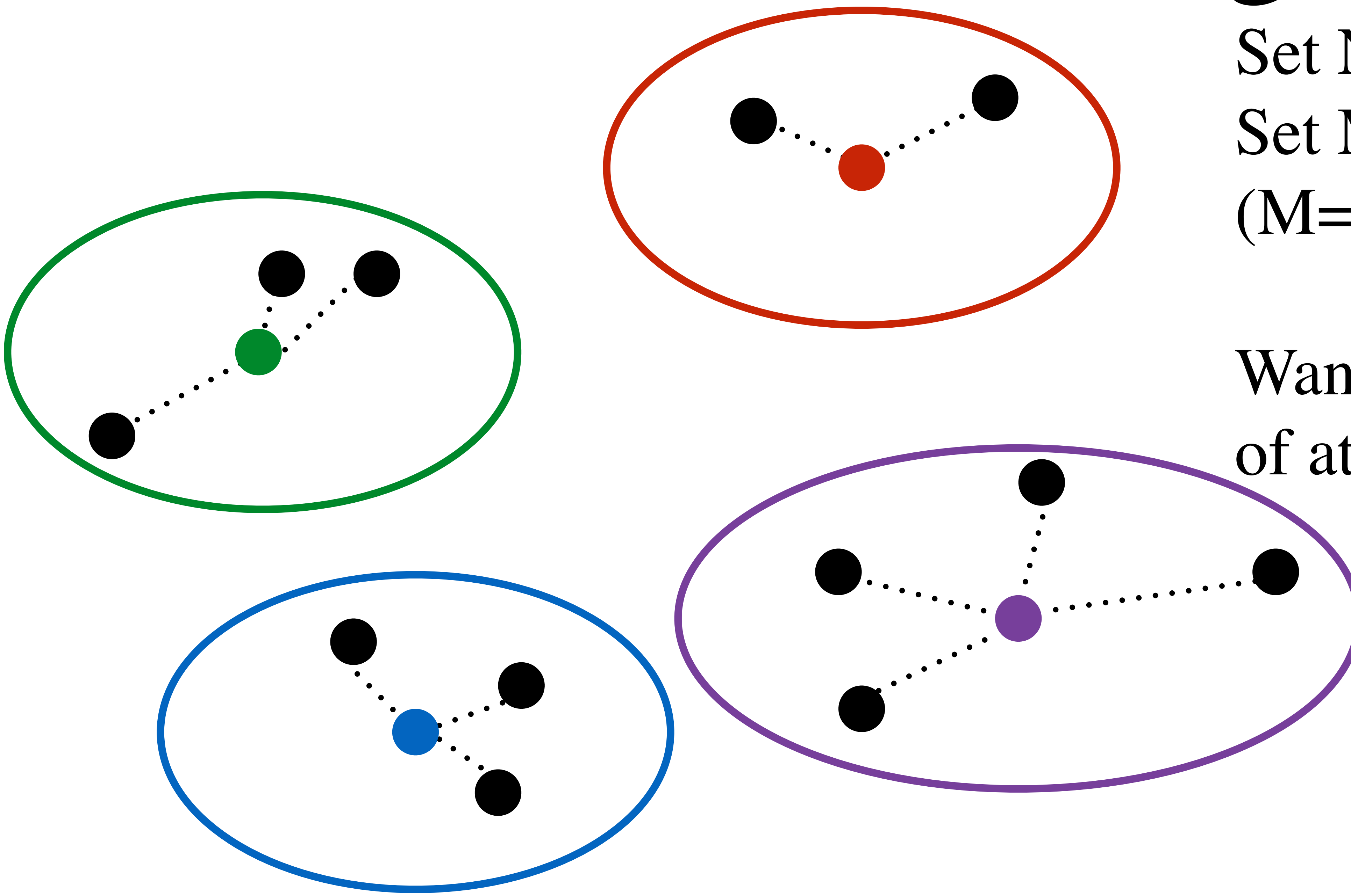
DUKE COMPUTER SCIENCE

Proportionally Fair Clustering

Xingyu Chen, **Brandon Fain**, Liang Lyu, Kamesh Munagala
Department of Computer Science, Duke University
ICML 2019

Centroid Clustering



Set N of n points
Set M of m centers.
(M=N is common)

Want to choose a set X
of at most k centers.

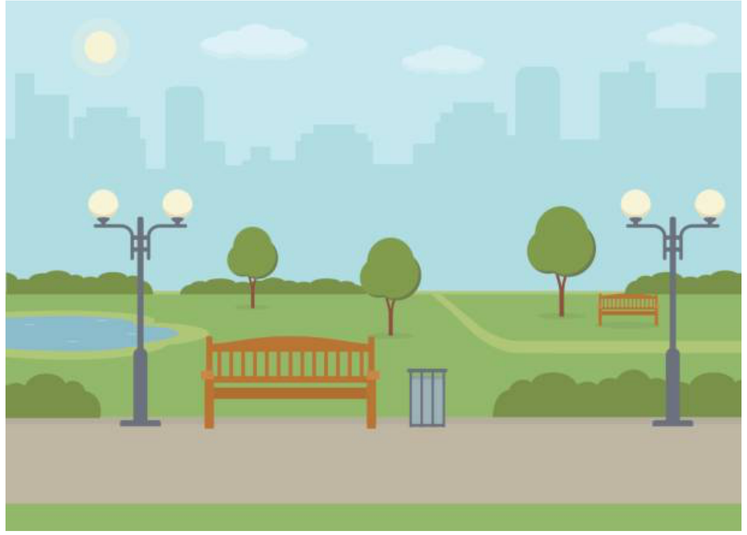
Point i has
cost d(i, x)
for center x.

Typically we want to minimize the sum of
costs (k-median) or squared costs (k-means).

How should we cluster if the data points represent individuals who care about how they are clustered?

Motivating Applications

Facility Location



For example, if we want to decide where to build public parks, we might cluster home locations, where points prefer to be closer to the centers.

Precision Medicine

Alternatively, when clustering medical data, we might want to ensure that we don't inaccurately cluster any large subgroup of agents.

Defining Proportionality

Entitlements. We assume that any *n*/*k* agents are entitled to choose their own center/cluster if they wish.

Let $D_i(X) = \min_{x \in X} d(i, x)$

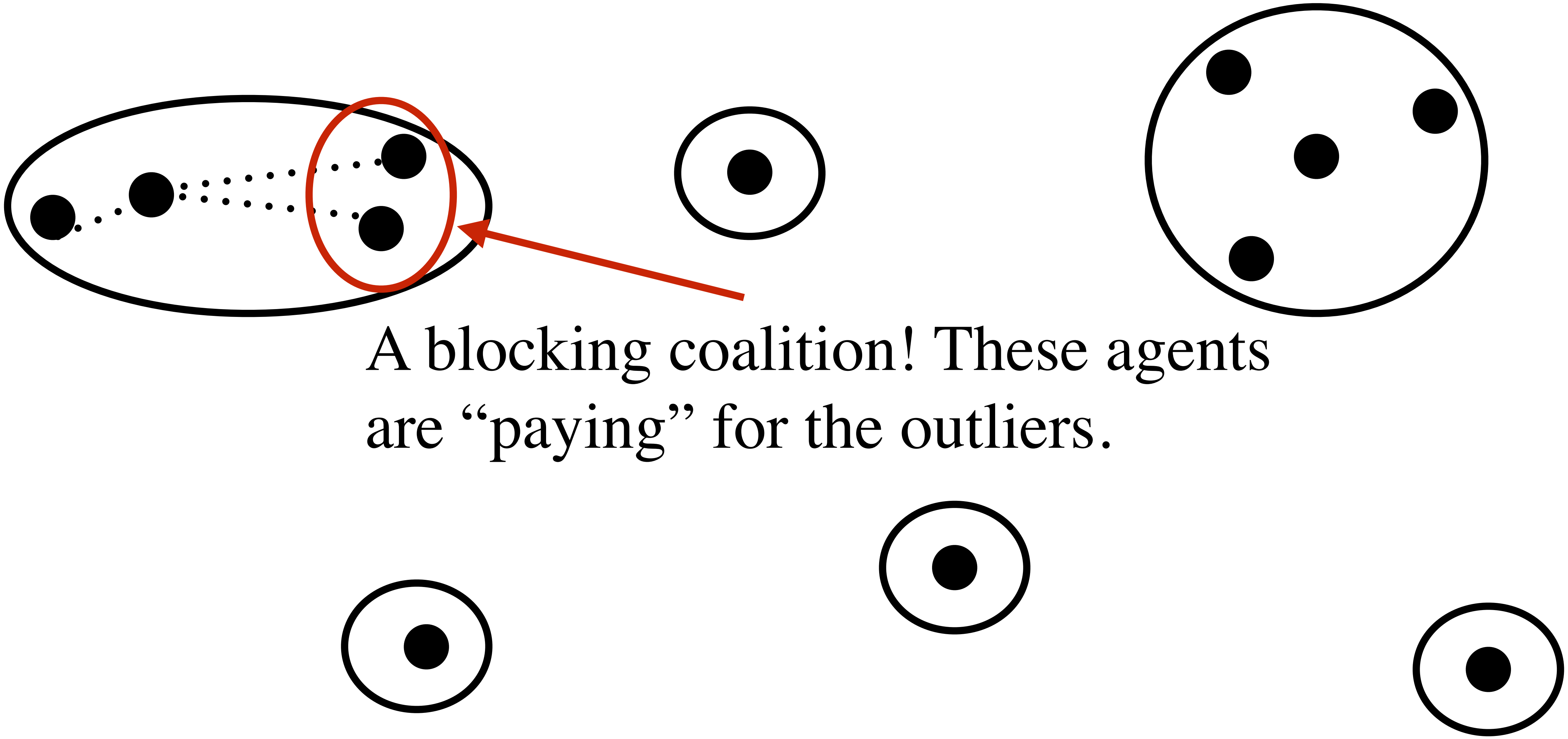
A **blocking coalition** against X is a set $S \subseteq N$ of at least n/k points and a center y such that $d(i, y) < D_i(X)$ for all $i \in S$.

A **proportional clustering** is a clustering for which there is no blocking coalition.

(This definition adapts the idea of fairness as **core** from the fair resource allocation literature [Fain et al., 2018]).

Defining Proportionality

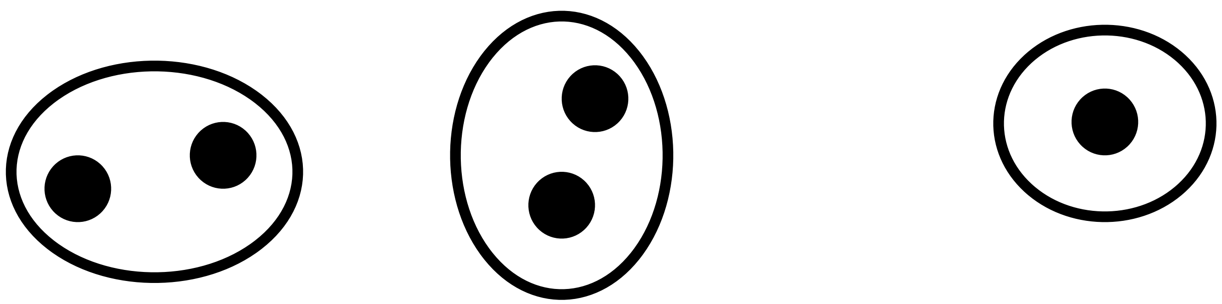
Example. Suppose k=6 and M = N.

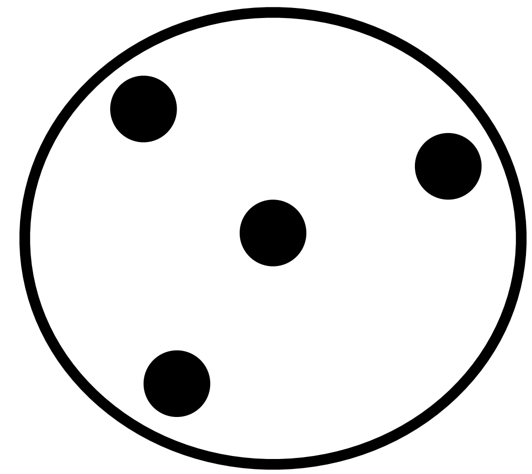


Defining Proportionality

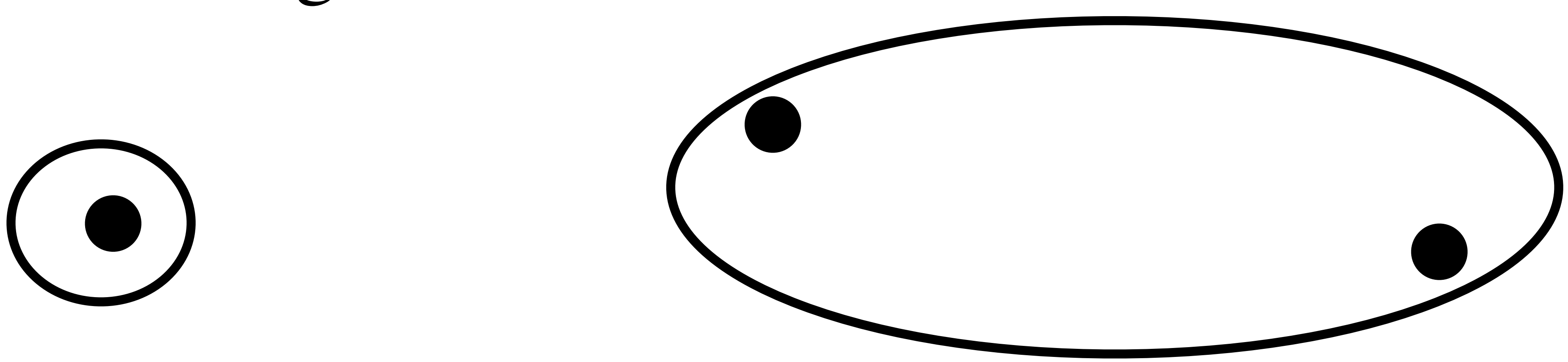
A **proportional clustering** is a clustering for which there is no blocking coalition.

Example. Suppose k=6.





This, instead, would be a proportional clustering.



Defining Proportionality

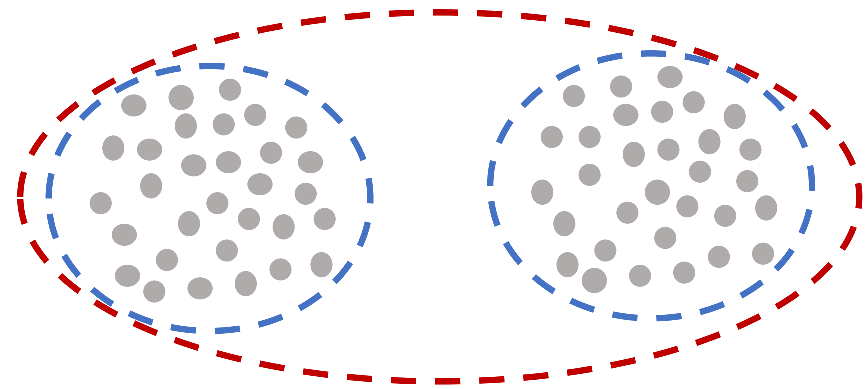
Some Advantages.

- Ensures a form of “no justified complaint” guarantee
- Is oblivious to protected/sensitive demographics (while still protecting such subgroups)
- Not sensitive to outliers
- Can be efficiently computed and audited (this paper)

Proportionality vs. Traditional Clustering

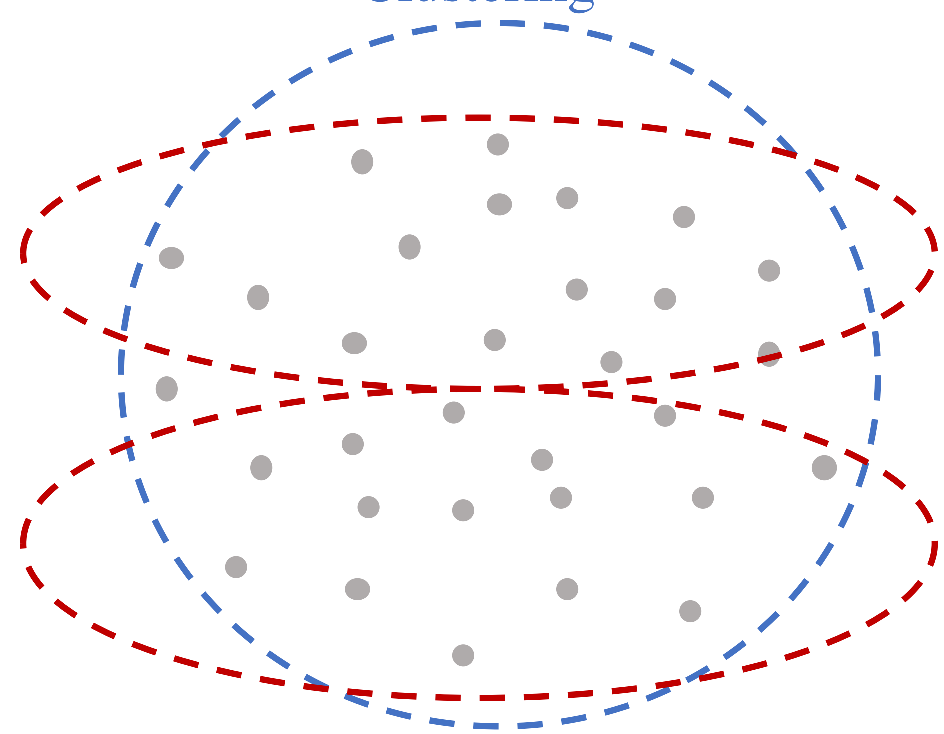
Traditional clustering, for example, k-means or k-median minimization, force some points to pay for the high variance in other regions of the data.

(One might see these kinds of instances as an independent motivation for proportionality)



Traditional Clustering

Proportional Clustering



Existence

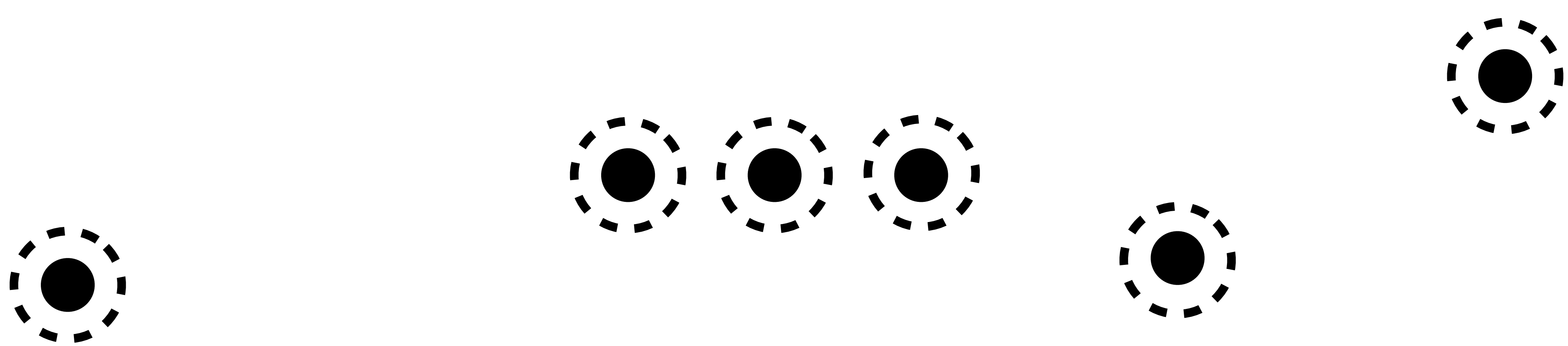
A proportional clustering may not exist. In that case, we need a notion of approximate proportionality.

X is ρ-proportional if for all $S \subseteq N$ with $|S| \geq \lceil \frac{n}{k} \rceil$, and for all $y \in M$, there exists $i \in S$ such that $\underline{\rho} \cdot d(i, y) \geq D_i(X)$.

Result 1. For $\rho < 2$, a ρ-proportional clustering may not exist. However, we can always compute a $(1 + \sqrt{2})$-proportional clustering in $\tilde{O}(n^2)$ time.

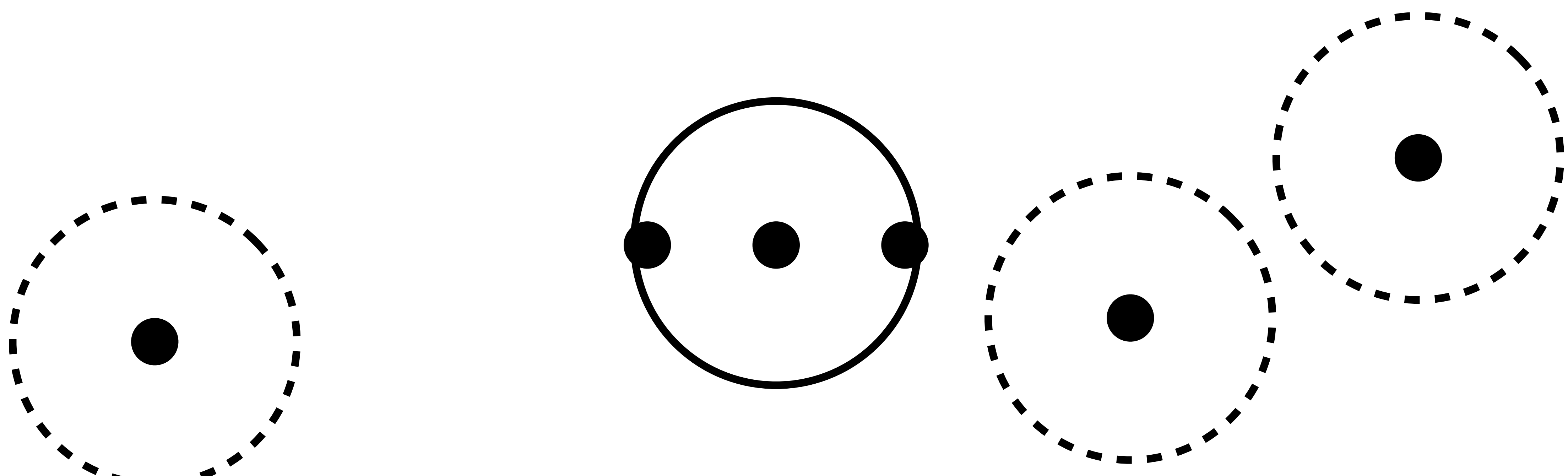
Greedy Capture Algorithm

- All points start out un-captured, and X is empty.
- Continuously grow balls around every center.
  - If there are n/k un-captured points in the ball around j:
    - Add j to X, which captures those points.
  - If an un-captured point is in the ball around j in X:
    - j captures the point.



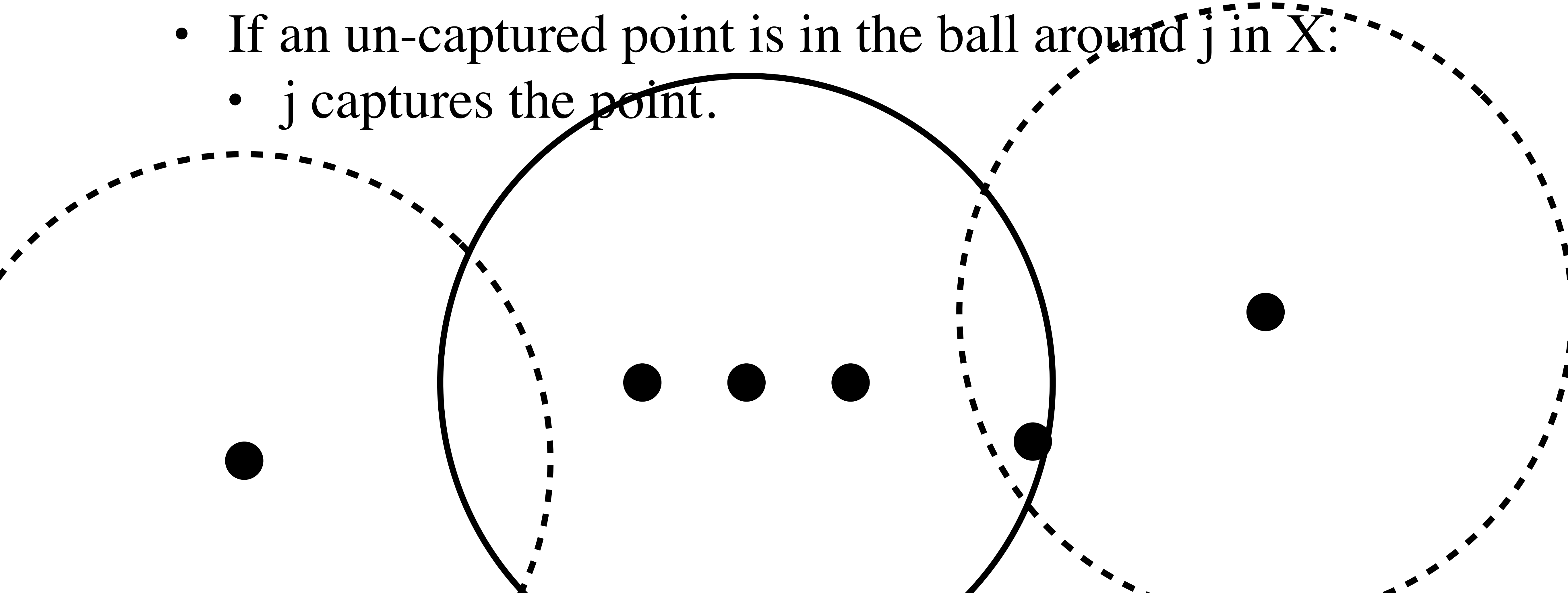
Greedy Capture Algorithm

- All points start out un-captured, and X is empty.
- Continuously grow balls around every center.
  - If there are n/k un-captured points in the ball around j:
    - Add j to X, which captures those points.
  - If an un-captured point is in the ball around j in X:
    - j captures the point.



Greedy Capture Algorithm

- All points start out un-captured, and X is empty.
- Continuously grow balls around every center.
  - If there are n/k un-captured points in the ball around j:
    - Add j to X, which captures those points.
  - If an un-captured point is in the ball around j in X:
    - j captures the point.



Upper Bound

Theorem. The greedy capture algorithm returns a $(1+\sqrt{2})$-proportional clustering.

Proof. Suppose the algorithm returns some X that is not $(1+\sqrt{2})$-proportional.

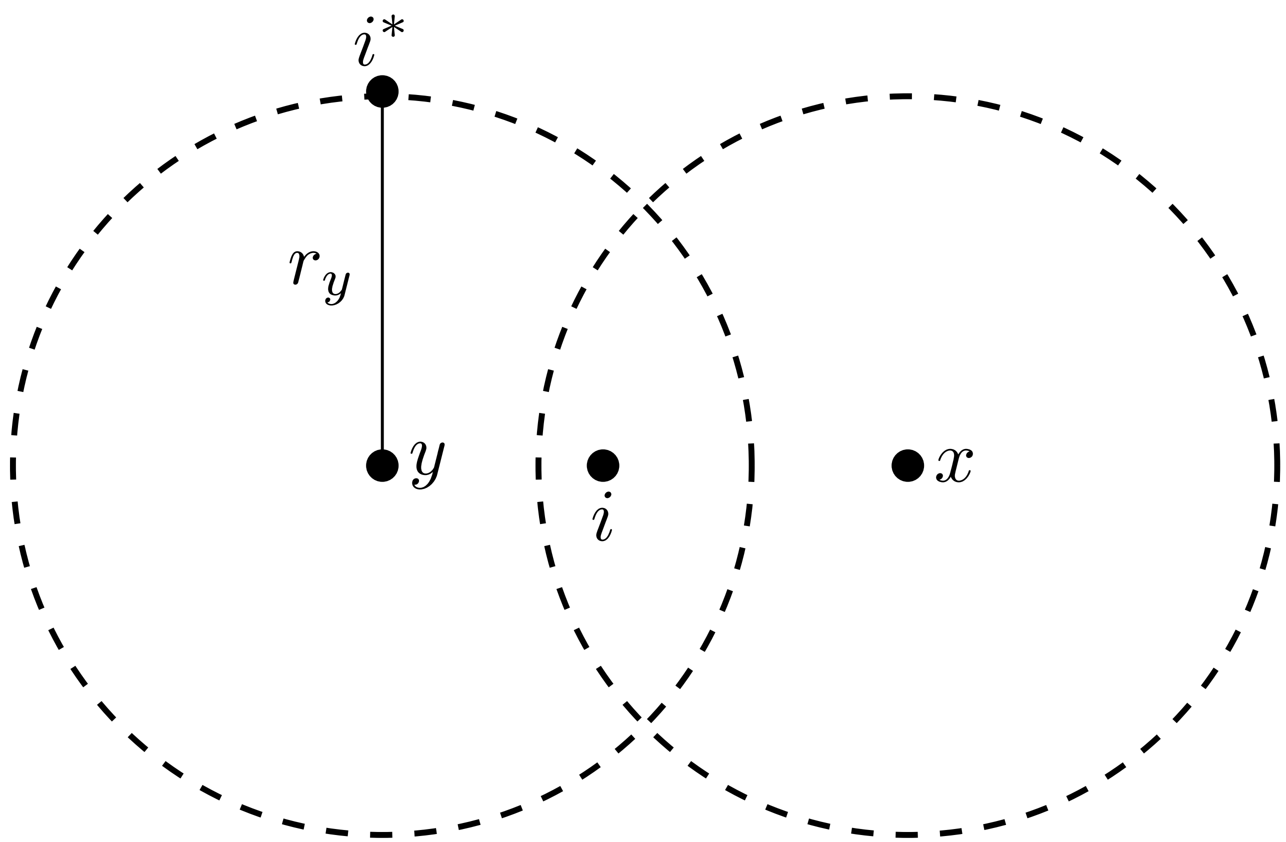
Then there are some n/k agents S and some $y \in M$ such that $\forall i \in S, (1+\sqrt{2}) \cdot d(i, y) < D_i(X)$.

Let $r_y = \max_{i \in S} d(i, y)$

There must be some $x \in X$ such that the radius r_y ball about x captured some $i \in S$.

Upper Bound

But then there must be some $i^* \in S$ for whom the distances to y and x are comparable.



The worst case bound works out to $1 + \sqrt{2}$.

Local Capture Algorithm

Problem. Greedy Capture may not find an exact proportional clustering, even when one exists.

Solution. We introduce Local Capture, a local search heuristic for finding more proportional solutions.

- Input a target value of ρ, and an arbitrary set X of k centers
- While the solution is still not ρ-proportional:
  - Add the center y of the blocking to X
  - Remove the center from X that is the least utilized (i.e., is the closest center for the fewest points)

Constrained Optimization

Problem. Although the greedy capture algorithm is approximately proportional, it may choose an inefficient clustering, even when there is an efficient proportional solution.

Result 2. Suppose there is a ρ-proportional clustering with total cost c. In polynomial time in n, we can compute a $O(\rho)$-proportional clustering with k-median objective at most $8c$.

(The approach is based on LP rounding, adapting methods from Charikar et al., 2002)

Sampling

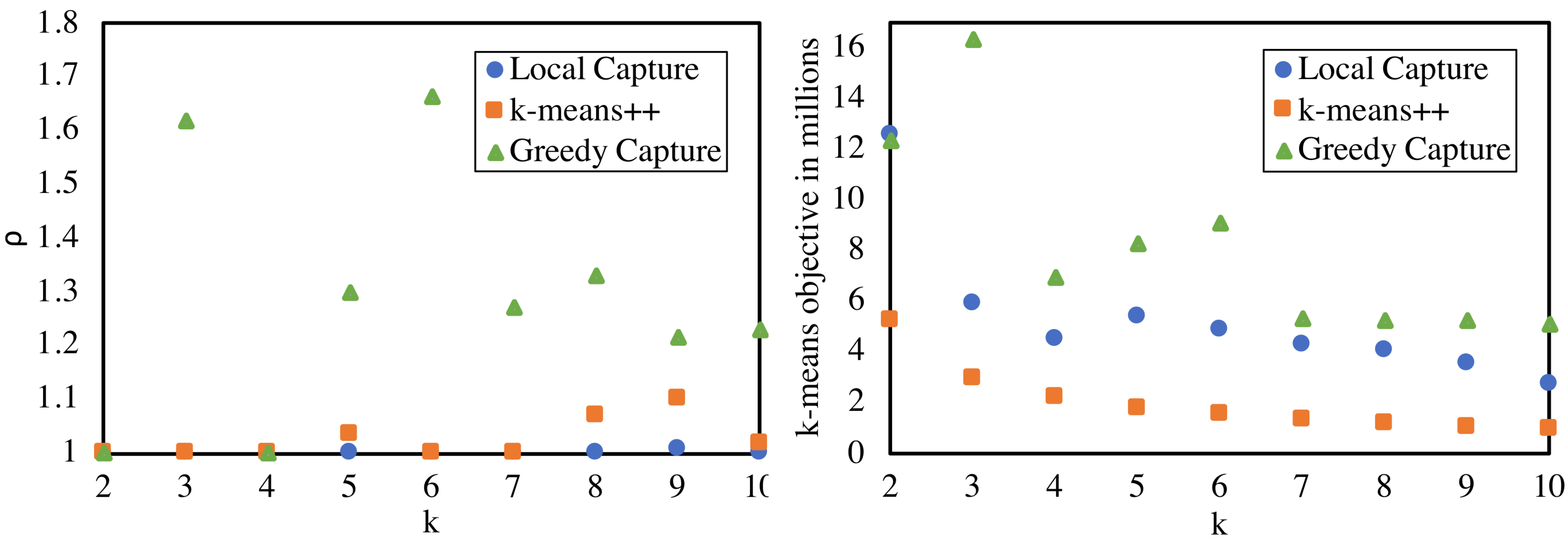
Problem. Running greedy capture, or even checking whether a clustering is proportional, takes $\Omega(n^2)$ time.

Observation. Proportionality is well preserved under random sampling.

Result 3. We design Monte Carlo style randomized algorithms for computing and auditing an approximately proportional clustering in $\tilde{O}\left(\frac{m}{\epsilon^2}\right)$ time (recall m is the number of centers, sometimes just n).

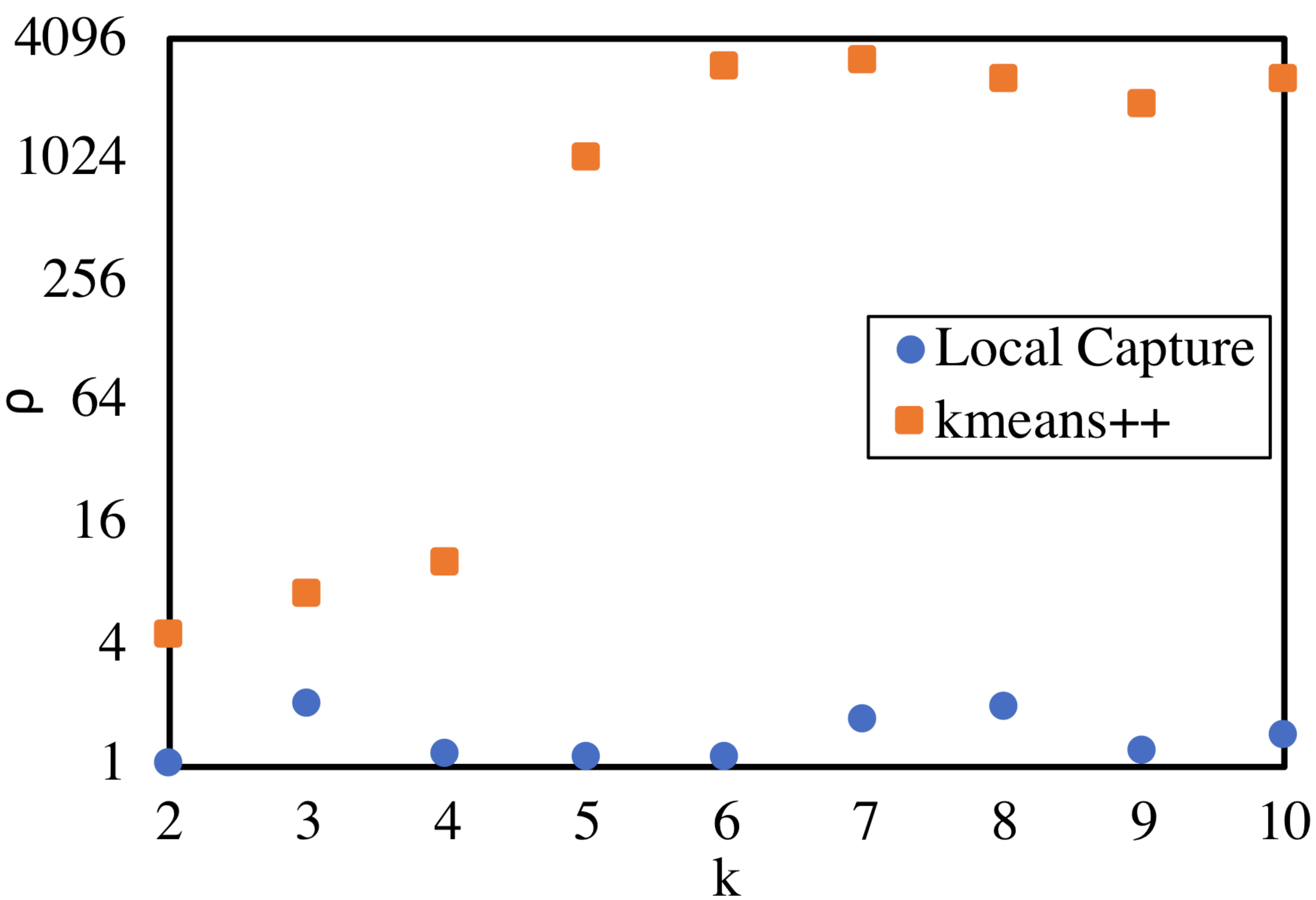
Experiment - Diabetes

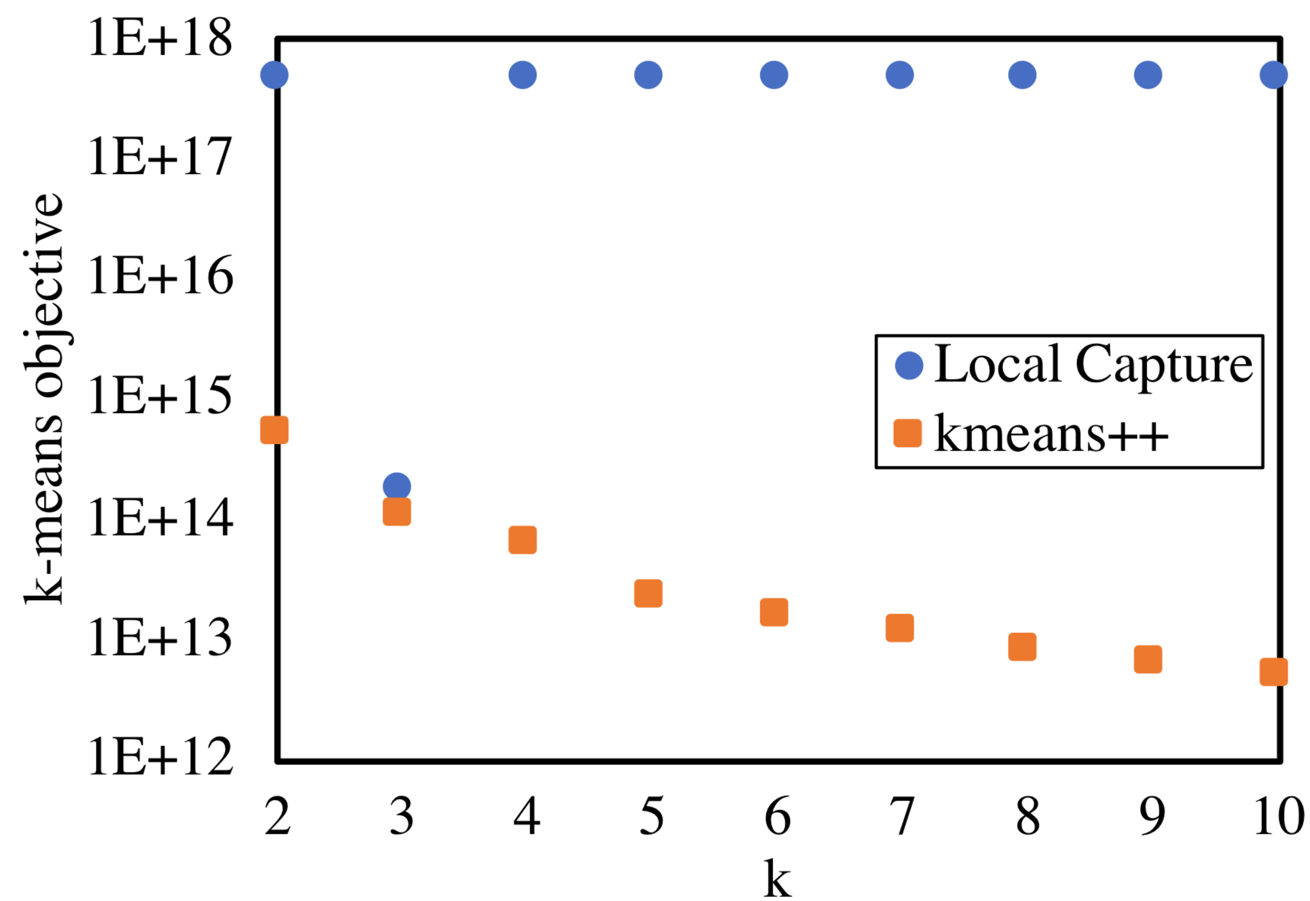
This data set contains 768 diabetes patients, recording features like glucose, blood pressure, age and skin thickness. These are our centers and data points, i.e., M = N.



Experiment - KDD

The KDD cup 1999 data set has information about sequences of TCP packets and contains many outliers. We work with a subsample of 100,000 data points, and a further subsample of 400 points for M.





Open Questions

- Can we close the approximation gap?
- Is there a more simple, efficient, and intuitive way to optimize the k-median objective subject to approximate proportionality?
- What are the right other competing fairness notions for clustering?
- Can fairness as proportionality be adapted for supervised learning tasks like classification?

DUKE
COMPUTER SCIENCE

Proportionally Fair Clustering

Xingyu Chen, **Brandon Fain**, Liang Lyu, Kamesh Munagala
Department of Computer Science, Duke University
ICML 2019

References.

- Charikar, M., Guha, S., va Tardos, and Shmoys, D. B. A constant-factor approximation algorithm for the k-median problem. *Journal of Computer and System Sciences*, 65 (1):129 – 149, 2002.
- Fain, B., Munagala, K., and Shah, N. Fair allocation of indivisible public goods. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC)*, pp. 575–592, 2018.
- Fain, B., Goel, A., and Munagala, K. The core of the par- ticipatory budgeting problem. In *Proceedings of the 12th International Conference on Web and Internet Economics (WINE)*, pp. 384–399, 2016.