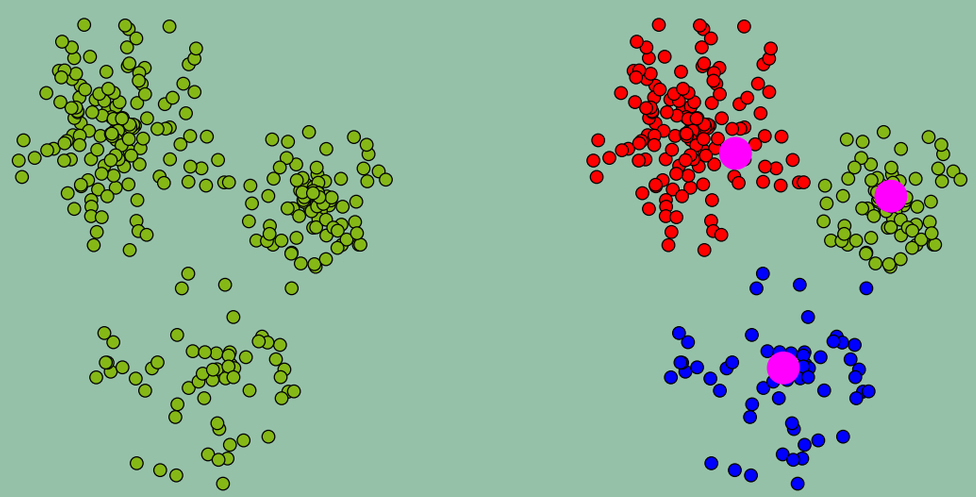
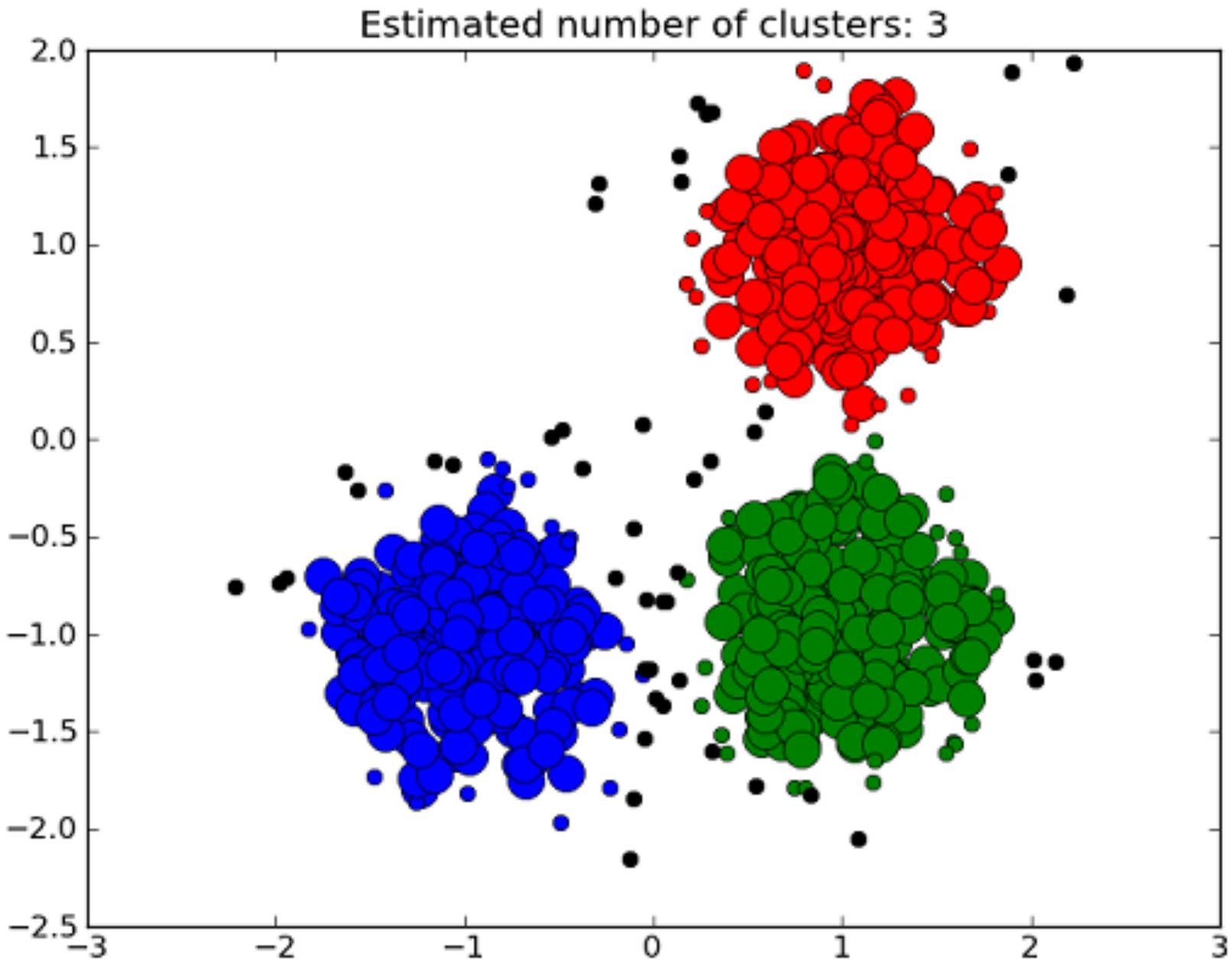
Local search for clustering-type problems

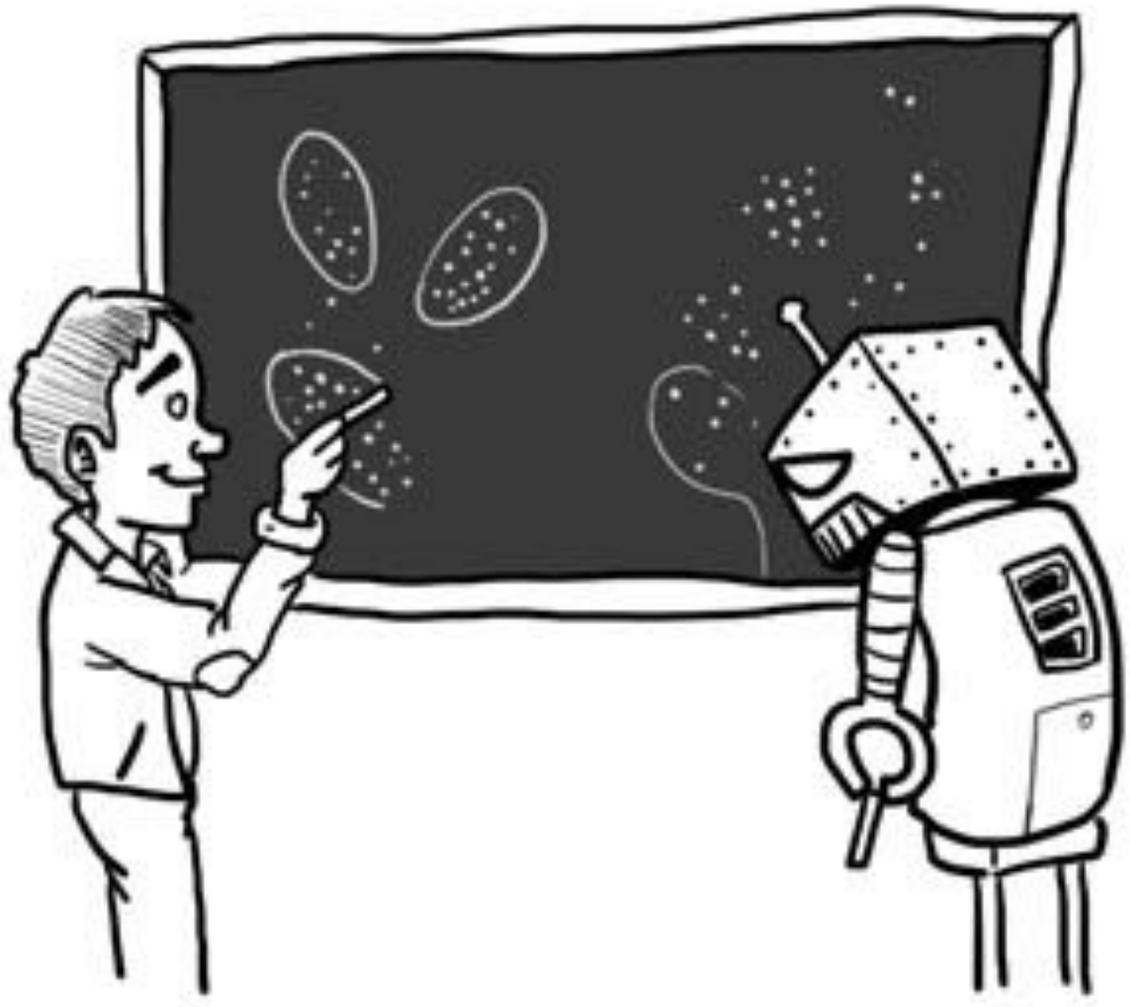


Clustering data



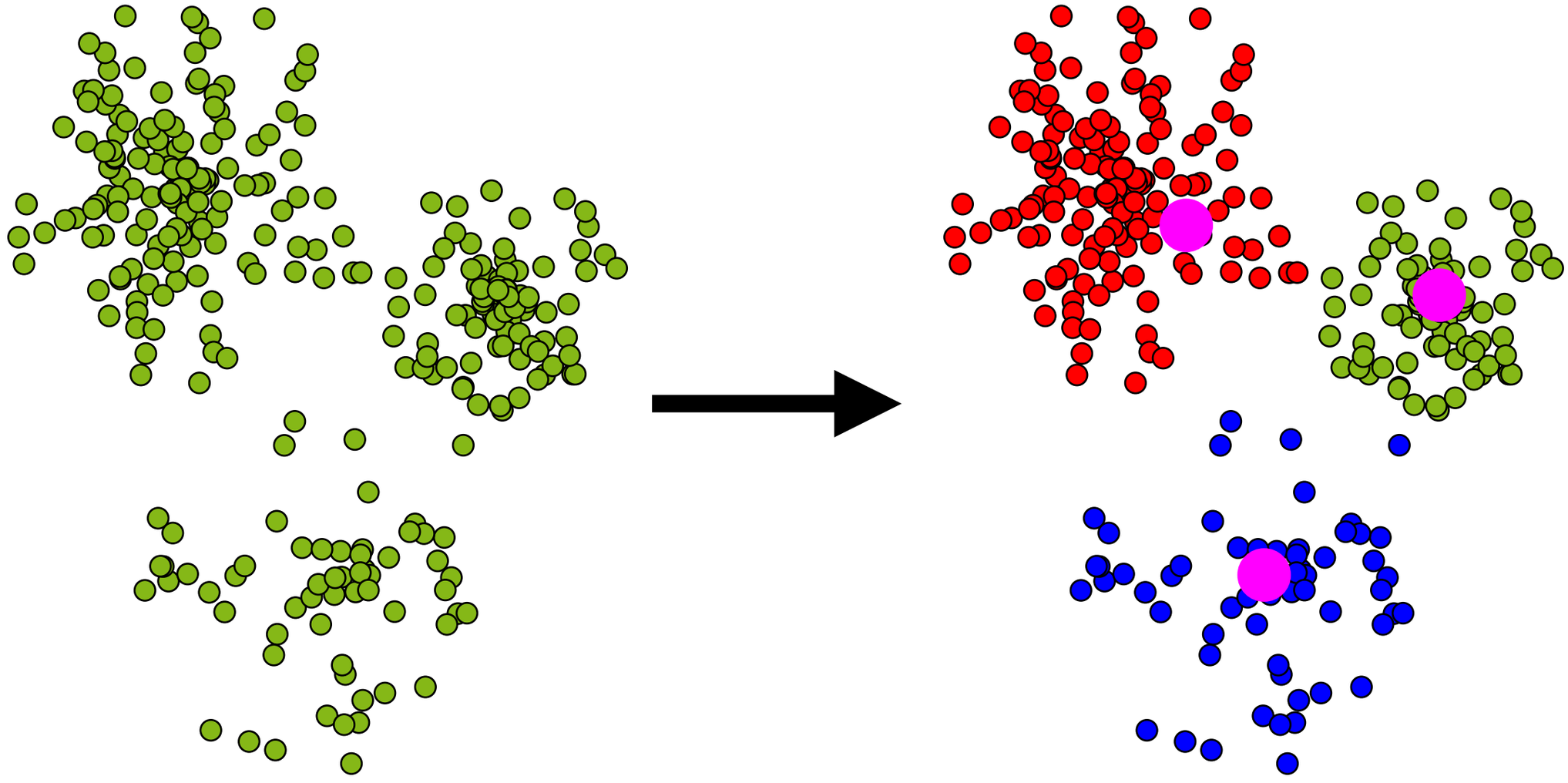
How do we find those clusters?

How do we find those clusters?



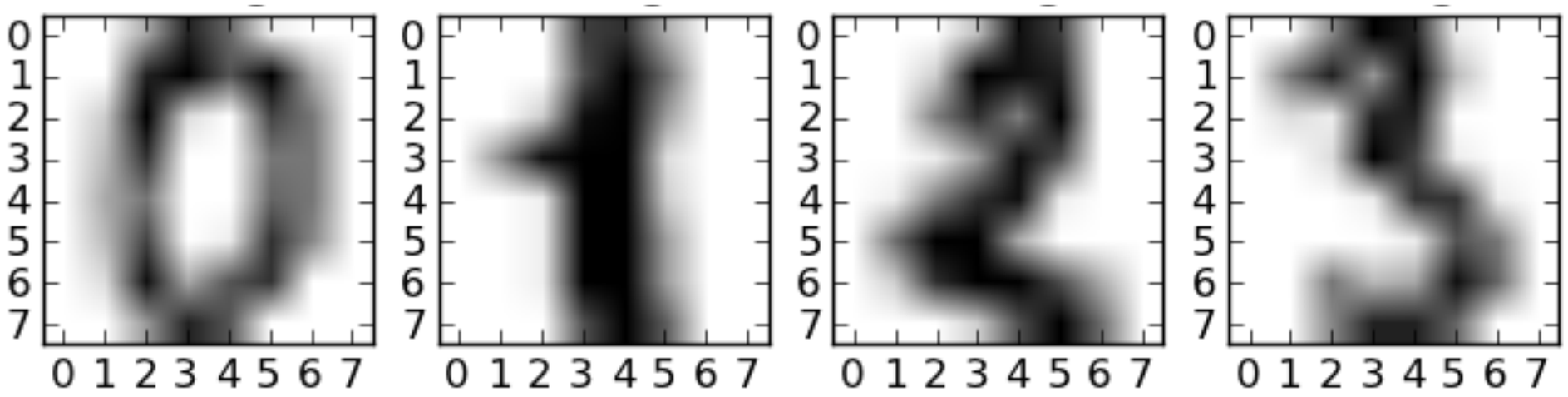
We eyeball them.

Center-based clustering



Recognizing digits

10 clusters: 0,1,2,3,4,5,6,7,8,9

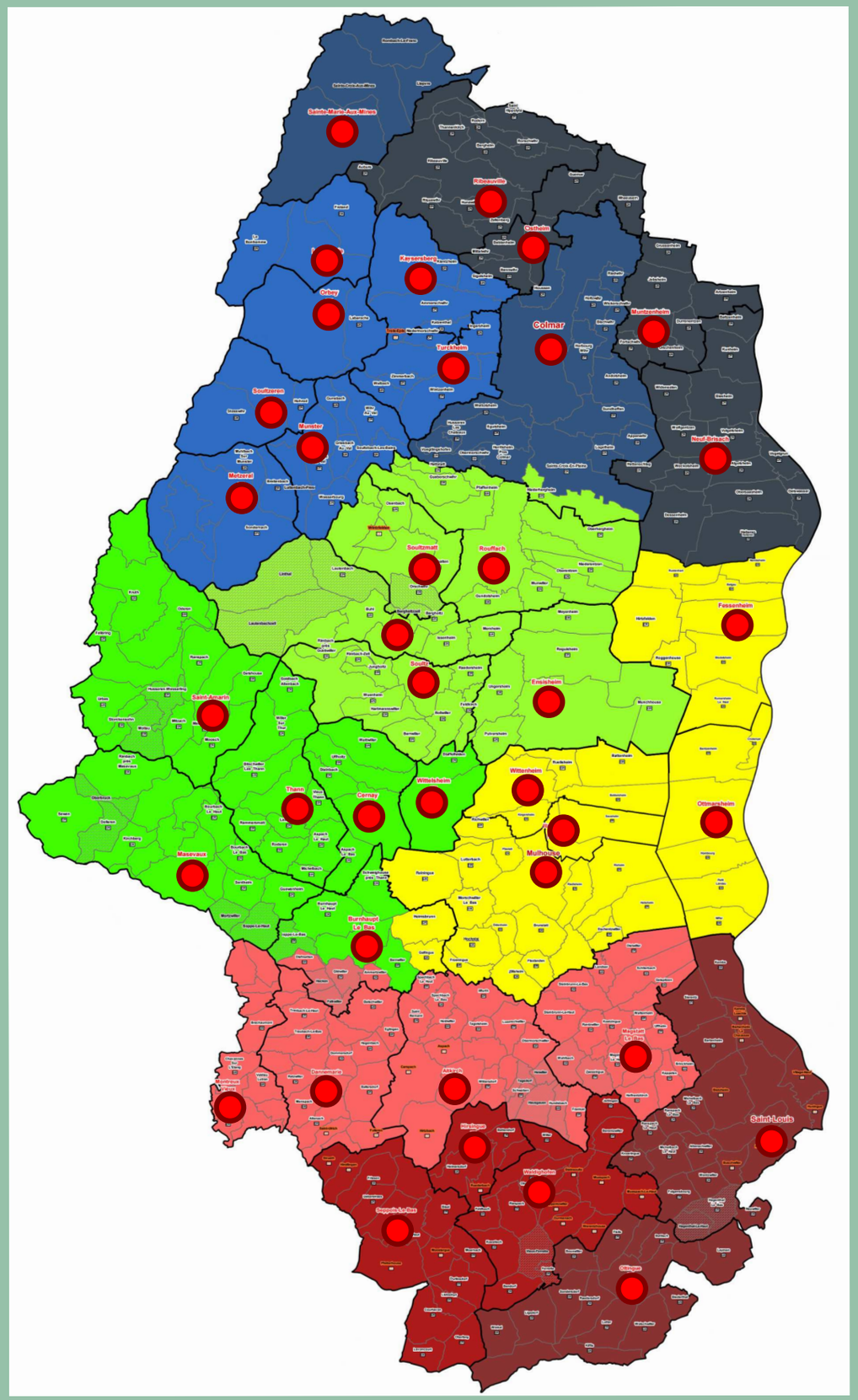


data "point" = image = 64 pixels

= element of [0,1]^64

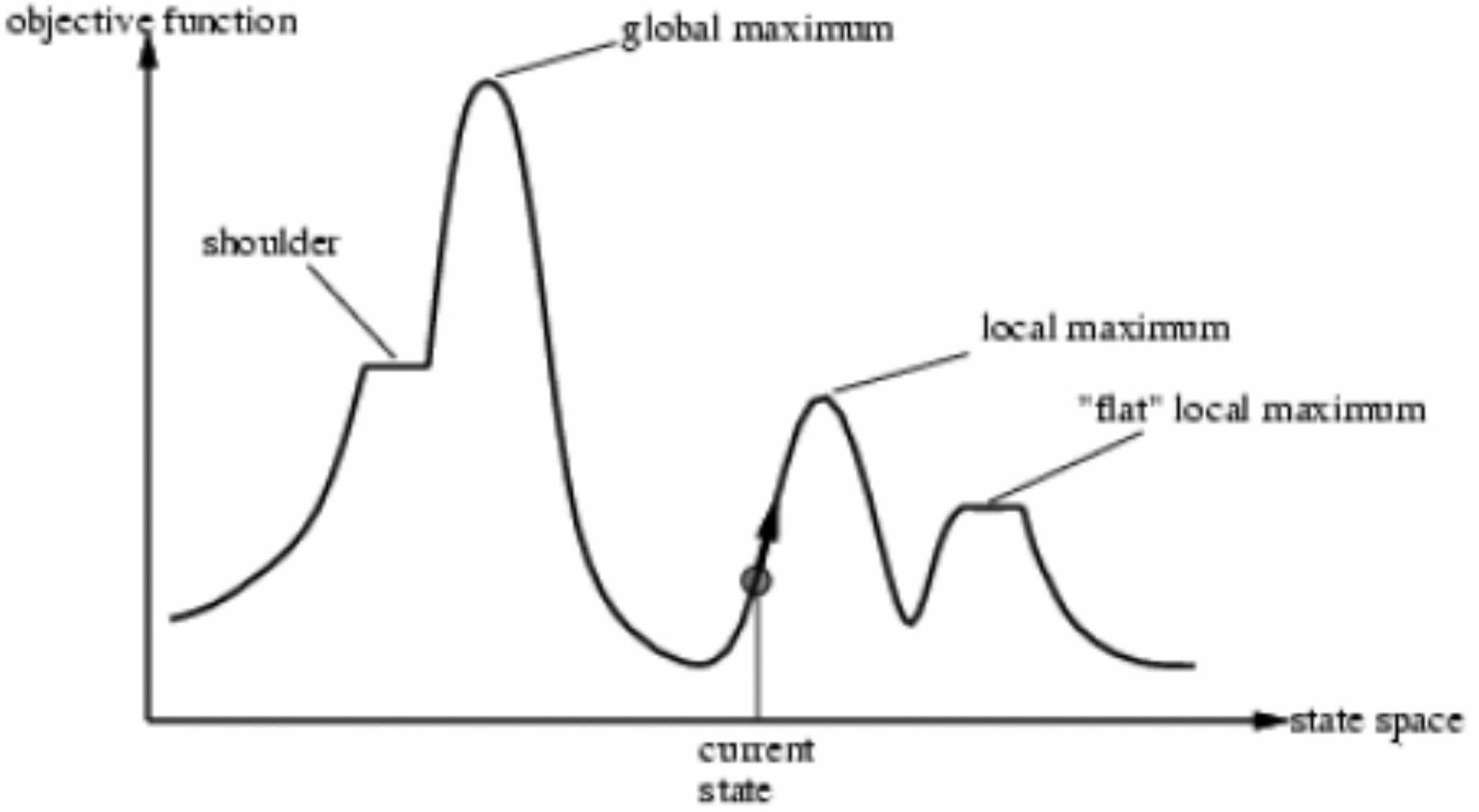
Project onto low dimension, then cluster

Clustering population: voting centers,...



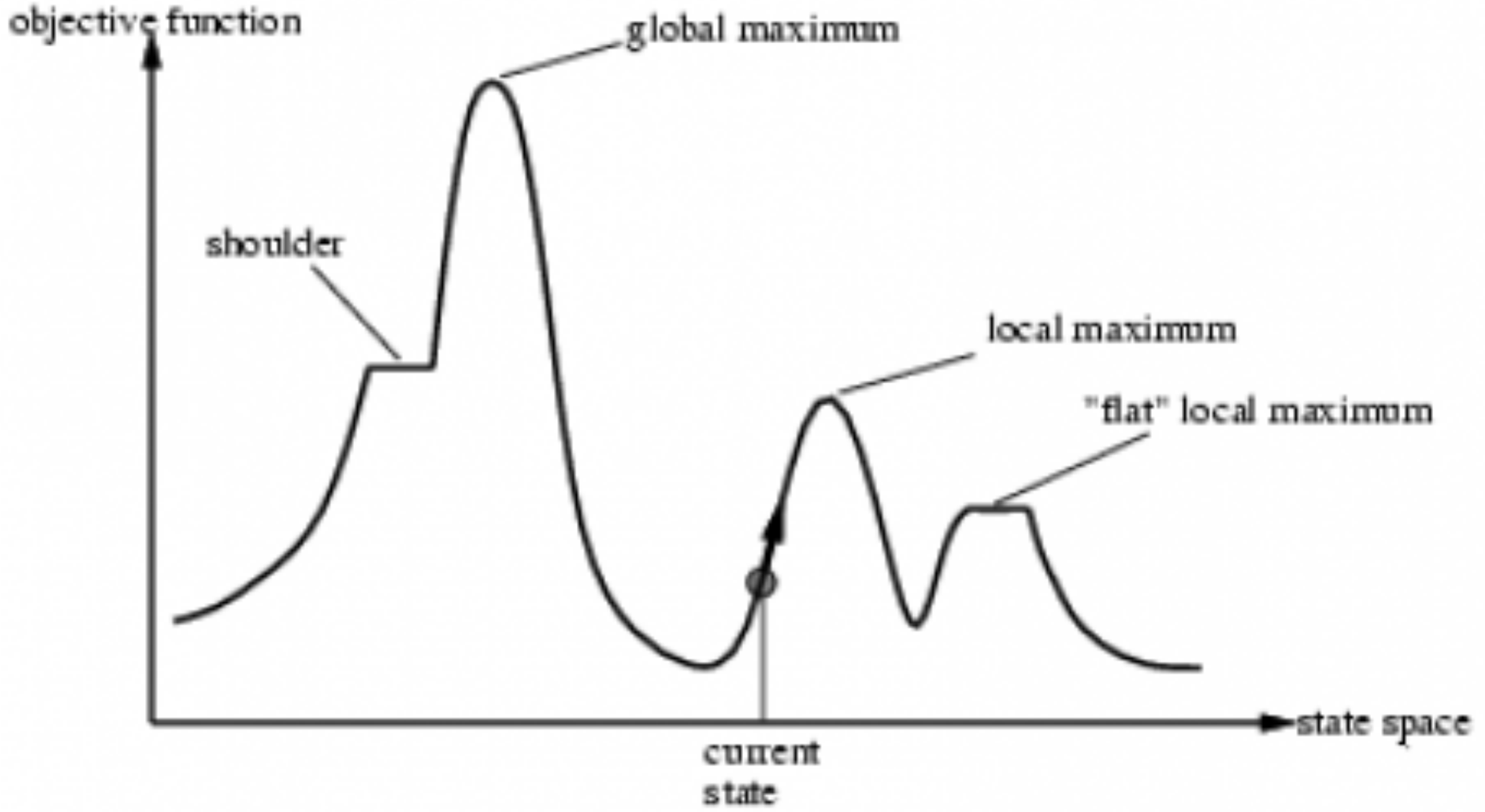
Clustering around centers

Local search heuristic



Example: hill-climbing methods

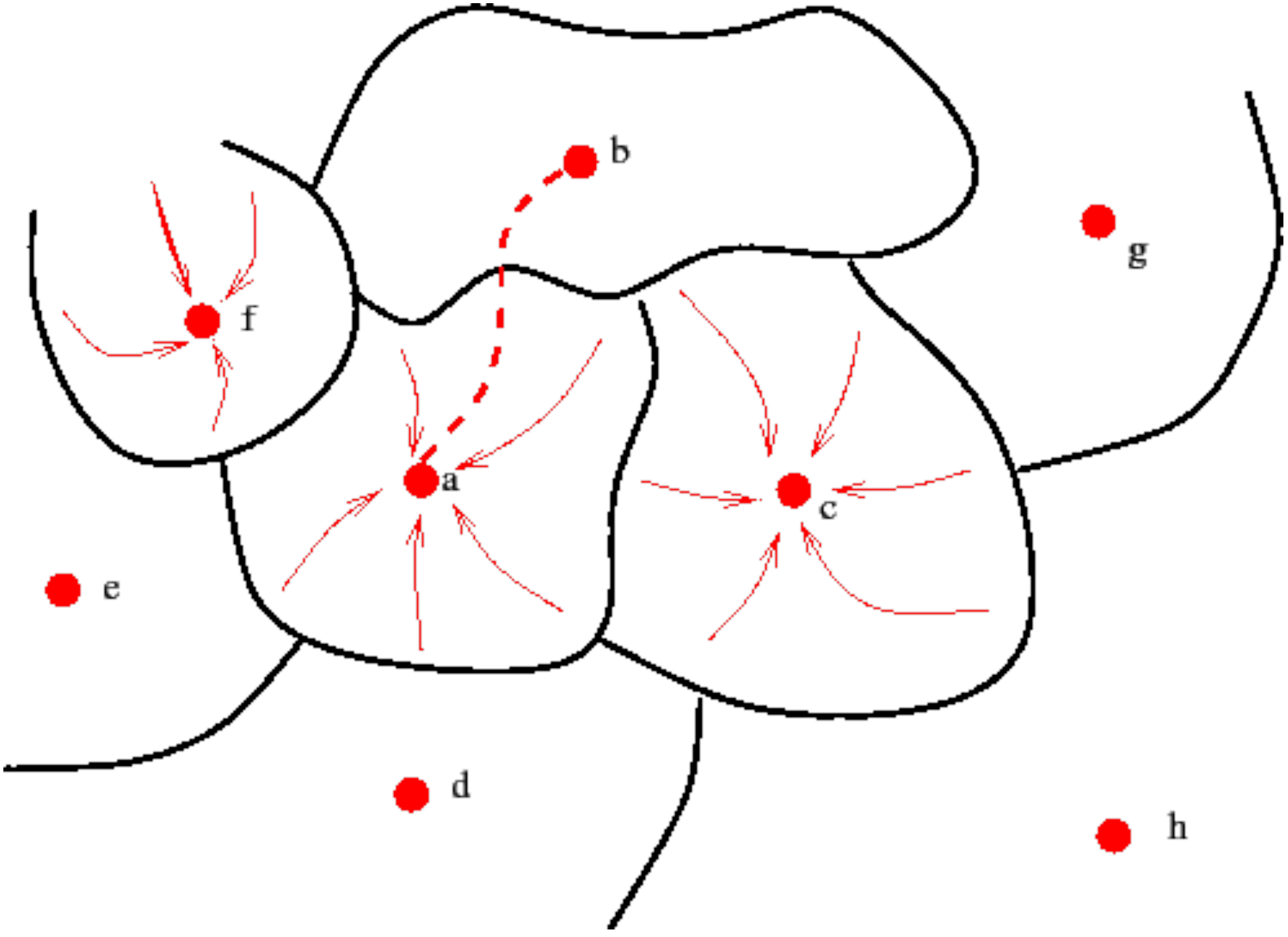
Comment: instead, put a photo of Massif de Belledonne



Local optimum can be good or bad

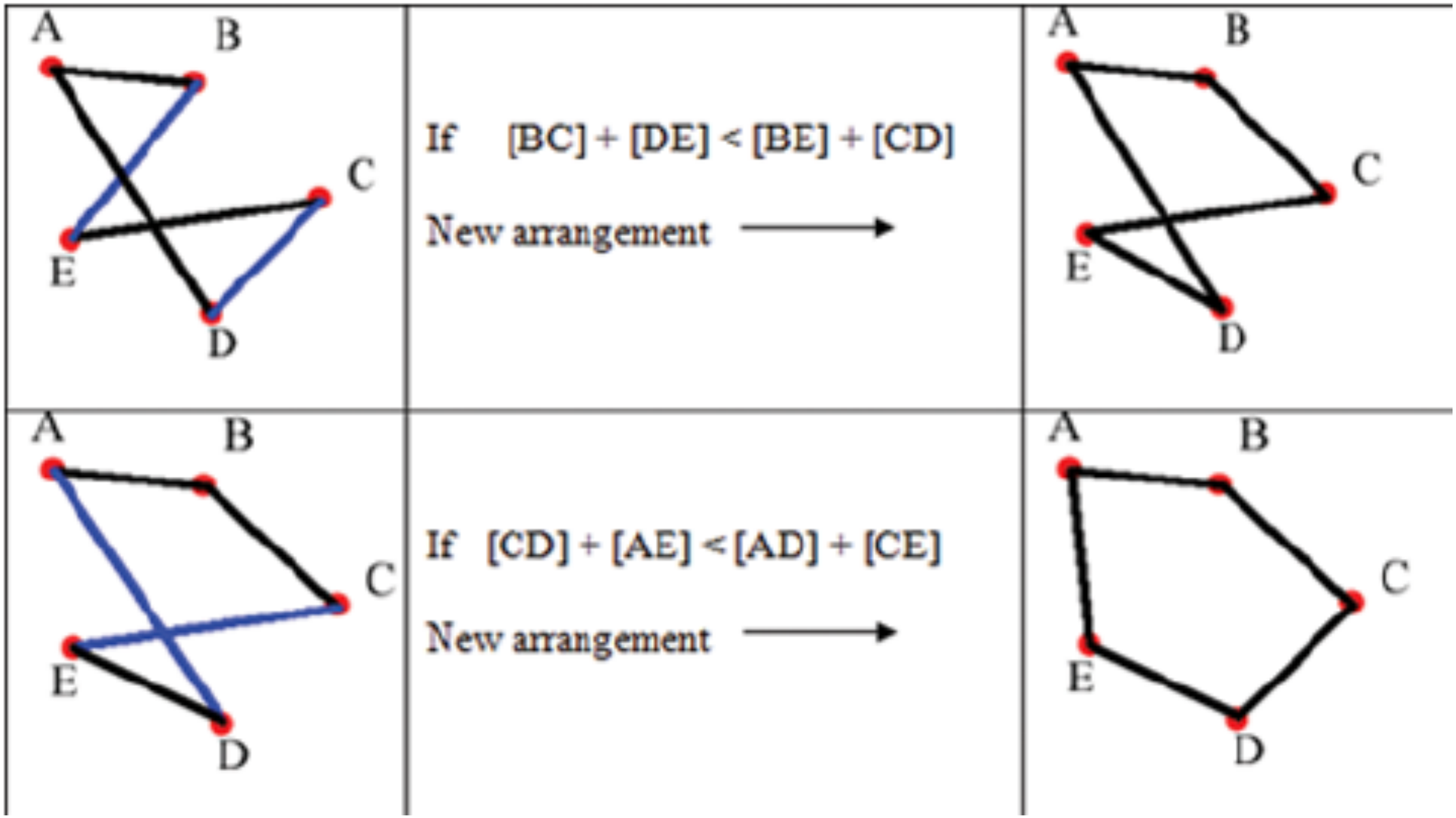
Where we end up depends

- where we start
- how we improve the solution



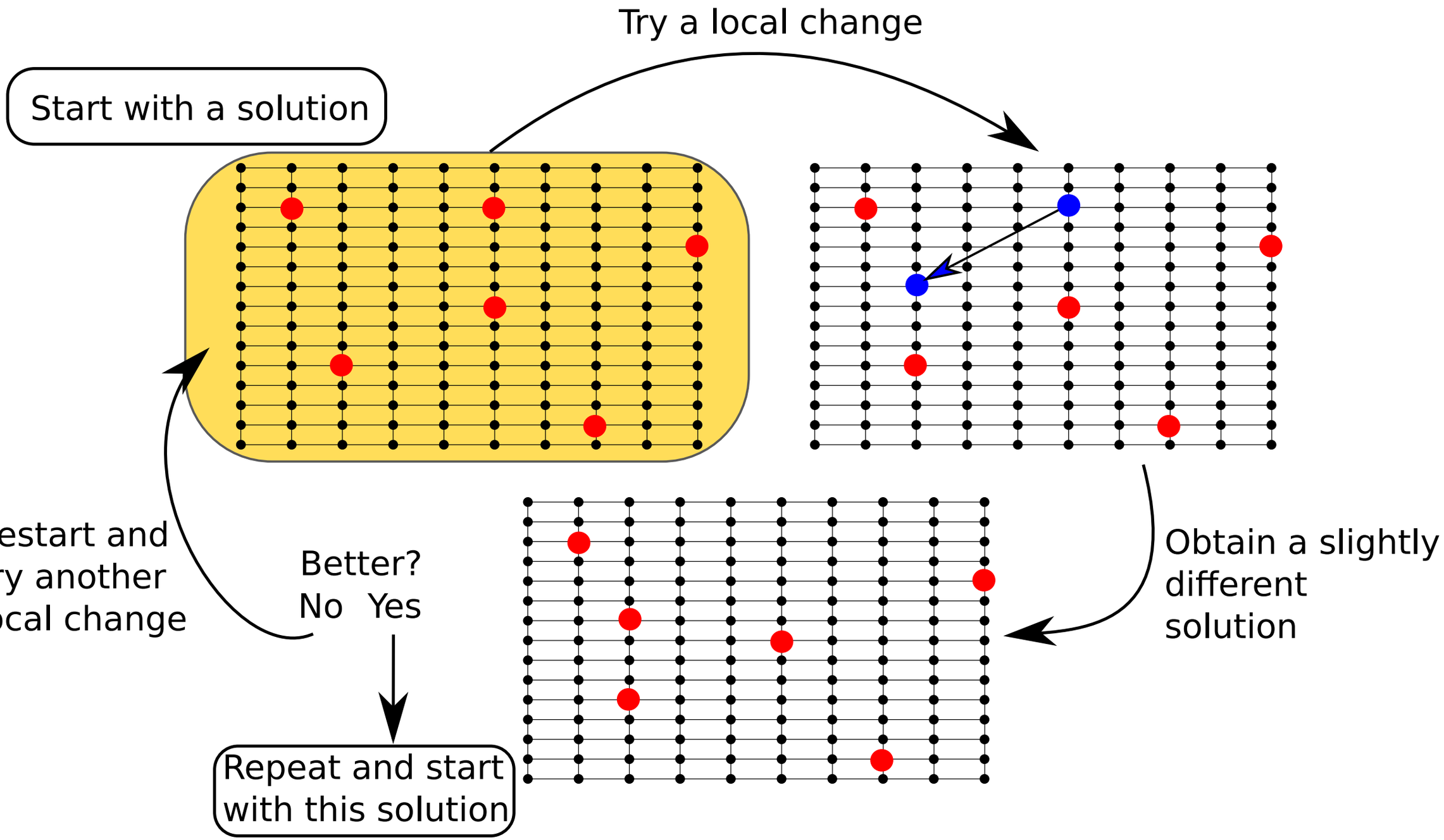
There can be many local optima

Example: Local search for TSP



On benchmark, local opt is within a few % of global opt

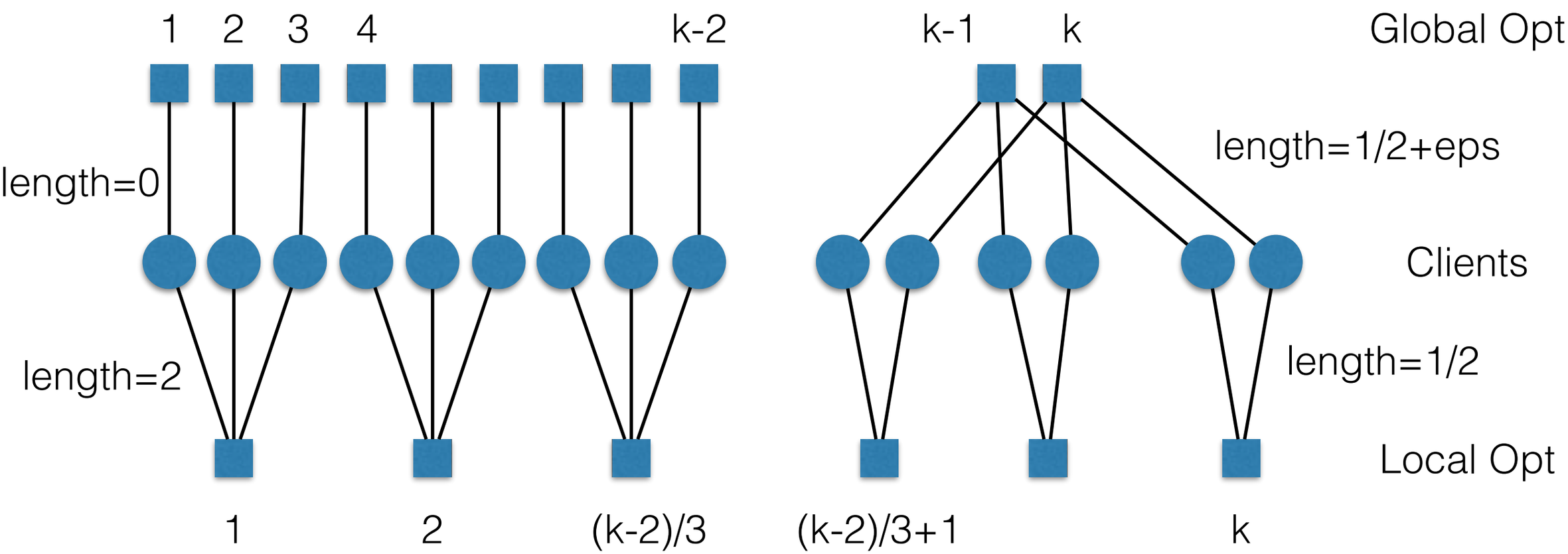
Local search heuristic



Cost=Sum of dist(client,closest center)

k centers

A local optimum



Cost of top solution = (1/2)*2*(k-k/3)=(2/3)*k

Cost of bottom solution = 2*k+(2/3)*k

Off by a factor of 4

Local opt even if we allow swapping 2 centers in and out

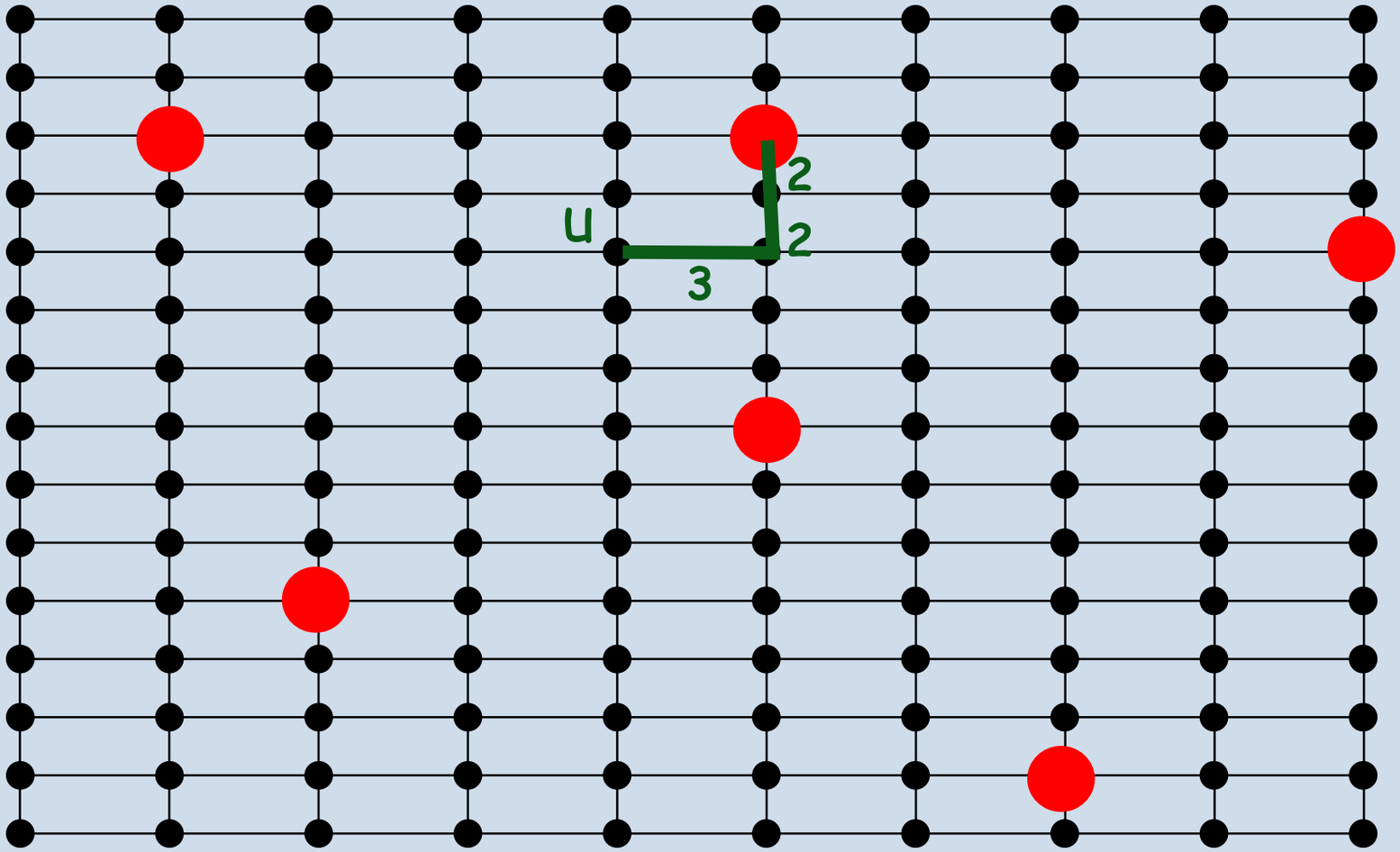
A local optimum that is not globally optimum

Even if we allow swapping up to p centers in and out, local opt can be off by a factor of 3+2/p

How can local search be better?
Only in special cases:
Restrict attention to planar metrics

Comment: put a picture of a roadmap here to motivate

Facility location, weighted planar graph



dist(u,closest center)=3+2+2=7

f: cost of opening a facility

Cost=#(facilities)*f+sum of dist(clients,closest center)

Extended local search

S : current set of centers

Repeat:

Find better solution S' among sets that differ from S in at most $1/\epsilon^2$ centers

Replace S by S'

Until: local optimum

Polynomial-time approximation scheme:
cost < (1+epsilon)*OPT

Extended local search yields a polynomial-time approximation scheme (PTAS) for facility location in edge-weighted planar graphs

Proof

S : current set of centers

Repeat:

Find better solution S' among sets that differ from S in at most $1/\epsilon^2$ centers

Replace S by S'

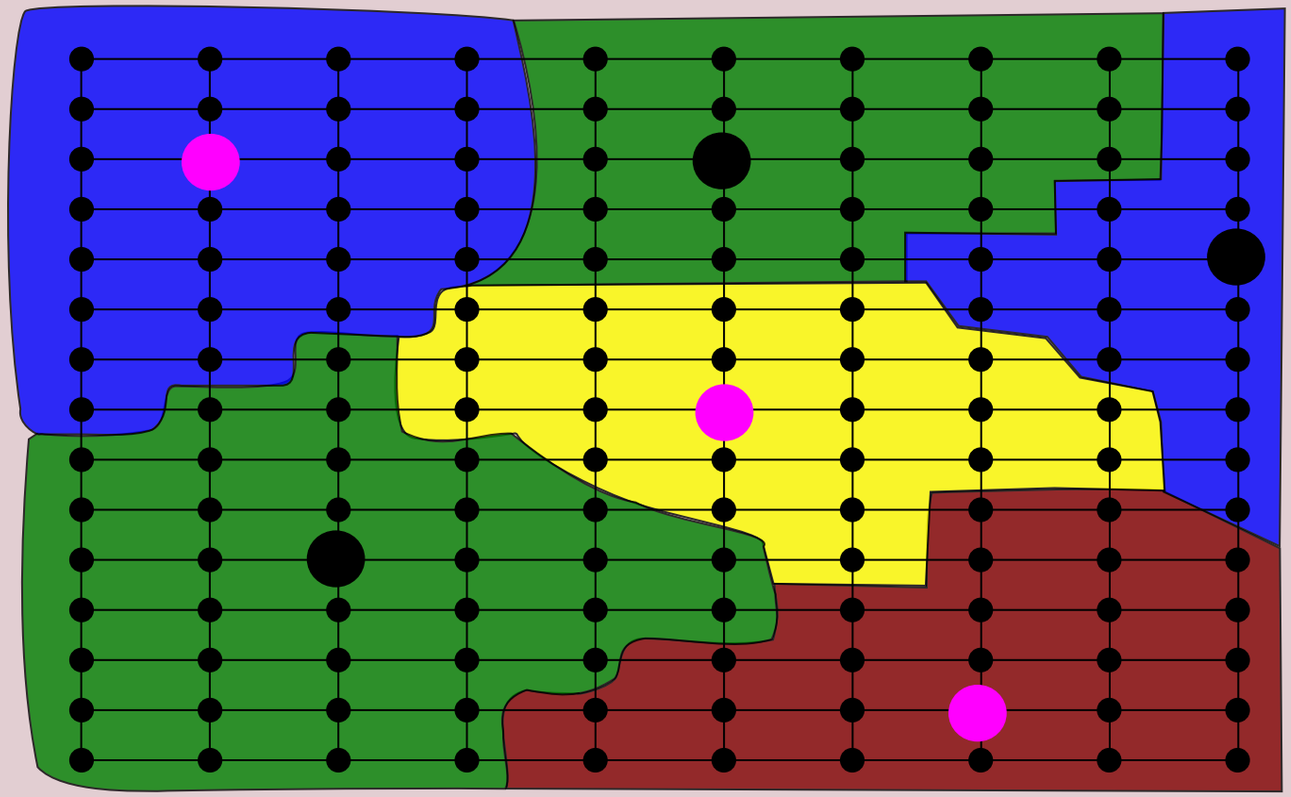
Until: local optimum

Cost analysis: Why does a locally optimal solution have cost at most $(1+\epsilon)\mathbf{OPT}$?

Voronoi cells

Notations for facilities
L: local optimum
C: global optimum
F: L union C

map client u —> closest facility
defines Voronoi cells



Contraction

Notations for facilities

L: local optimum

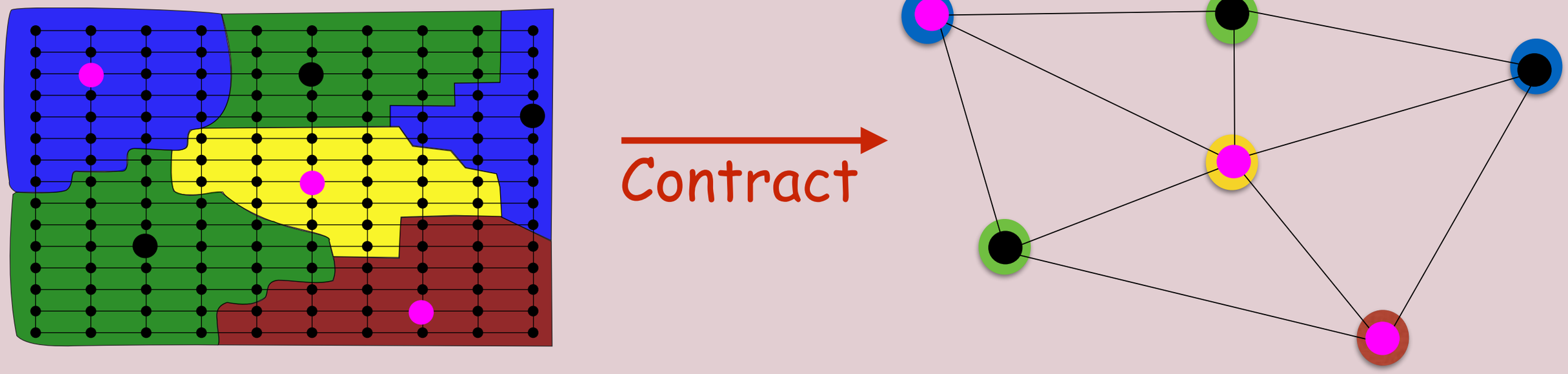
C: global optimum

F: L union C

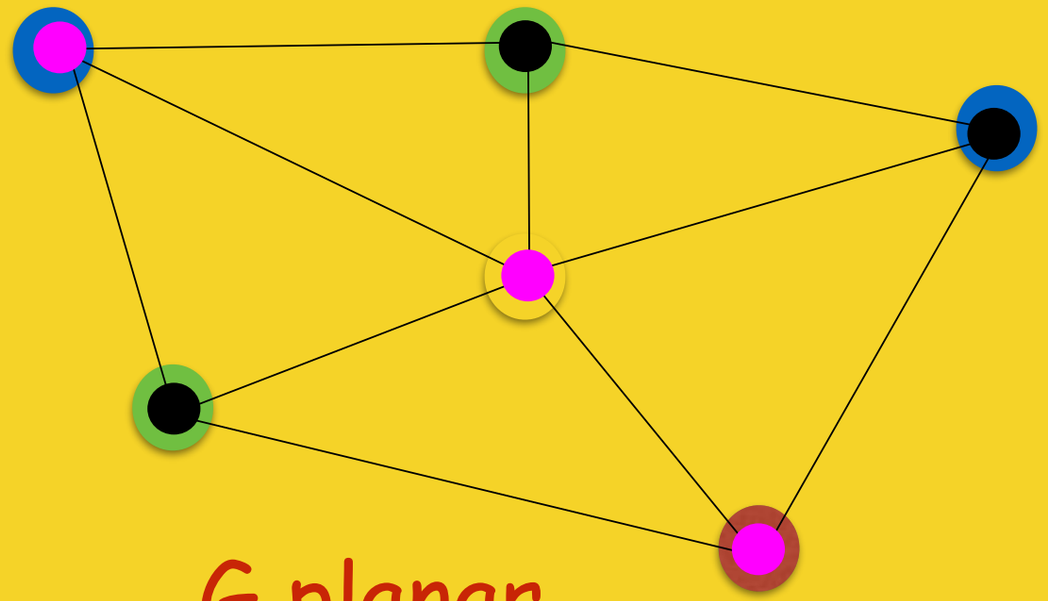
map client u —> closest facility

defines Voronoi cells

contract cells : graph G



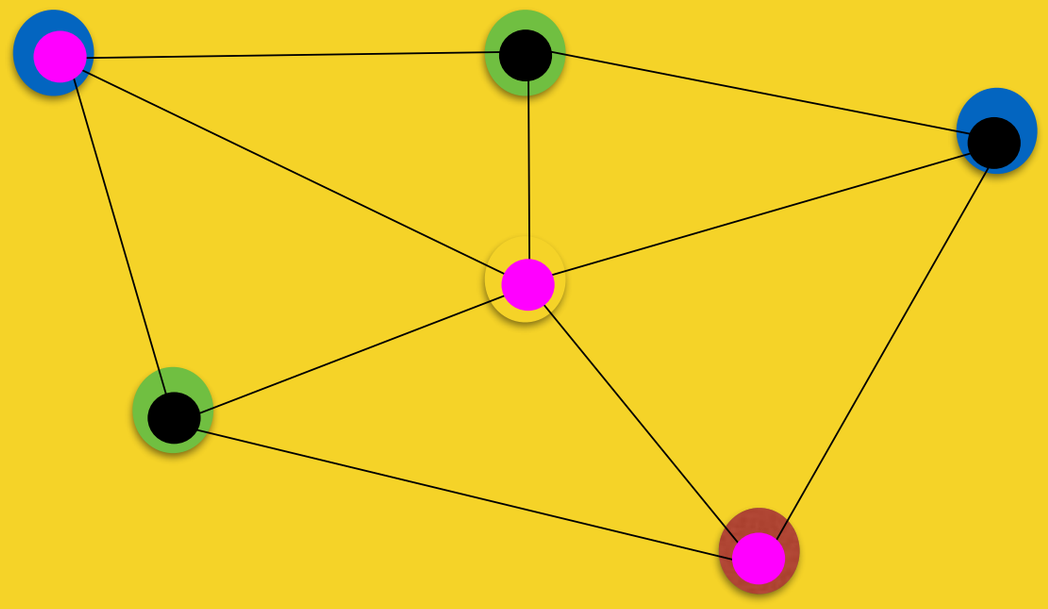
(Parenthesis on planar graphs, 1/3)



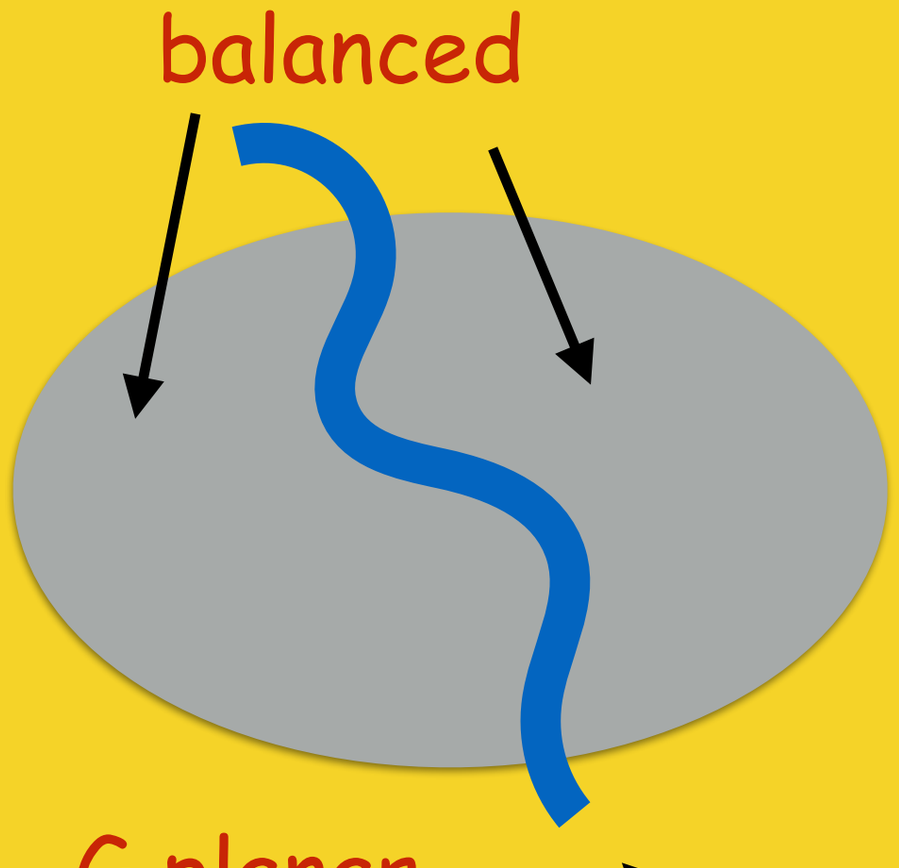
G planar

What do we know about planar graphs?

(Parenthesis on planar graphs, 2/3)



G planar



G planar

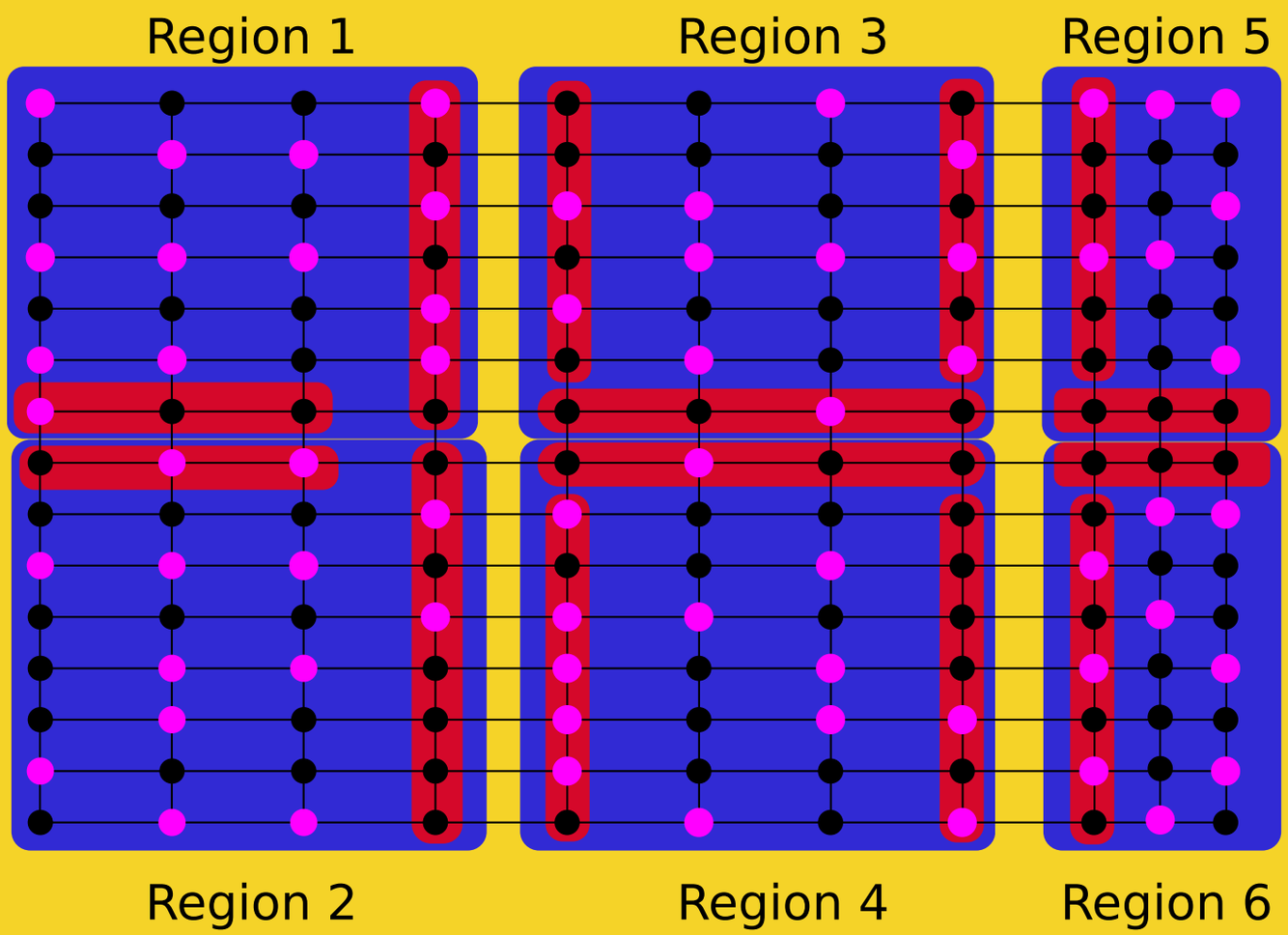
separator

Lipton-Tarjan: if G is planar then there is a small roughly balanced separator: $\mathbf{O}(\sqrt{\mathbf{n}})$ vertices

(Parenthesis on planar graphs, 3/3)

Corollary: if G planar then there is a partition into

- small regions: at most $1/\epsilon^2$ vertices in each region
- with small boundaries : at most ϵn boundary vertices

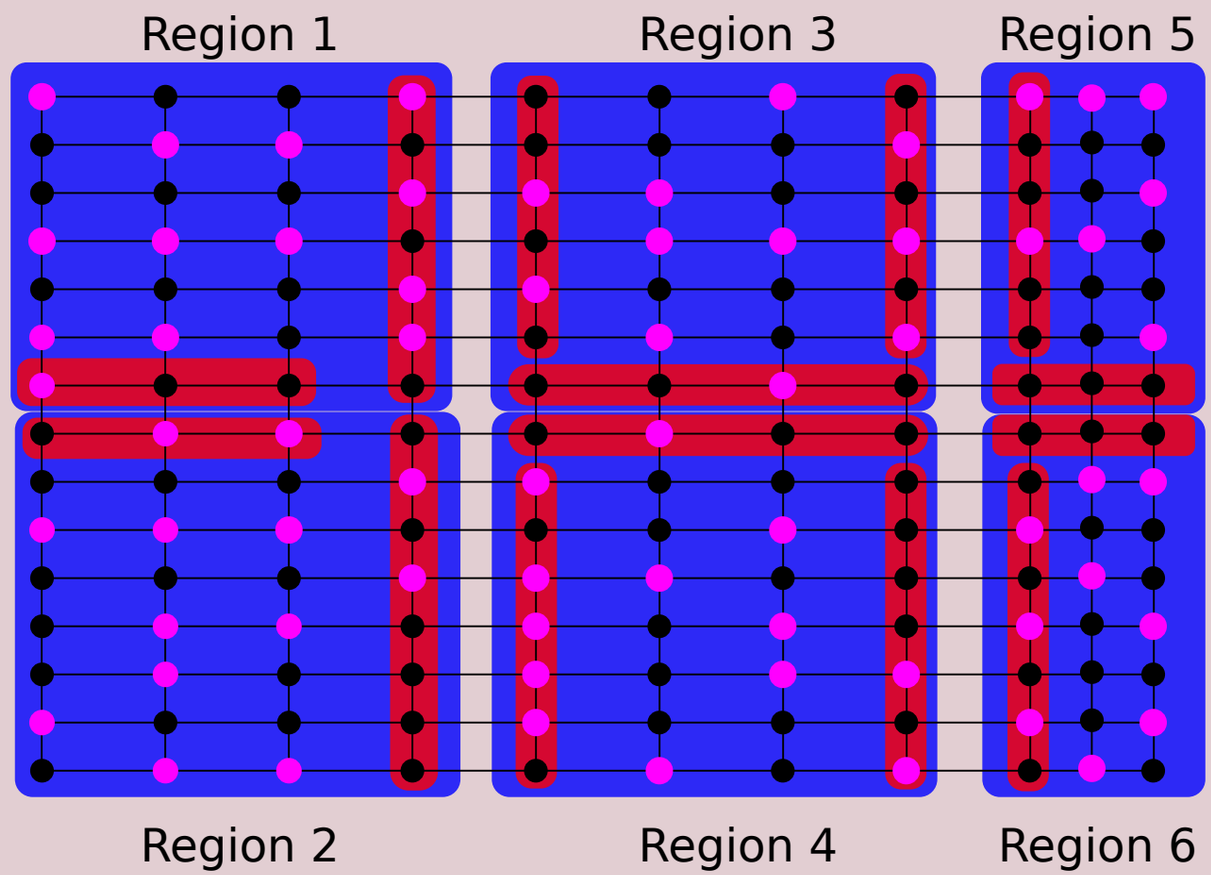


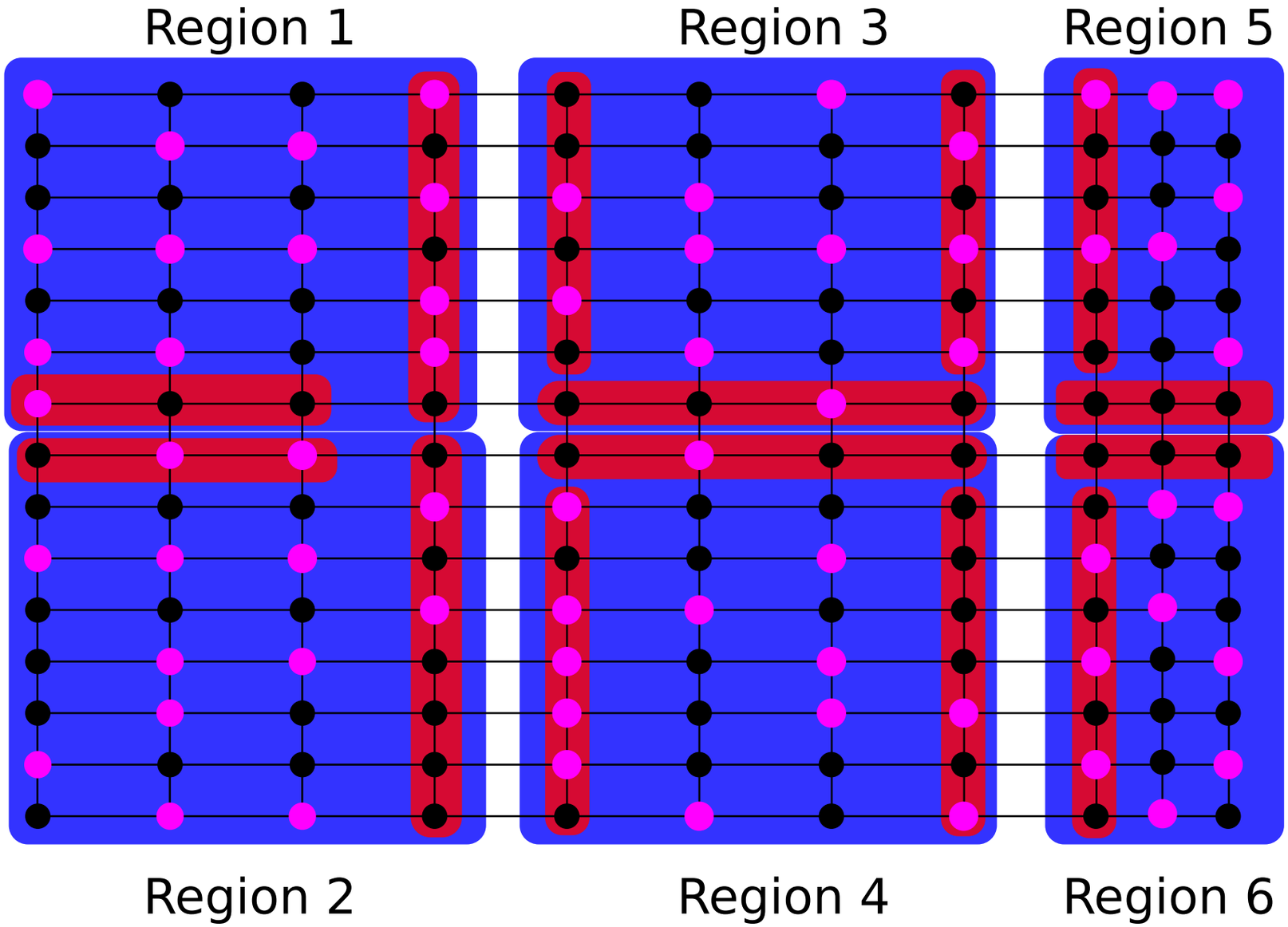
L: local optimum
C: global optimum
F: L union C

client u —> closest facility
Voronoi cells
G: after contracting cells

Partition G into small regions with small boundaries

B: centers of boundary regions



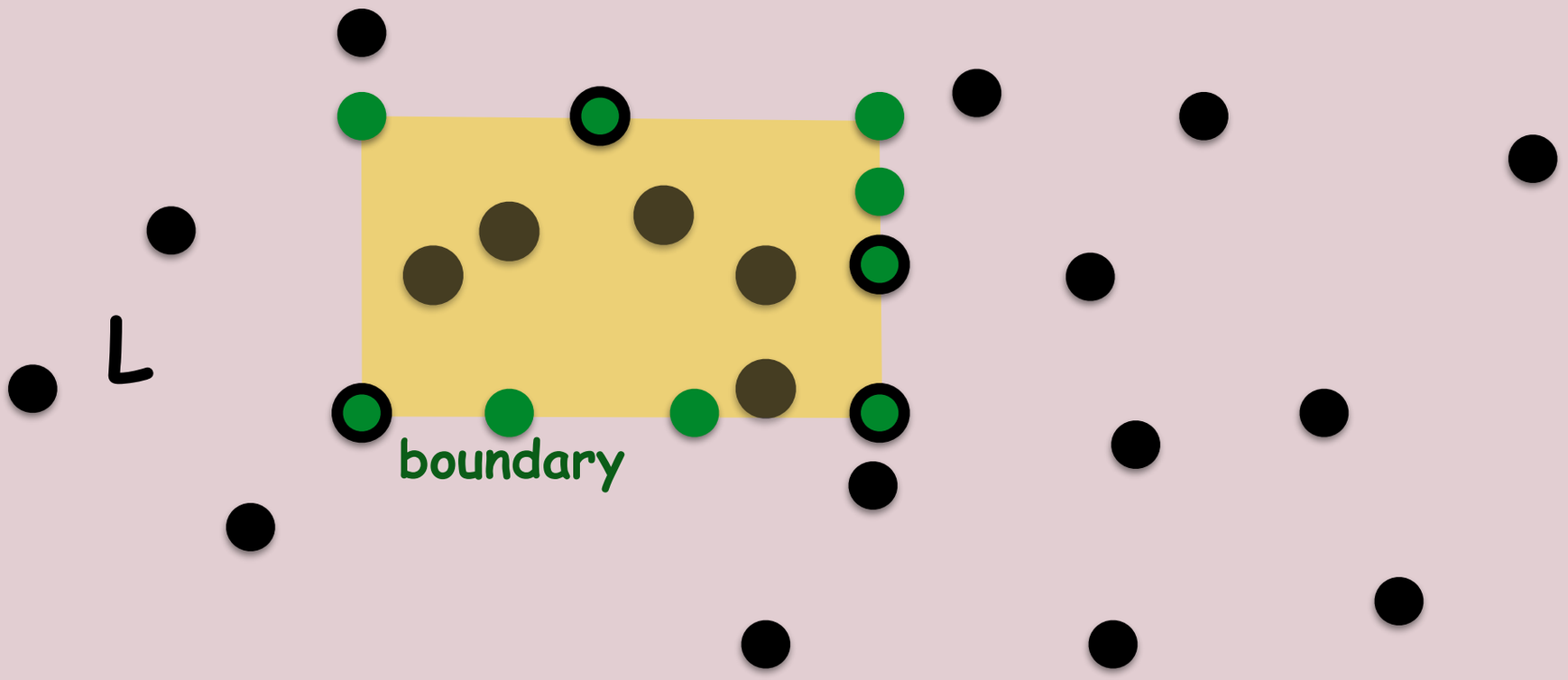

Region 1
Region 3
Region 5
Region 2
Region 4
Region 6

Comparing L (local) to C (global opt)

For each region,

define a mixed solution M

compare L to M



L

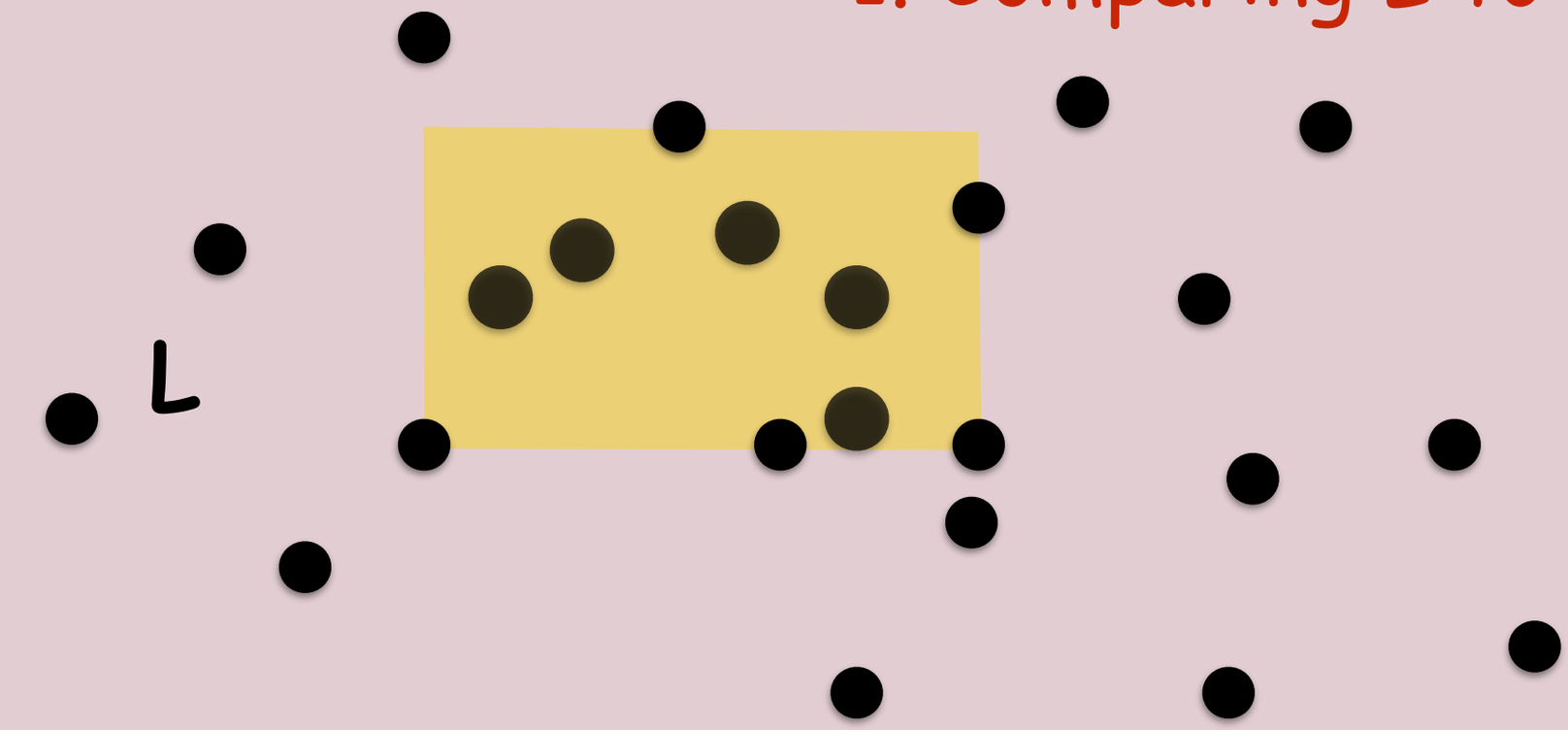
C

L

C

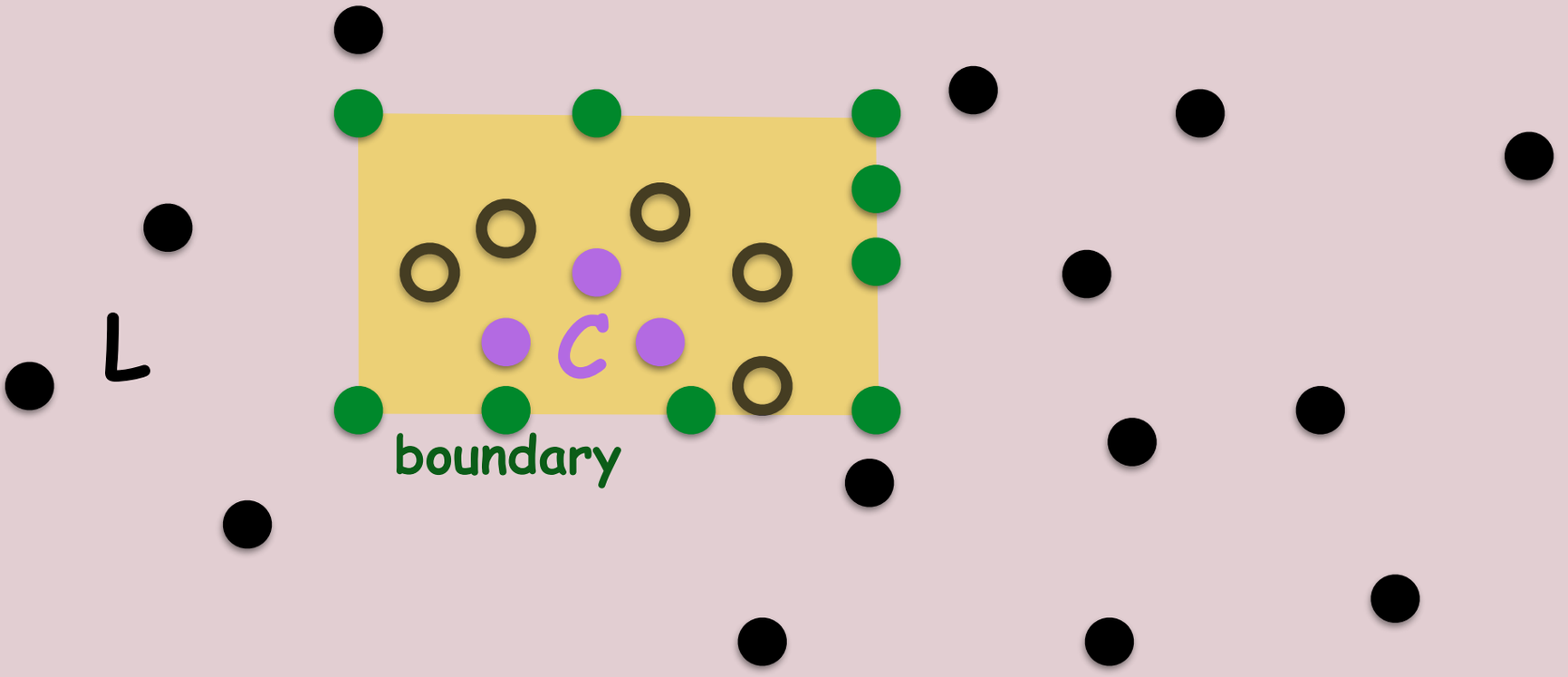
boundary

1. Comparing L to M



- Mixed solution M: facilities of C union B inside, of L outside

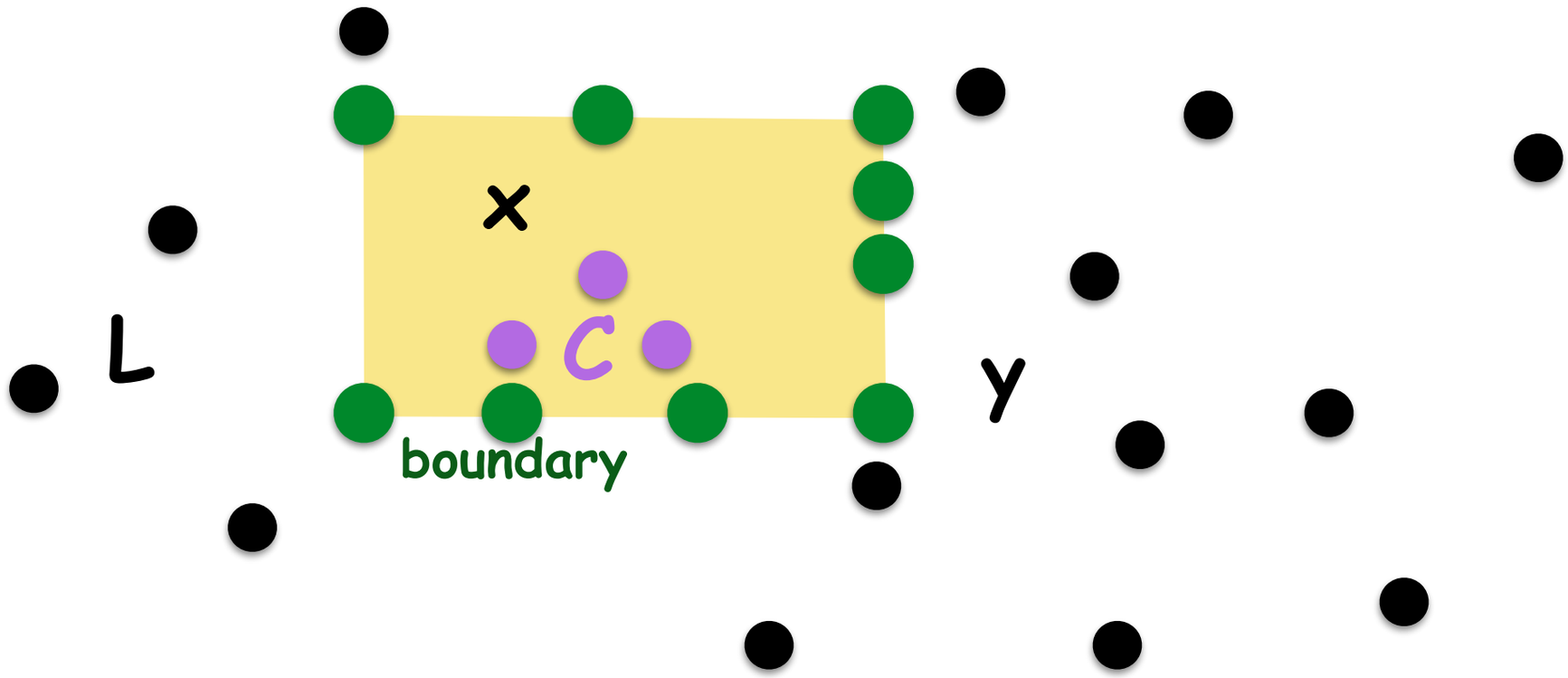
---



M, L only differ inside region: 1/eps^2 facilities

L local optimum, so: cost(L)<cost(M)

2. Client connection cost: internal clients



for internal client x: dist(x,M) is at most dist(x,C)

for outside client y: dist(y,M) is at most dist(y,L)

Comment: ugliest slide of the whole talk

1. **cost(L) is at most cost(M)**
2. **for internal client x: dist(x,M) is at most dist(x,C)**
3. **for outside client y: dist(y,M) is at most dist(y,L)**

$$\begin{aligned}
\text{cost}(L) &\leq \text{cost}(M) \\
\text{cost}(M) &\leq \sum_{x \text{ internal}} \text{dist}(x, C) + \sum_{y \text{ outside}} \text{dist}(y, L) + f|M| \\
\text{cost}(L) &= \sum_{x \text{ internal}} \text{dist}(x, L) + \sum_{y \text{ outside}} \text{dist}(y, L) + f|L| \\
&\Longrightarrow
\end{aligned}$$

$$\sum_{x \text{ internal}} \text{dist}(x, L) + f|L_{\text{internal}}| \leq \sum_{x \text{ internal}} \text{dist}(x, C) + f|M_{\text{internal}}|$$

Sum over regions

$$\text{cost}(L) \leq \text{cost}(C) + f|\text{Boundary facilities}|$$

$$\text{cost}(L) \leq (1 + O(\epsilon))\text{cost}(C)$$

Extended local search yields a PTAS for facility location in weighted planar graphs

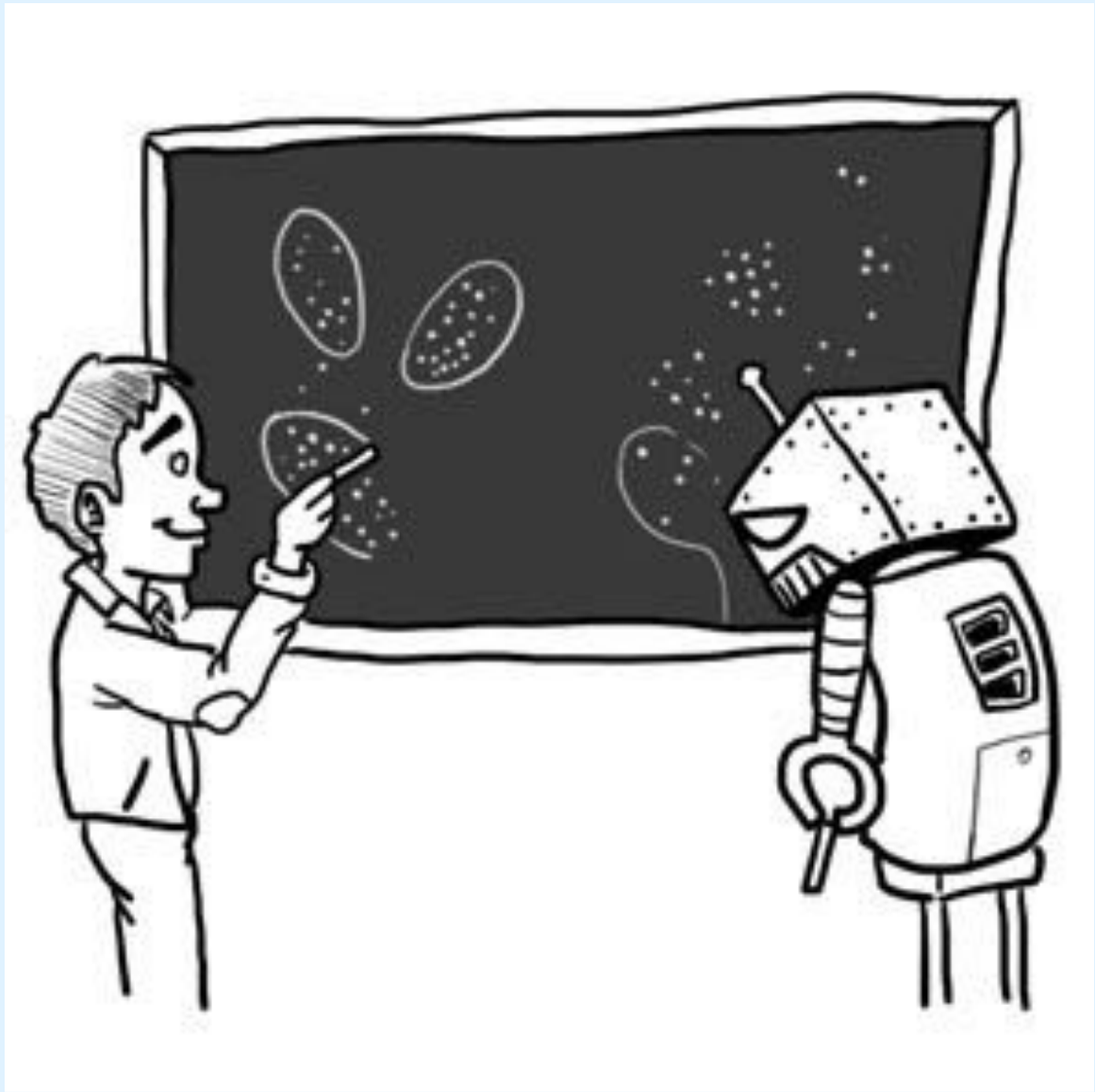
S : current set of centers

Repeat:

Find better solution S' among sets that differ from S in at most $1/\epsilon^2$ centers

Replace S by S'

Until: local optimum



Wait… what about the runtime?

S : current set of centers

Repeat:

Find better solution S' among sets that differ from S in at most $1/\epsilon^2$ centers

Replace S by S'

Until: local optimum

S : current set of centers

Repeat:

Find **much** better solution S' among sets that differ from S in at most $1/\epsilon^2$ centers

Replace S by S'

Until: $\text{cost}(\mathbf{S}) < (\mathbf{1} - \epsilon/\mathbf{n})\text{cost}(\mathbf{S'})$

Extensions

What if, in objective, dist(x,L) is replaced by dist(x,L)^2 Then: ok

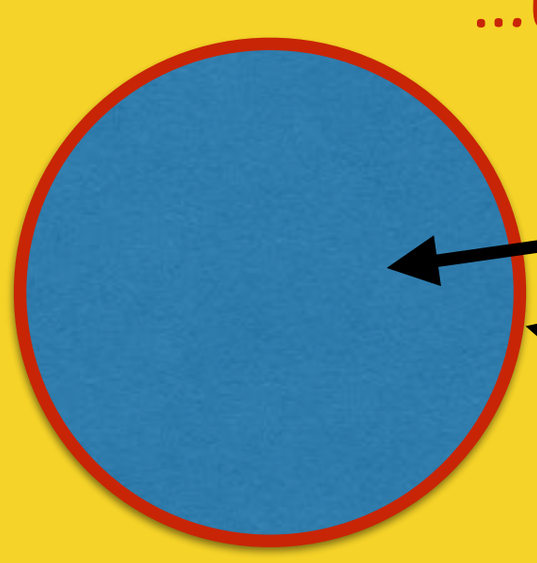
What if dist(x,L) is replaced by monotone function of dist(x,L)? Then: ok

What if, instead of planar, points in minor-closed graph family? Then: ok

What if, instead of planar, points are in Euclidean space?

Lipton-Tarjan: if G is planar then small balanced separator...

...Euclidean analog?



Add a few points so that the Voronoi cells can be partitioned into small, separated regions

Euclidean analog: Separating Voronoi diagrams,

Bhattiprolu Har-Peled 2014, so ok

What if, instead of uniform facility opening cost, the cost varies?

?

What if, instead of uncapacitated facility location, a facility can only serve k clients?

?

What if, instead of paying f to open a facility, we are only allowed k facilities?

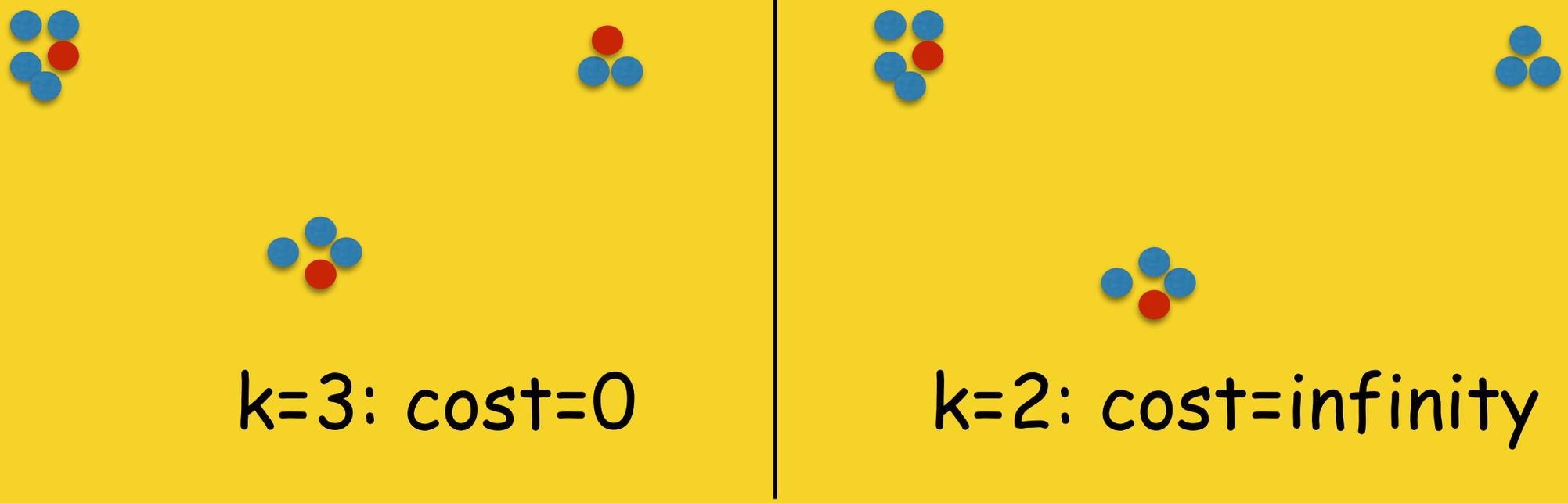
Hard constraint

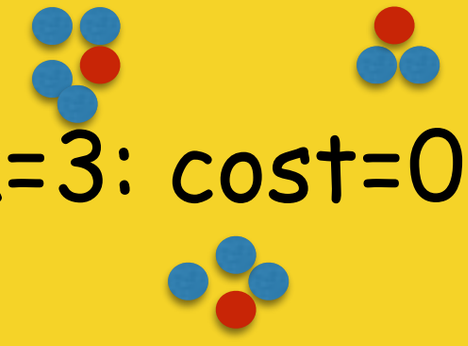
We are not allowed to add boundary facility

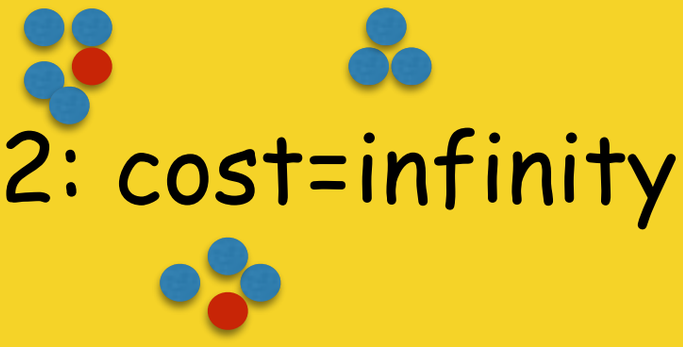
not even just a few, not even one!

But OPT(k) can be much larger than OPT(k+1):

problematic example







However: this example is obvious and local search for k=3 will work

(2) Extended local search yields a PTAS for k-median in edge-weighted planar graphs

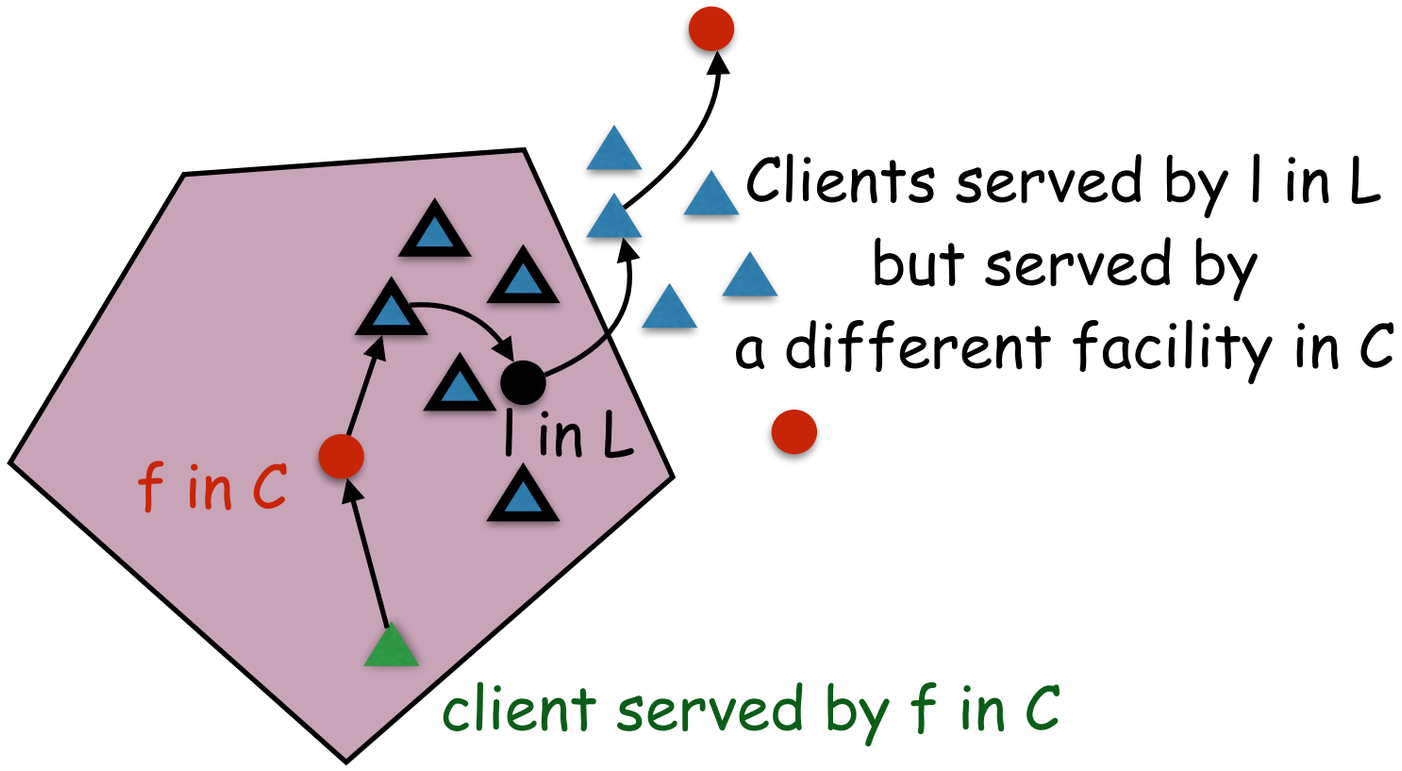
Proof

Definition (deferred): non-isolated facilities

Notation: k′ non-isolated facilities of C.

Robustness theorem: ϵ^3 k′ facilities can be removed from C with small additional connection cost.

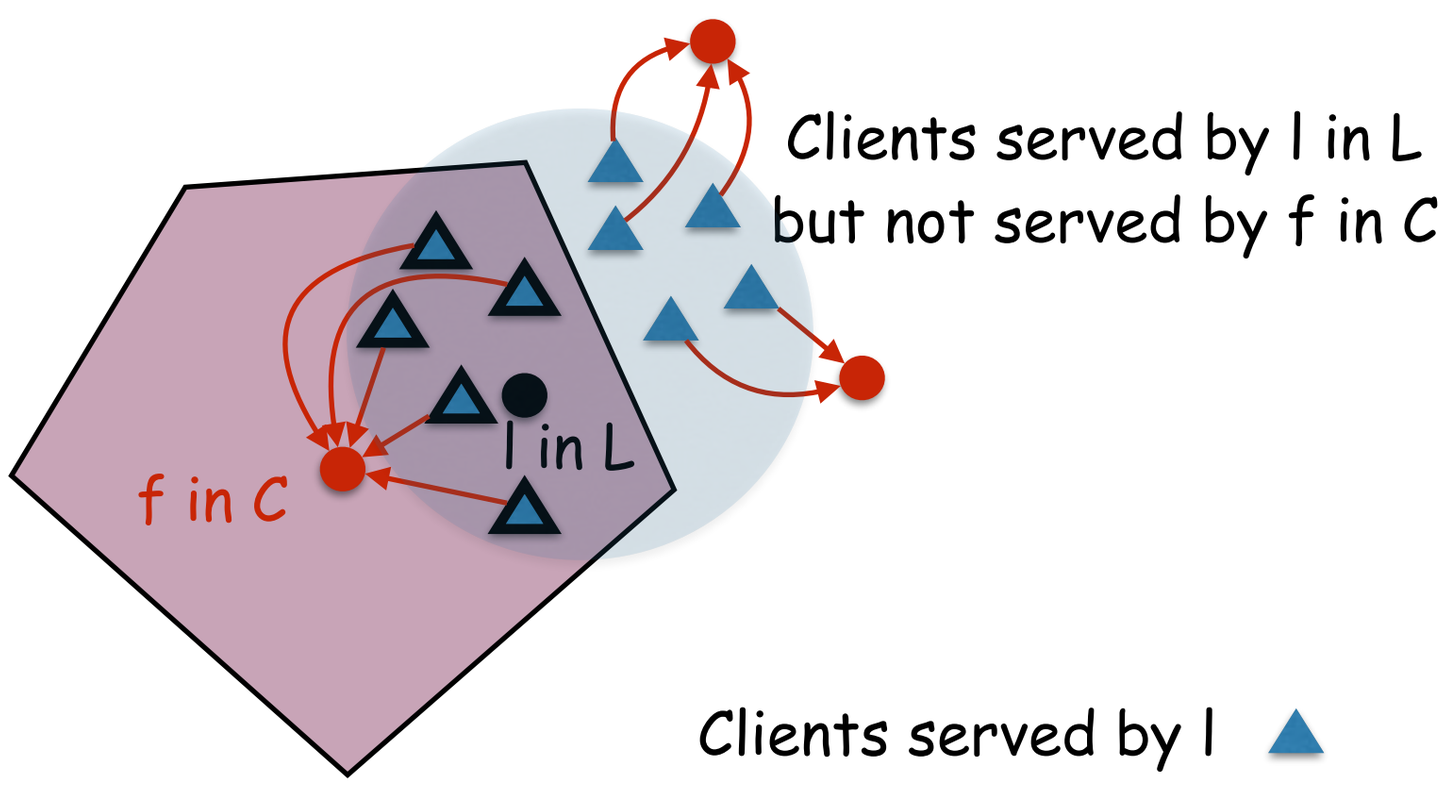
Proof idea: use non-isolation to re-route clients of f to a different facility of C atsmall cost, then remove f from C



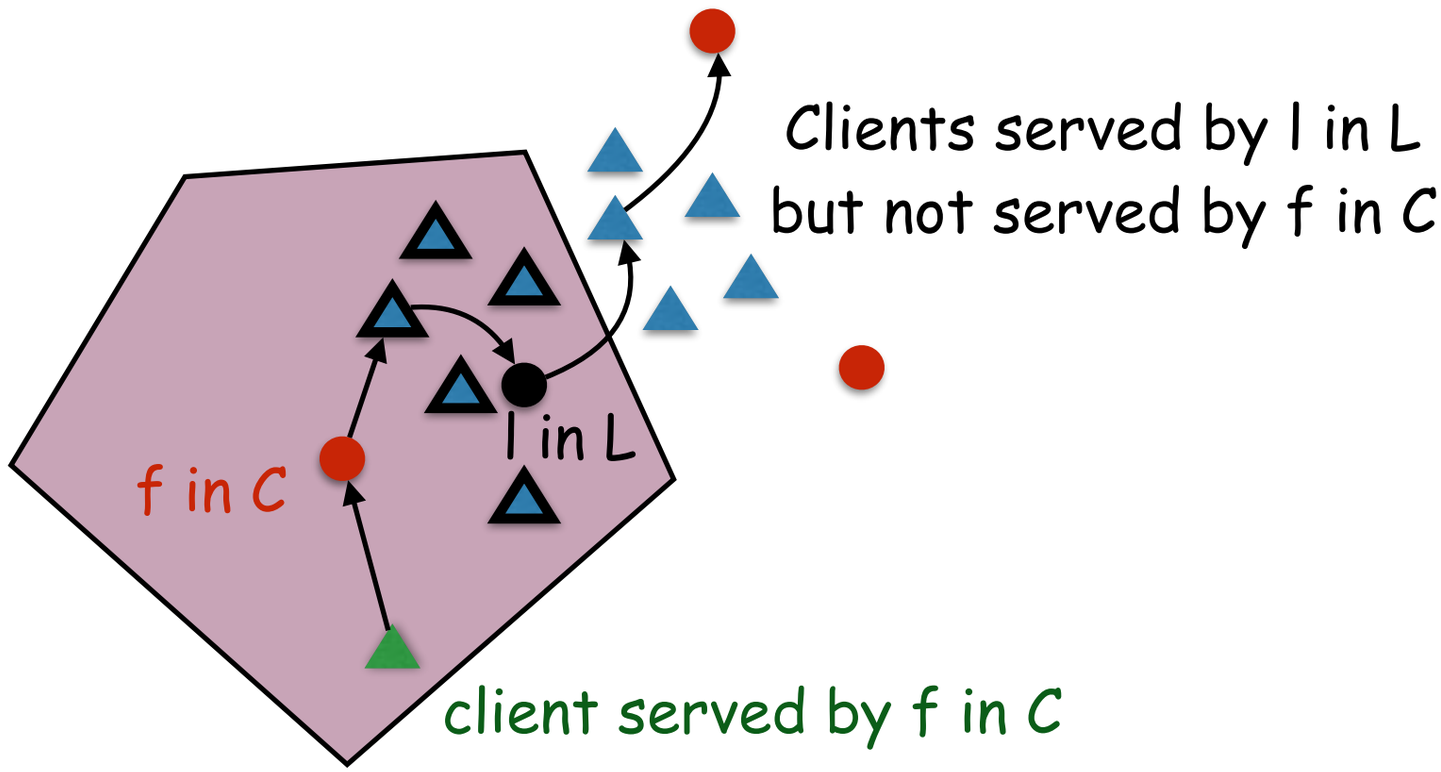
C: optimum L: local

Consider f in C. Mark its clients that, in L, are served by a facility that has a significant (ϵ) fraction of clients that, in C, are served elsewhere.

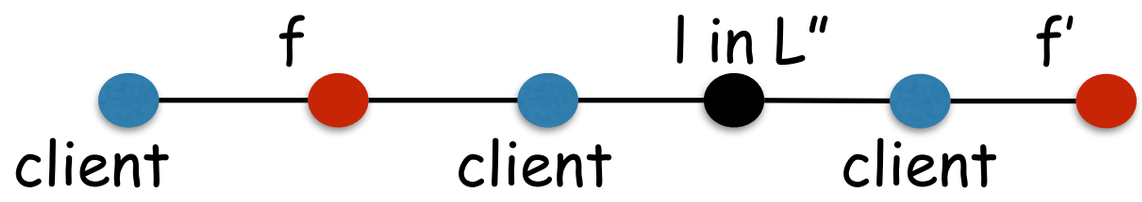
Definition: f is not isolated if a significant fraction of its clients are marked.



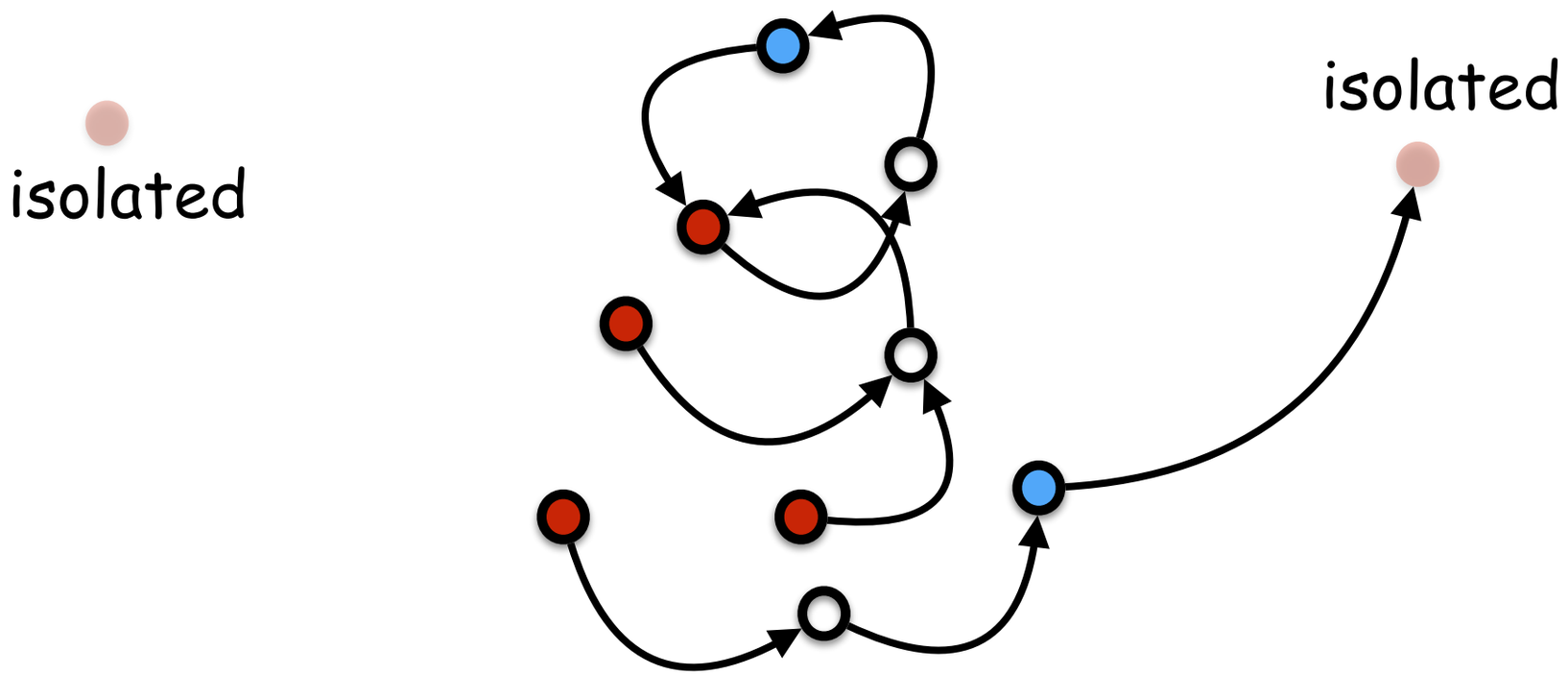
Step 1: use non-isolation to re-route clients of f to a different facility of C



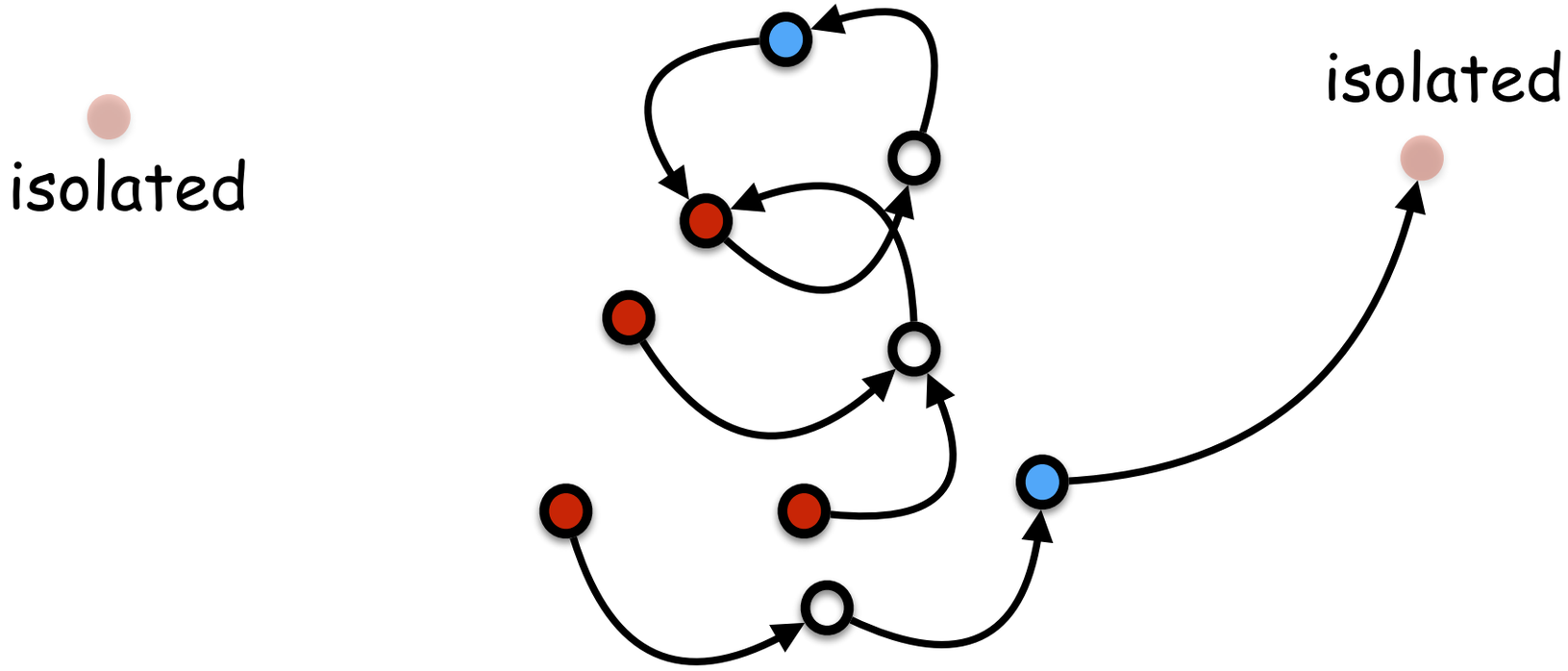
Exists map from non-isolated facilities of C to C, with no fixed point, s.t. reassigning f's clients to its image incurs additional connection cost $O(\mathrm{cost}(C) + \mathrm{cost}(L))/\epsilon^2$



Step 2: any map graph with no fixed point is 3-colorable, and some color class has at least 33% of facilities



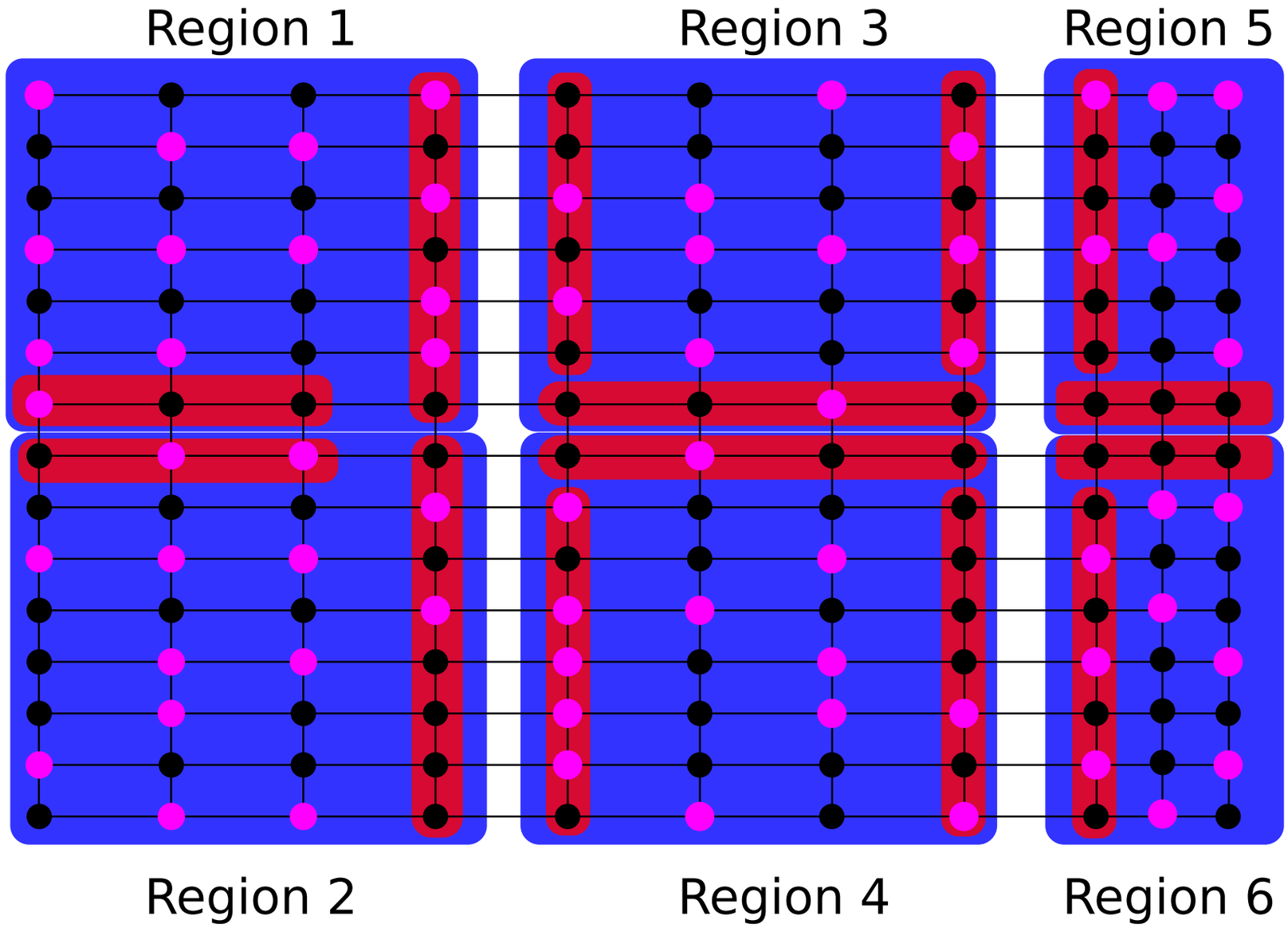
33% of facilities are red, and their images are not red.



Step 3: Partition that color class into many (1/eps^3) parts. For one of them, S, re-assigning its clients to f(x) incurs small additional cost.

After re-assignment, facilities of S now serve no one and can be removed. QED

Extension of r-division



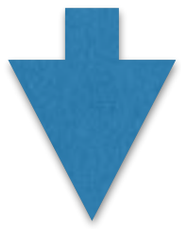
Replacing L by C in region 1 would change #facilities by 3
Replacing L by C in region 2 would change #facilities by -1
Replacing L by C in region 3 would change #facilities by -2
Replacing L by C in region 4 would change #facilities by 4
Replacing L by C in region 5 would change #facilities by -4
Replacing L by C in region 6 would change #facilities by 0

Super-region 1
Super-region 2
Super-region 3

From robustness theorem to PTAS

Robustness theorem: o(k') non-isolated facilities can be removed from C w/ small additional connection cost.

Apply extension of r-division approach to that solution and L



Local search yields a PTAS for k-median in edge-weighted planar graphs

(1) Local search yields a PTAS for uniform facility location in edge-weighted planar graphs

(2) Local search yields a PTAS for k-median in edge-weighted planar graphs

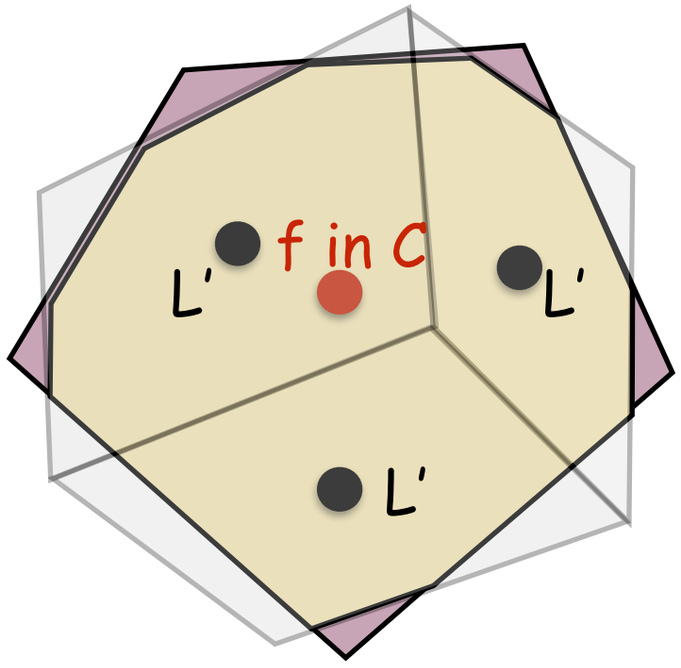
Take-home message: local search works well for clustering type problems in structured instances s.a. planar graphs.

THE END

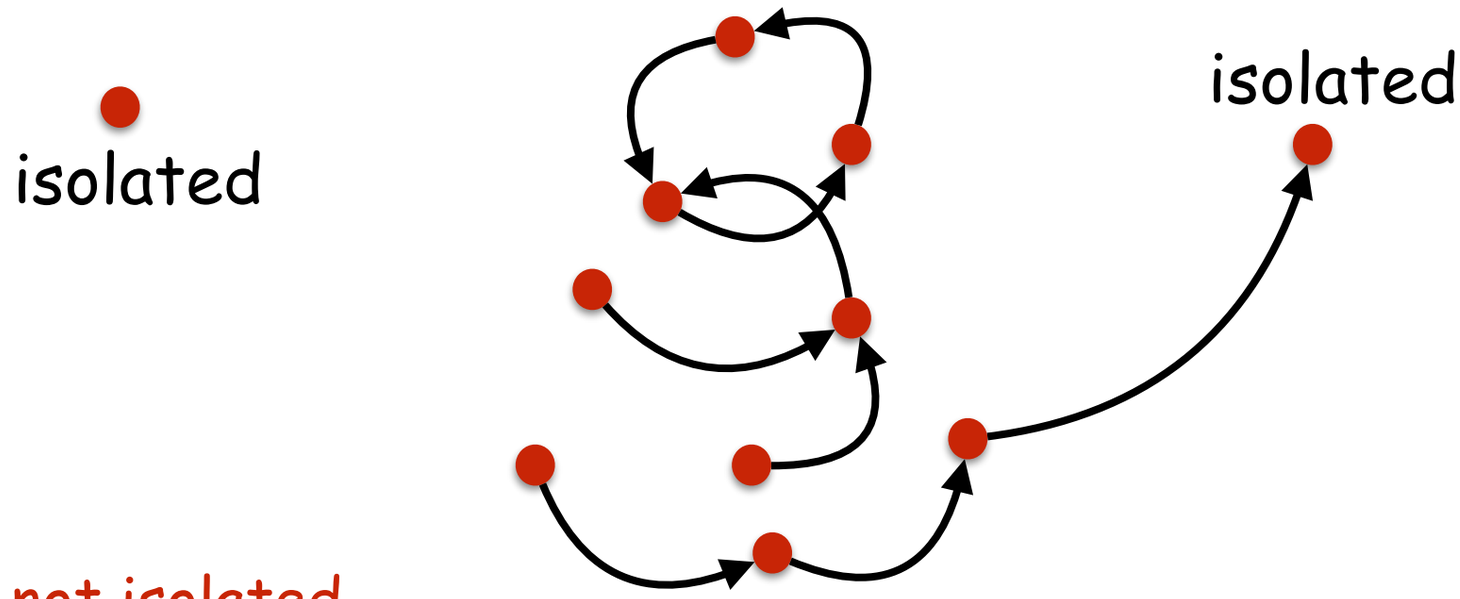
C: optimum L: local

Definition: Consider f in C and L' subset of L.
(f,L') is isolated if

- for each l in L', most clients served by l in L are served by f in C &
- most clients served by f in C are served by L'



Proving the key lemma (1/2)



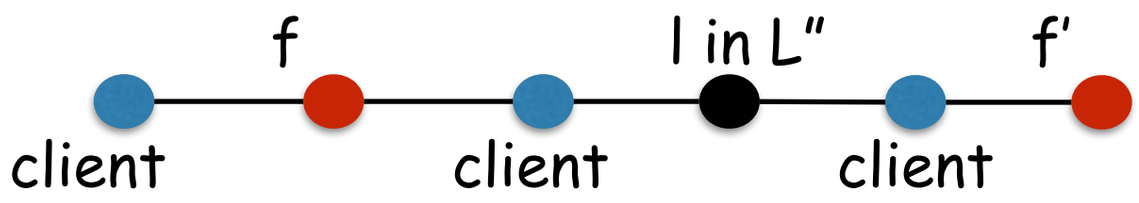
Consider f in C that is not isolated.

Let L" be the subset of L consisting of l s.t.

NOT: "most clients served by l in L are served by f in C"

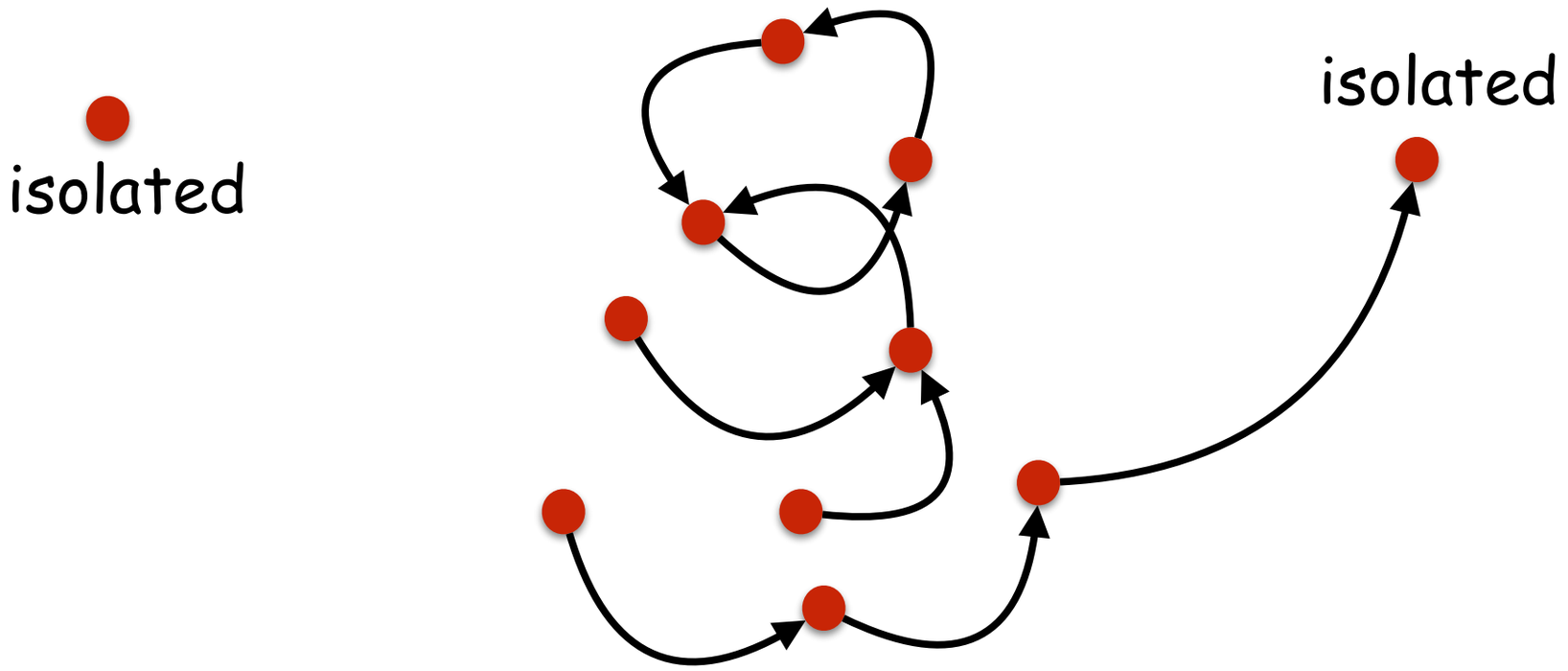
i.e.: at least eps % of clients served by l are served by f' other than f in C

Then at least eps % of clients served by f in C are served by L"

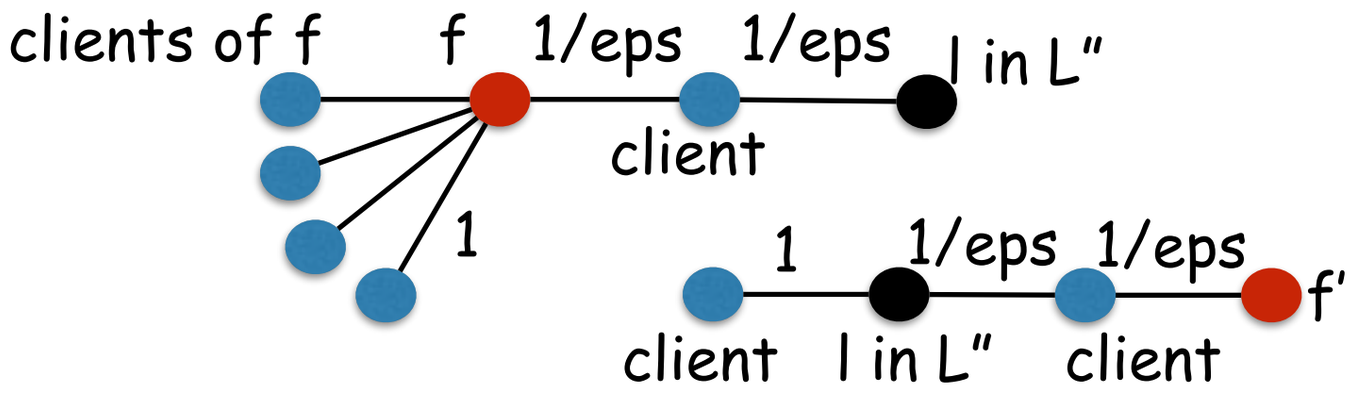


Use that to re-route f's clients

Step 1: Exists map from non-isolated facilities of C to C s.t. reassigning every client of f to its image incurs additional connection cost $O(\mathrm{cost}(C) + \mathrm{cost}(L))/\epsilon^2$



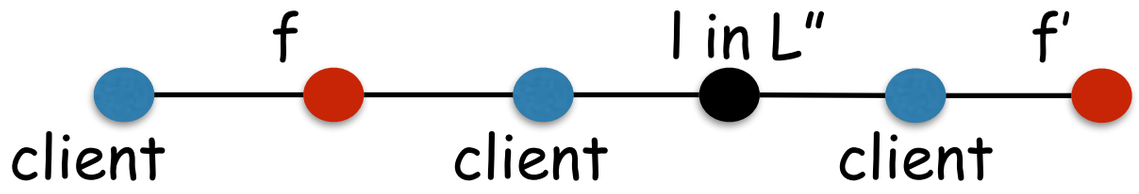
To re-route f's clients: fractional assignment.



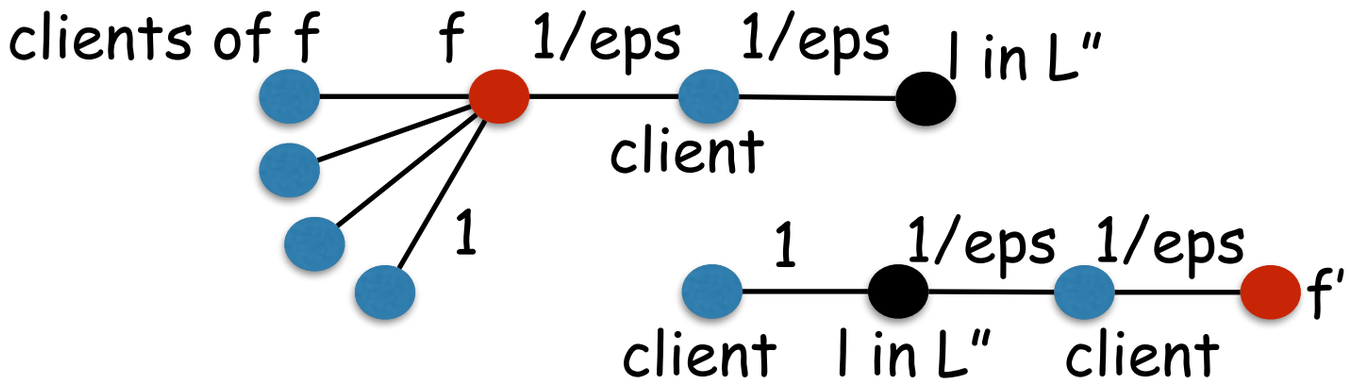
Additional connection cost: edges of C and L carry at most 1/eps^2 connections, QED.

Proving the key lemma (2/2)

Consider f in C that is not isolated.



To re-route f's clients: fractional assignment.



Additional connection cost:
edges of C and L carry at most
1/eps^2 connections, QED.