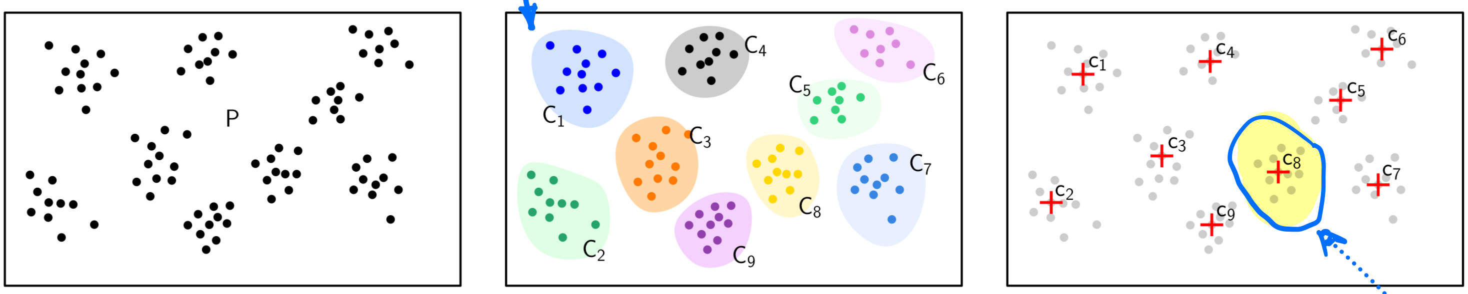
# CMSC 451 - Algorithm Design
# Lecture 6 - k-Center Clustering + Gonzalez's Algorithm

Greedy algorithms often used to approximate NP-hard problems

Clustering - Given a set of points P + distance function, partition it into similar groups, called clusters $\{C_1, \dots, C_k\}$




Center-based clustering -

Compute a set of cluster centers $\{c_1, \dots, c_k\}$ and clusters are implicitly defined by distance

$N(c_i)$ = subset of P closest to $c_i$

Two varieties -

- Centers must be chosen from P (discrete clustering)
- Centers can be any point in space

## Three Common Center-Based Clusterings:

Compute k center points to minimize...

$\boxed{\text{sum of}}$ k-median

$\boxed{\text{sum of squared}}$ k-means

$\boxed{\text{max of}}$ k-center

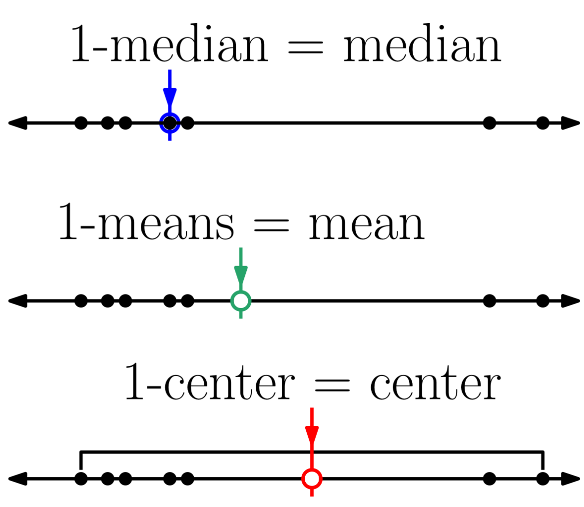
...Euclidean distances to nearest center

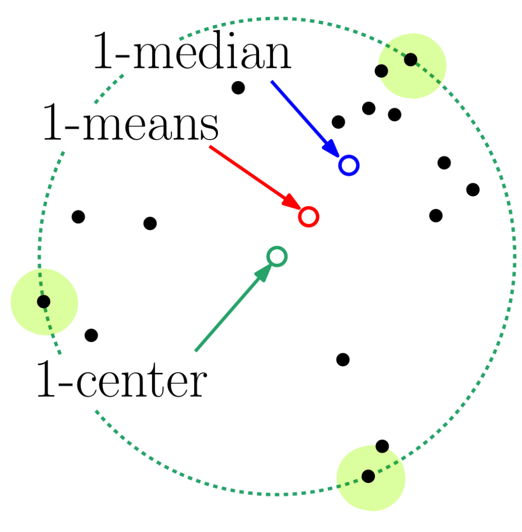
Which is best? Depends on application (e.g., sensitity to outliers)

Helps to understand the single-cluster case (k=1)

1-median - 1-D → median

d-D → hard! Fermat-Weber problem



1-means - 1-D → mean

d-D → centroid (center of mass)

Easy to compute in any dimension!

Take mean coord. value in each dim.

k-Means is very popular - Lloyd's Algorithm

1-center - 1-D → midpoint of min + max

d-D → center of min enclosing ball

(Can compute in $O(n)$ time,

but tricky algorithm - take CMSC 754)

## Metric Space:

Distance function $\delta : P \times P \to \mathbb{R}^{\geq 0}$

- $\delta(p,q) \geq 0$ + $\delta(p,p) = 0$ - Positive
- $\delta(p,q) = \delta(q,p)$ - Symmetric
- $\delta(p,r) \leq \delta(p,q) + \delta(q,r)$ - $\Delta$-Inequality

## k-Center Problem:

Given point set $P$ in a metric space and $k \geq 1$, compute $C \subseteq P$ of size $k$ to minimize max distance to closest center in $C$.

Note: Centers must be drawn from $P$

More formally - Given $C \subseteq P$, define objective fn.

$$\Delta_P(C) = \max_{p \in P} \min_{c \in C} d(p,c)$$

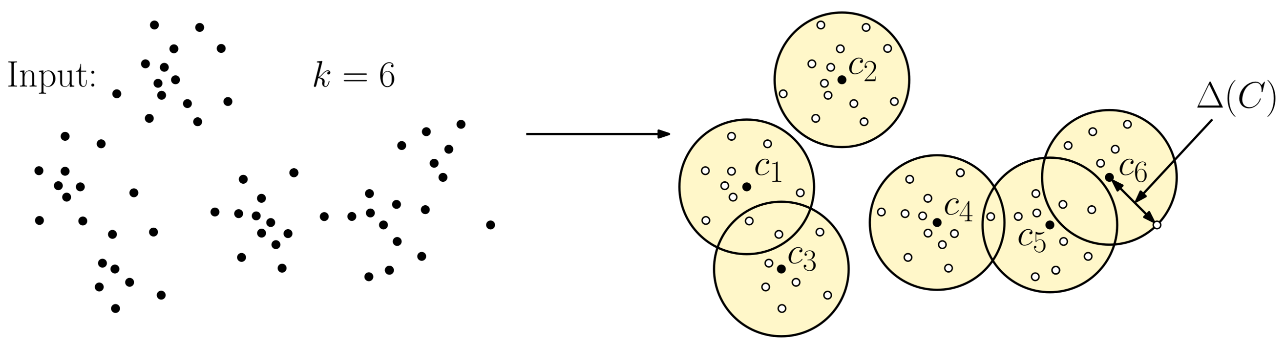
Problem - Compute a k-element set to minimize this:

$$\min_{\substack{C \subseteq P \\ |C| = k}} \Delta_C(P)$$

Geometric interpretation -

Cover all pts of $P$

- k balls (centered at pts of $P$)
- minimum radius $r = \Delta_P(C)$




Gonzalez's Algorithm -

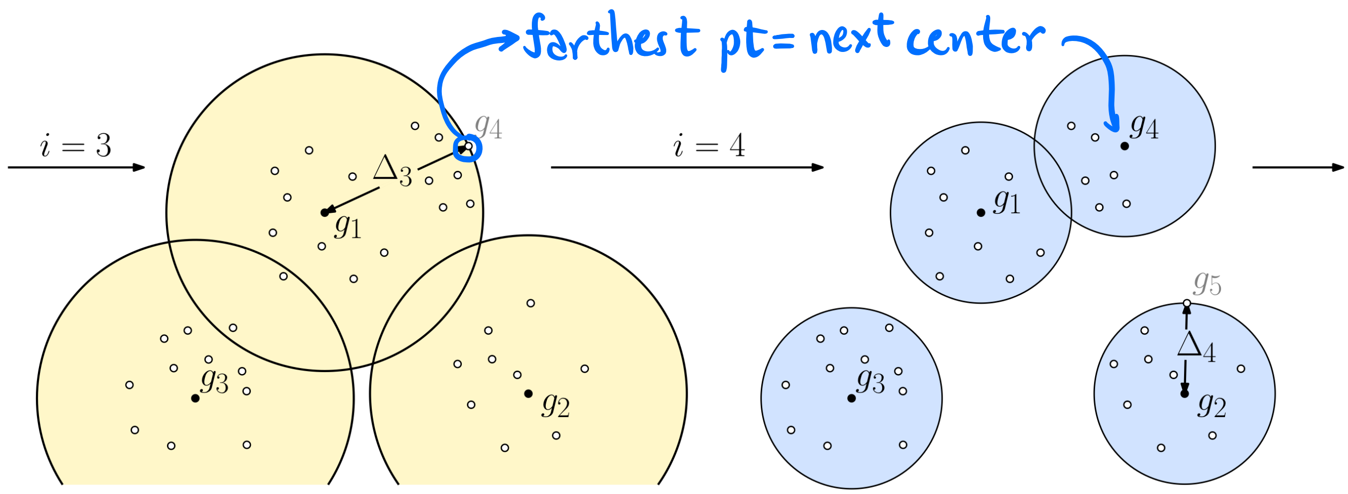
- Greedy + very simple
- 2x-approx. to k-center
- $O(k \cdot n)$ time

## Intuitive Explanation:

Repeatedly add the point that is farthest from its closest center

```
gonzalez(P, k)                        // Gonzalez's k-center
    G ← ∅
    for each (p ∈ P) d[p] ← +∞        // init. dists
    for (i ← 1 to k)
        p ← pt of P s.t. d[p] is max  // farthest pt
        add p to G                    // ... is next center
        for each (q ∈ P)              // update dist to closest
            d[q] ← min(d[q], dist(p,q))
    return G                          // final centers
```

Example – $i = 4$   $\Delta_i$ = max distance to closest center = ball radius



**Correctness** –

**Feasibility** – Clearly the algorith returns a valid clustering (provided $|P| \geq k$)

**(Approx.) Optimality** – Will show that our final radius $\leq 2 \cdot$ opt radius

Given any set $C \subseteq P$, recall that obj. fn. is

$$\Delta_P(C) = \max_{p \in P} \min_{c \in C} d(p,c)$$

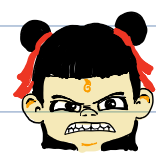
Let $G$ = output of Gonzalez

$O$ = opt. k-center solution

$$\boxed{\textbf{Thm: } \Delta_P(G) \leq 2 \cdot \Delta_P(O)}$$

we'll drop this subscript

At first glance this seems hopeless!



- k-center is NP-hard
- We cannot know what $\Delta(O)$ is!

**Strategy** –

- Derive an easily computable estimate, $\Delta_{min}$
- Show: $\Delta(O) \geq \Delta_{min}$
- Show: $\Delta(G) \leq 2 \cdot \Delta_{min}$

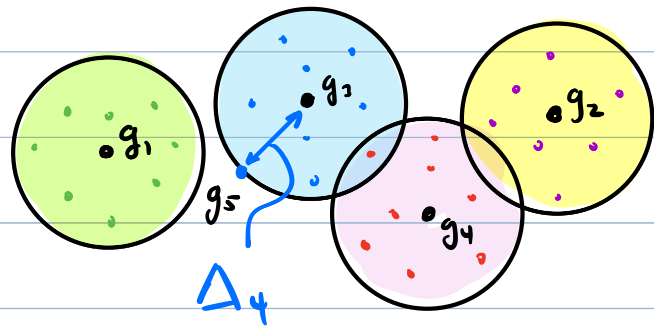
$\Rightarrow$ $\Delta(G) \leq 2 \cdot \Delta_{min} \leq 2 \cdot \Delta(O)$ ✓

## Notation:

$G_i = \{g_1, \ldots, g_i\}$ – the first $i$ greedy ctrs.

$\Delta_i = \Delta(G_i)$ – farthest dist to these ctrs.

$i=4$




Imagine that we ran one additional iteration to get $k+1$ centers $G_{k+1}$

## The theorem follows from 3 claims:

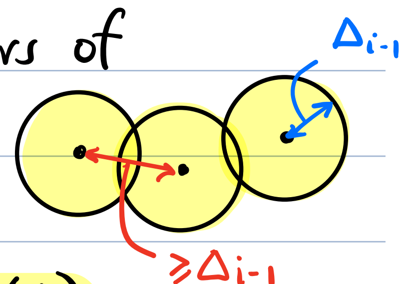
**Claim 1:** (Greedy distances decrease)

For $1 \le i \le k+1$, $\Delta_{i+1} \le \Delta_i$

**Pf:** As we add more centers, the dist to each pt's closest ctr. can never increase.

**Claim 2:** (Greedy centers are never close)

For $1 \le i \le k+1$, every pair of centers of $G_i$ are at dist $\ge \Delta_{i-1}$




**Corollary:** $g, g' \in G_{k+1} \Rightarrow \delta(g, g') \ge \Delta_k = \Delta(G)$

**Pf.** By induction on $i$.

- At stage $i-1$, by induction, all old ctrs. are sep. by dist $\geq \Delta_{i-2}$.
- by Claim 1, $\Delta_{i-2} \geq \Delta_{i-1}$ ✓
- New center is at dist $\Delta_{i-1}$ from its closest center $\Rightarrow$ it's at dist $\geq \Delta_{i-1}$ from all centers ✓

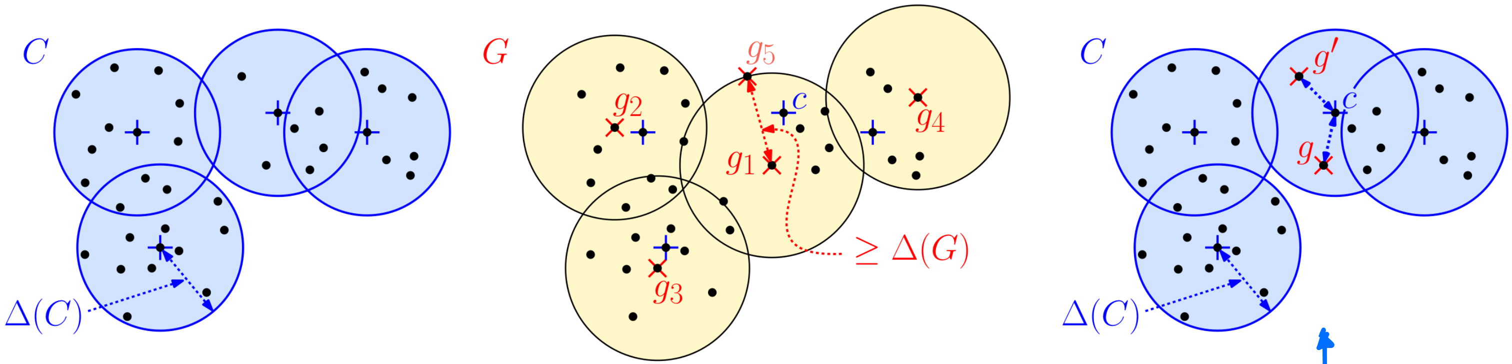
**Define:** $\Delta_{min} = \Delta(G)/2$

**Claim 3:** ($\Delta_{min}$ is a lower bound)

For any set $C \subseteq P$ of size $k$, $\Delta(G) \geq \Delta_{min}$

**Pf:** By def, every pt of $P$ lies within dist $\Delta(C)$ of some pt of $C$.

- Since $G_{k+1} \subseteq P$, every pt. of $G_{k+1}$ is within dist $\Delta(C)$ of some pt of $C$.



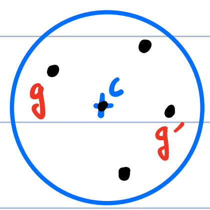
- Since $|G_{k+1}| = k+1$ + $|C| = k$, at least two pts of $G_{k+1}$ are within dist $\Delta(C)$ of same pt of $C$.

"Pigeonhole principle"

$\Rightarrow \exists\, g, g' \in G_{k+1}$ $c \in C$ s.t.

Ⓑ $\delta(g,c) \leq \Delta(C)$ + $\delta(g',c) \leq \Delta(C)$

- We have:




$$\begin{aligned}
\Delta(G) &\overset{ⓐ}{\leq} \delta(g,g') \\
&\leq \delta(g,c) + \delta(c,g') \quad (\Delta\text{-inequality}) \\
&\leq \delta(g,c) + \delta(g',c) \quad (\text{symmetry}) \\
&\overset{ⓑ}{\leq} \Delta(C) + \Delta(C) \\
&\leq 2\Delta(C)
\end{aligned}$$

$\Rightarrow \Delta(C) \geq \Delta(G)/2 \overset{\text{by Def of } \Delta_{min}}{=} \Delta_{min}$ ✓

In conclusion: Applying Claim 3 to opt, $O$,

$$\Delta(G) = 2 \cdot \Delta_{min} \leq 2 \cdot \Delta(O)$$

$\therefore$ Greedy is within factor 2 of opt. □

Summary
- k-center - NP-hard clustering problem
- Gonzalez - Greedy alg. for k-center
- Factor 2 approx. to optimum