Clustering: The k-means Problem
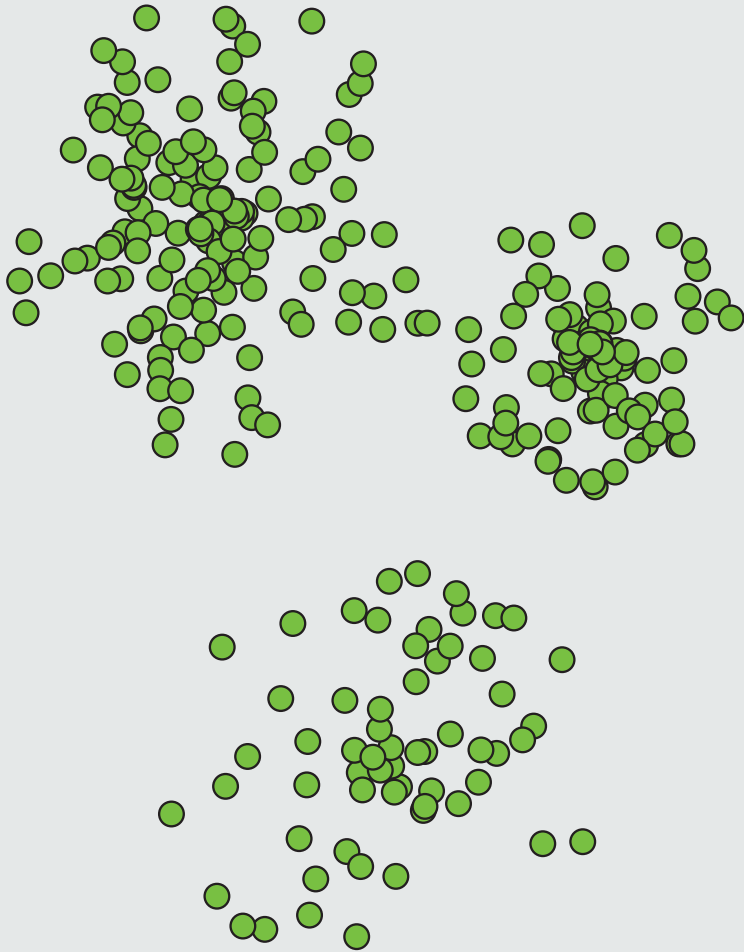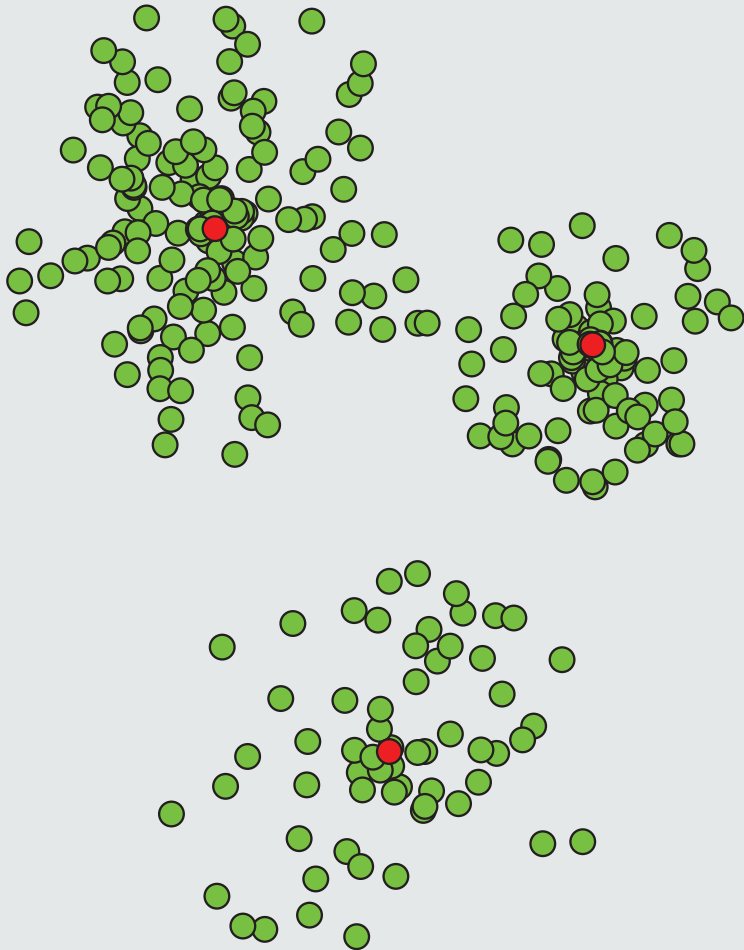
Example

The Problem

- Input points $P \subseteq \mathbb{R}^d$

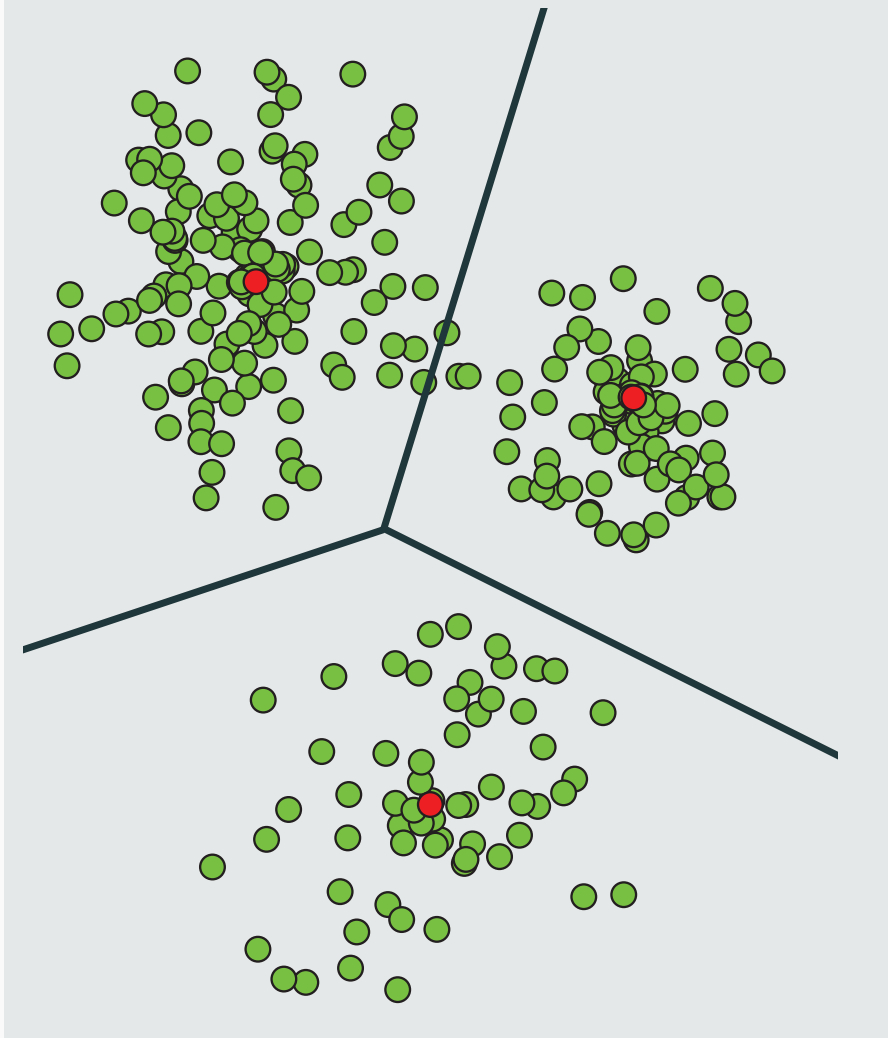# Definition: The k-means Problem

Example

The Problem

- Input points $P \subseteq \mathbb{R}^d$

- Compute a set $C \subseteq \mathbb{R}^d$ with $|C| = k$ centers such that the sum of squared distances

$$\sum_{x \in P} \min_{c \in C} ||c - x||^2$$

is minimized.

Definition: The k-means Problem
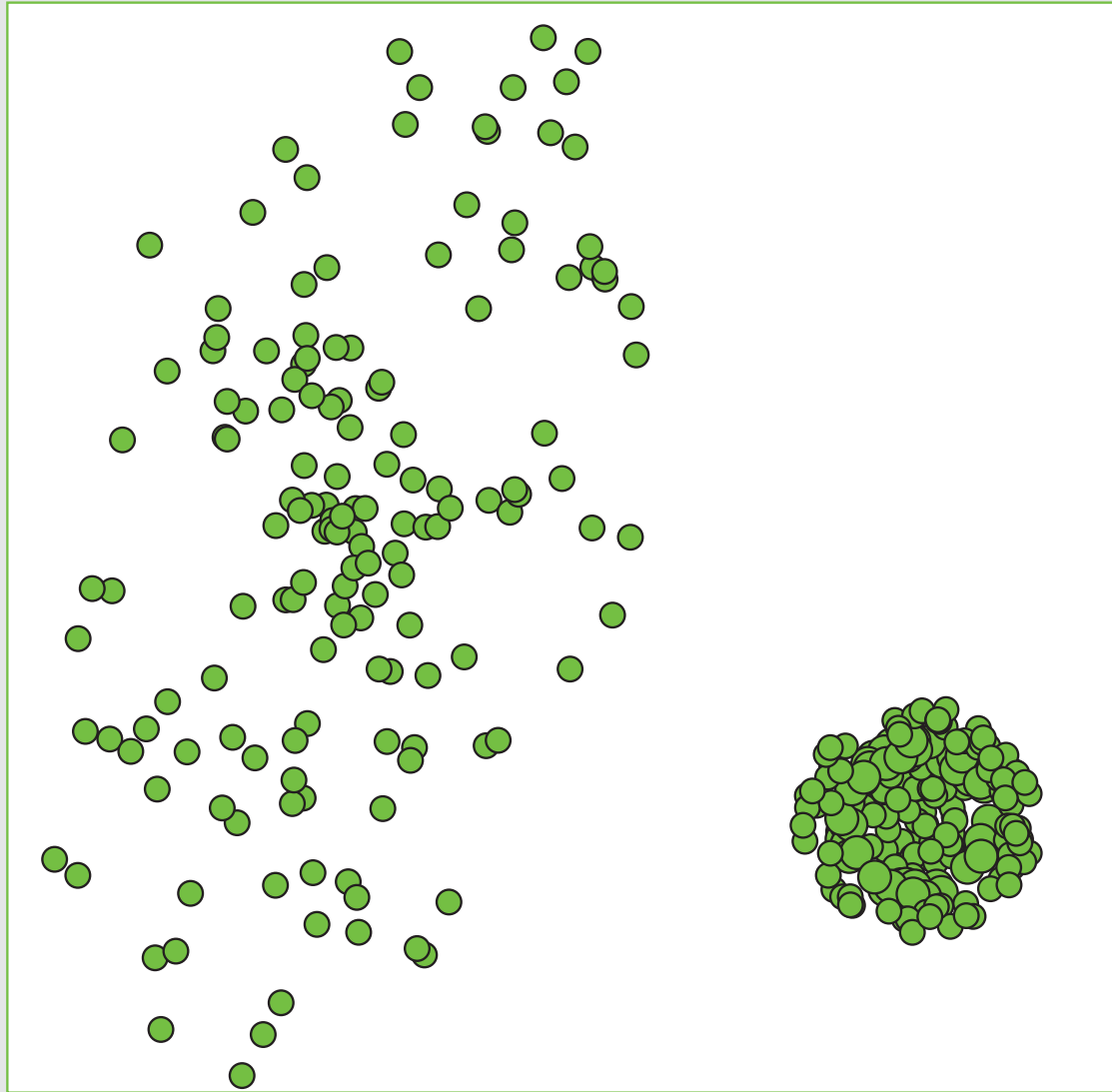
Example

The Problem

- Input points $P \subseteq \mathbb{R}^d$
- Compute a set $C \subseteq \mathbb{R}^d$ with $|C| = k$ centers such that the sum of squared distances

$$\sum_{x \in P} \min_{c \in C} ||c - x||^2$$
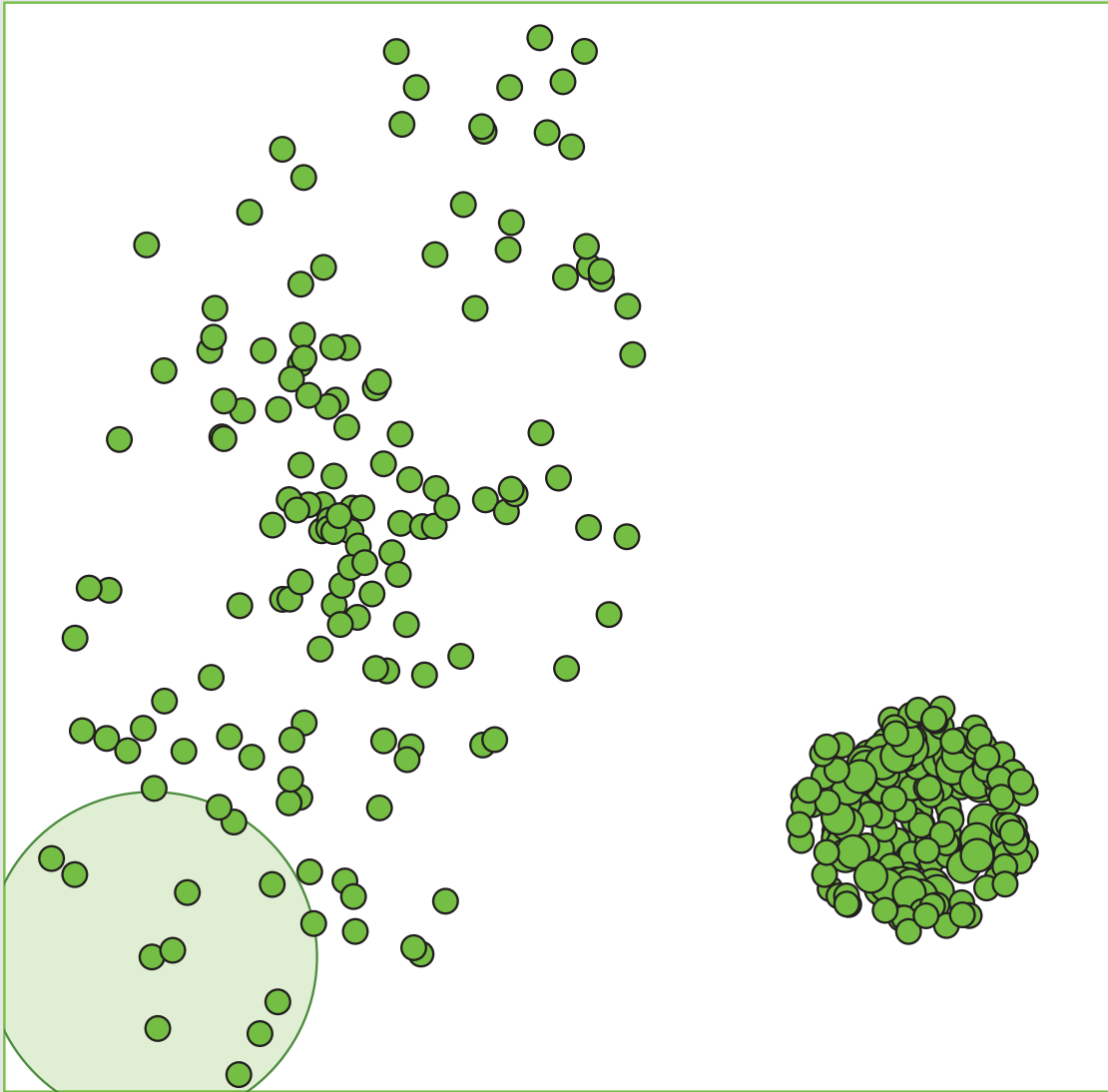
is minimized.

Reduction Idea: Smart Covering

BICO – BIrch meets COresets

Reduction Idea: Smart Covering
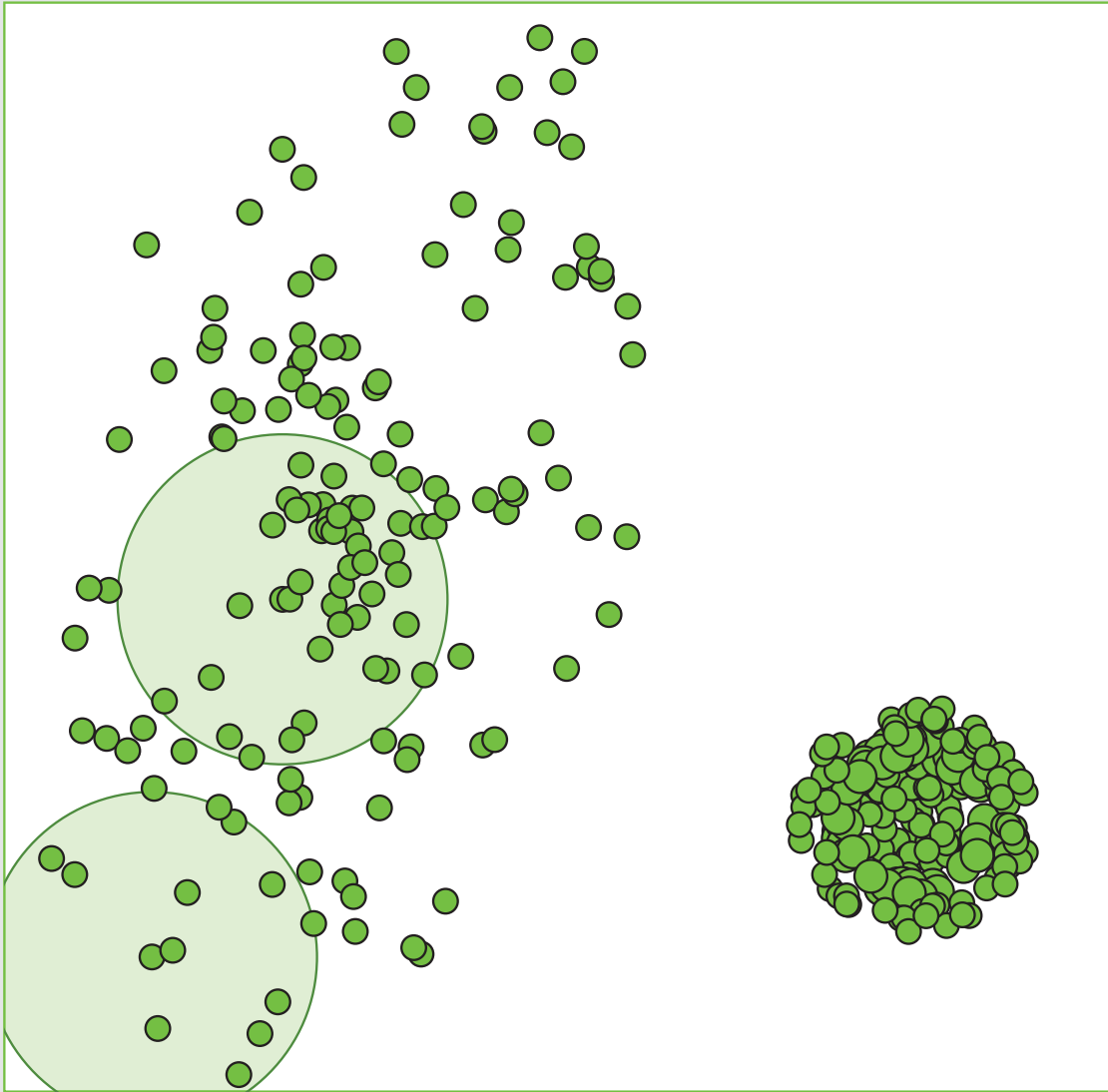
BICO – BIrch meets COresets

Idea

- Compute ball cover

Reduction Idea: Smart Covering

BICO – BIrch meets COresets

Idea

- Compute ball cover

Reduction Idea: Smart Covering

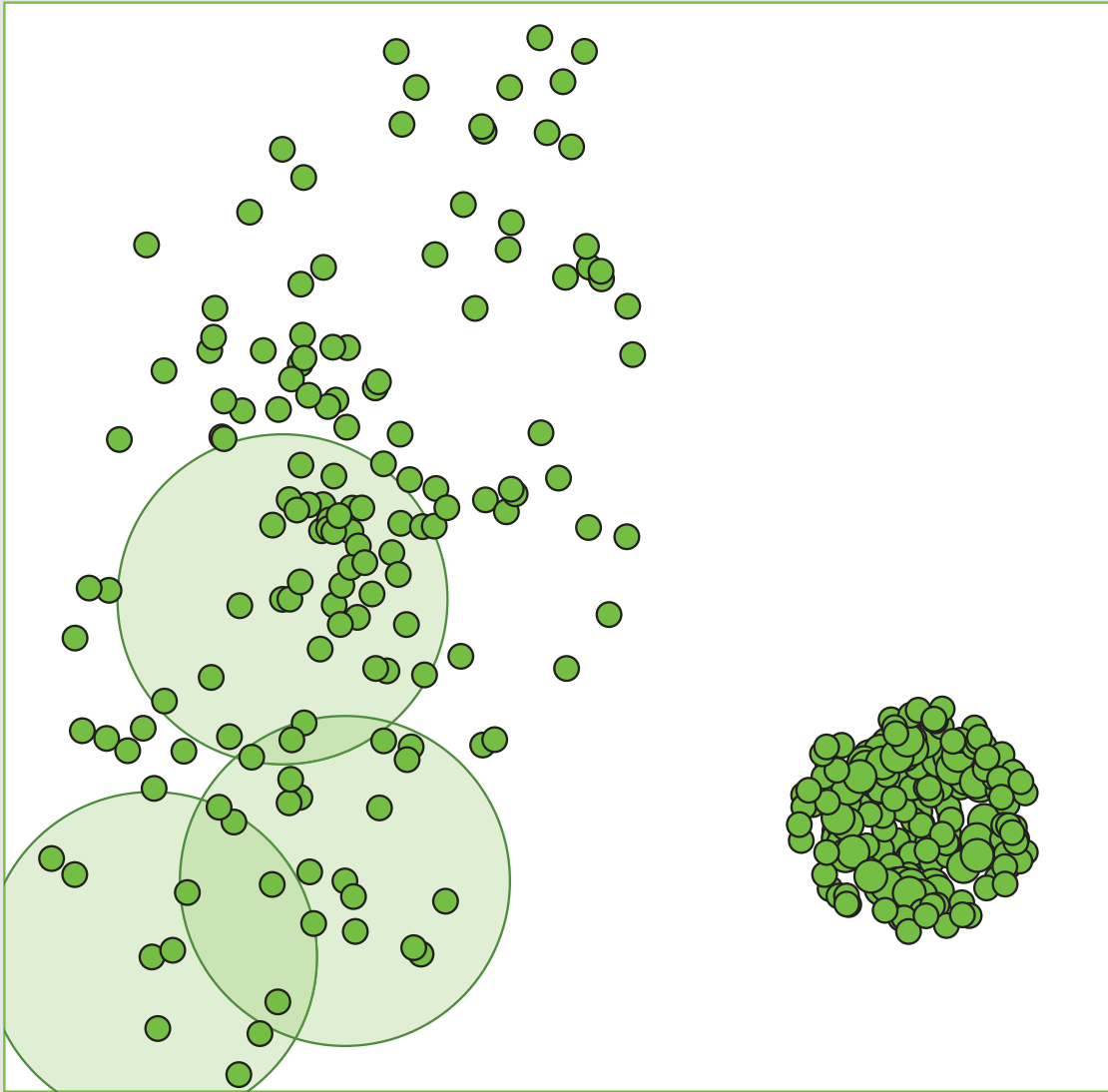## BICO – BIrch meets COresets
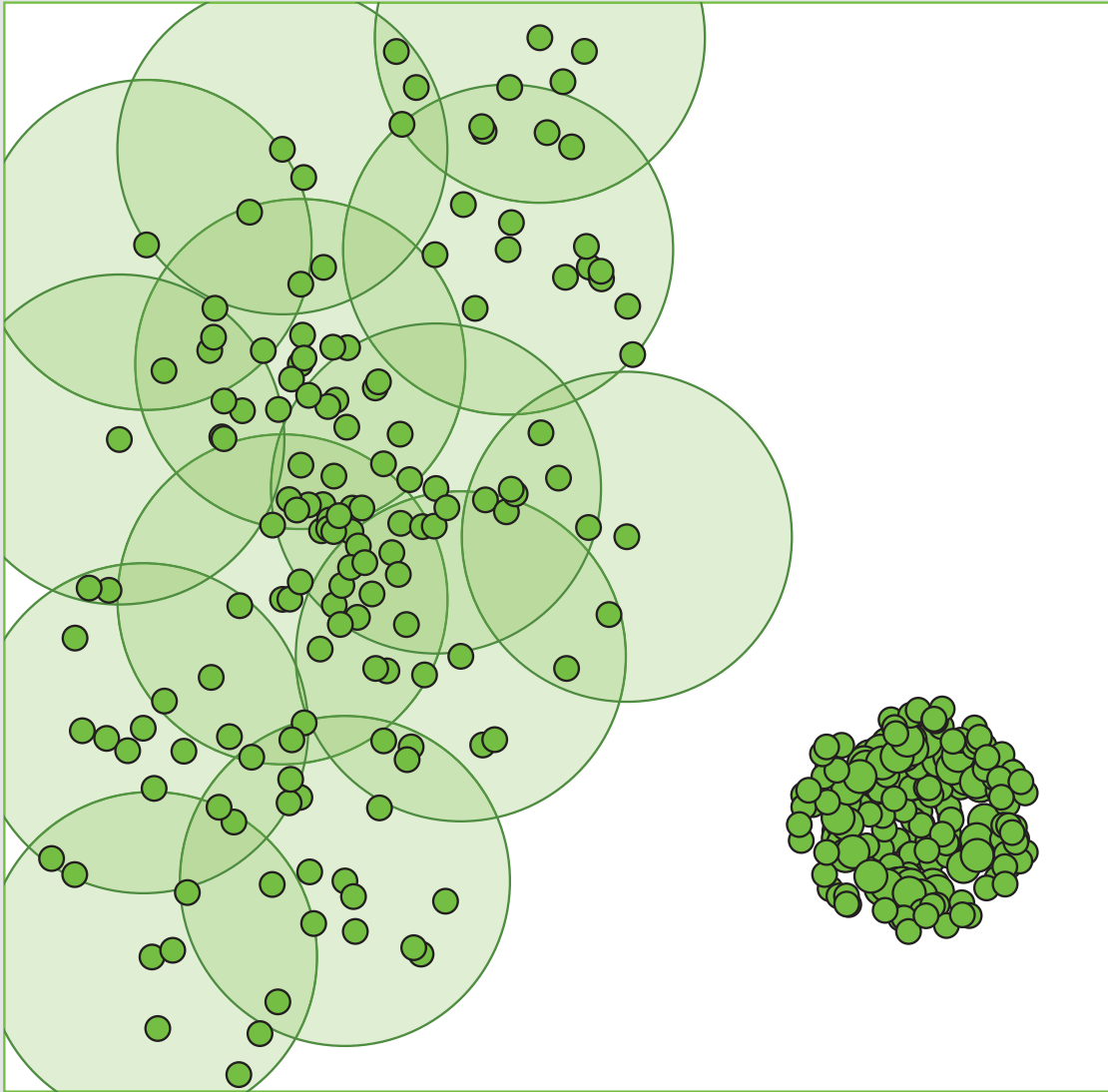

Idea

- Compute ball cover

Reduction Idea: Smart Covering
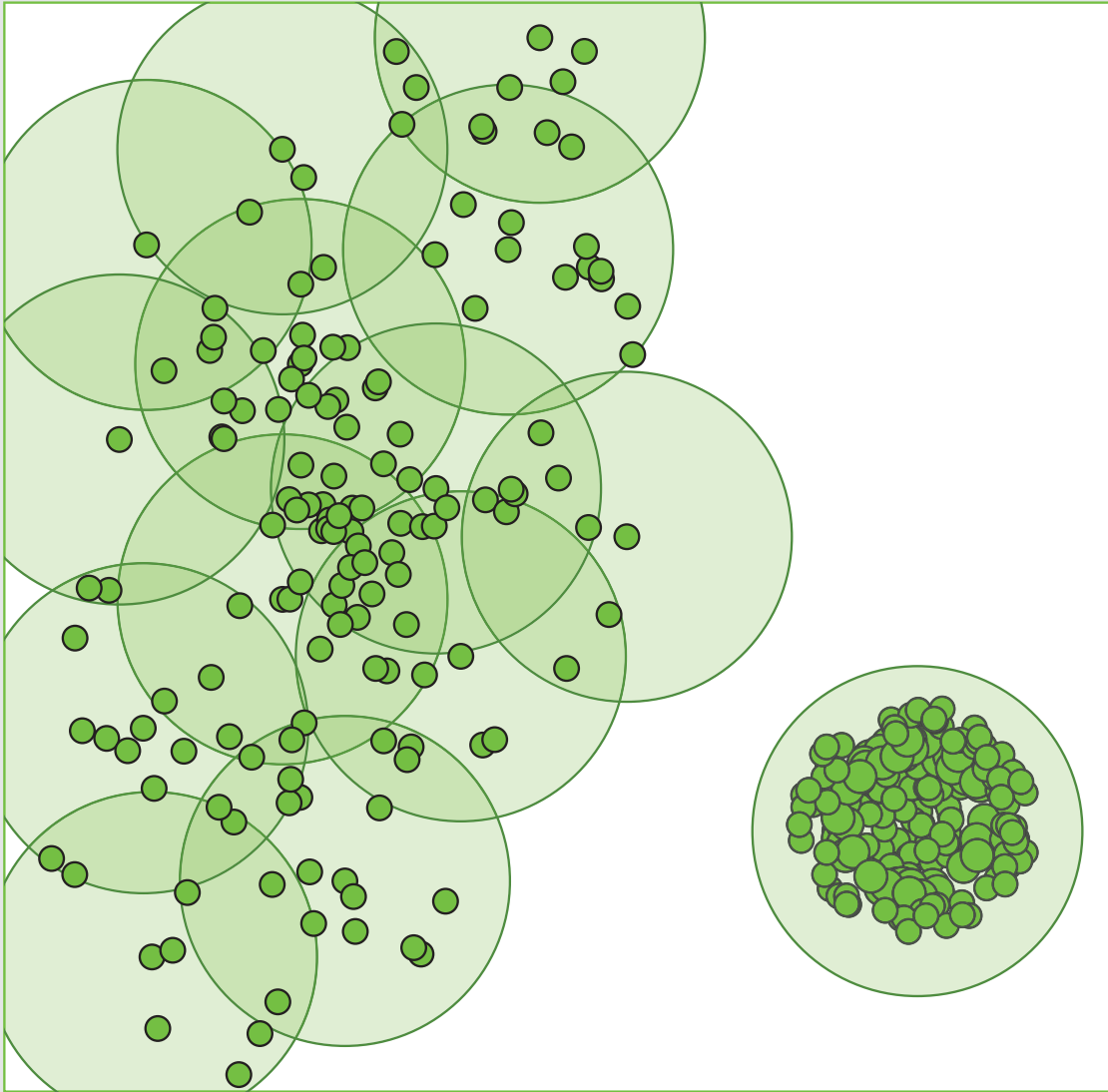
BICO – BIrch meets COresets

Idea

- Compute ball cover

Reduction Idea: Smart Covering

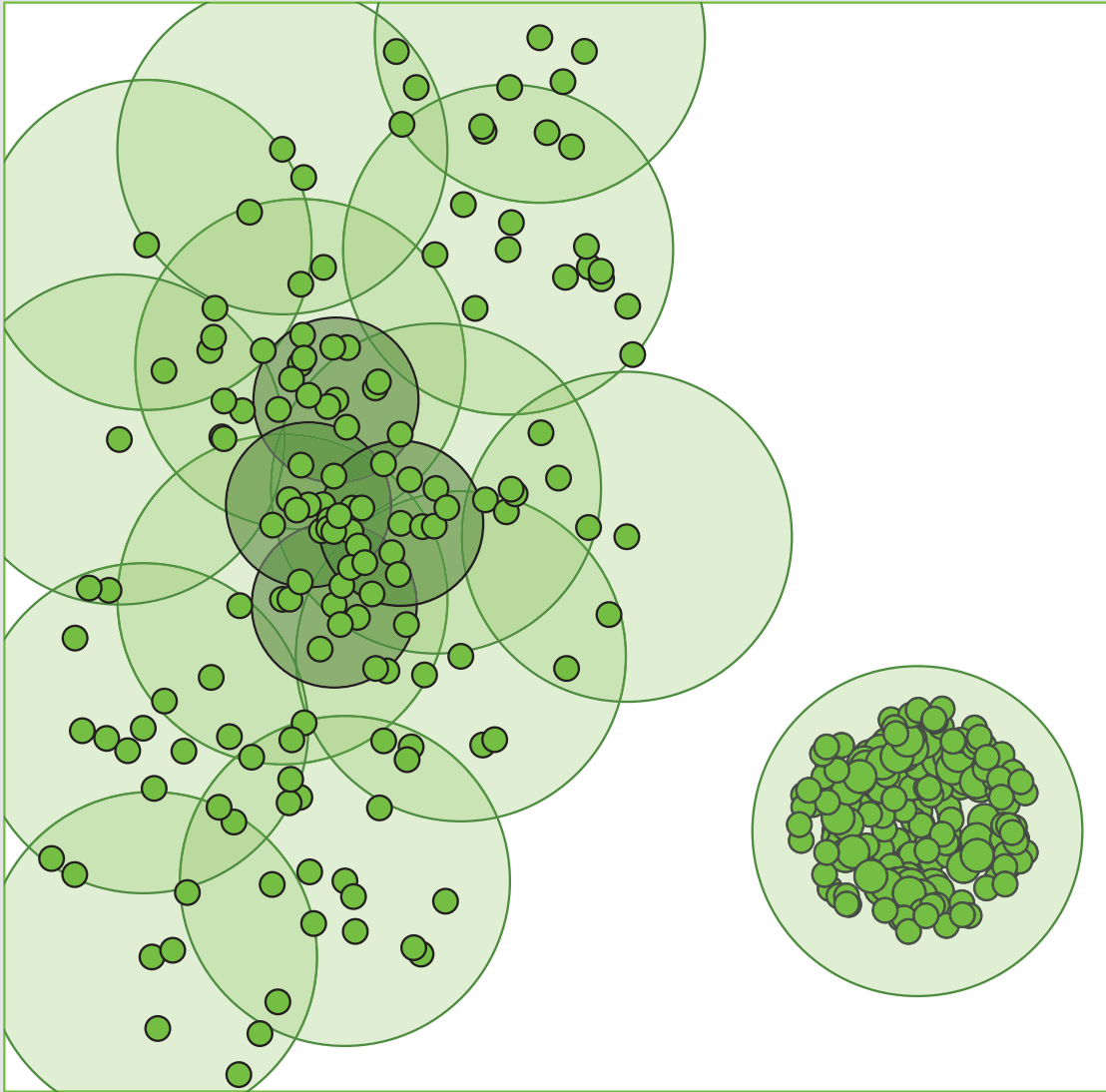## BICO – BIrch meets COresets

Idea

- Compute ball cover
- Replace points within each ball by center

Reduction Idea: Smart Covering

BICO – BIrch meets COresets

Idea

- Compute ball cover
- Replace points within each ball by center
- If ball gets too *heavy*, open new ball with smaller radius

Reduction Idea: Smart Covering
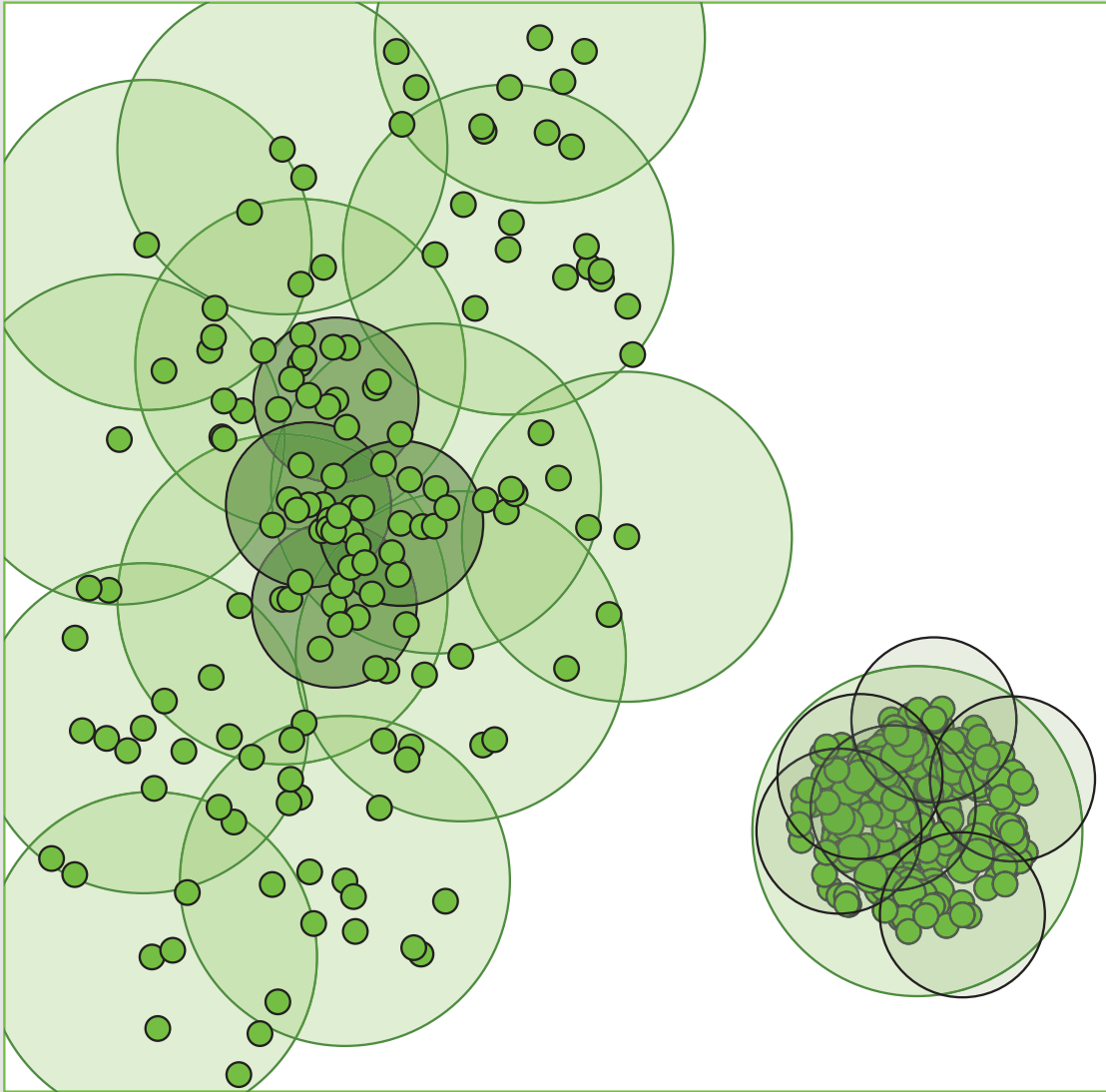
BICO – BIrch meets COresets

Idea

- Compute ball cover
- Replace points within each ball by center
- If ball gets too *heavy*, open new ball with smaller radius

Reduction Idea: Smart Covering
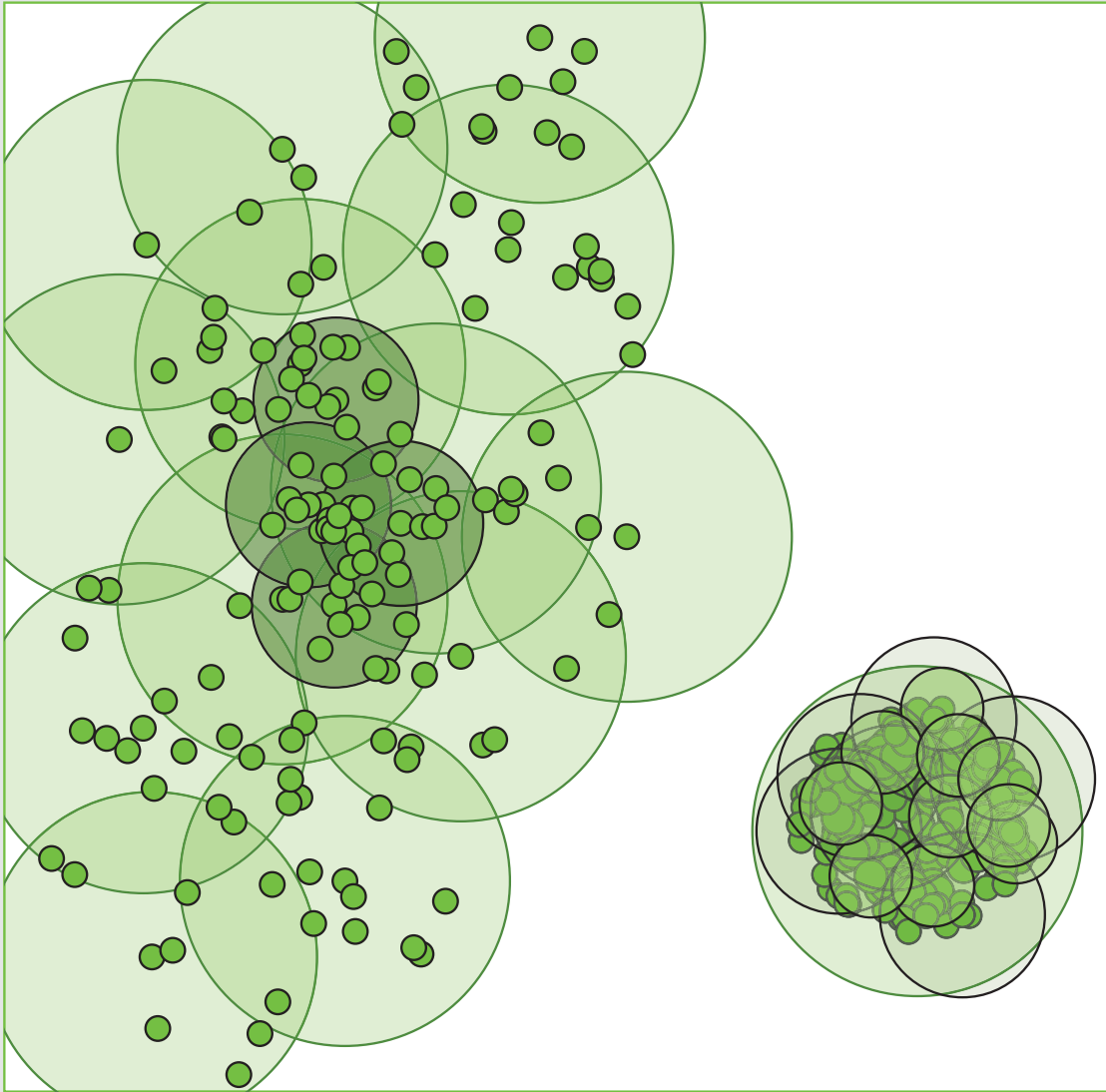
BICO – BIrch meets COresets
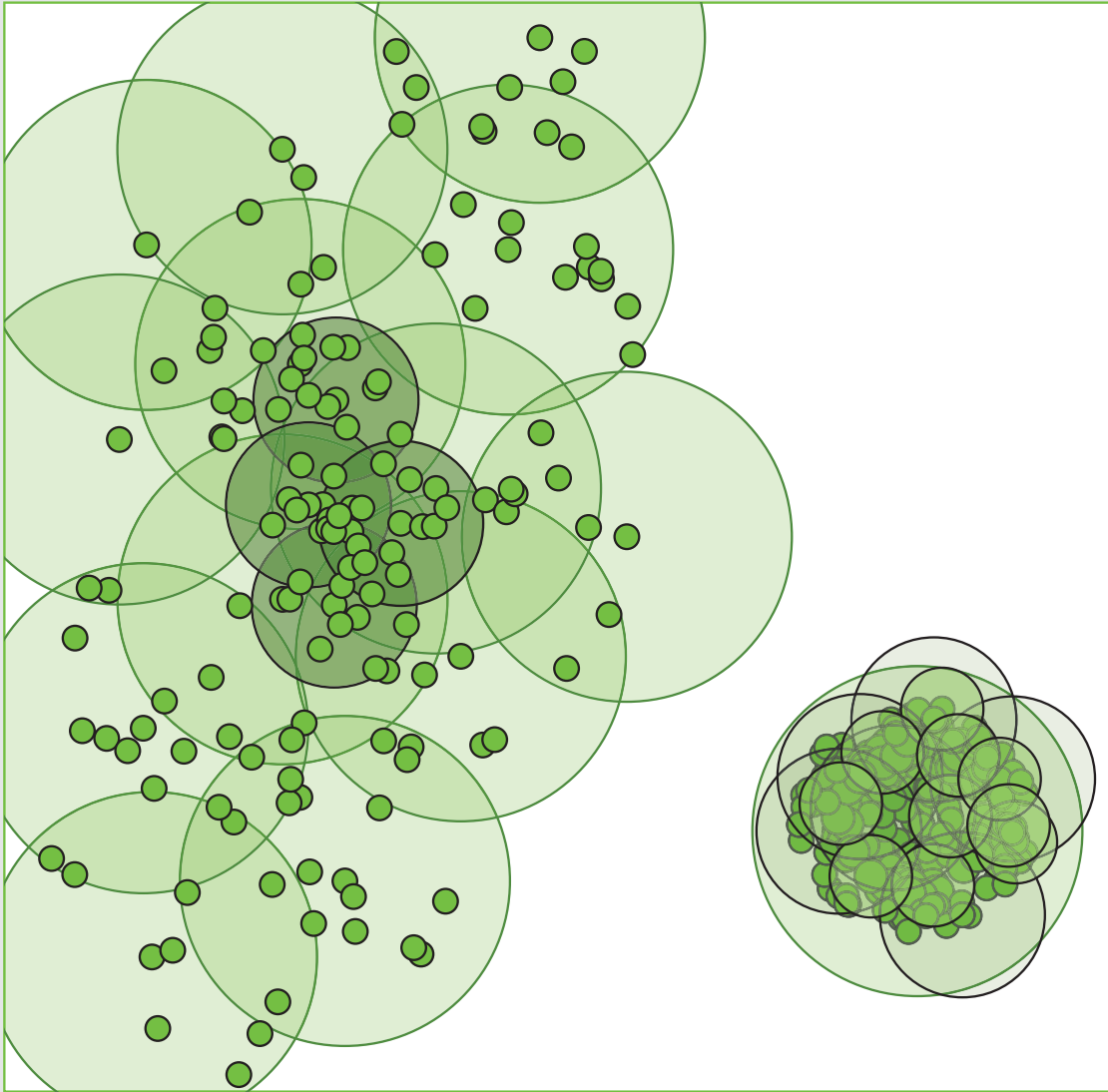

Idea

- Compute ball cover
- Replace points within each ball by center
- If ball gets too *heavy*, open new ball with smaller radius

Reduction Idea: Smart Covering

BICO – BIrch meets COresets

Idea

- Compute ball cover
- Replace points within each ball by center
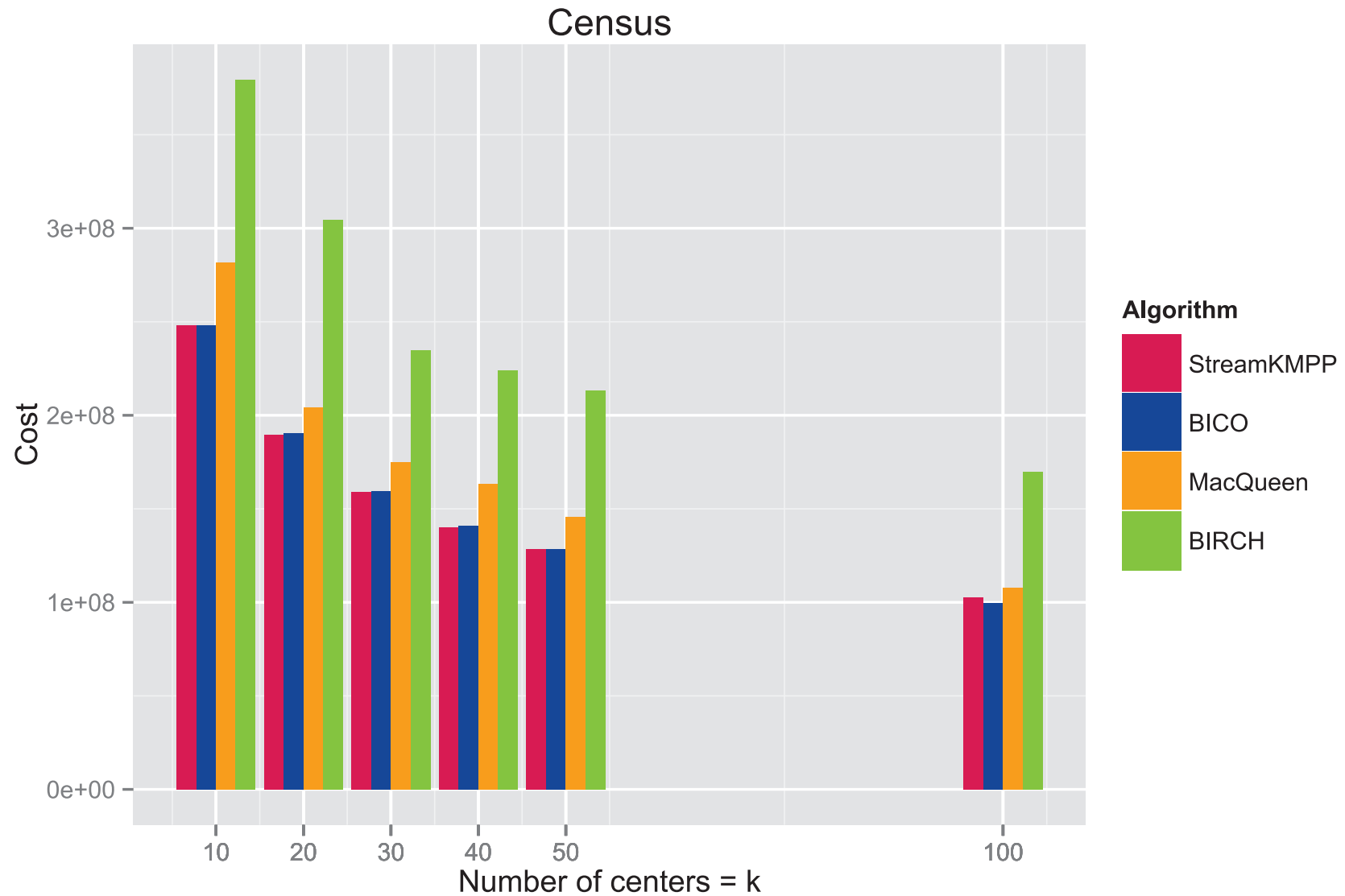- If ball gets too *heavy*, open new ball with smaller radius

Theoretical Result

Low dimensions:
Size $\approx k \log n \ll n$

Implementation

Good results with size $200k$

Census

Census