k-Means algorithm

How to choose initial cluster representatives?

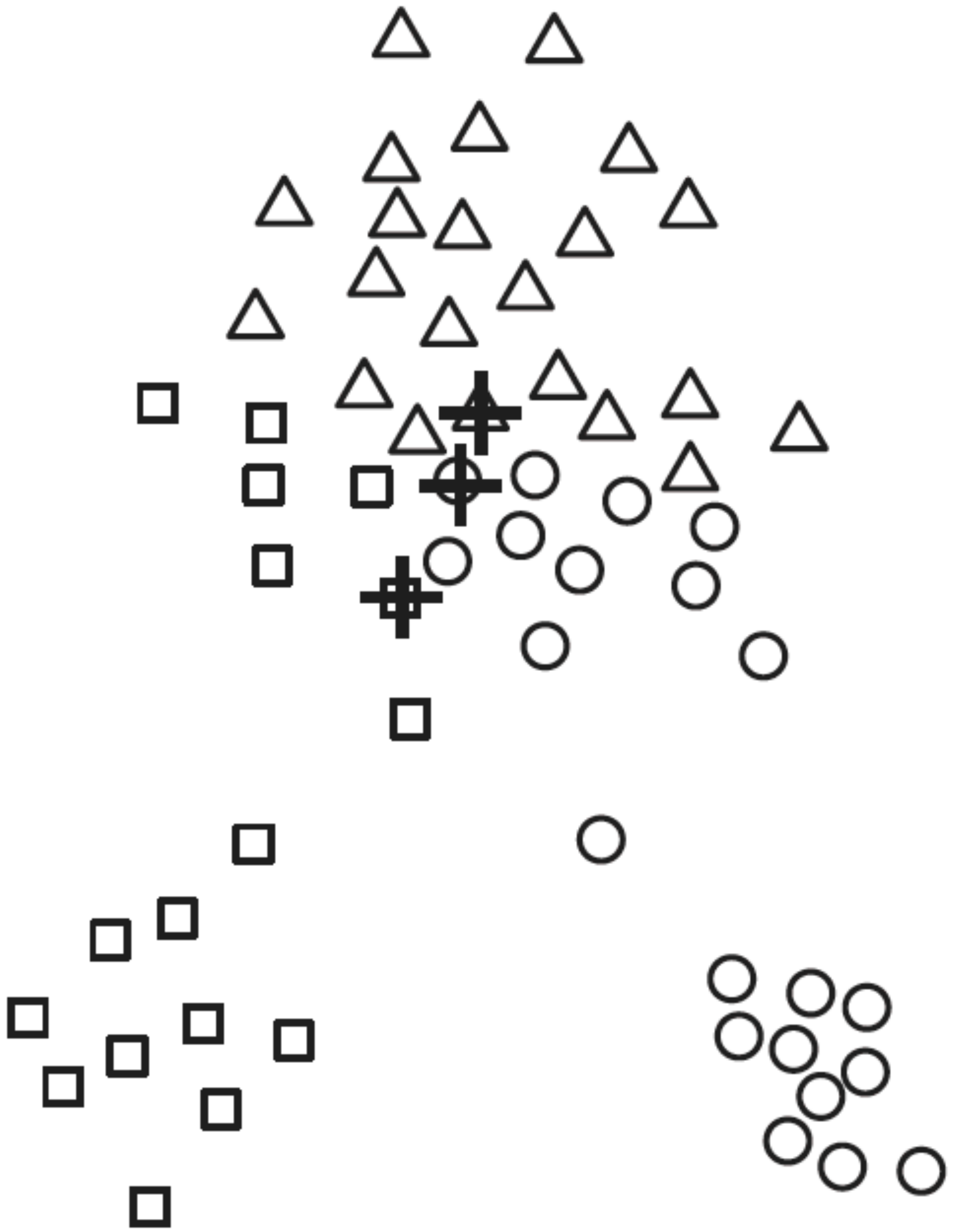
Procheta Sen

UNIVERSITY OF LIVERPOOL

How to choose initial cluster representatives

1. Randomly
2. Randomly repeat several times
3. Sampling + hierarchical clustering
4. Furthest points
5. k-means++

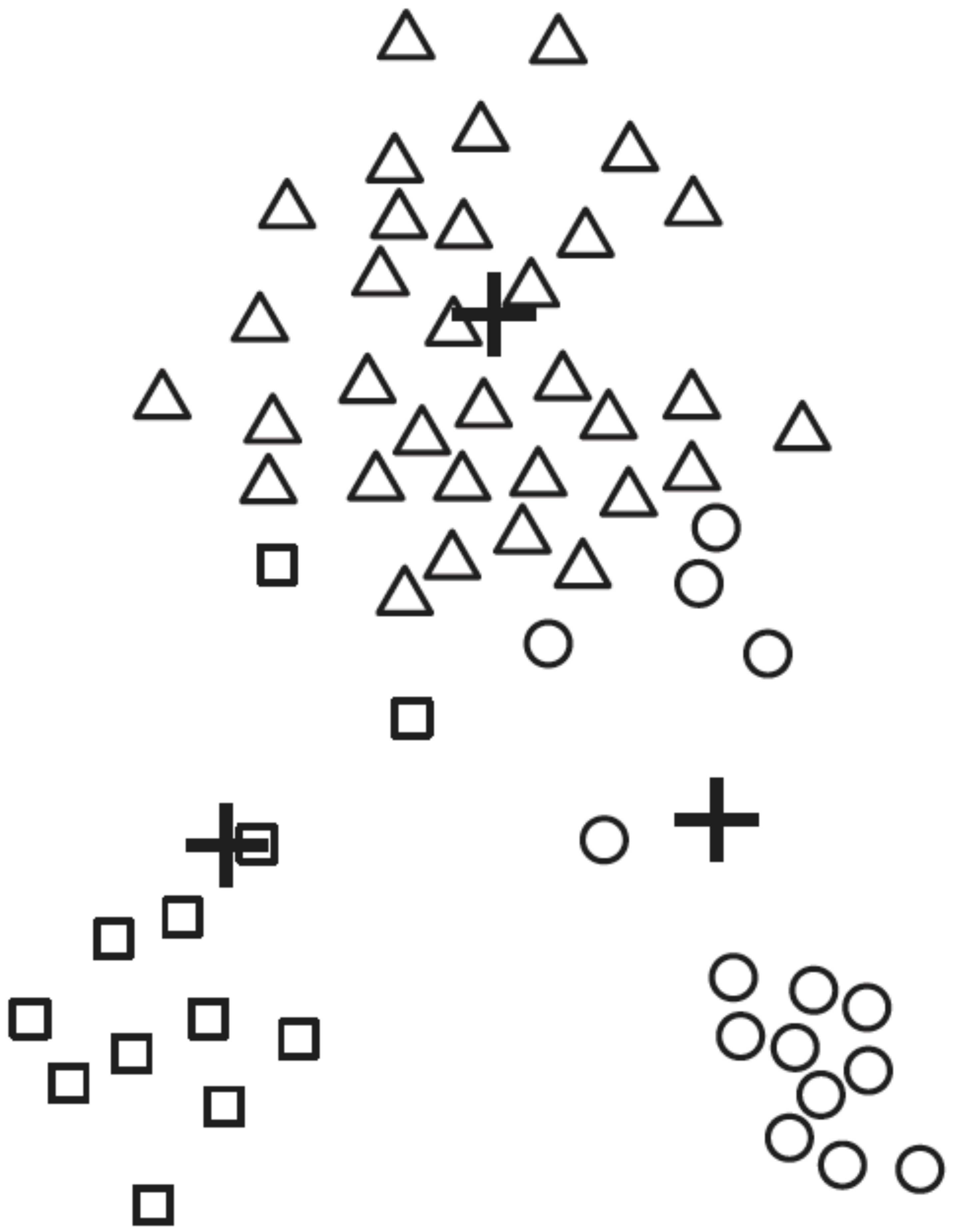
Choosing randomly: lucky choice



Iteration 1

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

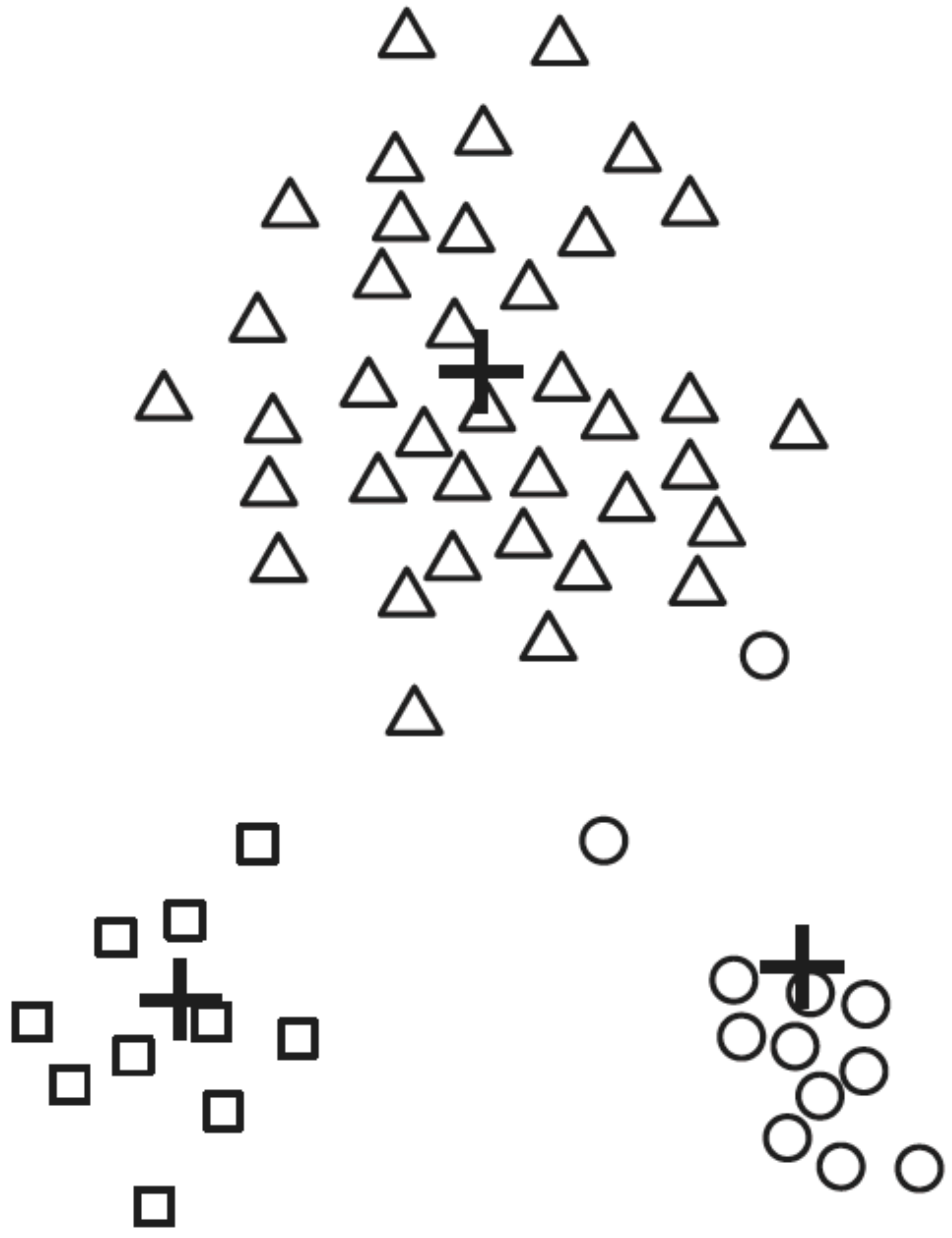
Choosing randomly: lucky choice



Iteration 2

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

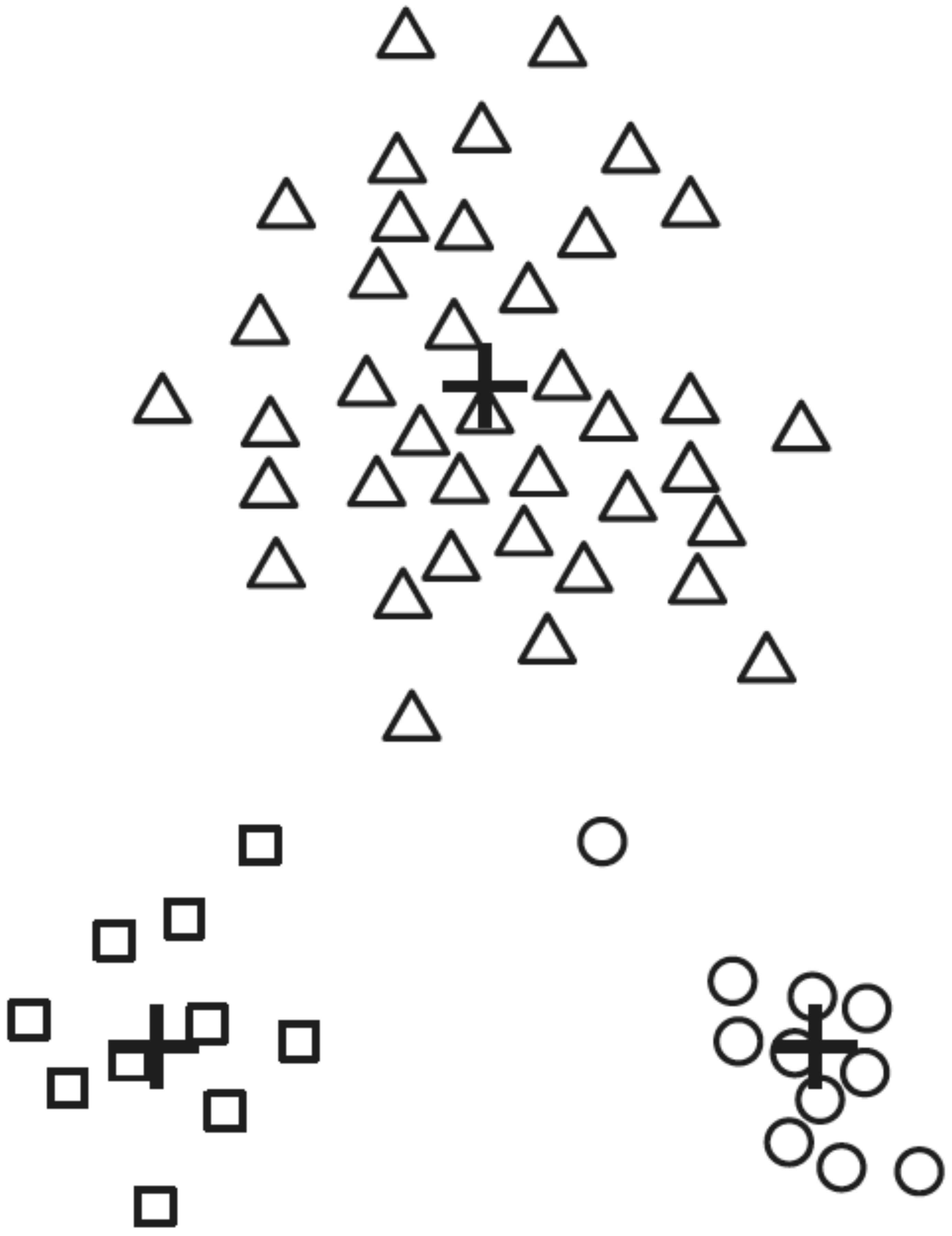
Choosing randomly: lucky choice



Iteration 3

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

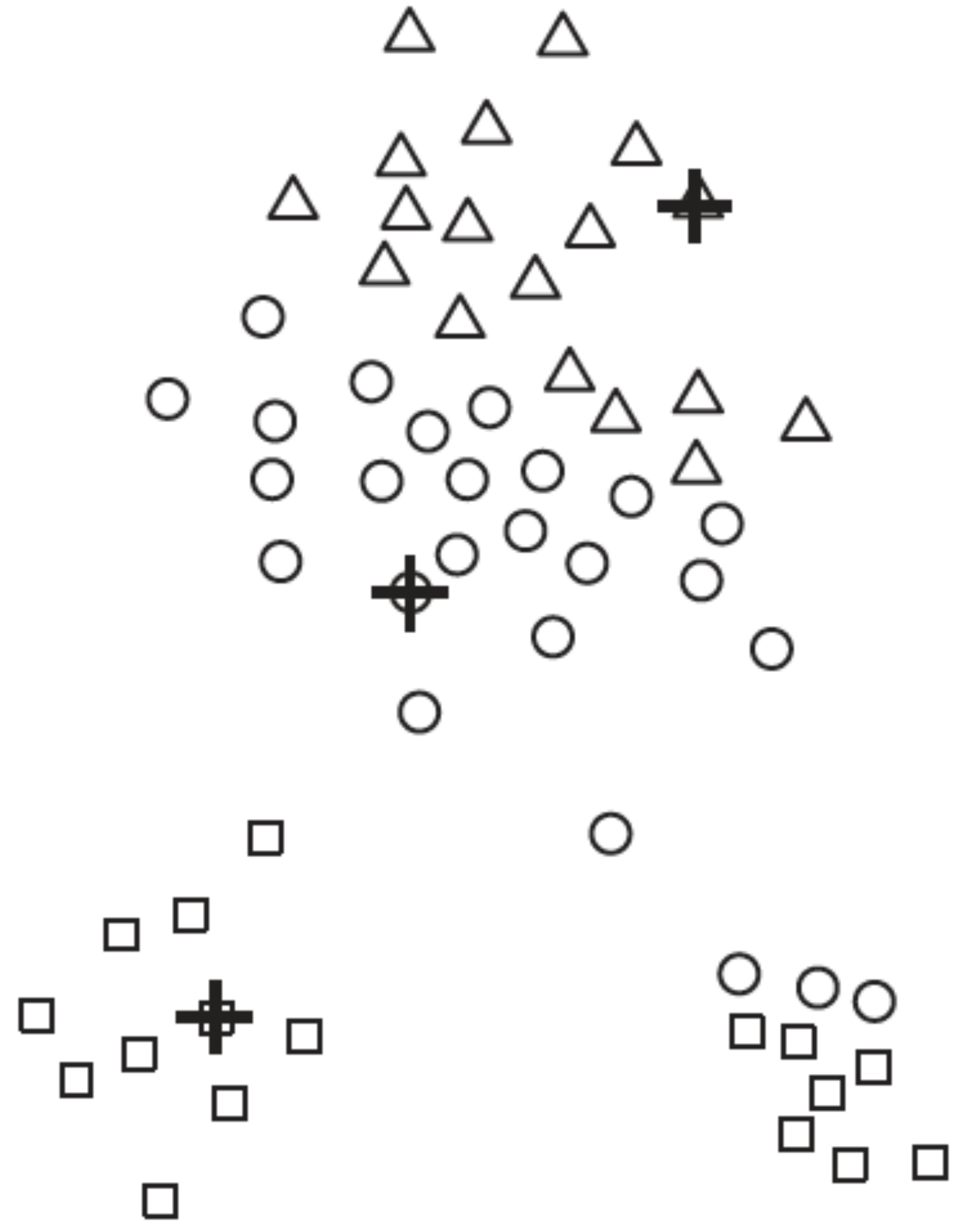
Choosing randomly: lucky choice



Iteration 4

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

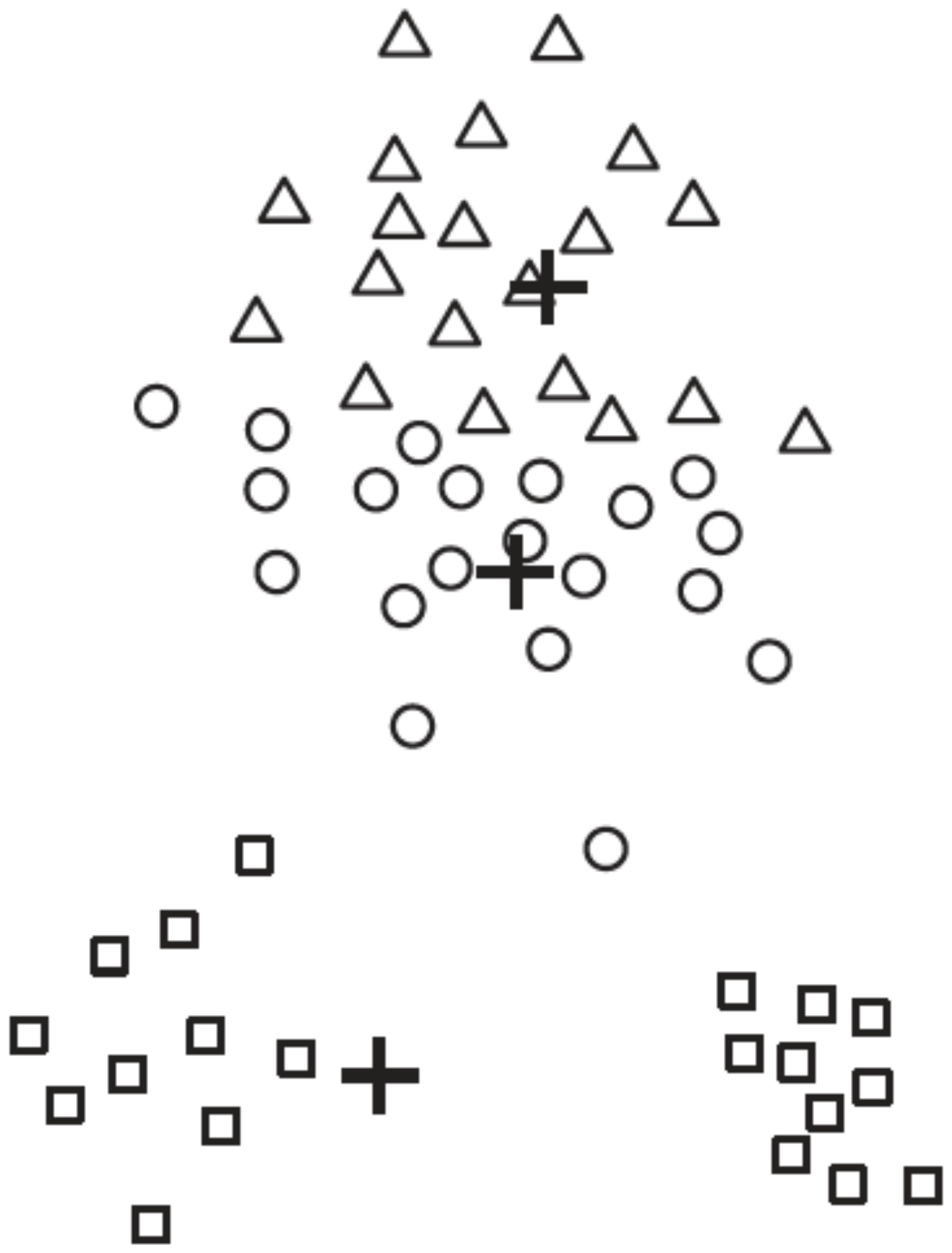
Choosing randomly: **un**lucky choice



Iteration 1

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

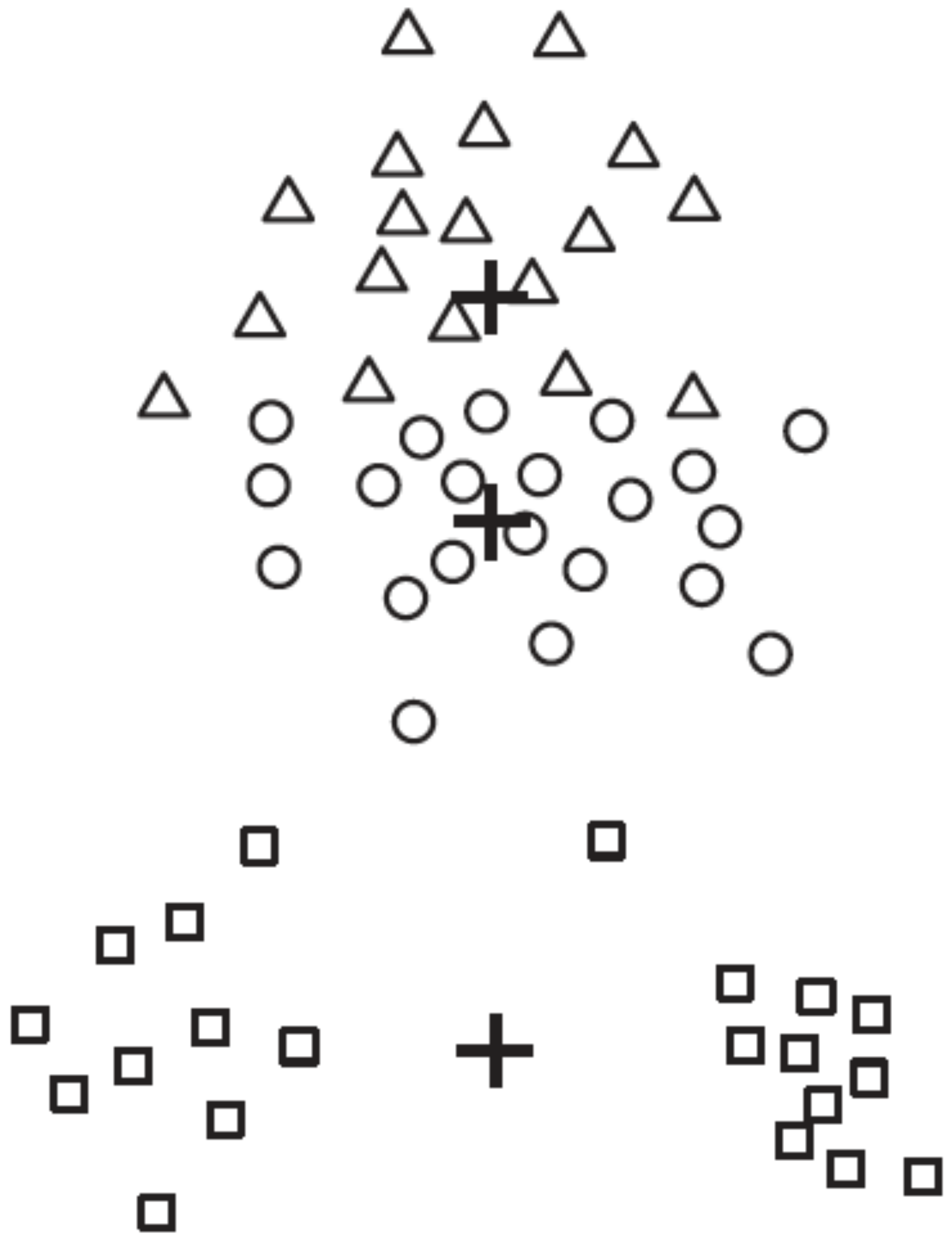
Choosing randomly: **un**lucky choice



Iteration 2

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

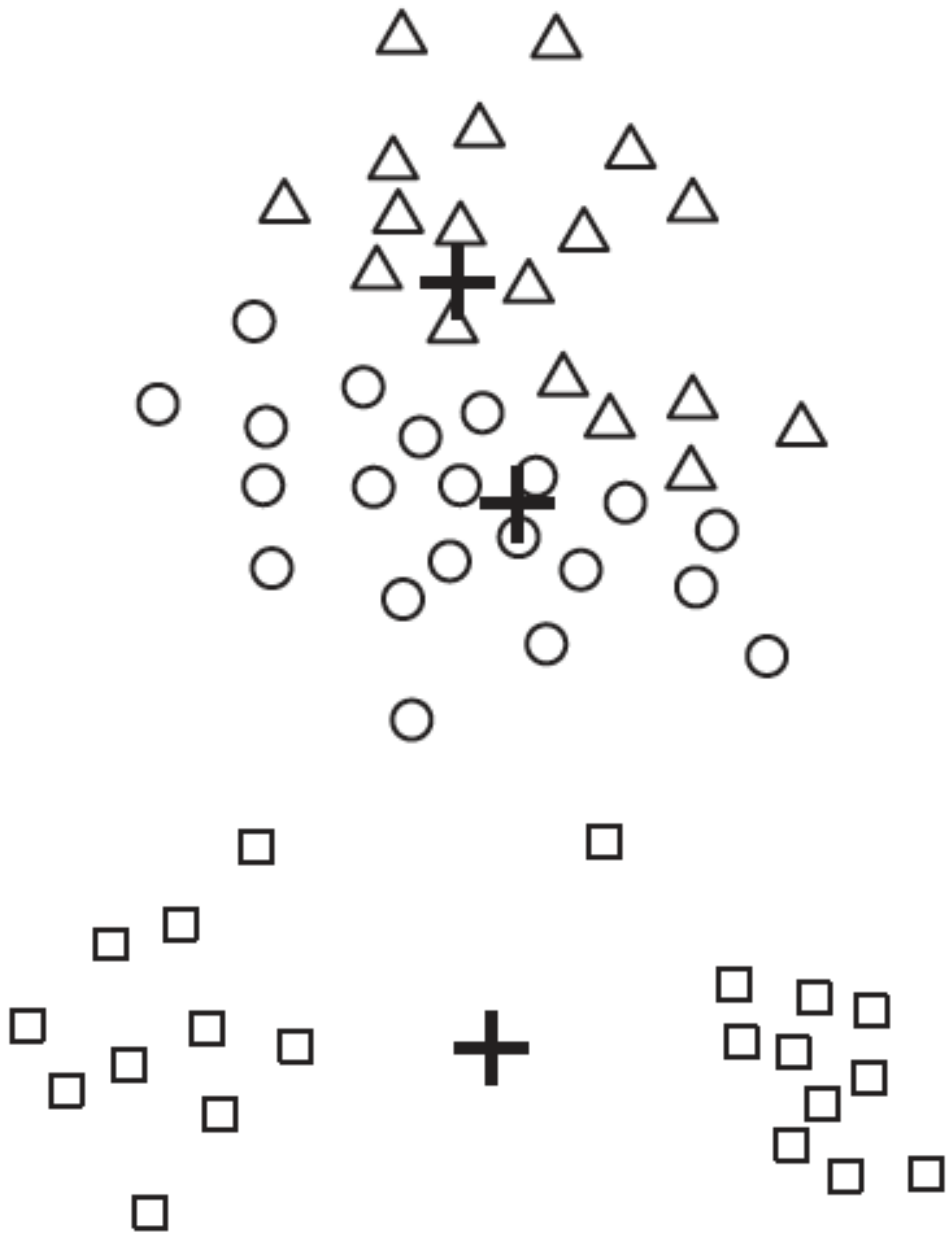
Choosing randomly: **un**lucky choice



Iteration 3

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

Choosing randomly: **un**lucky choice



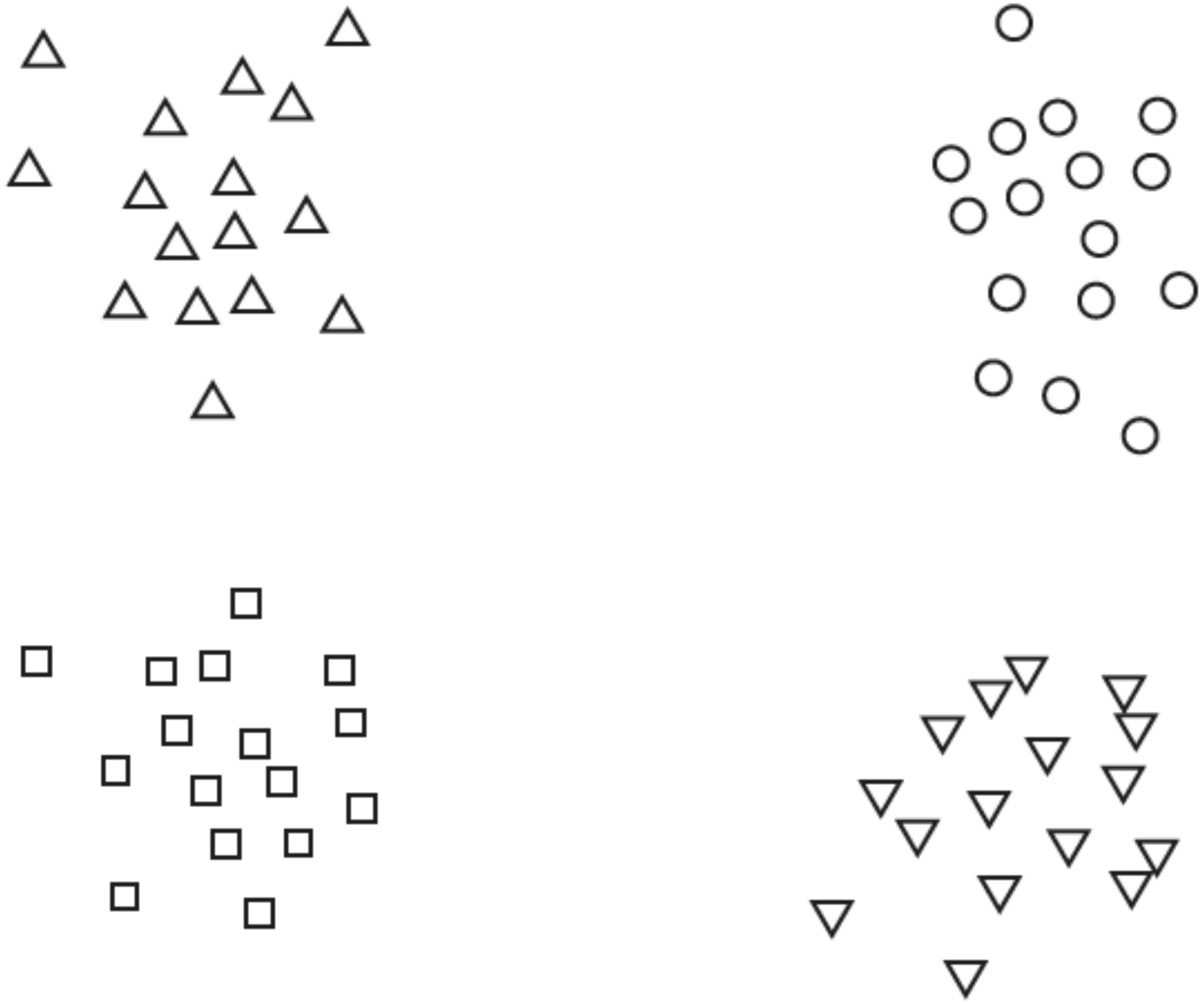
Iteration 4

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

Choosing randomly several times

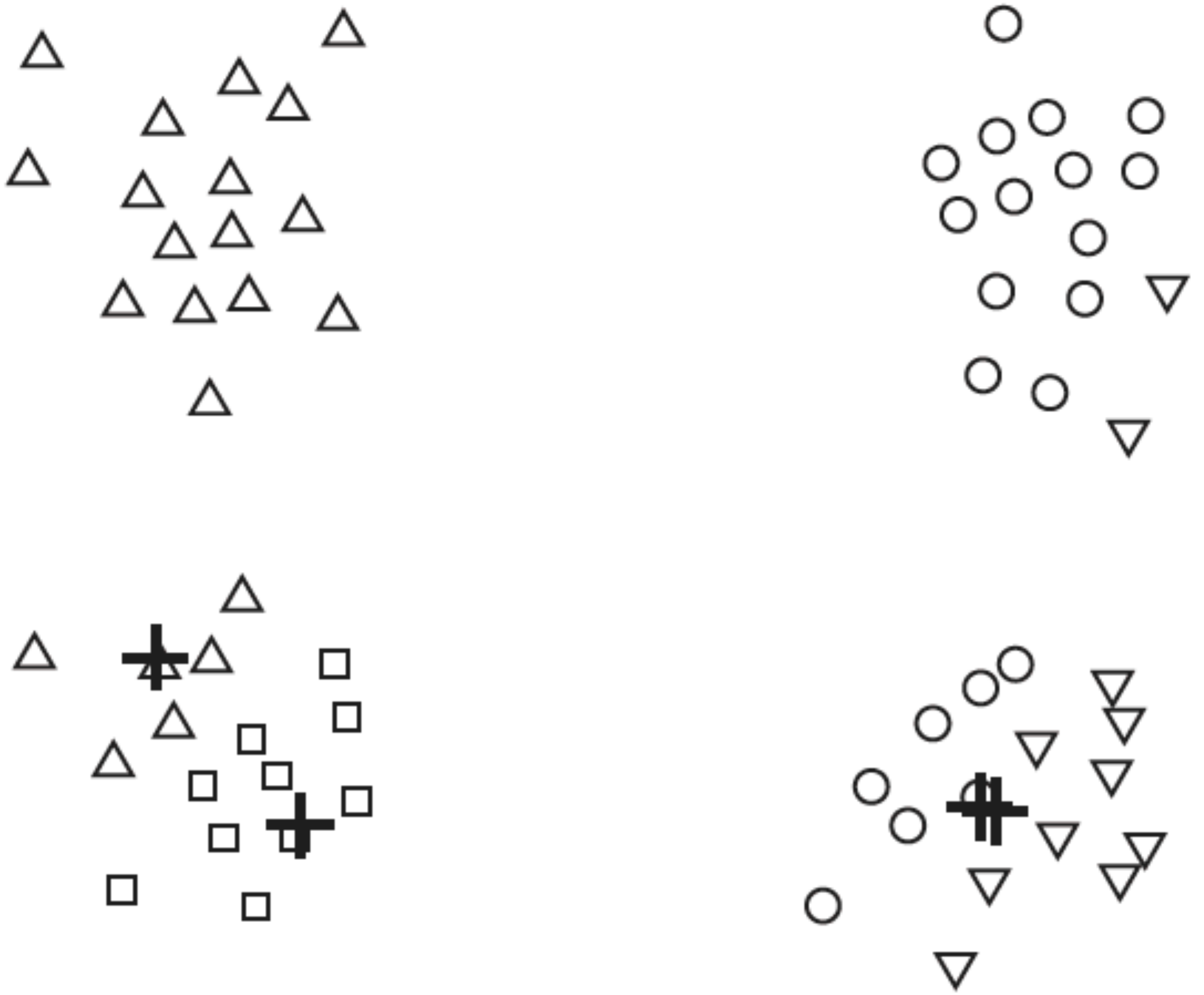
1. Select initial cluster representatives randomly
2. Run k-means
3. Repeat steps 1-2 several times and choose the best clustering

Choosing randomly several times: lucky choice



Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

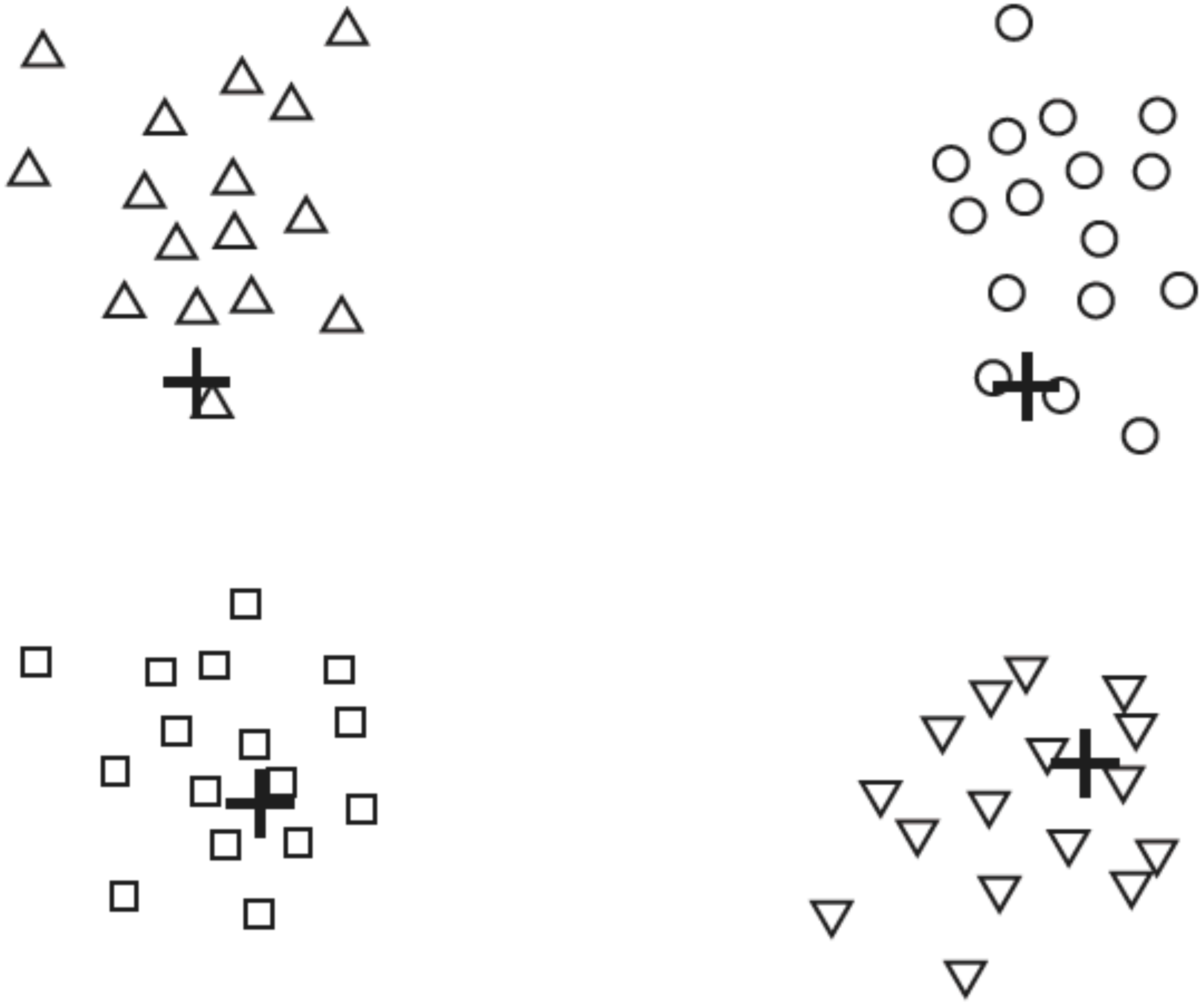
Choosing randomly several times: lucky choice



Iteration 1

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

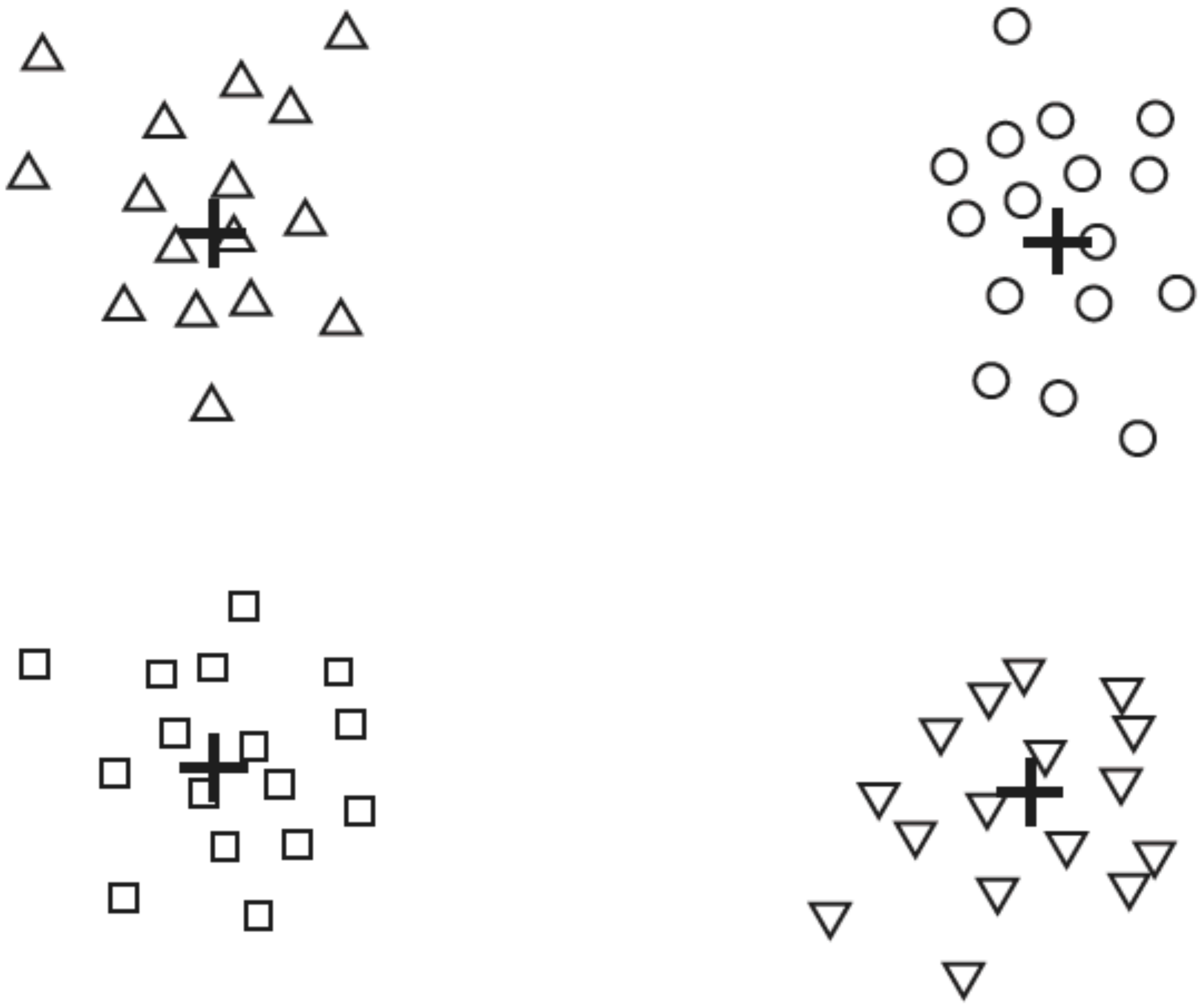
Choosing randomly several times: lucky choice



Iteration 2

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

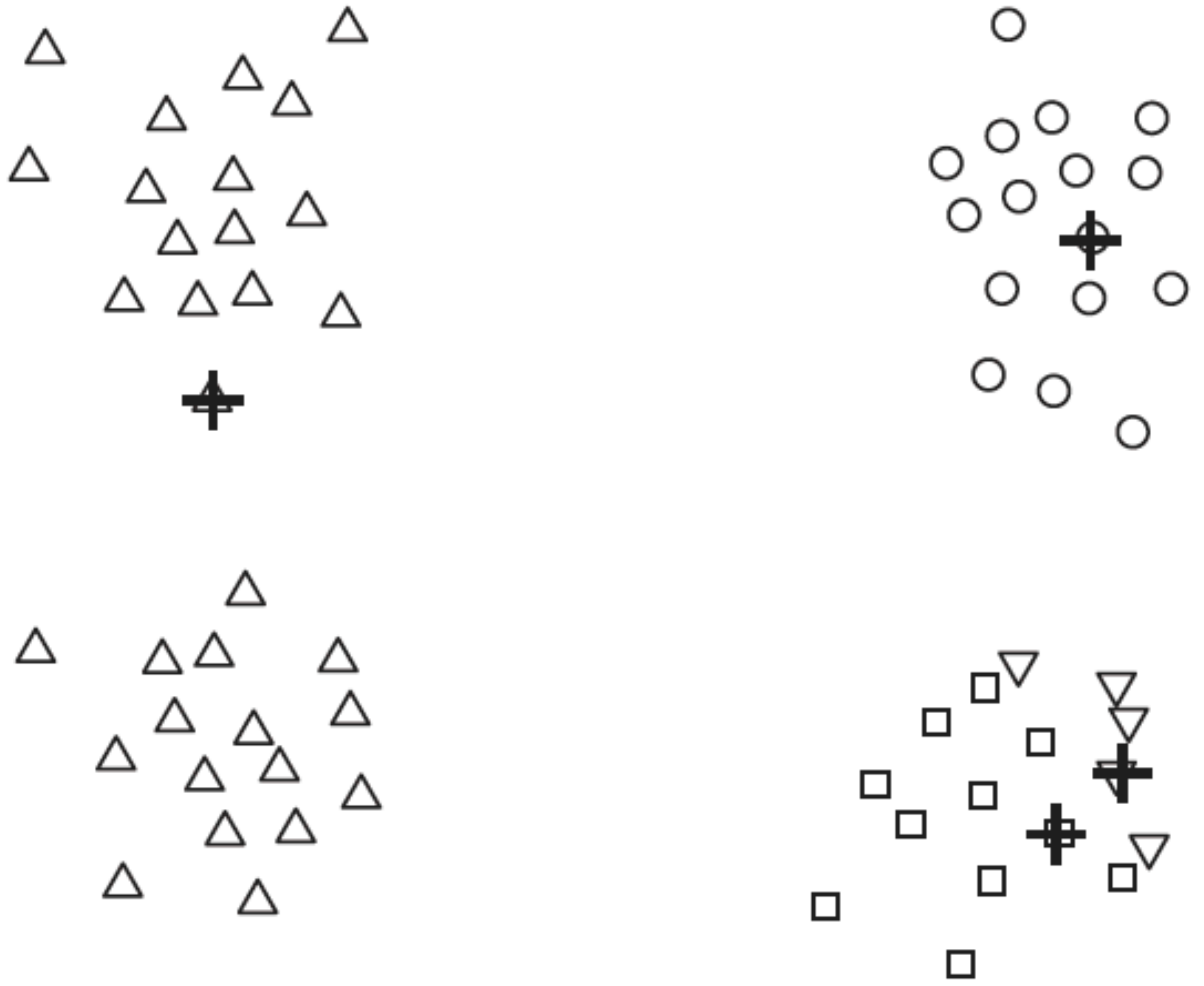
Choosing randomly several times: lucky choice



Iteration 3

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

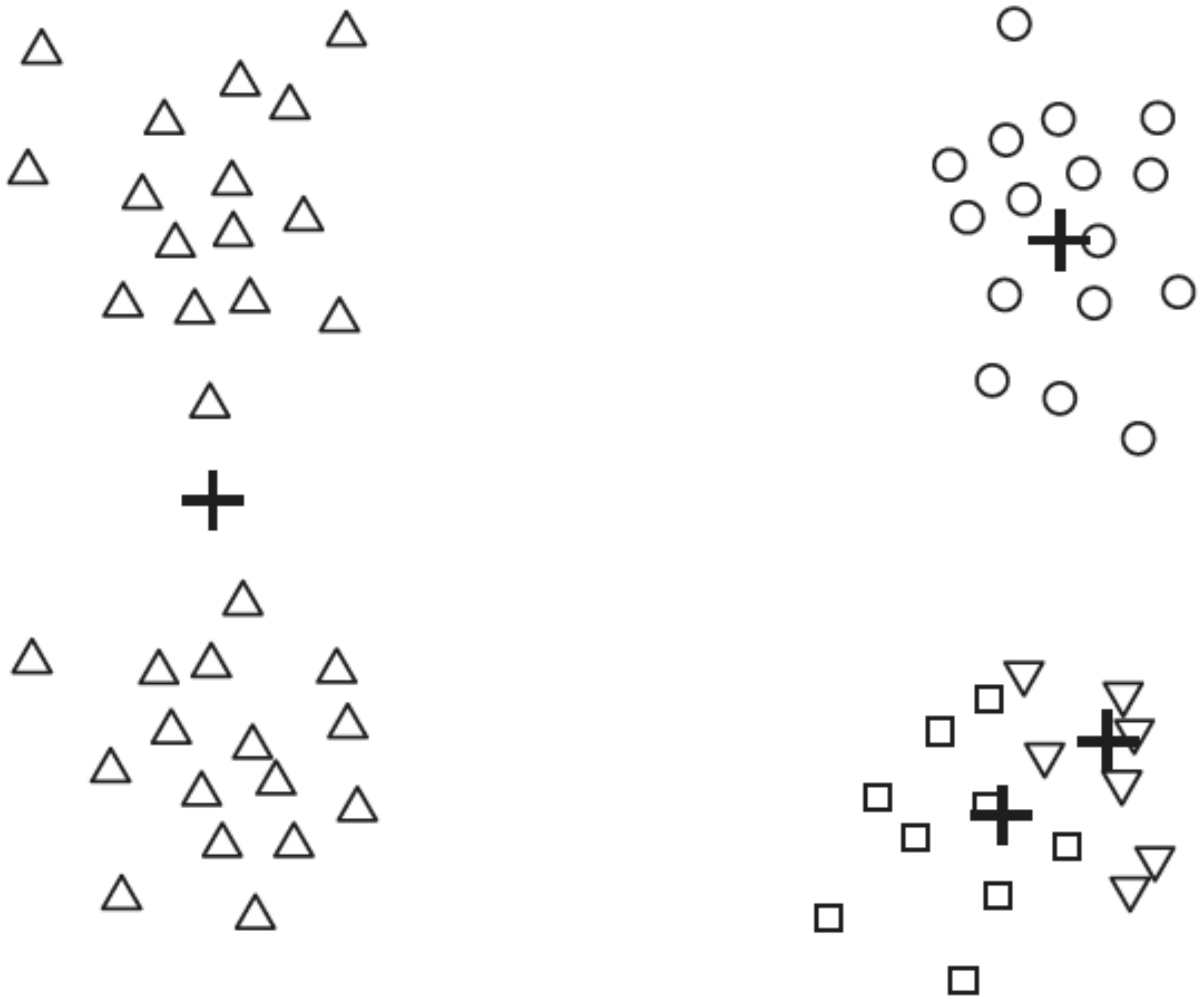
Choosing randomly several times: **un**lucky choice



Iteration 1

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

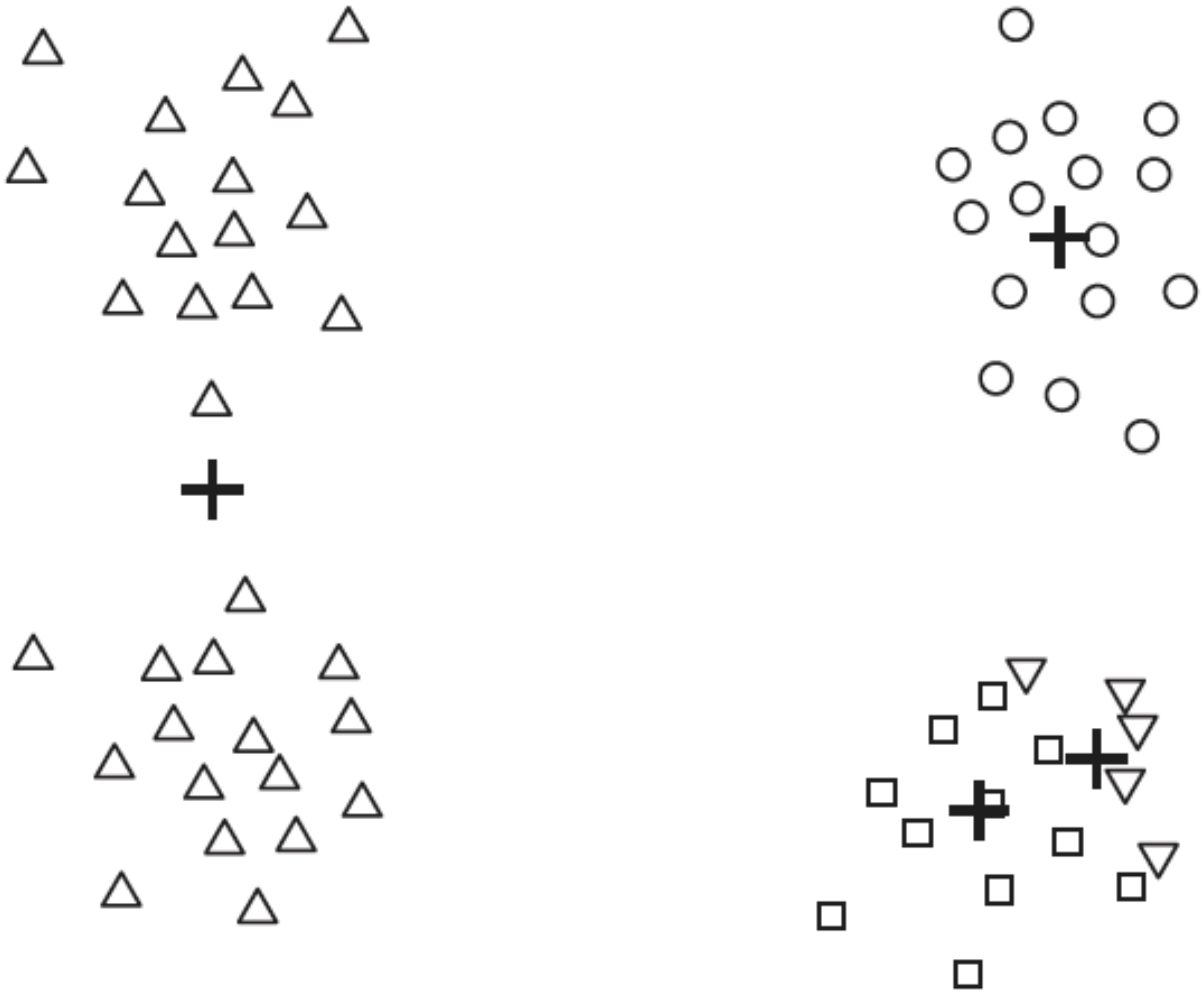
Choosing randomly several times: **un**lucky choice



Iteration 2

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

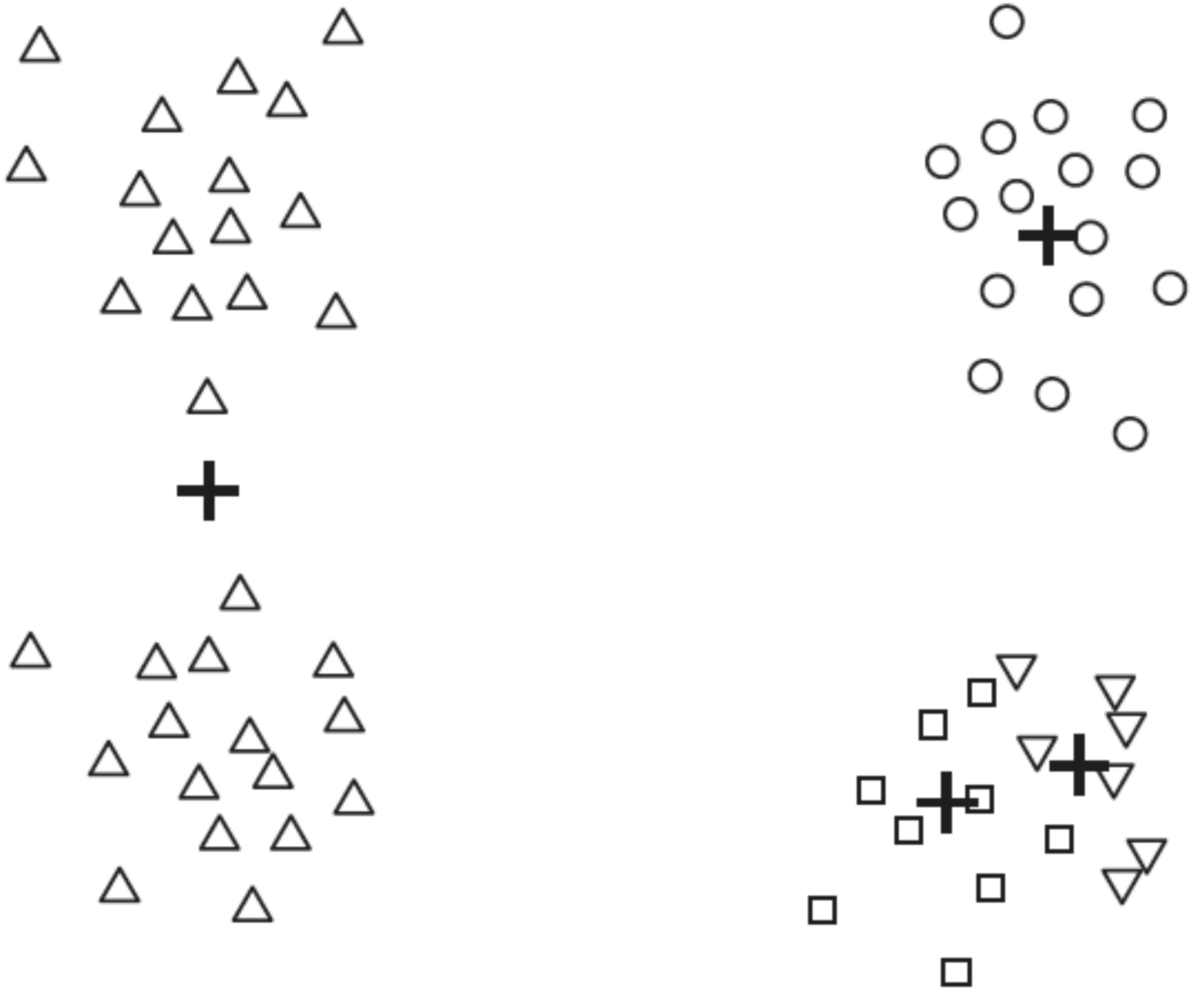
Choosing randomly several times: **un**lucky choice



Iteration 3

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

Choosing randomly several times: **un**lucky choice



Iteration 4

Tan, P., M. Steinbach, and V. Kumar. "Introduction to data mining . Hoboken." (2019).

Sampling + hierarchical clustering

1. Sample subset $\mathscr{D}'$ of points from the dataset $\mathscr{D}$
2. Cluster $\mathscr{D}'$ using a hierarchical clustering technique.
3. Extract k clusters from the hierarchical clustering.
4. Compute the means of these k clusters, and use them as the initial cluster representatives
5. Proceed with the standard k-means with these cluster representatives

Sampling + hierarchical clustering

Often works well, but it is practical only if

1. The sampled subset $\mathscr{D}'$ is relatively small (a few hundred to a few thousand), as hierarchical clustering is expensive

2. k is relatively small compared to the size of the sampled set $\mathscr{D}'$

Selecting furthest points

For an object $\overline{X}$ in the dataset $\mathcal{D}$ let $R(\overline{X})$ be the distance from $\overline{X}$ to the closest cluster representative we have already chosen.

1. Select one representative $\overline{Y}_1$ uniformly at random from $\mathcal{D}$.
2. For every $i = 2,..,k$
    1. Select representative $\overline{Y}_i$ from $\mathcal{D}$ with the maximum value of $R(\cdot)$
3. Proceed with the standard k-means using $\overline{Y}_1, \ldots, \overline{Y}_k$ as initial cluster representatives

Selecting furthest points

The main drawback:
such an approach can select outliers, rather than points in clusters

k-means++

For an object $\overline{X}$ in the dataset $\mathscr{D}$ let $R(\overline{X})$ be the distance from $\overline{X}$ to the closest cluster representative we have already chosen.

1. Select one representative $\overline{Y}_1$ uniformly at random from $\mathscr{D}$.
2. For every $i = 2,..,k$
    1. Select representative $\overline{Y}_i$ from $\mathscr{D}$ with probability $\overline{Y}_i = \overline{X}$ being equal to $\frac{R(\overline{X})^2}{\sum_{\overline{X} \in \mathscr{D}} R(\overline{X})^2}$
3. Proceed with the standard k-means using $\overline{Y}_1, \ldots, \overline{Y}_k$ as initial cluster representatives

S. Vassilvitskii, D. Arthur. "k-means++: The advantages of careful seeding." SODA 2006.

k-means++

Within cluster sum of squares: $\phi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} \|x - c\|^2.$

Theorem 3.1. *If $\mathcal{C}$ is constructed with* **`k-means++`**, *then the corresponding potential function ϕ satisfies,* $E[\phi] \leq 8(\ln k + 2)\phi_{\mathrm{OPT}}$.

S. Vassilvitskii, D. Arthur. "k-means++: The advantages of careful seeding." SODA 2006.

k-means++ **vs** k-means: experimental study

- Synthetic datasets Norm-10 and Norm-25.
  - Chose 10 (respectively 25) “real” centers uniformly at random from the hypercube of side length 500.
  - Then added points from a Gaussian distribution of variance 1, centered at each of the real centers.
  - Thus, we have a number of well separated Gaussians with the the real centers providing a good approximation to the optimal clustering.

S. Vassilvitskii, D. Arthur. "k-means++: The advantages of careful seeding." SODA 2006.

k-means++ **vs** k-means: experimental study

- Synthetic datasets Norm-10.

k	Average ϕ		Minimum ϕ	
	`k-means`	`k-means++`	`k-means`	`k-means++`
10	10898	5.122	2526.9	5.122
25	787.992	4.46809	4.40205	4.41158
50	3.47662	3.35897	3.40053	3.26072

Table 1: Experimental results on the *Norm-10* dataset (n = 10000, d = 5)

Within cluster sum of squares: $\phi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} \|x - c\|^2.$

S. Vassilvitskii, D. Arthur. "k-means++: The advantages of careful seeding." SODA 2006.

k-means++ **vs** k-means: experimental study

- Synthetic datasets Norm-25.

	Average ϕ		Minimum ϕ	
k	`k-means`	`k-means++`	`k-means`	`k-means++`
10	135512	126433	119201	111611
25	48050.5	15.8313	25734.6	15.8313
50	5466.02	14.76	14.79	14.73

Table 2: Experimental results on the *Norm-25* dataset (n = 10000, d = 15)

Within cluster sum of squares: $\phi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} \|x - c\|^2.$

S. Vassilvitskii, D. Arthur. "k-means++: The advantages of careful seeding." SODA 2006.

k-means++ **vs** k-means: experimental study

- The synthetic datasets: the k-means method does not perform well, because
  - the random seeding will inevitably merge clusters together, and the algorithm will never be able to split them apart.
  - the careful seeding method of k-means++ avoids this problem altogether, and it almost always attains the optimal results on the synthetic datasets

S. Vassilvitskii, D. Arthur. "k-means++: The advantages of careful seeding." SODA 2006.

k-means++ **vs** k-means: experimental study

- Real datasets: Cloud dataset

	Average ϕ		Minimum ϕ	
k	`k-means`	`k-means++`	`k-means`	`k-means++`
10	7553.5	6151.2	6139.45	5631.99
25	3626.1	2064.9	2568.2	1988.76
50	2004.2	1133.7	1344	1088

Table 3: Experimental results on the *Cloud* dataset (n = 1024, d = 10)

Within cluster sum of squares: $\phi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} \|x - c\|^2.$

S. Vassilvitskii, D. Arthur. "k-means++: The advantages of careful seeding." SODA 2006.

k-means++ **vs** k-means: experimental study

- Real datasets: Intrusion dataset

k	Average ϕ		Minimum ϕ	
	k-means	k-means++	k-means	k-means++
10	$3.45 \cdot 10^8$	$2.31 \cdot 10^7$	$3.25 \cdot 10^8$	$1.79 \cdot 10^7$
25	$3.15 \cdot 10^8$	$2.53 \cdot 10^6$	$3.1 \cdot 10^8$	$2.06 \cdot 10^6$
50	$3.08 \cdot 10^8$	$4.67 \cdot 10^5$	$3.08 \cdot 10^8$	$3.98 \cdot 10^5$

Table 4: Experimental results on the *Intrusion* dataset (n = 494019, d = 35)

Within cluster sum of squares: $\phi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} \|x - c\|^2.$

S. Vassilvitskii, D. Arthur. "k-means++: The advantages of careful seeding." SODA 2006.

k-means++ **vs** k-means: experimental study

- The real-world datasets
- On the **Cloud** dataset, k-means++ terminates almost **twice** as fast while achieving potential function (within cluster sum of squares) values about **20% better**.
- On the larger **Intrusion** dataset: the potential value obtained by k-means++ is better by factors of **10 to 1000** and is also obtained **up to 70% faster**.

S. Vassilvitskii, D. Arthur. "k-means++: The advantages of careful seeding." SODA 2006.