



# Spatial Data Mining: A Deep Dive into Cluster Analysis

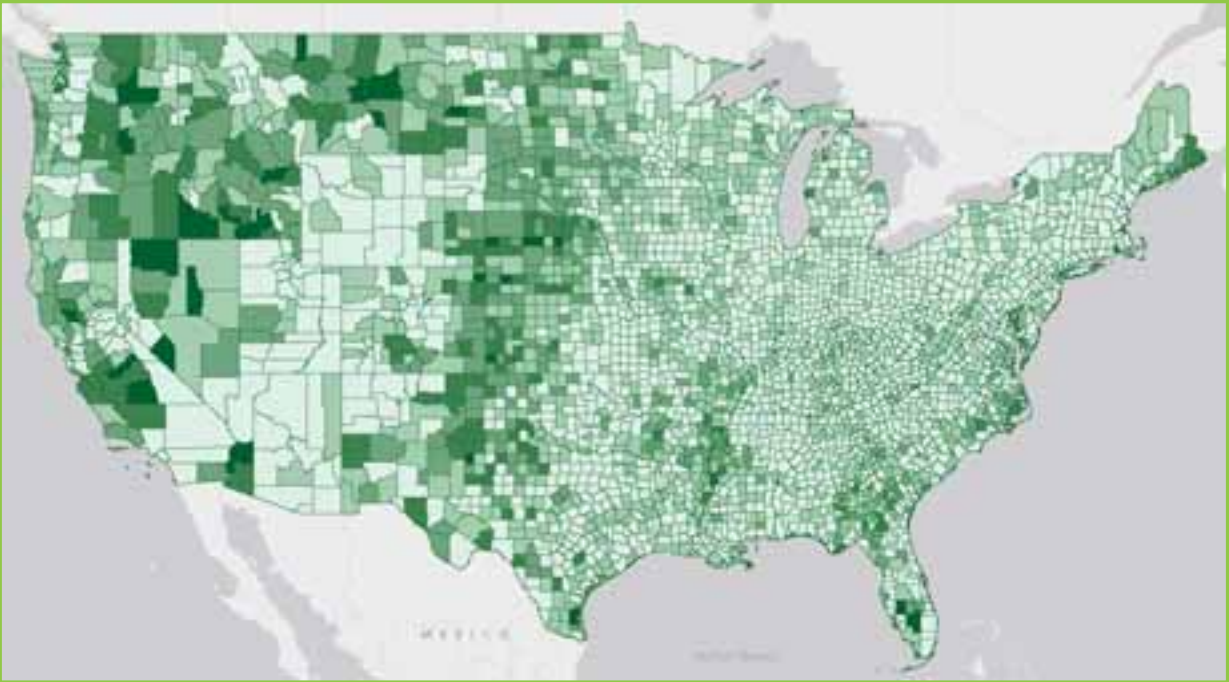
Lauren Bennett

Flora Vale

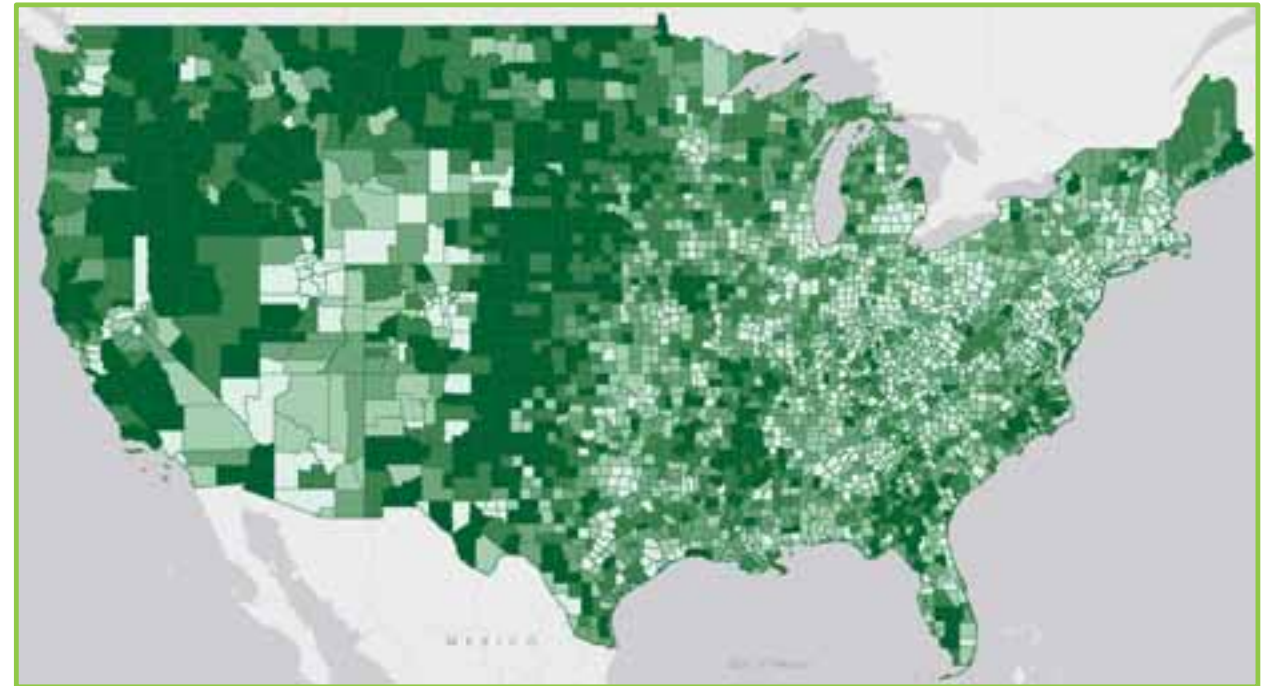
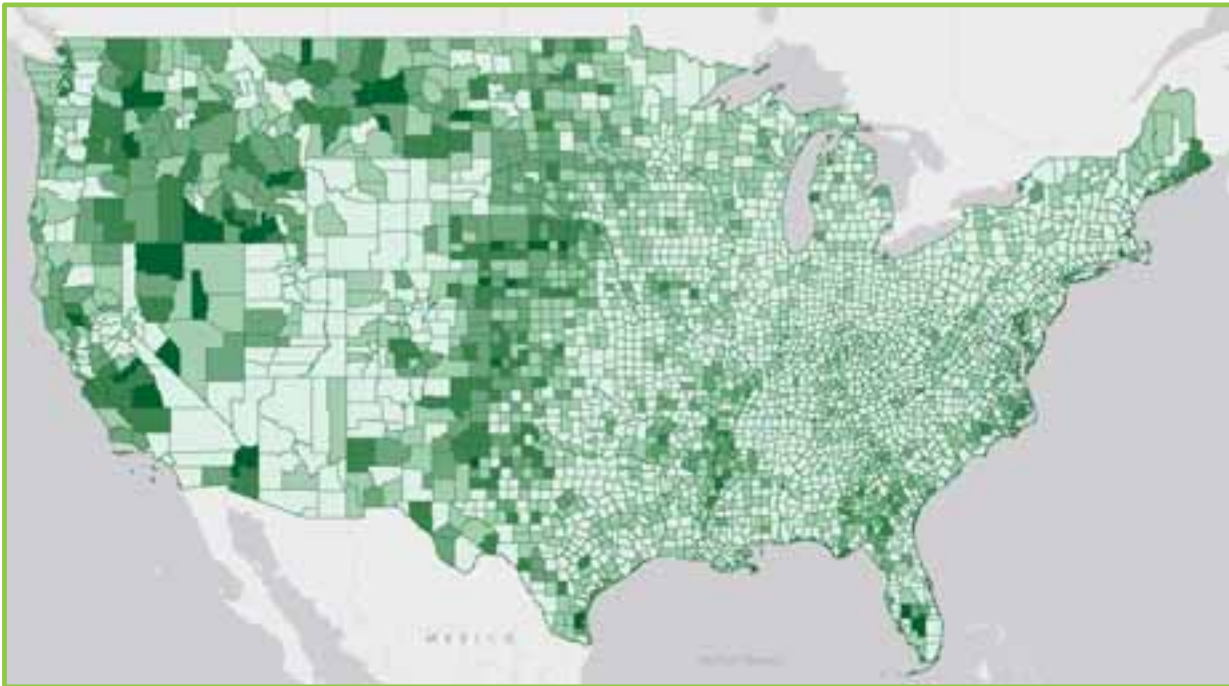
Technical Workshop

# Subjectivity of Maps

# Percent of Employment in Agriculture

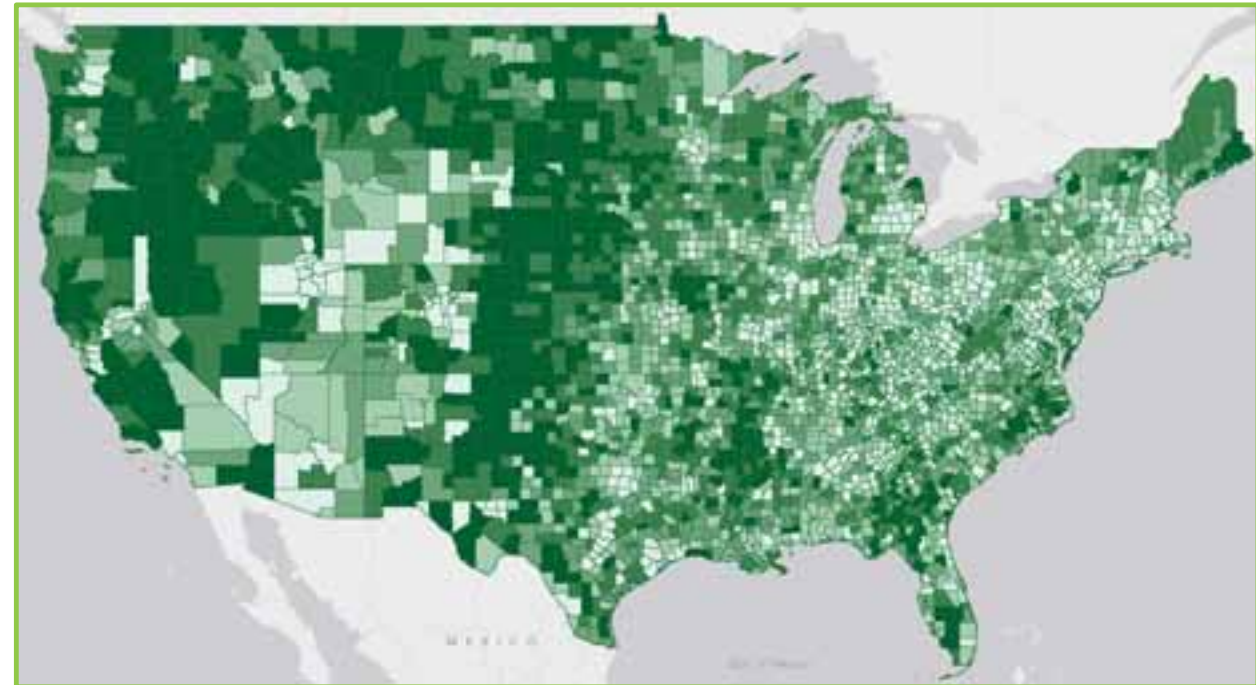
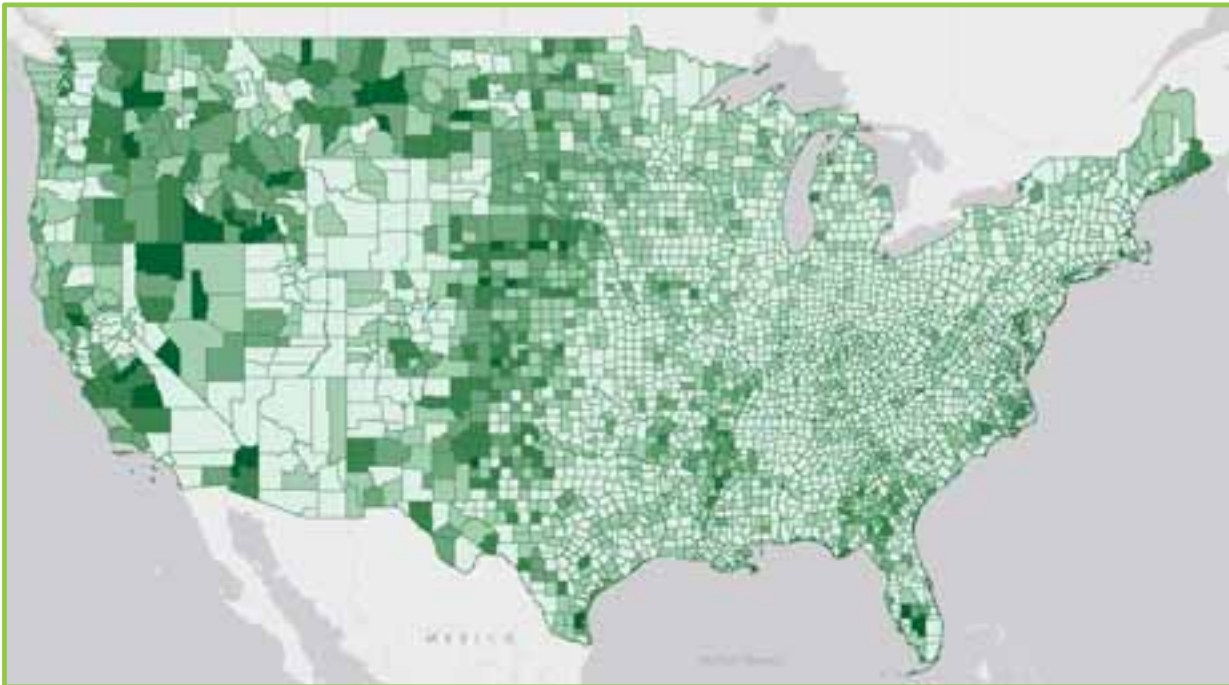


# Percent of Employment in Agriculture





## Percent of Employment in Agriculture

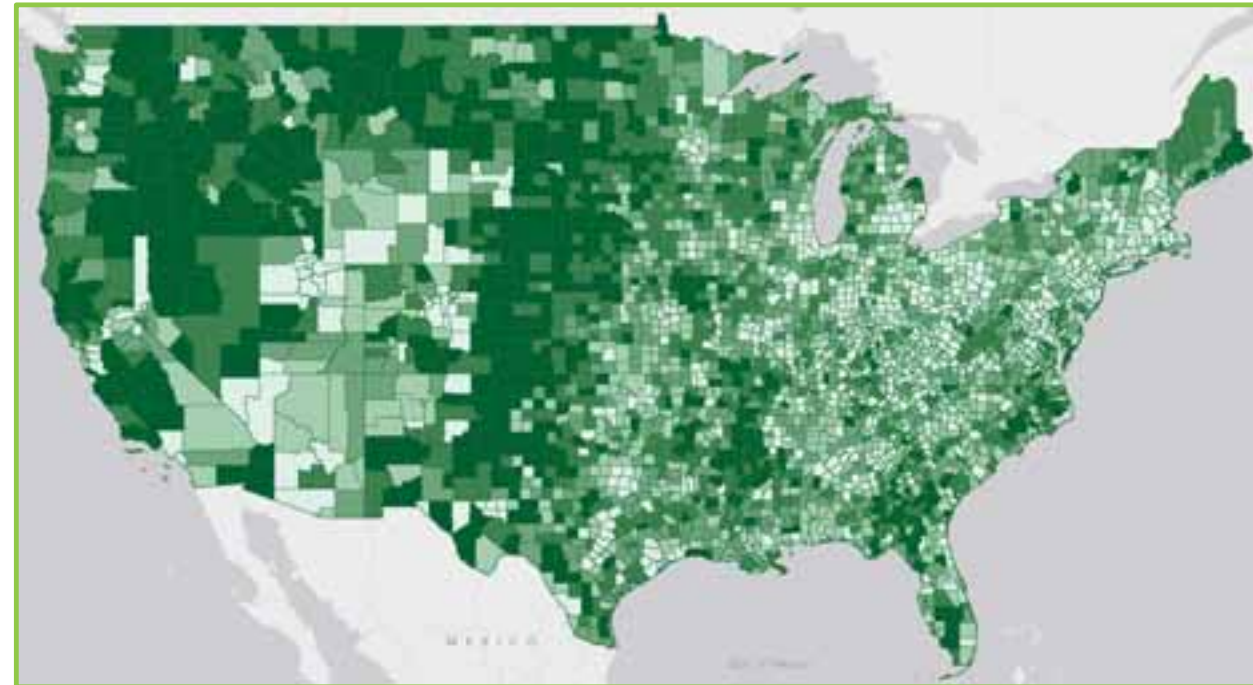
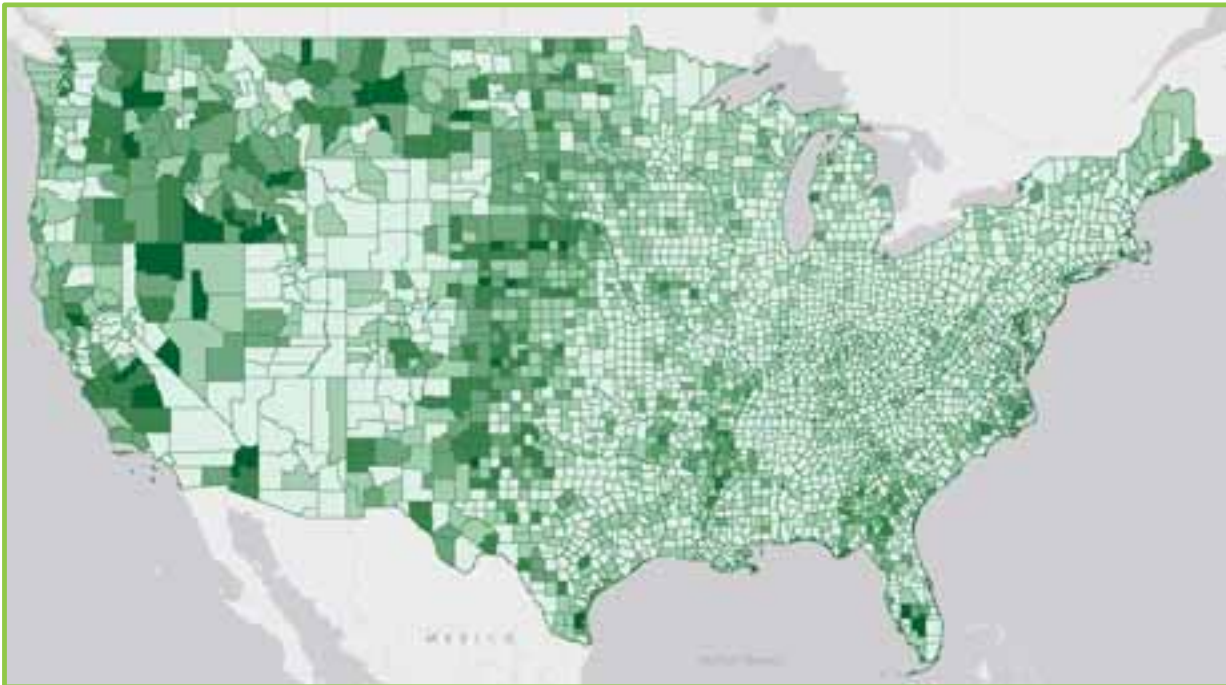


Where are the hot spots? Where is the variation greater?

# Percent of Employment in Agriculture

## Natural Breaks

## Quantile



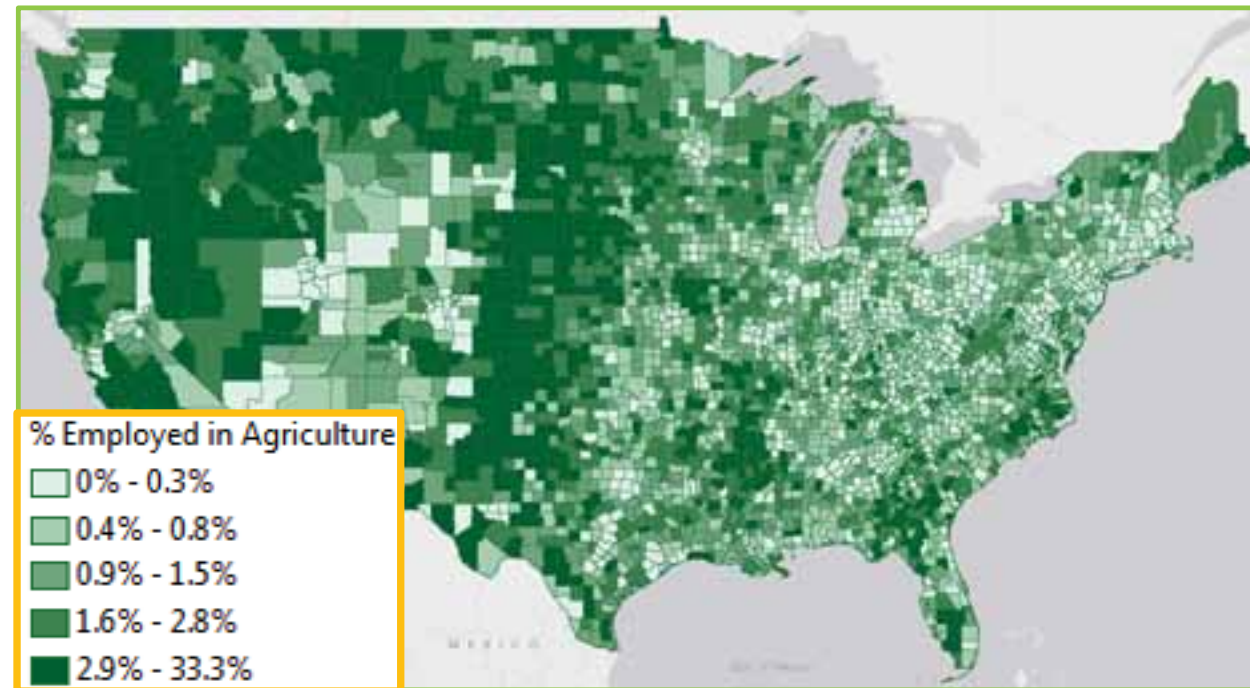
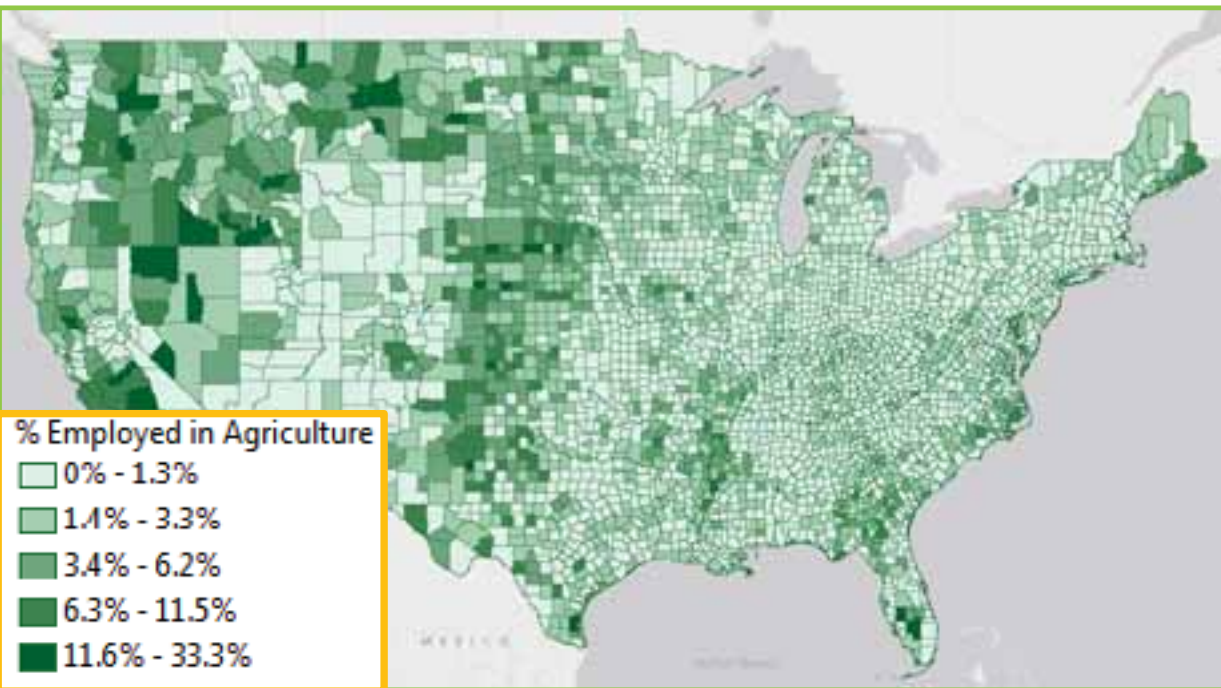
Where are the hot spots? Where is the variation greater?



# Percent of Employment in Agriculture

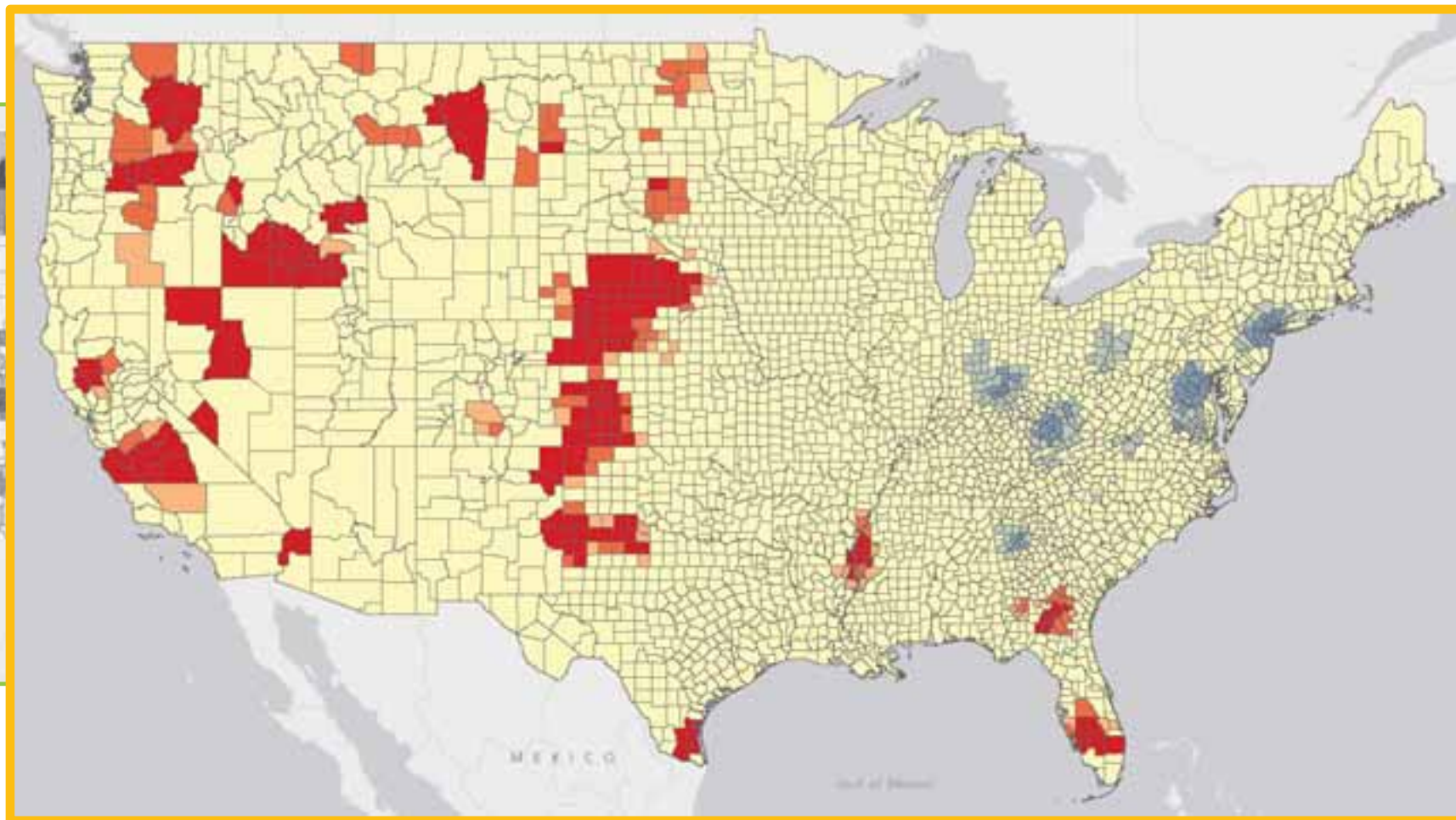
## Natural Breaks

## Quantile



Where are the hot spots? Where is the variation greater?

# Minimizing the Subjectivity





# Inferential Statistics



# Complete Spatial RANDOMNESS

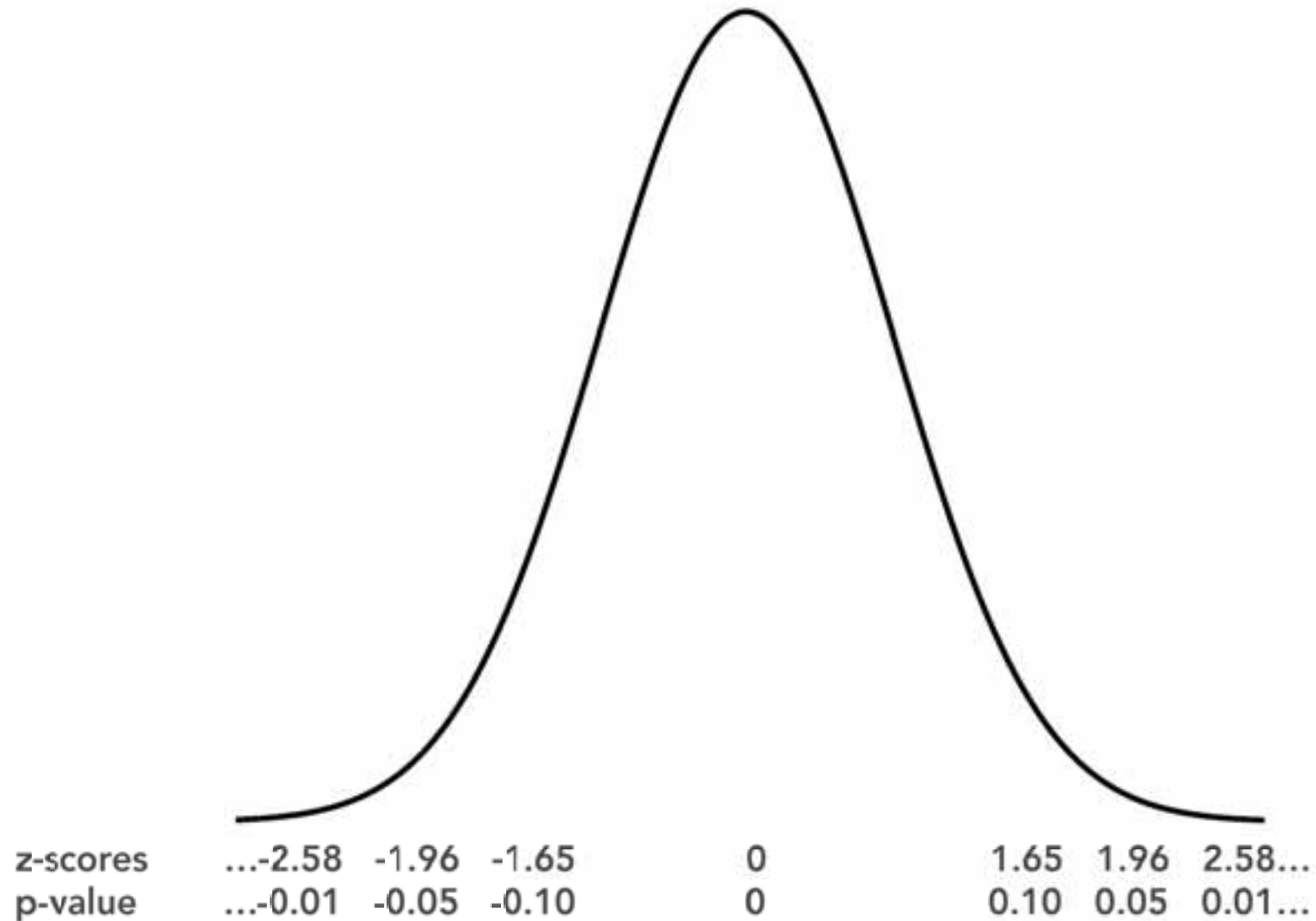
Is there a **PATTERN**?



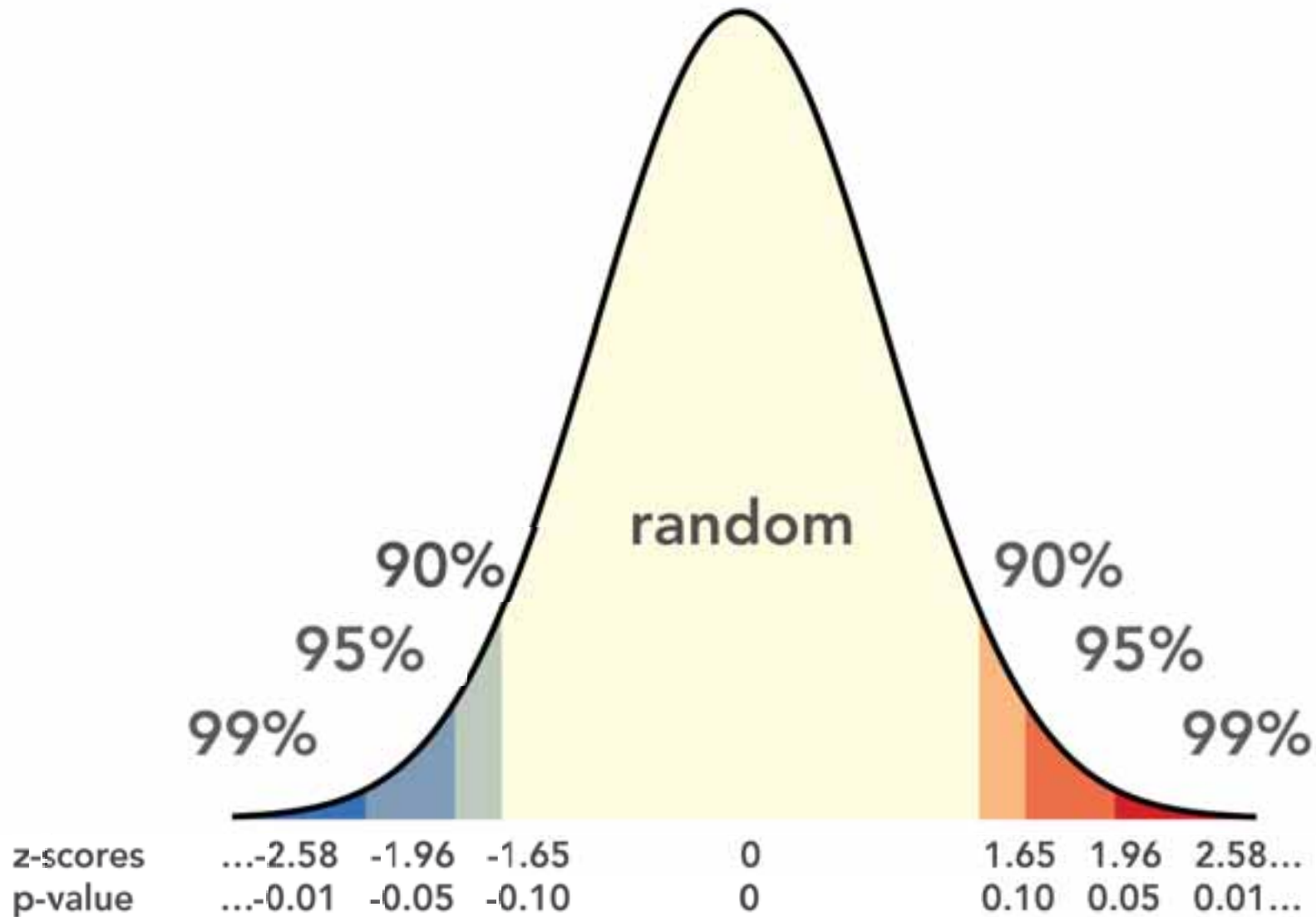
**z-scores**

**p-values**

# z-scores and p-values




# z-scores and p-values





[-]  Mapping Clusters

 Cluster and Outlier Analysis (Anselin Local Morans I)

 Grouping Analysis

 Hot Spot Analysis (Getis-Ord  $G_i^*$ )

 Optimized Hot Spot Analysis

 Similarity Search


[-]  Mapping Clusters

 Cluster and Outlier Analysis (Anselin Local Morans I)

 Grouping Analysis

 Hot Spot Analysis (Getis-Ord  $G_i^*$ )

 Optimized Hot Spot Analysis

 **Similarity Search**



**python  
script!**





```
OptimizedHotSpotAnalysis - Notepad
File Edit Format View Help

#### Remove Locations Outside Boundary FC ####
featureLayer = "ClippedPointFC"
DM.MakeFeatureLayer(tempFC, featureLayer)
if self.boundaryFC:
    msg = ARCPY.GetIDMessage(84454)
    ARCPY.SetProgressor("default", msg)
    DM.SelectLayerByLocation(featureLayer, "INTERSECT",
                             self.boundaryFC, "#",
                             "NEW_SELECTION")
    DM.SelectLayerByLocation(featureLayer, "INTERSECT",
                             "#", "#", "SWITCH_SELECTION")
    DM.DeleteFeatures(featureLayer)
else:
    if additionalZeroDistScale == "ALL":
        msg = ARCPY.GetIDMessage(84455)
        ARCPY.SetProgressor("default", msg)
        DM.SelectLayerByAttribute(featureLayer, "NEW_SELECTION",
                                  "'Join_Count' = 0'")
        DM.DeleteFeatures(featureLayer)
    else:
        distance = additionalZeroDistScale * fish.quadLength
        distanceStr = self.ssd.distanceInfo.linearUnitString(distance,
                                                               convert = True)
        nativeStr = self.ssd.distanceInfo.printDistance(distance)
        msg = "Removing cells further than %s from input pointsd...."
        ARCPY.AddMessage(msg % nativeStr)
        DM.SelectLayerByLocation(featureLayer, "INTERSECT",
                                 self.ssd.inputFC, distanceStr,
                                 "NEW_SELECTION")
        DM.SelectLayerByLocation(featureLayer, "INTERSECT",
                                 "#", "#", "SWITCH_SELECTION")
        DM.DeleteFeatures(featureLayer)

DM.Delete(featureLayer)
del collSSDO

ARCPY.env.extent = oldExtent
```



# Hot Spot Analysis

(Getis-Ord  $G_i^*$ )

**Statistically  
Significant Clusters  
of High and Low  
Values.**

**Statistically  
Significant Clusters  
of High and Low  
Values.**

Statistically  
Significant **Clusters**  
of High and Low  
Values.



Statistically  
Significant Clusters  
of High and Low  
Values.

$$G_i^* = \sum_{j=1}^n w_{i,j} x_j - \bar{X} \sum_{j=1}^n w_{i,j}$$

$$S = \sqrt{\frac{\left[ \sum_{j=1}^n w_{i,j}^2 - \left( \sum_{j=1}^n w_{i,j} \right)^2 \right]}{n-1}}$$

**RANDOMNESS**

**PROBABILITY**

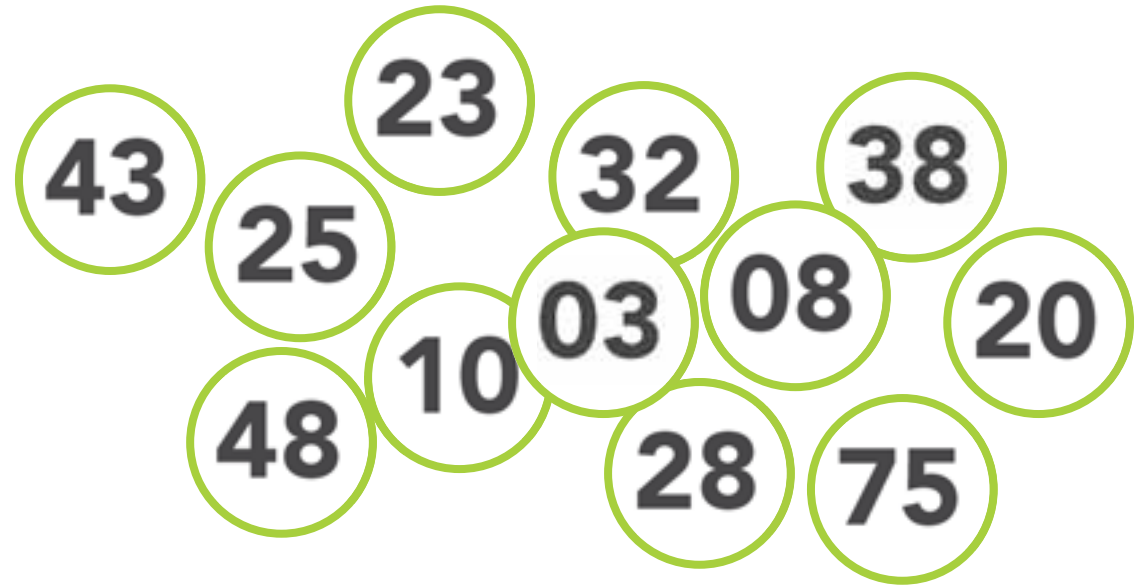
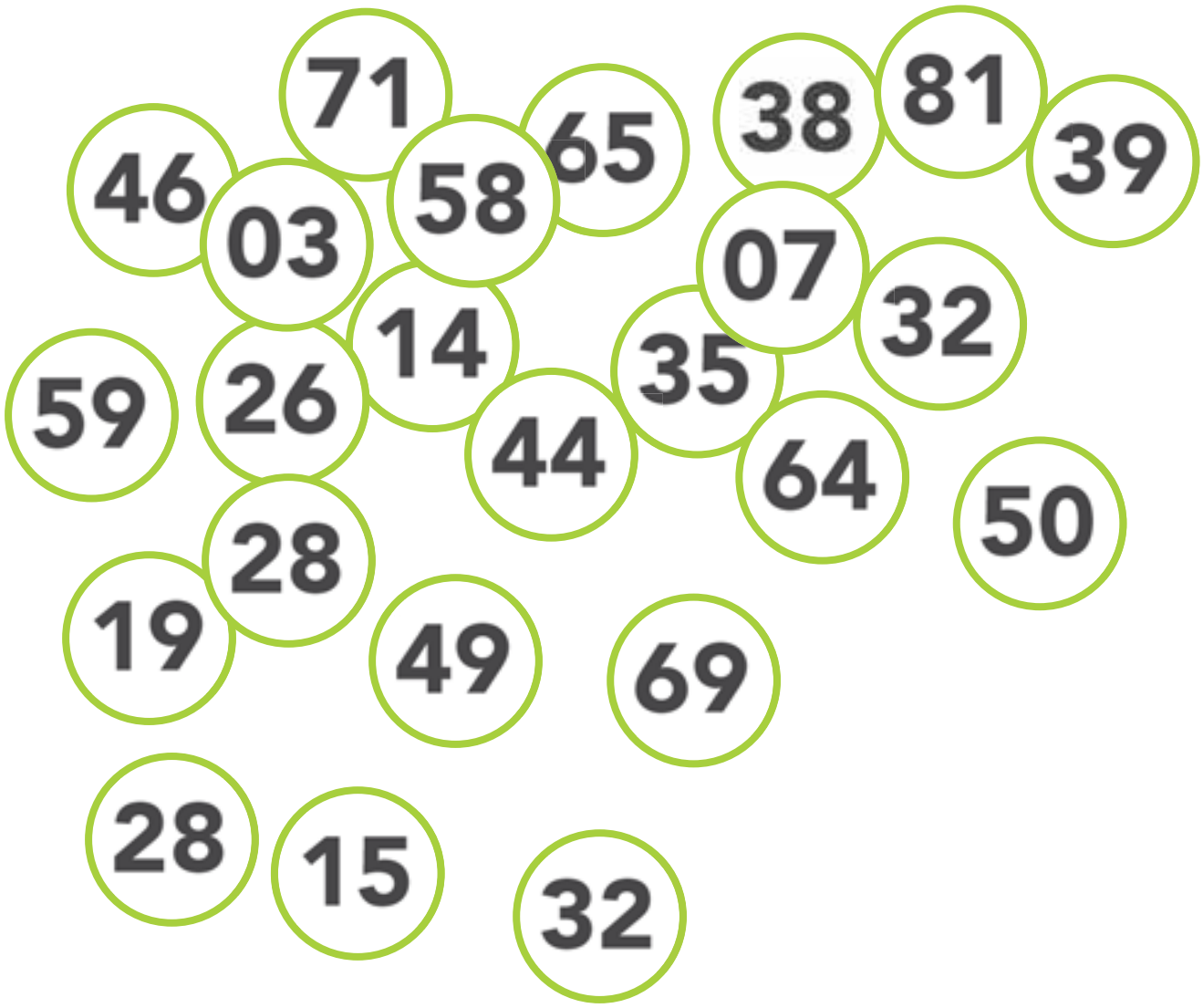
Basically, we're asking:

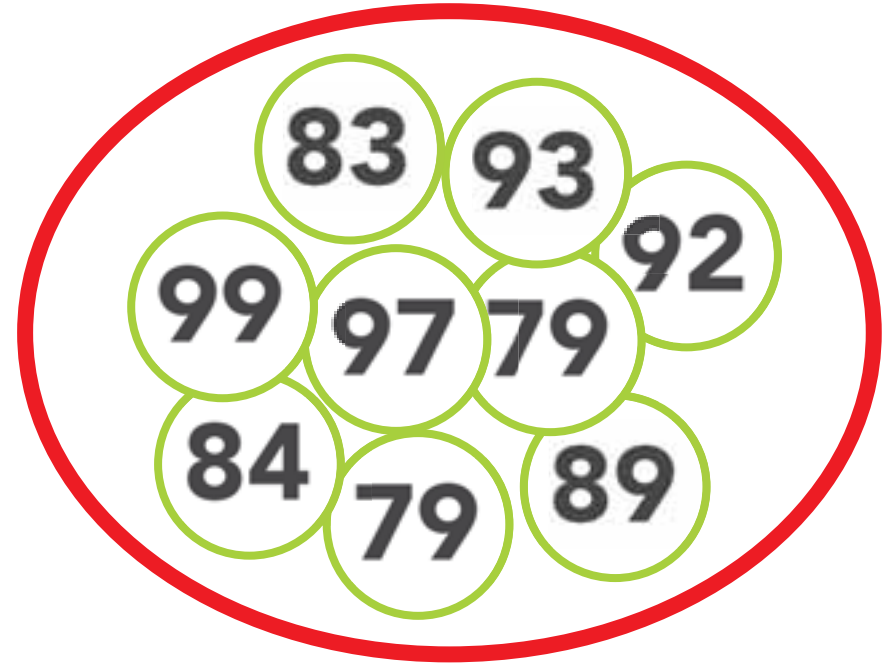
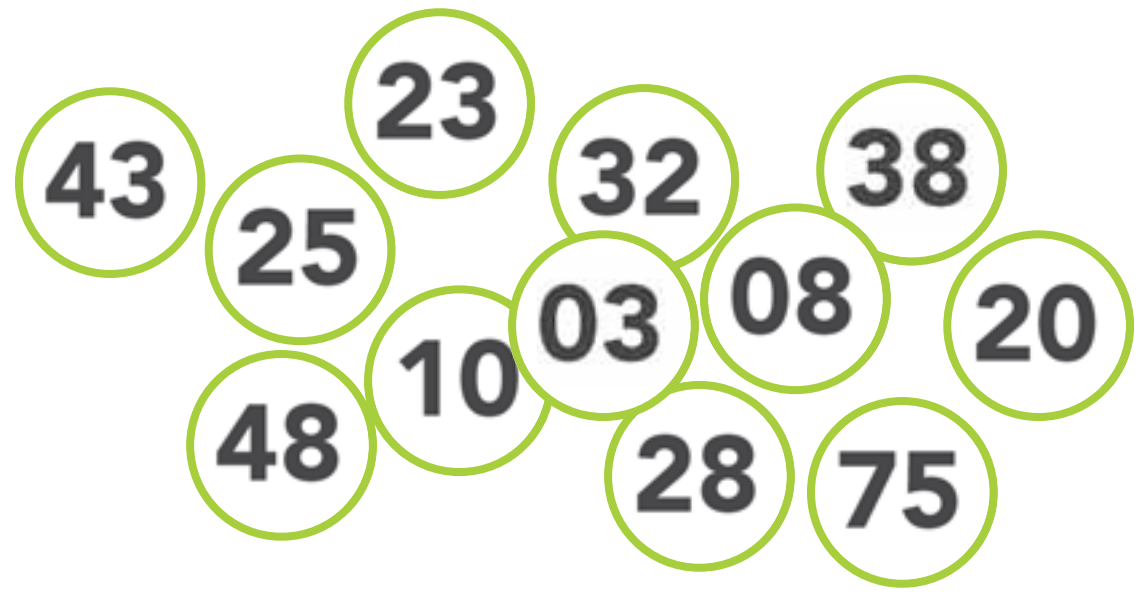
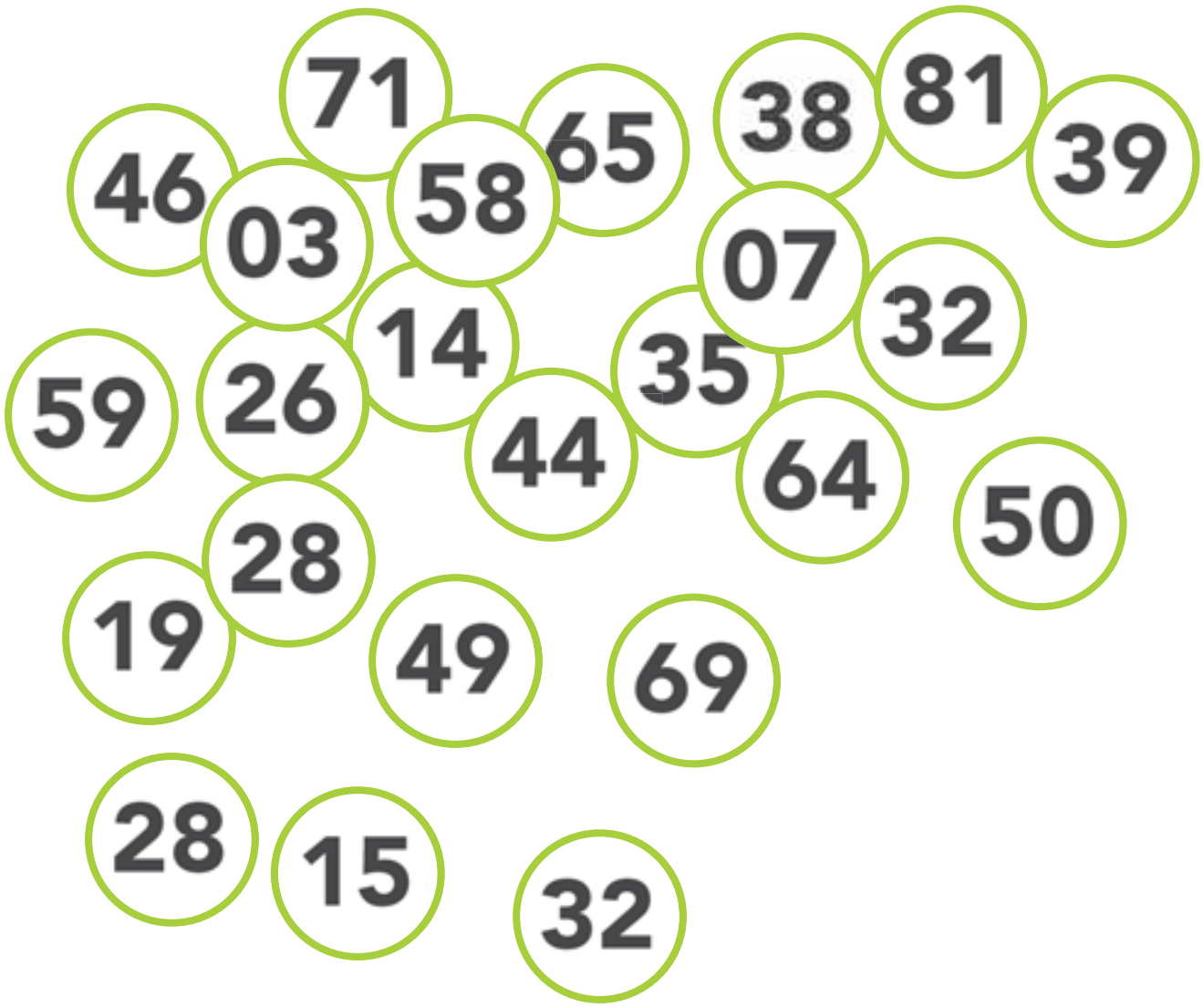
“what is the **probability** that  
a spatial distribution of  
values is **RANDOM?**”

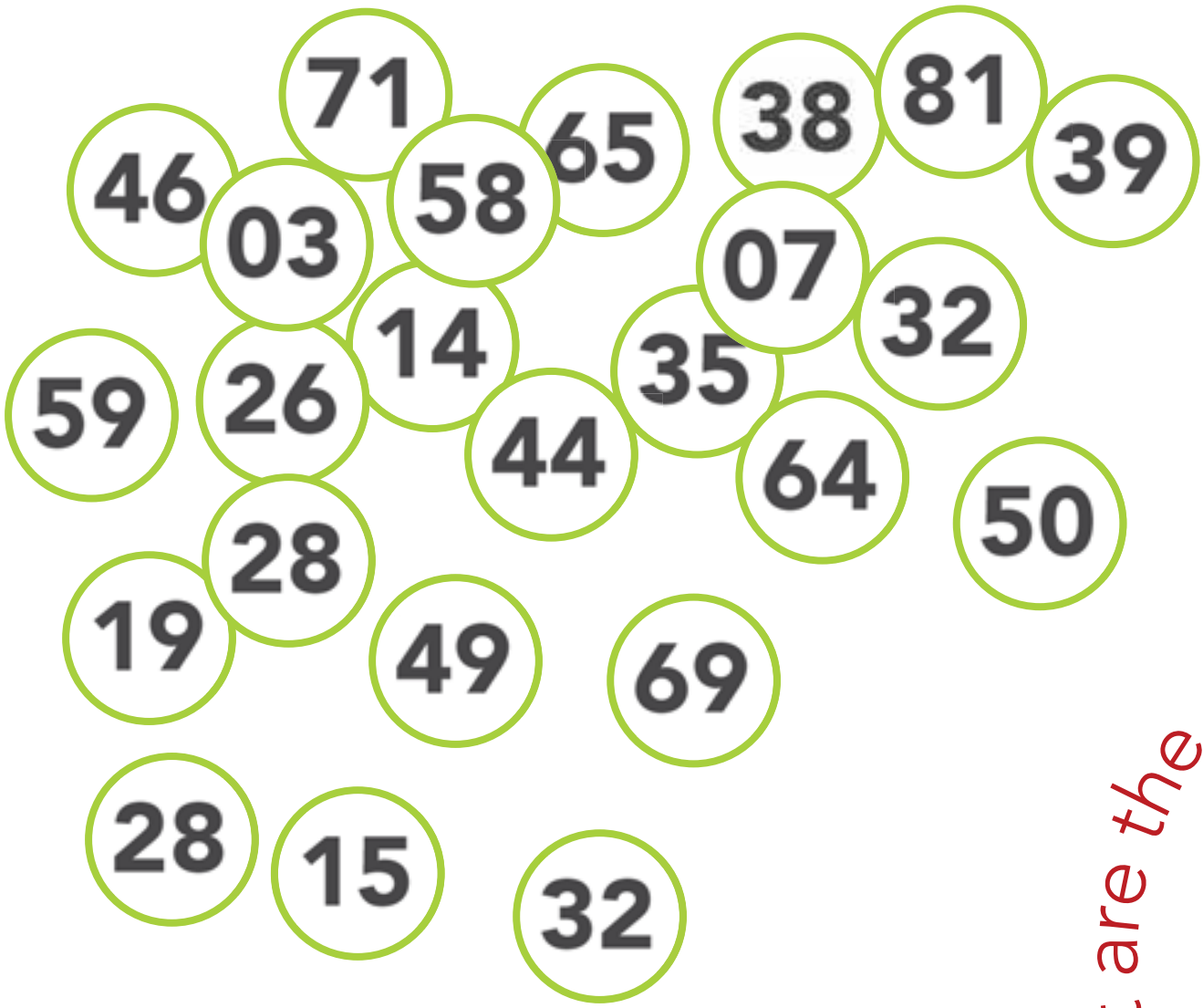
**...distribution of values???**



14	15	92	65	35	89	79	32	38
46	26	43	38	32	79	50	28	84
19	71	69	39	93	75	10	58	20
97	49	44	59	23	07	81	64	03
28	32	08	99	83	28	03	48	25







What are the chances this happened RANDOMLY???



# Polygons







feature



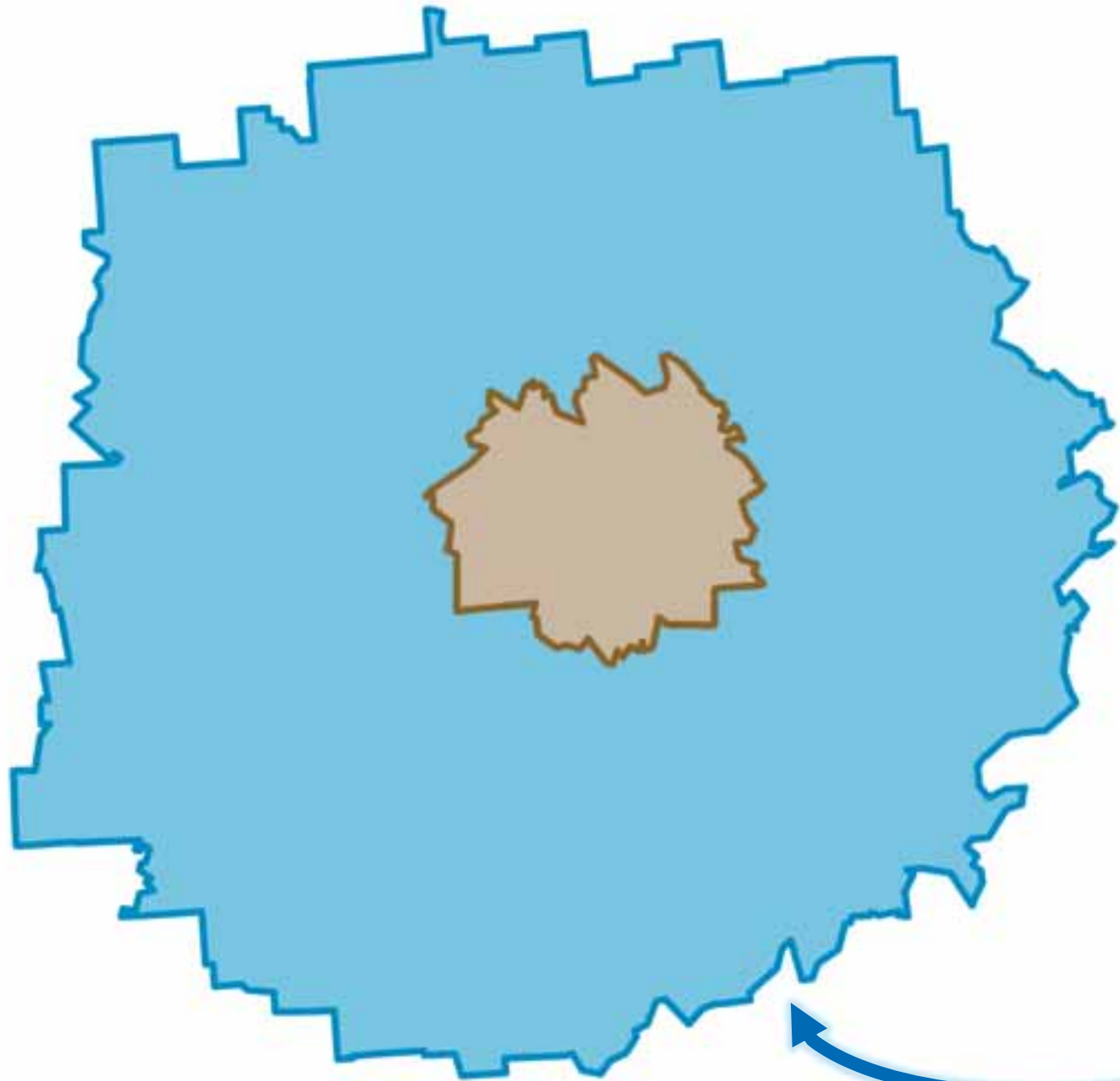
each  
feature  
has a  
value





**Neighborhood**

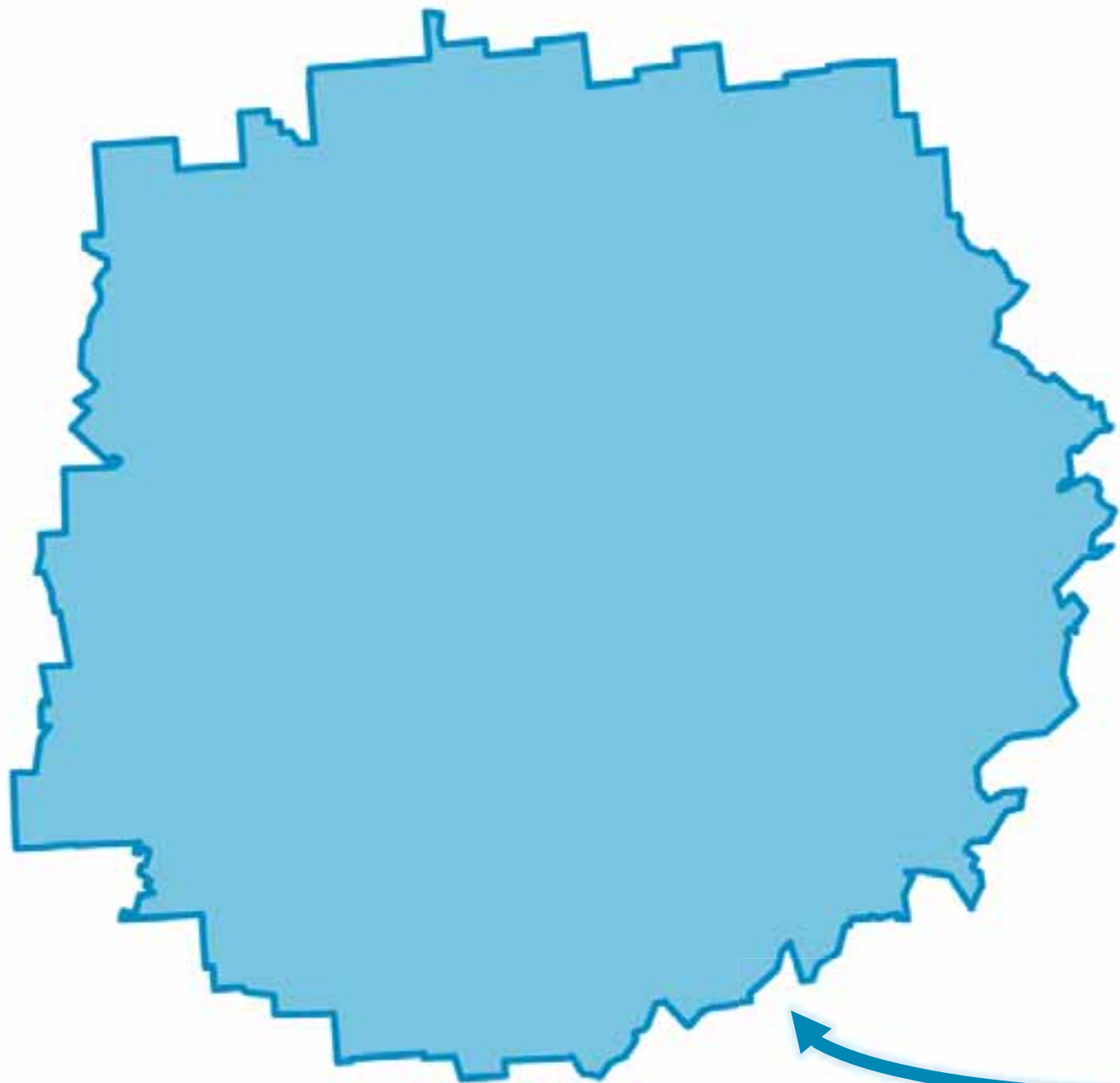




**Study  
Area**





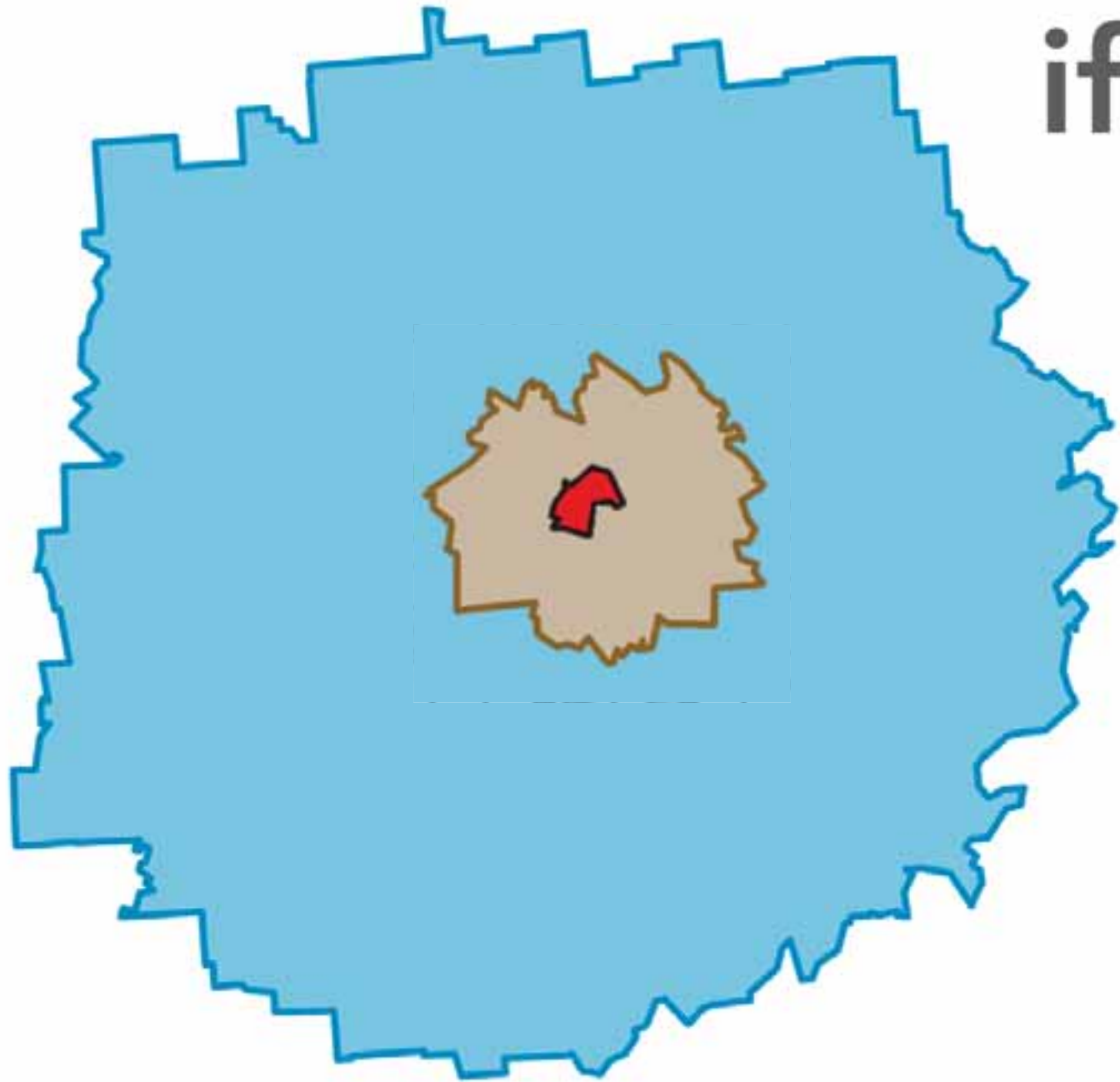


is this



significantly  
different from  
this?



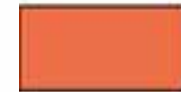


if significantly  
higher...

feature is  
marked as a  
**hot spot!**



Hot Spot - 90% Confidence

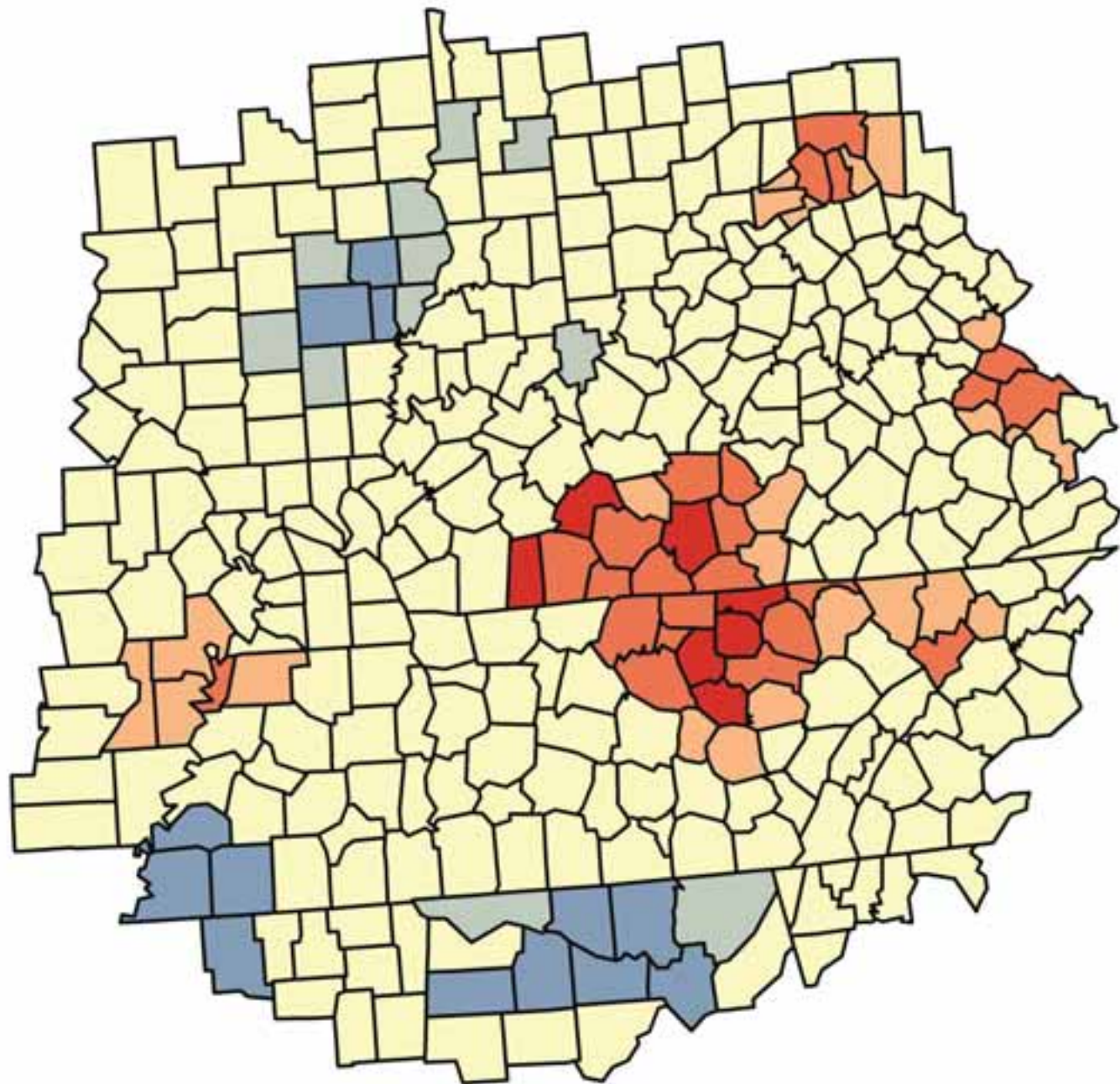





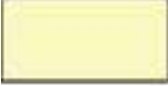



Hot Spot - 95% Confidence



Hot Spot - 99% Confidence





-  Cold Spot - 99% Confidence
-  Cold Spot - 95% Confidence
-  Cold Spot - 90% Confidence
-  Not Significant
-  Hot Spot - 90% Confidence
-  Hot Spot - 95% Confidence
-  Hot Spot - 99% Confidence

...how do we know if it's  
**SIGNIFICANTLY**  
different???

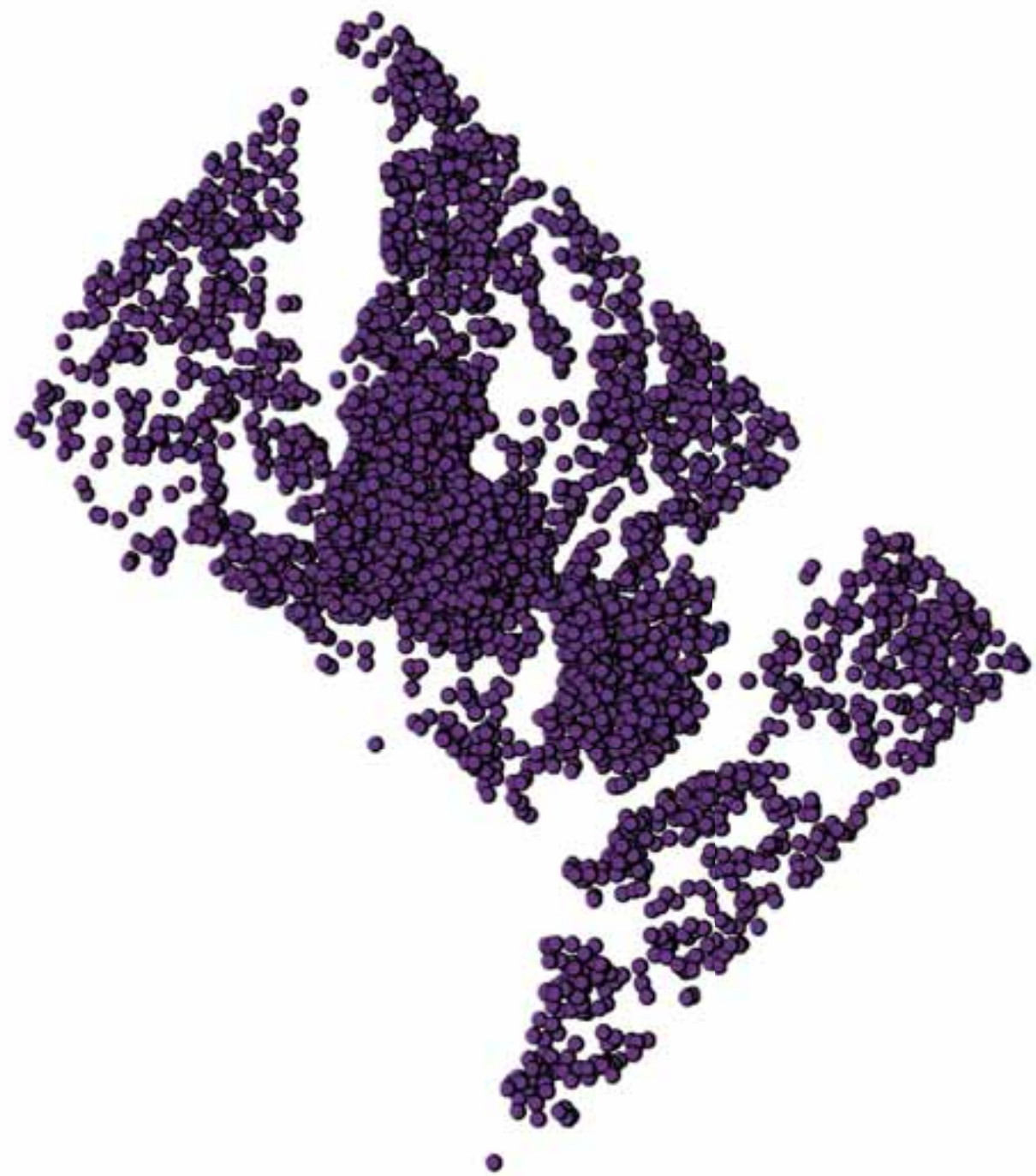


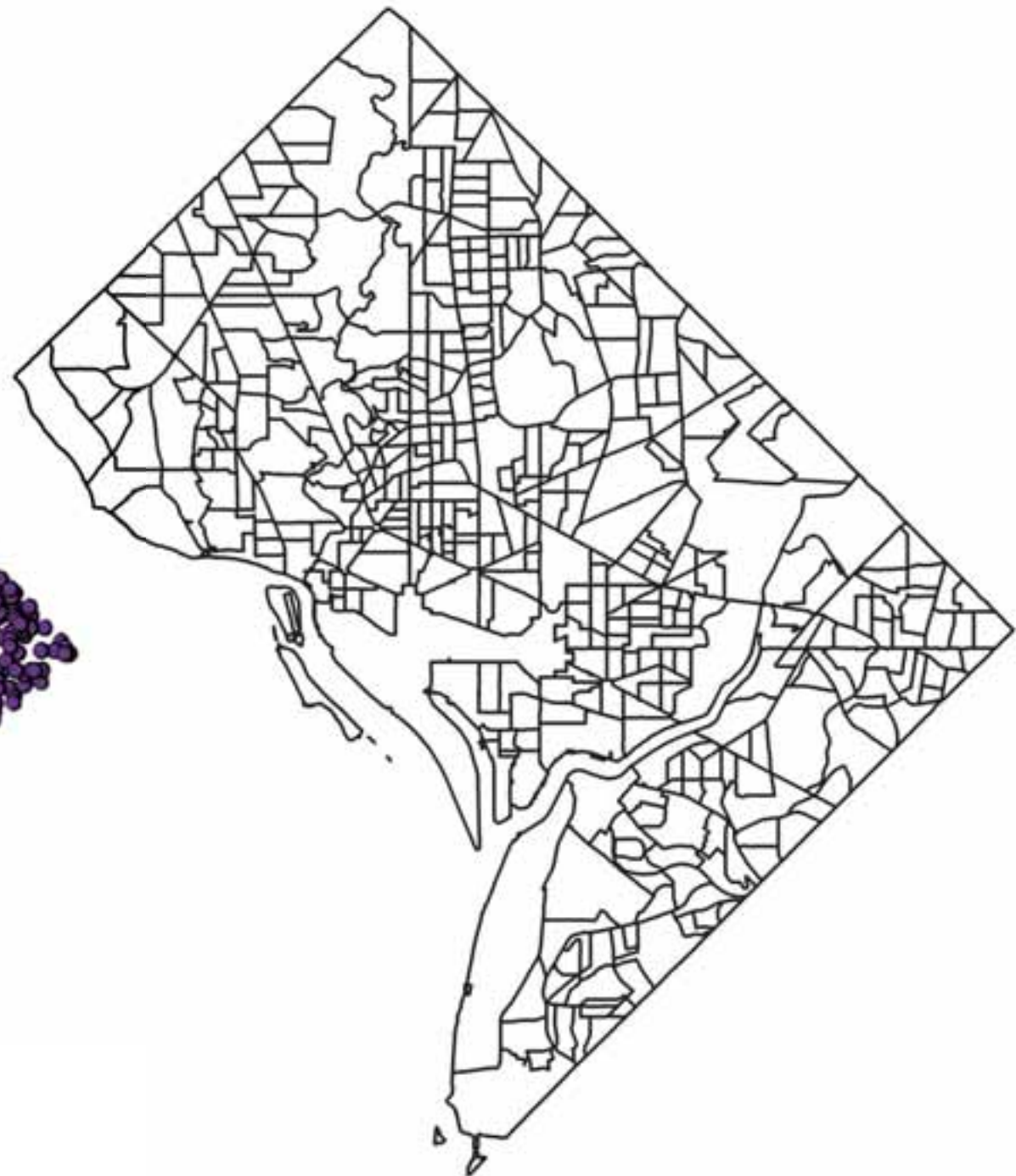
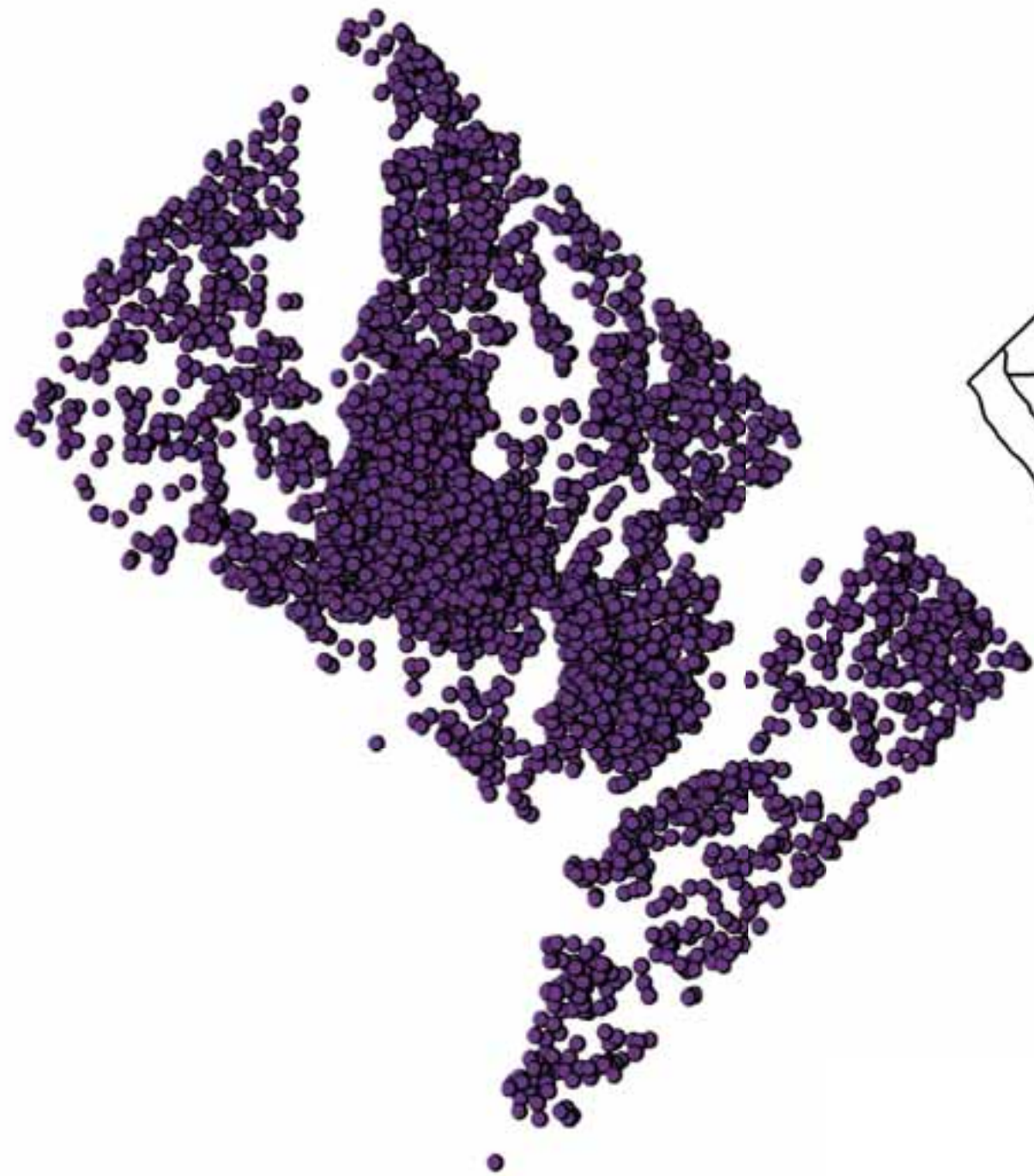
**МАТН!**

**Getis-Ord  $G_i^*$**

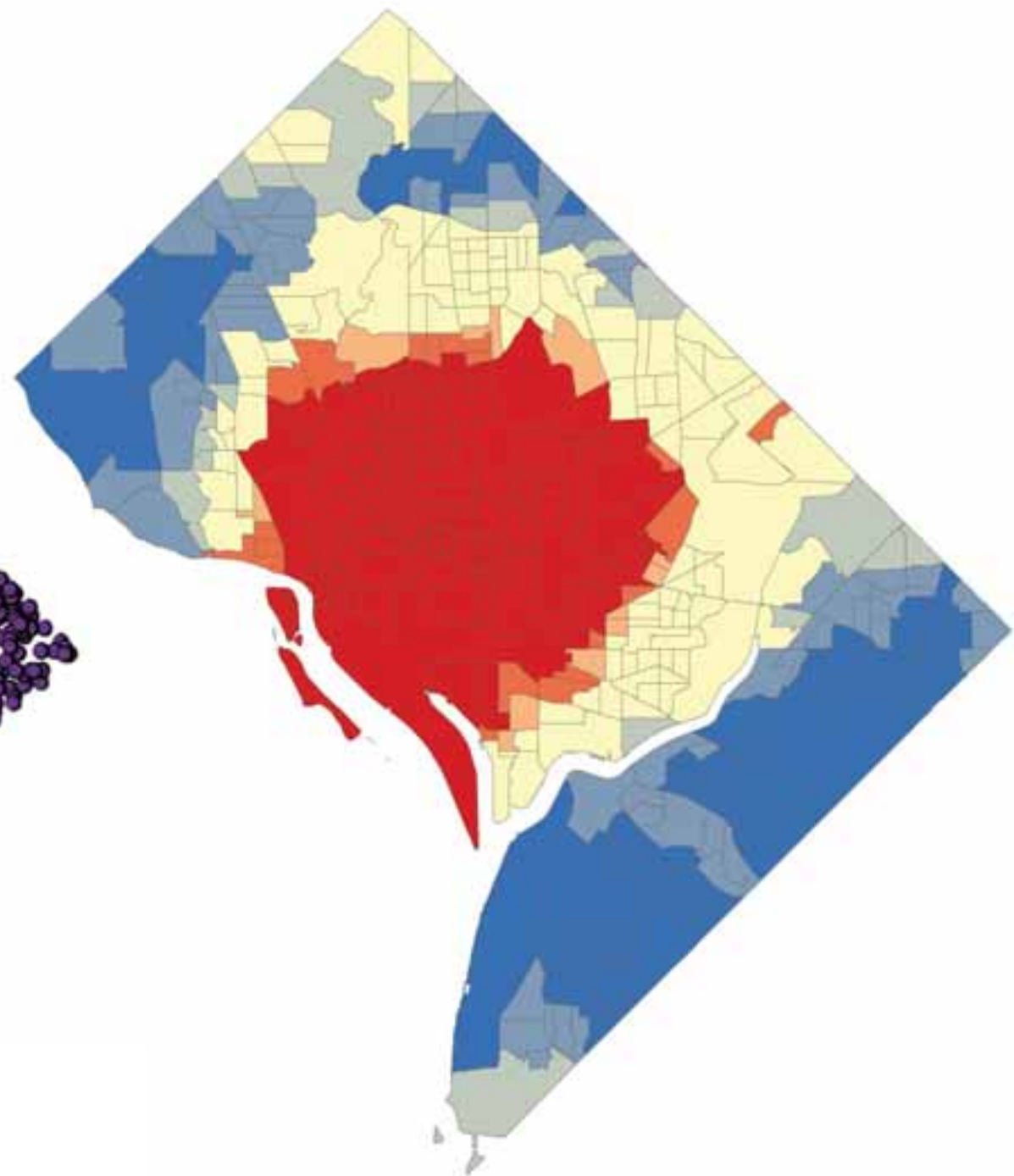
**Statistic**

# Points



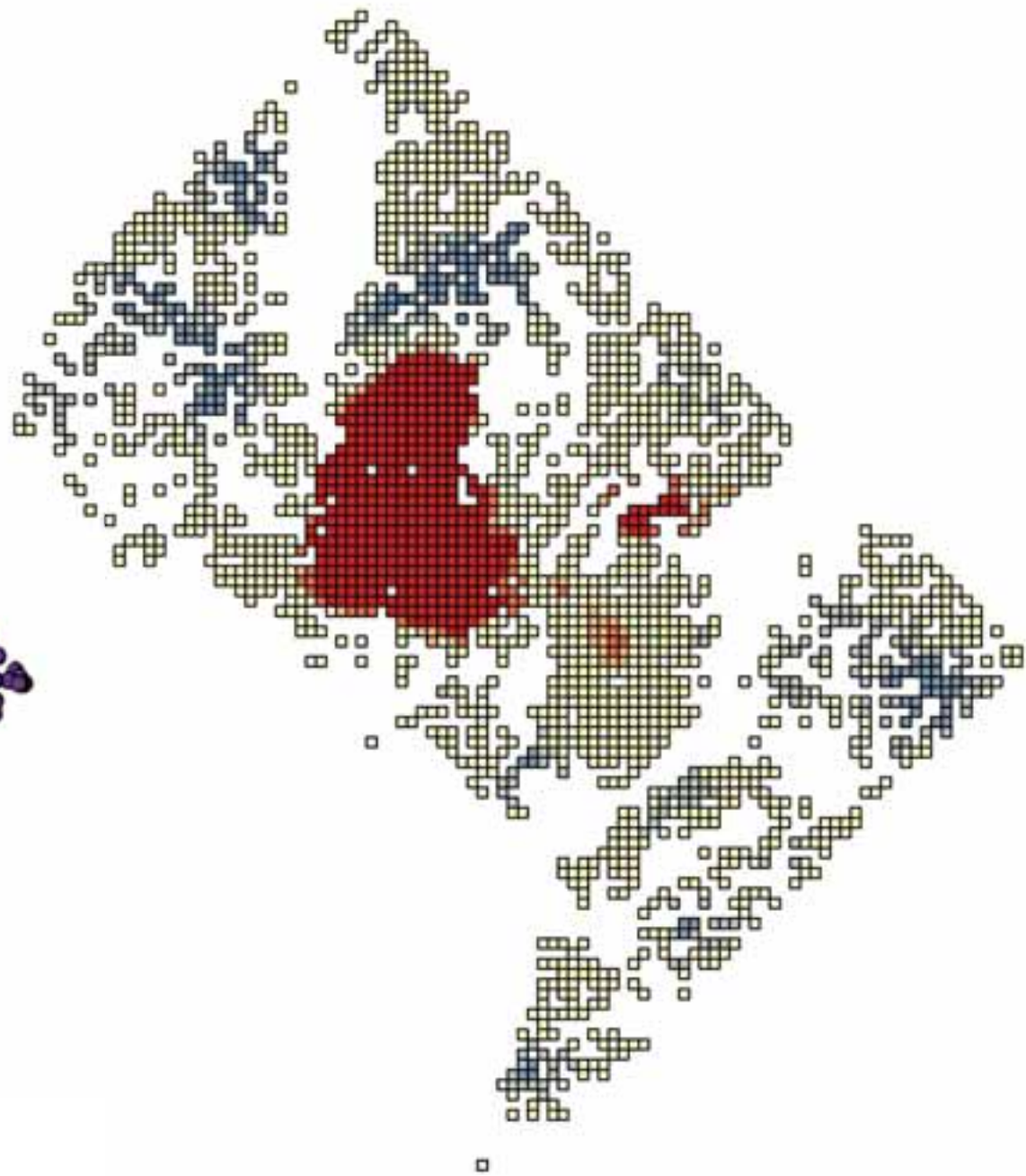


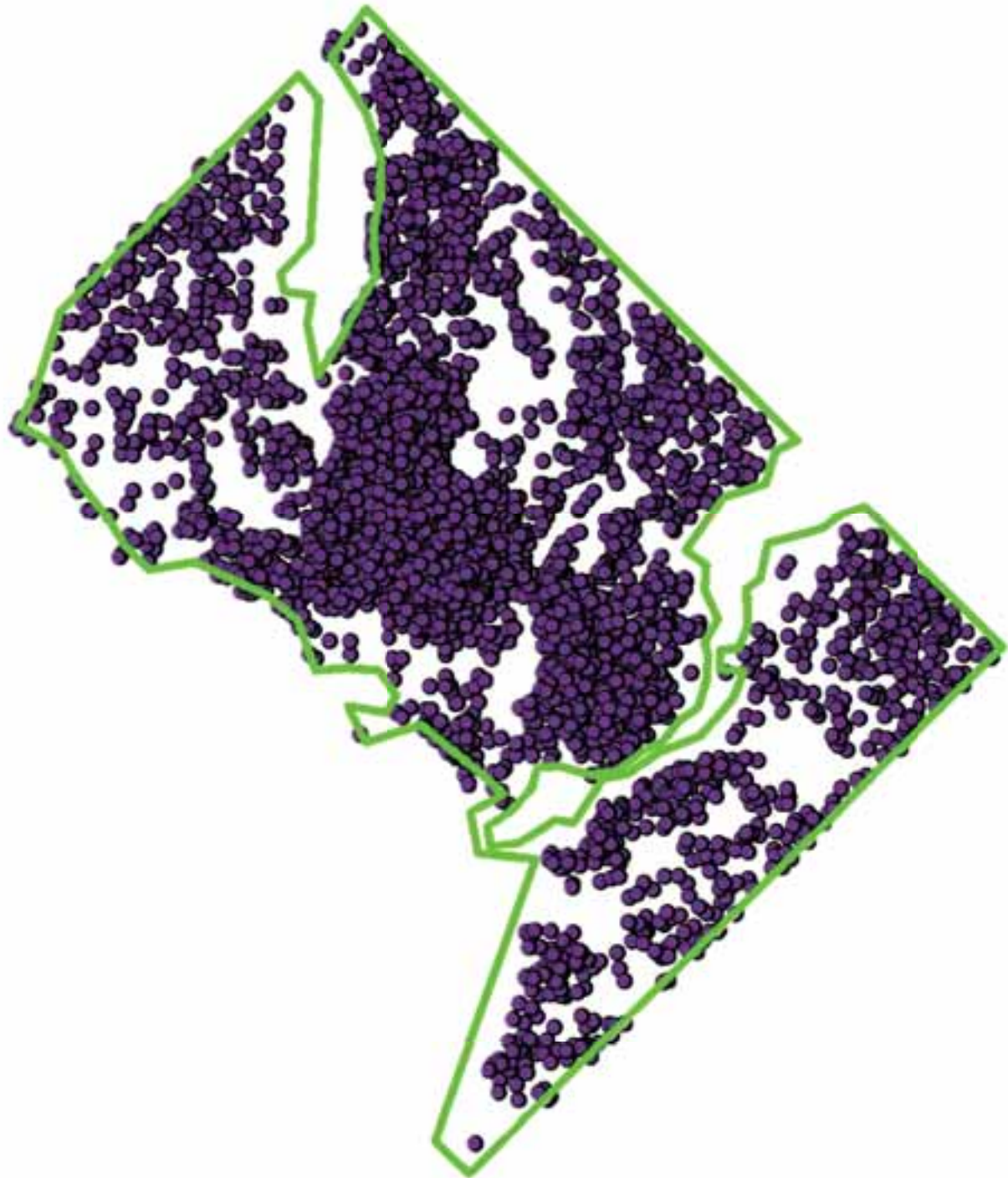




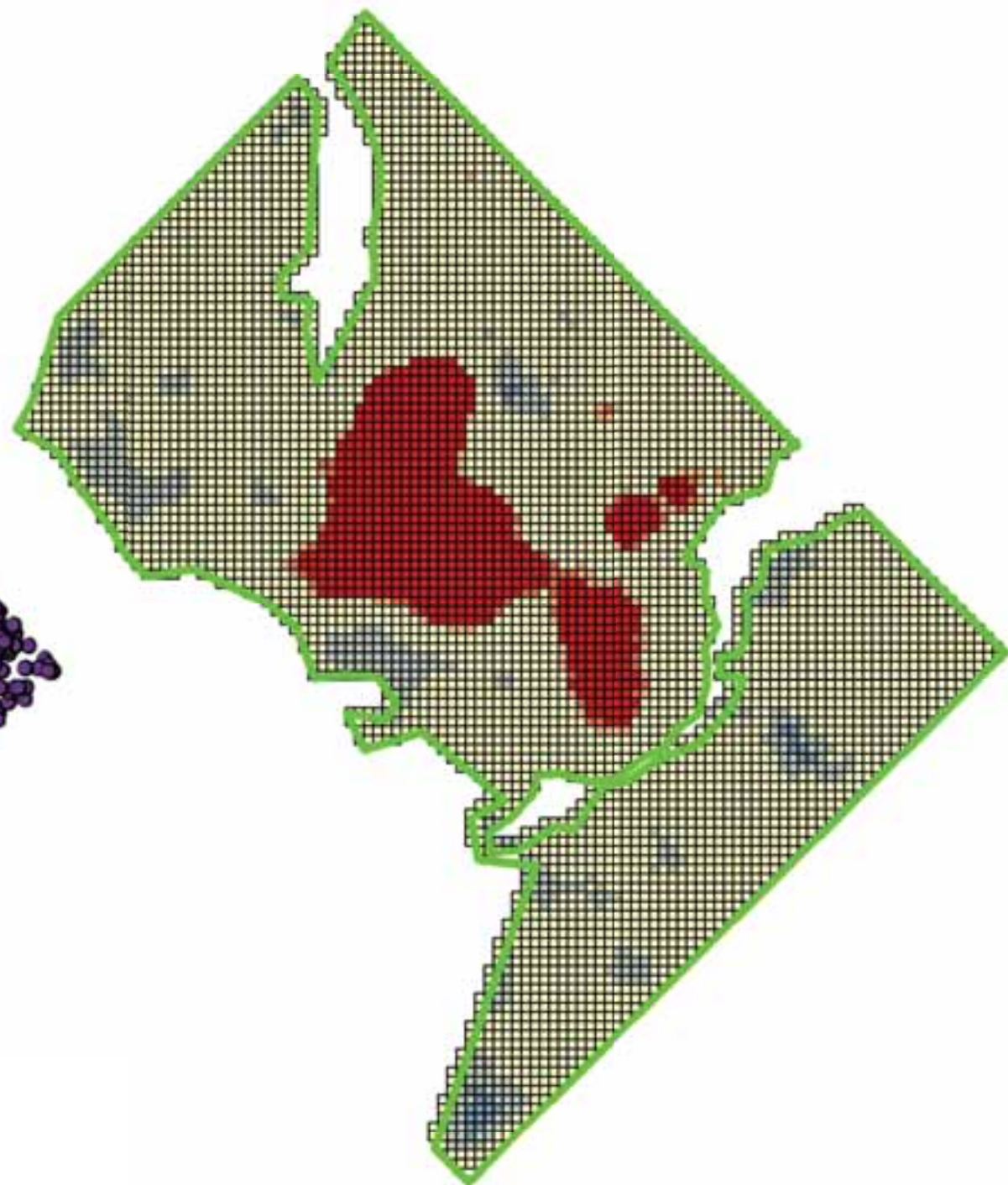












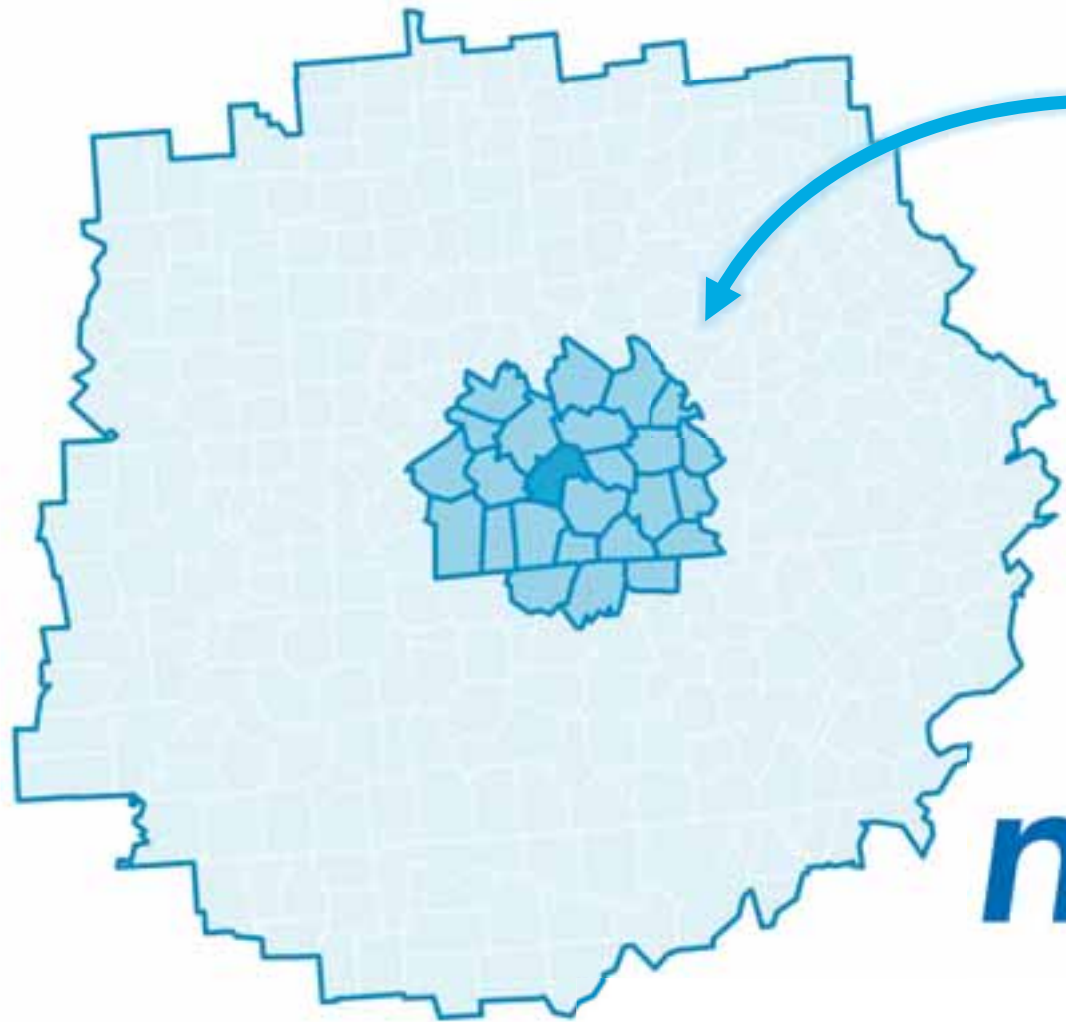


**“Where are  
the Hot  
Spots?”**

not (necessarily)

the same as

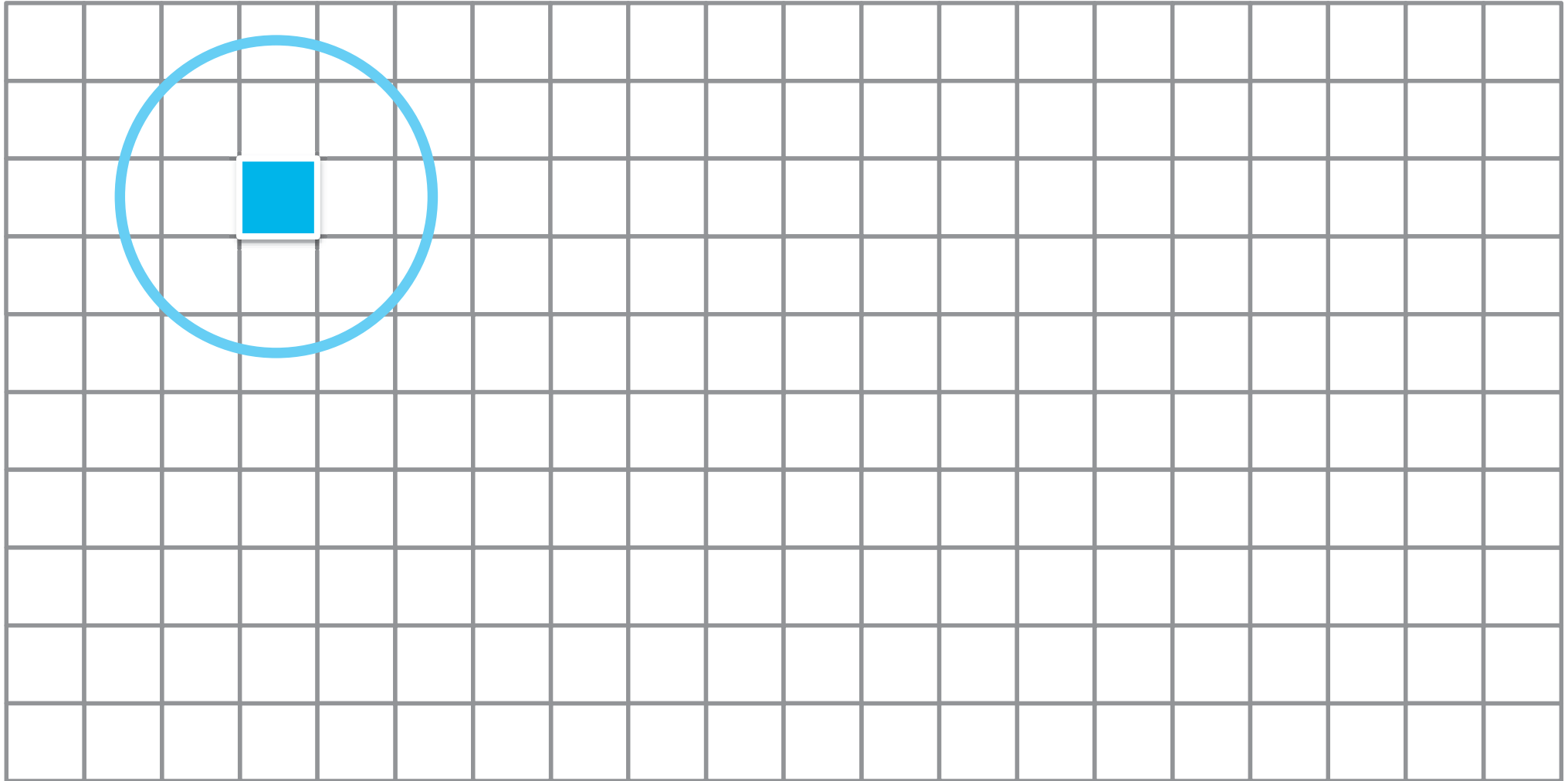
**“Where are  
the highest  
values?”**



How are  
*neighborhood*  
sizes determined?

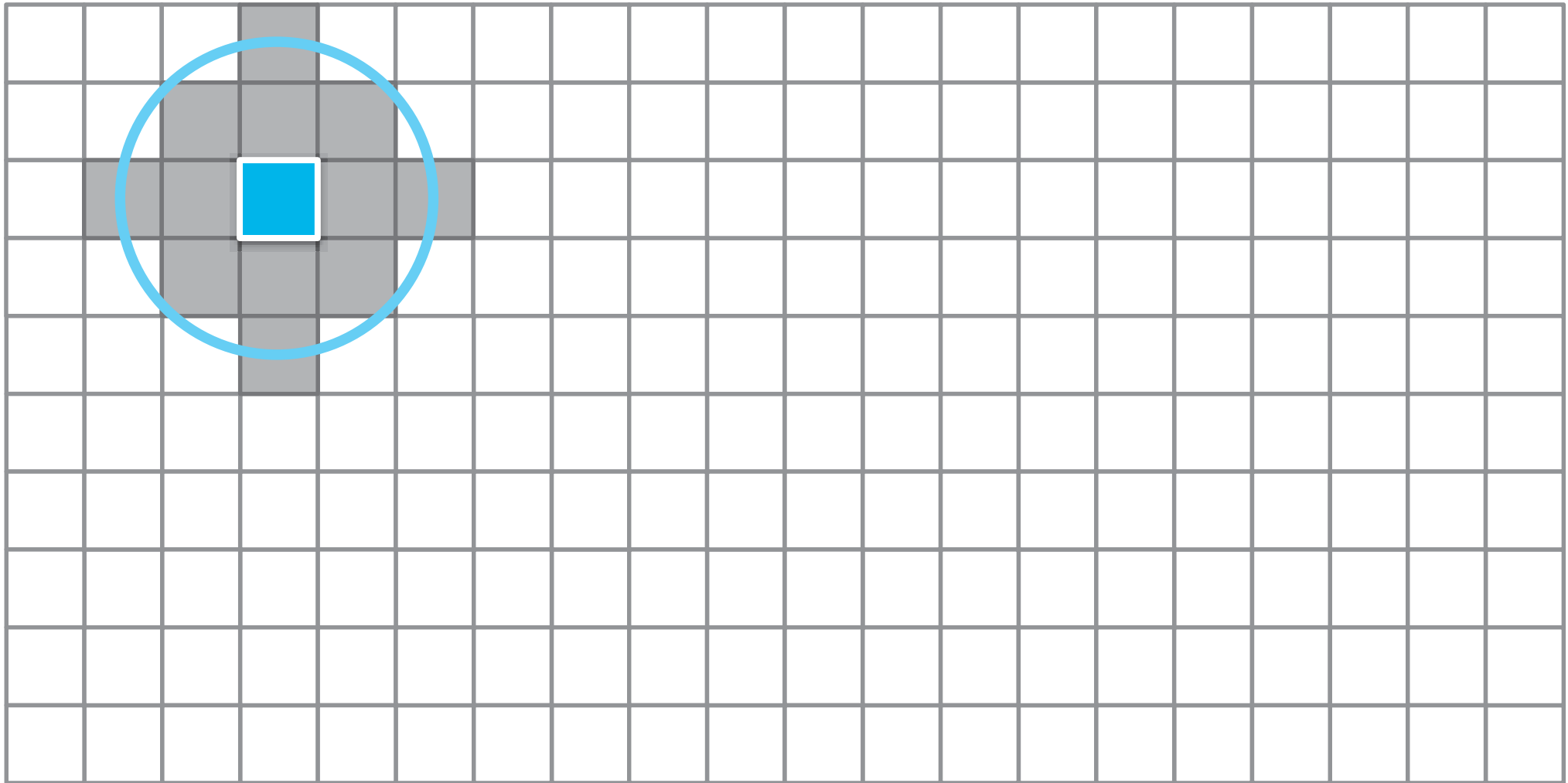


# Fixed Distance Band

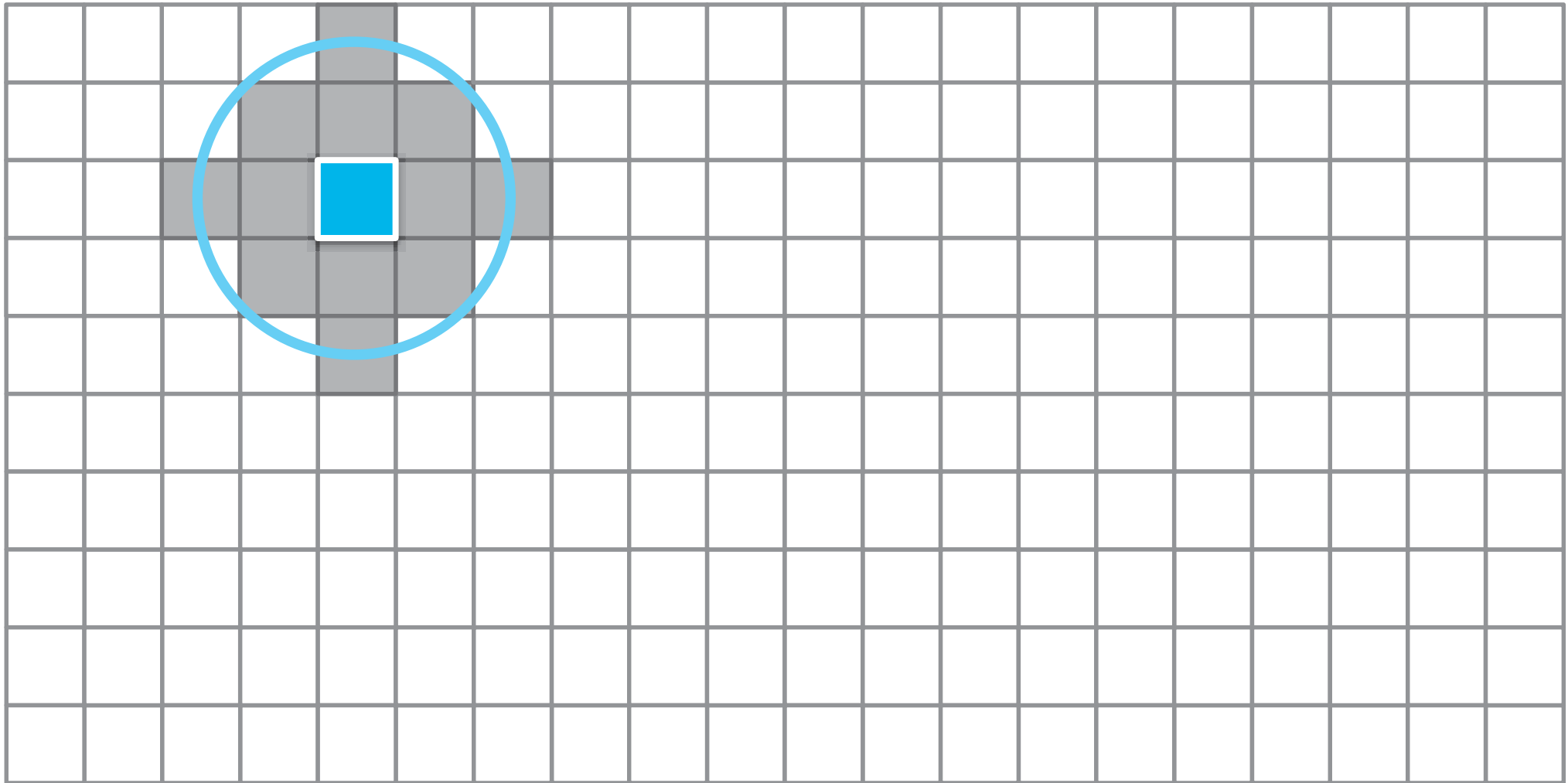




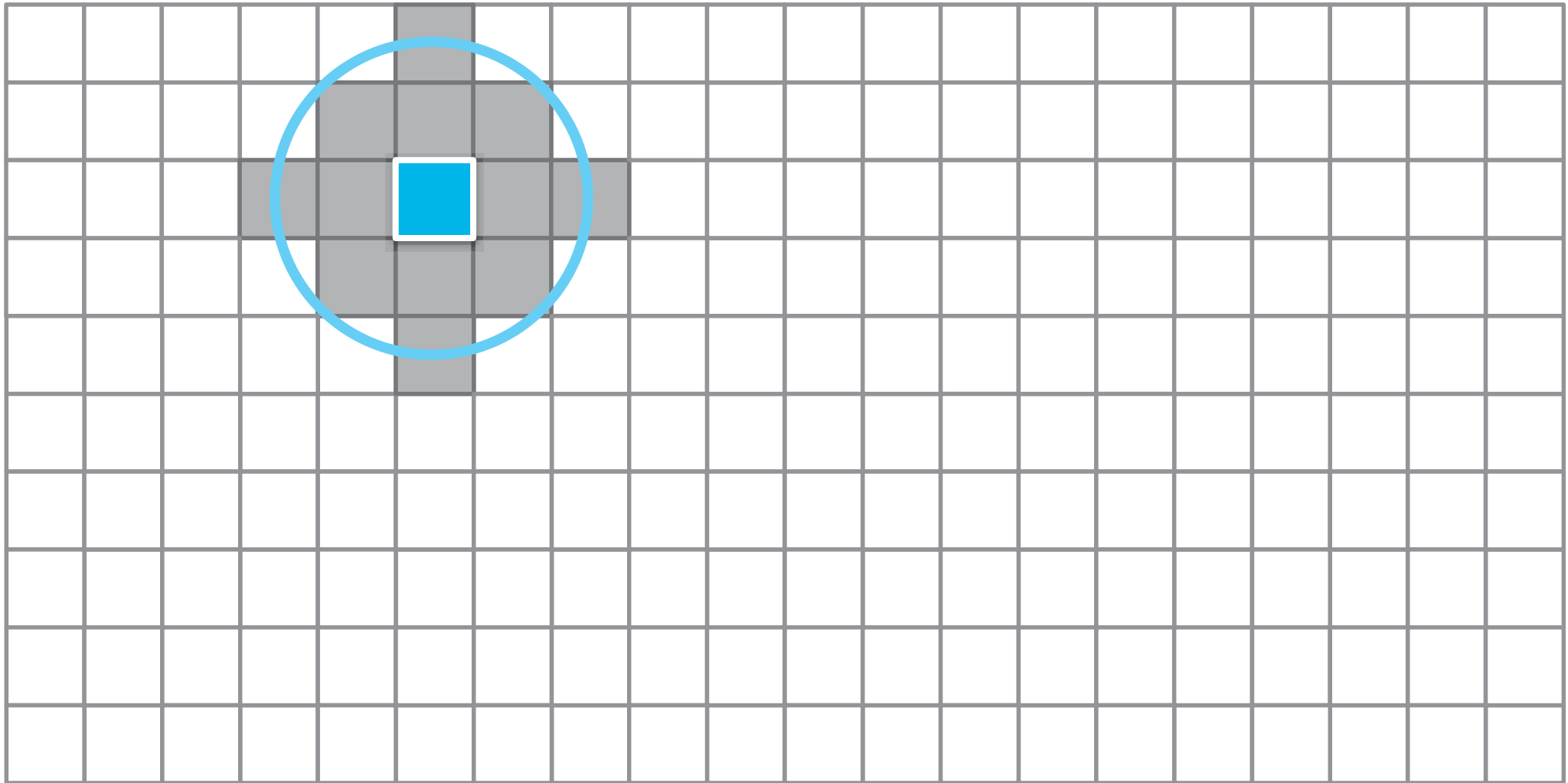
# Fixed Distance Band



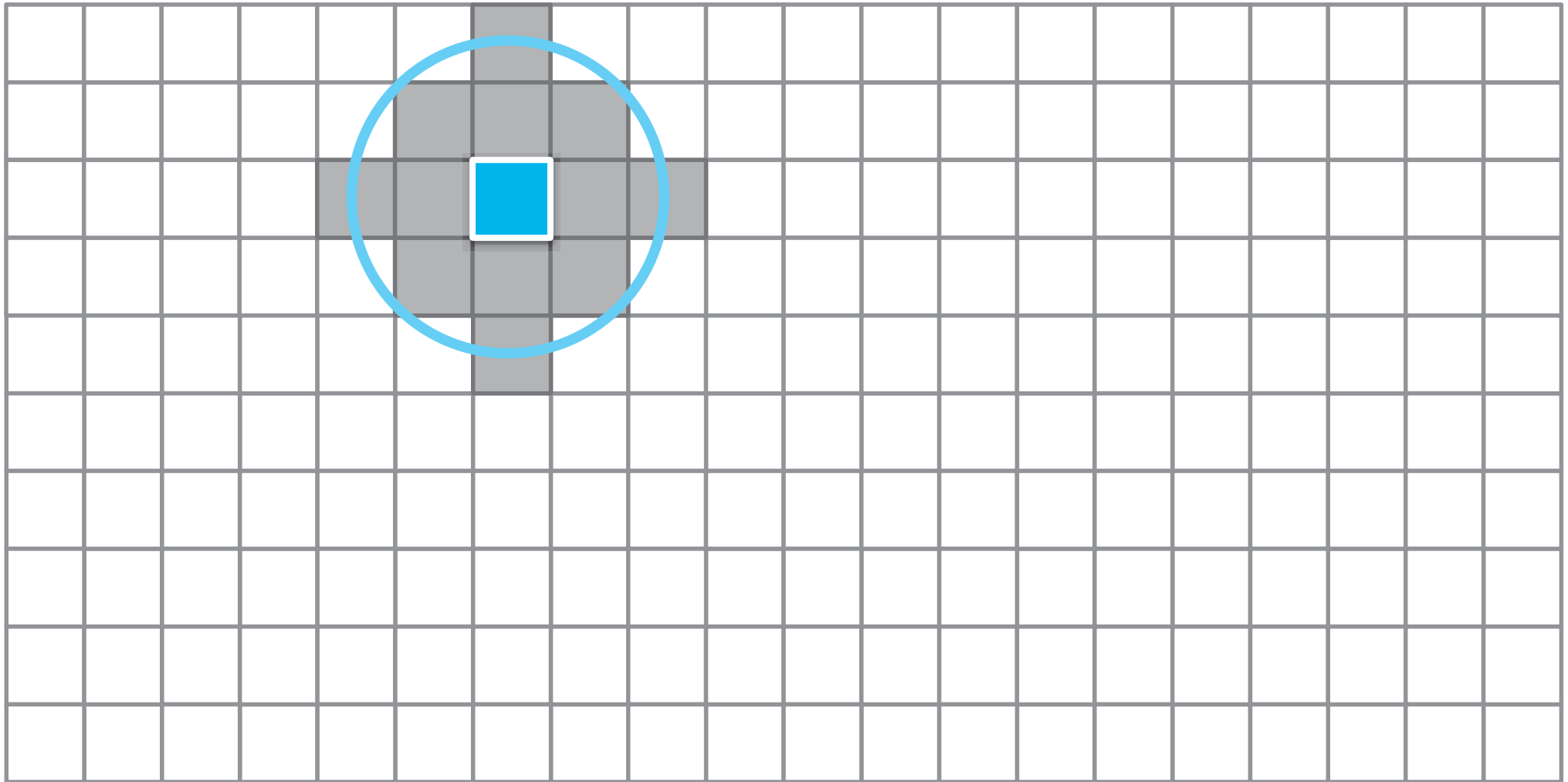
# Fixed Distance Band



# Fixed Distance Band



# Fixed Distance Band





# Inverse Distance

$1/\text{distance}$

# Inverse Distance

$1/\text{distance}$

$\frac{1}{4}$  meters = 0.25

# Inverse Distance

$1/\text{distance}$

$$\frac{1}{4} \text{ meters} = 0.25$$

$$\frac{1}{20} \text{ meters} = 0.05$$

# Inverse Distance

$1/\text{distance}$

$$1/4 \text{ meters} = 0.25$$

$$1/20 \text{ meters} = 0.05$$

farther  
distances  
have  
smaller  
weights



# Inverse Distance<sup>2</sup>

1/distance

$$1/4 \text{ meters} = 0.25$$

$$1/20 \text{ meters} = 0.05$$

# Inverse Distance<sup>2</sup>

$$1/\text{distance}^2$$

$$1/4 \text{ meters} = 0.25$$

$$1/20 \text{ meters} = 0.05$$

# Inverse Distance<sup>2</sup>

$$1/\text{distance}^2$$

$$1/4^2 \text{ meters} = 0.25$$

$$1/20 \text{ meters} = 0.05$$

# Inverse Distance<sup>2</sup>

$$1/\text{distance}^2$$

$$1/4^2 \text{ meters} = 0.0625$$

$$1/20 \text{ meters} = 0.05$$

# Inverse Distance<sup>2</sup>

$$1/\text{distance}^2$$

$$1/4^2 \text{ meters} = 0.0625$$

$$1/20^2 \text{ meters} = 0.05$$



# Inverse Distance<sup>2</sup>

$$1/\text{distance}^2$$

$$1/4^2 \text{ meters} = 0.0625$$

$$1/20^2 \text{ meters} = 0.0025$$

# Inverse Distance<sup>2</sup>

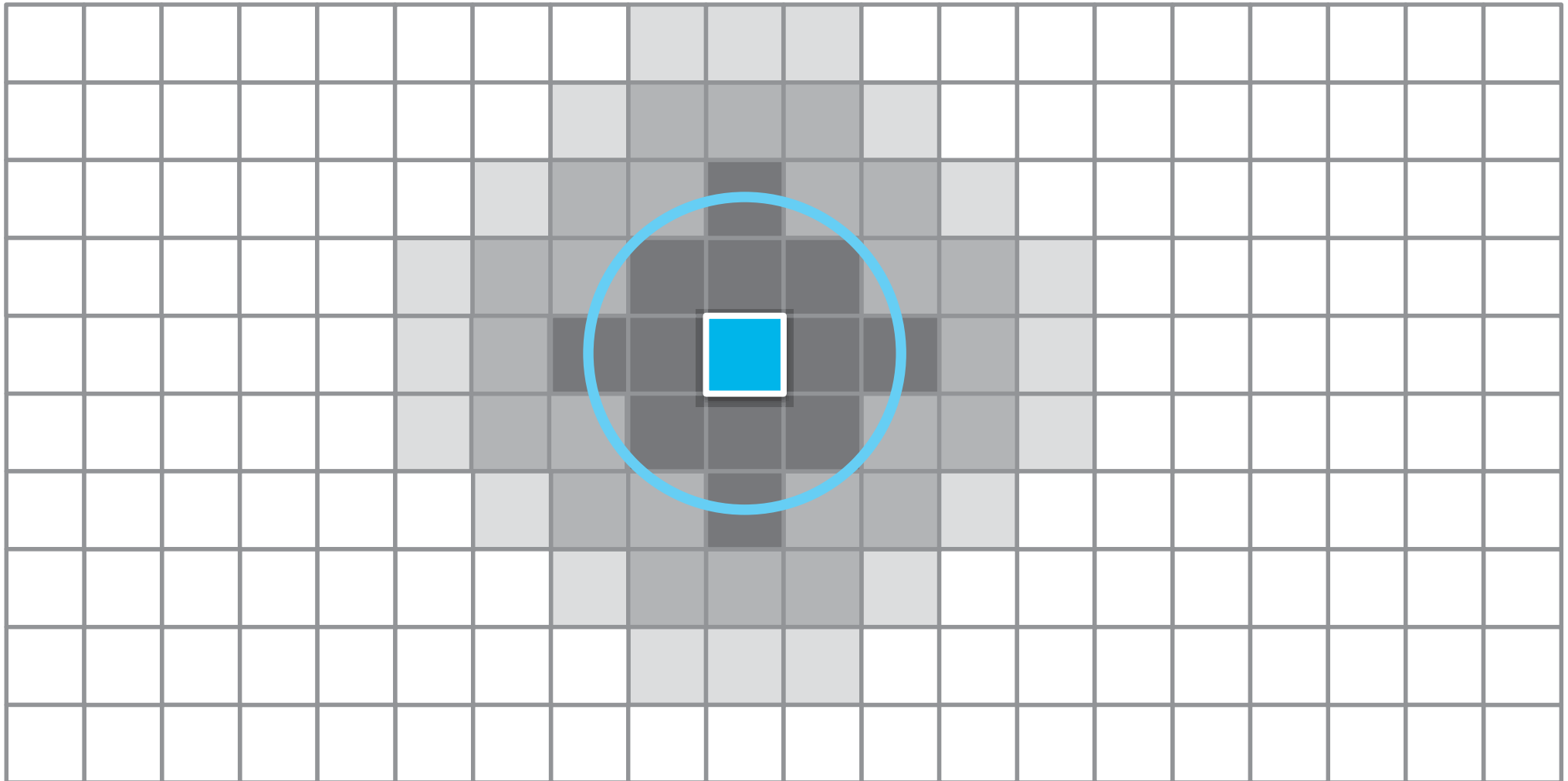
$$1/\text{distance}^2$$

$$1/4^2 \text{ meters} = 0.0625$$

$$1/20^2 \text{ meters} = 0.0025$$

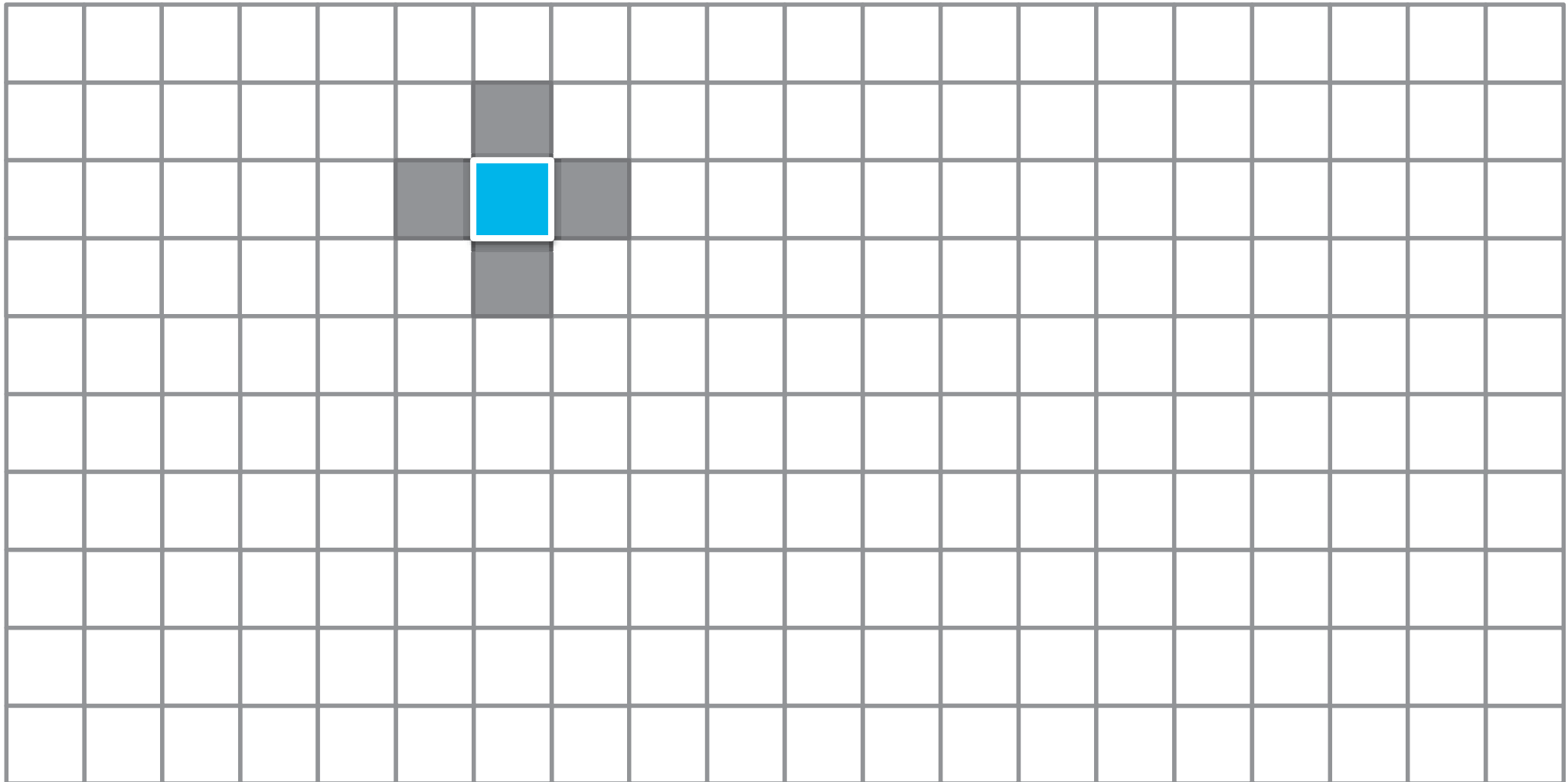
weights  
drop-off  
more  
quickly

# Zone of Indifference



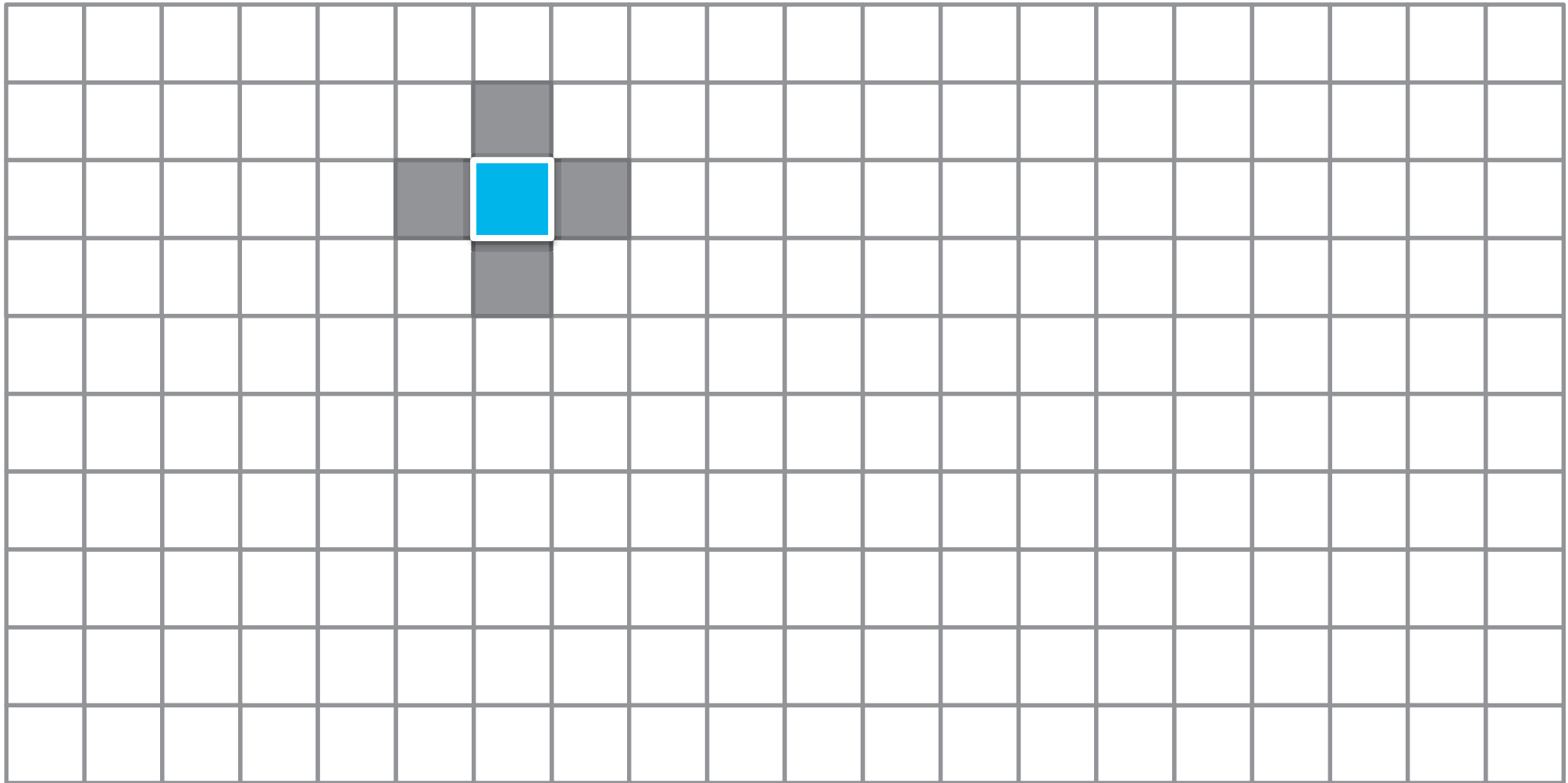
# Contiguity

Edges



# Contiguity

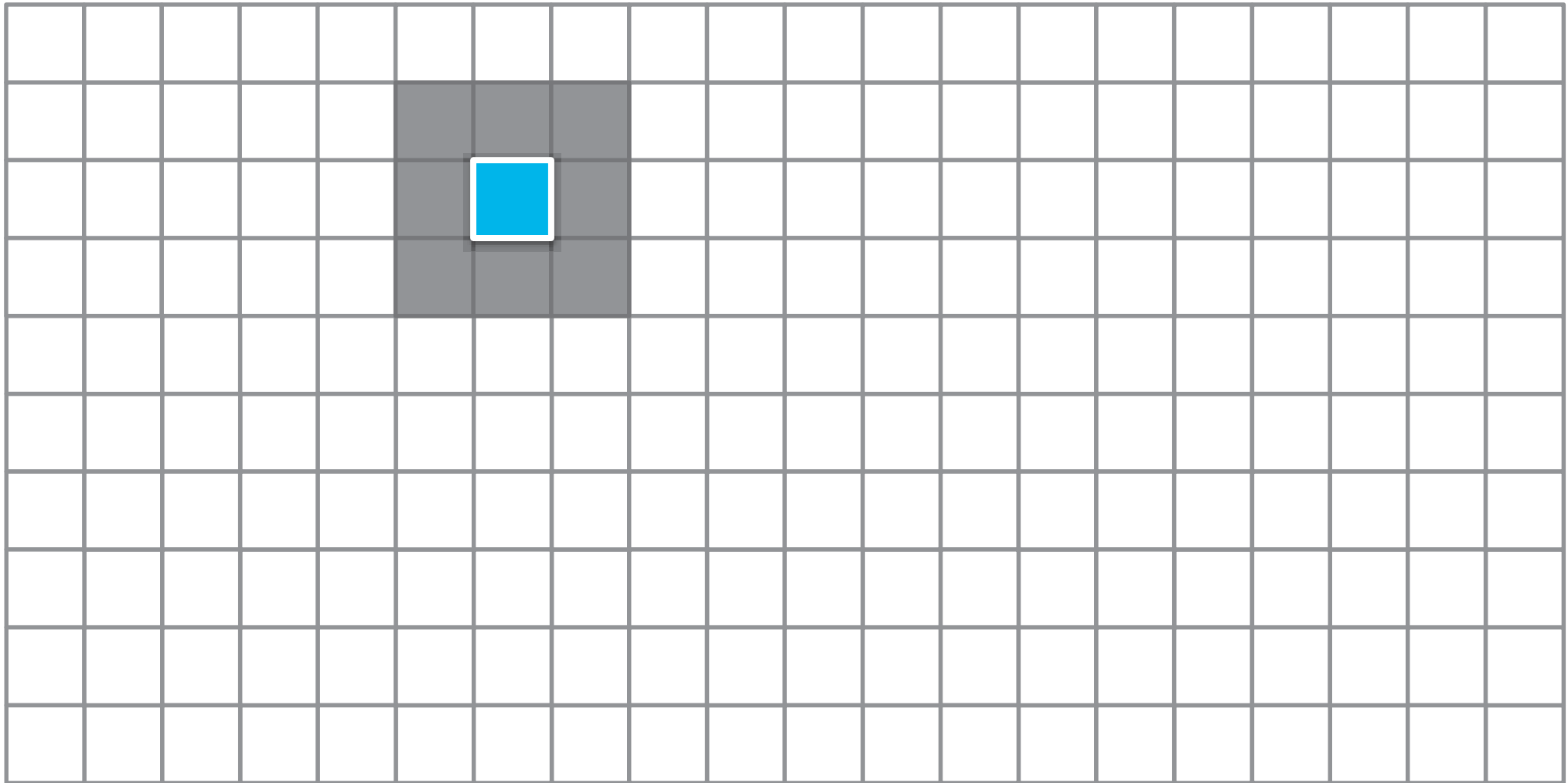
## Rook's Case





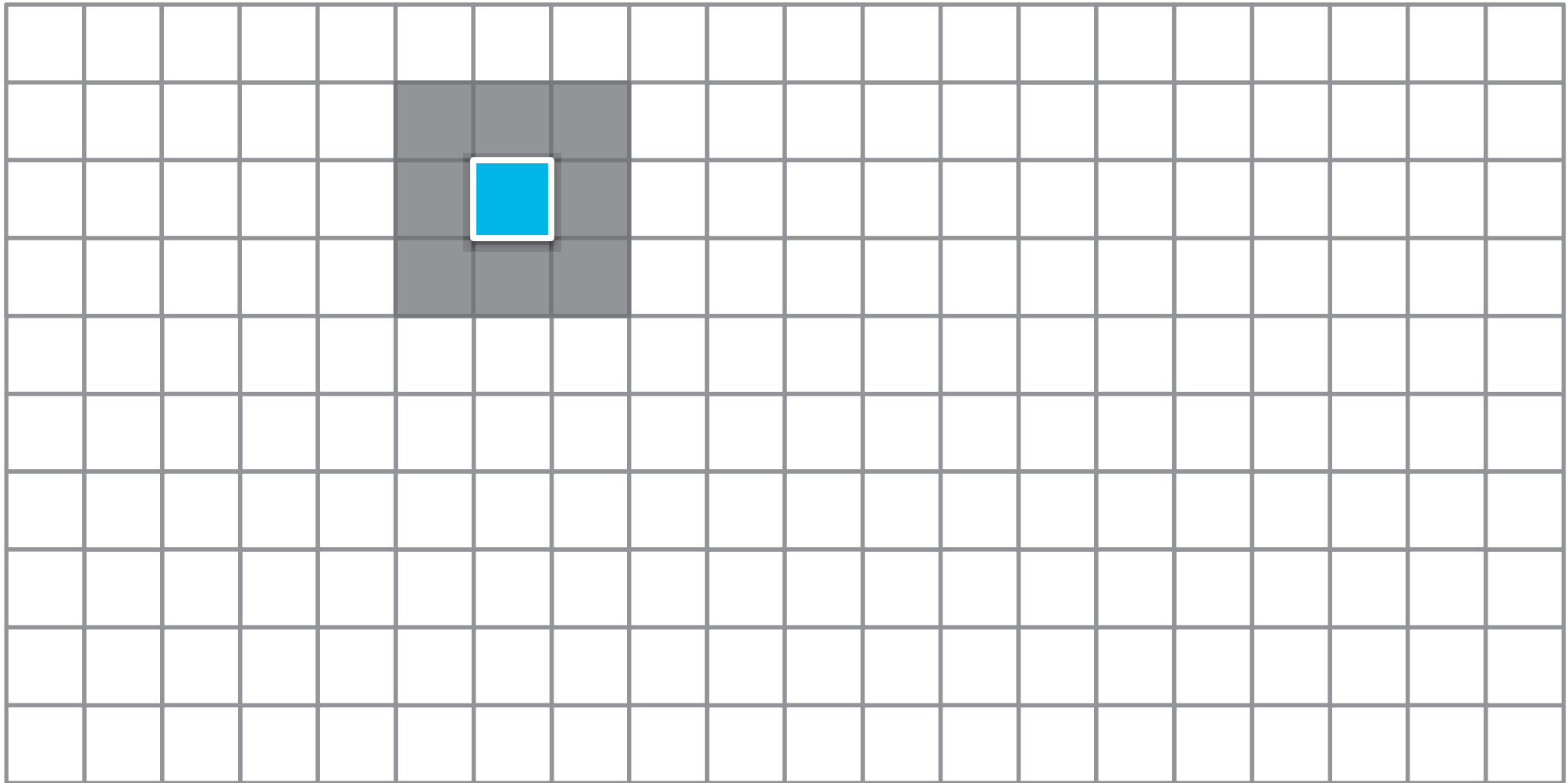
# Contiguity

Edges/Corners



# Contiguity

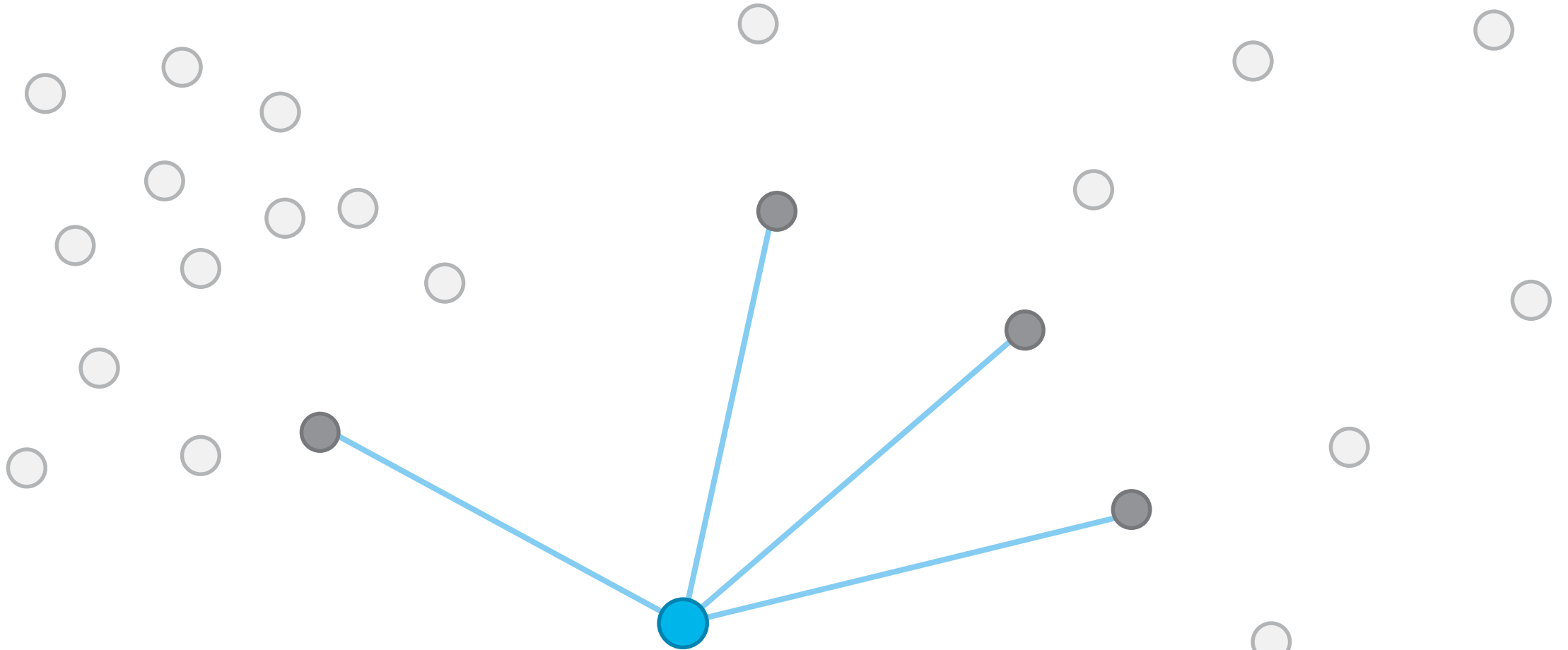
## Queen's Case



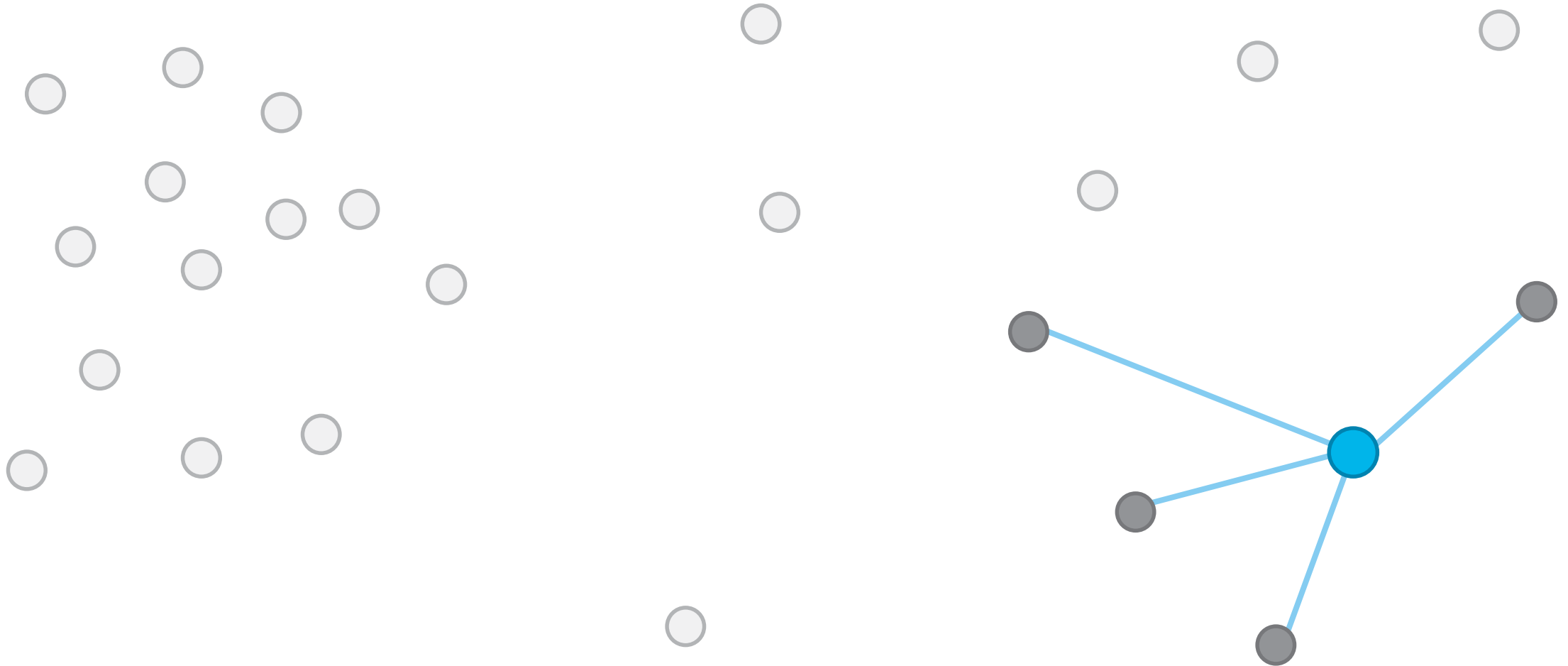
# K Nearest Neighbors



# K Nearest Neighbors

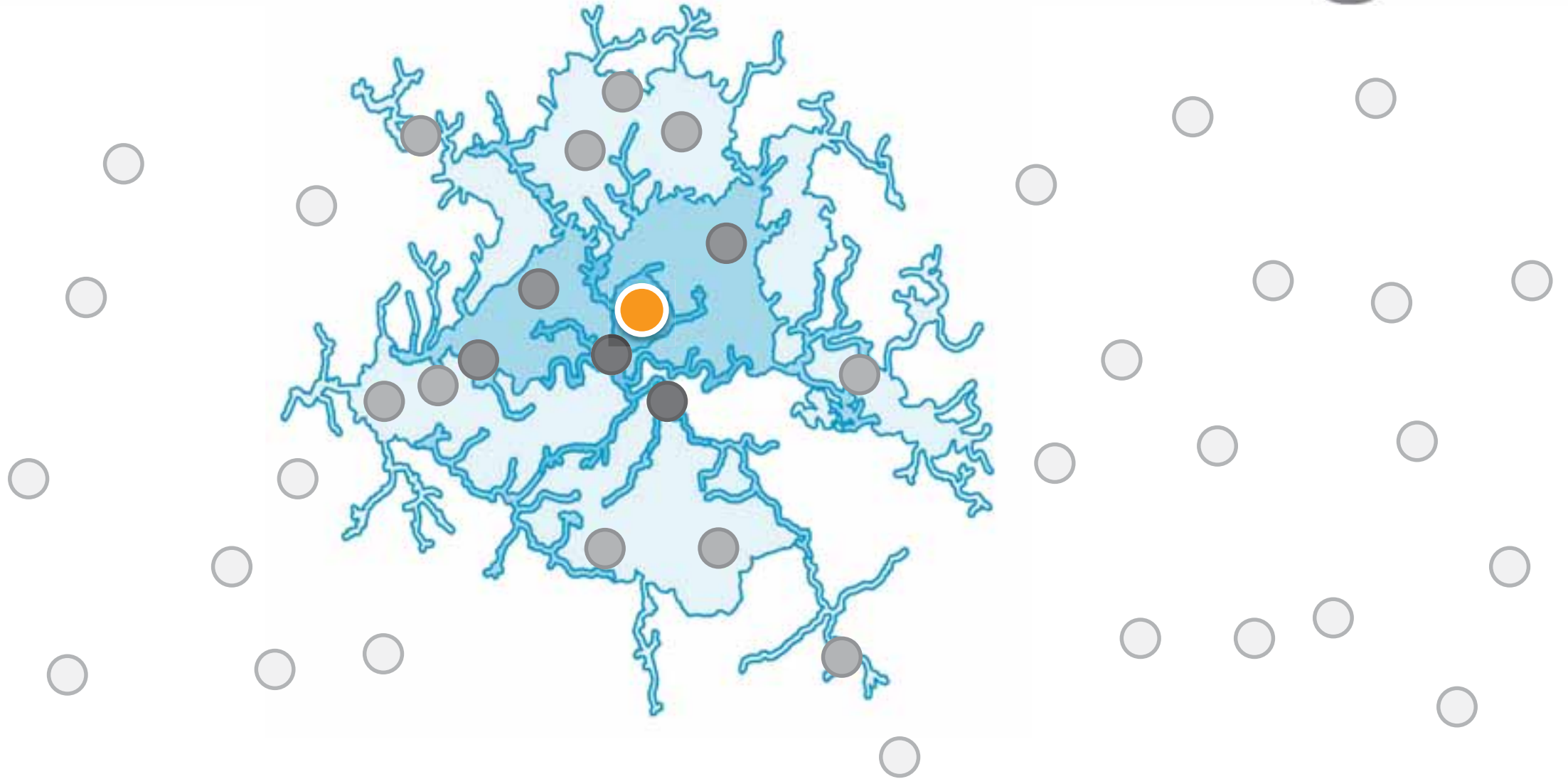


# K Nearest Neighbors

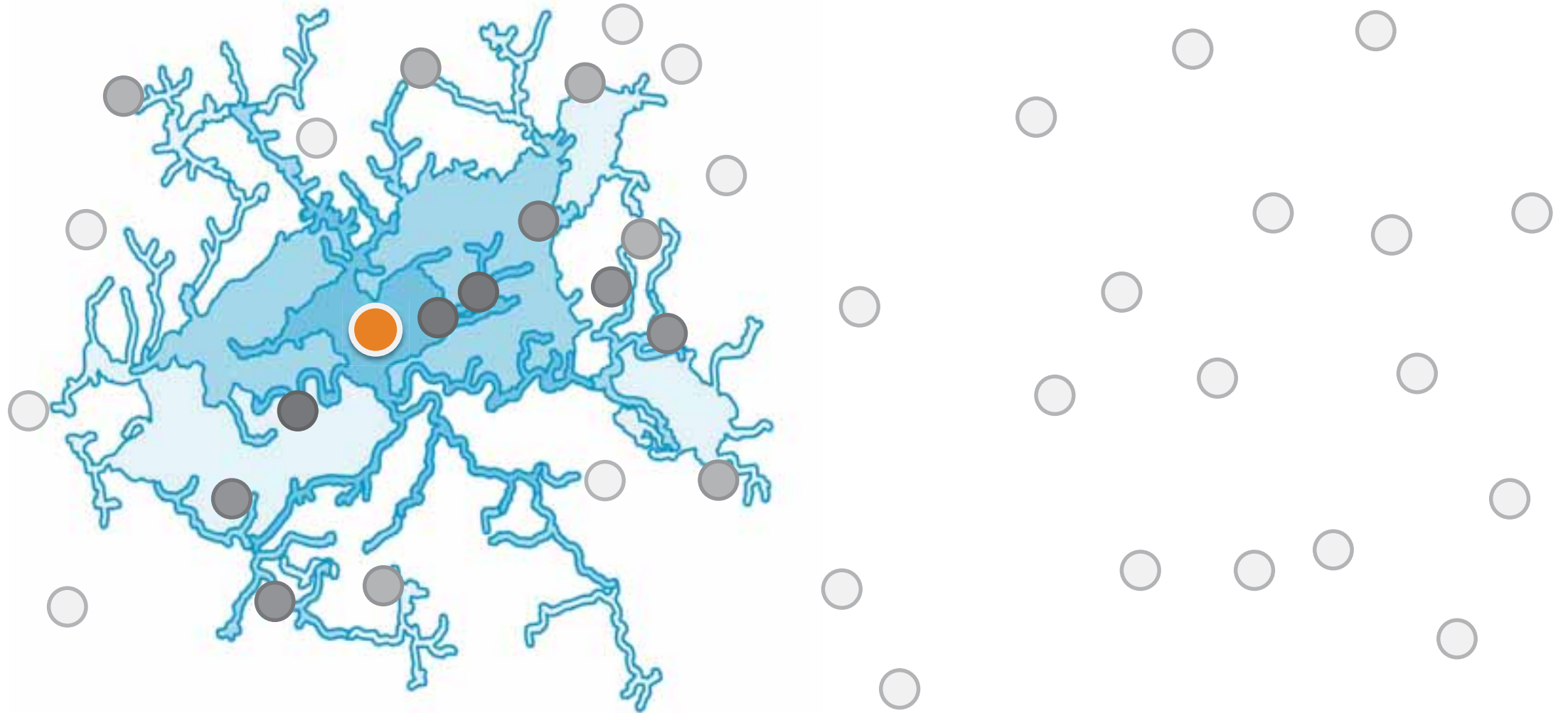




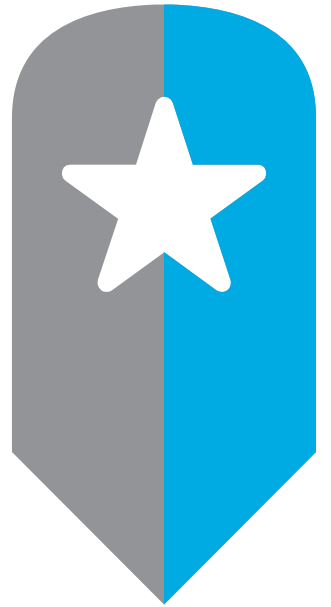
# Network Spatial Weights



# Network Spatial Weights



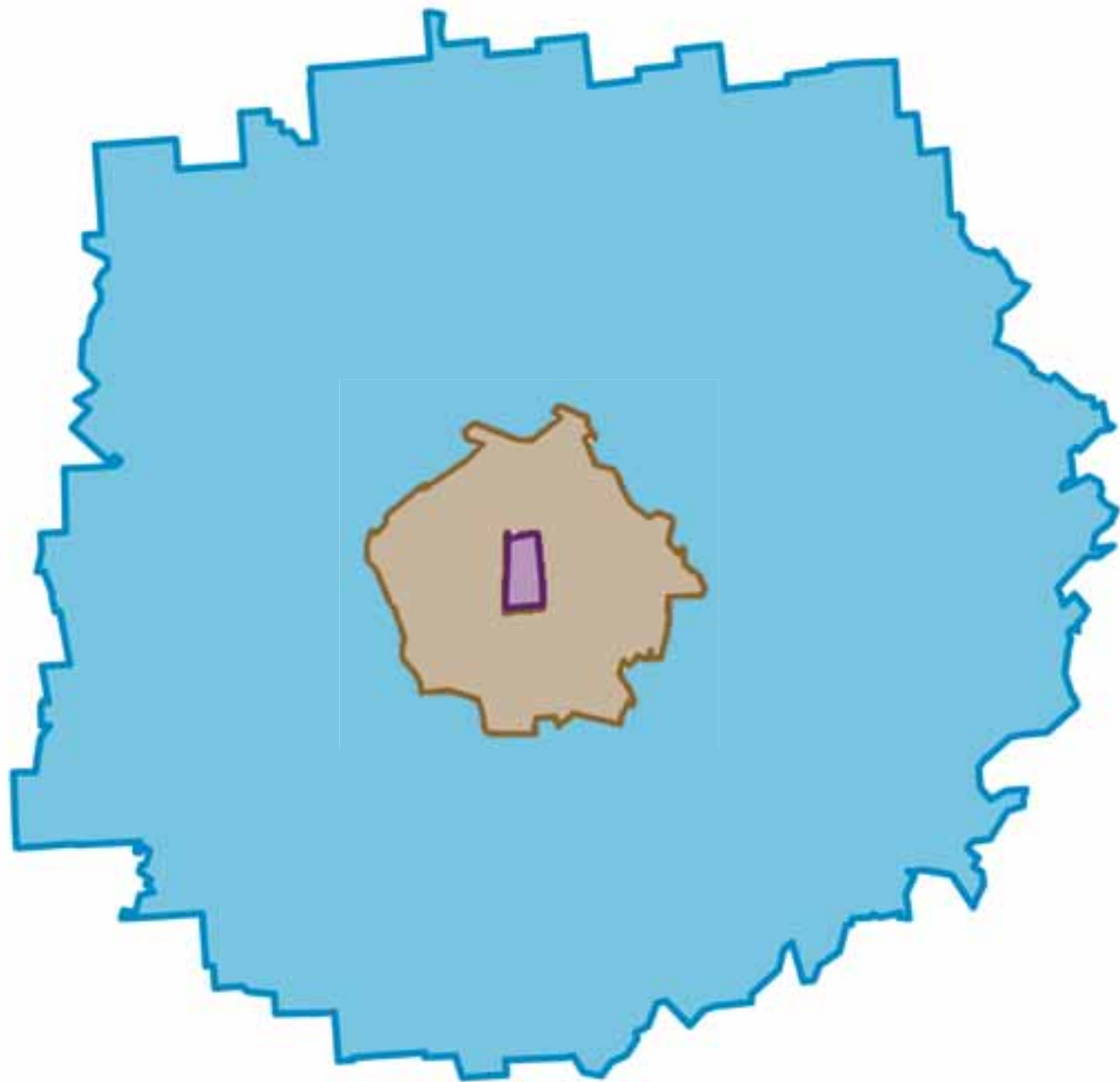
demo



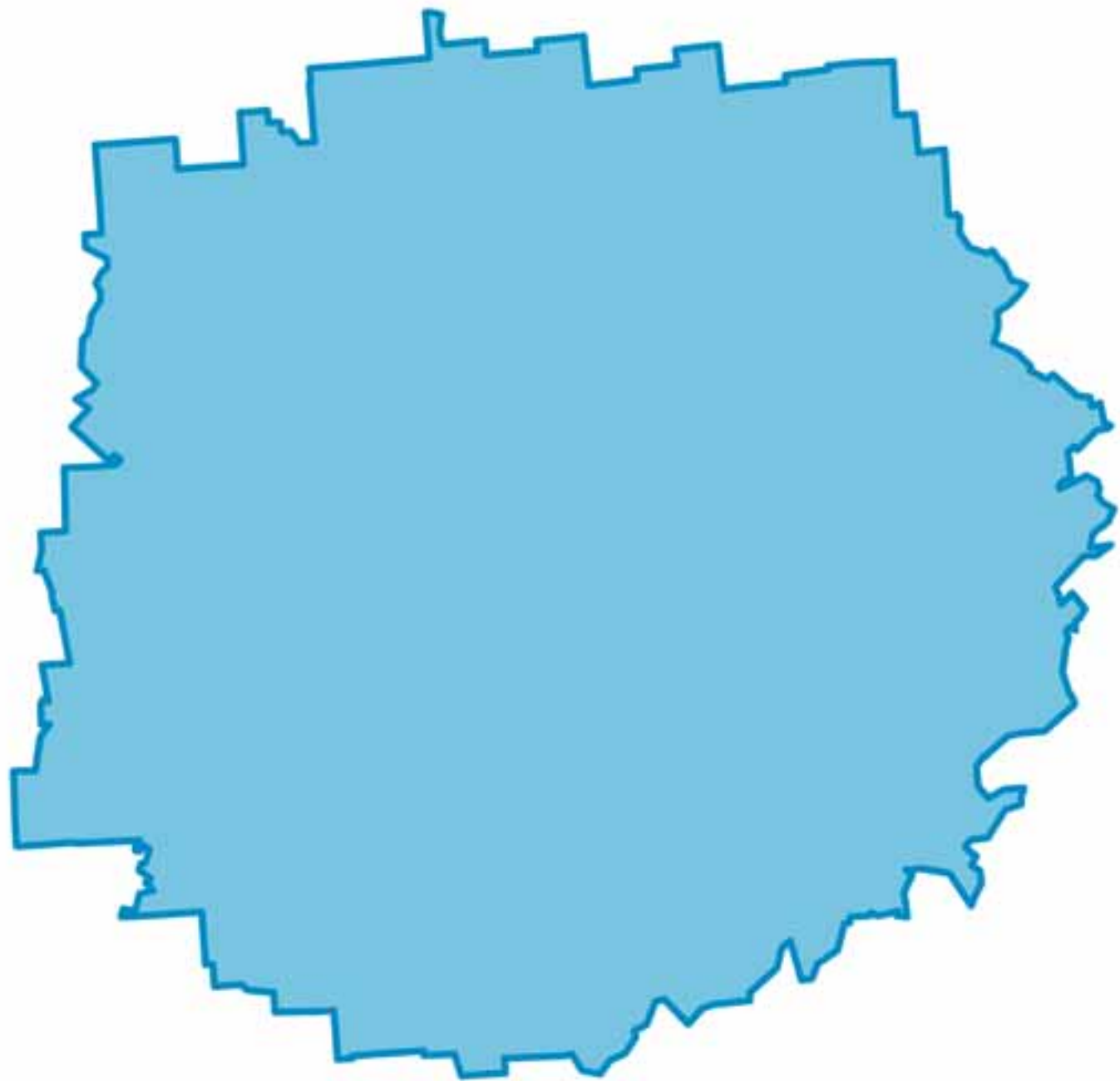
**Cluster and Outlier**

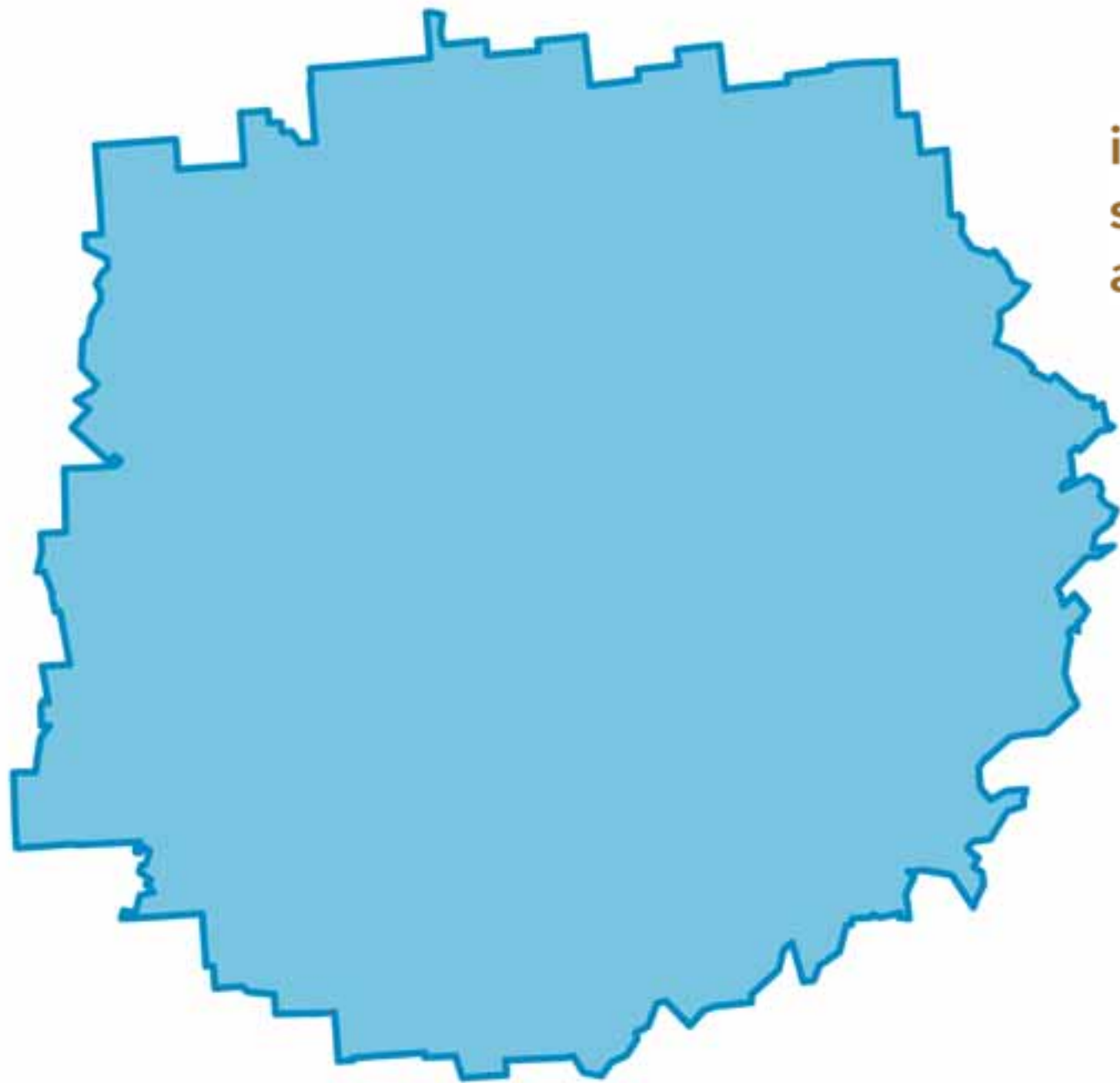
**Analysis**

(Anselin Local Moran's I)



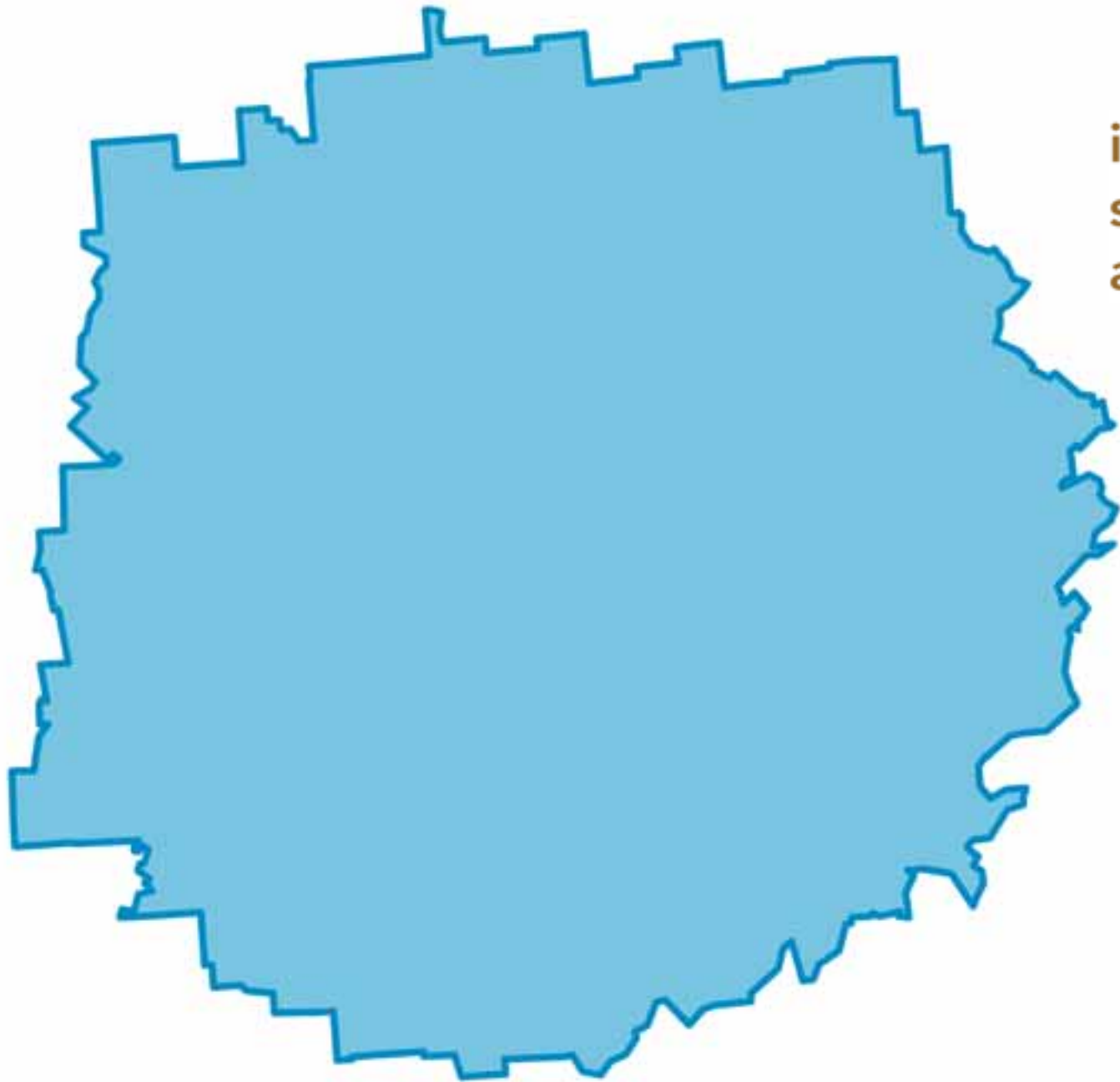




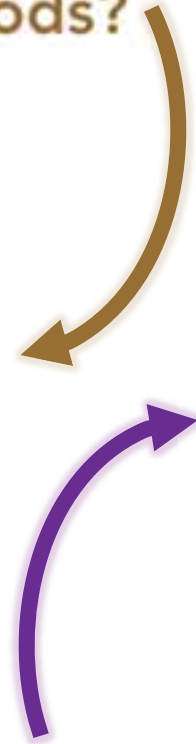


is the neighborhood  
significantly different from  
all other neighborhoods?





is the neighborhood significantly different from all other neighborhoods?



is the feature significantly different from all other features?

feature is **higher** than other features, neighborhood is **lower** than other neighborhoods

**HL**

feature is **higher** than other features, neighborhood is **higher** than other neighborhoods

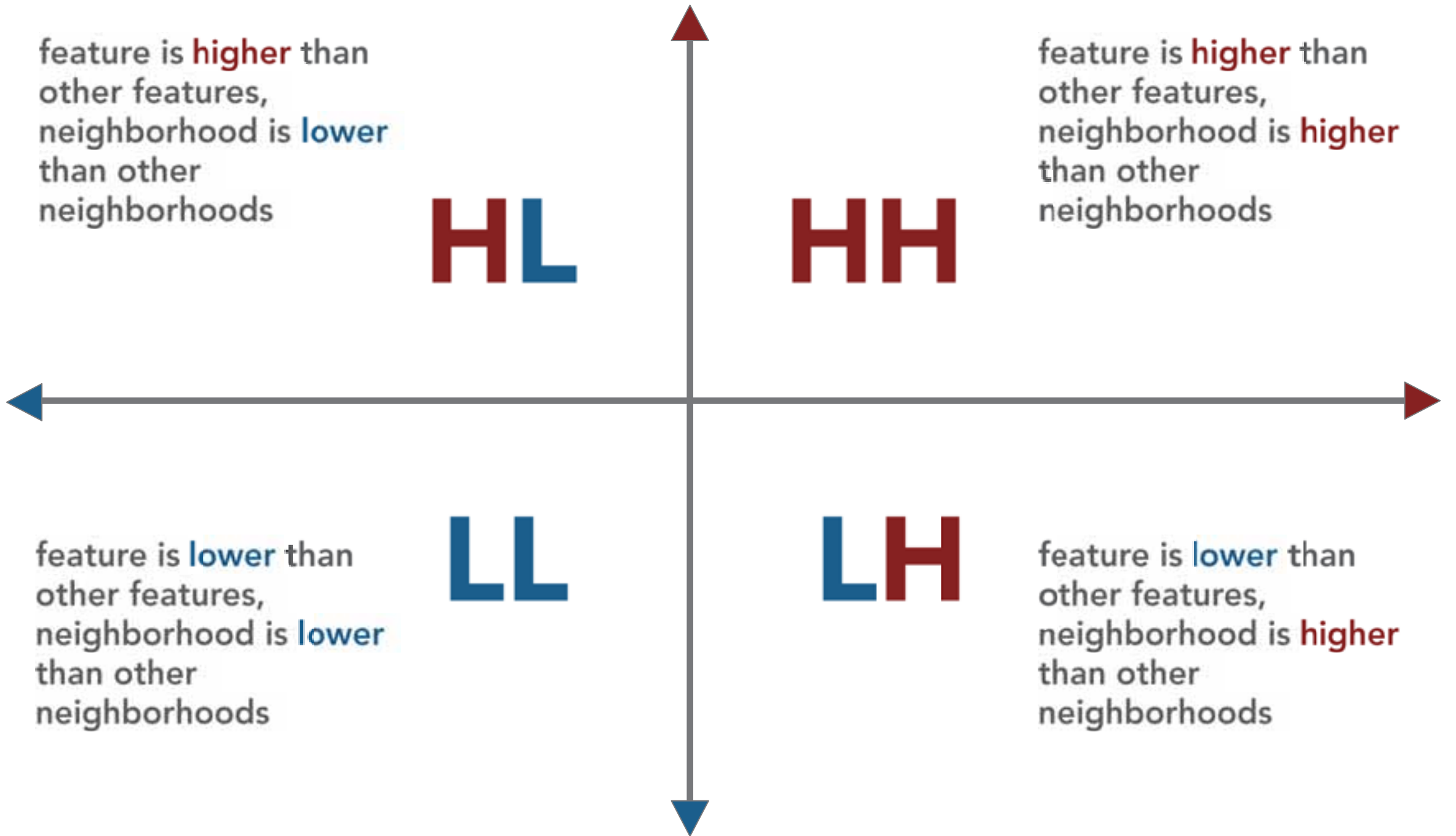
**HH**

feature is **lower** than other features, neighborhood is **lower** than other neighborhoods

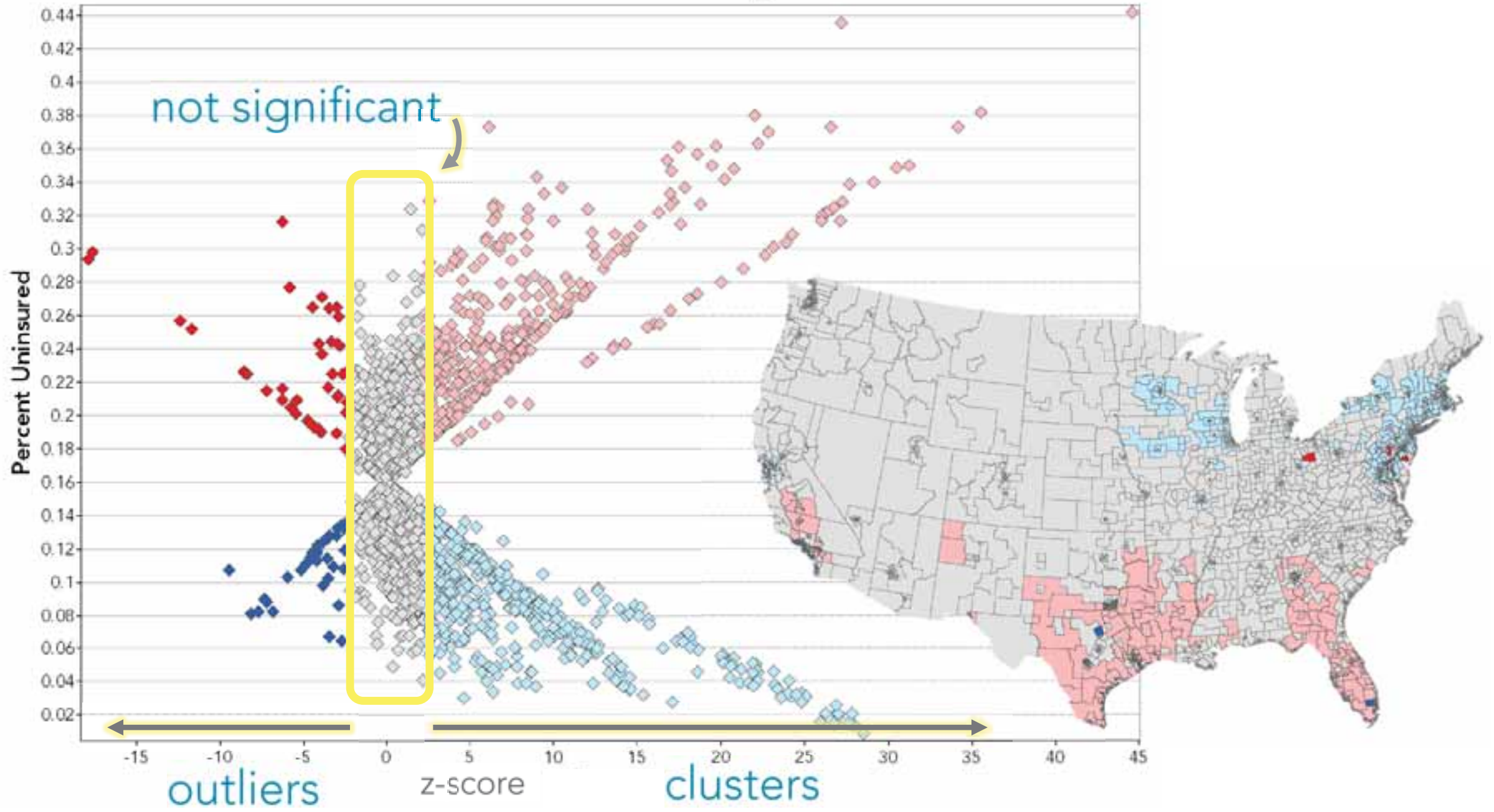
**LL**

feature is **lower** than other features, neighborhood is **higher** than other neighborhoods

**LH**



# Cluster and Outlier Analysis





demo



# Considerations

Must have at least  
**30 features**

Each feature has a  
**value**

There is

**variance**

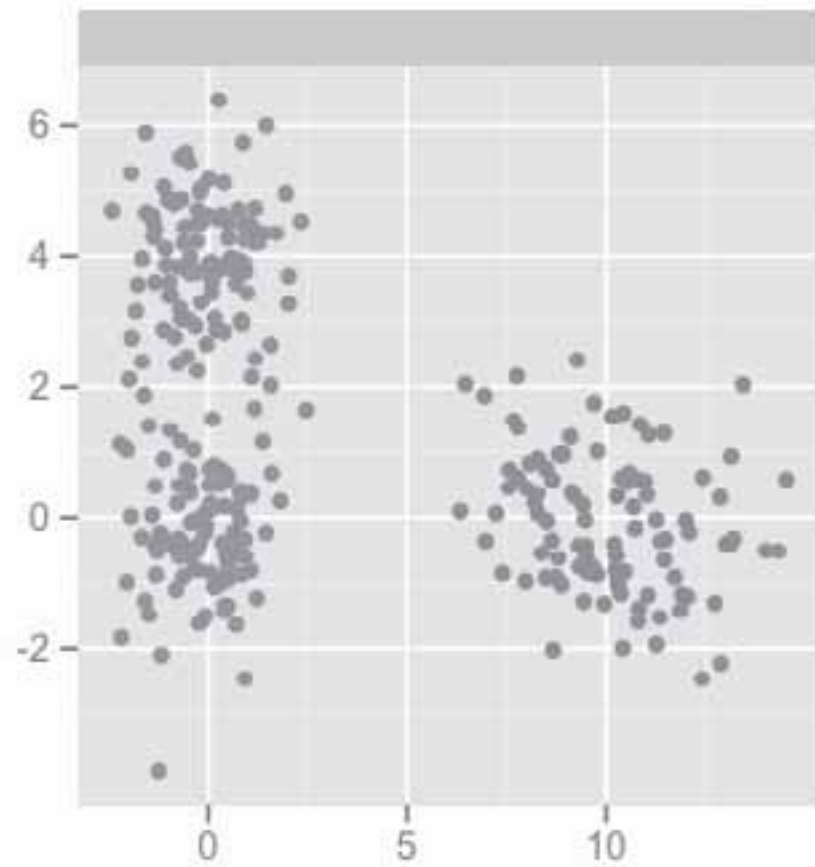
between values

# Grouping Analysis





# K Means



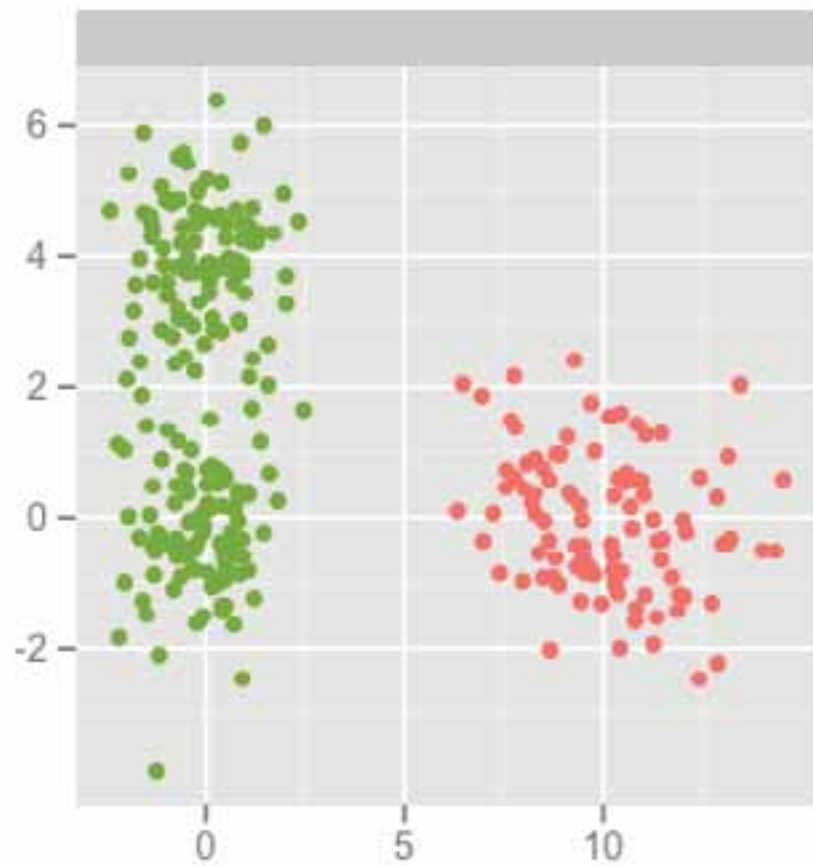
# K Means

2 groups

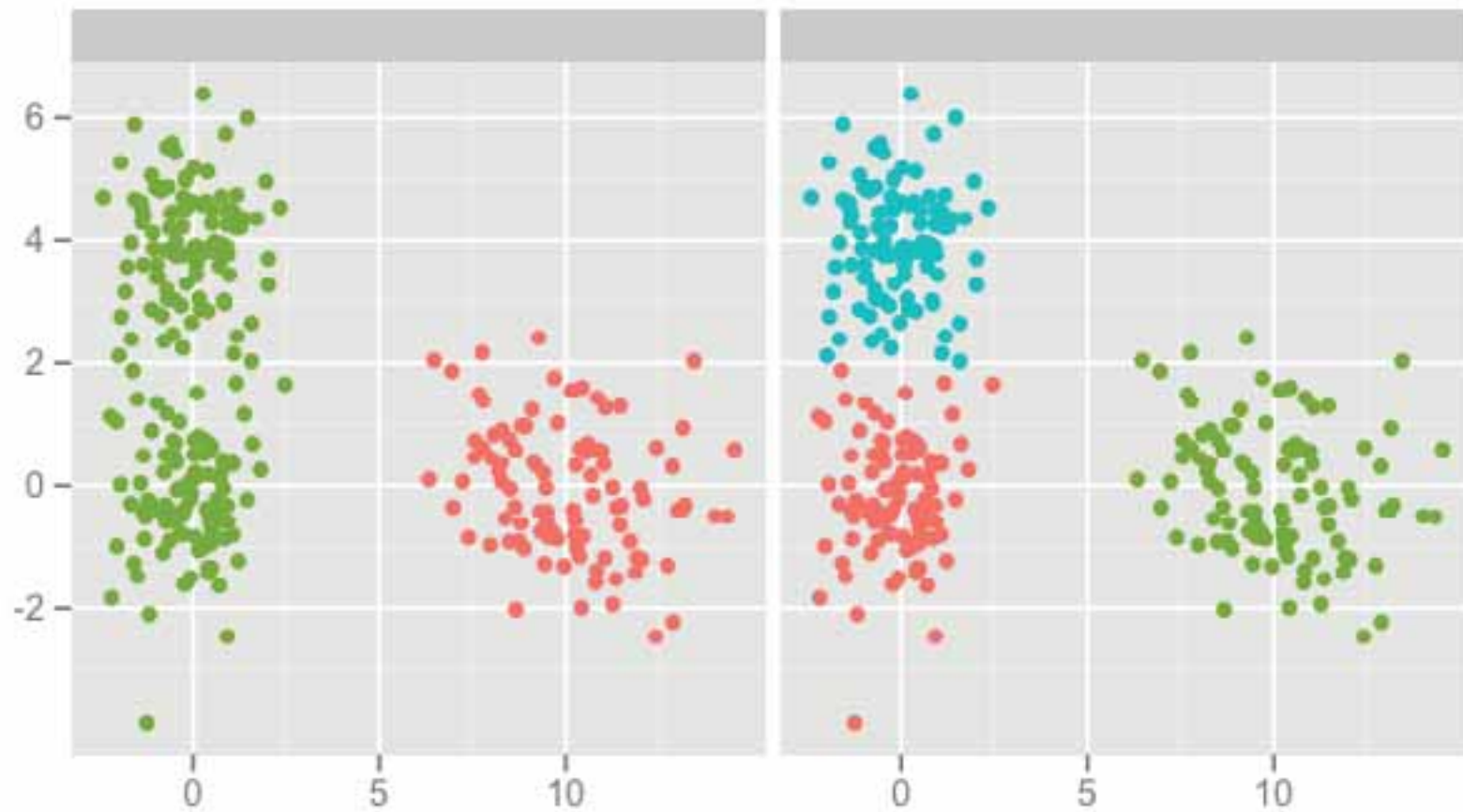


# K Means

2 groups

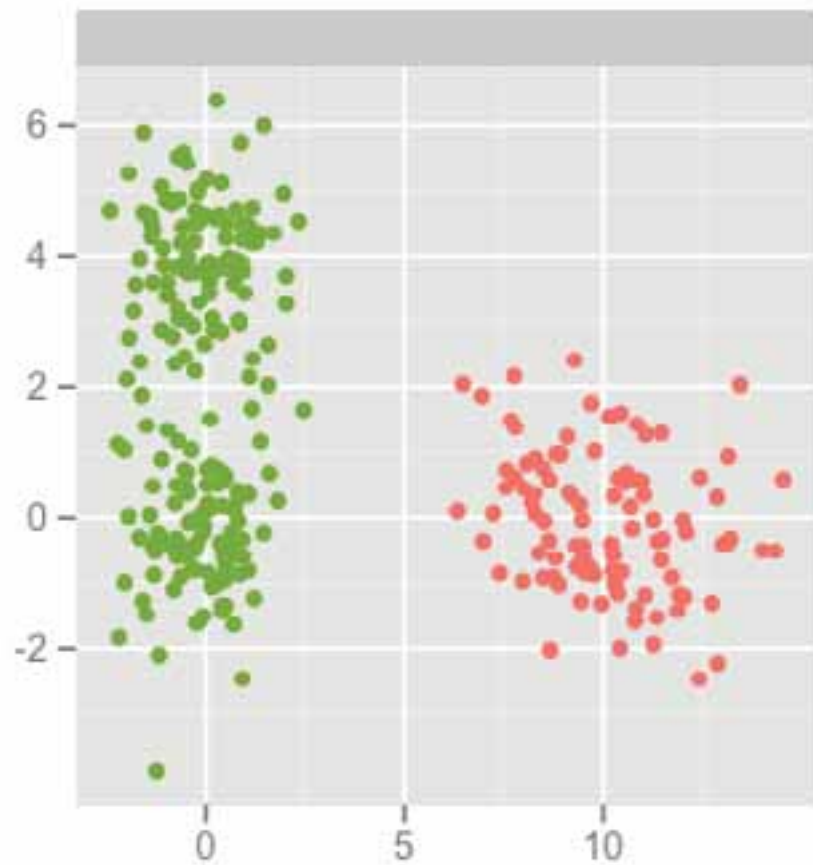


3 groups



# K Means

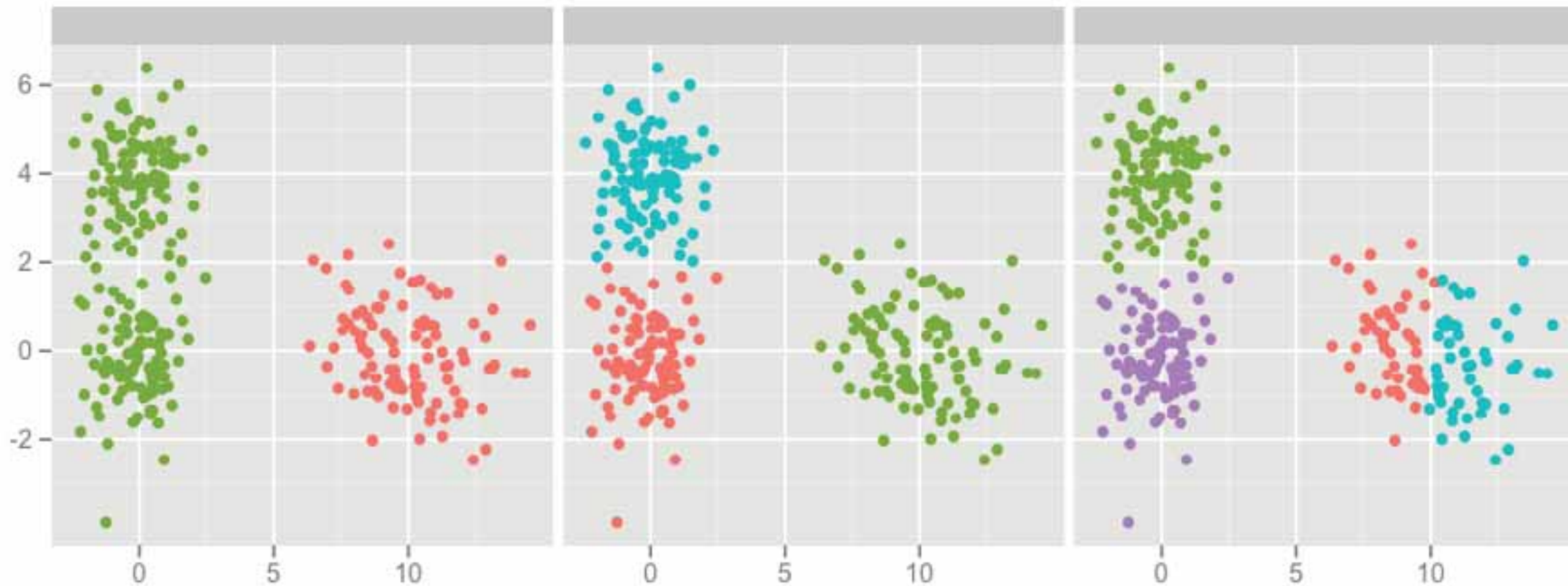
2 groups



3 groups

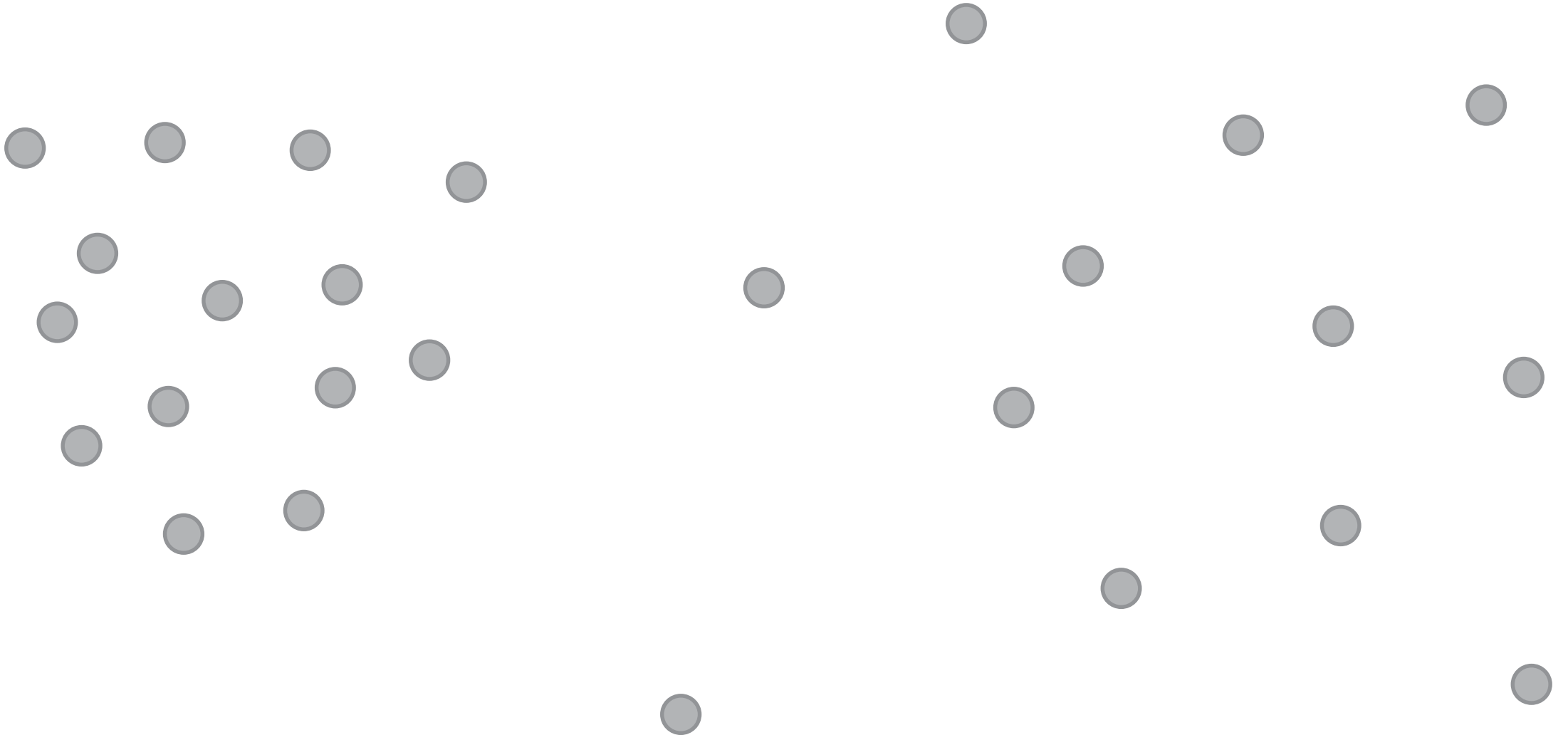


4 groups

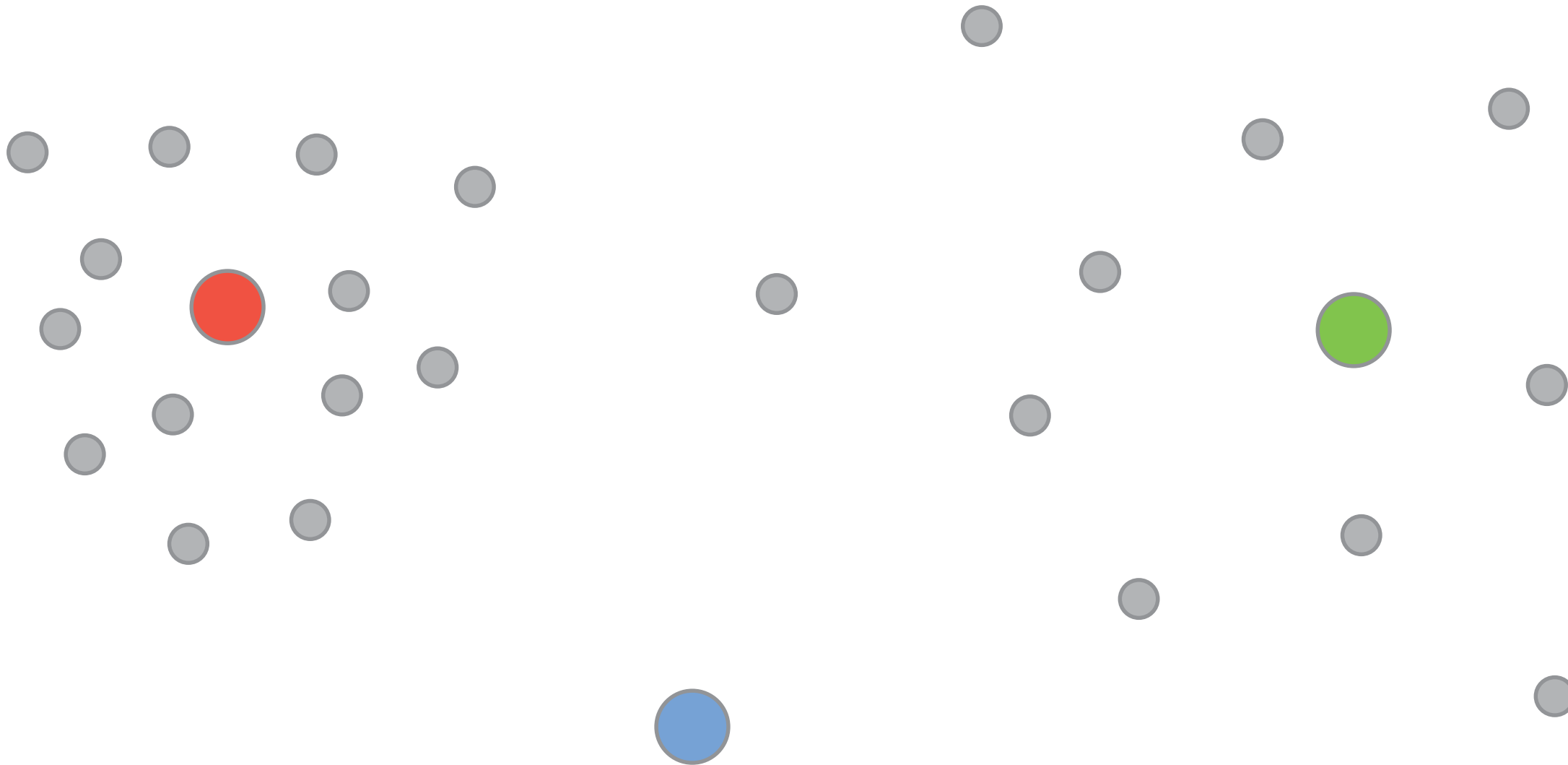




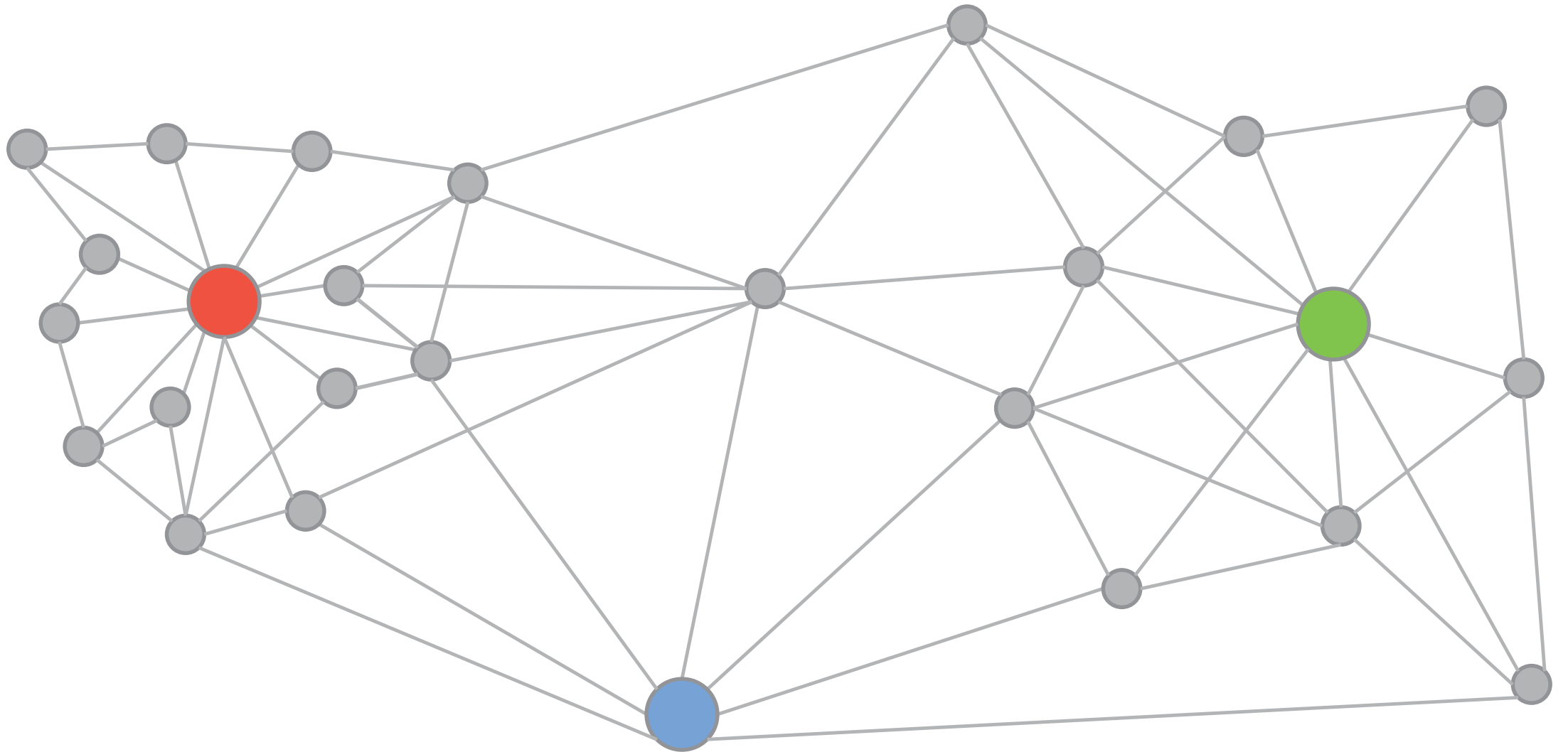
# Minimum Spanning Tree



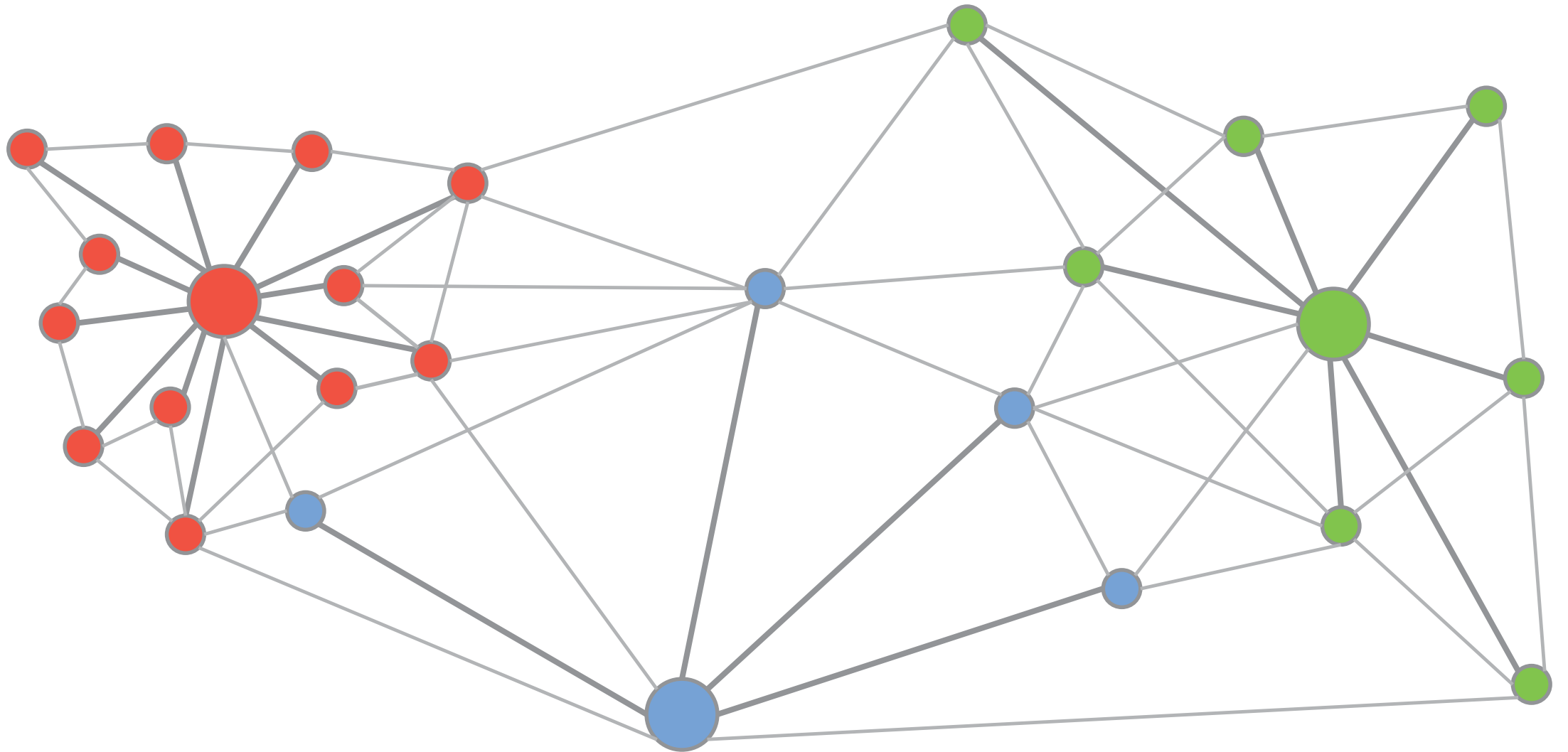
# Minimum Spanning Tree



# Minimum Spanning Tree

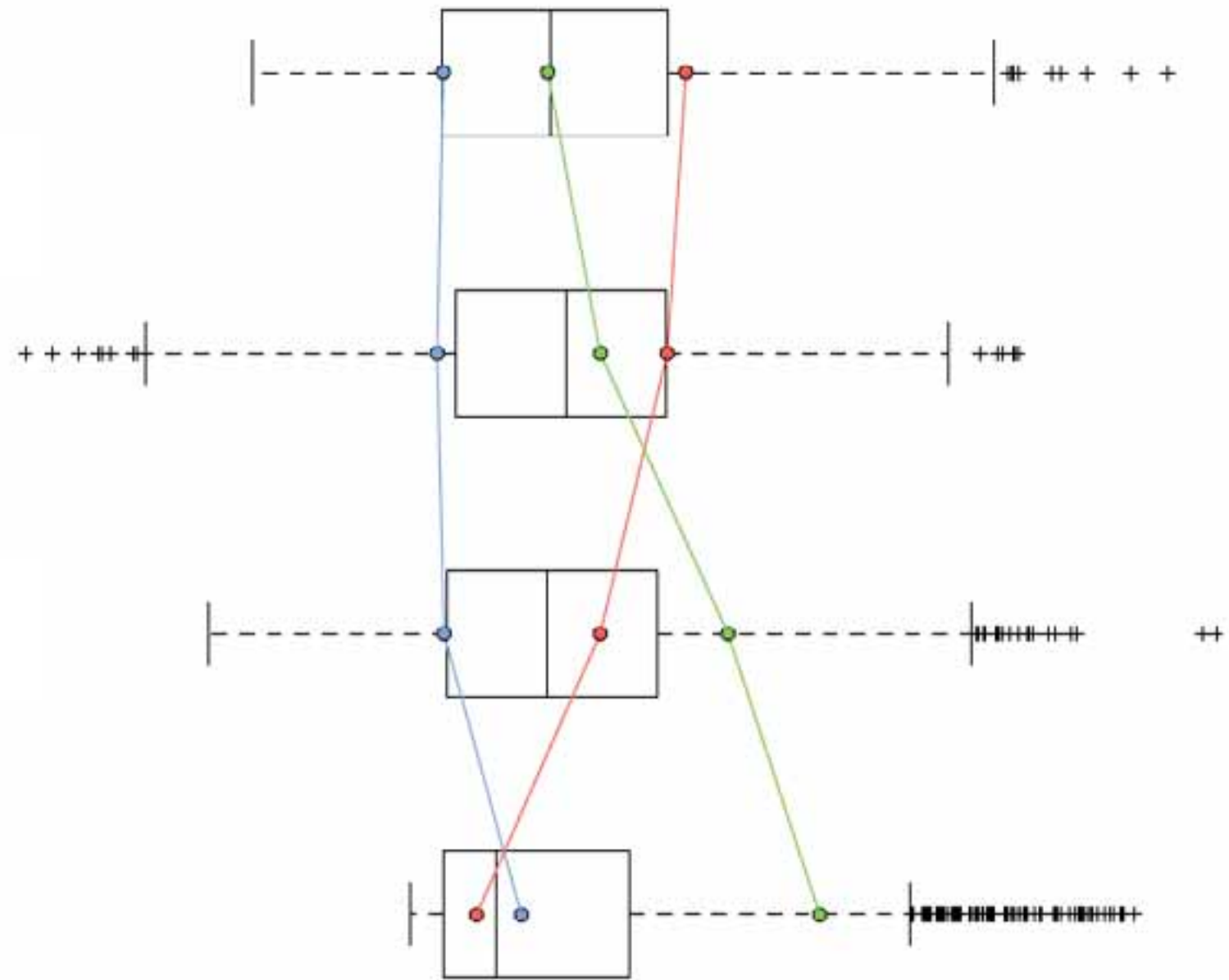


# Minimum Spanning Tree



**МАТН!**

interpret  
results  
through  
box plots





demo



# Similarity Search



**potential  
store  
locations**



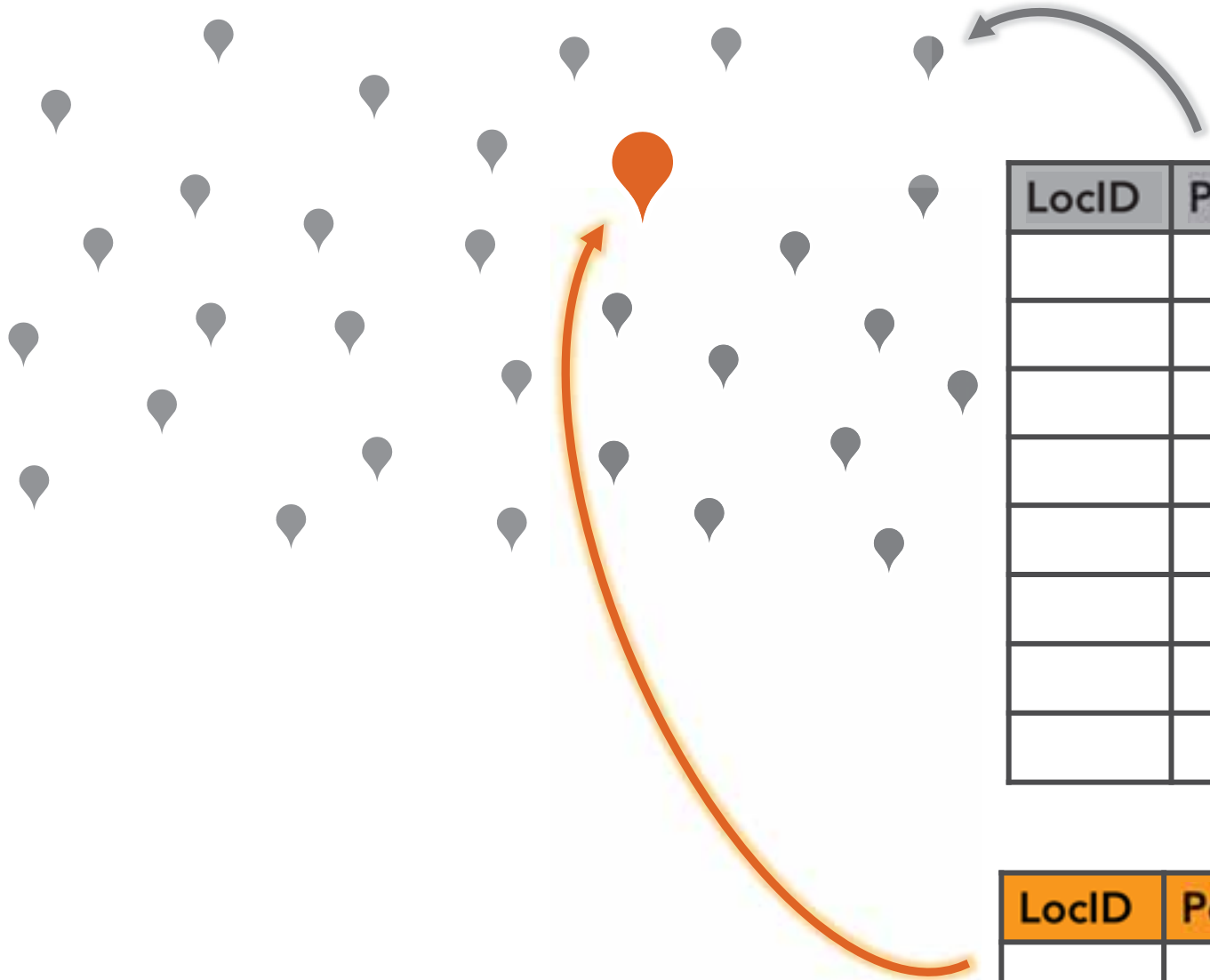
high  
performing  
store

potential  
store  
locations



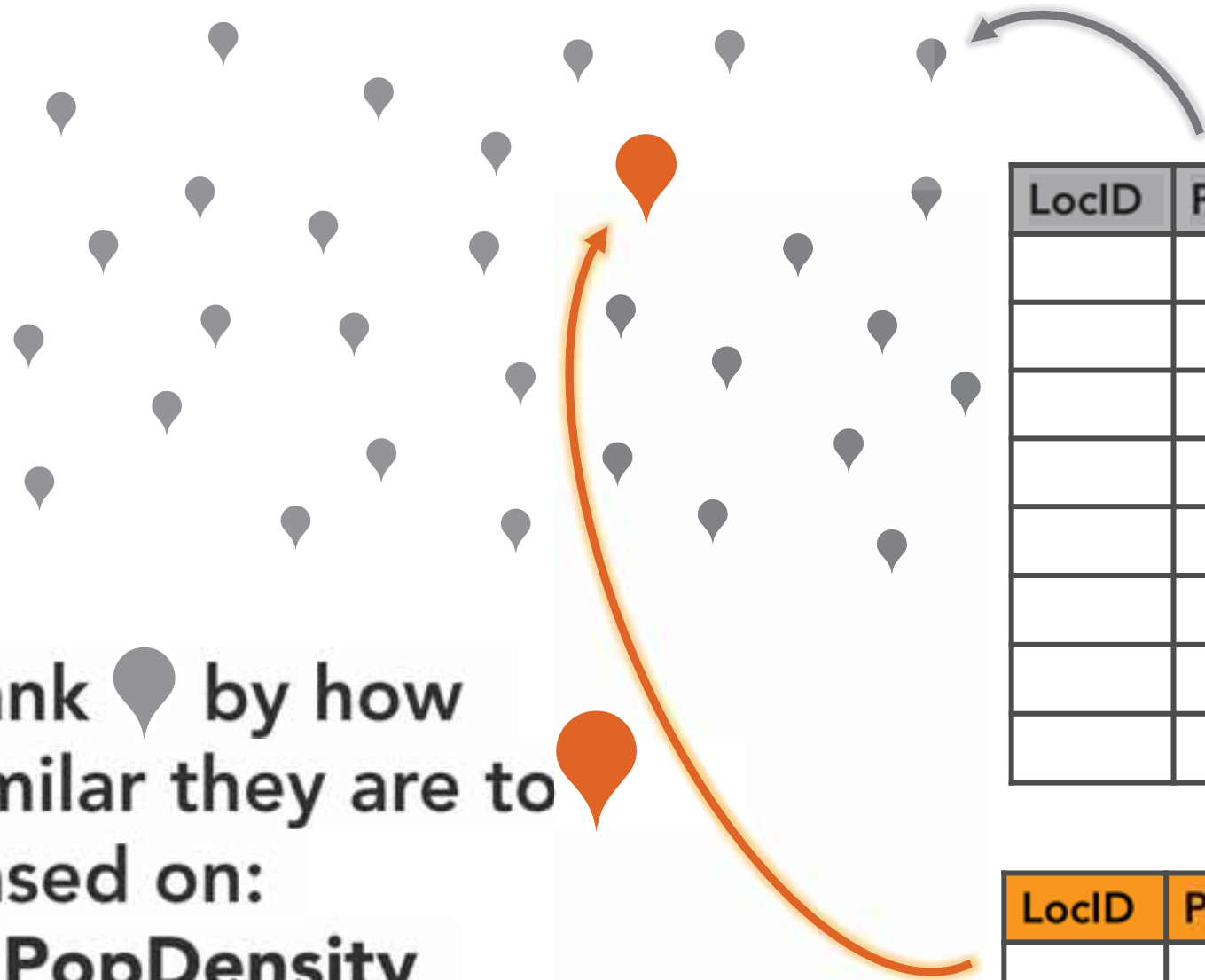






LocID	PopDensity	AvgIncome	DistToCompetition

LocID	PopDensity	AvgIncome	DistToCompetition



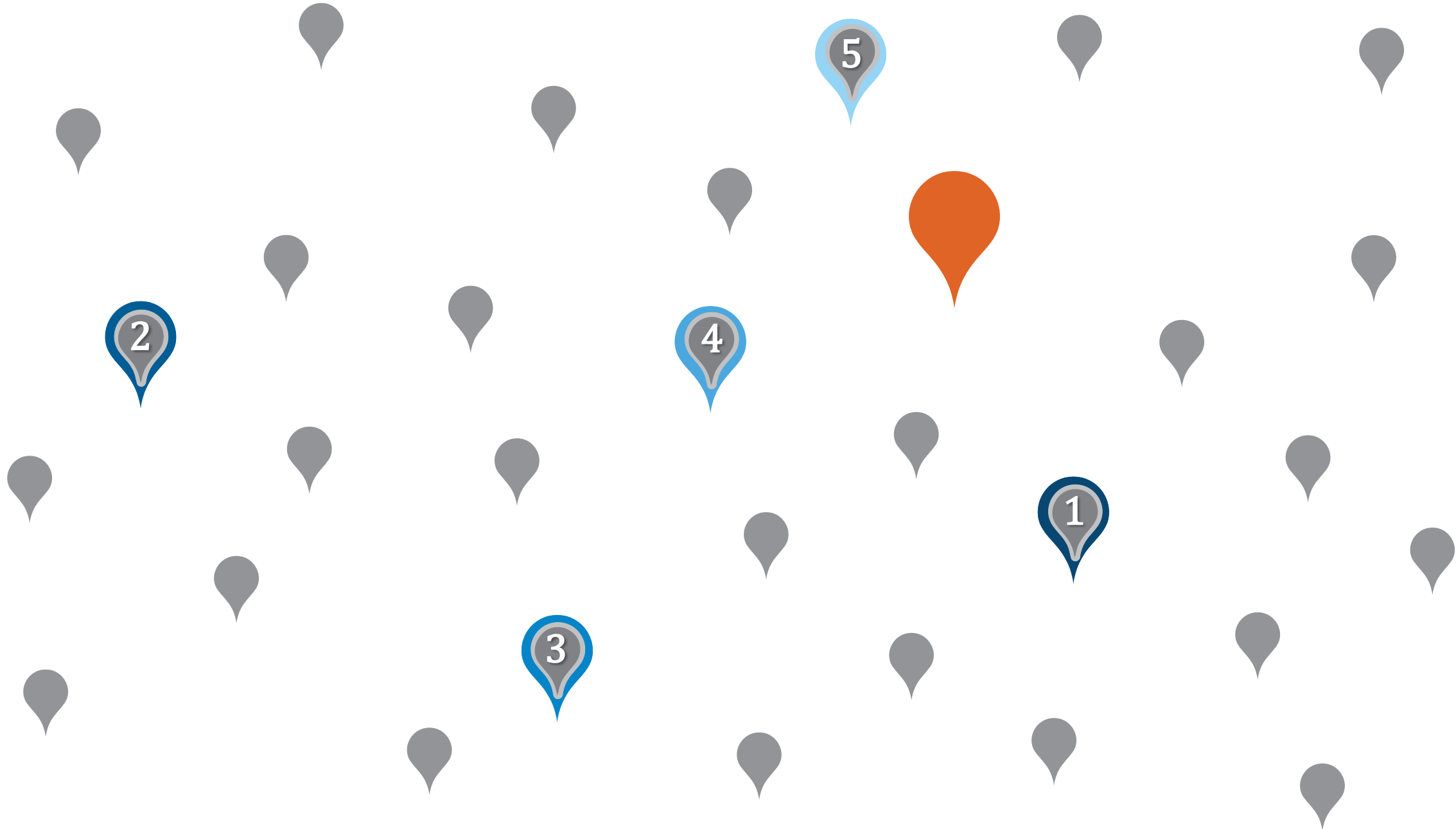
Rank 📍 by how similar they are to 📍 based on:

- **PopDensity**
- **AvIncome**
- **DistToCompetition**

LocID	PopDensity	AvIncome	DistToCompetition

LocID	PopDensity	AvIncome	DistToCompetition





# Input Feature(s) to Match



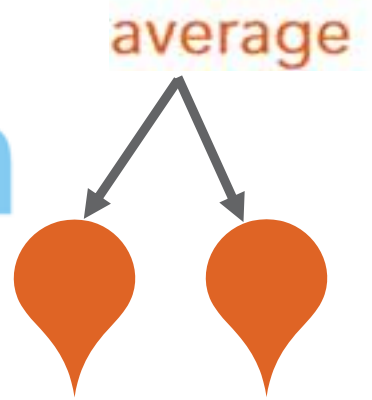
# Candidate Features



# Attributes of Interest

PopDensity	AvIncome	DistToCompetition

# Input Feature(s) to Match



# Candidate Features



# Attributes of Interest

PopDensity	AvIncome	DistToCompetition



# **3 Match Methods**

# 3 Match Methods

**Attribute Values**

# 3 Match Methods

Attribute Values

Ranked Attribute Values

# 3 Match Methods

Attribute Values

Ranked Attribute Values

Attribute Profiles

# Attribute Values

# Attribute Values



standardize  
attributes

Z-transform:

$$(x - \bar{x}) / SD$$



# Attribute Values

standardize  
attributes



Population = 14,159

% Uninsured = .26

Distance (km) = 535.89

# Attribute Values

standardize  
attributes

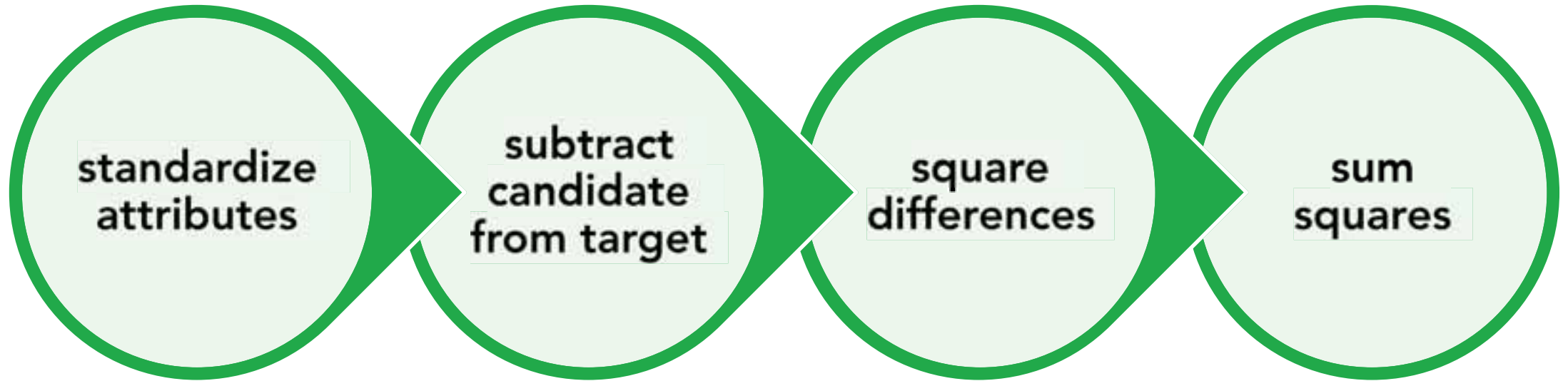


Population =  $-.7932$

% Uninsured =  $3.8462$

Distance (km) =  $.6433$

# Attribute Values



# Ranked Attribute Values

# Ranked Attribute Values



# Ranked Attribute Values



9.5

8.8

8.3

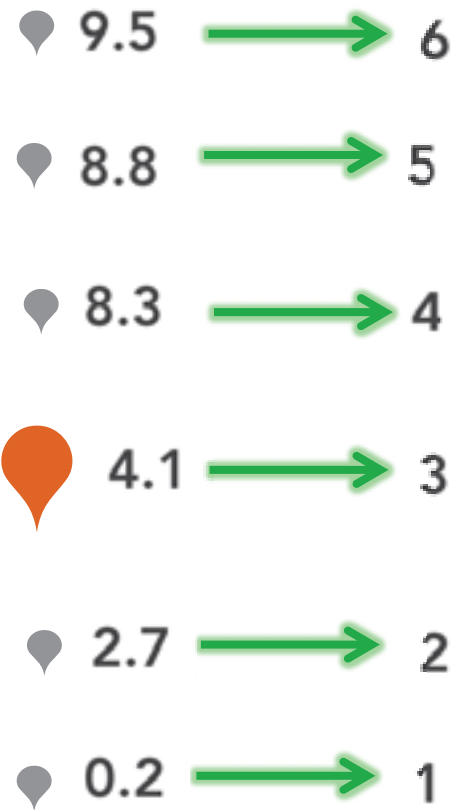
4.1

2.7

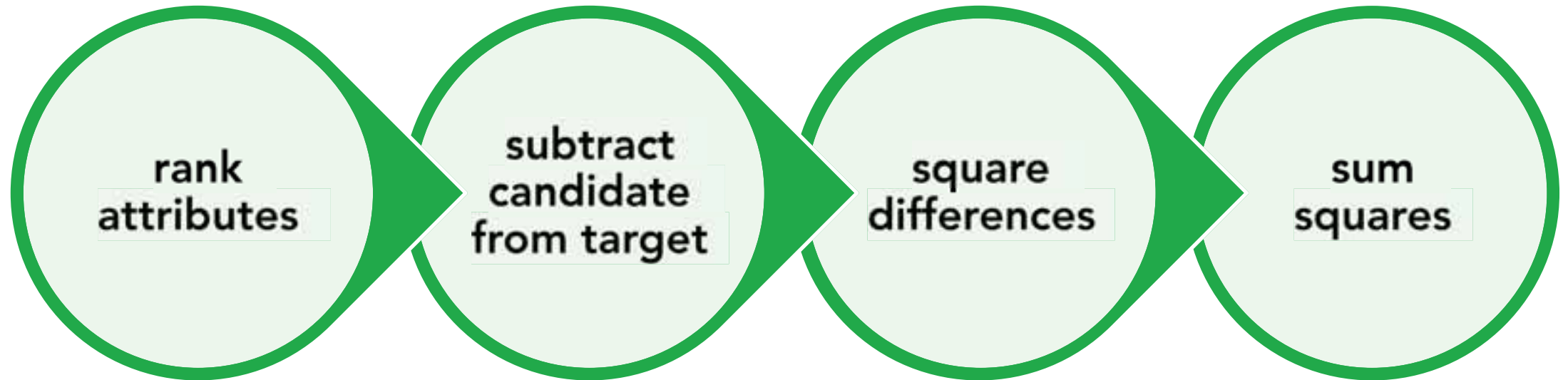
0.2



# Ranked Attribute Values

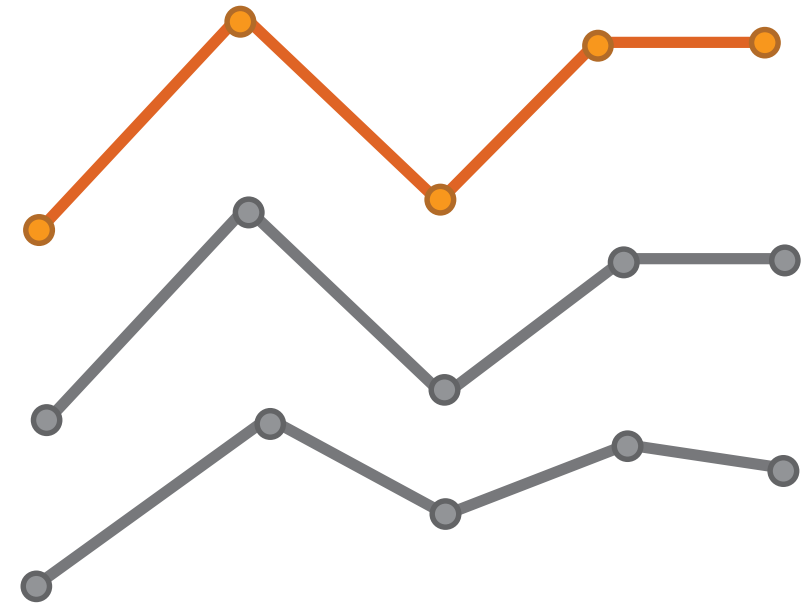
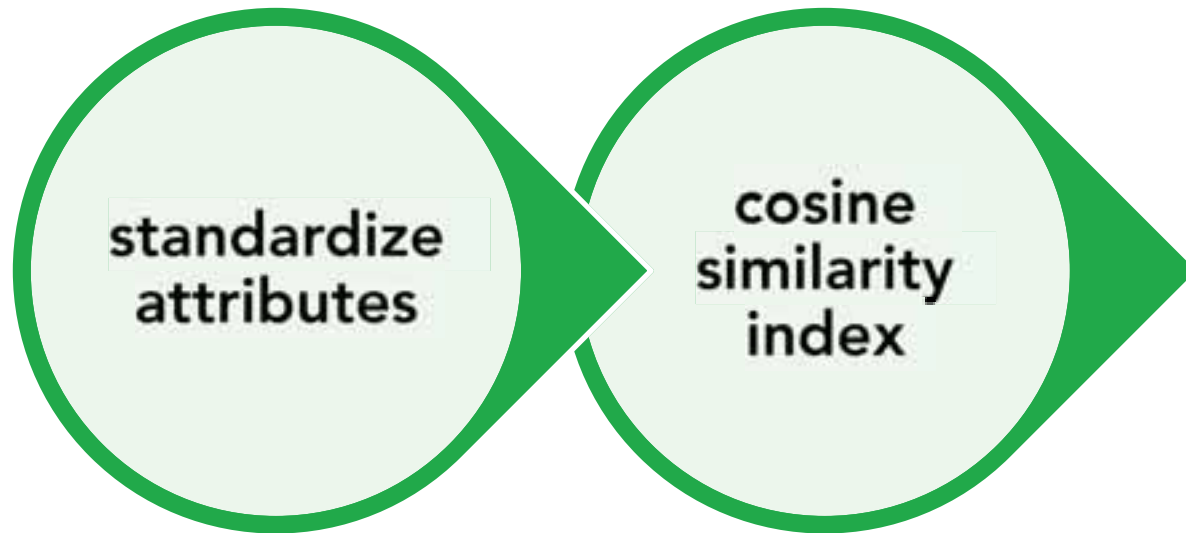


# Ranked Attribute Values



# Attribute Profiles

# Attribute Profiles



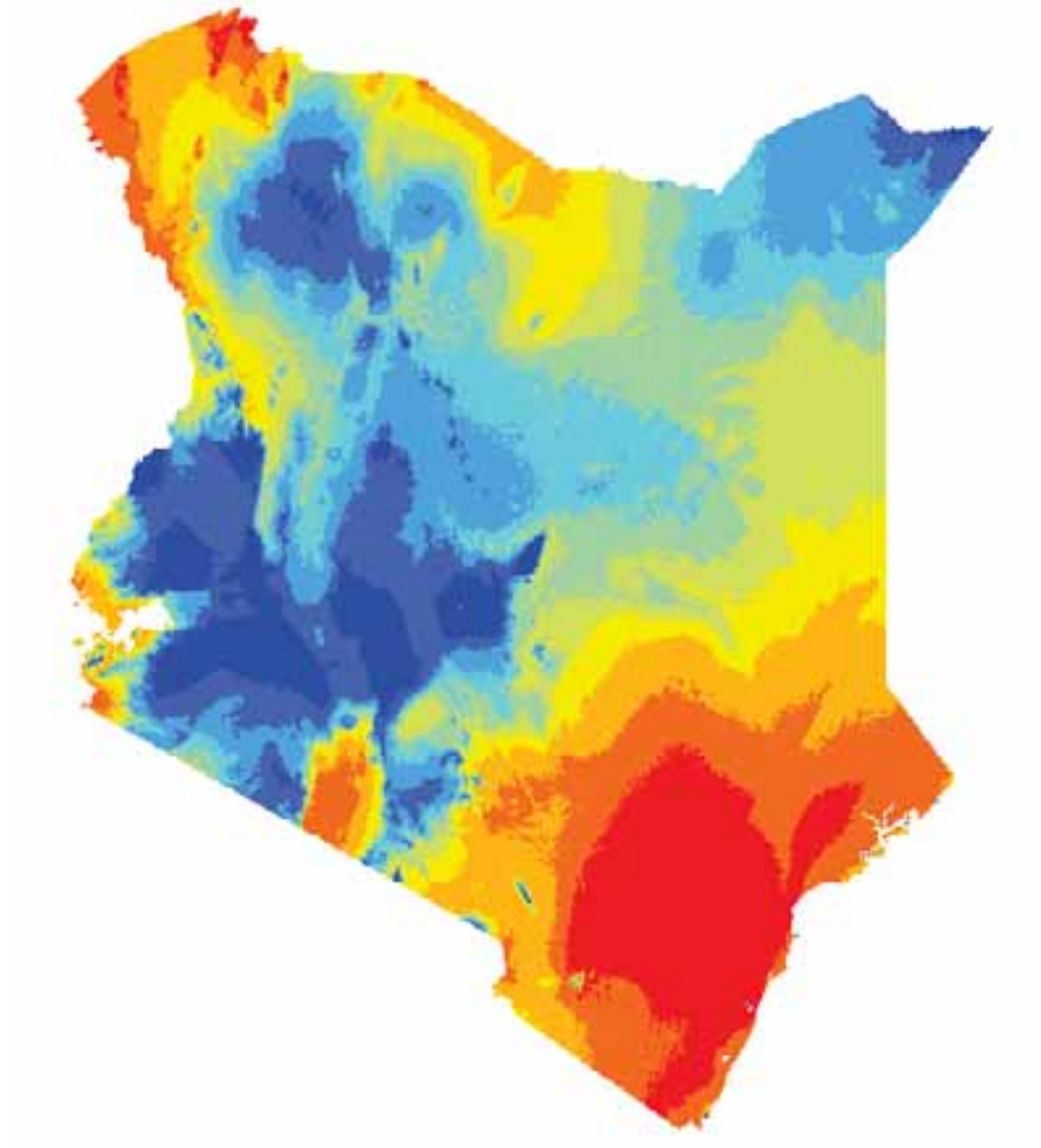
$$\text{cosine similarity index} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \sqrt{\sum_{i=1}^n (B_i)^2}}$$

\* Must have at least 2 attributes of interest

demo



# Dengue Fever Risk in Kenya







lbennett@esri.com  
fvale@esri.com

Want to learn more???

[esriurl.com/spatialstats](https://esriurl.com/spatialstats)

