Clustering

Sujay Sanghavi

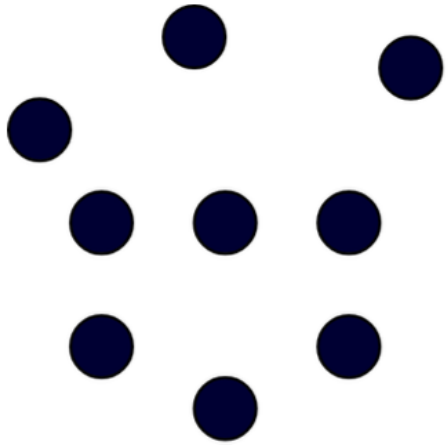
k-means Clustering

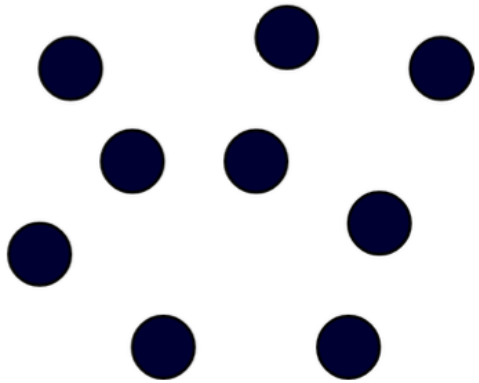
"Given data points $x_1, \ldots, x_n$, partition them into clusters"

... so that 2 points in the same cluster are "more like each other" as compared to 2 points in different clusters ...

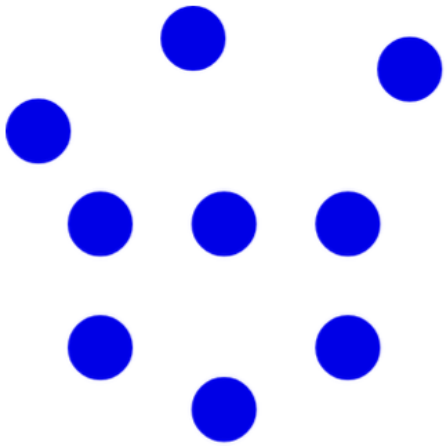
This is an **unsupervised learning** problem: there is no response / label y to predict.

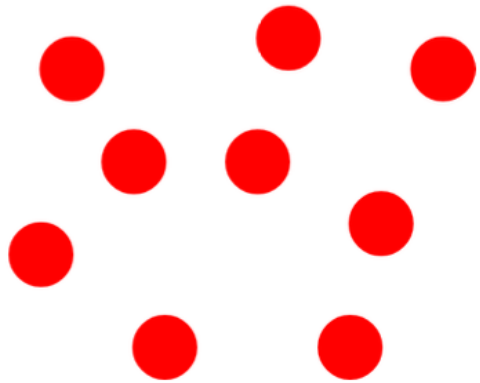
Clustering





Clustering





k-Means setup

Data: $x_1, x_2, \cdots, x_n \in \mathcal{R}^d$

Distance Metric: $d(x_i, x_j) = ||x_i - x_j||_2$

Cluster Objective: Points in same cluster closer

k-Means Objective

Find centers $\mu_1, \mu_2, \cdots, \mu_k$

and sets $S_1, S_2, \cdots, S_k$ to minimize

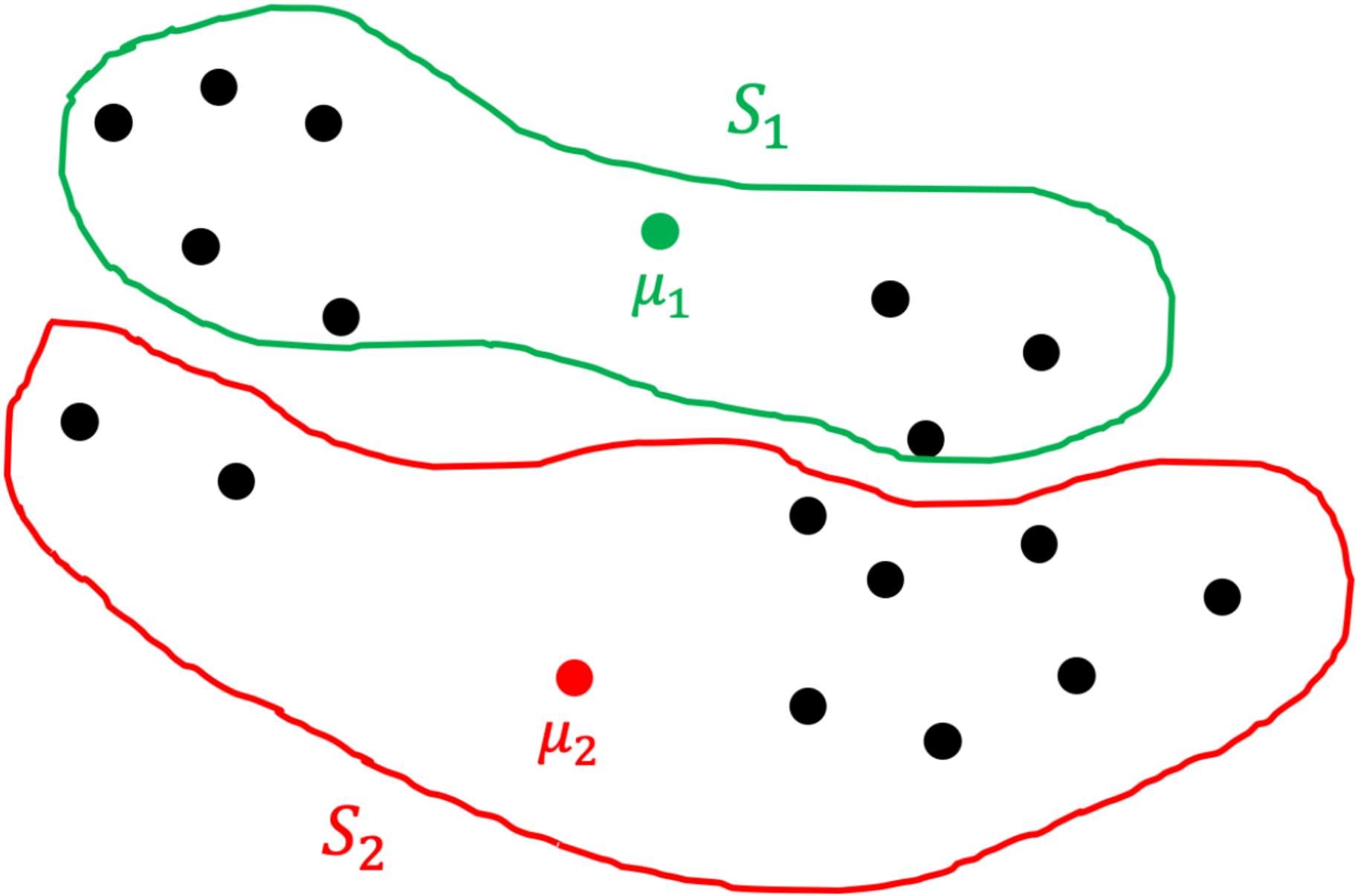
$$\sum_{i=1}^{k} \sum_{j \in S_i} ||x_j - \mu_i||_2^2$$

"Variance"/"Within Cluster Sum of Squares (WCSS)

$$\min_{\substack{\mu_1, \mu_2, \cdots, \mu_k \\ S_1, S_2, \cdots, S_k}} \sum_{i=1}^{k} \sum_{j \in S_i} ||x_j - \mu_i||_2^2$$

NP-Hard even for $k = 2$ clusters. We now study a popular heuristic.

Towards k-means algorithm



Towards k-means algorithm

$$\min_{\substack{\mu_1,\mu_2,\cdots,\mu_k \\ S_1,S_2,\cdots,S_k}} \sum_{i=1}^{k} \sum_{j \in S_i} ||x_j - \mu_i||_2^2$$

Given partition $S_1, S_2, \cdots, S_k$,

Best μ_i

$$= \arg\min_{\mu} \sum_{j \in S_i} ||x_j - \mu_i||^2$$

$$= \frac{1}{|S_i|} \sum_{j \in S_i} x_j$$

Or, mean of S_i

Towards k-means algorithm

$$\min_{\substack{\mu_1,\mu_2,\cdots,\mu_k \\ S_1,S_2,\cdots,S_k}} \sum_{i=1}^{k} \sum_{j \in S_i} ||x_j - \mu_i||_2^2$$

Given centers $\mu_1, \mu_2, \cdots, \mu_k$,

Best $S_1, S_2, \cdots, S_k$?

Towards k-means algorithm

$$\min_{\substack{\mu_1,\mu_2,\cdots,\mu_k \\ S_1,S_2,\cdots,S_k}} \sum_{i=1}^{k} \sum_{j \in S_i} ||x_j - \mu_i||_2^2$$

Given centers $\mu_1, \mu_2, \cdots, \mu_k$,

Assign each point x_j to its closest μ_i

k-Means Algorithm

$$\min_{\substack{\mu_1,\mu_2,\cdots,\mu_k \\ S_1,S_2,\cdots,S_k}} \sum_{i=1}^{k} \sum_{j \in S_i} ||x_j - \mu_i||_2^2$$

- Initialize $\mu_1^{(0)}, \mu_2^{(0)}, \cdots, \mu_k^{(0)}$
- Until Convergence, alternate between:

Assign every point to closest center

$$S_i^{(t)} = \left\{ j \mid ||x_j - \mu_i^{(t)}|| \leq ||x_j - \mu_{i'}^{(t)}|| \; \forall \, i' \right\}$$

Update centers to center of their assigments

$$\mu_i^{(t+1)} = \frac{1}{S_i^{(t)}} \left(\sum_{j \in S_i^{(t)}} x_j \right)$$

Convergence == when assignments do not change from one step to next

k-Means Algorithm

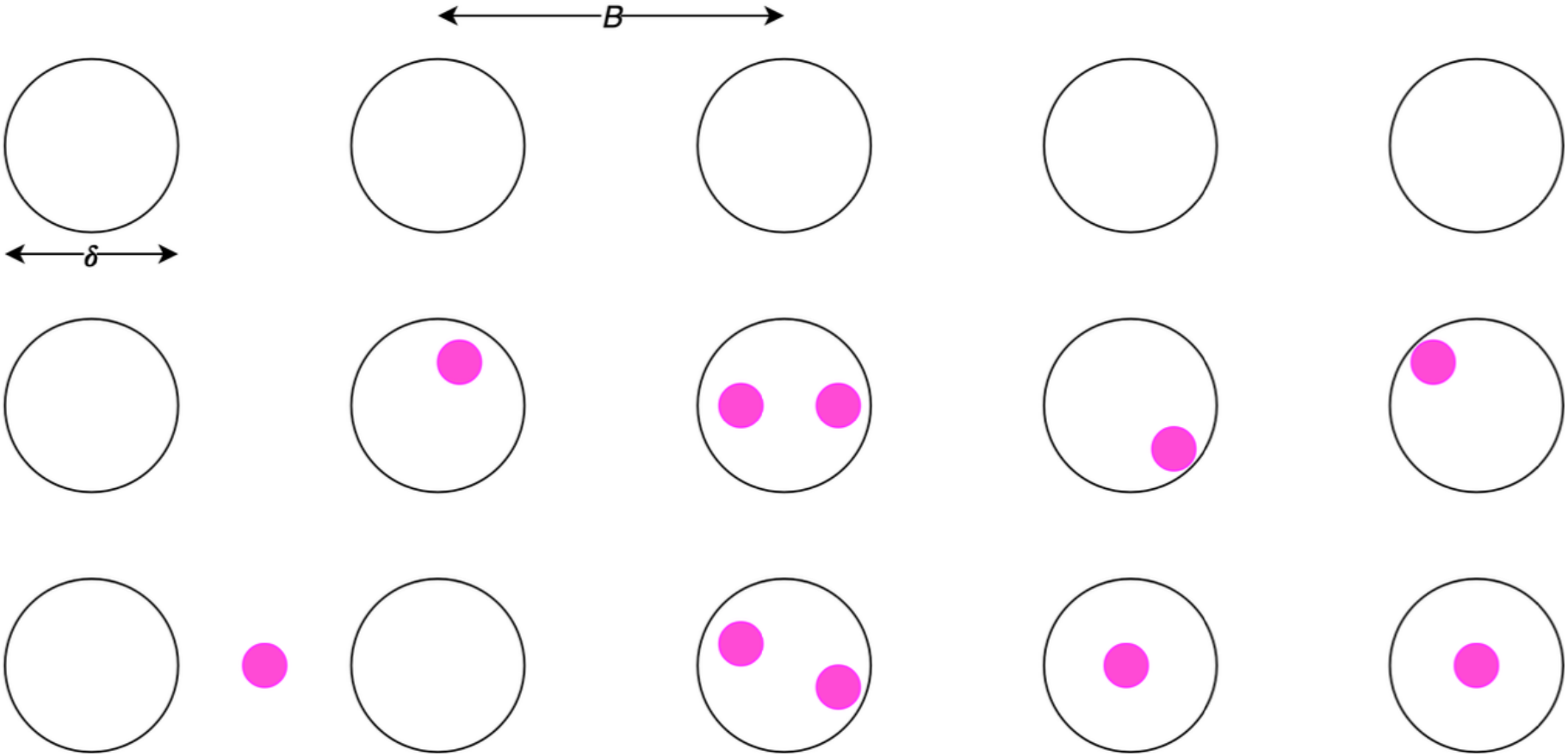
Guaranteed to converge

- because we are always decreasing a positive function
- and there are a finite number of choices for the $S_1, \ldots, S_k$

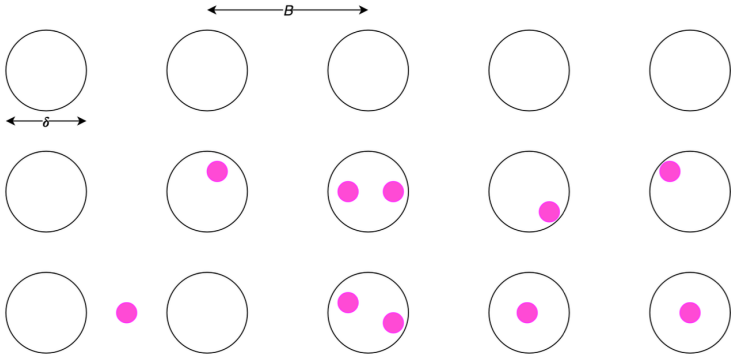
But final answer can be (arbitrarily) bad...

...it is a local optimum

Bad Example



Bad Example



- Data is well-separated into 5 parts, distance between each is B, distance within each is $\delta < B$.
- In first step, μ_is are chosen as in row 2. Points are assigned to the closest μ_i.
- In the next step, new μ_is are chosen as in row 3. In further iterations, they do not change.
- Resulting loss is high. **Initialization is important**

k-means ++

k-means ++ is a way to get a good initialization

- Pick μ_0: uniformly random from data points
- $t = 0$
- While $t < k$ sequentially pick centers:
  - Assign a cost to each remaining point

$$\text{cost}(x_j) = \min_{\mu \in \{\mu_1, \cdots, \mu_{t-1}\}} ||x_j - \mu||^2$$

  - Pick a point as the next center, proportional to its cost

$$P(\mu_t = x_j) \propto \text{cost}(x_j)$$

  - $t = t + 1$

k-means ++

Theoretical guarantee on the initialization:

$$E[\text{cost}(\mu_1, \cdots, \mu_k)] \leq E[\text{cost}(\mu_1^*, \cdots, \mu_k^*)]O(\log k)$$

... and note that cost decreases with each iteration of k-means

k-medoids

Sometimes the data points $x^{(i)}$ are not in euclidean space.

E.g. task: organize all articles of the New York Times this year into 8 clusters.

Now each x is a text document, and we can define a **similarity** between any pair x and x' of documents to be

$$\mathcal{S}(x, x') = \frac{\text{number of unique words common to both } x \text{ and } x'}{\text{total number of unique words in } x \text{ and } x'}$$

Cannot do k-means because we cannot make an "average document"

k-medoids

Similar idea to k-means, except that the **cluster centers have to be one of the original data points**

- Initialize medoids $\mu_1^{(0)}, \mu_2^{(0)}, \cdots, \mu_k^{(0)}$ (e.g. by choosing k random data points)
- Until Convergence, alternate between:

Assign every data point to the most-similar current medoid

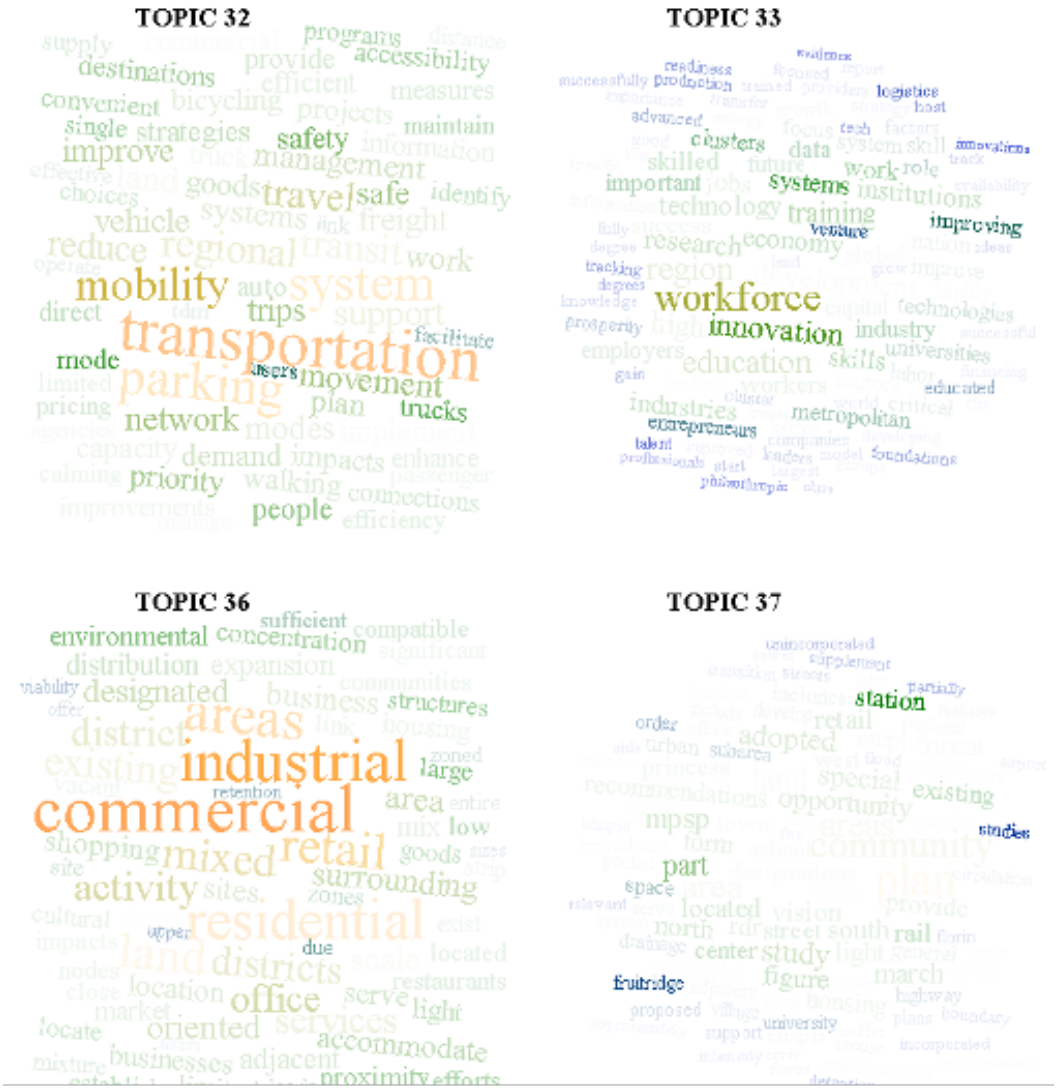
$$S_i^{(t)} = \left\{ j \mid \mathcal{S}(x_j, \mu_i^{(t)}) \geq \mathcal{S}(x_j, \mu_{i'}^{(t)}) \; \forall \, i' \right\}$$

Update the medoid of each cluster to the point most similar to all the other points in that cluster

$$\mu_i^{(t+1)} = \arg \max_{x \in S_i^{(t)}} \left(\sum_{x' \in S_i^{(t)} - x} \mathcal{S}(x, x') \right)$$

Convergence == when assignments do not change from one step to next

k-medoids


TOPIC 32
programs
accessibility
destinations
provide
efficient
measures
convenient
bicycling
projects
single
strategies
safety
maintain
improve
management
travel
safe
identify
goods
choices
vehicle
systems
freight
reduce
regional
transit
work
mobility
auto
system
direct
trips
support
transportation
facilitate
mode
parking
users
movement
pricing
plan
trucks
network
modes
capacity
demand
impacts
enhance
priority
walking
connections
people
efficiency
TOPIC 33
readiness
production
logistics
host
advanced
clusters
tech
data
skilled
future
work
role
important
systems
institutions
technology
training
improving
venture
research
economy
region
workforce
innovation
industry
technologies
universities
employers
education
skills
educated
industries
metropolitan
entrepreneurs
TOPIC 36
sufficient
environmental
concentration
compatible
distribution
expansion
designated
business
structures
areas
district
industrial
large
commercial
retention
area
mix
low
shopping
mixed
retail
goods
activity
surrounding
residential
zones
upper
due
land
districts
located
restaurants
location
office
serve
light
oriented
services
accommodate
businesses
adjacent
proximity
efforts
TOPIC 37
station
order
adopted
retail
special
existing
opportunity
mpsp
studies
community
part
space
plan
located
vision
provide
north
street
south
rail
center
study
figure
march
fruitridge
university