How *Slow* is the k-means Method?

Sariel Har-Peled

Bardia Sadri

UIUC, Urbana, IL

1: Who is the terrorist?

1: Who is the terrorist?

Bardia Sadri

Sariel Har-Peled

2: Geometric Clustering

- **Input:** A $P \subseteq \mathbb{R}^d$, k.
- Partition P into k "good" clusters.
- k**-Median**: $\min_C \sum_{p \in P} \mathrm{dist}(p, C)$

  k**-Means**: $\min_C \sum_{p \in P} \Big(\mathrm{dist}(p, C)\Big)^2$

  $\mathrm{dist}(p, C) = \min_{c \in C} \|pc\|$.

3: k-Median clustering

- **k-Median $(1+\varepsilon)$-aprx:**
  - Low dim: **[Arora et al. (1998)]**, **[Kolliopoulos and Rao (1999)]**...
    $O\left(n + \rho k^{O(1)} \log^{O(1)} n\right)$
    ρ - func. of ε, d
  - High dim: **[Bădoiu et al. (2002); Kumar et al. (2004)]**
    $O(\tau \cdot nd)$: linear time
    τ - function of ε, k

4: k-Means clustering

- **k-Median $(1+\varepsilon)$-aprx:**
  - Low dim: **[Matoušek (2000)]**...
    $O(n + \mathrm{poly}(k, \log n, 1/\varepsilon) + \mathrm{func}(k, \varepsilon))$
  - High dim: **[de la Vega et al. (2003); Kumar et al. (2004)]**
    $O(\tau \cdot nd)$: linear time
    τ - function of ε, k
- Algorithms are useless in practice.
- There is a simple hueristic for k-means!

5: k-Means method names

- k-means algorithm.
- k-means method.
- k-means.
- Lloyd's k-means method.
- k-means heuristic.
- Axis of evil.

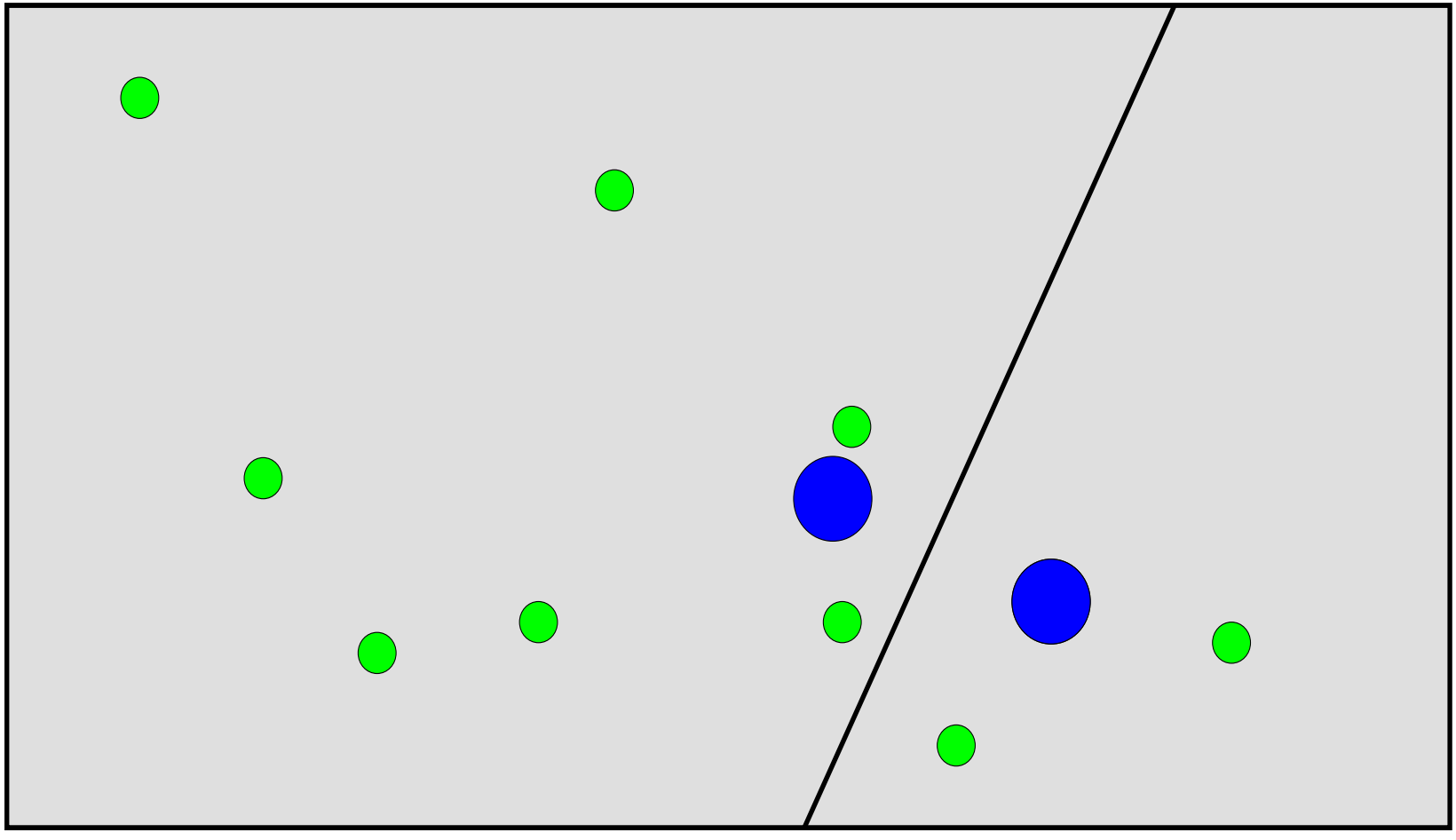
6: k-Means method

- C - set of centers
- $\text{Price}_C(P) = \sum_{p \in P} (\text{dist}(p, C))^2$
- **Observation:** If center c serves a cluster $Q \Rightarrow$ min price when $c =$ center of mass of Q.
- **Observation:** $p \in P$ then p uses NN in C.

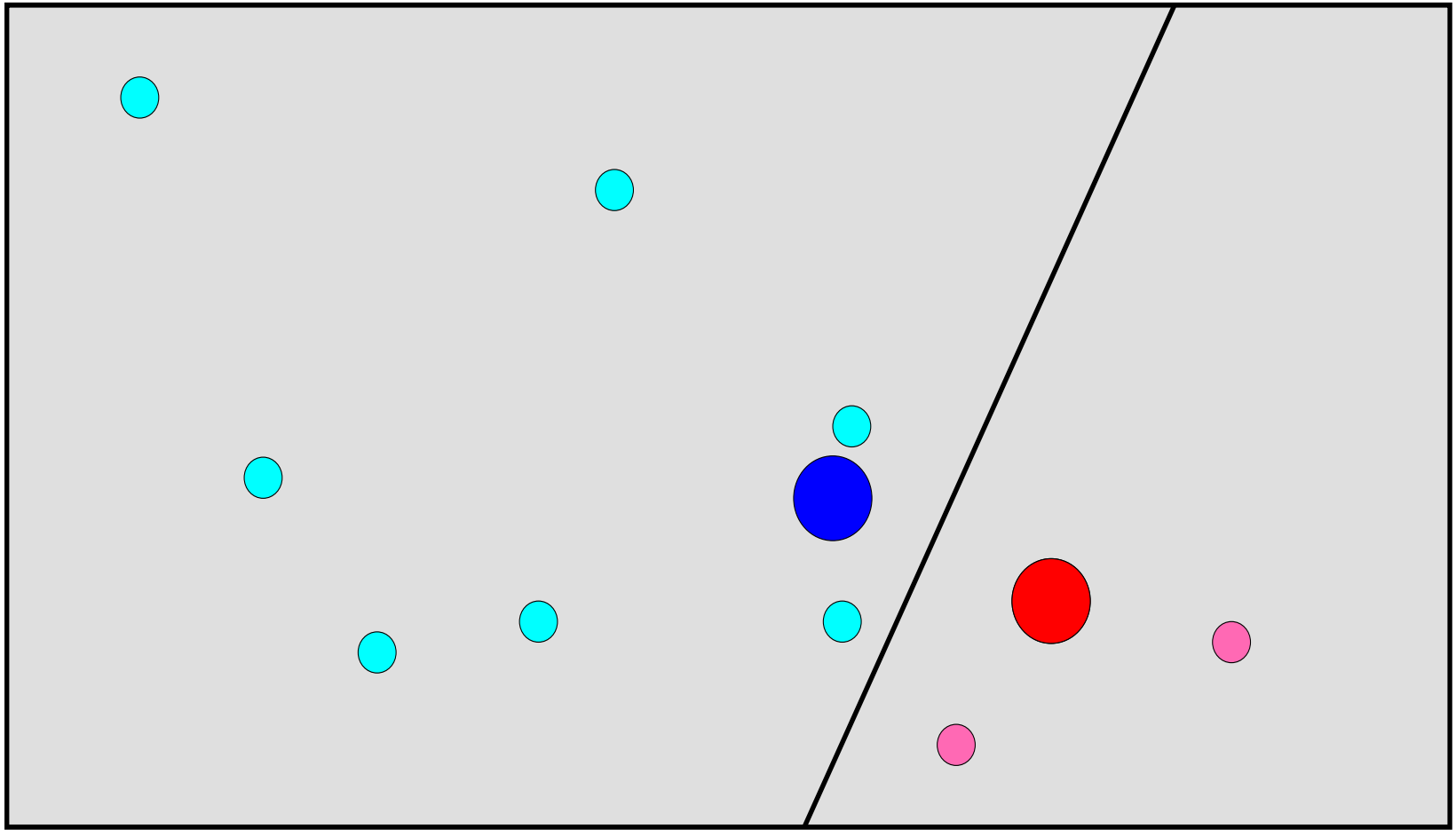
k-means method:

- Partition P into clusters using C
- Compute centers of mass of every cluster
- Set C to be new set of centers. Repeat.

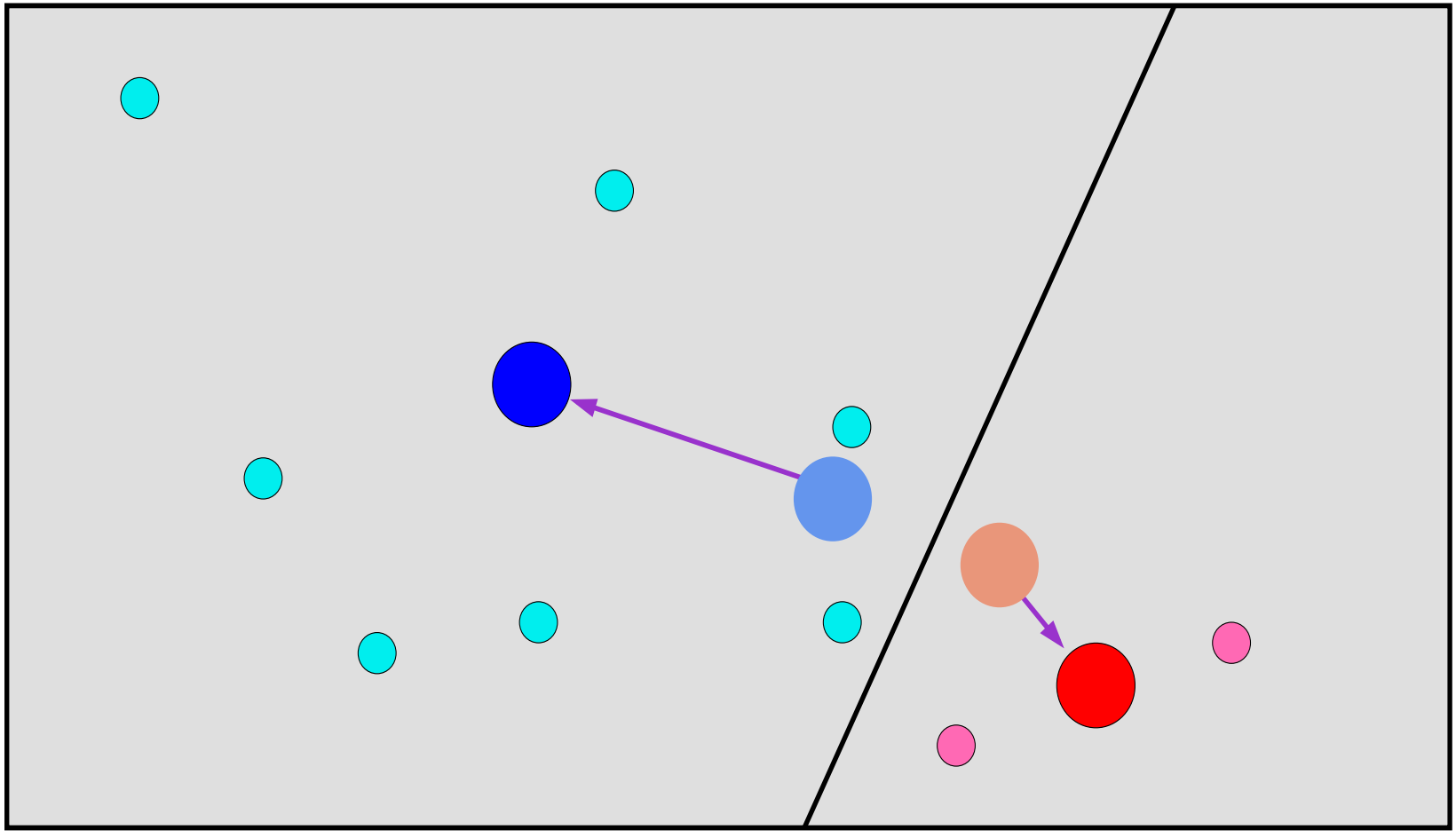
7: k-Means method - Demo



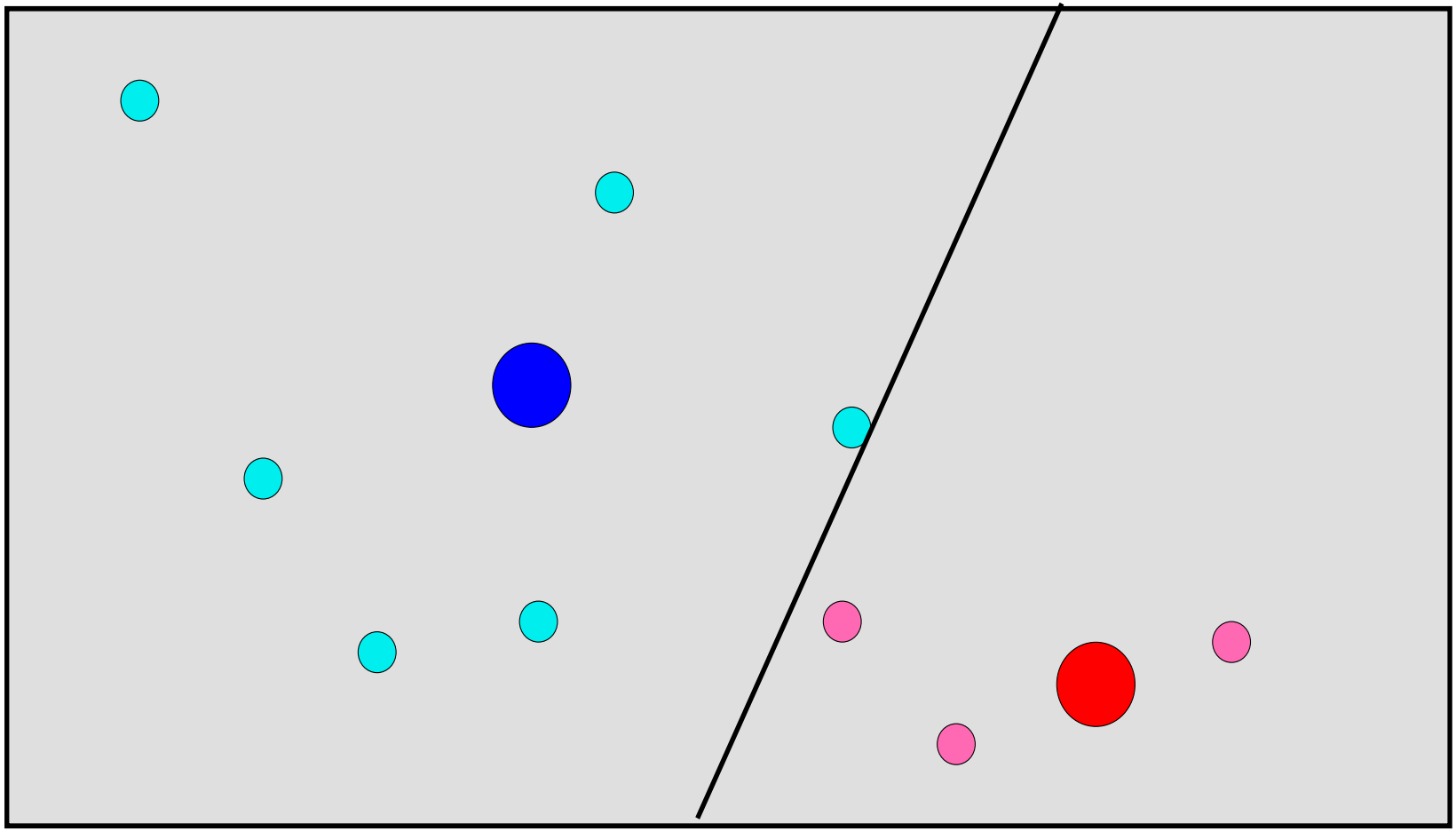
7: k-Means method - Demo



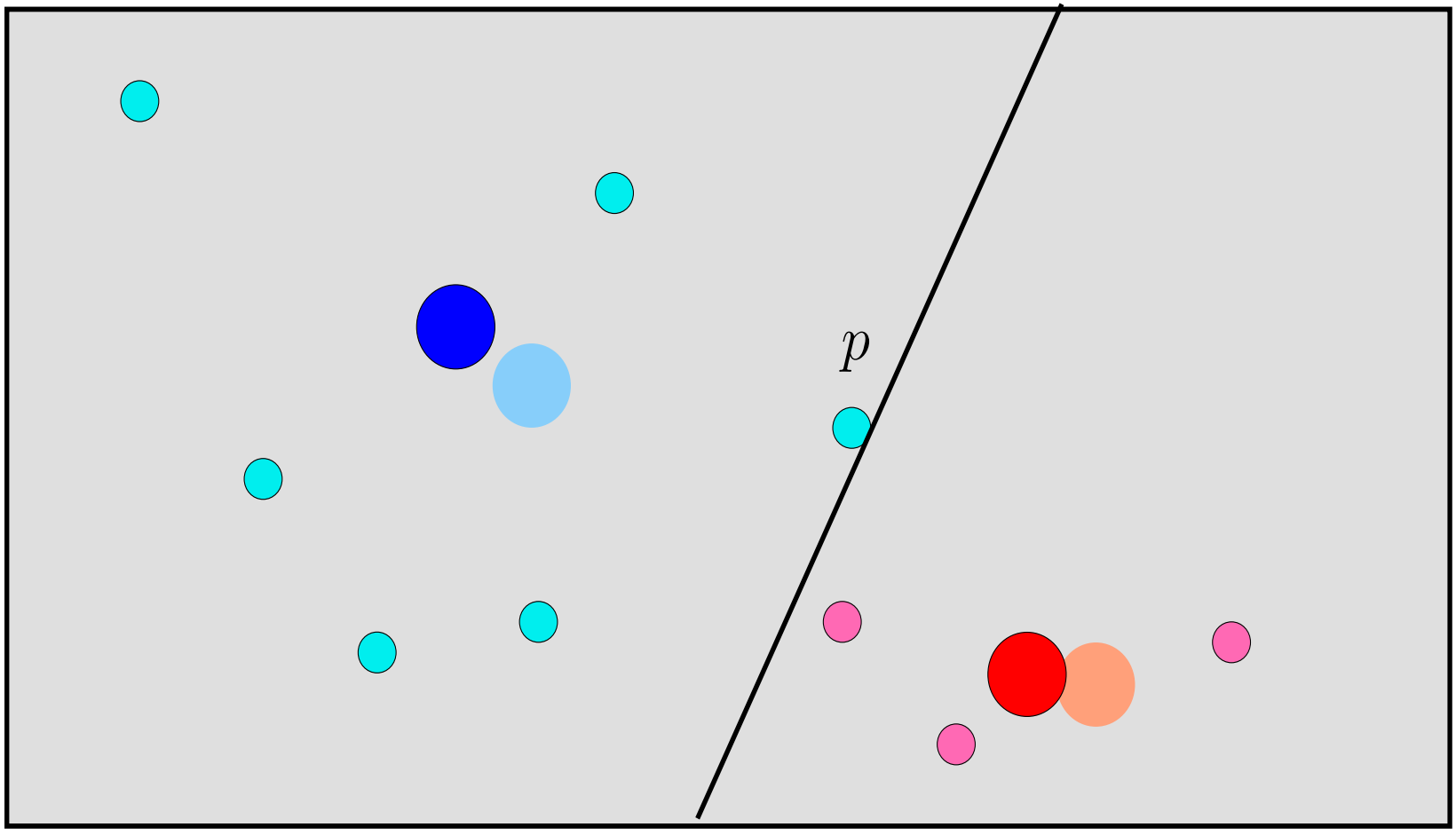
7: k-Means method - Demo



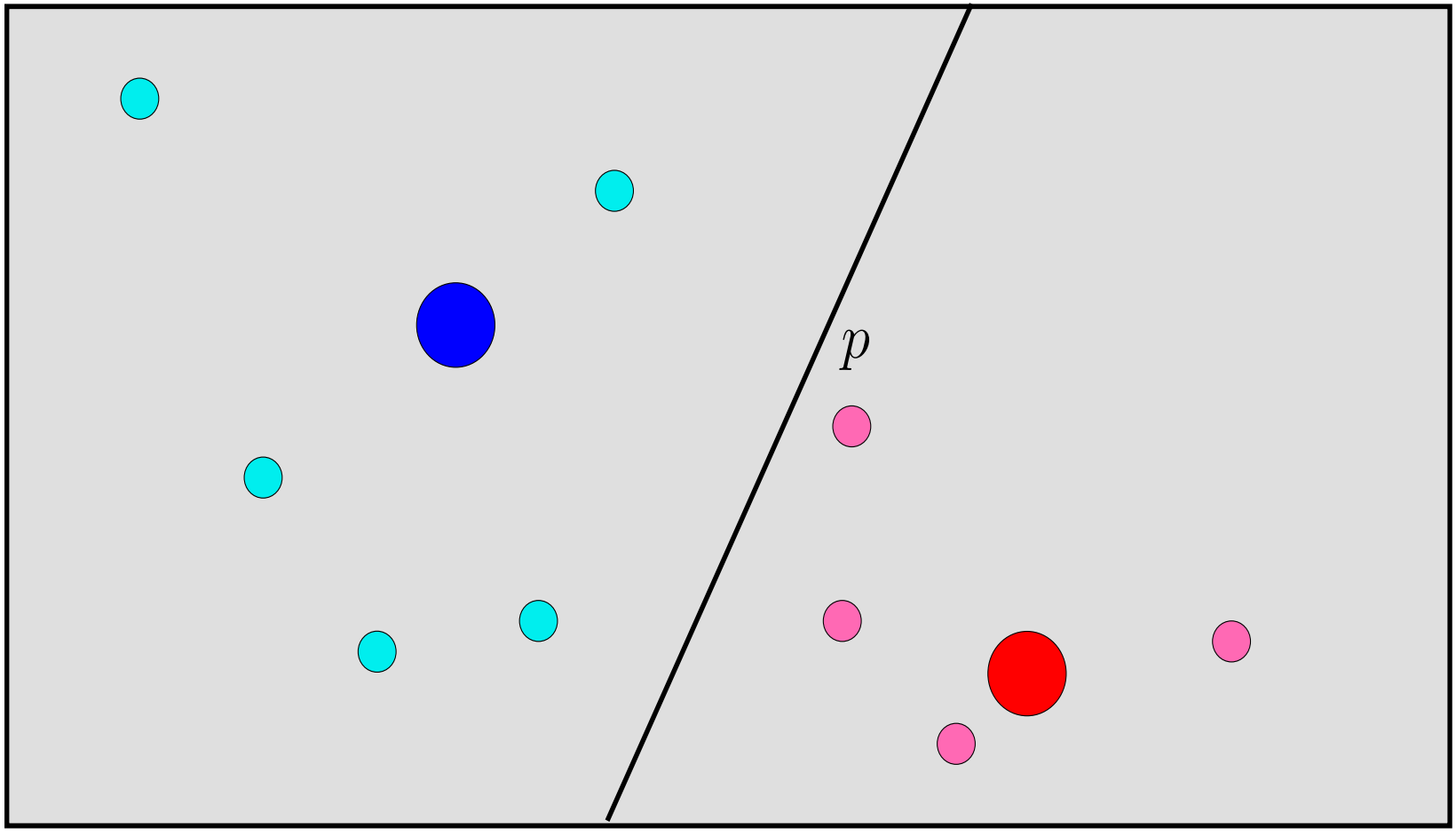
7: k-Means method - Demo



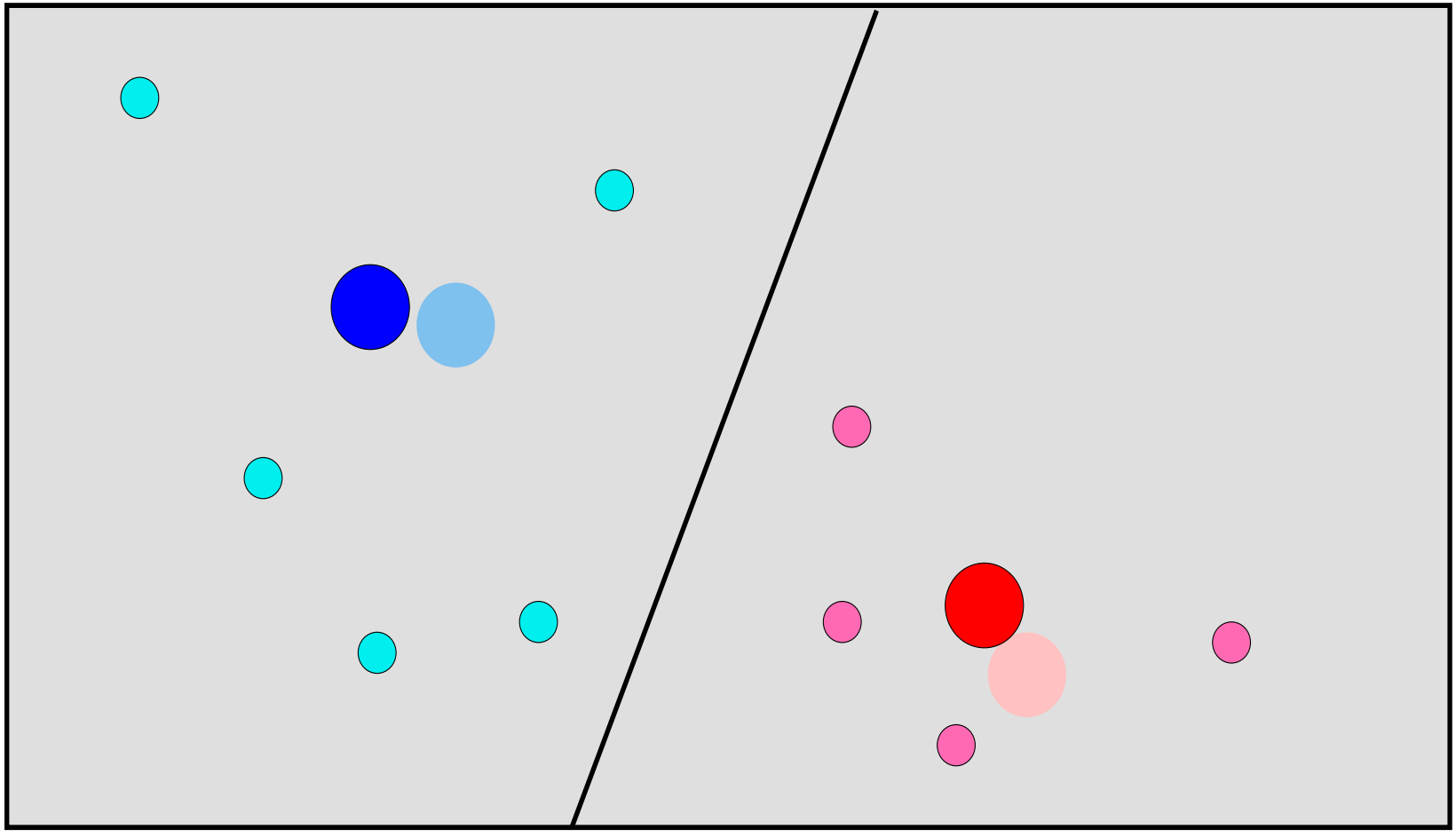
7: k-Means method - Demo



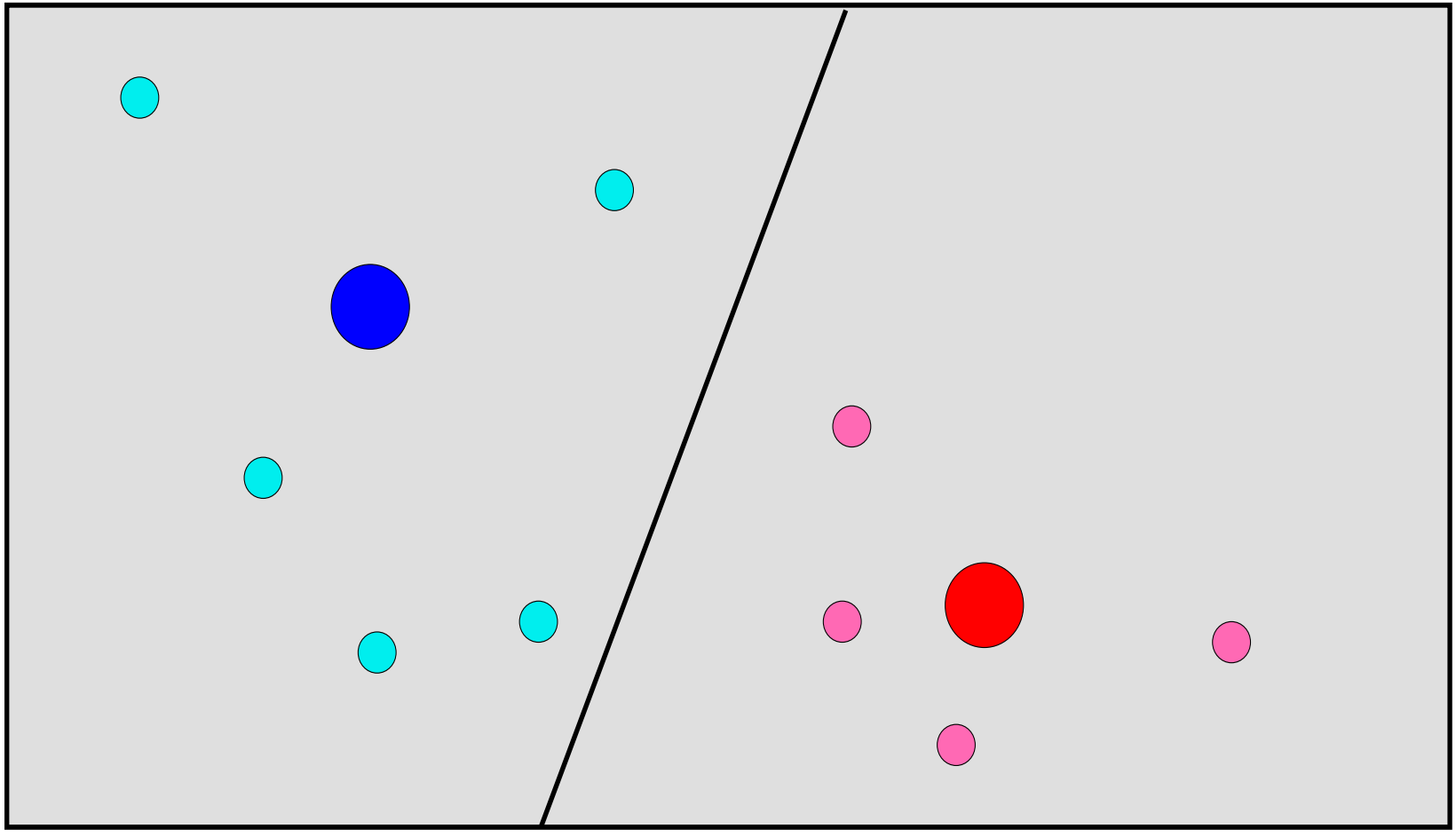
7: k-Means method - Demo



7: k-Means method - Demo



7: k-Means method - Demo



8: k-Means method

- Every iteration improves price of clustering.
- Alg. walks on Voronoi partitions of point set.
- Alg. does not cycle.
- k-means method always terminates.
- **Observation [Inaba et al. (1994)]**:
  # iterations $O\left(n^{kd}\right)$.
- Bound too big $\Rightarrow$ meaningless.
- No quality guarantee...

9: k-Means method

- **Q:** (raised by Pankaj Agarwal): Give polynomial bound on the number of iterations.
- **Motivation:** Better understand k-means method.
- **Our results:** Initial and partial answer to this question.

10: k-Means method - lower bound

- For $k = 2$
- Exist P - n points on the real line
- **Result:**
  k-means method takes $n - 2$ iterations on P.
- **Bad news...** n can be quite big...
- Spread of P is polynomial!

k-Means method

11: Upper bound $d = 1$

- $X \subset \mathbb{R}$ - set of n points.
- Δ - spread of X.
  (Ratio between longest distance to shortest distance.)
- **Result:** The number of steps of k-MeansMtd is $O(n\Delta^2)$.

k-Means method

12: Upper bound for grid

- M - integer number.
- $X \subseteq \{1, \ldots, M\}^d$ - set of n points.
- Number of iters of k-MeansMtd is $\leq dn^5M^2$.
- Covers the case of images
  - $M = 256$
  - $d = 1024 \times 768$.

SINGLEPNT
13: Alternative Algorithm

- X - set of points
- C - set of centers
- Every point maintain current center.
- Centers are centroids of points assigned to them.
- Scan the points of X
- If $x \in X$ is misclassified then
  - Reassign x to its closest center.
  - Update the two centers involved.
    (i.e., recompute centroids)

Difference between 14: SinglePnt and k-MeansMtd

- k-MeansMtd scan all the points
  $\Rightarrow$ Then update centroids.
  (i.e., batch mode)
- SinglePnt - update centroids whenever finding a misclassified points.
  (i.e., “online” mode)
- **“Conjecture”**:
  k-MeansMtd and SinglePnt have similar # of iters.

15: SINGLEPNT Performance

- $X \subset \mathbb{R}^d$ - n points.
- Δ - spread of X.
- **Result:**
  SINGLEPNT makes at most $O(kn^2\Delta^2)$ iters.
- Dimension independent!

Yet Another Variant

16: The LAZY-k-MEANS algorithm

- $\varepsilon > 0$ - parameter.
- LAZY-k-MEANS reassigns only *substantially* misclassified points.
  - x associated with center c
  - c' = Nearest center to x
  - $\|xc\| \geq (1+\varepsilon)\|xc'\|$
- **Result:**
  # of iters of LAZY-k-MEANS is $O(n\Delta^2\varepsilon^{-3})$.=

17: Why spread does not matter

- Spread tends to be small in high dimensions. (i.e., random distributions)
- Snapping to grid and breakup input into several chunks.
- Each chunk has small spread.
- analyze algorithm inside each chunk.
- Reasonable assumption.

18: Technique used

- Consider the clustering price:

$$\min_{C} \sum_{p \in P} \Big(\mathbf{dist}(p, C) \Big)^2$$

- Initial price is at most $L = n\Delta^2$
- Argue that in every k iterations prices decreases by at least $\delta = \dfrac{1}{128n}$.
- # iters $\leq \frac{L}{\delta}$.
- Natural argument.

19: Conclusions

- Preliminary results about the k-means method.
- Good bounds for variants.
- Further improvement should be possible...

References

Arora, S., Raghavan, P., and Rao, S. (1998). Approximation schemes for Euclidean k-median and related problems. In *Proc. 30th Annu. ACM Sympos. Theory Comput.*, pages 106–113.

Bădoiu, M., Har-Peled, S., and Indyk, P. (2002). Approximate clustering via coresets. In *Proc. 34th Annu. ACM Sympos. Theory Comput.*, pages 250–257.

de la Vega, W. F., Karpinski, M., Kenyon, C., and Rabani, Y. (2003). Approximation schemes for clustering problems. In *Proc. 35th Annu. ACM Sympos. Theory Comput.*, pages 50–58.

Har-Peled, S. and Kushal, A. (2004). Smaller coresets for k-median and k-means clustering. http://www.uiuc.edu/~sariel/papers/04/small_coreset/.

Har-Peled, S. and Mazumdar, S. (2004). Coresets for k-means and k-median clustering and their applications. In *Proc. 36th Annu. ACM Sympos. Theory Comput.*, pages 291–300.

Inaba, M., Katoh, N., and Imai, H. (1994). Applications of weighted voronoi diagrams and randomization to variance-based k-clustering. In *Proc. 10th Annu. ACM Sympos. Comput. Geom.*, pages 332–339.

Kolliopoulos, S. G. and Rao, S. (1999). A nearly linear-time approximation scheme for the euclidean κ-median problem. In *Proc. 7th Annu. European Sympos. Algorithms*, pages 378–389.

Kumar, A., Sabharwal, Y., and Sen, S. (2004). Linear time algorithms for clustering problems in any dimension. manuscript.

Matoušek, J. (2000). On approximate geometric k-clustering. *Discrete Comput. Geom.*, 24:61–84.