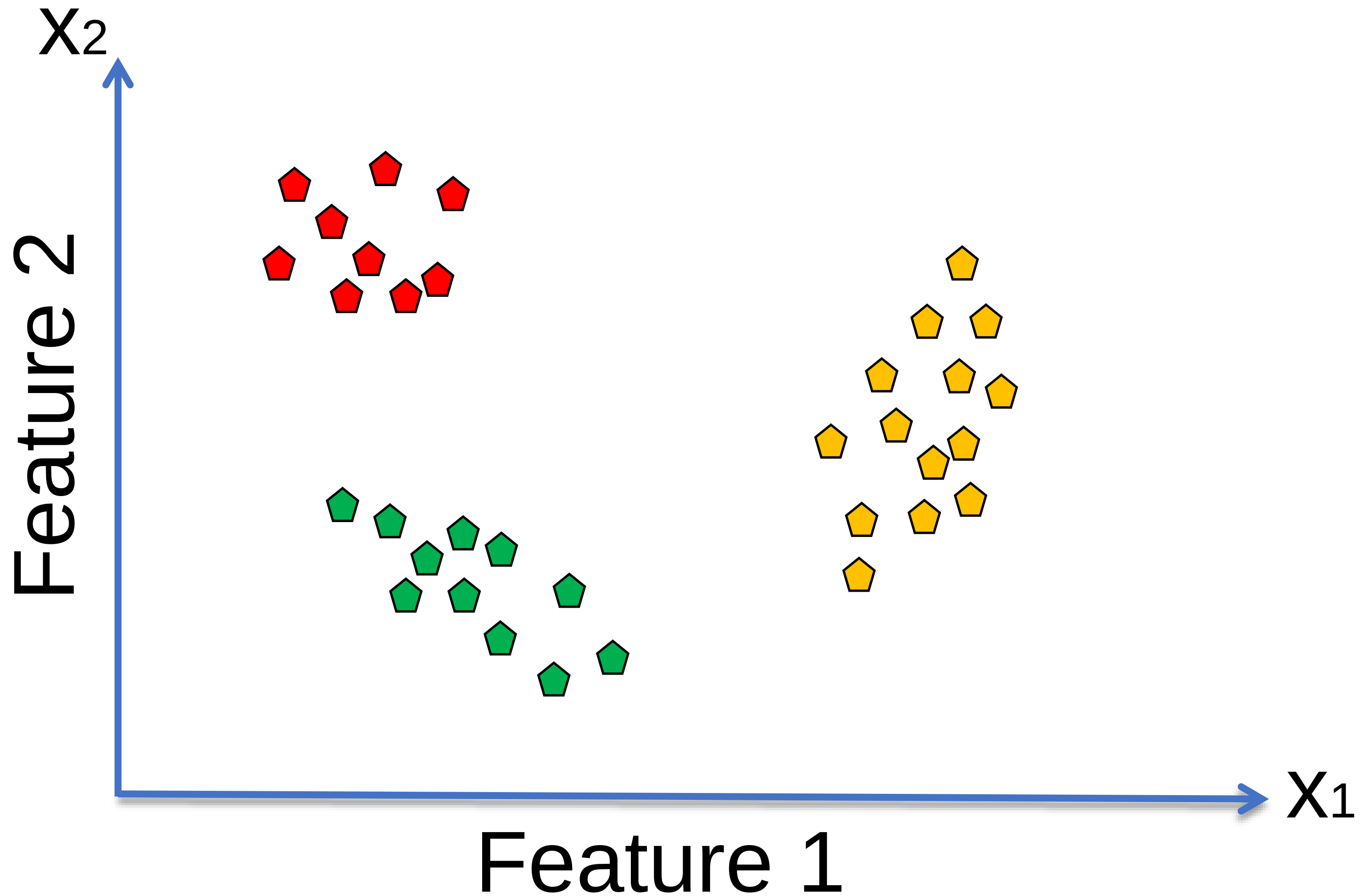
Clustering



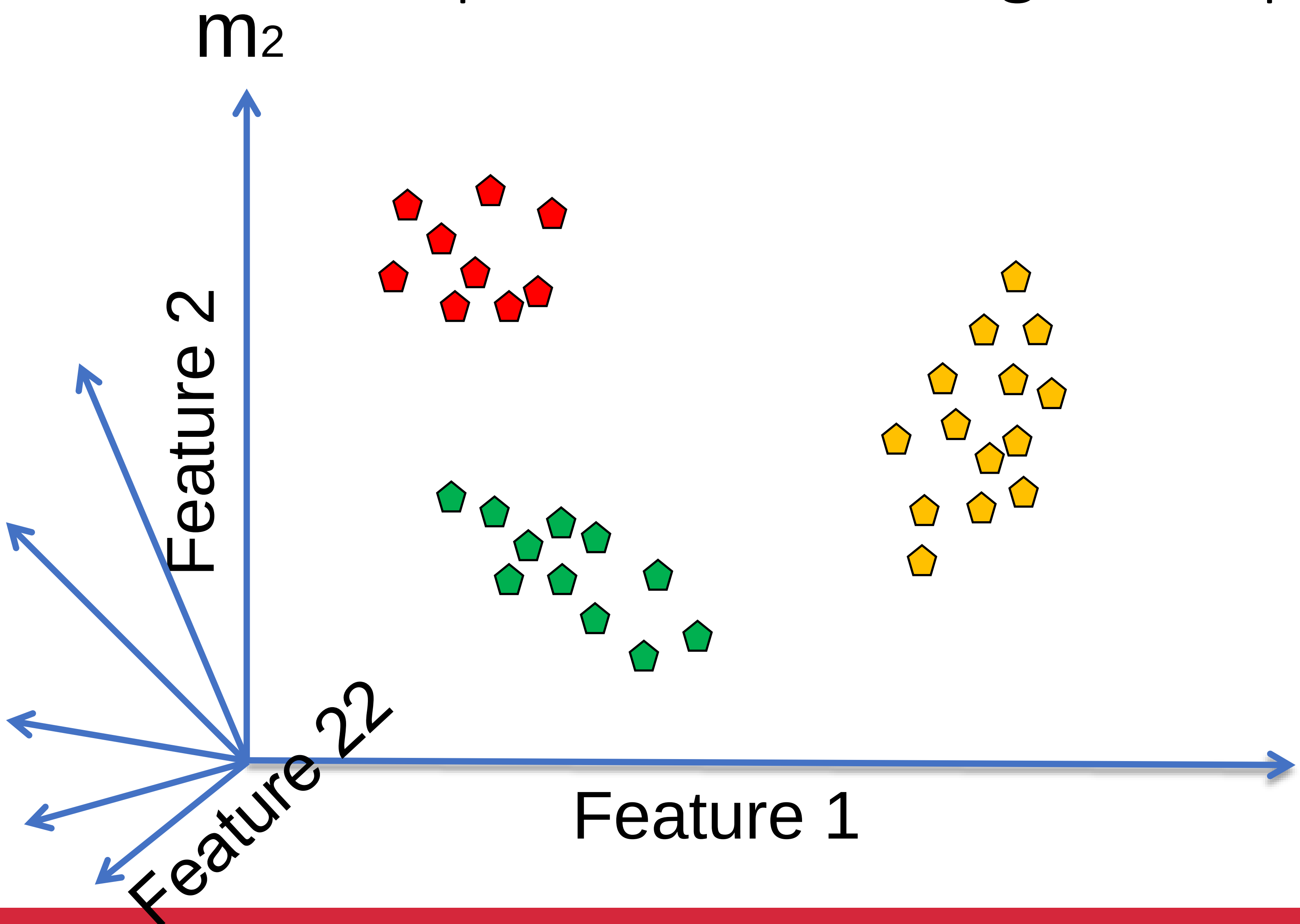
Clustering



Clustering



Unsupervised learning: example



https://news.google.com/?hl=en-US&gl=US&ceid=US:en

Google News

Search for topics, locations & sources

- Top stories
- For you
- Favorites
- Saved searches
- U.S.
- World
- Local
- Business
- Technology
- Entertainment
- Sports
- Science
- Health

Language & region
English | United States

Headlines

More Headlines

Southwest Boeing 737 Max makes emergency landing in Orlando; FAA cites engine issue unrelated to recent crashes

CNBC • 1 hour ago

- NYT: Pilots had 40 seconds to avert disaster in test of Boeing 737 Max plane
  CNN • 4 hours ago
- Boeing CEO says company is 'humbled and learning' from deadly Ethiopian Airlines crash
  CNBC • 2 hours ago
- Boeing is handling the 737 Max crisis all wrong
  Quartz • 3 hours ago • Opinion
- After Boeing Crashes, Sharp Questions About Industry Regulating Itself
  The New York Times • 3 hours ago

View full coverage

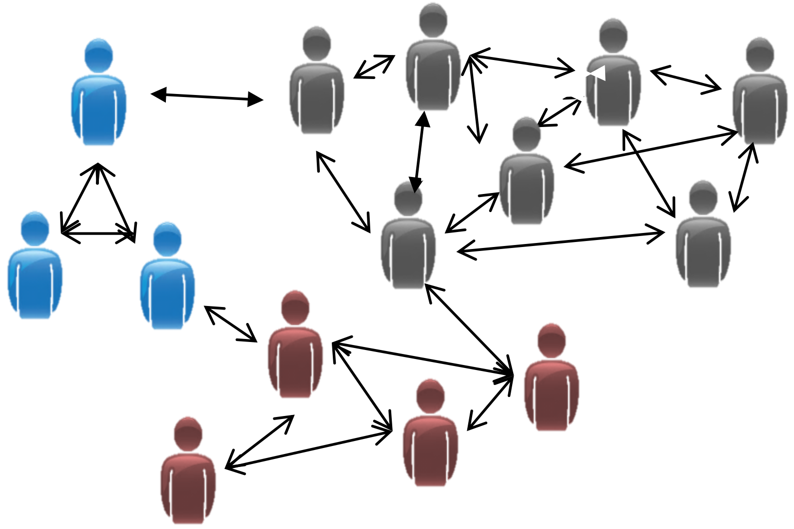
Howard Kurtz: Media tried to convict President Trump, Mueller's findings show they were wrong

Fox News • 1 hour ago

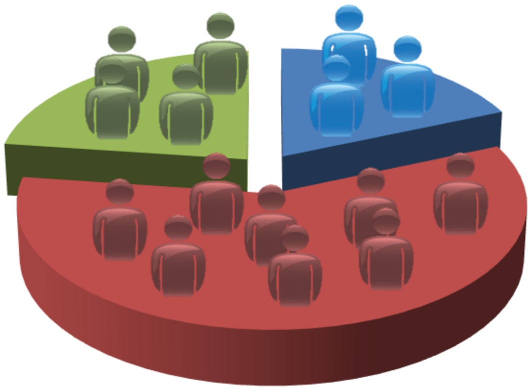
- Trump hands Democrats a gift with new effort to kill Obamacare
  POLITICO • 5 hours ago

View full coverage

Unsupervised learning: applications

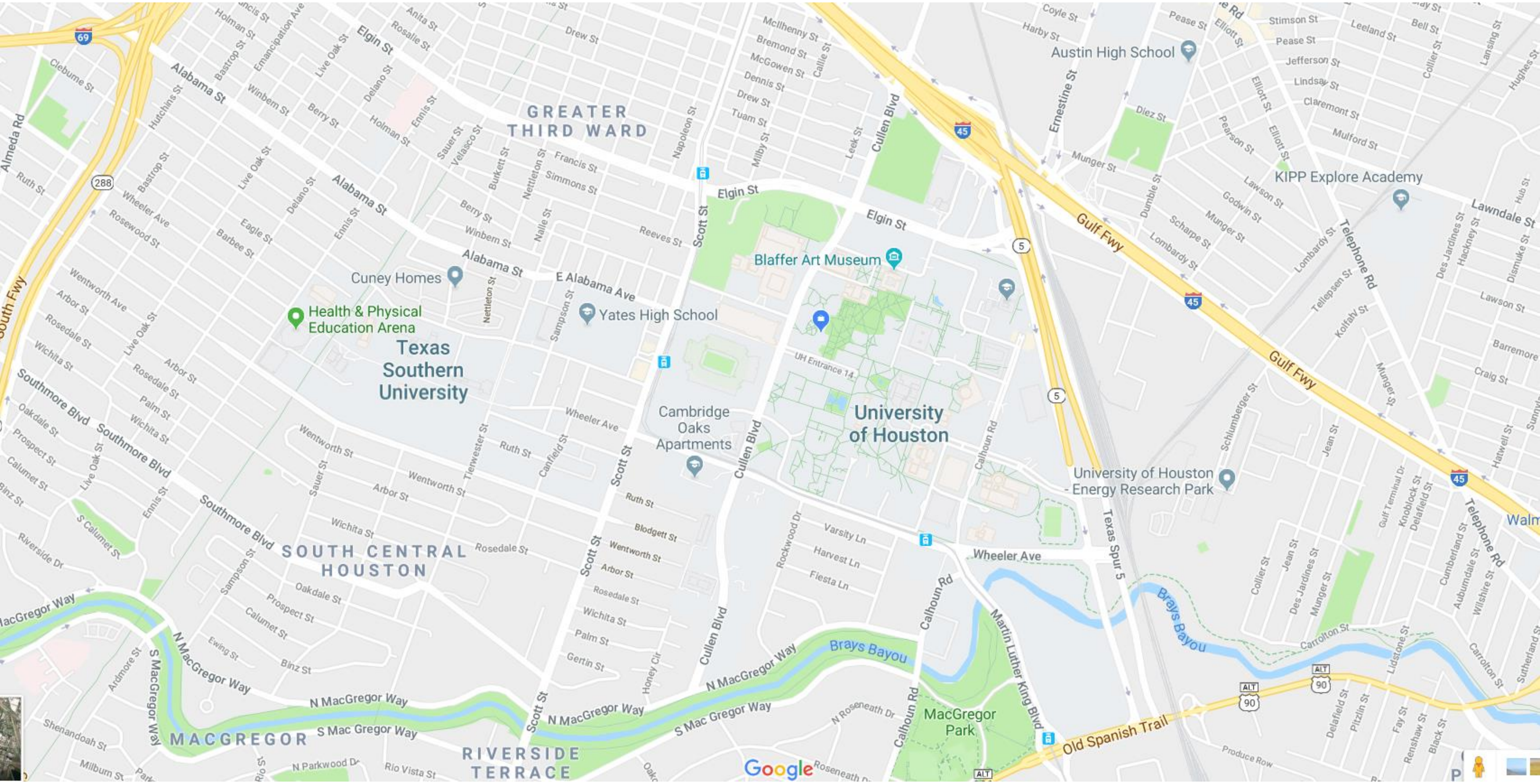


Social network analysis

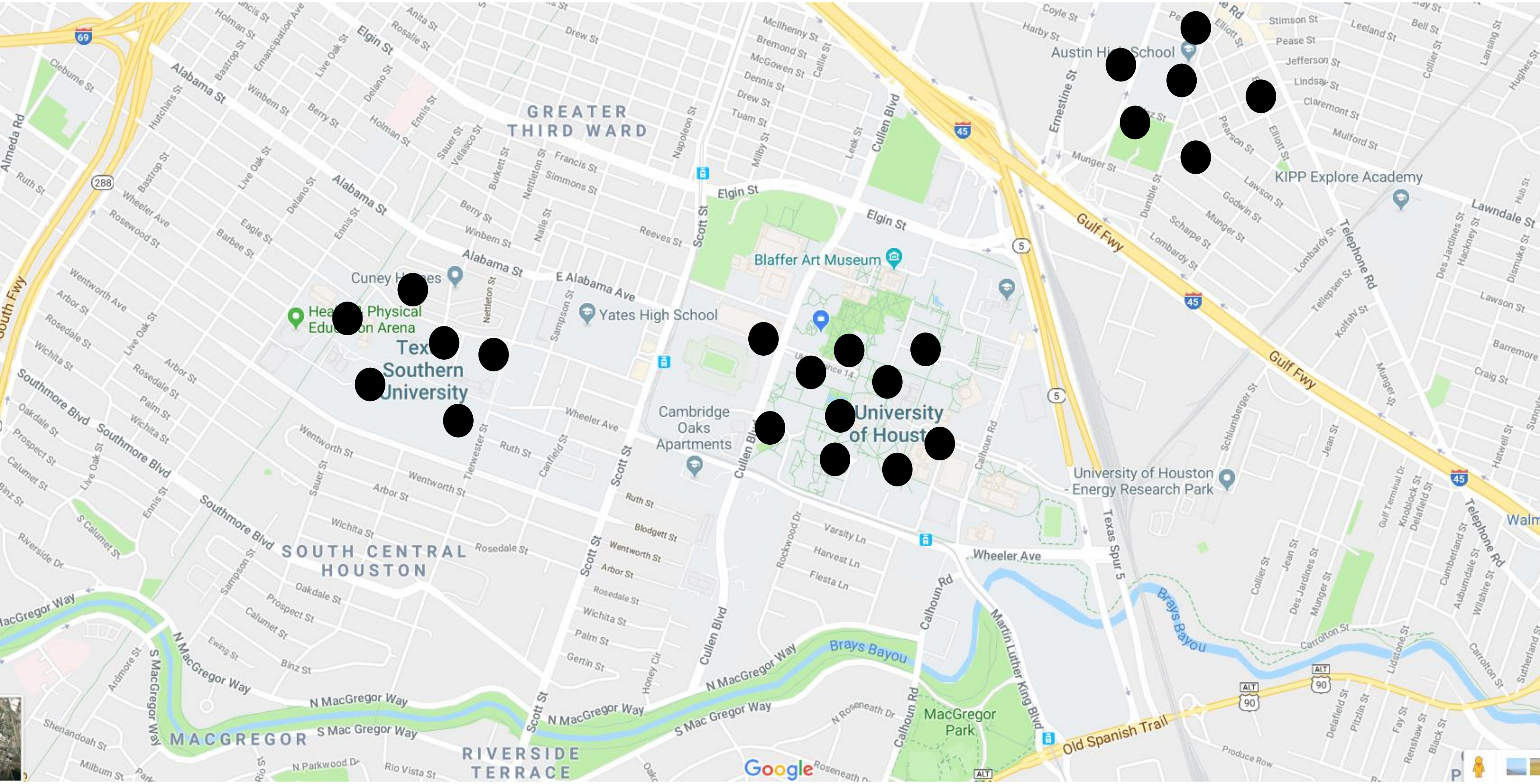


Market segmentation

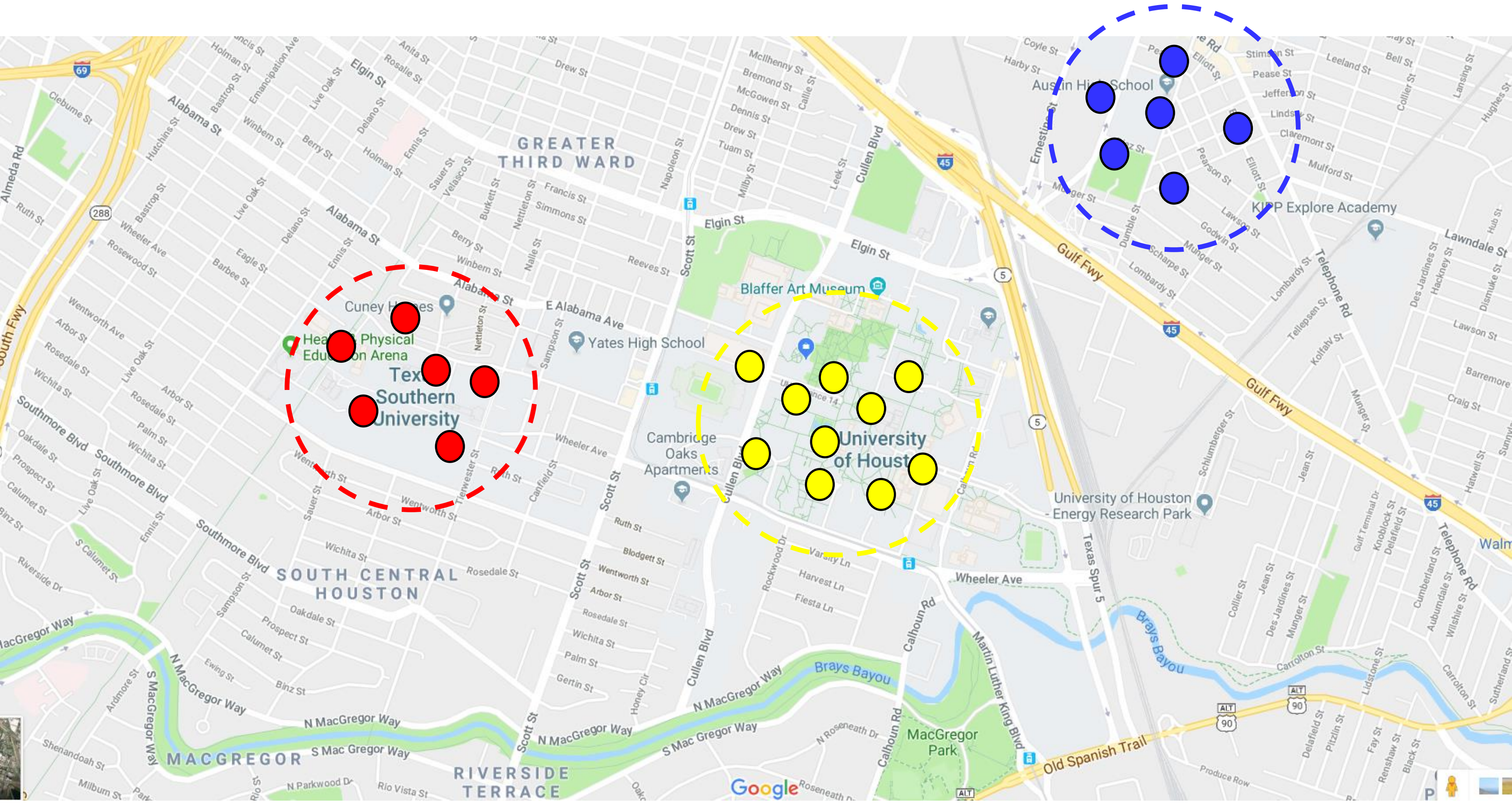
Investing in pizza stores in Houston



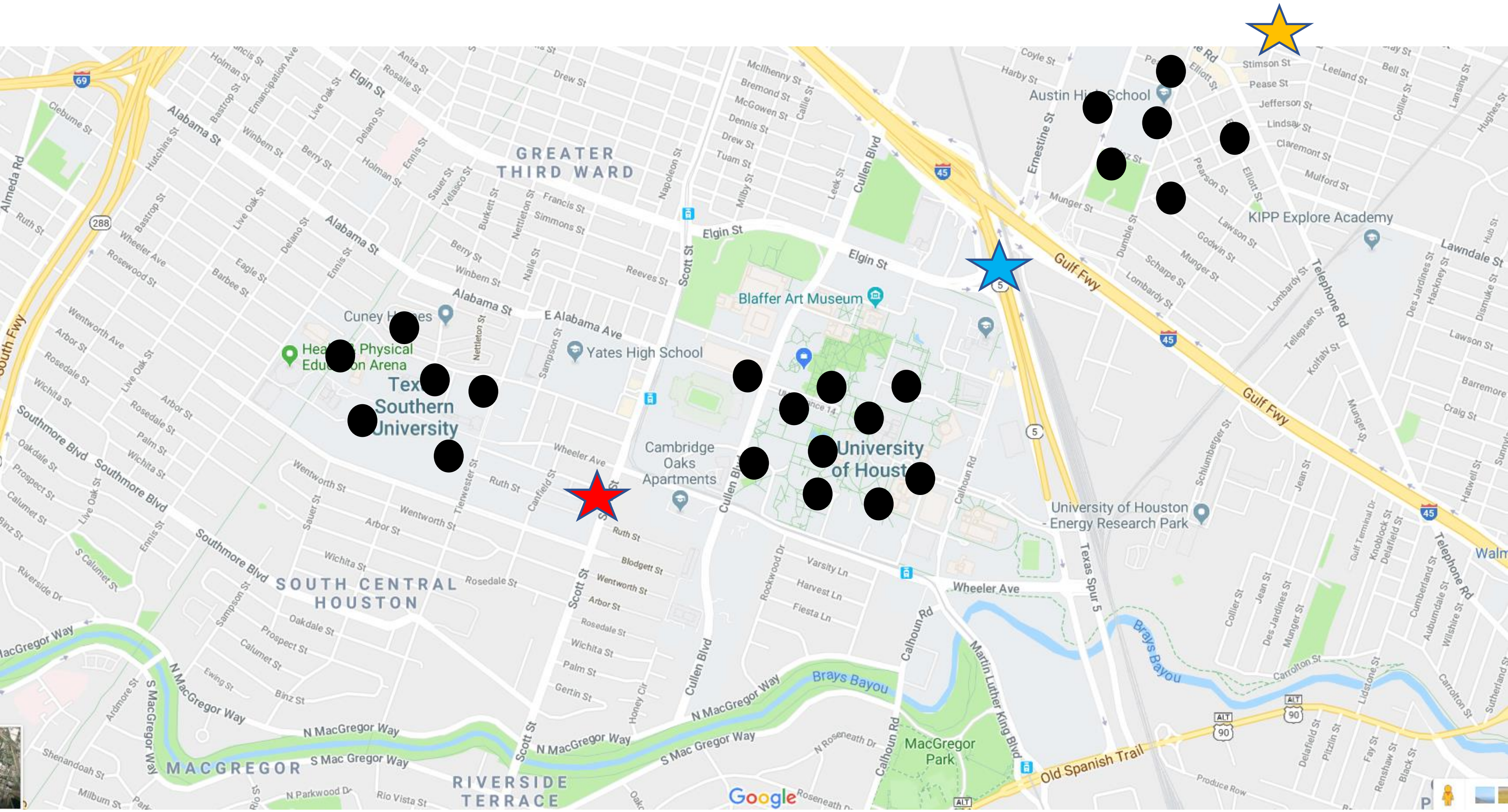
Customer data



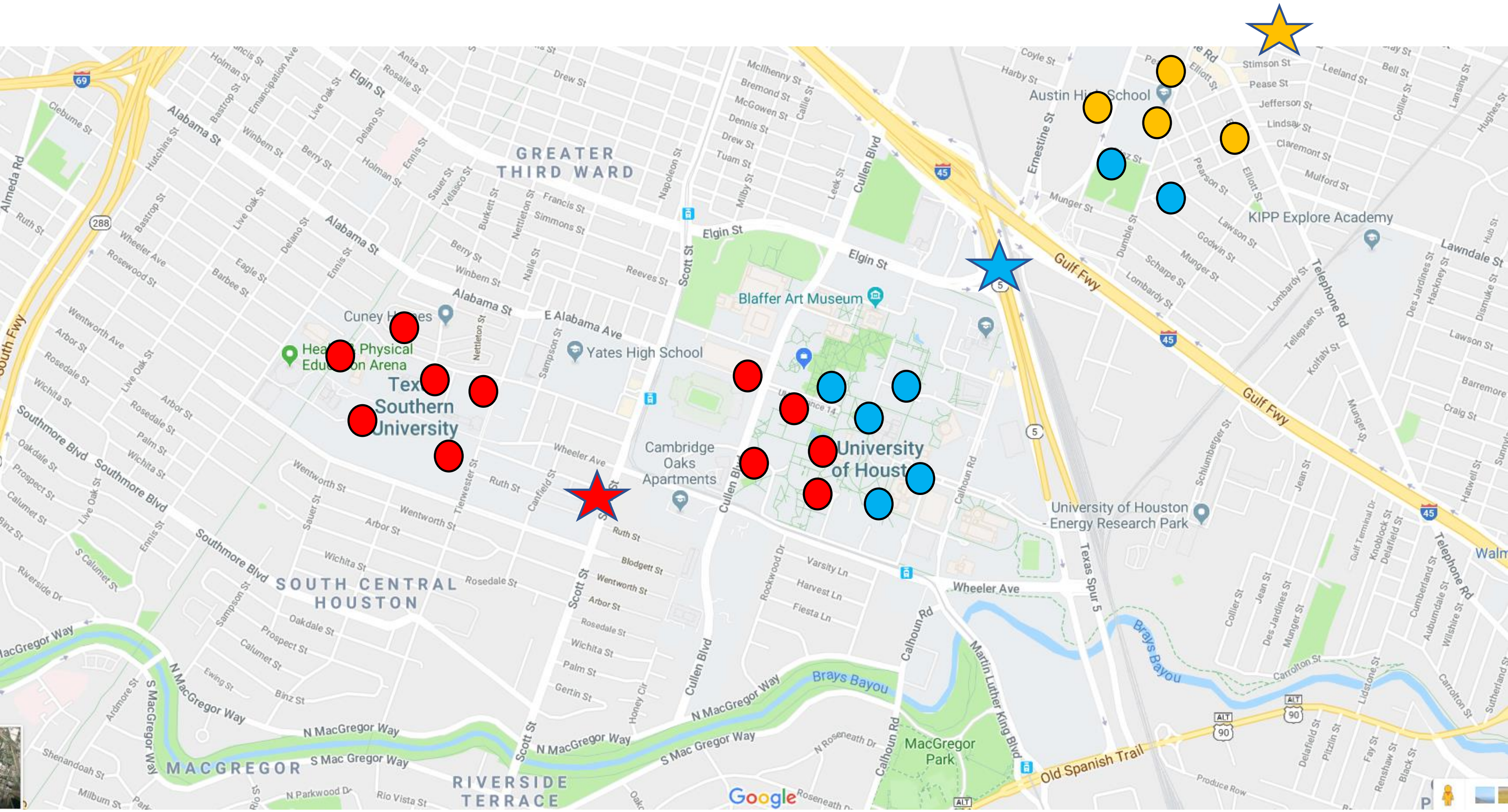
Human wisdom



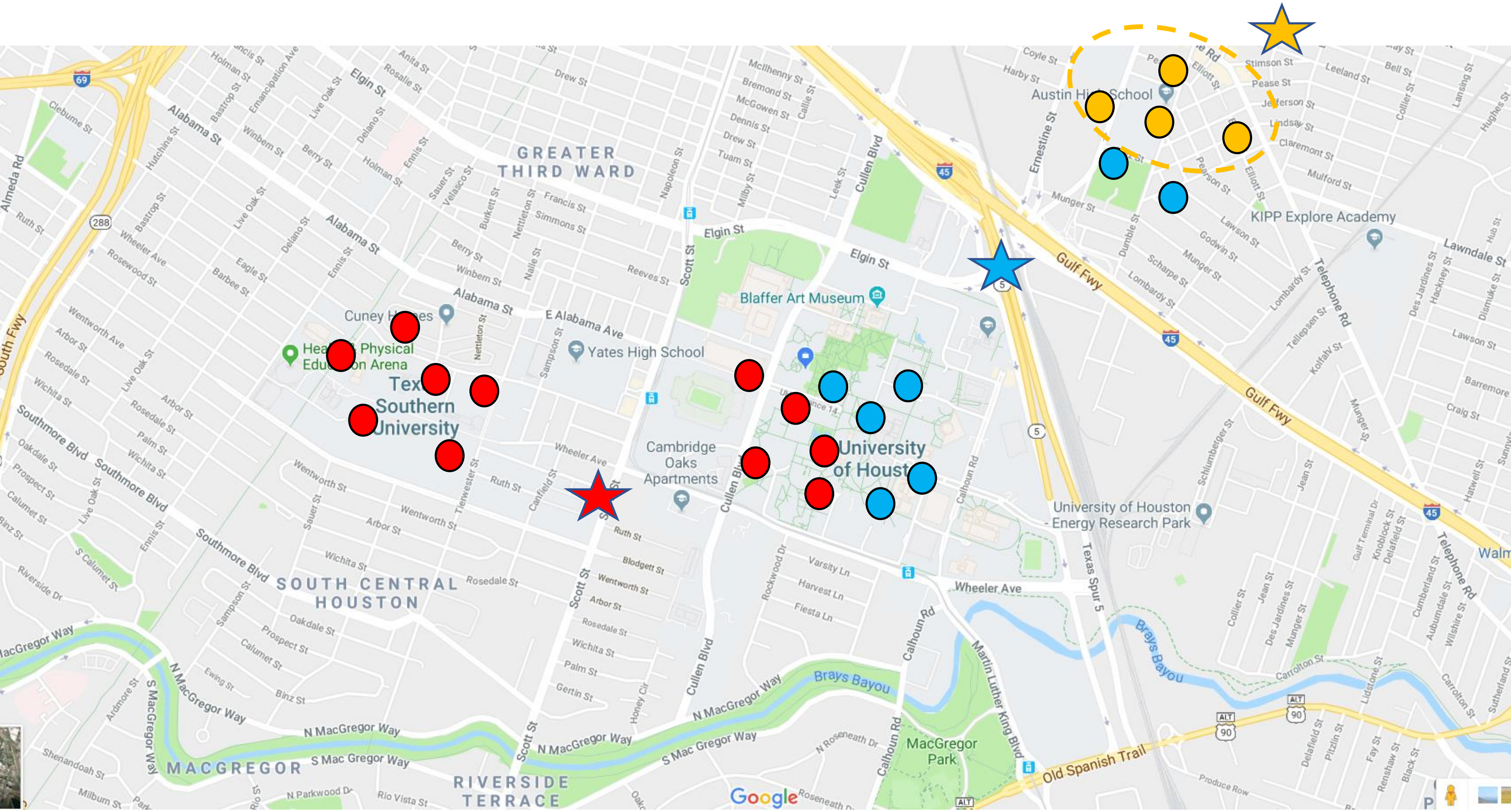
Start with random locations



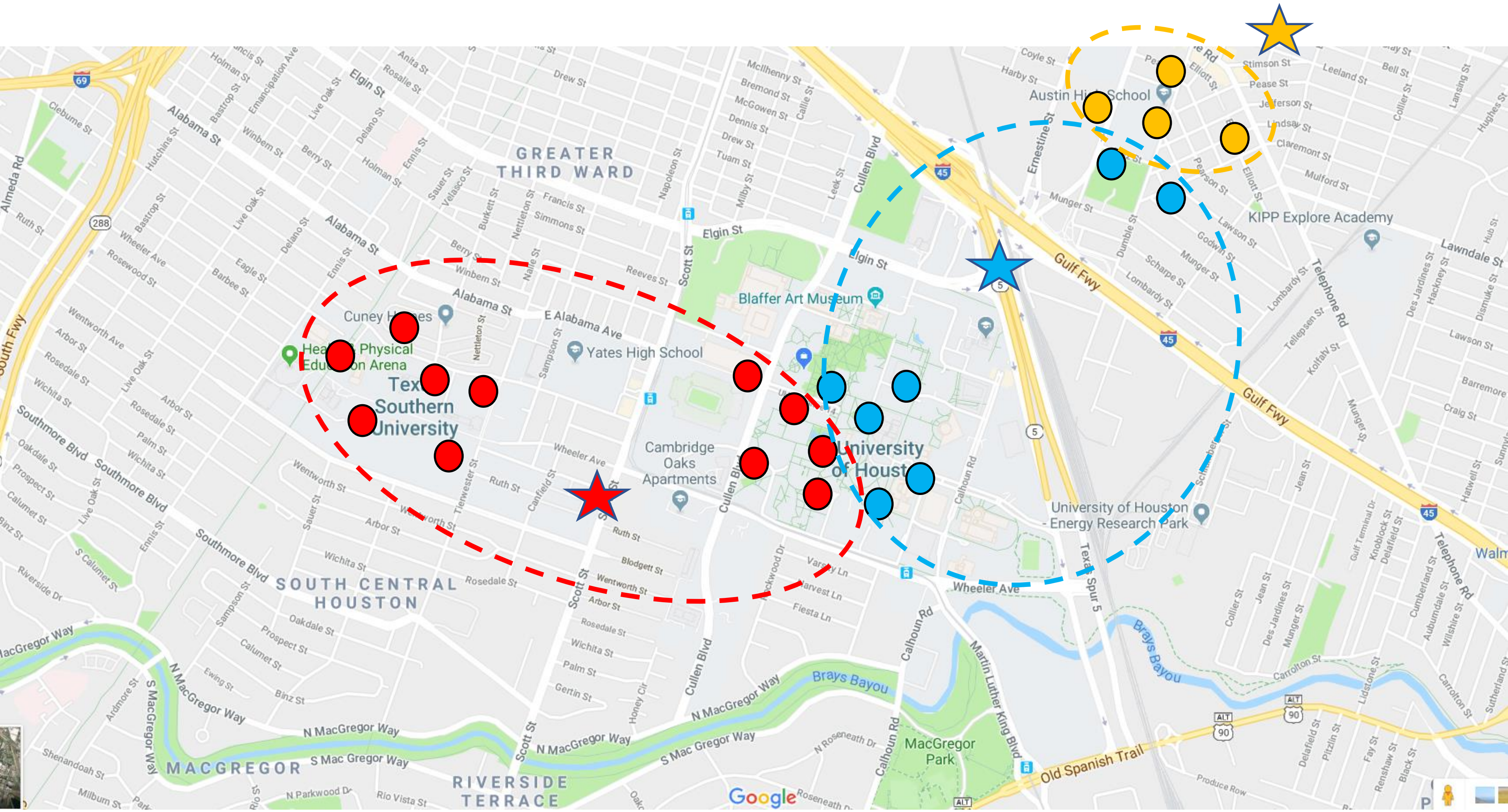
Distribution of customers



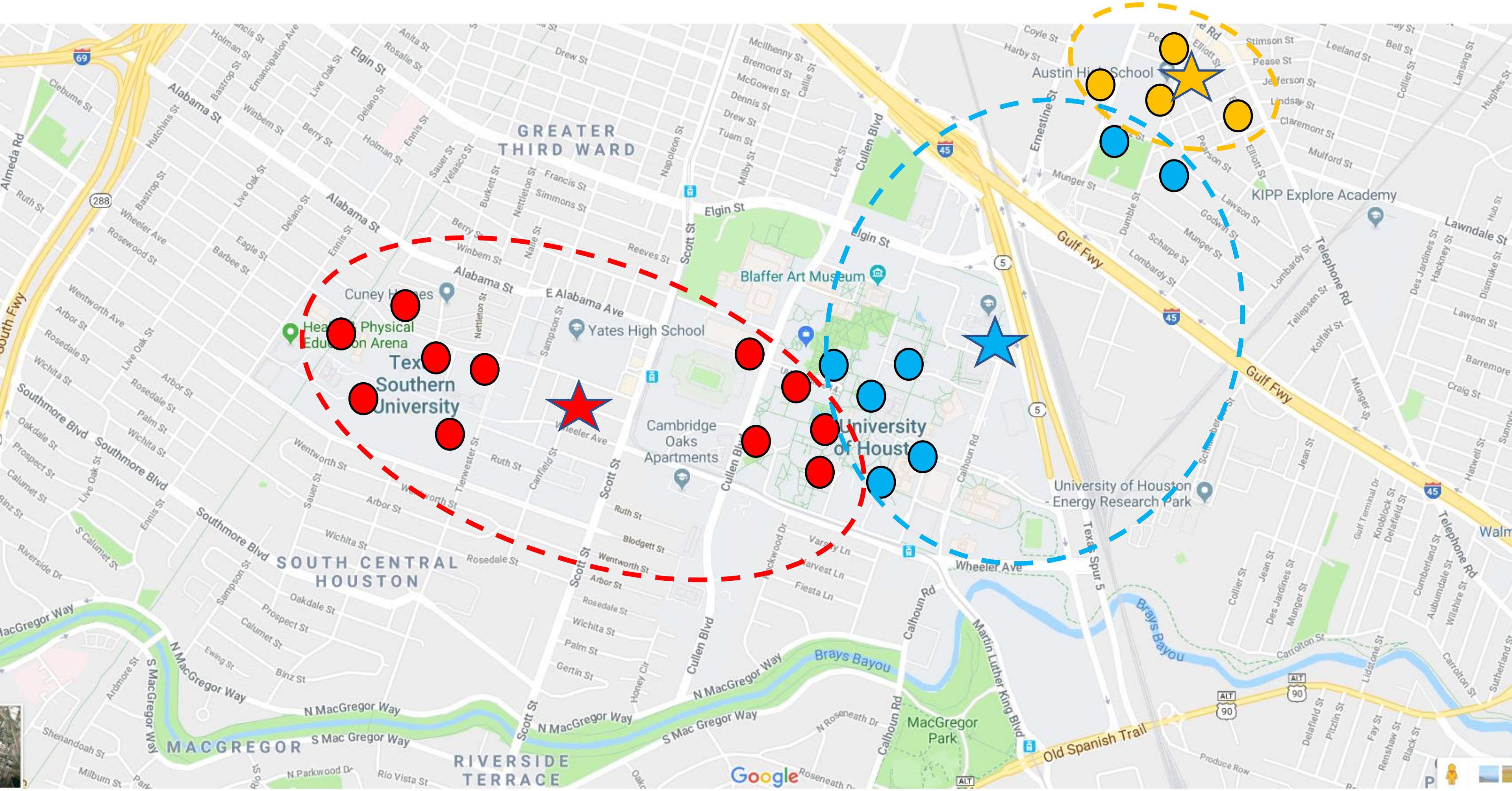
Distribution of customers



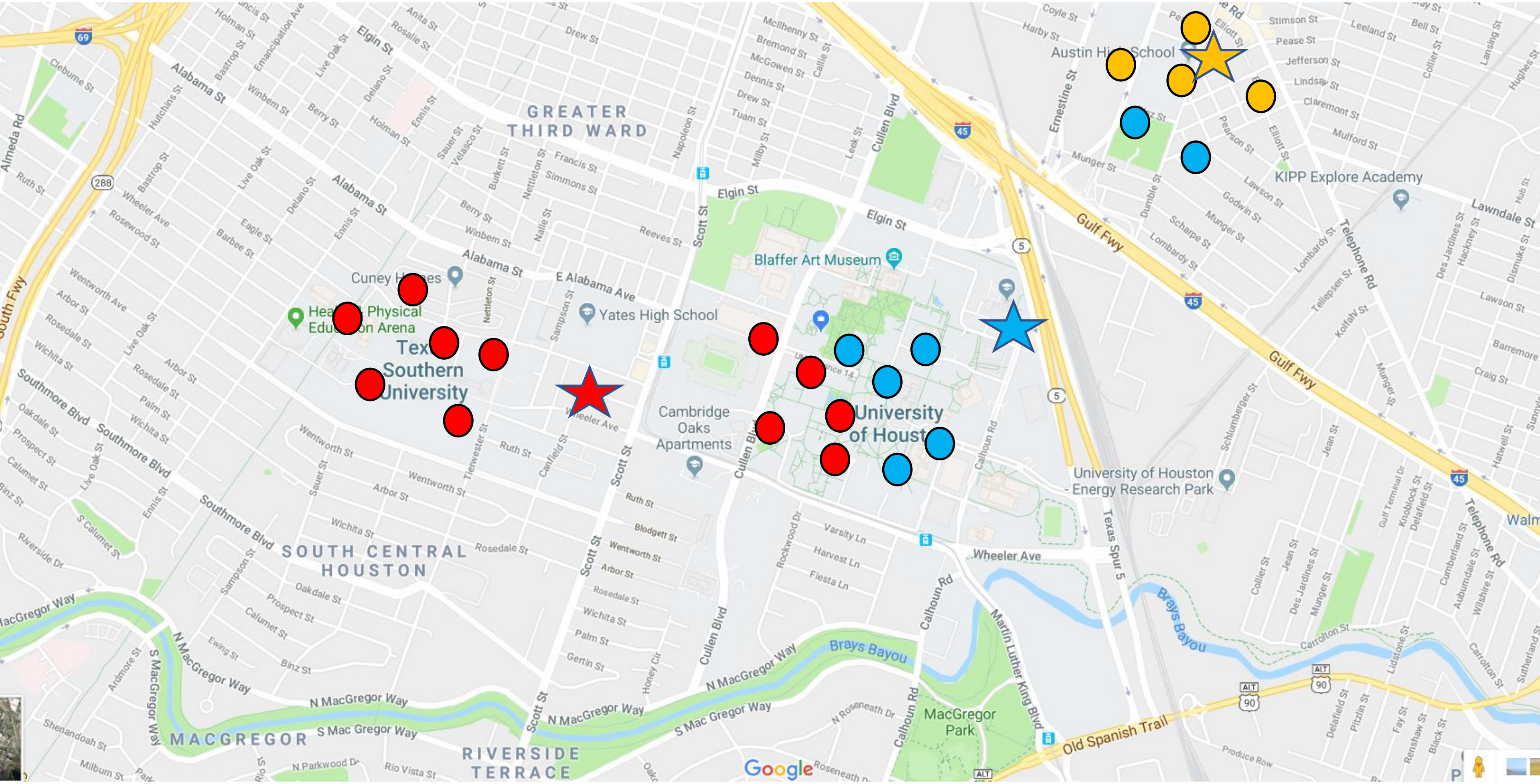
Distribution of customers



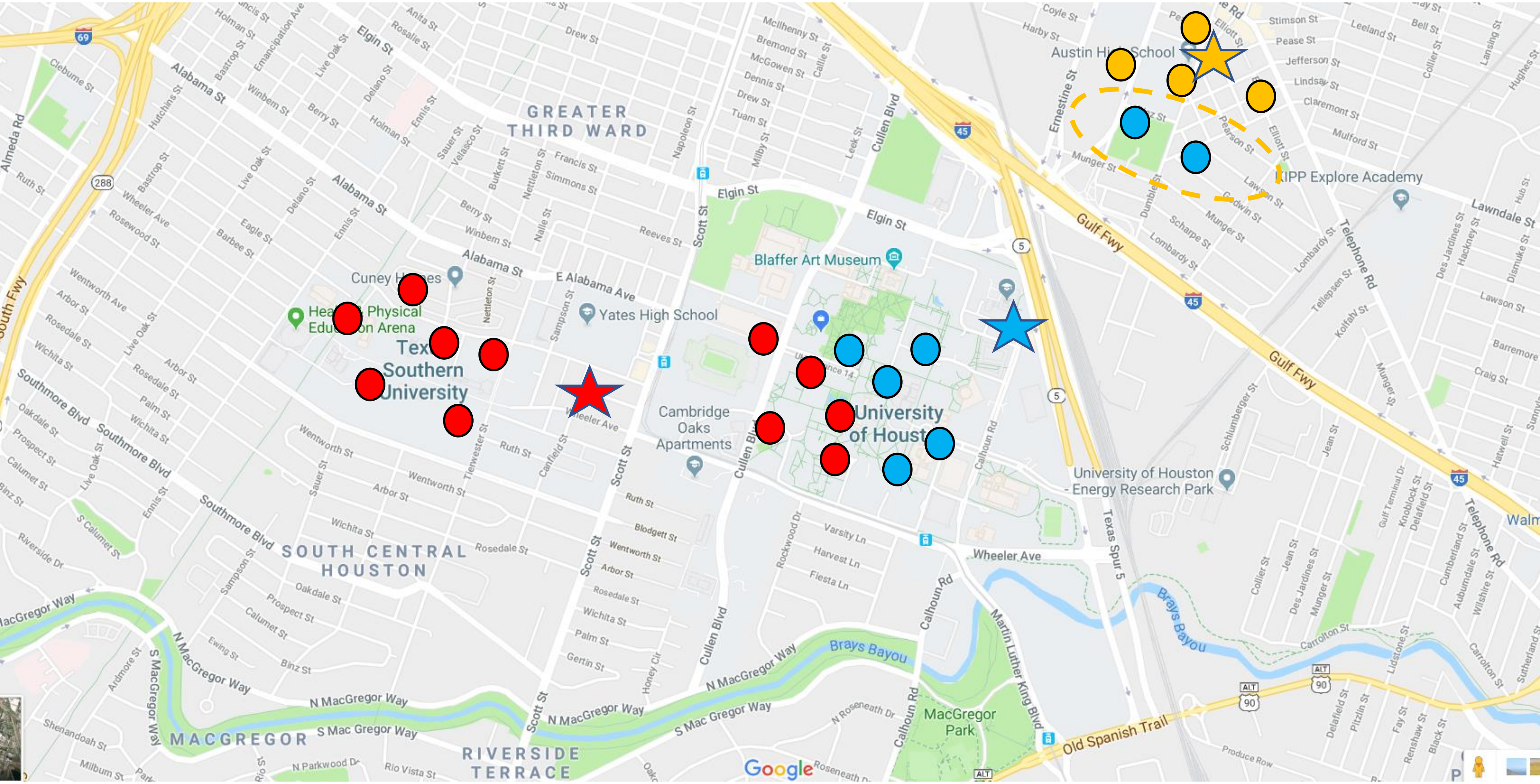
Update store locations



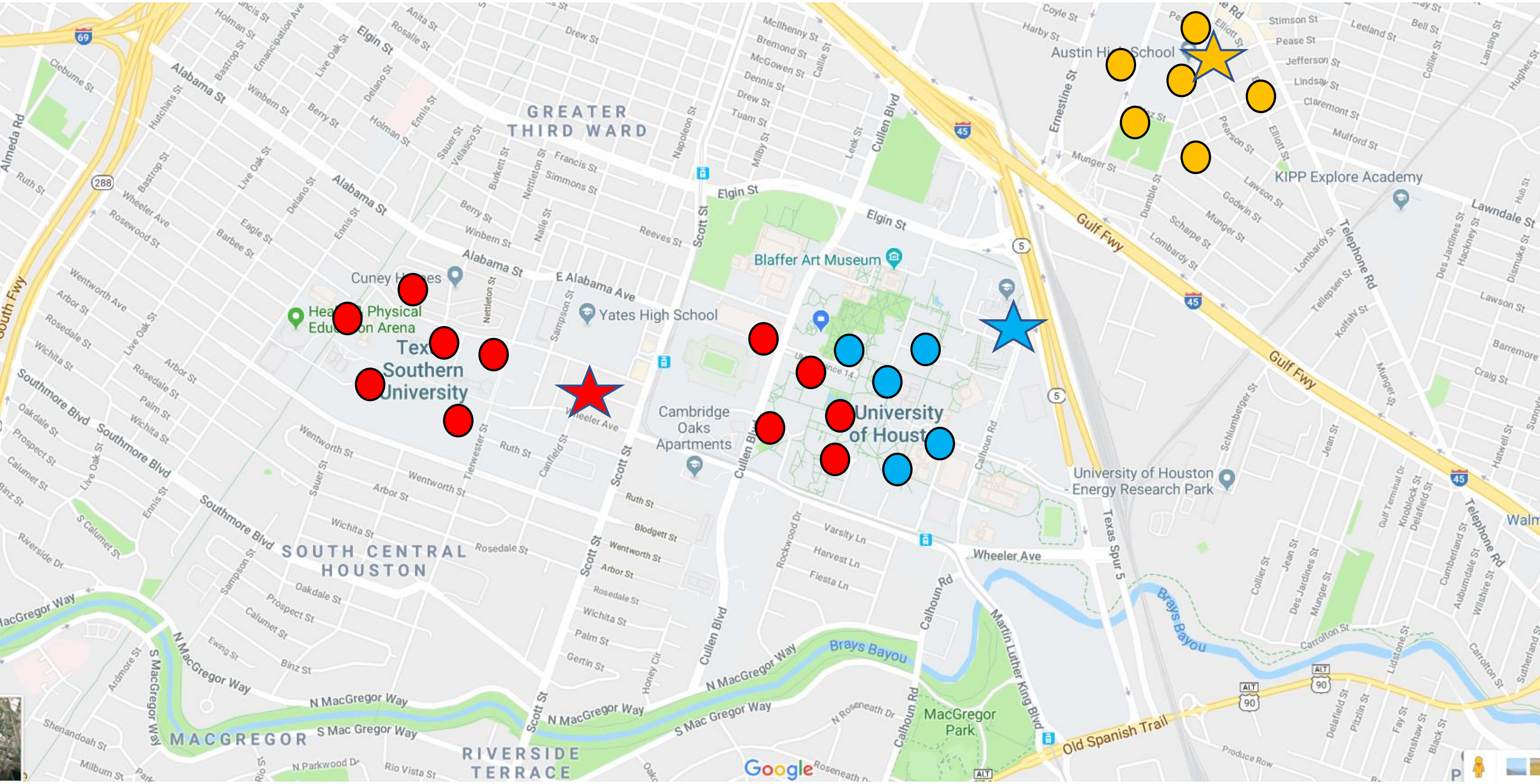
Update store locations



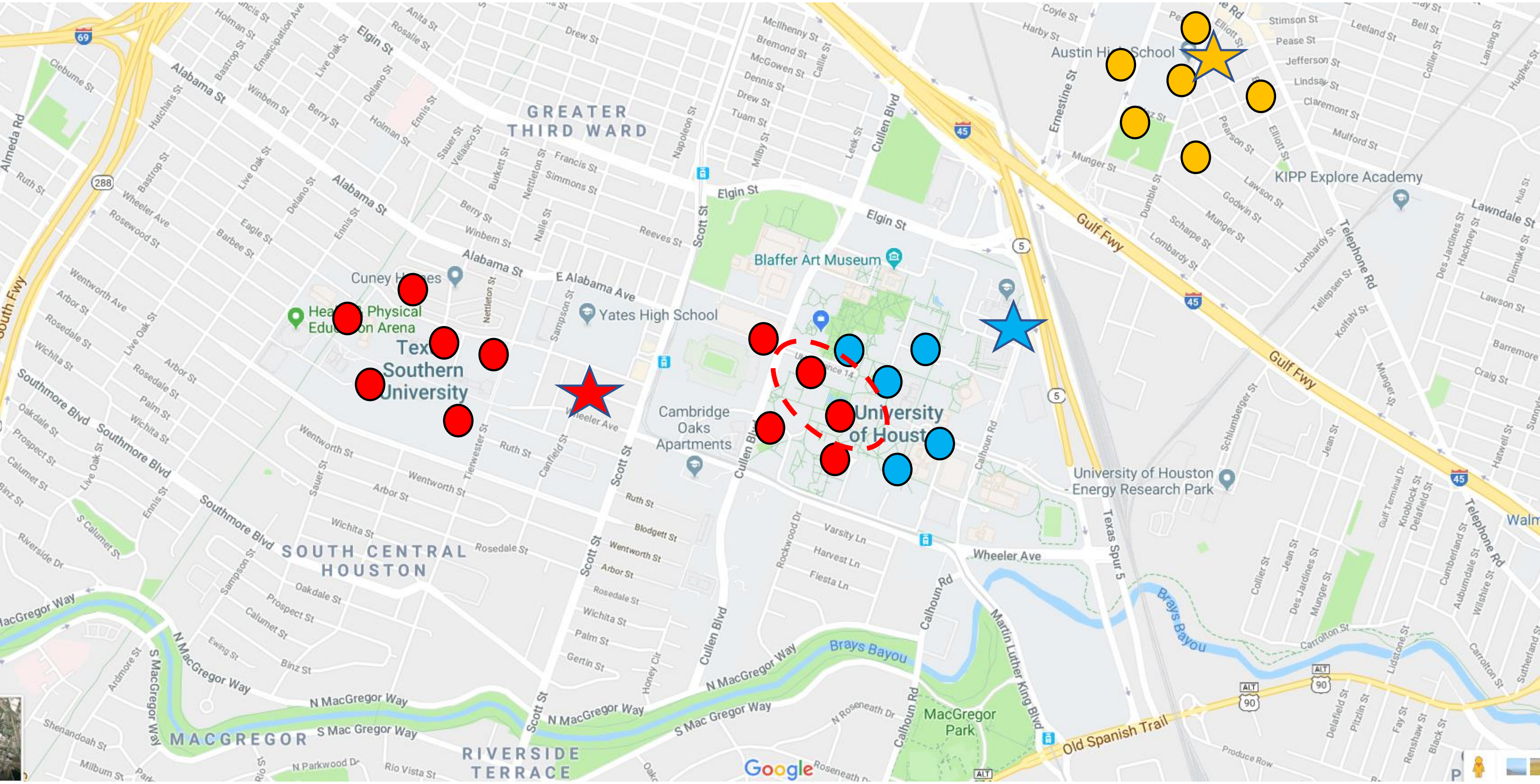
Customer distribution update



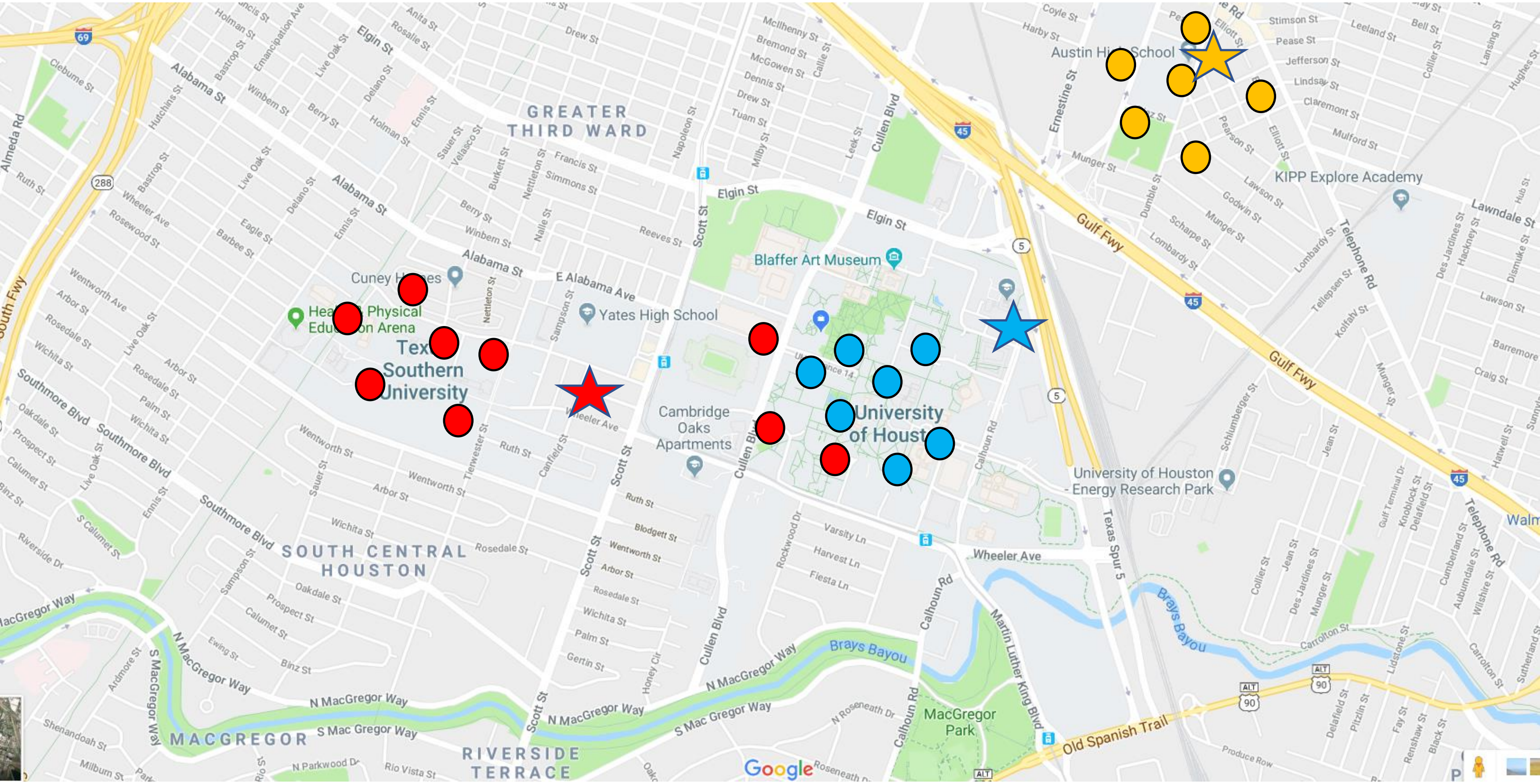
Customer distribution update



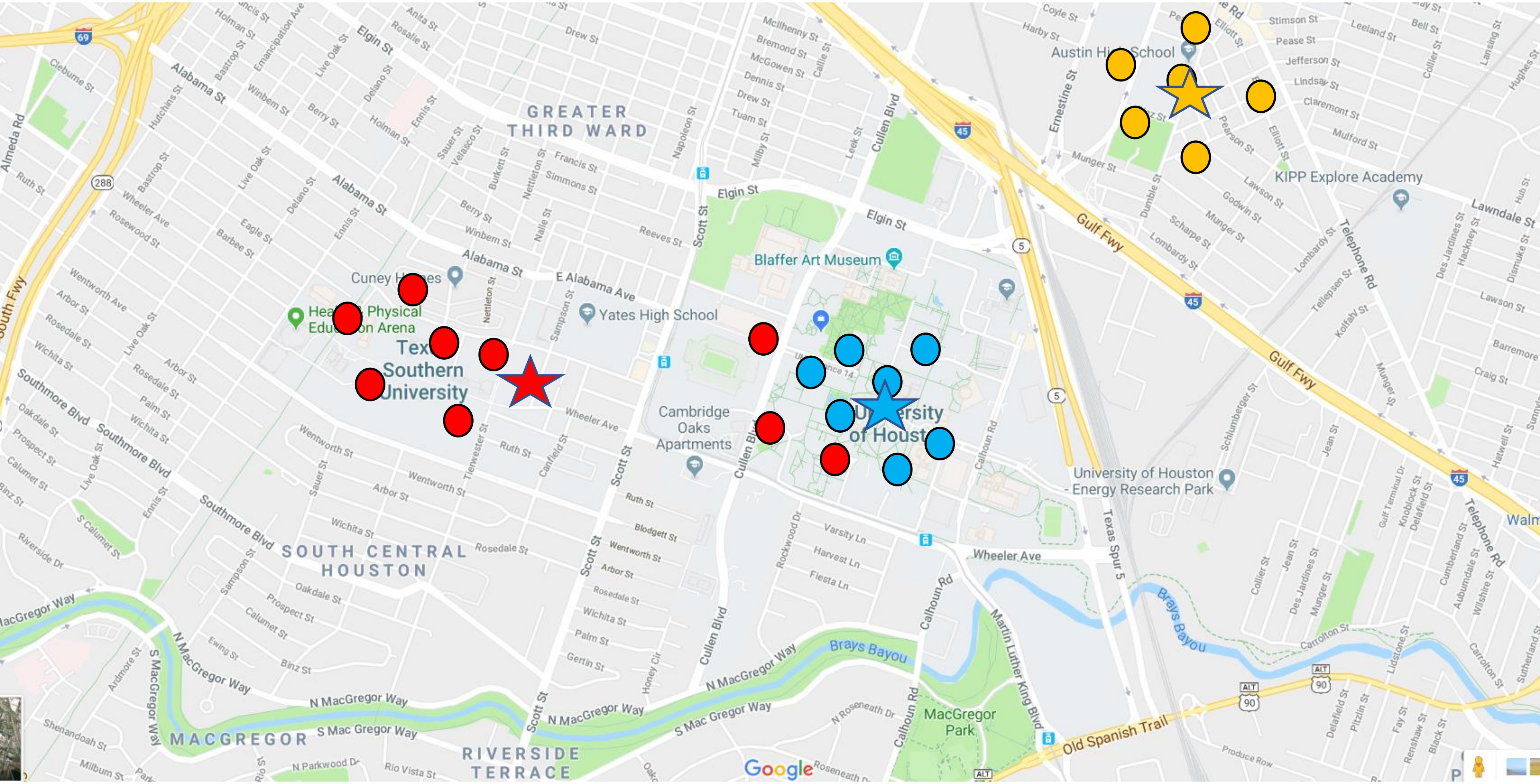
Customer distribution update



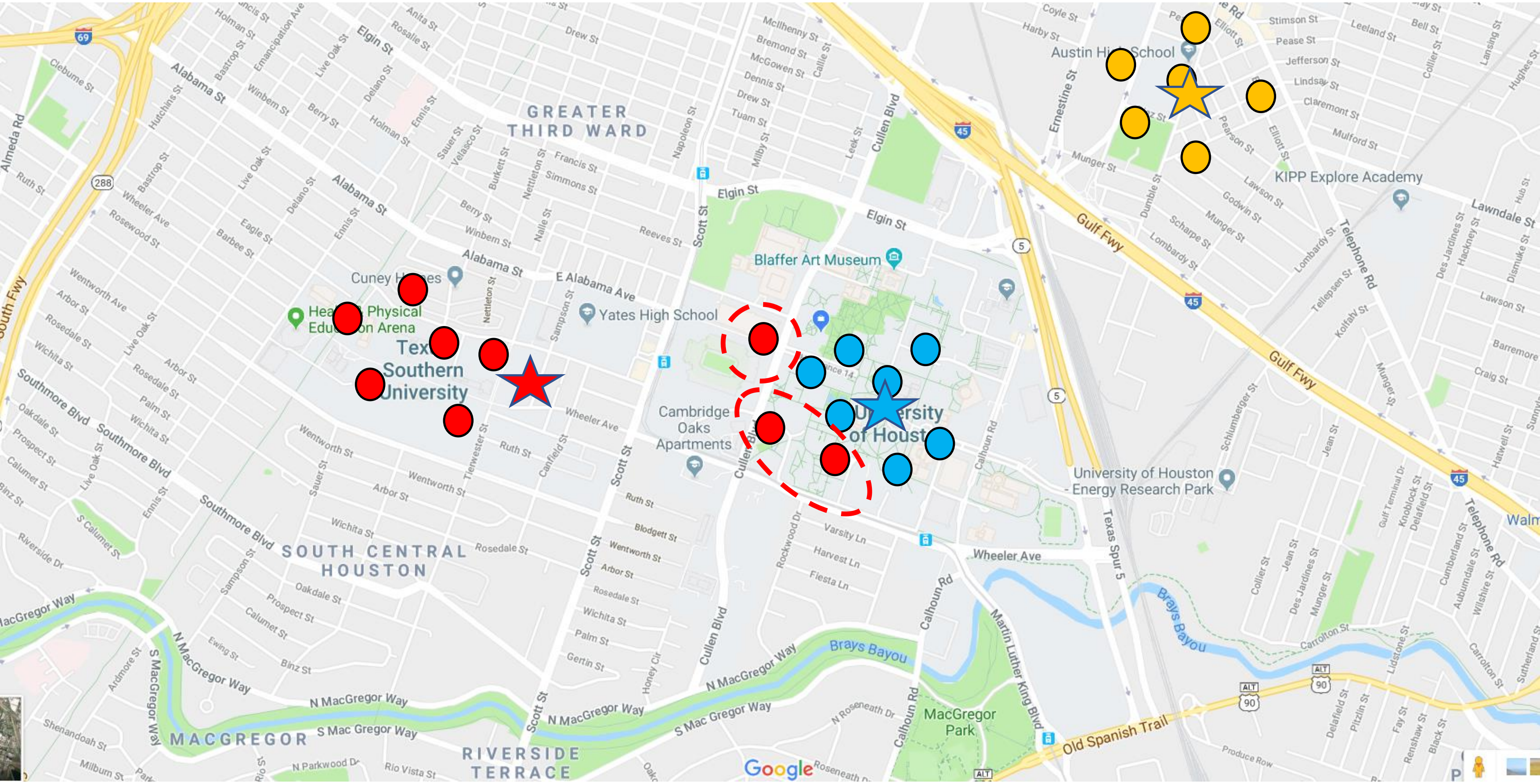
Customer distribution update



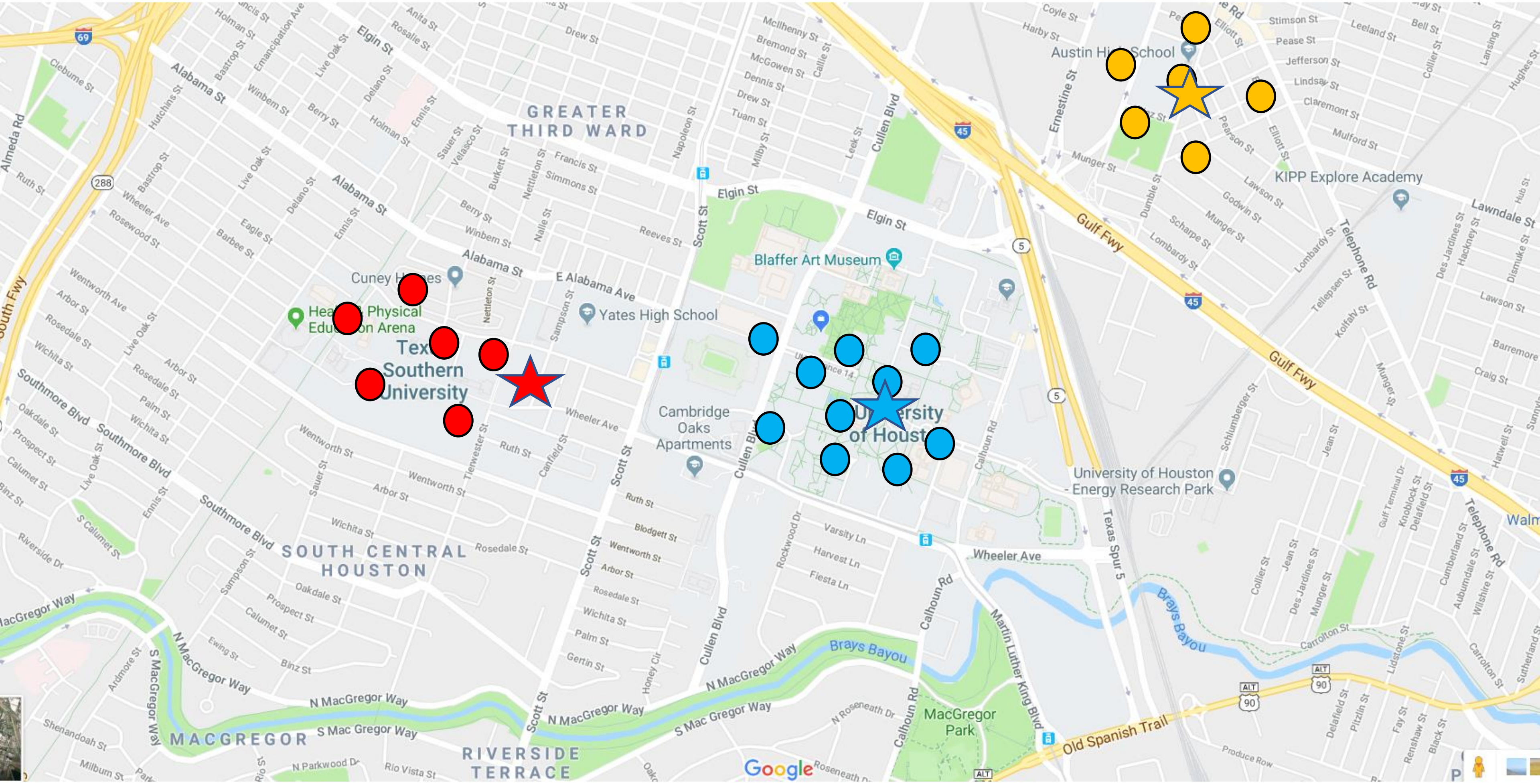
Update store locations



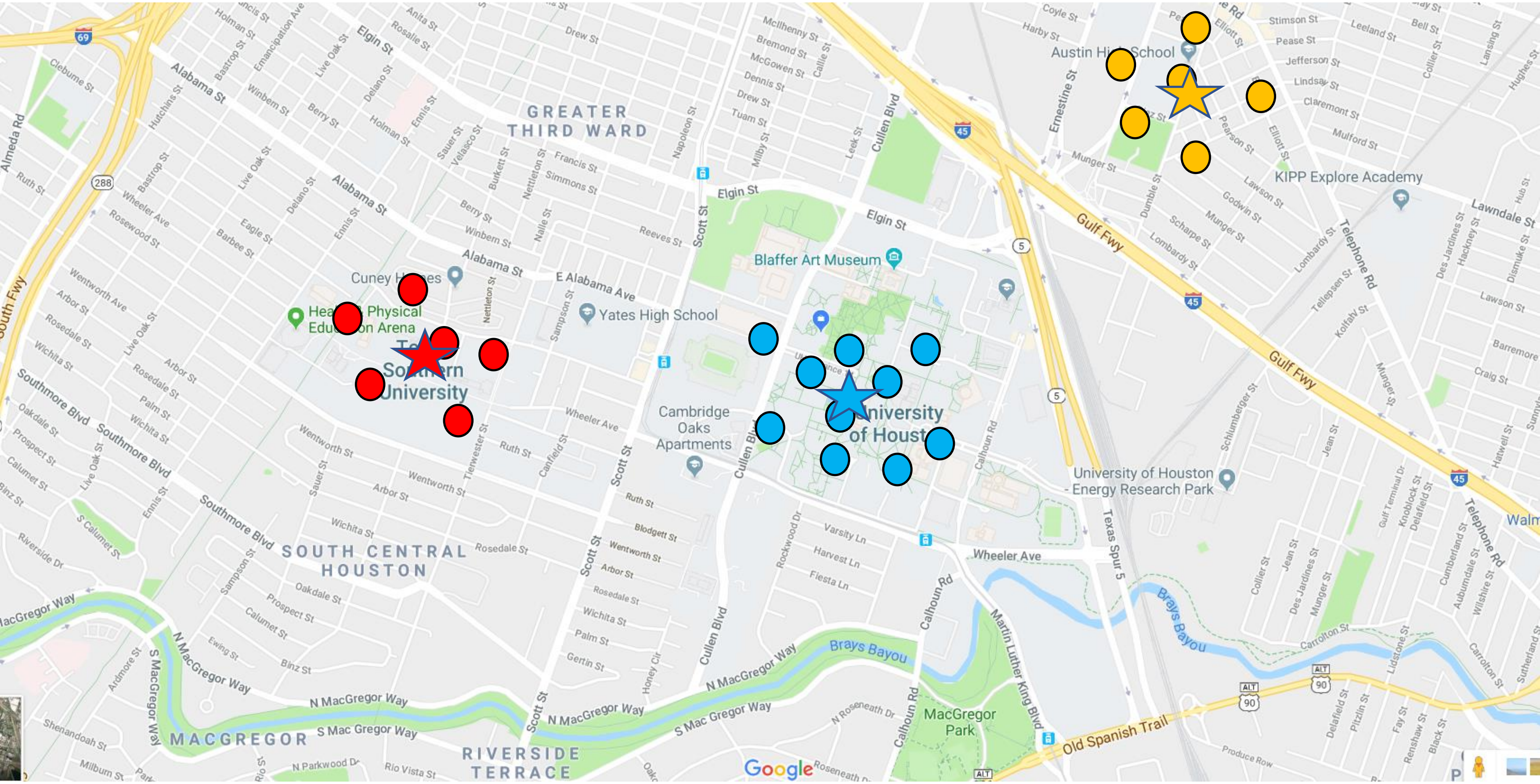
Update customer distributions



K-means clustering



Update store locations



K-means clustering: terminology

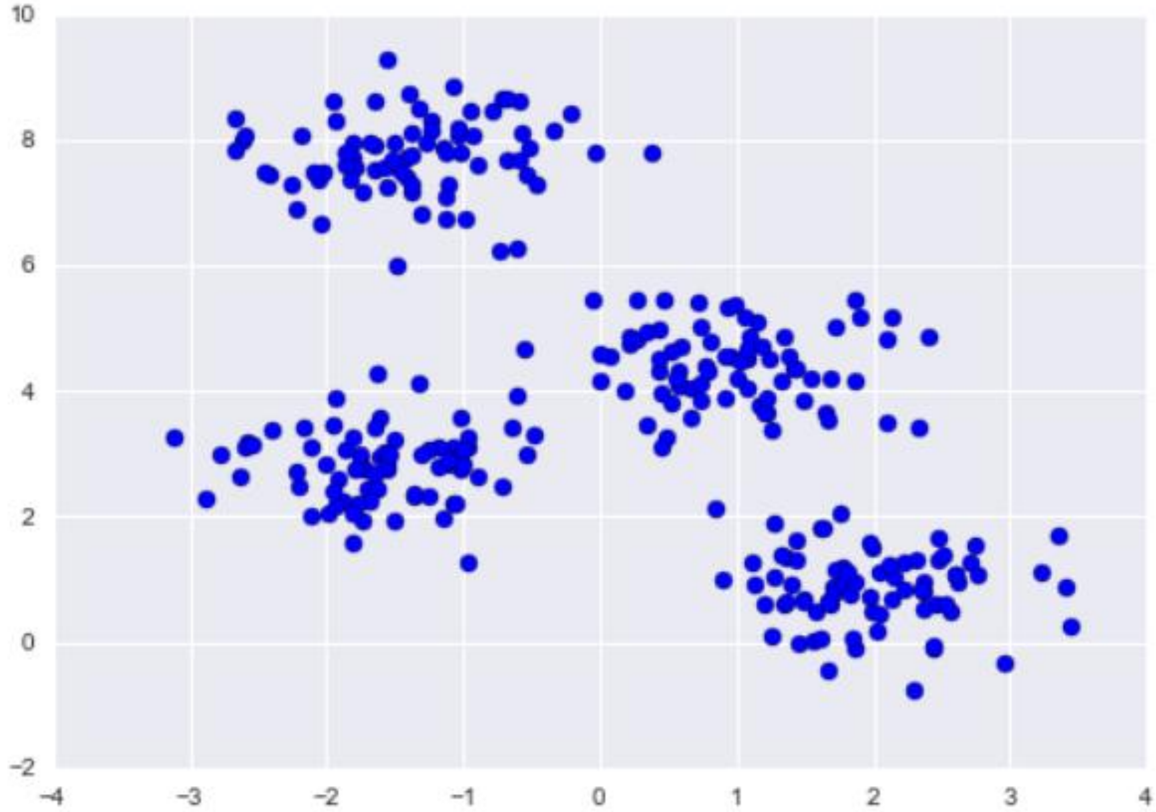
- The pizza store locations are called cluster centers or centroids.
- Each group of customers is called a cluster
- Each location or building is called an observation (or an object/instance).

K-means clustering: how it works

- Start with random initial cluster centers
- While (not converge)
  - Calculate distance between each cluster center and each instance.
  - Assign each instance to the nearest cluster center
  - Update cluster centers by calculating the mean of all the instances assigned to the same cluster
- end

Implementation in Scikit-learn

```
from sklearn.datasets.samples_generator import make_blobs
X, y_true = make_blobs(n_samples=300, centers=4,
                       cluster_std=0.60, random_state=0)
plt.scatter(X[:, 0], X[:, 1], s=50);
```



Implementation in Scikit-learn

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=4)
kmeans.fit(X)
y_kmeans = kmeans.predict(X)
```

Implementation in Scikit-learn

```
plt.scatter(X[:, 0], X[:, 1], c=y_kmeans, s=50, cmap='viridis')

centers = kmeans.cluster_centers_
plt.scatter(centers[:, 0], centers[:, 1], c='black', s=200, alpha=0.5);
```