CSCE 689: Special Topics in Modern Algorithms for Data Science

Lecture 25

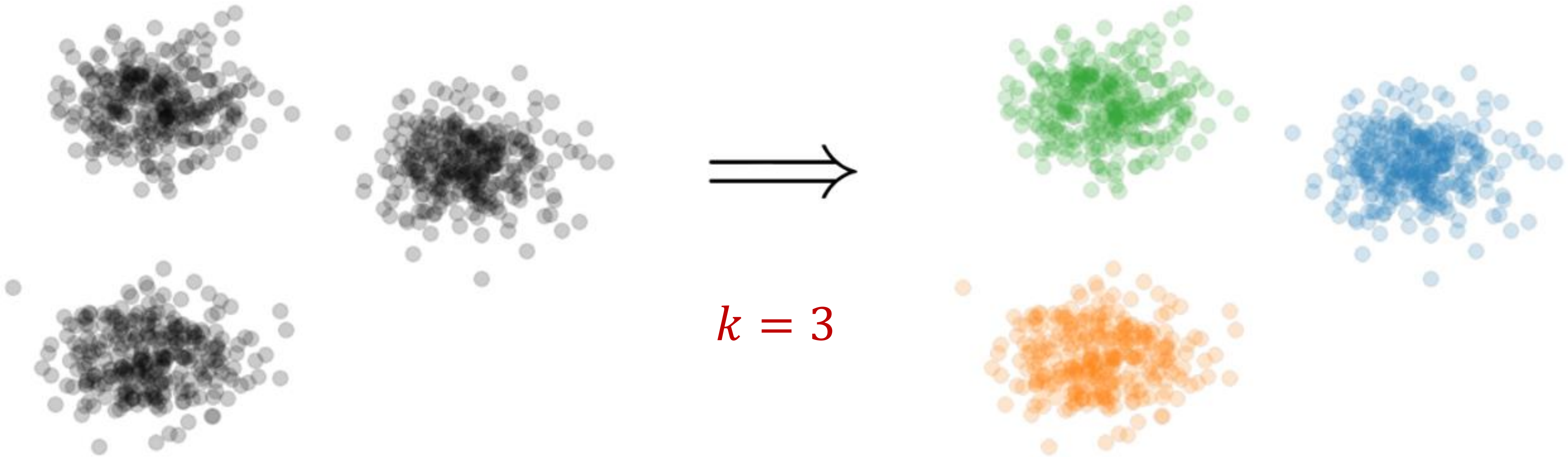
Samson Zhou

Presentation Schedule

- November 27: Chunkai, Jung, Galaxy AI
- November 29: STMI, Anmol, Jason
- December 1: Bokun, Ayesha, Dawei, Lipai

Previously: k-Clustering

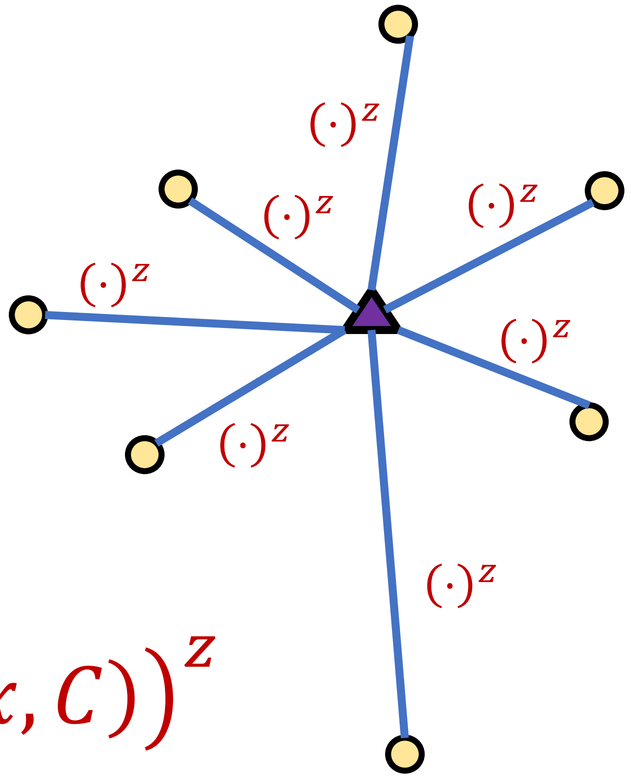
- Goal: Given input dataset X, partition X so that "similar" points are in the same cluster and "different" points are in different clusters
- There can be at most k different clusters



Previously: k-Clustering

- Define clustering cost $\text{Cost}(X, C)$ to be a function of $\{\text{dist}(x, C)\}_{x \in C}$

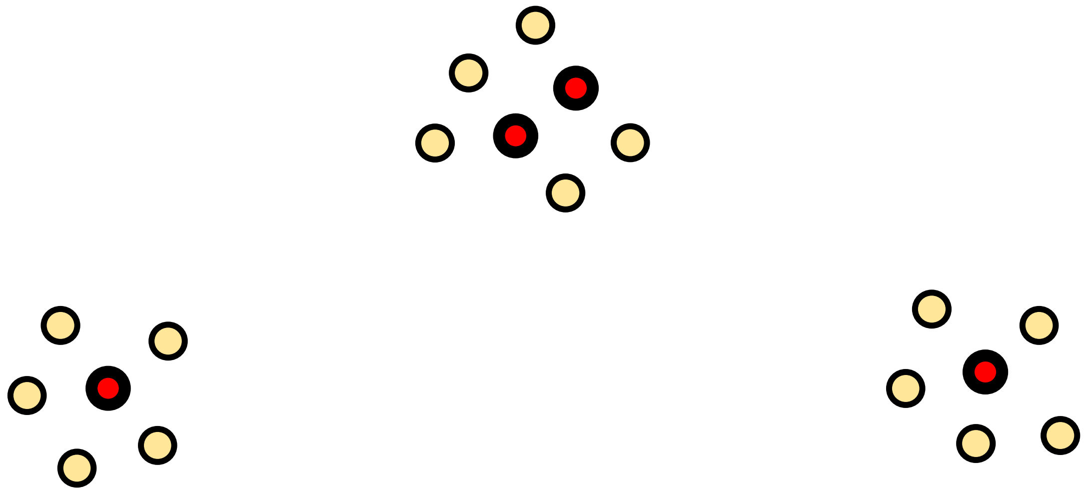
- k-center: $\text{Cost}(X, C) = \max_{x \in X} \text{dist}(x, C)$
- k-median: $\text{Cost}(X, C) = \sum_{x \in X} \text{dist}(x, C)$
- k-means: $\text{Cost}(X, C) = \sum_{x \in X} \left(\text{dist}(x, C)\right)^2$
- (k, z)-clustering: $\text{Cost}(X, C) = \sum_{x \in X} \left(\text{dist}(x, C)\right)^z$



Previously: Coreset

- Subset X' of representative points of X for a specific clustering objective

- $\text{Cost}(X, C) \approx \text{Cost}(X', C)$ for all sets C with $|C| = k$



Previously: Coreset

- Given a set X and an accuracy parameter $\varepsilon > 0$, we say a set X' with weight function w is an $(1+\varepsilon)$-*multiplicative coreset* for a cost function Cost, if for all queries C with $|C| \leq k$, we have

$$(1-\varepsilon)\mathrm{Cost}(X,C) \leq \mathrm{Cost}(X',C,w) \leq (1+\varepsilon)\mathrm{Cost}(X,C)$$

(k,z)-clustering: $\mathrm{Cost}(X',C,w) = \sum_{x\in X'} w(x)\cdot\left(\mathrm{dist}(x,C)\right)^z$

Previously: Bernstein's Inequality

- Bernstein's inequality: Let $X_1, \dots, X_n \in [-M, M]$ be independent random variables and let $X = X_1 + \cdots + X_n$ have mean μ and variance σ^2. Then for any $t \geq 0$:

$$\Pr[|X - \mu| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2 + \frac{4}{3}Mt}}$$

- Example: Suppose $M = 1$ and let $t = k\sigma$. Then

$$\Pr[|X - \mu| \geq k\sigma] \leq 2\exp\left(-\frac{k^2}{4}\right)$$

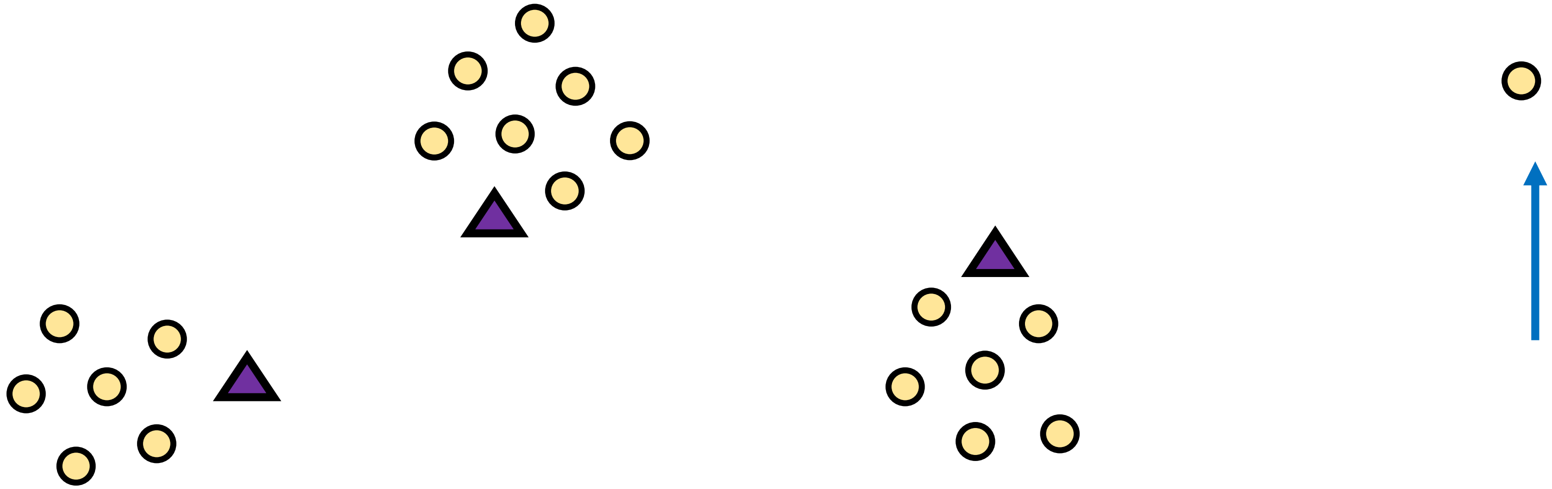
Previously: Importance Sampling for Sum Estimation

- Suppose $x_1, \ldots, x_n \in [1, n]$
- Suppose $p_i = \frac{x_i}{x}$ for all $i \in [n]$

- Can get a 2-approximation for importance sampling
- How many samples do we expect? $\frac{x_1}{x} + \cdots + \frac{x_n}{x} = 1$, so just a constant number of samples!

Last Time: Coreset Construction and Sampling

- Consider a fixed set X and a fixed set C of k centers, which induces a fixed cost $\text{Cost}(X, C)$
- Uniform sampling needs a lot of samples if there is a single point that greatly contributes to $\text{Cost}(X, C)$

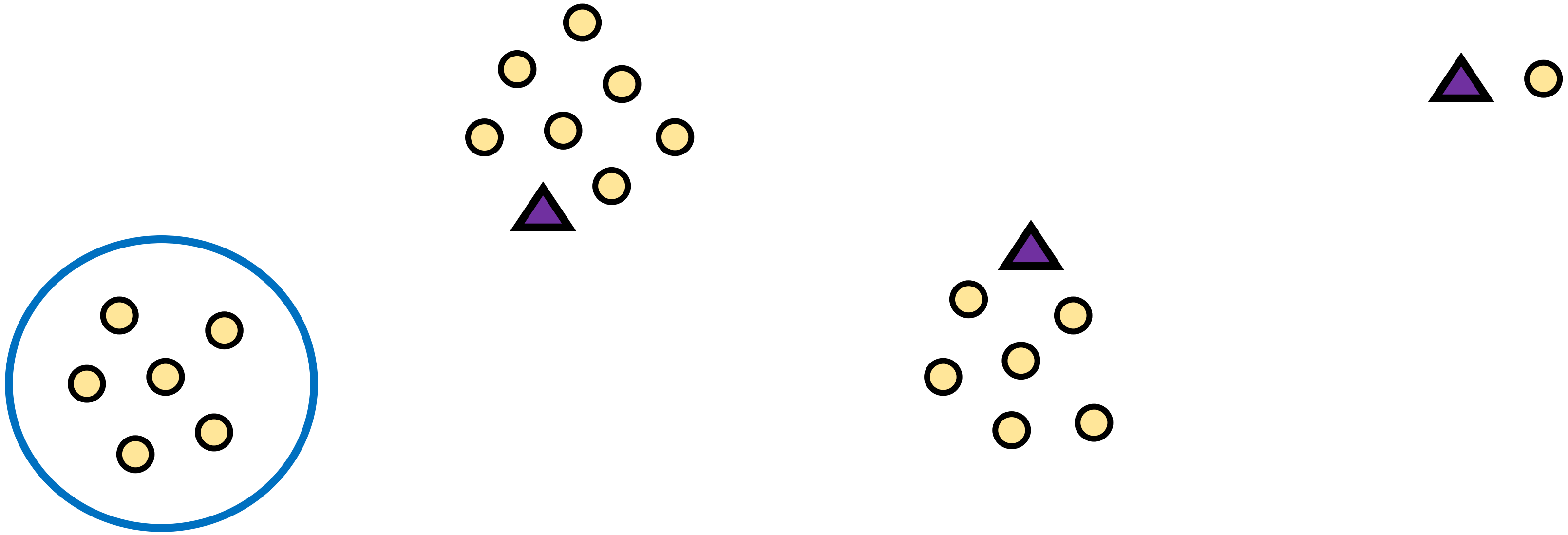


Last Time: Coreset Construction and Sampling

- Importance sampling, sample each point $x \in X$ into X' with probability proportional $\text{Cost}(x, C)$, i.e., $\text{Cost}(x, C)/\text{Cost}(X, C)$
- Importance sampling only needs X' to have size $O\left(\frac{1}{\varepsilon^2}\right)$ to achieve $(1 + \varepsilon)$-approximation to $\text{Cost}(X, C)$

Last Time: Coreset Construction and Sampling

- Importance sampling only needs X' to have size $O\left(\frac{1}{\varepsilon^2}\right)$ to achieve $(1+\varepsilon)$-approximation to $\text{Cost}(X, C)$
- What about a different choice C of k centers?



Last Time: Coreset Construction and Sampling

- Importance sampling only needs X' to have size $O\left(\frac{1}{\varepsilon^2}\right)$ to achieve $(1+\varepsilon)$-approximation to $\text{Cost}(X, C)$
- To handle all possible sets of k centers:
  - Need to sample each point x with probability $\max_C \frac{\text{Cost}(x,C)}{\text{Cost}(X,C)}$ instead of $\frac{\text{Cost}(x,C)}{\text{Cost}(X,C)}$
  - Need to union bound over a net of all possible sets of k centers

Net with size $\left(\frac{n\Delta}{\varepsilon}\right)^{O(kd)}$

Last Time: Sensitivity Sampling

- The quantity $s(x) = \max_C \frac{\text{Cost}(x,C)}{\text{Cost}(X,C)}$ is called the *sensitivity* of x and intuitively measures how "important" the point x is

- The *total sensitivity* of X is $\sum_{x \in X} s(x)$ and quantifies how many points will be sampled into X' through importance/sensitivity sampling (before the union bound)

Putting Things Together

- Consider a fixed set X and a fixed set C of k centers, which induces a fixed cost $\text{Cost}(X, C)$

- If we sample each point with probability $p(x) := \min\left(\frac{s(x)}{\varepsilon^2}\log\frac{1}{\delta}\right)$, then we get achieve $(1+\varepsilon)$-approximation to $\text{Cost}(X, C)$ with probability $1-\delta$

- What should δ be? How many points are sampled?

Putting Things Together

- What should δ be? How many points are sampled?

- Can union bound over multiple choices of C

- Recall: Net with size $\left(\frac{n\Delta}{\varepsilon}\right)^{O(kd)}$

Putting Things Together

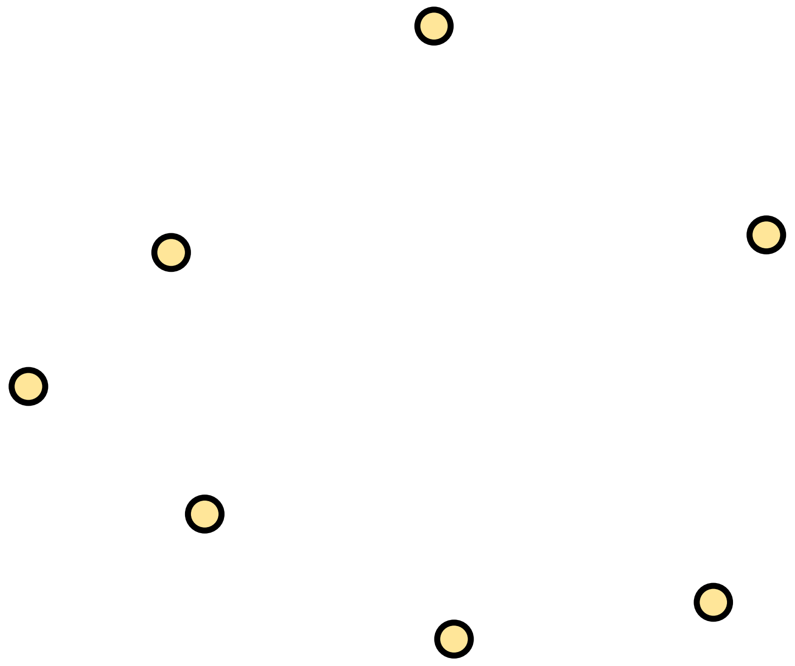
- Recall: Net with size $\left(\frac{n\Delta}{\varepsilon}\right)^{O(kd)}$
- Correctness on net implies correctness everywhere, so we set $\delta = \frac{1}{100} \cdot \left(\frac{\varepsilon}{n\Delta}\right)^{O(kd)}$ and by a union bound, our algorithm succeeds with probability 0.99
- $\log\frac{1}{\delta} = kd \cdot \log\frac{n\Delta}{\varepsilon}$

Putting Things Together

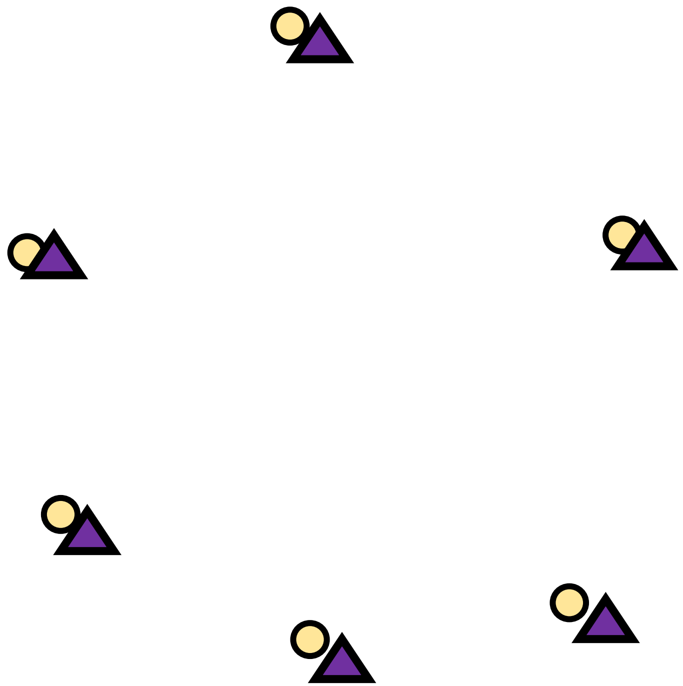
- $p(x) \coloneqq \min\left(\frac{s(x)}{\varepsilon^2}\log\frac{1}{\delta}\right)$, so we sample $\sum_{x\in X} p(x)$ points in expectation

- At most $\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\sum_{x\in X} s(x)$ points in total
- Since $\log\frac{1}{\delta} = kd \cdot \log\frac{n\Delta}{\varepsilon}$, then $\frac{kd}{\varepsilon^2} \cdot \log\frac{n\Delta}{\varepsilon} \cdot \sum_{x\in X} s(x)$ points
- What is $\sum_{x\in X} s(x)$? Total sensitivity!

$$s(x_t) = \max_{C:|C|\leq k} \frac{\mathrm{Cost}(x_t, C)}{\mathrm{Cost}(X, C)} = \max_{C:|C|\leq k} \frac{\mathrm{Cost}(x_t, C)}{\sum_{i=1}^{n} \mathrm{Cost}(x_i, C)}$$

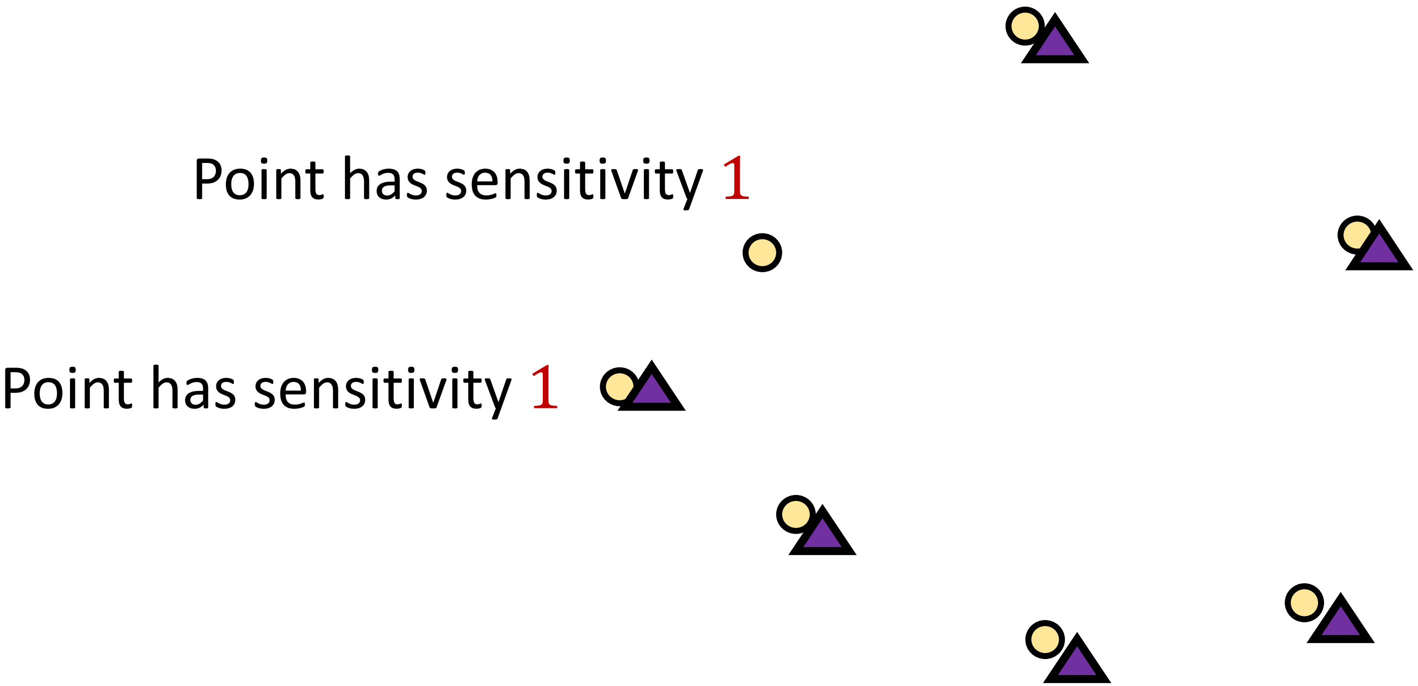


$$s(x_t) = \max_{C:|C|\leq k} \frac{\text{Cost}(x_t, C)}{\text{Cost}(X, C)} = \max_{C:|C|\leq k} \frac{\text{Cost}(x_t, C)}{\sum_{i=1}^{n} \text{Cost}(x_i, C)}$$

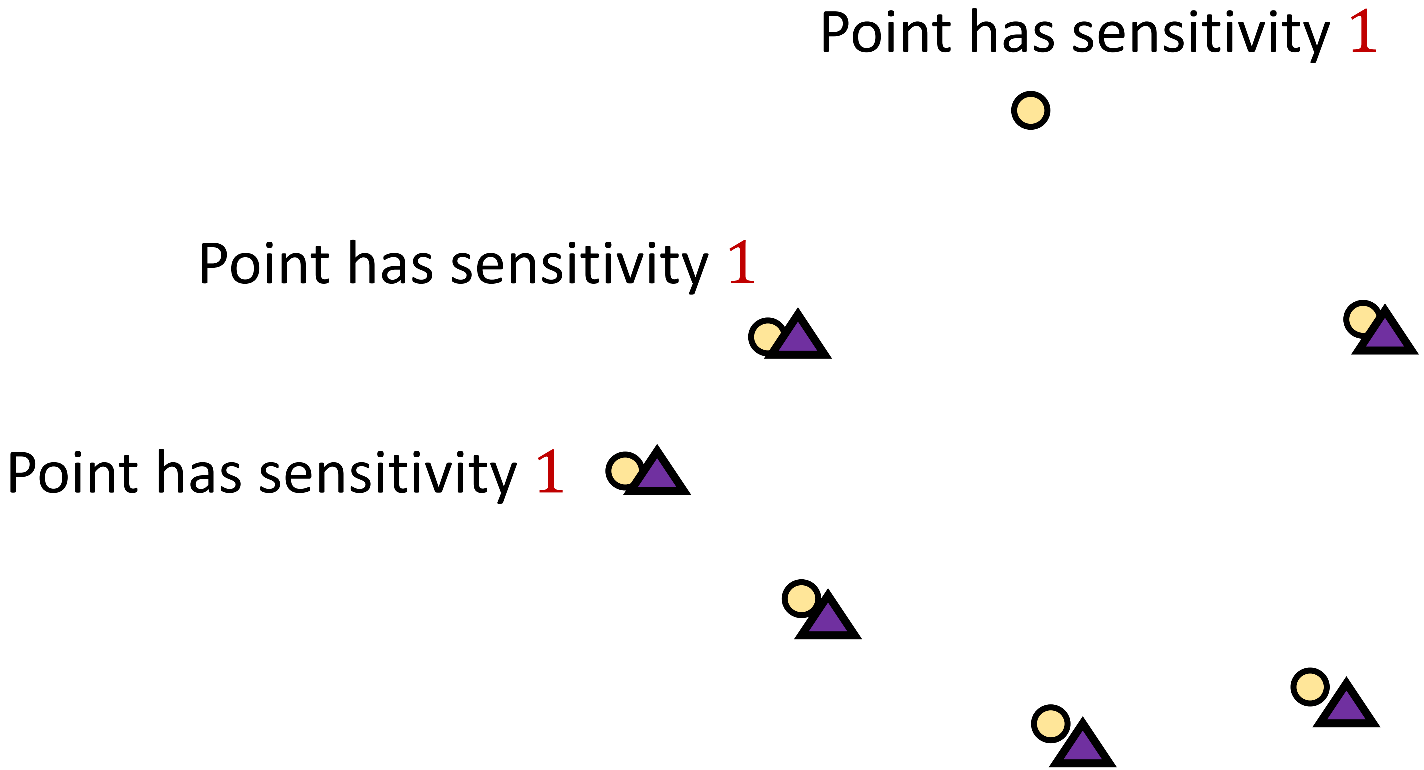


Point has sensitivity 1

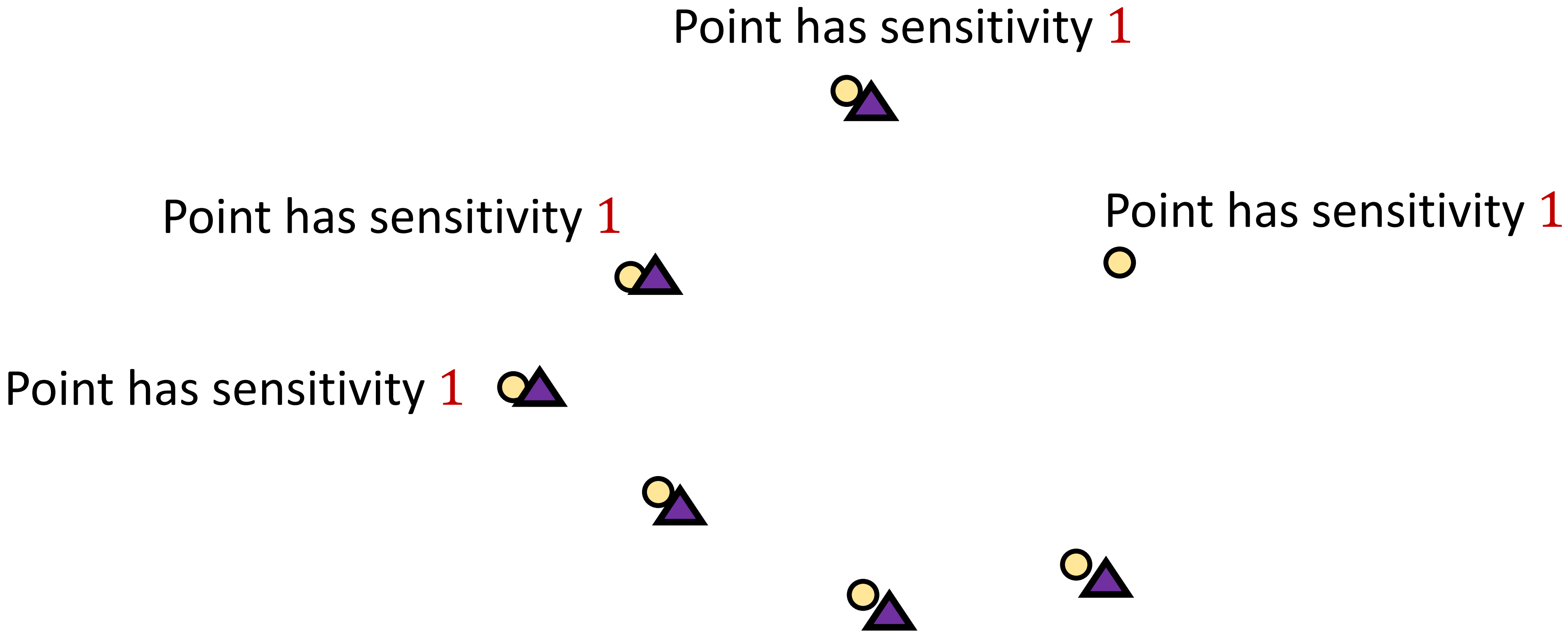
$$s(x_t) = \max_{C:|C|\leq k} \frac{\text{Cost}(x_t, C)}{\text{Cost}(X, C)} = \max_{C:|C|\leq k} \frac{\text{Cost}(x_t, C)}{\sum_{i=1}^{n} \text{Cost}(x_i, C)}$$



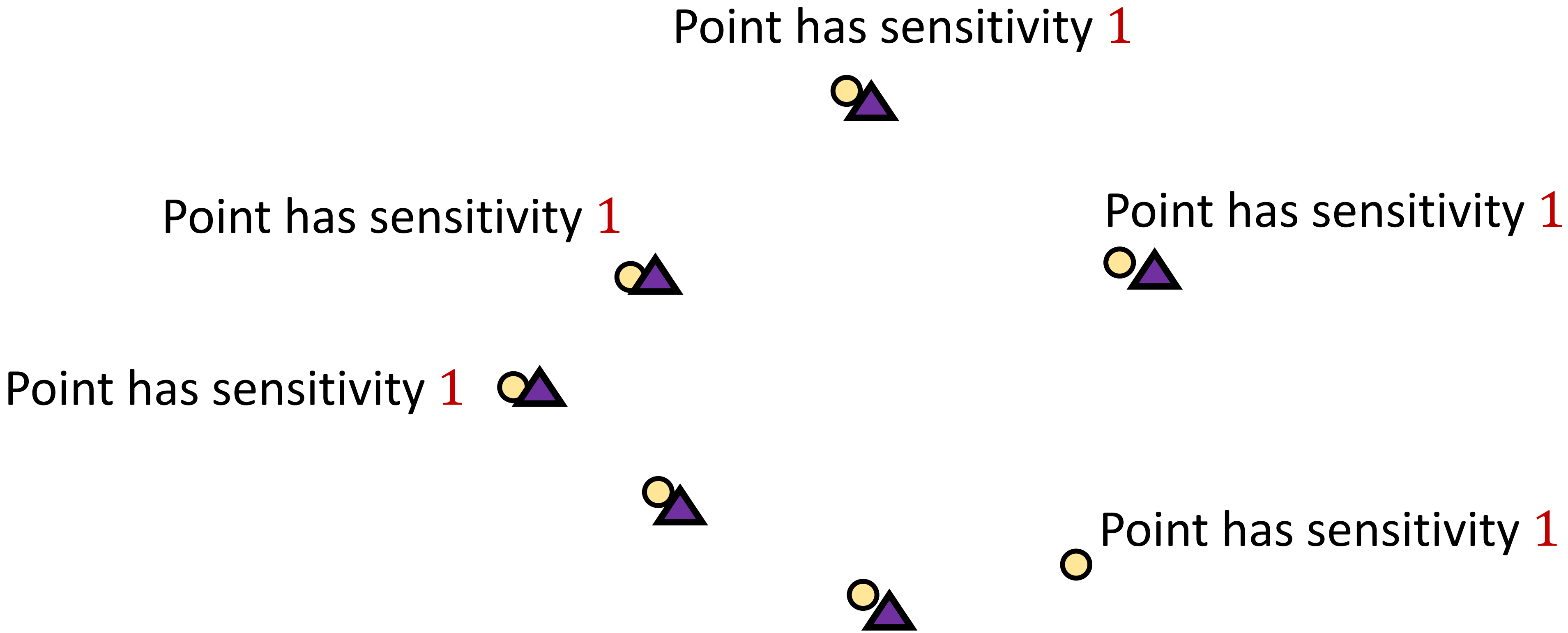
$$s(x_t) = \max_{C:|C|\leq k} \frac{\text{Cost}(x_t, C)}{\text{Cost}(X, C)} = \max_{C:|C|\leq k} \frac{\text{Cost}(x_t, C)}{\sum_{i=1}^{n} \text{Cost}(x_i, C)}$$



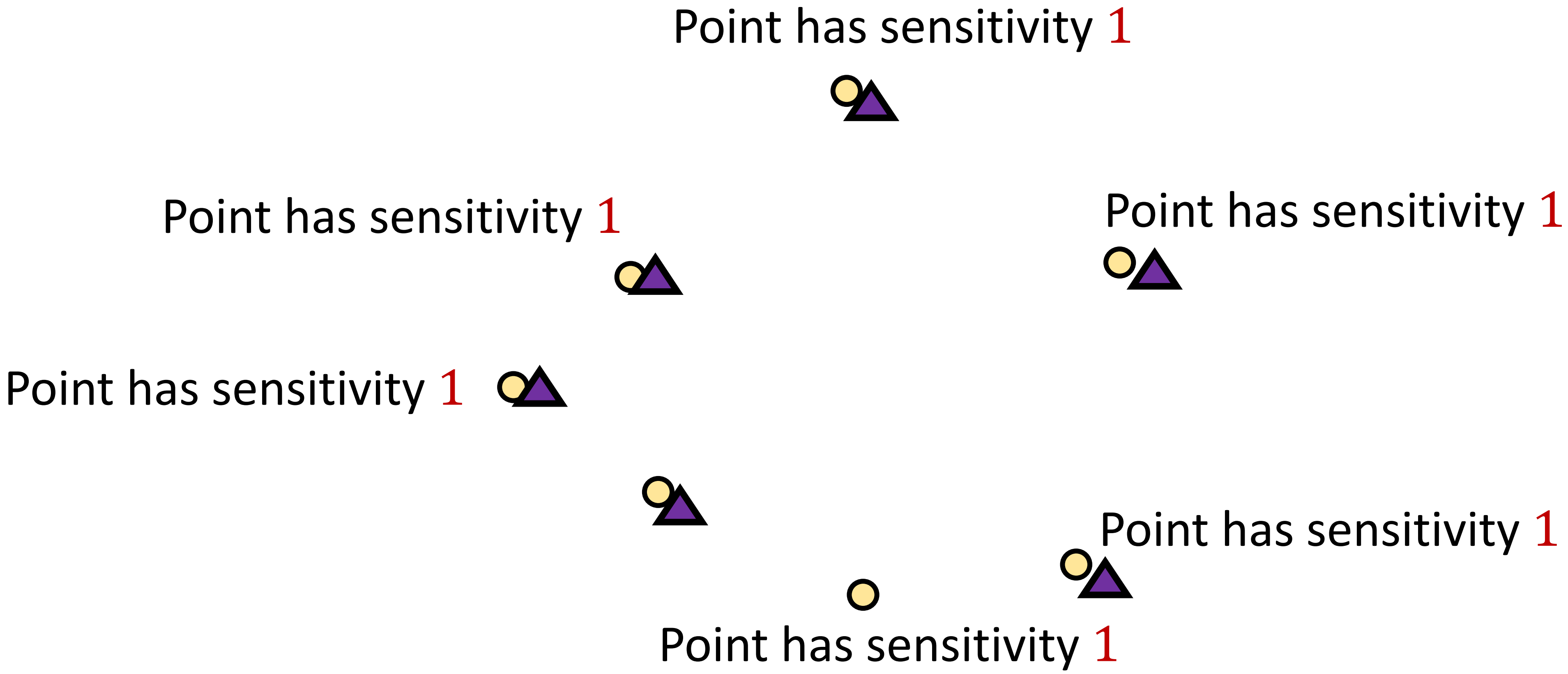
$$s(x_t) = \max_{C:|C|\leq k} \frac{\text{Cost}(x_t, C)}{\text{Cost}(X, C)} = \max_{C:|C|\leq k} \frac{\text{Cost}(x_t, C)}{\sum_{i=1}^{n} \text{Cost}(x_i, C)}$$



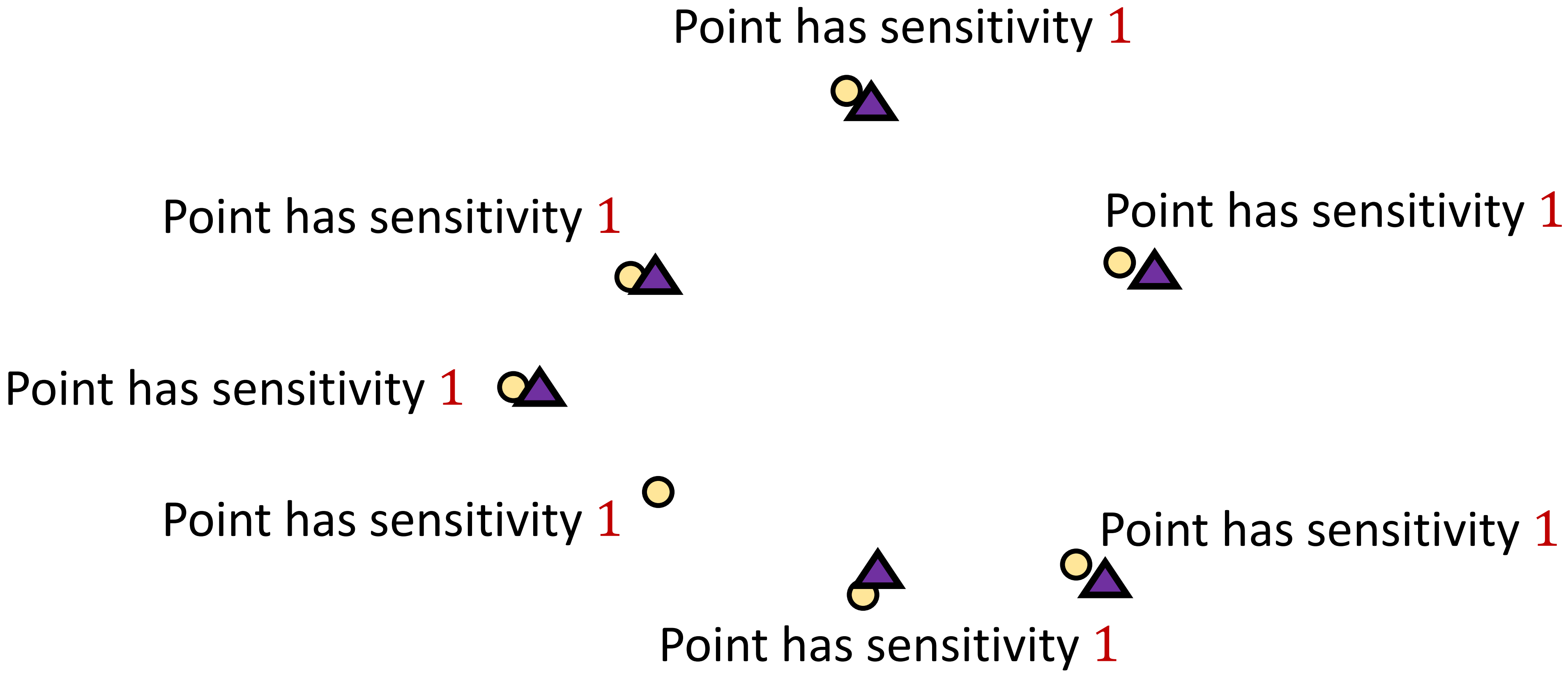
$$s(x_t) = \max_{C:|C|\leq k} \frac{\text{Cost}(x_t, C)}{\text{Cost}(X, C)} = \max_{C:|C|\leq k} \frac{\text{Cost}(x_t, C)}{\sum_{i=1}^{n} \text{Cost}(x_i, C)}$$



$$s(x_t) = \max_{C:|C|\le k} \frac{\text{Cost}(x_t, C)}{\text{Cost}(X, C)} = \max_{C:|C|\le k} \frac{\text{Cost}(x_t, C)}{\sum_{i=1}^{n} \text{Cost}(x_i, C)}$$



$$s(x_t) = \max_{C:|C|\le k} \frac{\text{Cost}(x_t, C)}{\text{Cost}(X, C)} = \max_{C:|C|\le k} \frac{\text{Cost}(x_t, C)}{\sum_{i=1}^{n} \text{Cost}(x_i, C)}$$

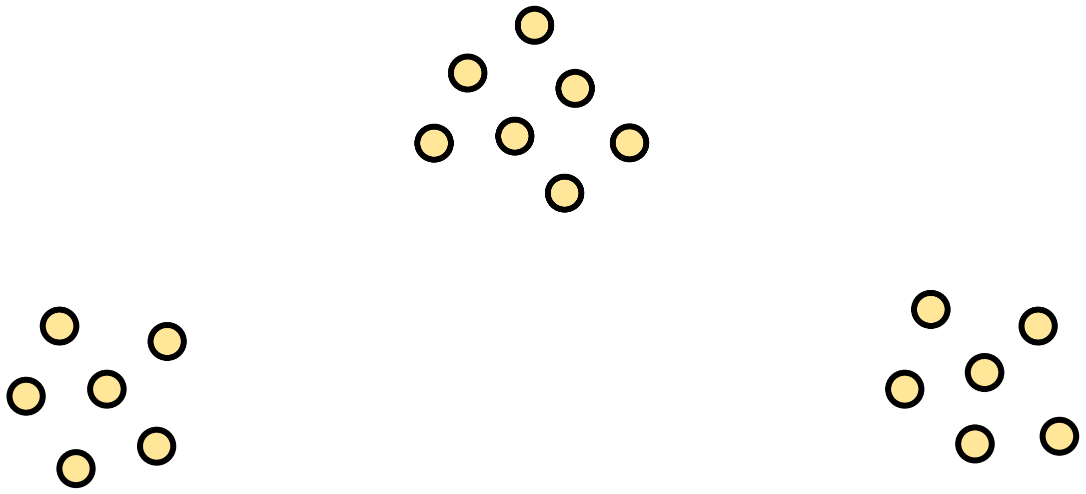


Total Sensitivity

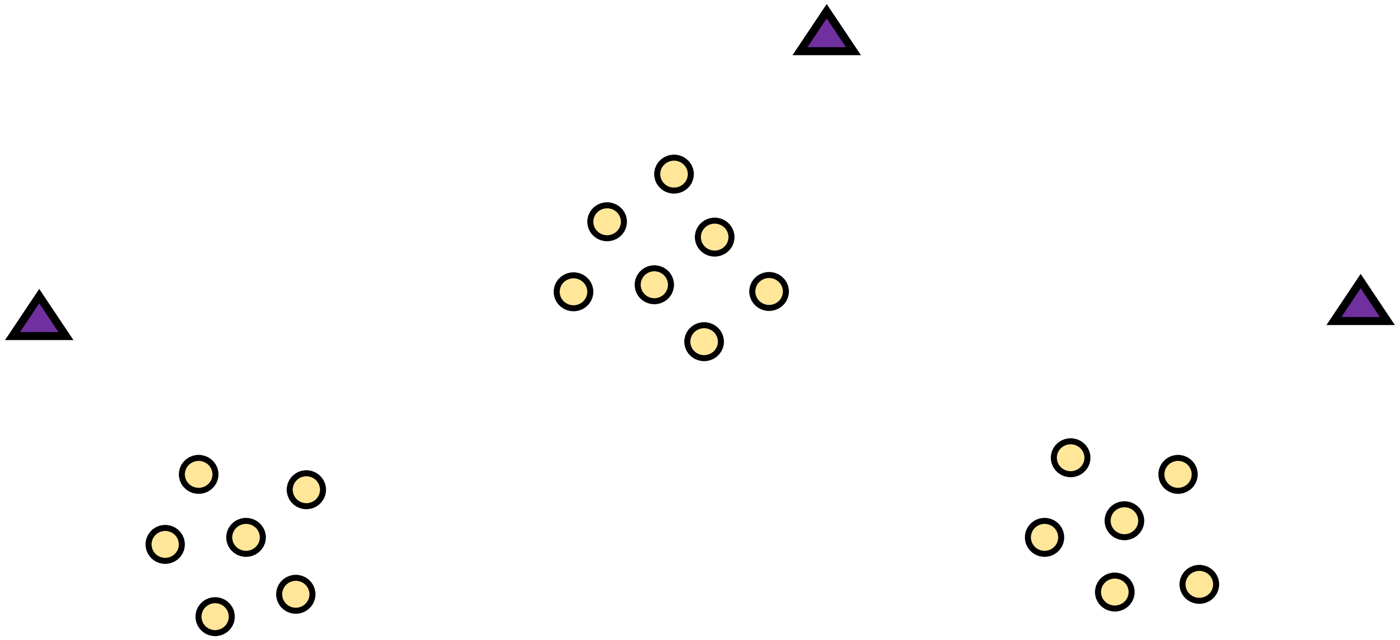
- Total sensitivity = Sum of sensitivities can be at least k

- How large can it be?

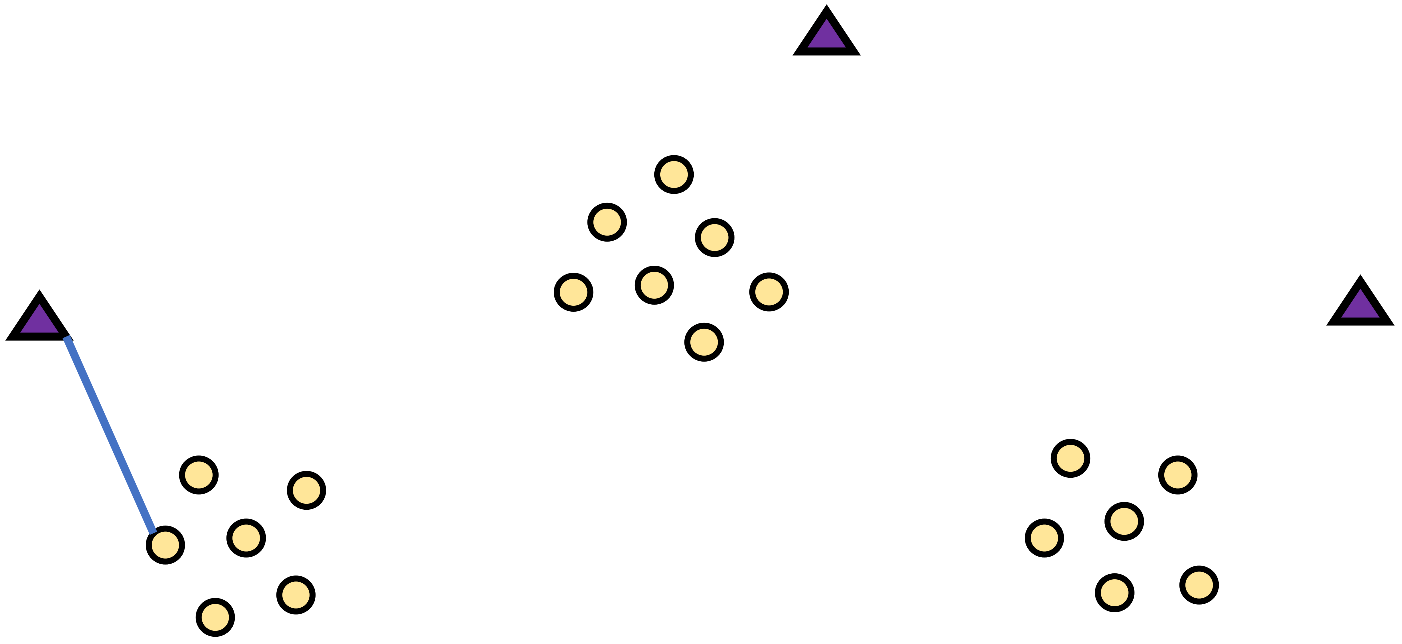
$$s(x_t) = \max_{C:|C|\leq k} \frac{\mathrm{Cost}(x_t, C)}{\mathrm{Cost}(X, C)} = \max_{C:|C|\leq k} \frac{\mathrm{Cost}(x_t, C)}{\sum_{i=1}^{n} \mathrm{Cost}(x_i, C)}$$



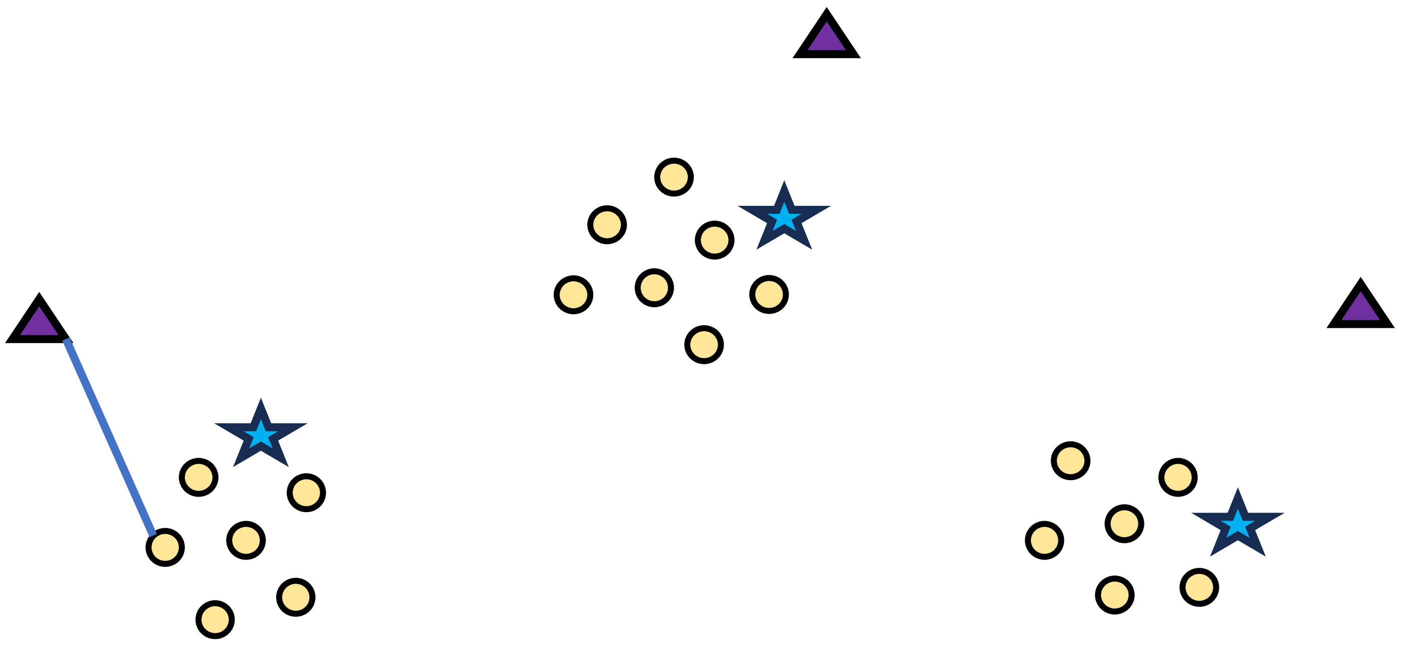
$$s(x_t) = \max_{C:|C|\leq k} \frac{\text{Cost}(x_t, C)}{\text{Cost}(X, C)} = \max_{C:|C|\leq k} \frac{\text{Cost}(x_t, C)}{\sum_{i=1}^{n} \text{Cost}(x_i, C)}$$



$$s(x_t) = \max_{C:|C|\leq k} \frac{\text{Cost}(x_t, C)}{\text{Cost}(X, C)} = \max_{C:|C|\leq k} \frac{\text{Cost}(x_t, C)}{\sum_{i=1}^{n} \text{Cost}(x_i, C)}$$

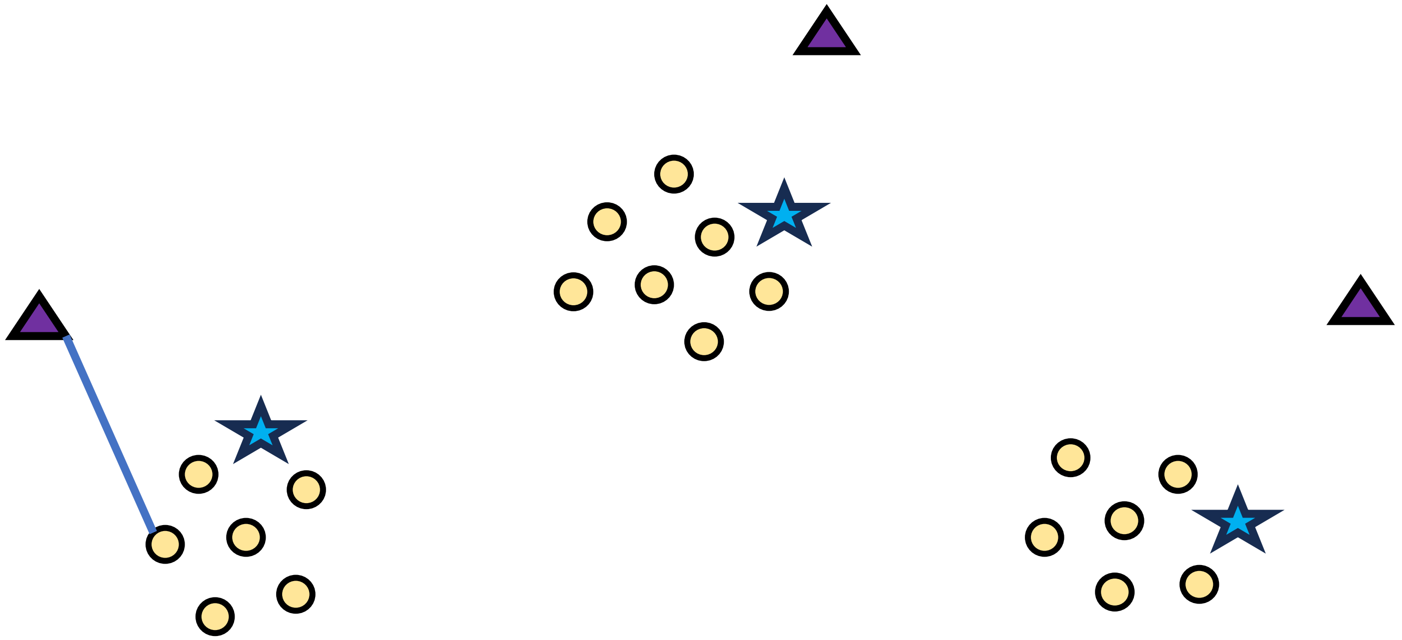


$$s(x_t) = \max_{C:|C|\leq k} \frac{\mathrm{Cost}(x_t, C)}{\mathrm{Cost}(X, C)} = \max_{C:|C|\leq k} \frac{\mathrm{Cost}(x_t, C)}{\sum_{i=1}^{n} \mathrm{Cost}(x_i, C)}$$

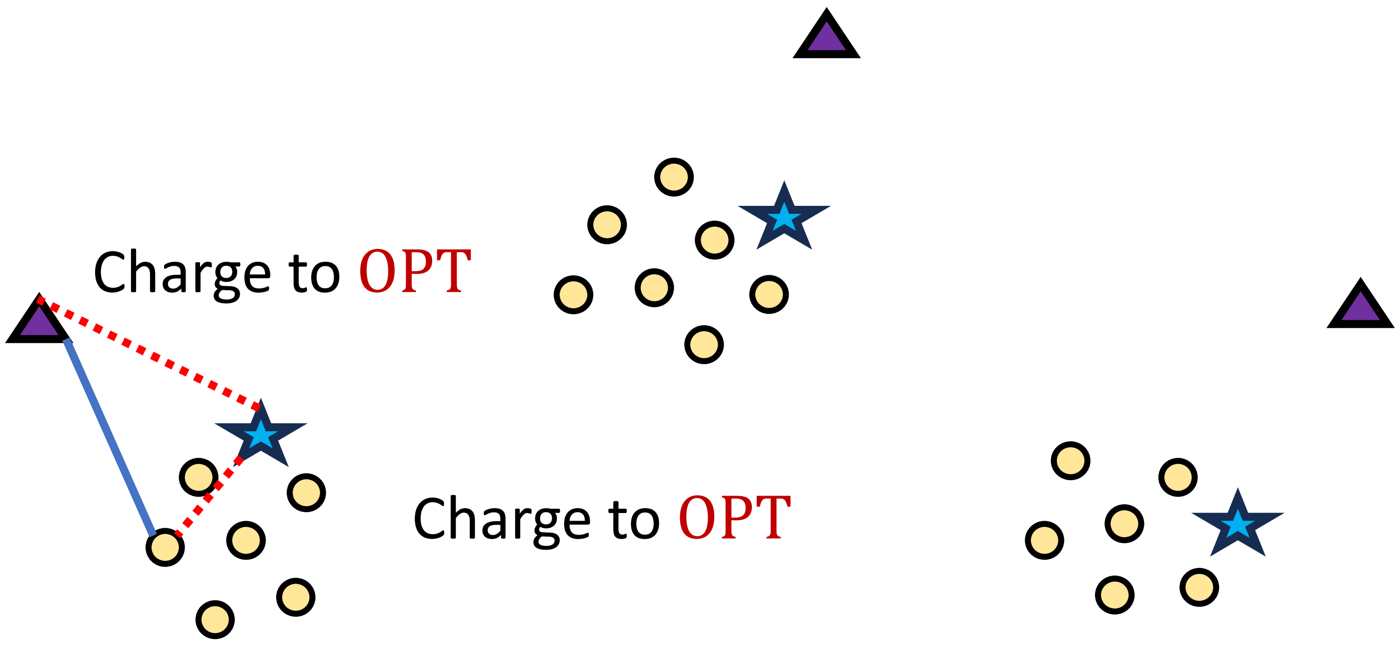


$$s(x_t) = \max_{C:|C|\le k} \frac{\mathrm{Cost}(x_t, C)}{\mathrm{Cost}(X, C)} = \max_{C:|C|\le k} \frac{\mathrm{Cost}(x_t, C)}{\sum_{i=1}^{n} \mathrm{Cost}(x_i, C)}$$

Partition the sum of the sensitivities by each cluster



$$s(x_t) = \max_{C:|C|\le k} \frac{\text{Cost}(x_t, C)}{\text{Cost}(X, C)} = \max_{C:|C|\le k} \frac{\text{Cost}(x_t, C)}{\sum_{i=1}^{n} \text{Cost}(x_i, C)}$$



Total Sensitivity

- Intuition: The sum of the sensitivities in each cluster induced by OPT is at most 1

- Since there are k clusters, the sum of the sensitivities is $O_z(k)$

Putting Things Together

- Recall: $\frac{kd}{\varepsilon^2} \cdot \log \frac{n\Delta}{\varepsilon} \cdot \sum_{x \in X} s(x)$ points sampled
- $\sum_{x \in X} s(x) = O_z(k)$

- In total, roughly $\frac{k^2 d}{\varepsilon^2} \cdot \log \frac{n\Delta}{\varepsilon}$ points sampled in expectation

How to Compute Sensitivities?

- Estimations to sensitivities suffice

- Bicriteria algorithms, e.g., online facility location