



Spatial Data Mining I: Essentials of Cluster Analysis

Ankita Bakshi

Alberto Nieto

Flora Vale

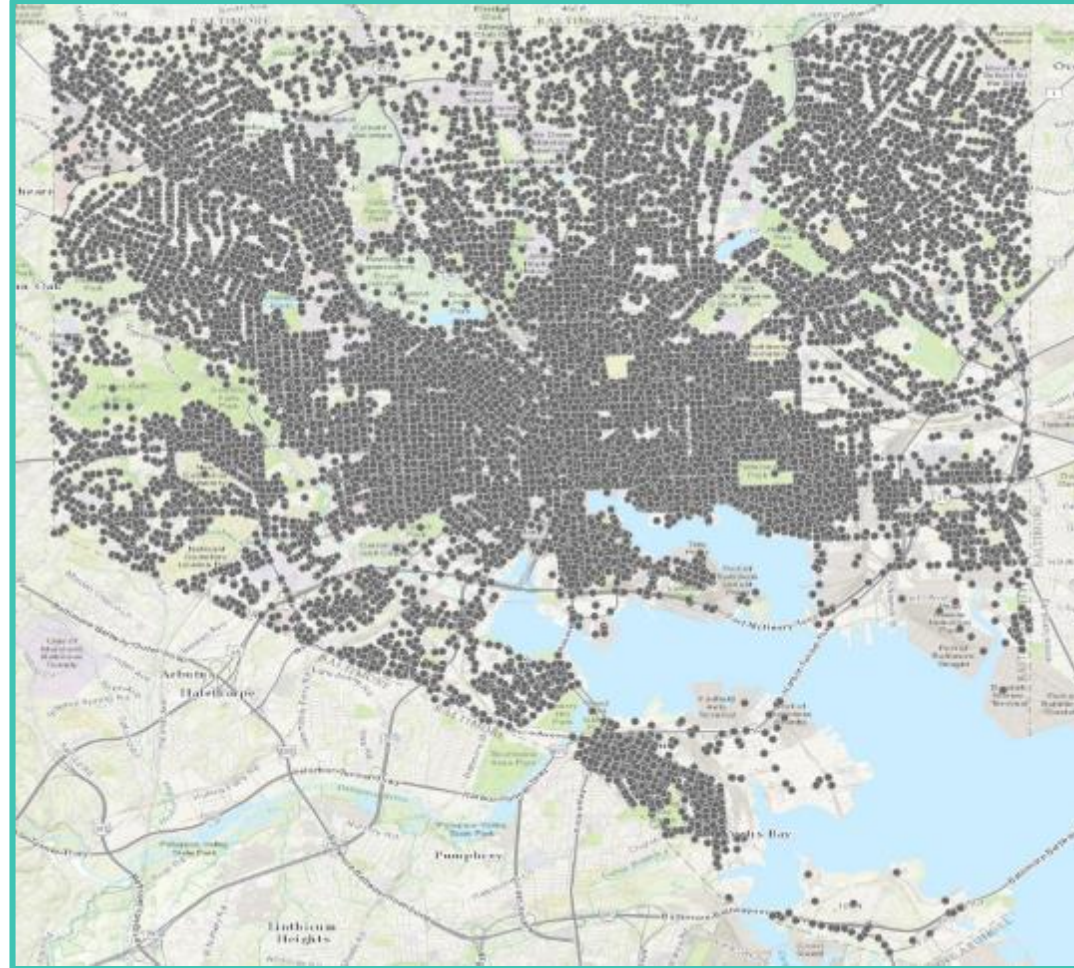
esriurl.com/spatialstats

SEE
WHAT
OTHERS
CAN'T

Subjectivity of Maps

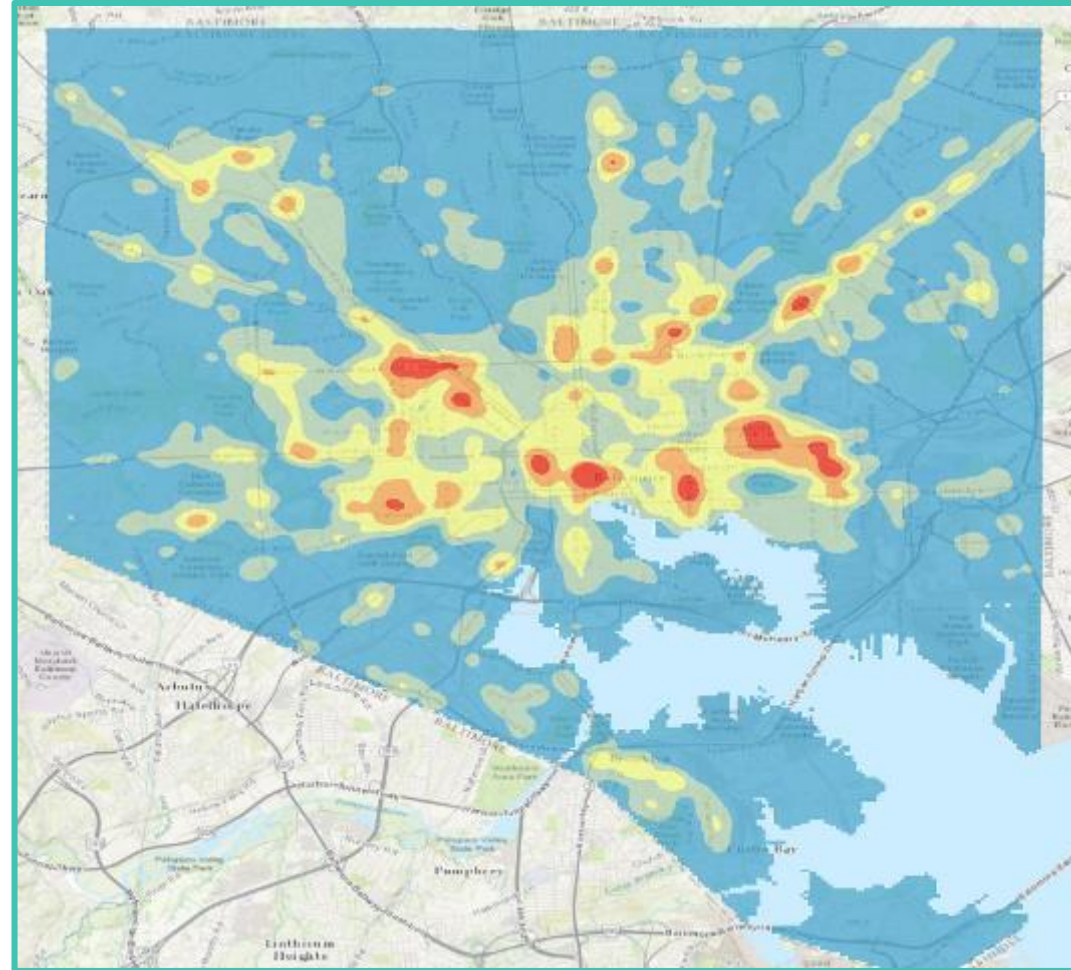
The map as data

High Priority 911 Calls in Baltimore



The map as data

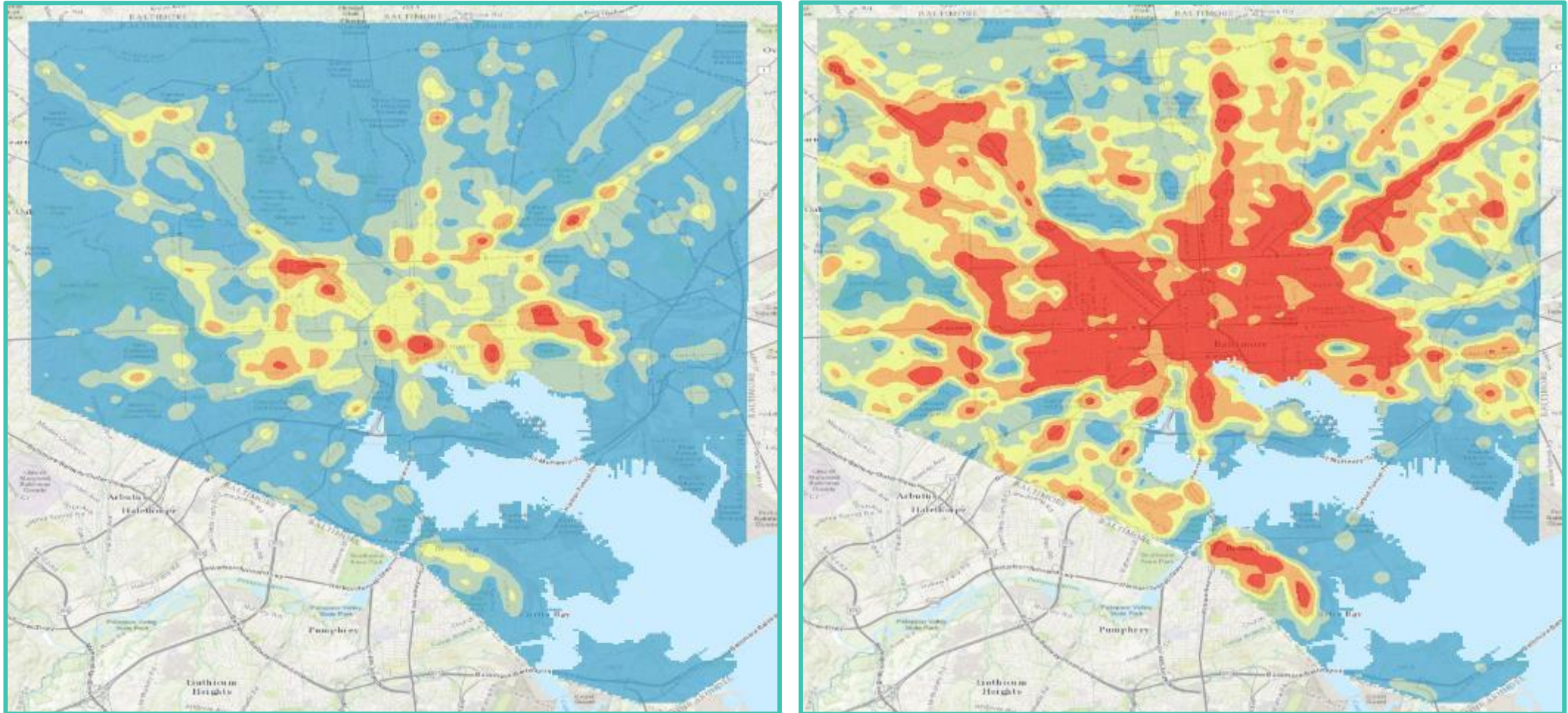
High Priority 911 Calls in Baltimore



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

The map as data

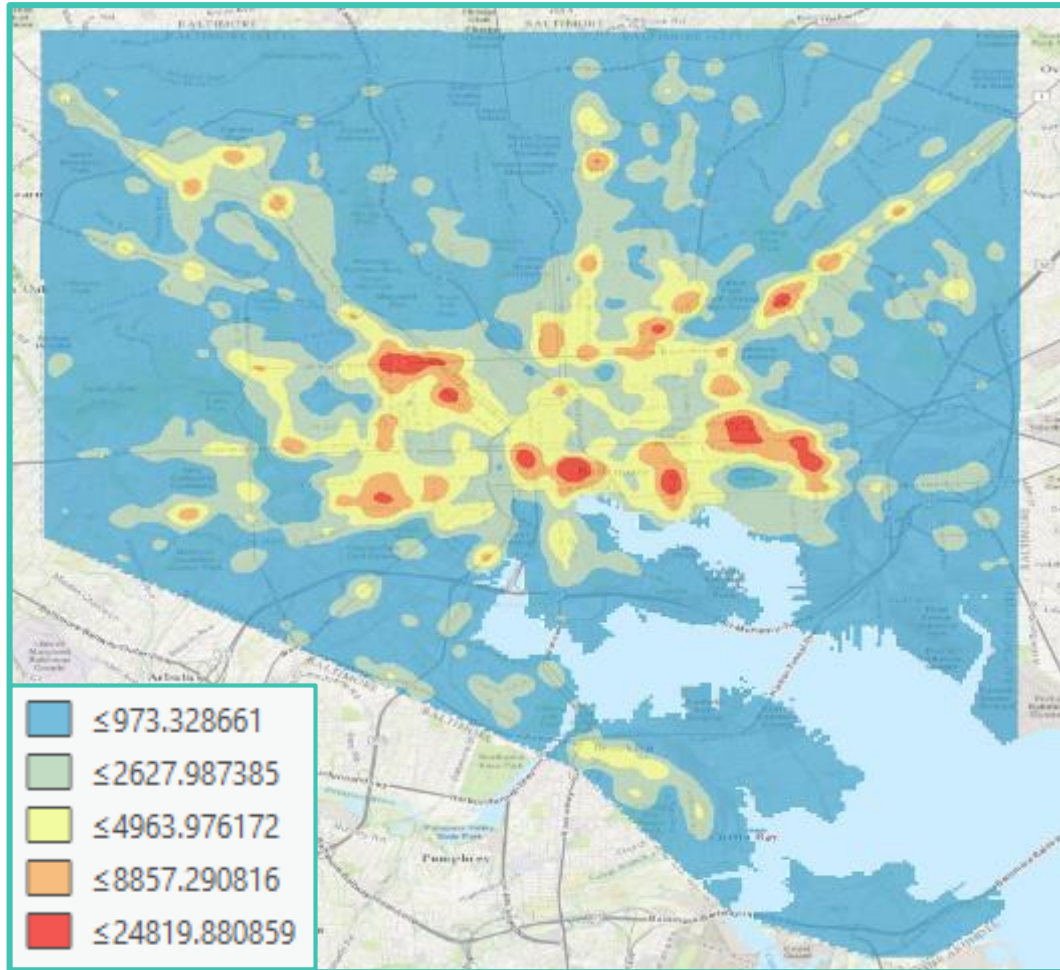
High Priority 911 Calls in Baltimore



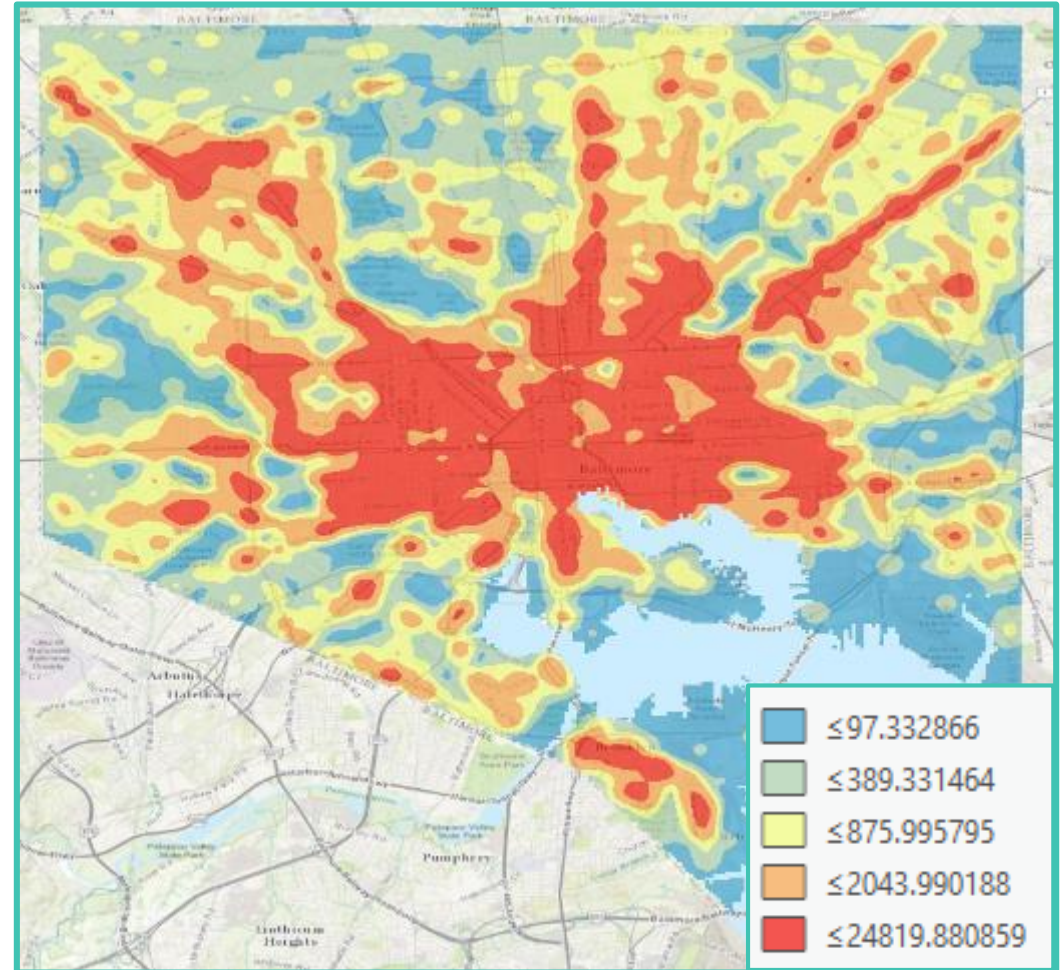
Where are the hot spots? Where is the variation greater?

The **subjectivity** of visual pattern analysis

Natural Breaks



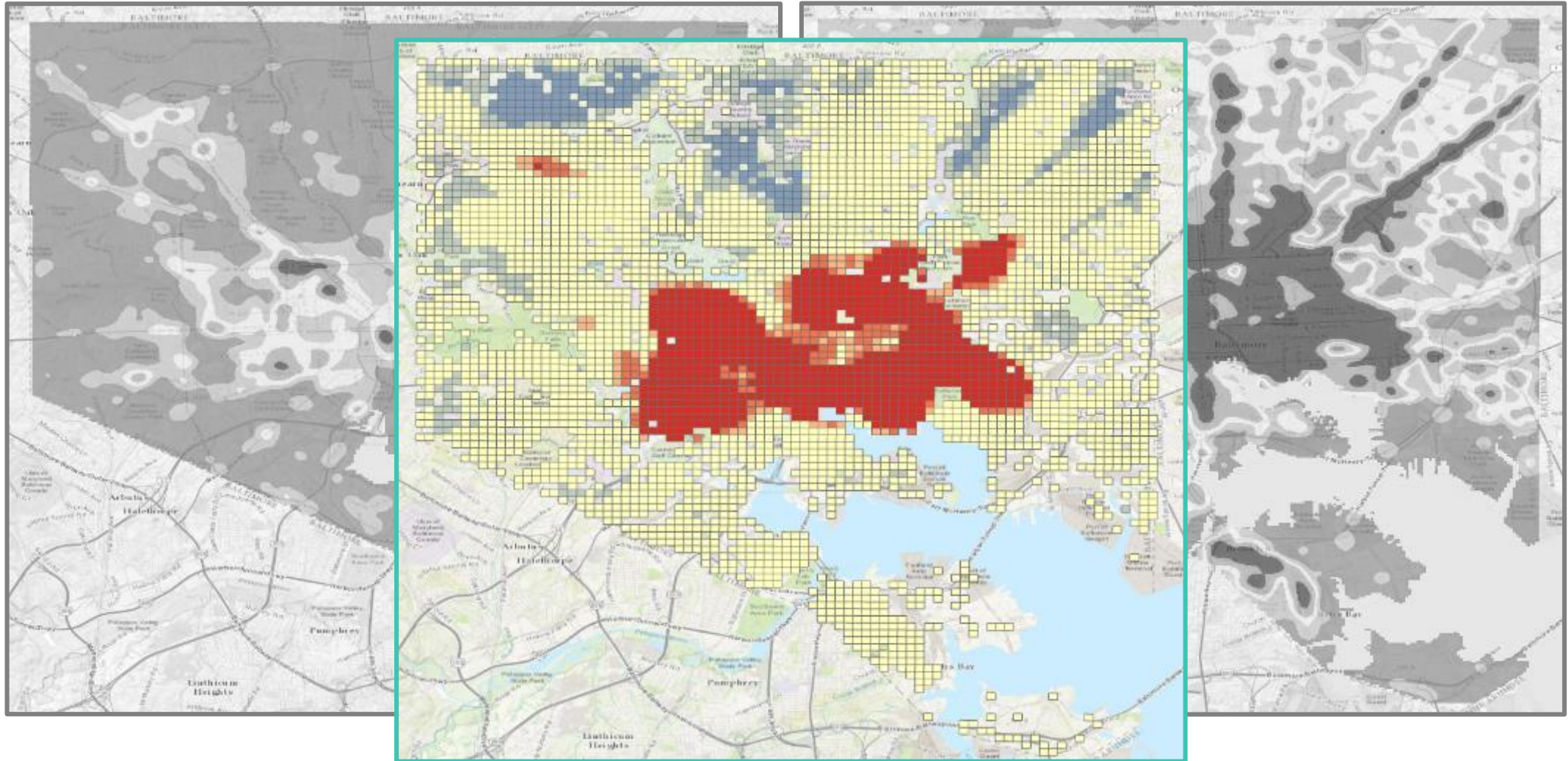
Quantile



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

Minimizing the subjectivity

Turning the map into **information**



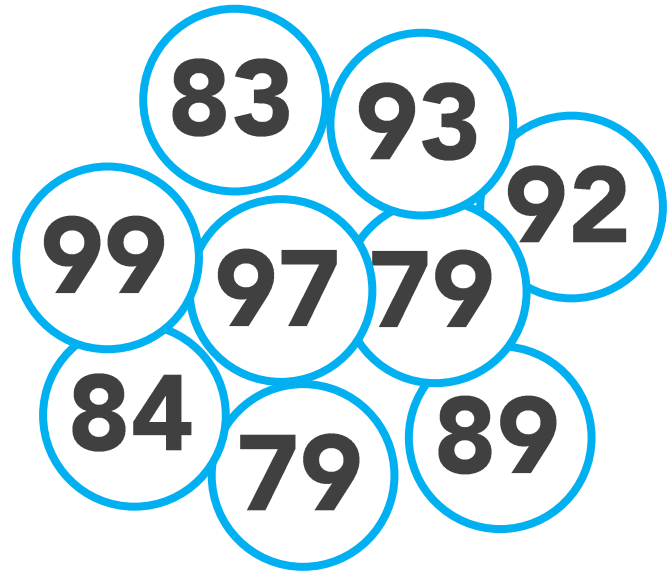
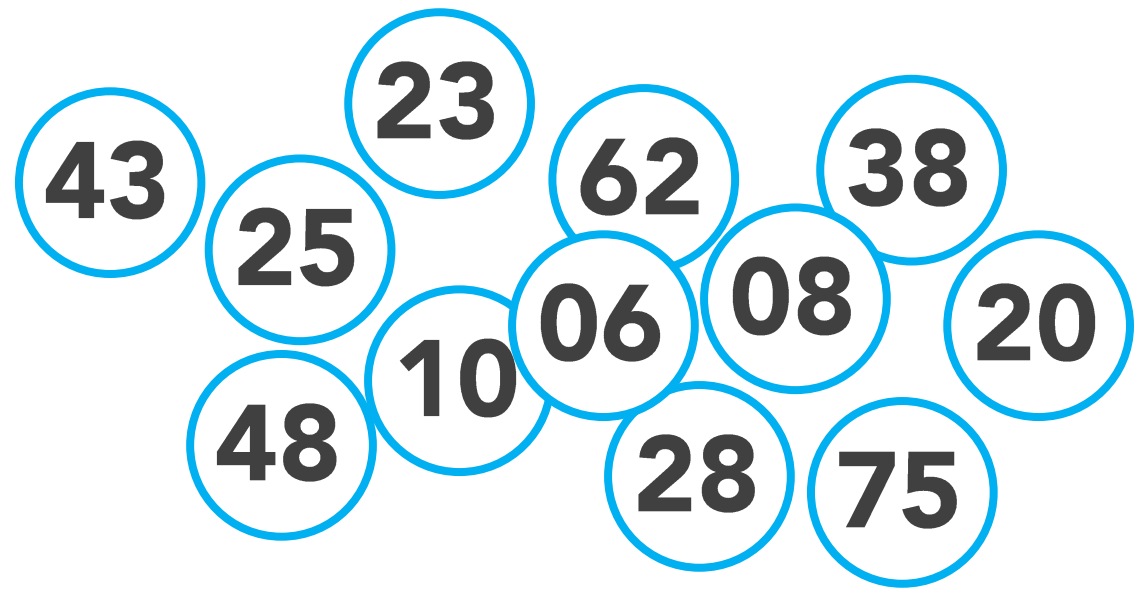
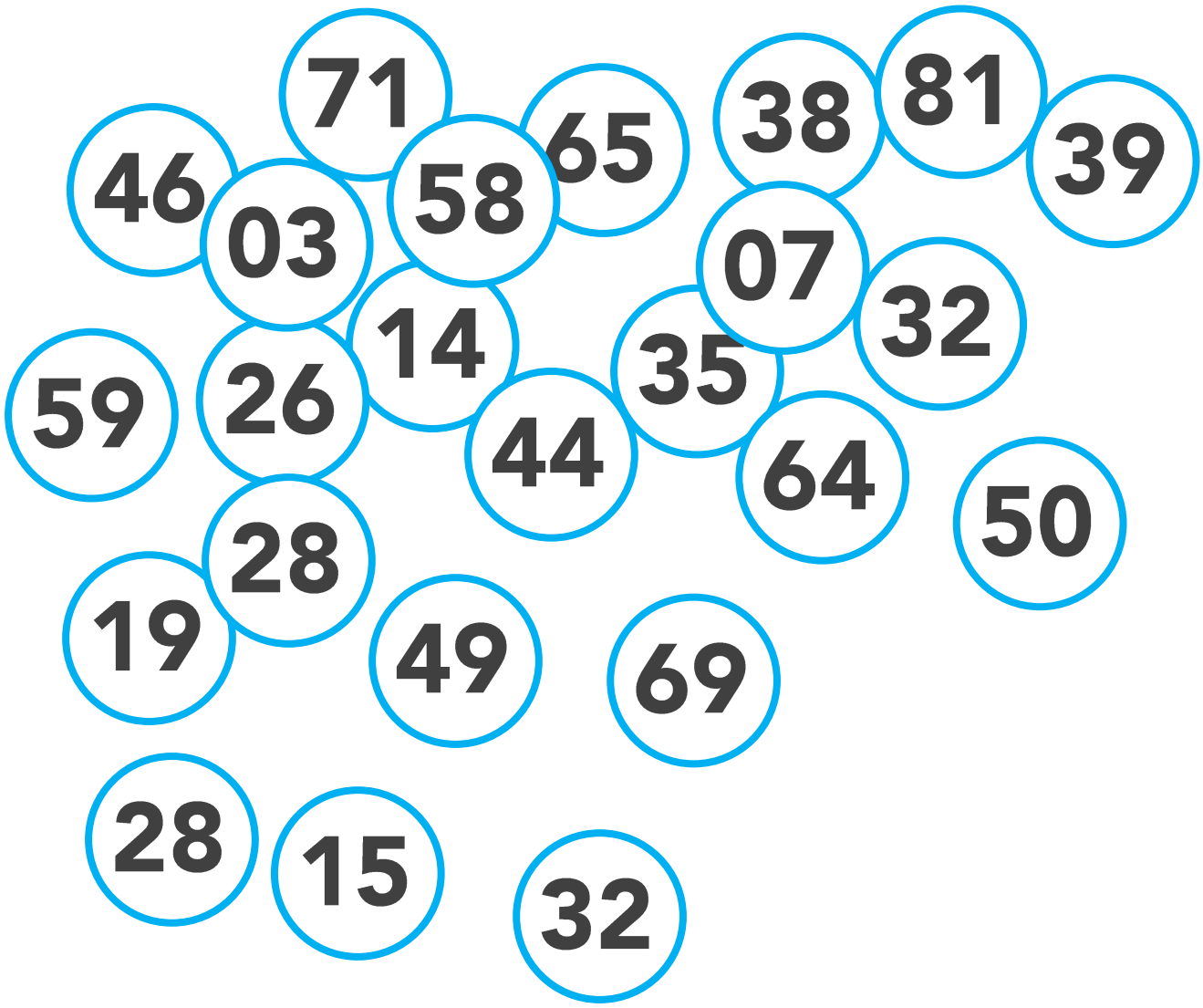
Inferential Statistics

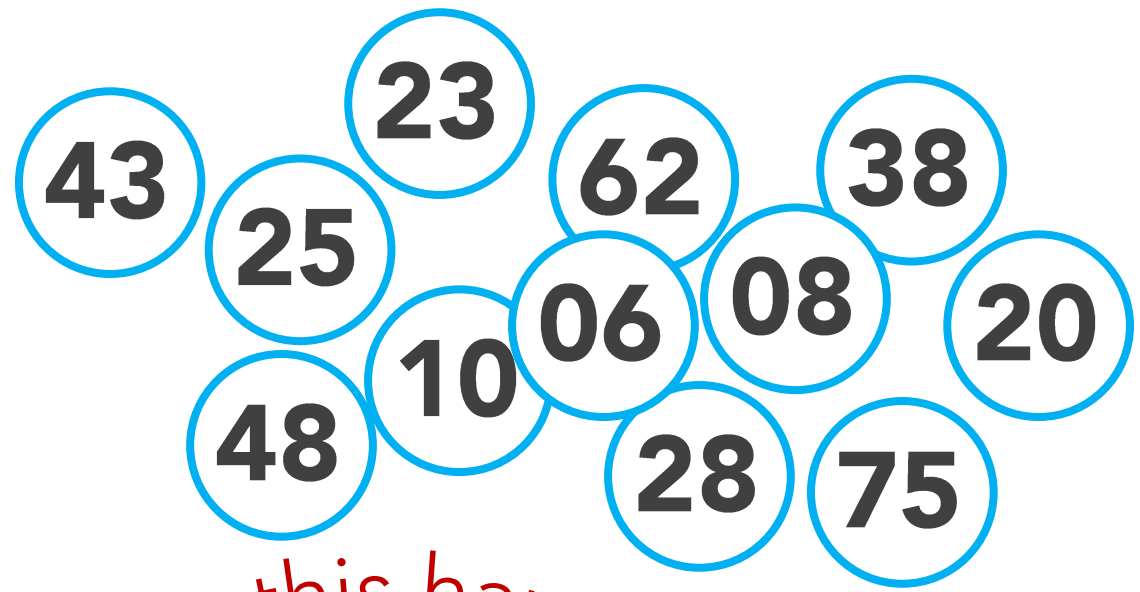
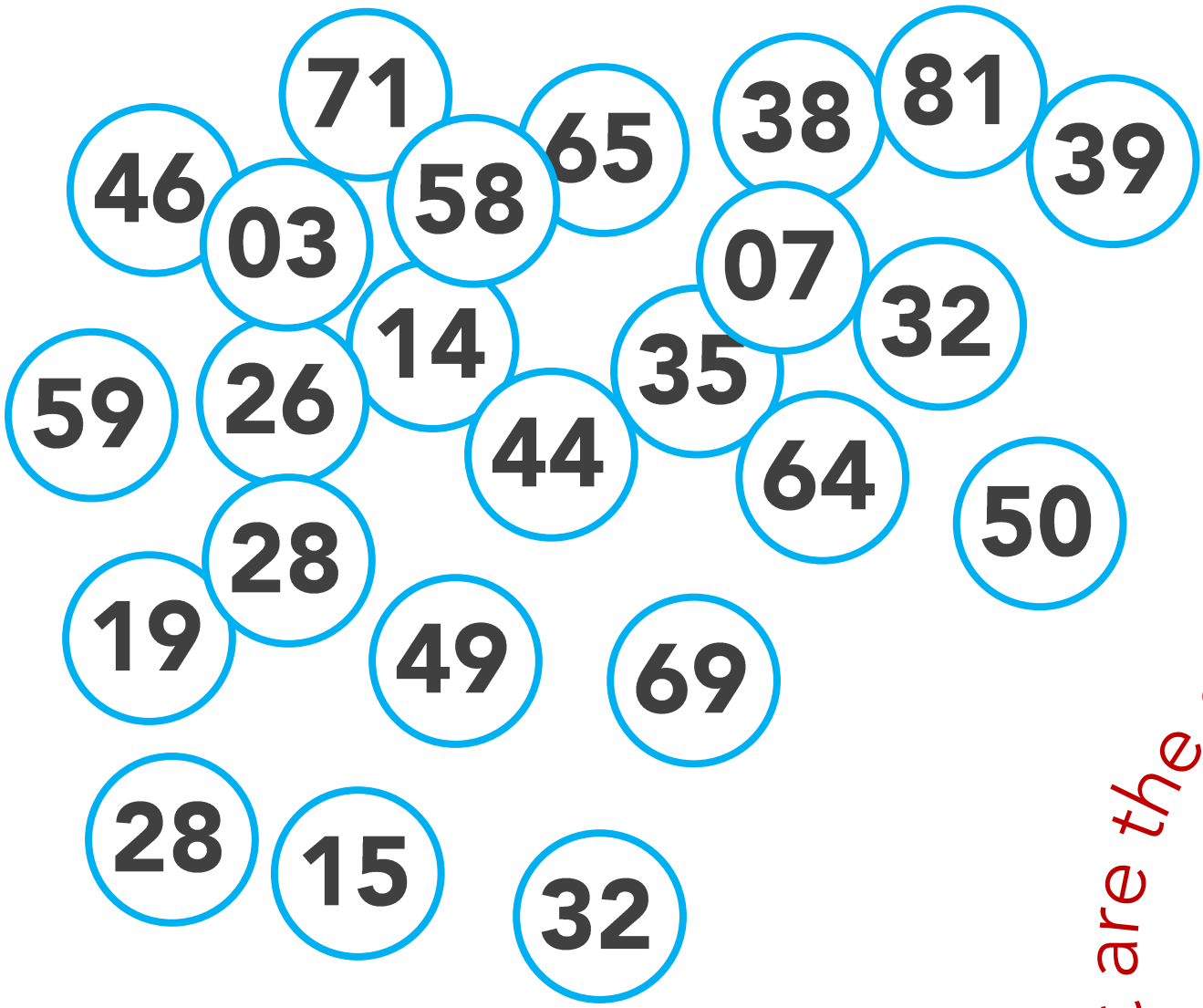


Complete Spatial RANDOMNESS

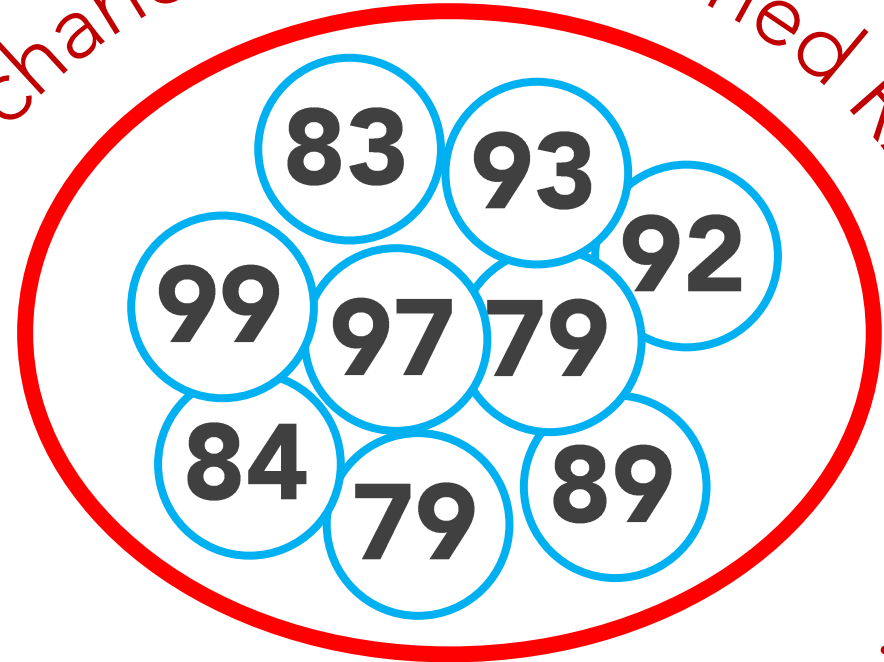
Is there a **PATTERN**?

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	06
28	62	08	99	86	28	03	48	25





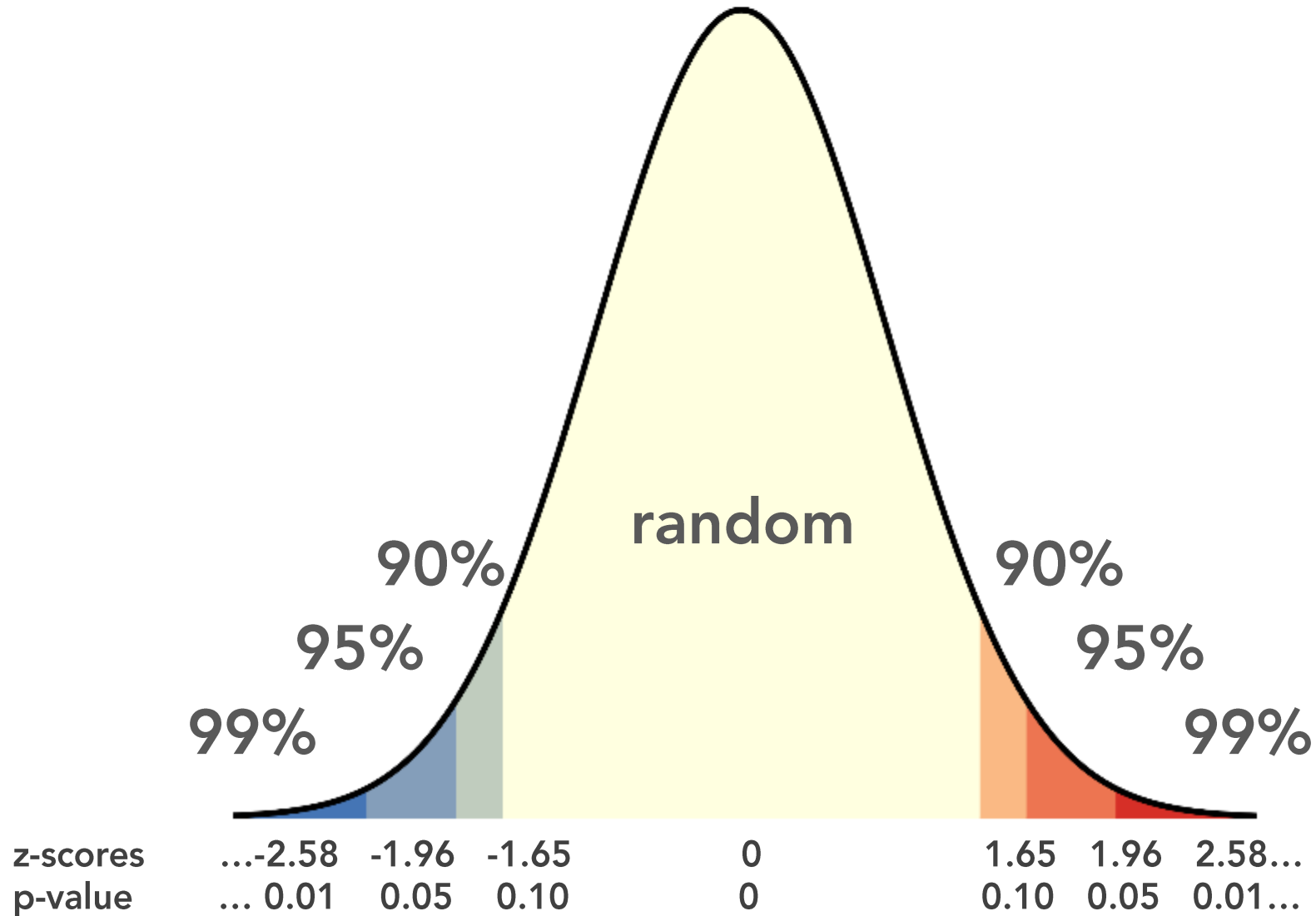
What are the chances this happened RANDOMLY???

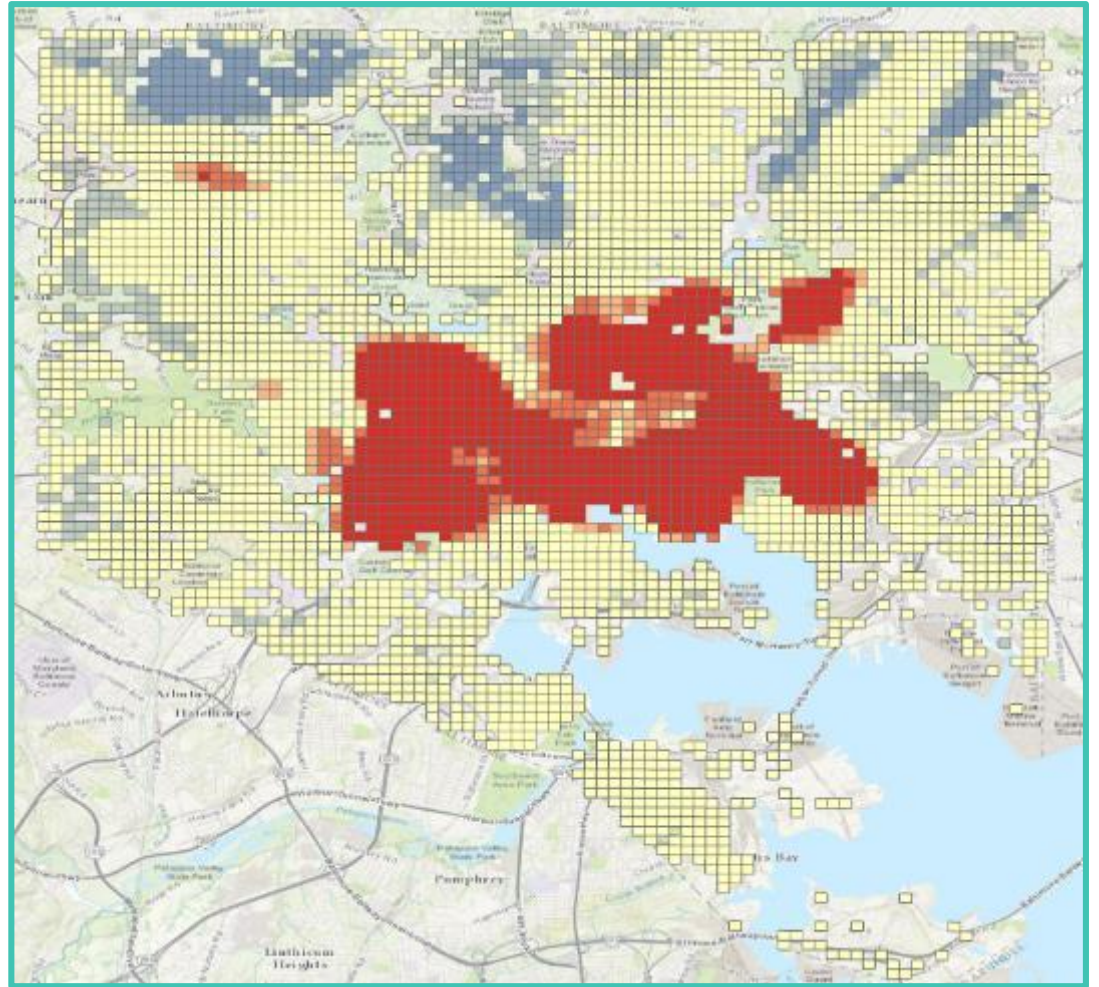
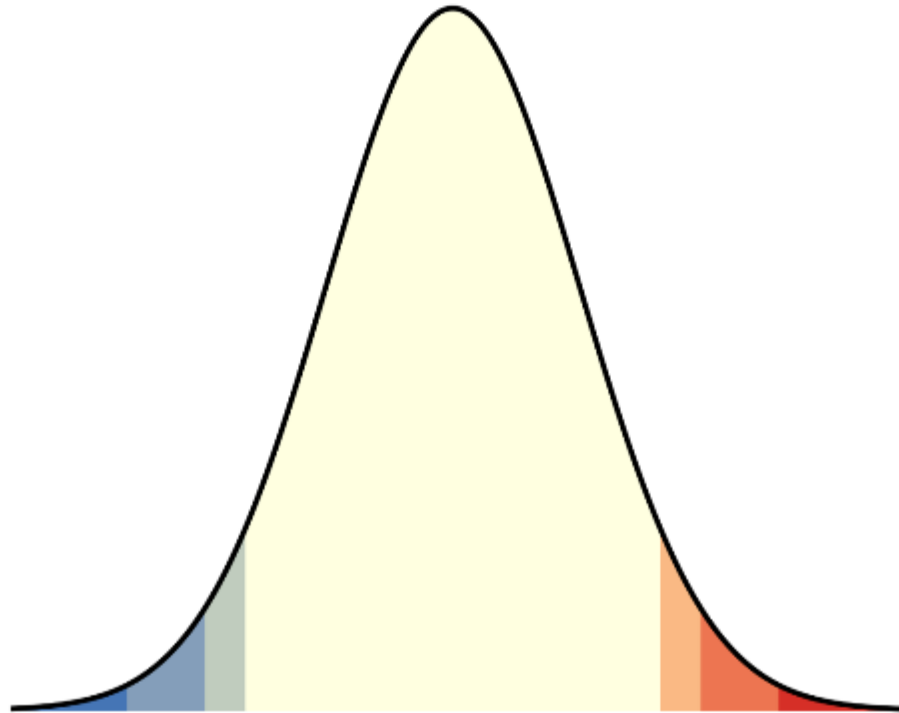


z-scores

p-values

z-scores and p-values

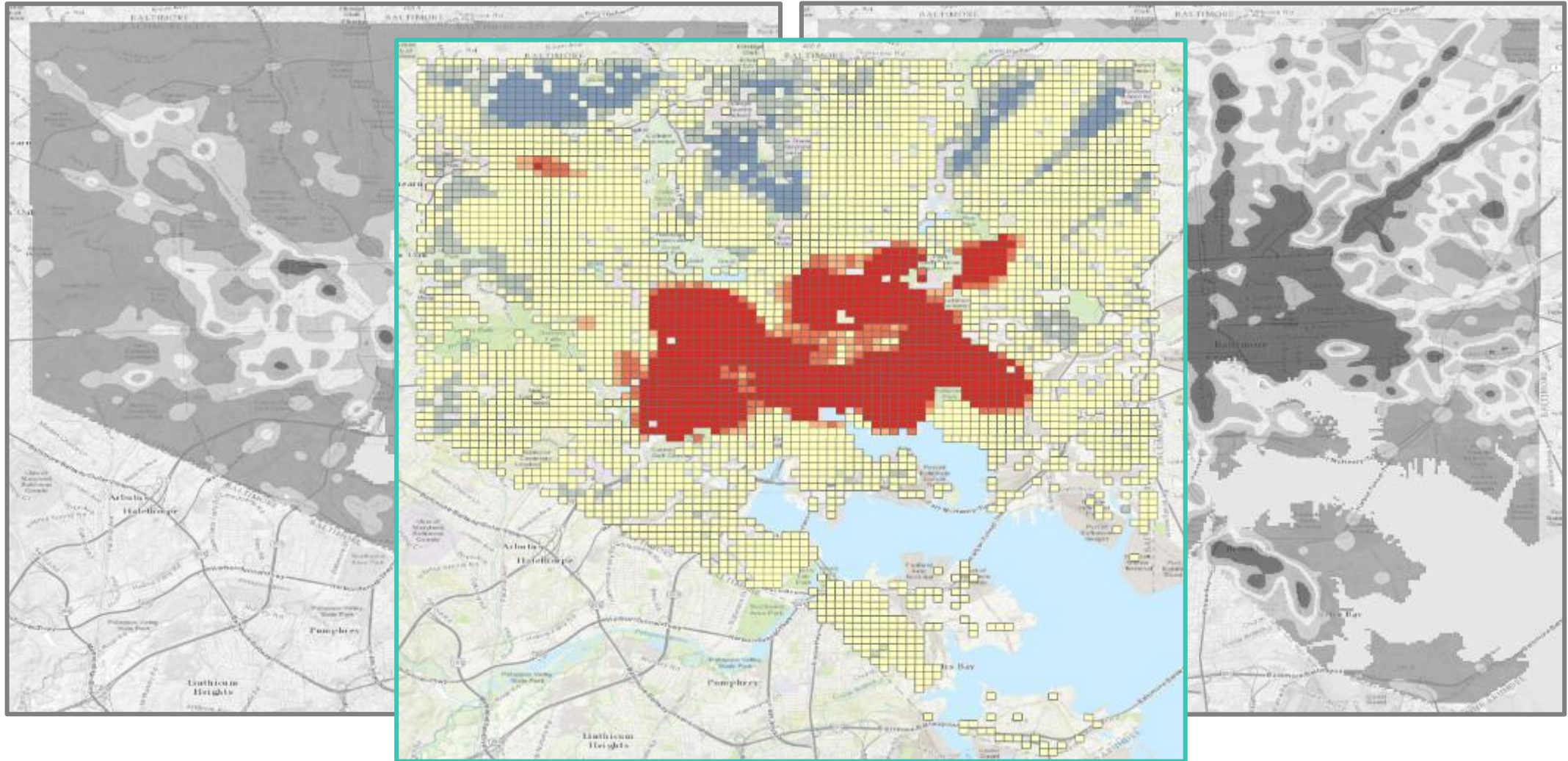




z-scores	...-2.58	-1.96	-1.65	0	1.65	1.96	2.58...
p-values	...0.01	0.05	0.10	0	0.10	0.05	0.01...

Minimizing the subjectivity

Turning the map into **information**



"...everything is related to everything else, but near things are more related than distant things."

Hot Spot Analysis

given a set of weighted features, identifies statistically significant hot spots and cold spots using the Getis-Ord G_i^* statistic

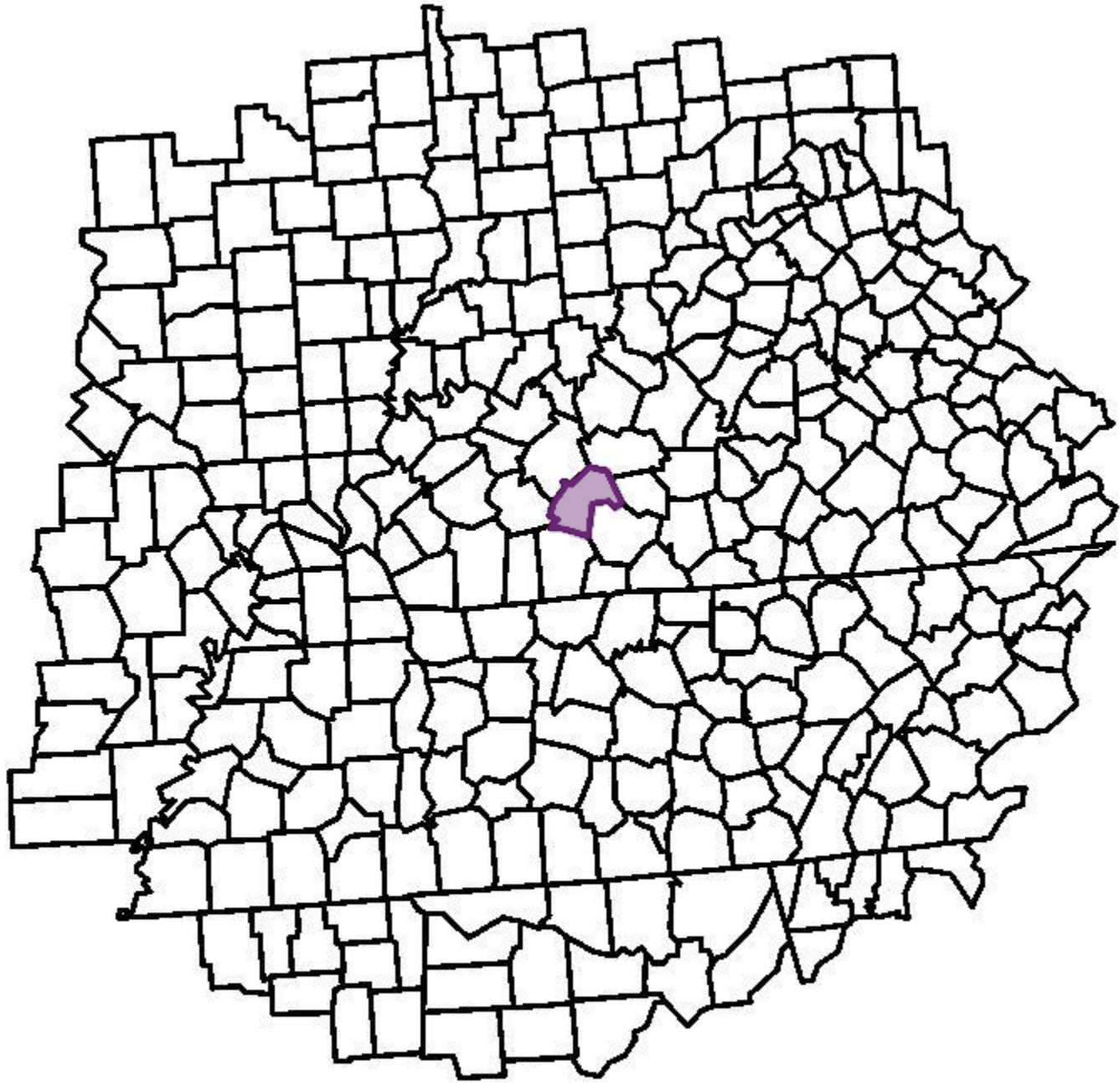
Polygons





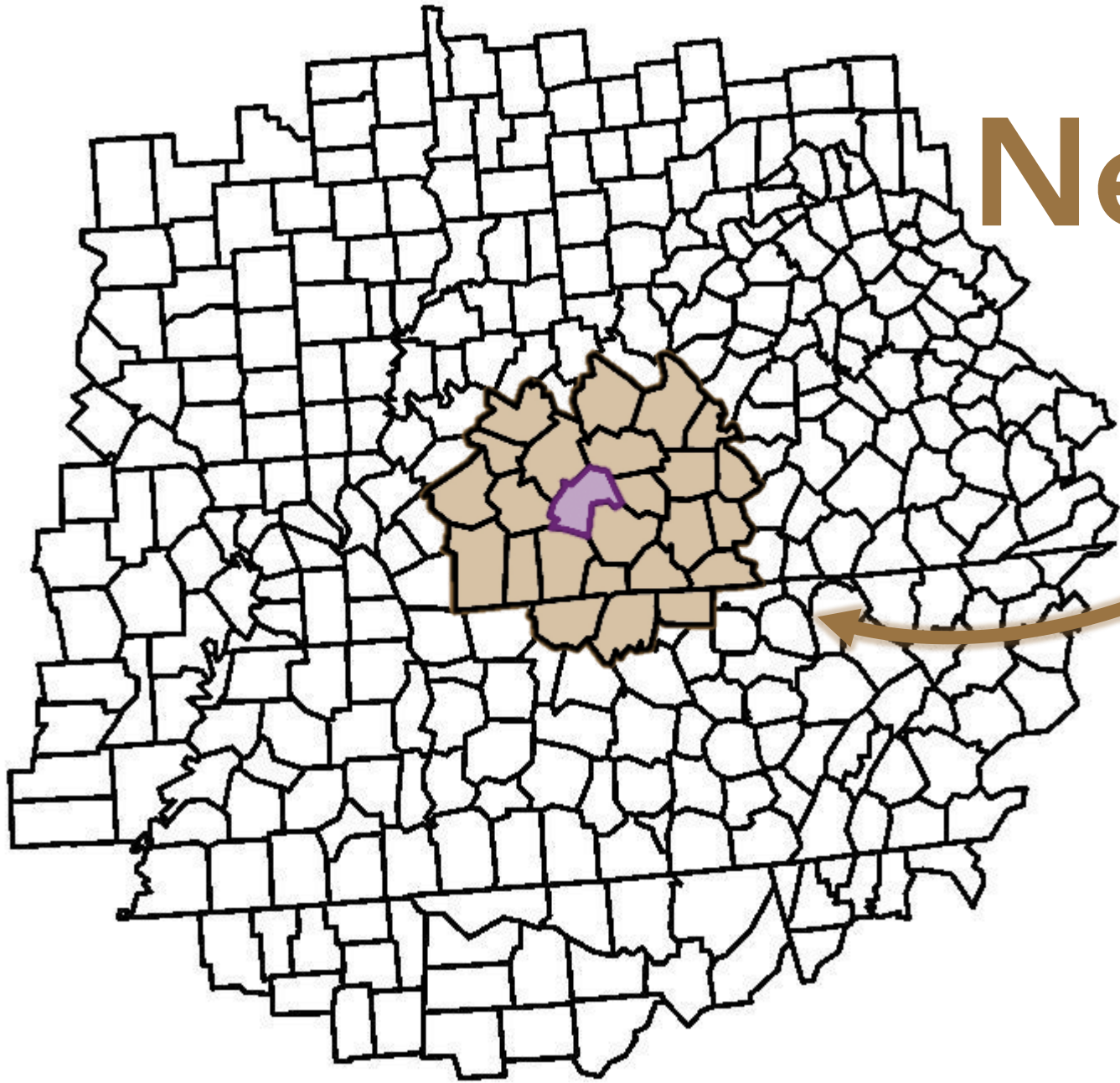
feature

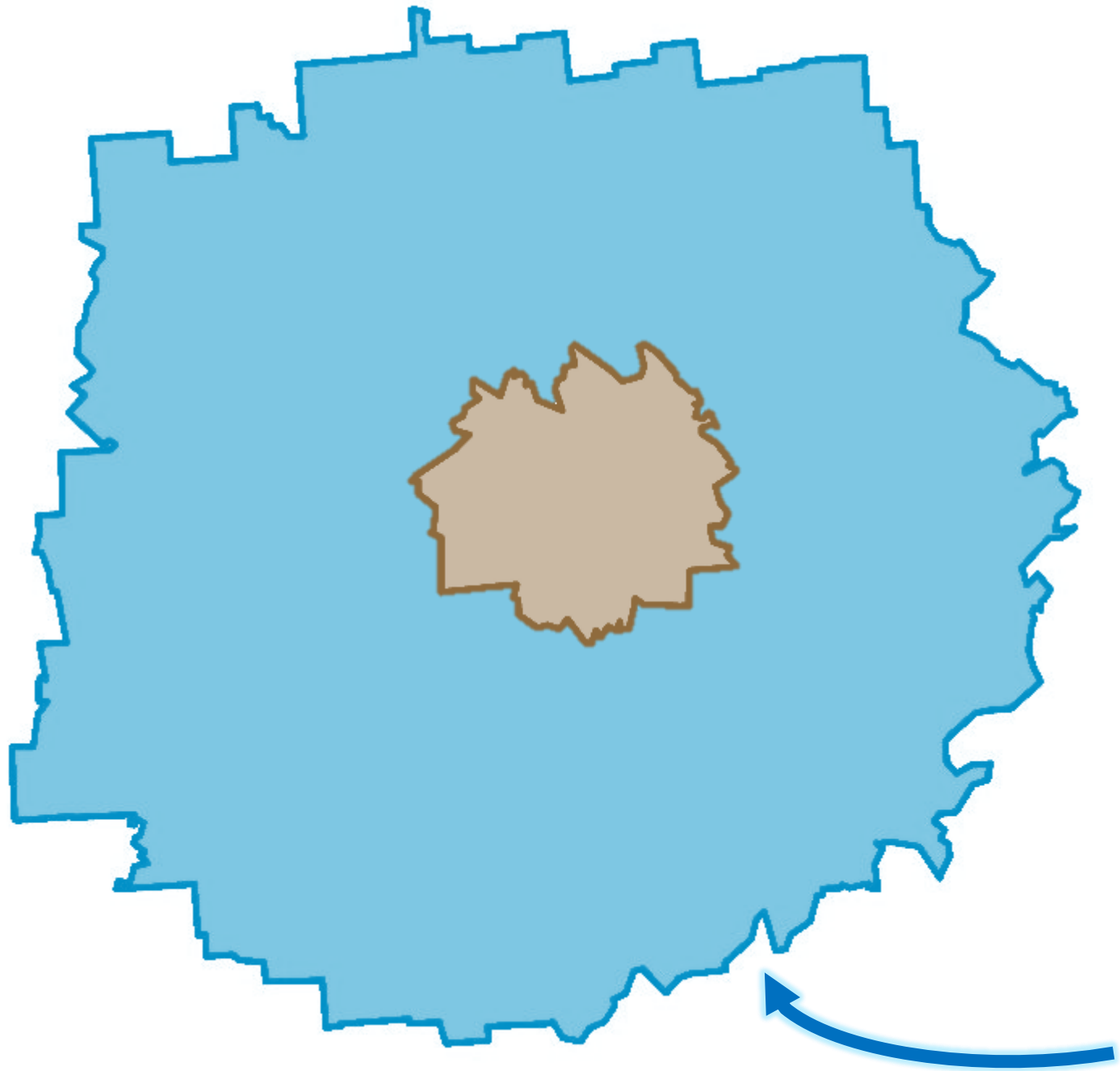




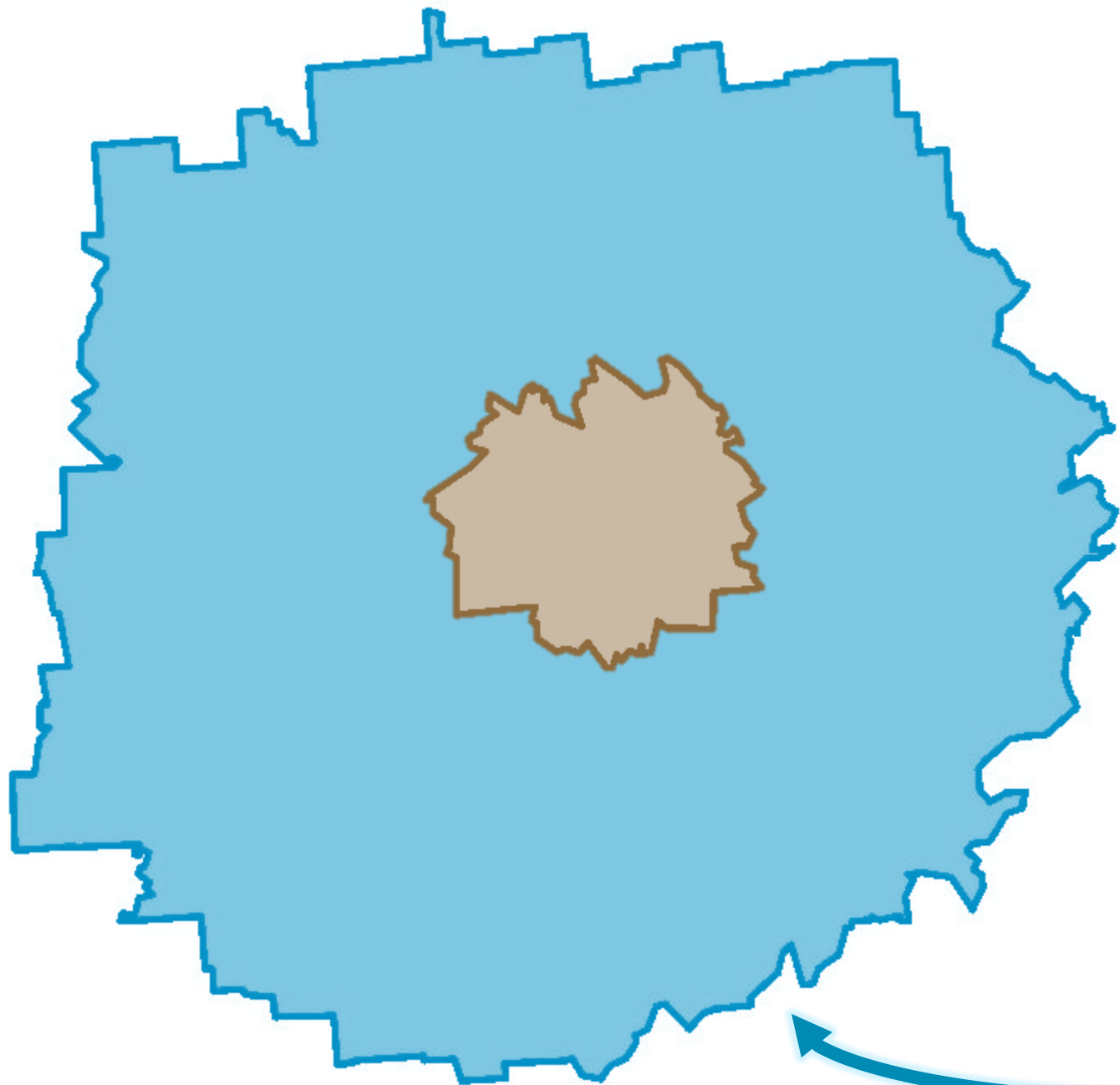
each
feature
has a
value

Neighborhood





**Study
Area**

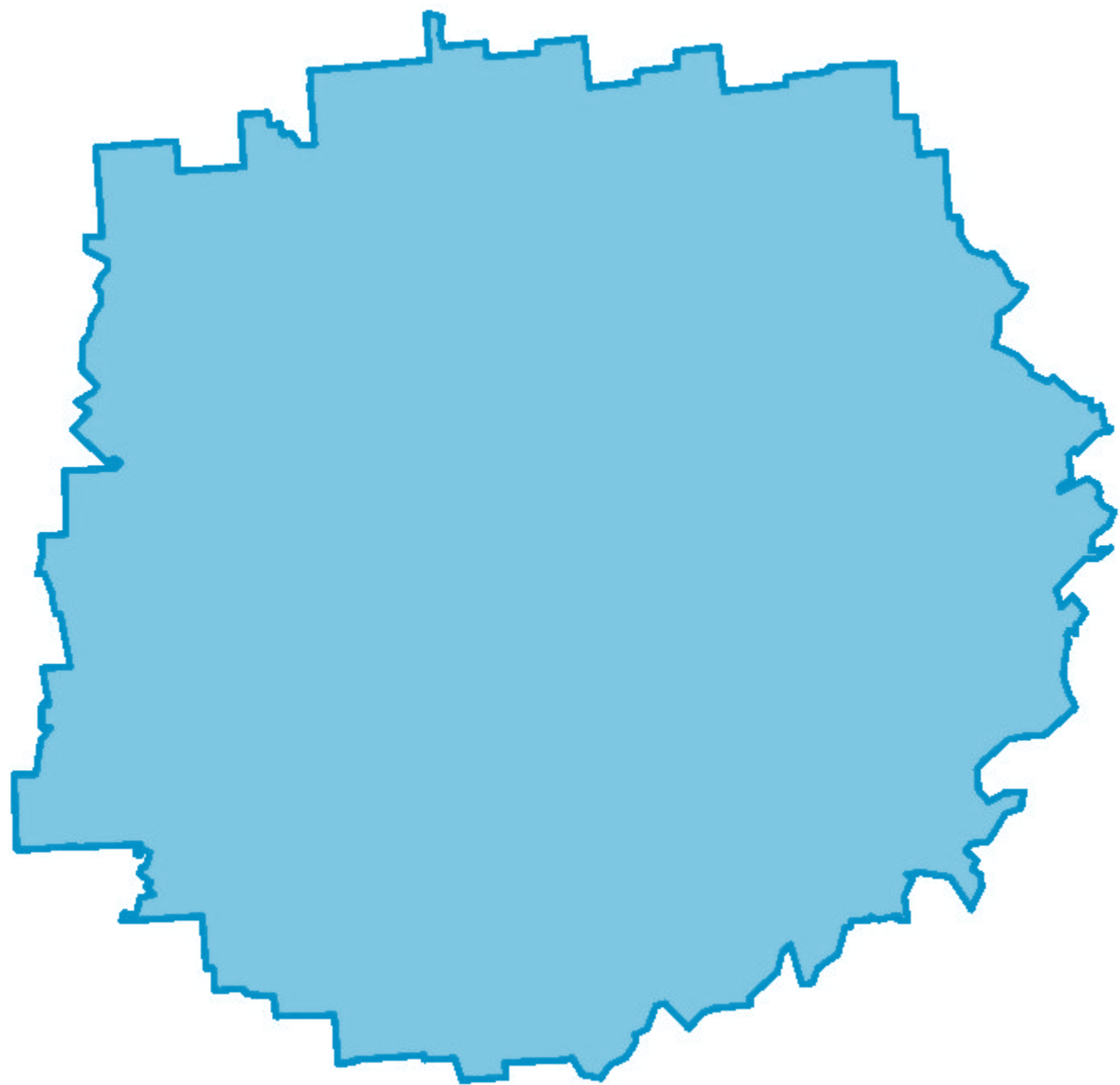


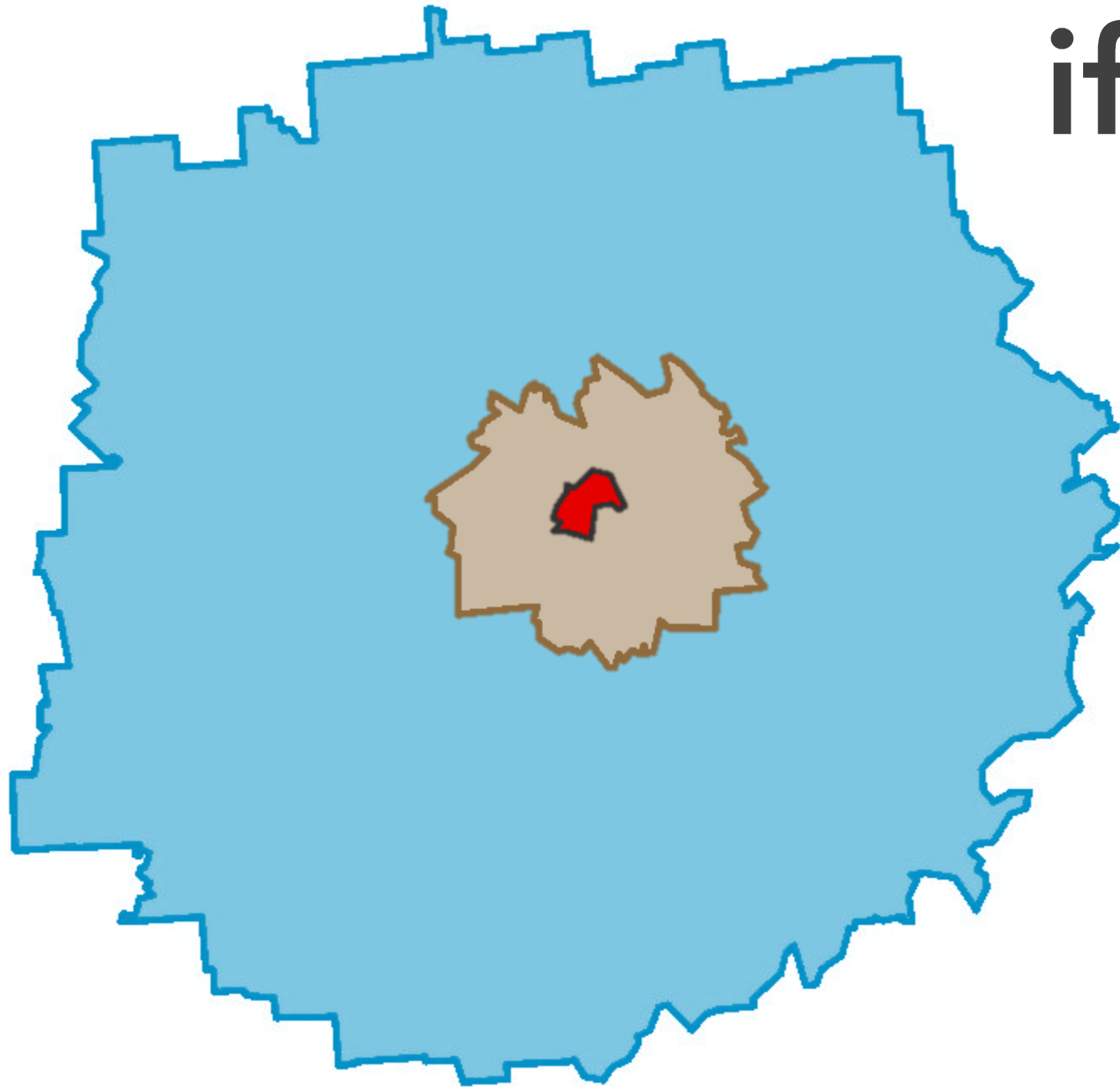
is this



significantly
different from
this?





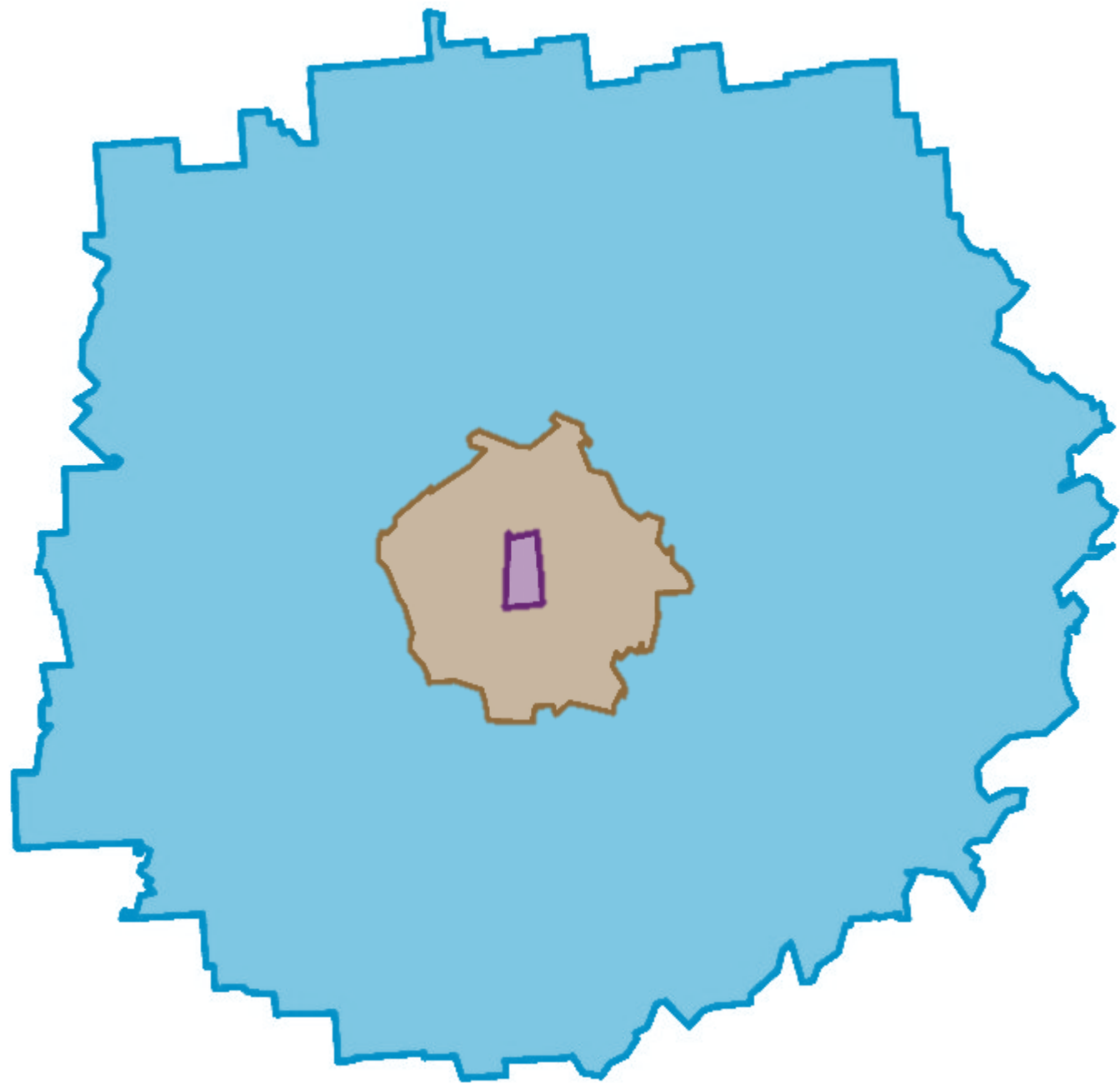


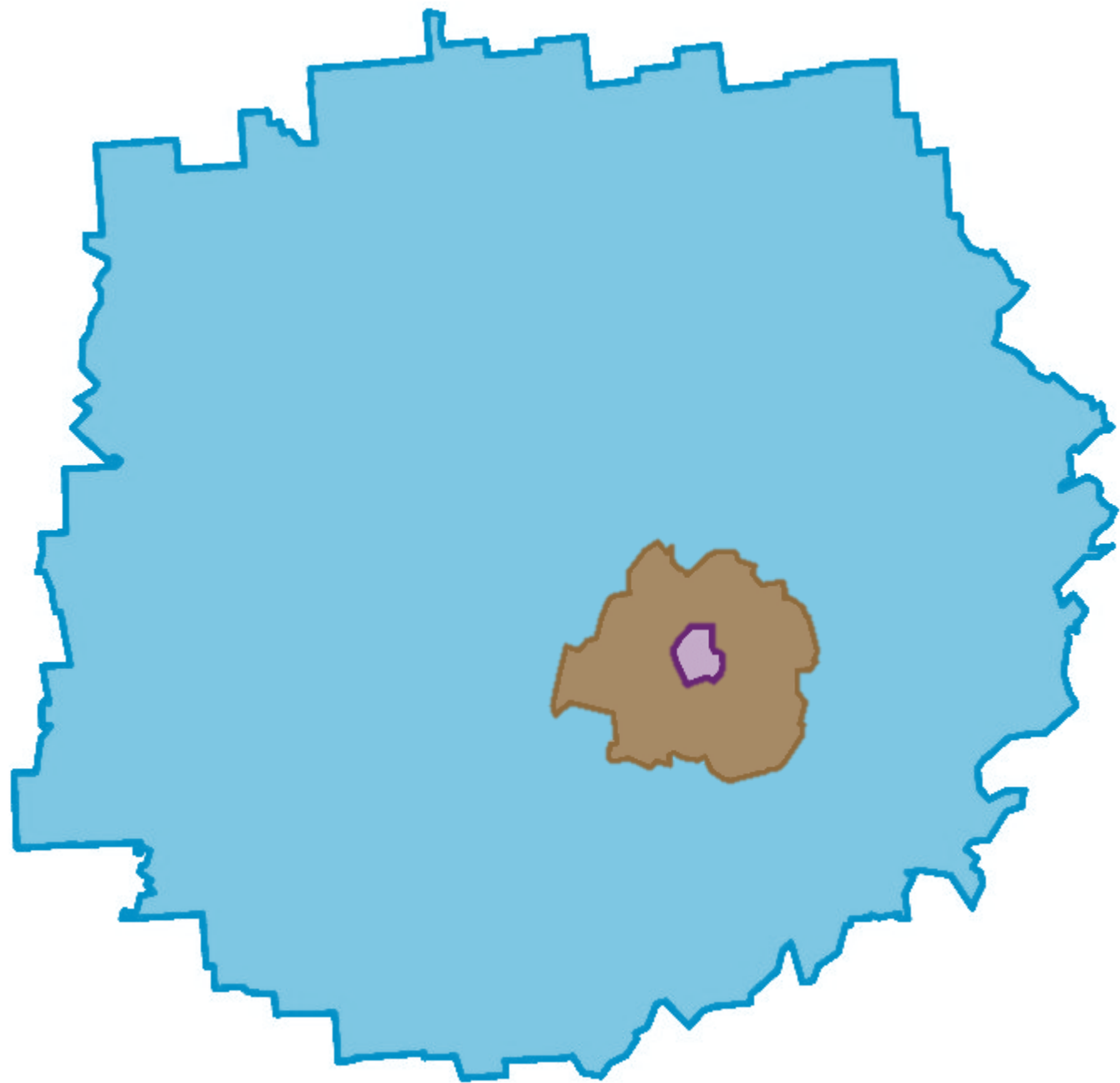
if significantly
higher...

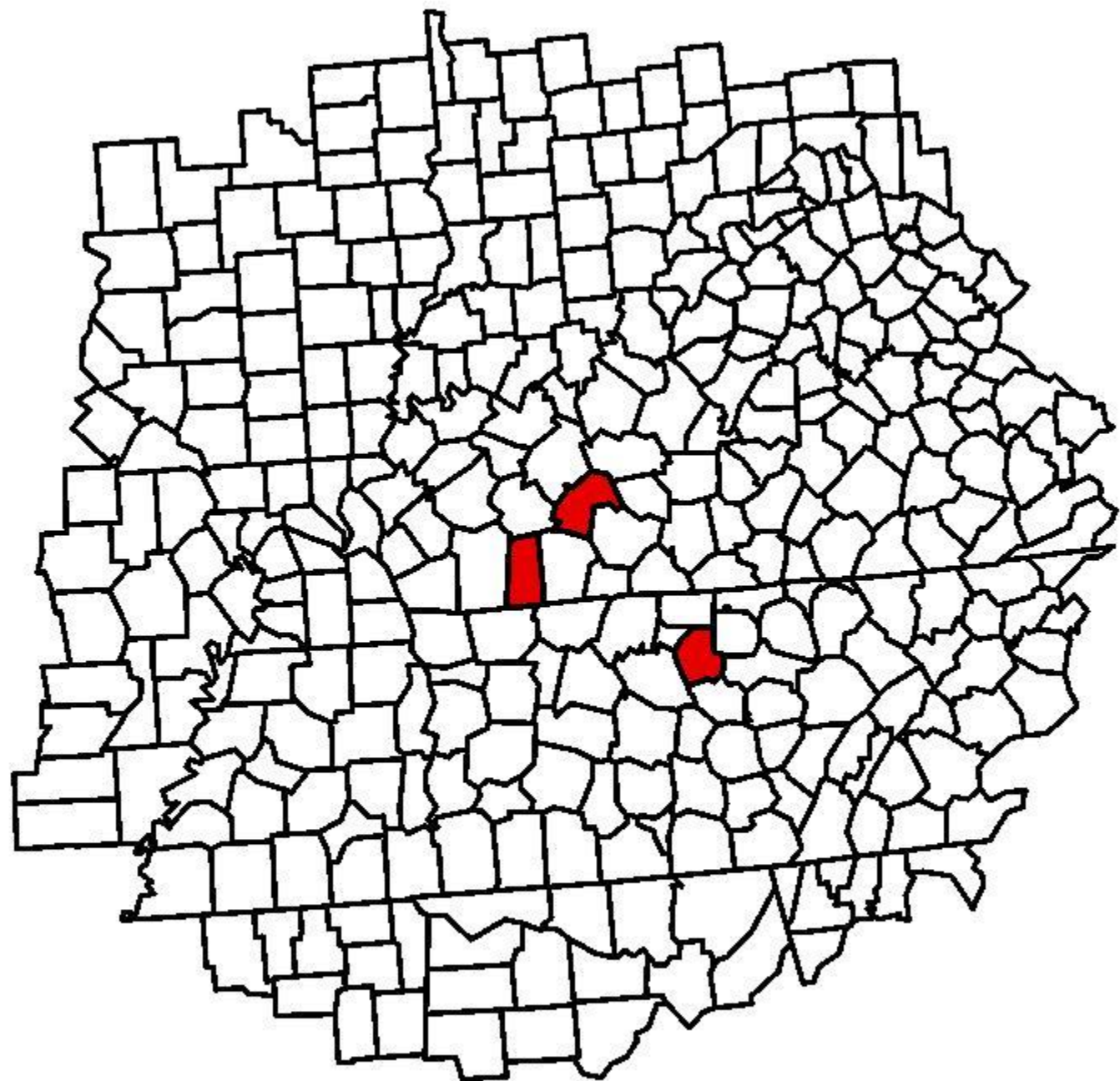
feature is
marked as a
hot spot!

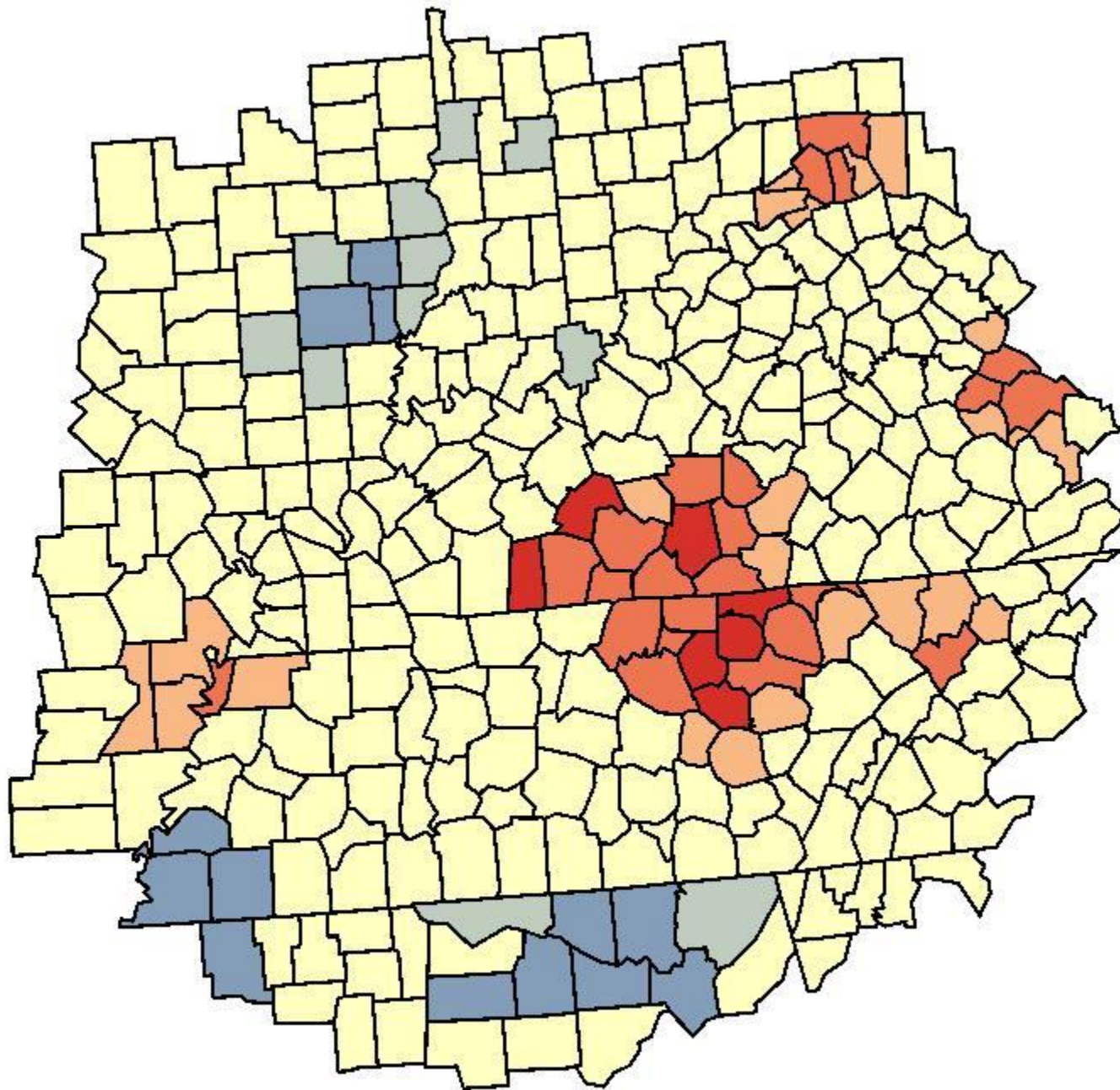





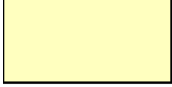



-  Hot Spot - 90% Confidence
-  Hot Spot - 95% Confidence
-  Hot Spot - 99% Confidence









-  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

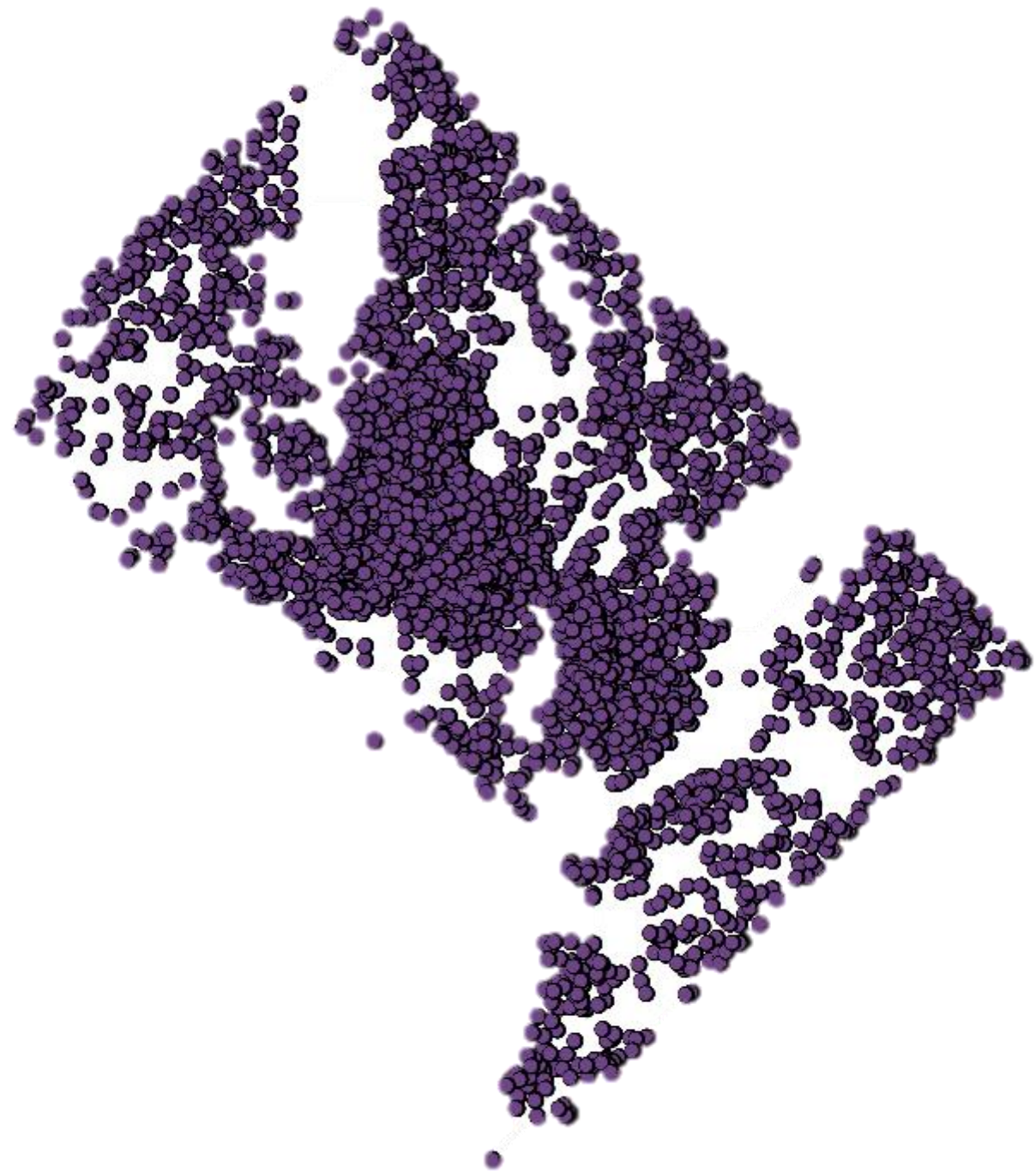
...how do we know if it's
SIGNIFICANTLY
different???

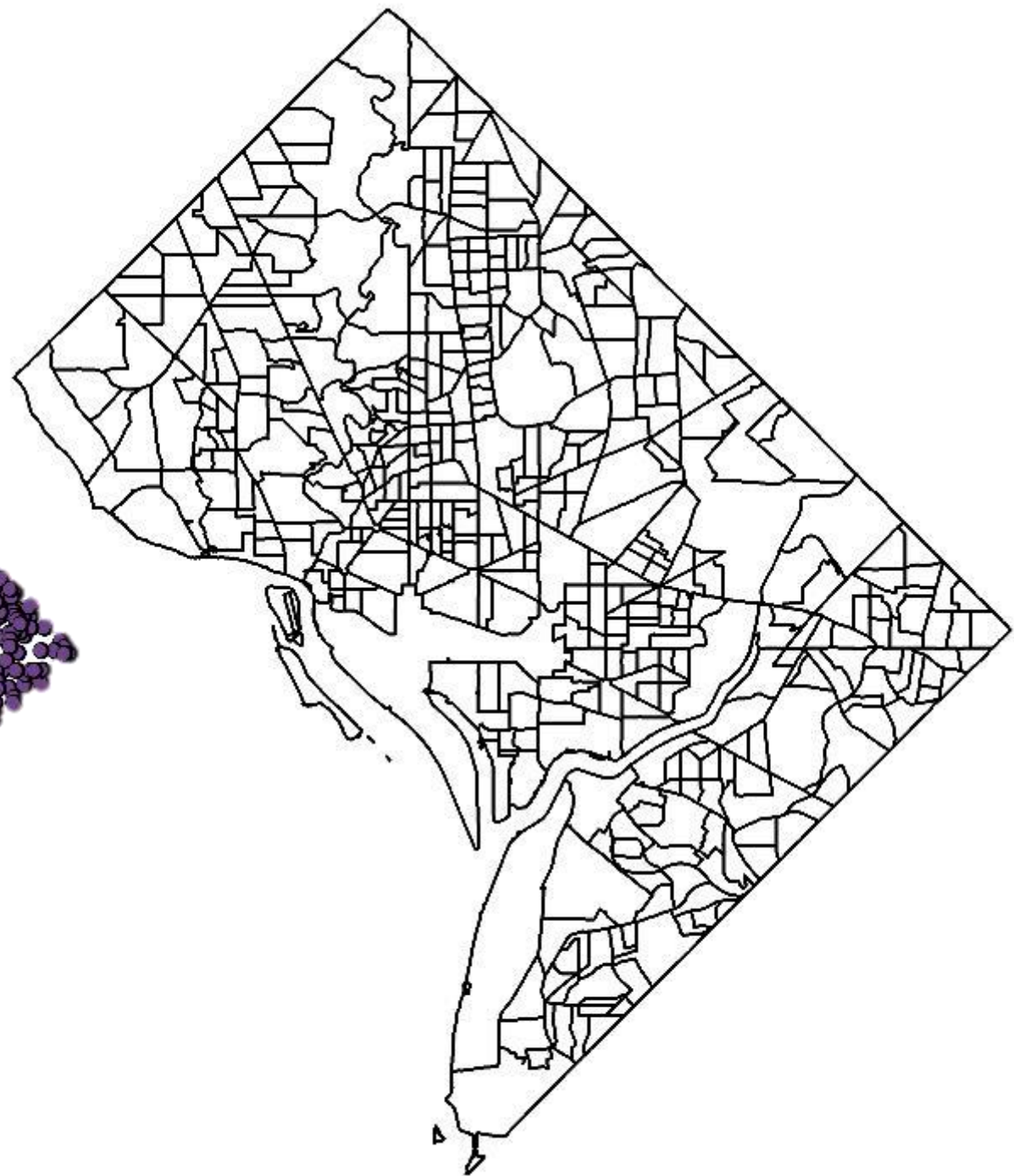
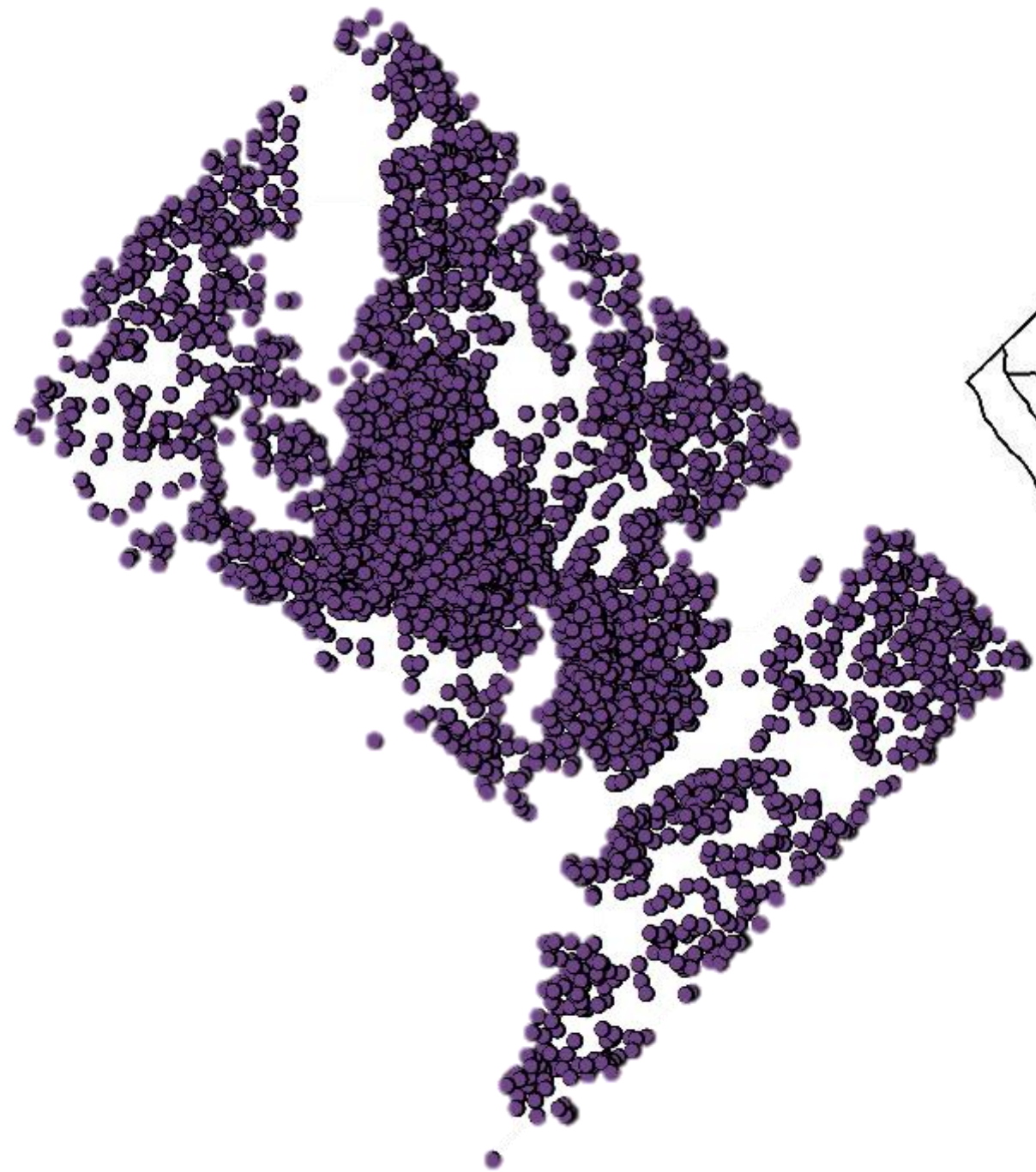
MATH!

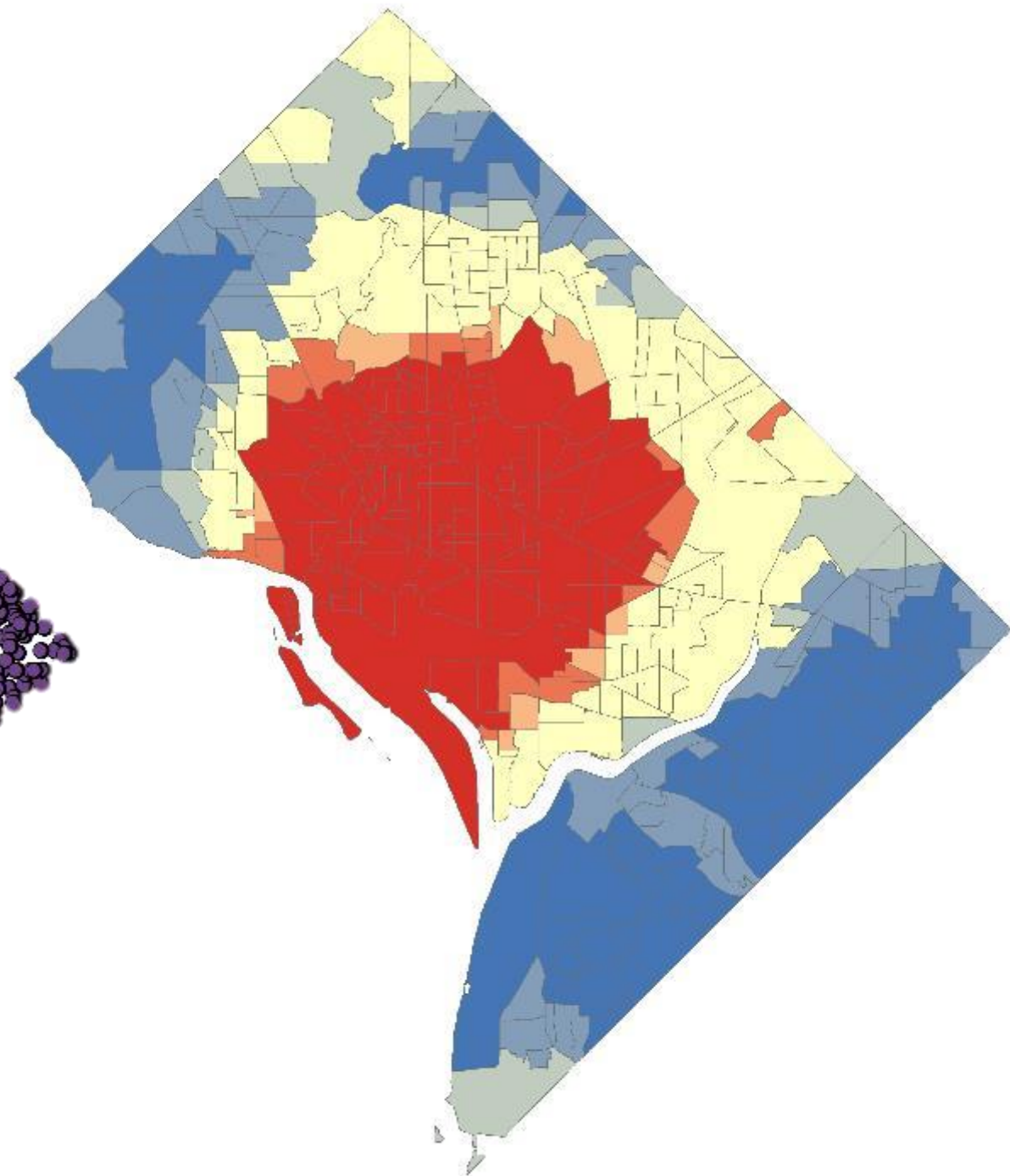
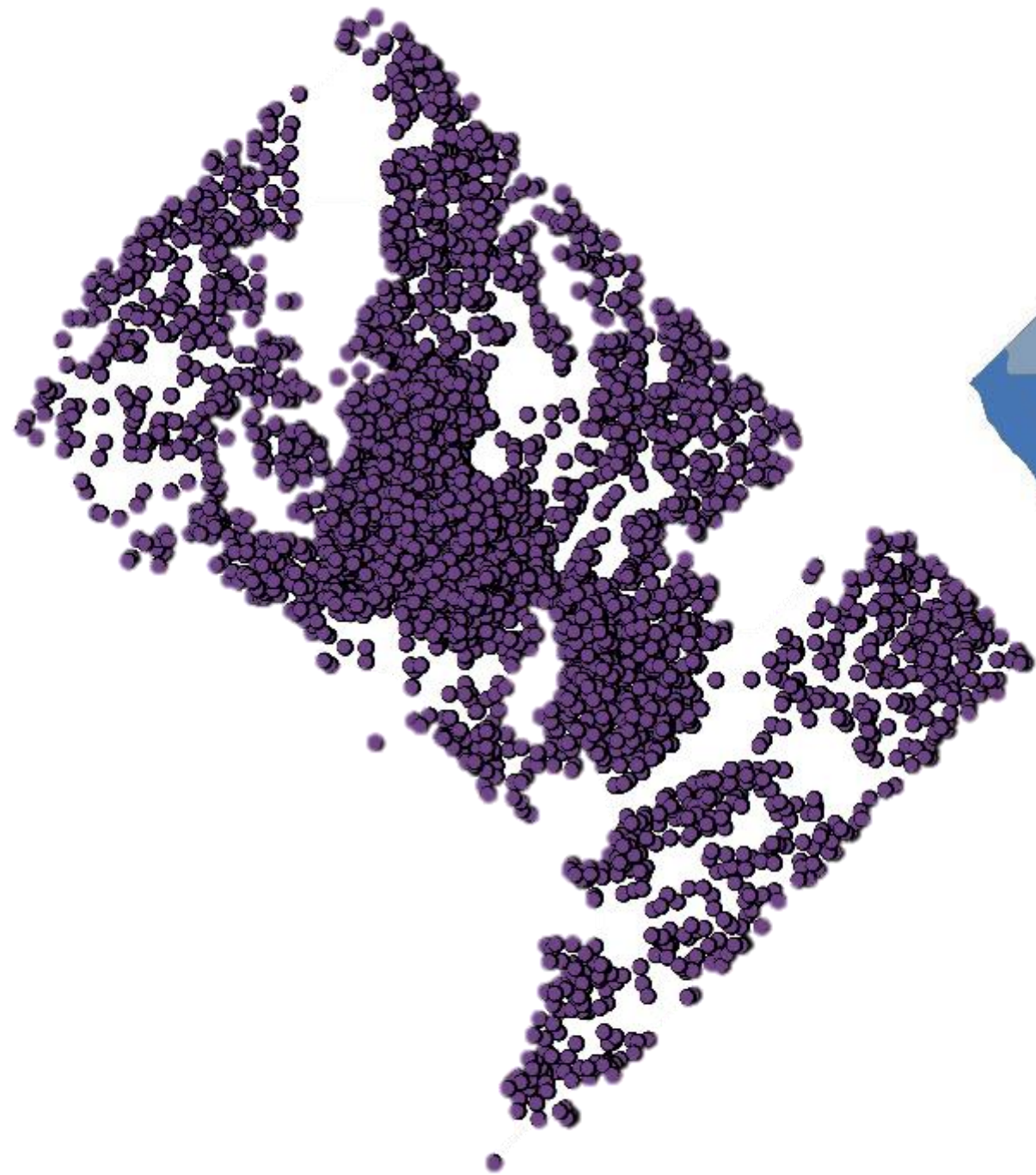
Getis-Ord G_i^*

Statistic

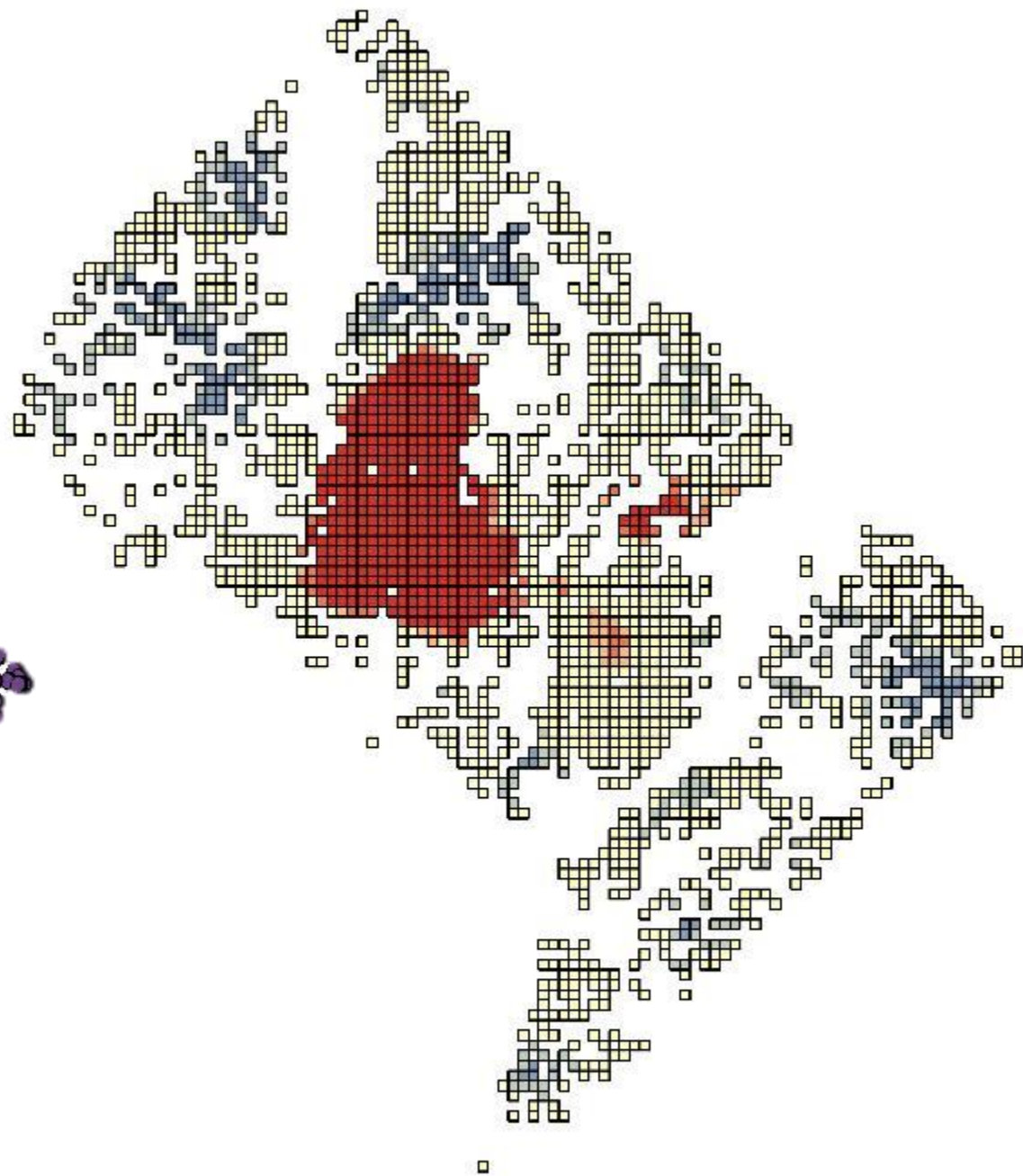
Points

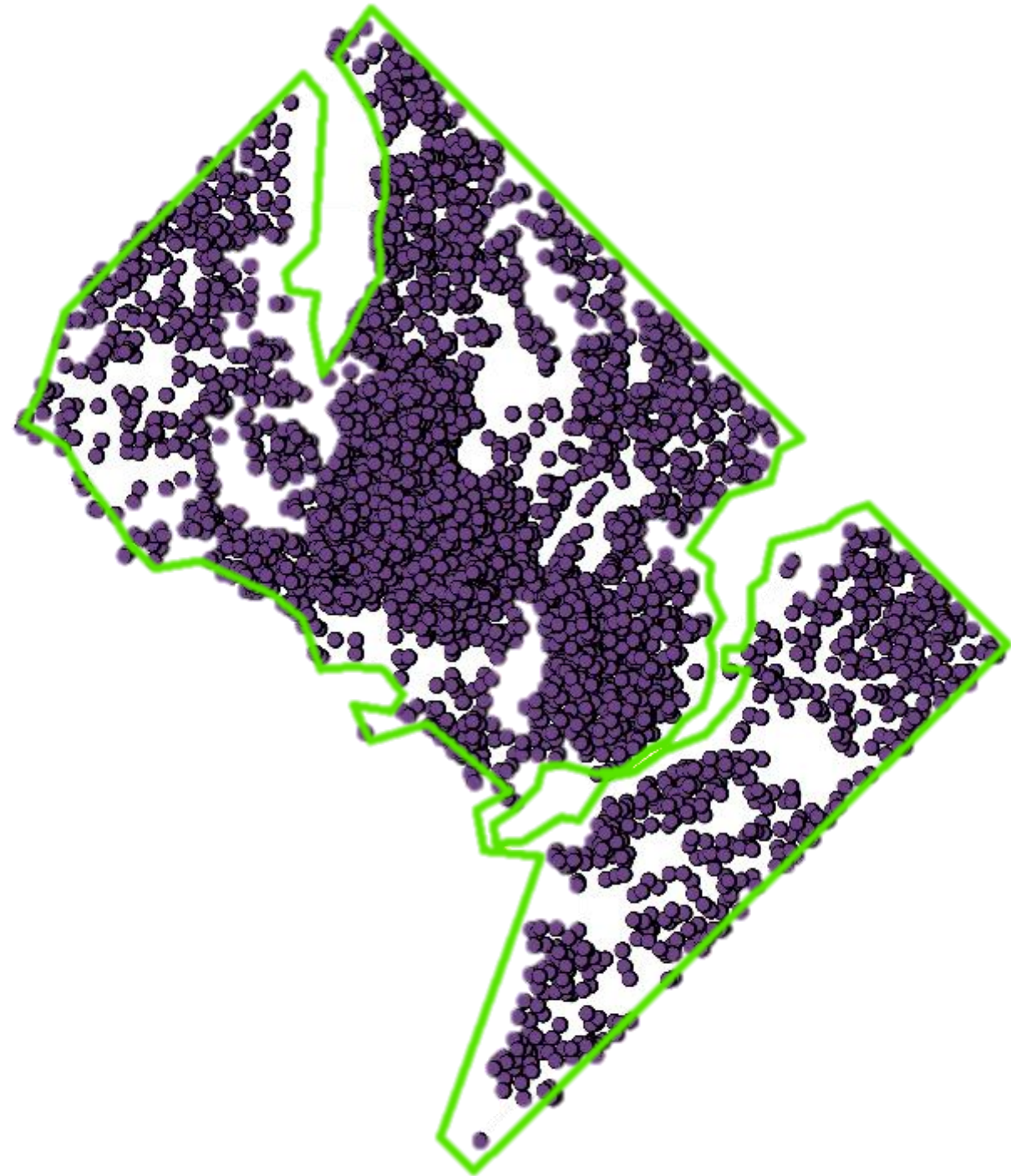


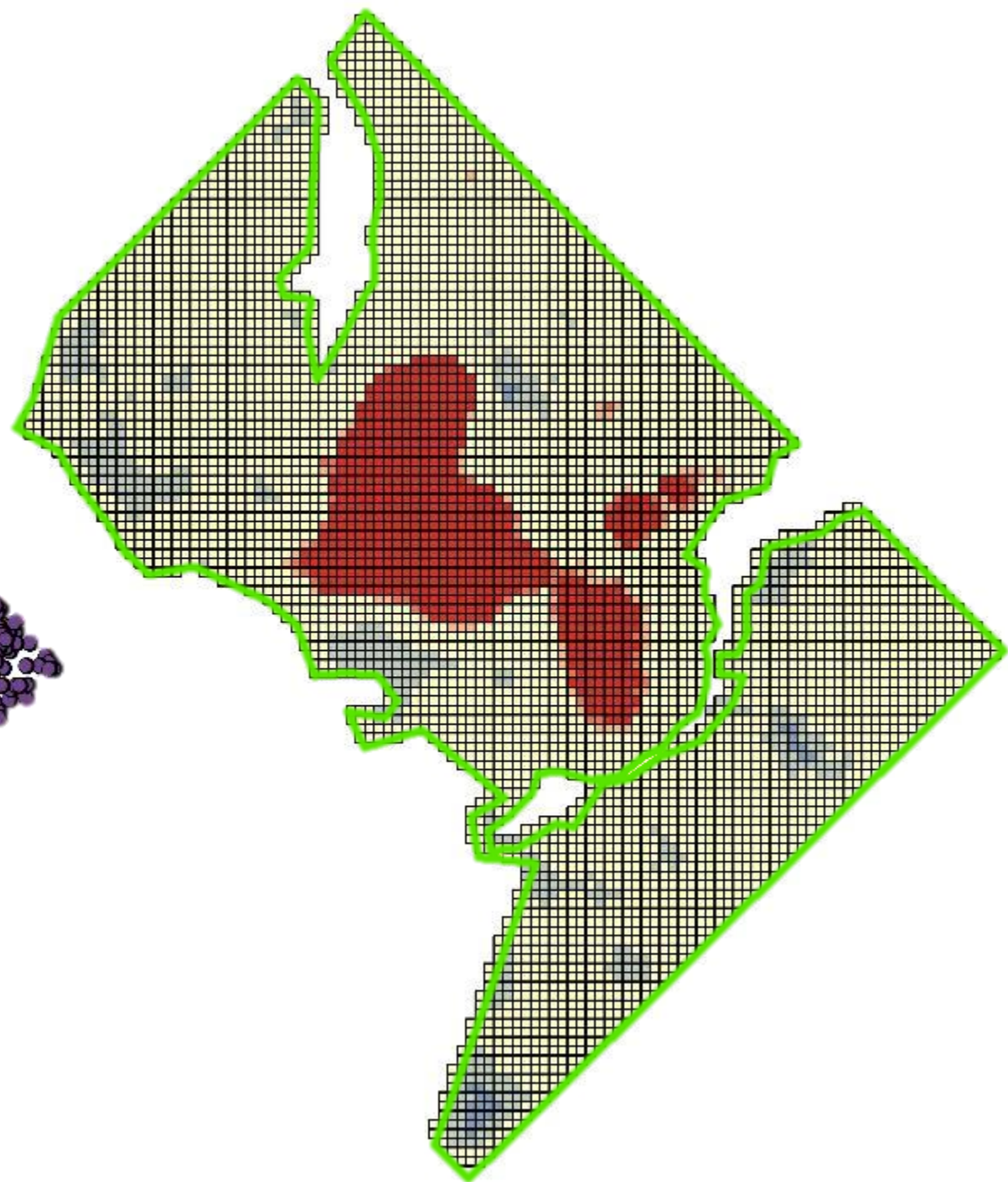
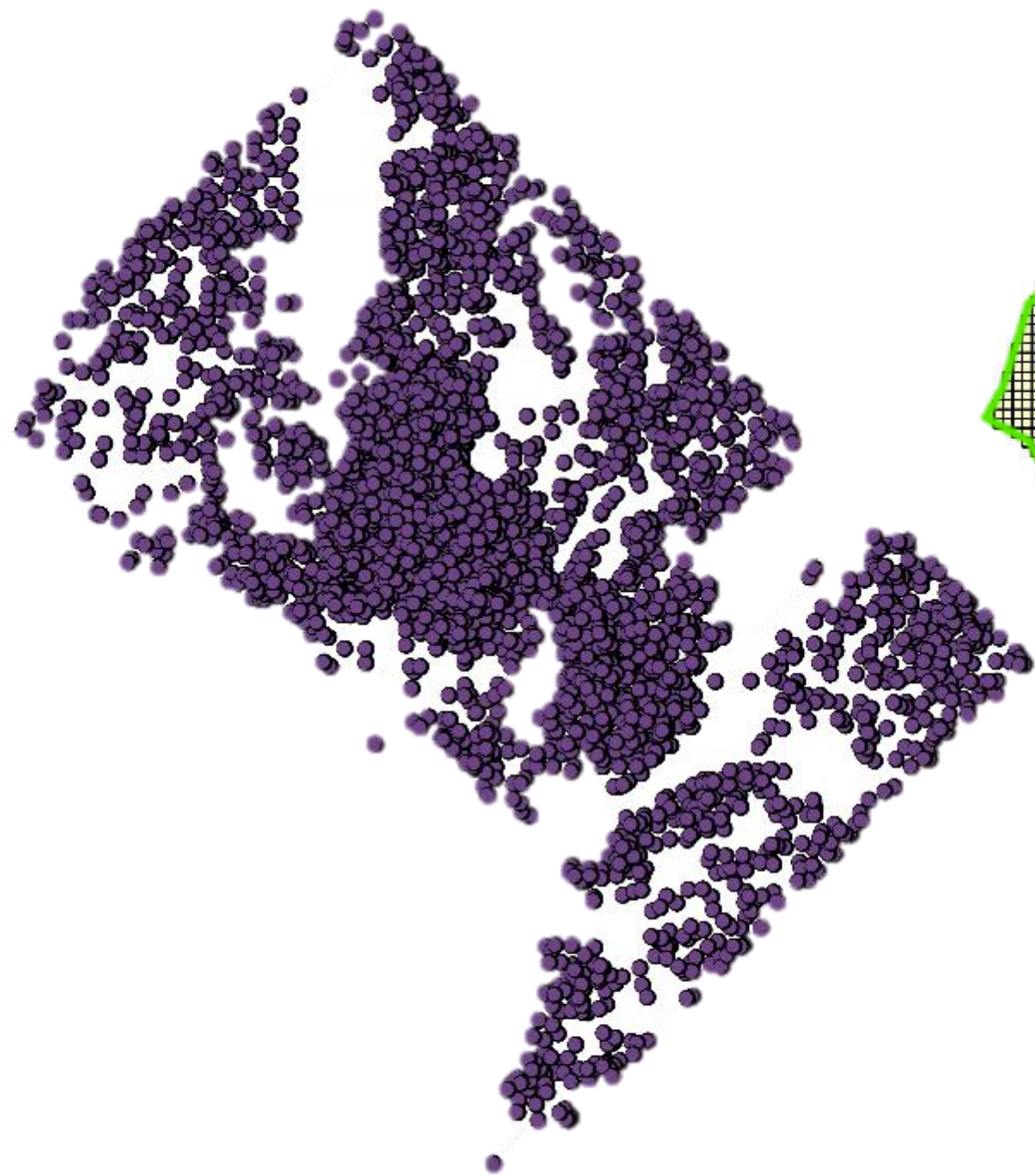










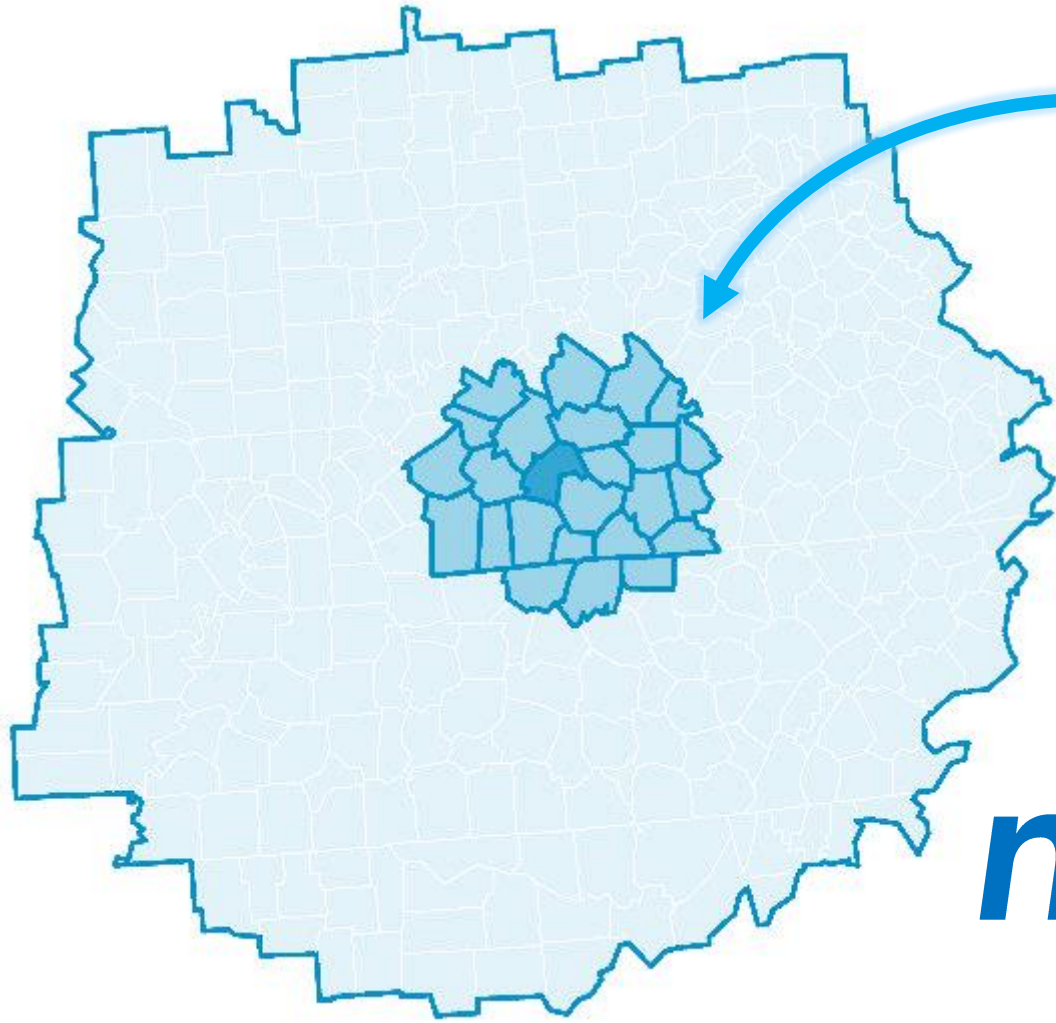


“Where are
the **Hot**
Spots?”

not (necessarily)

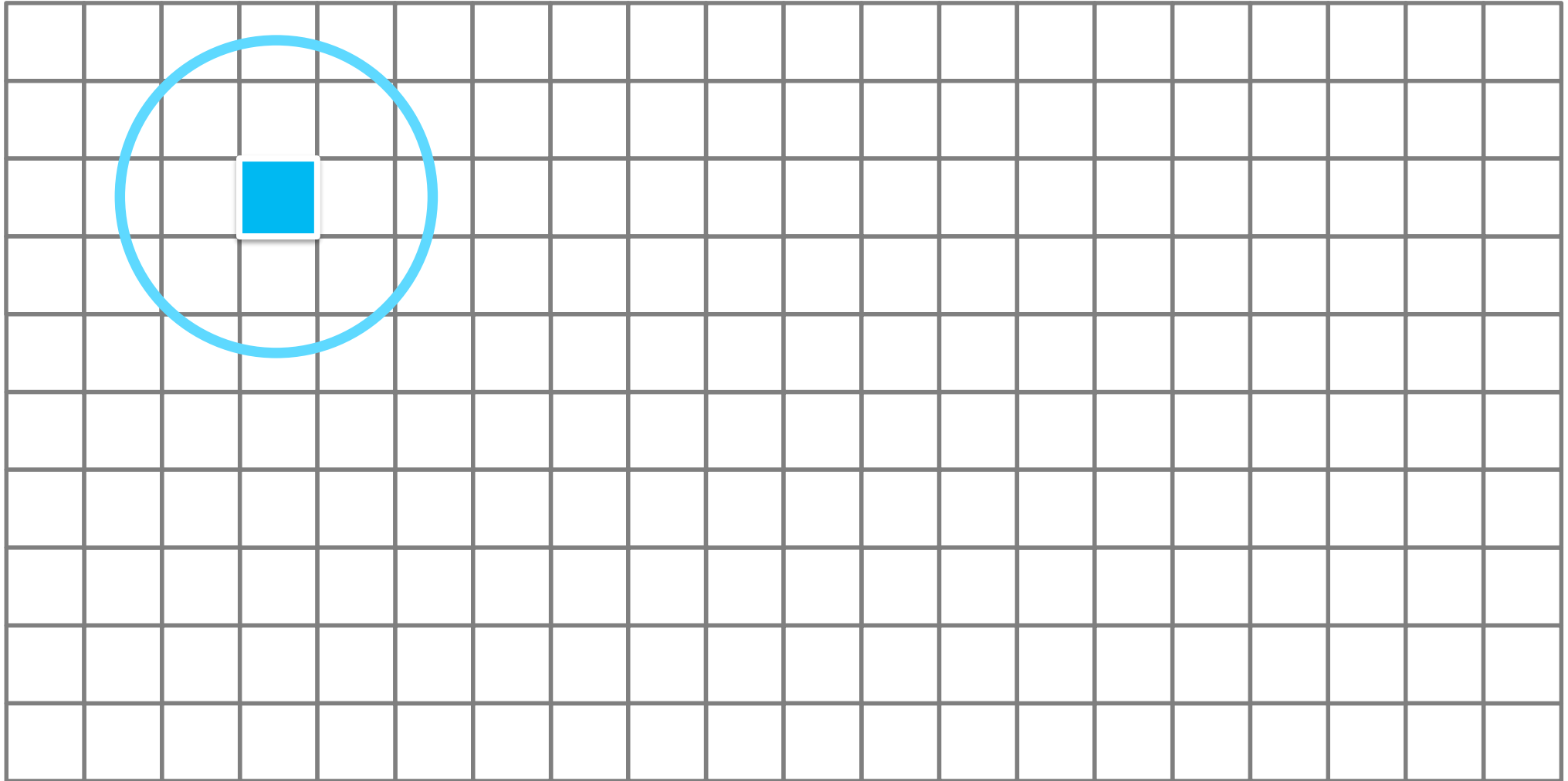
the same as

“Where are
the **highest**
values?”

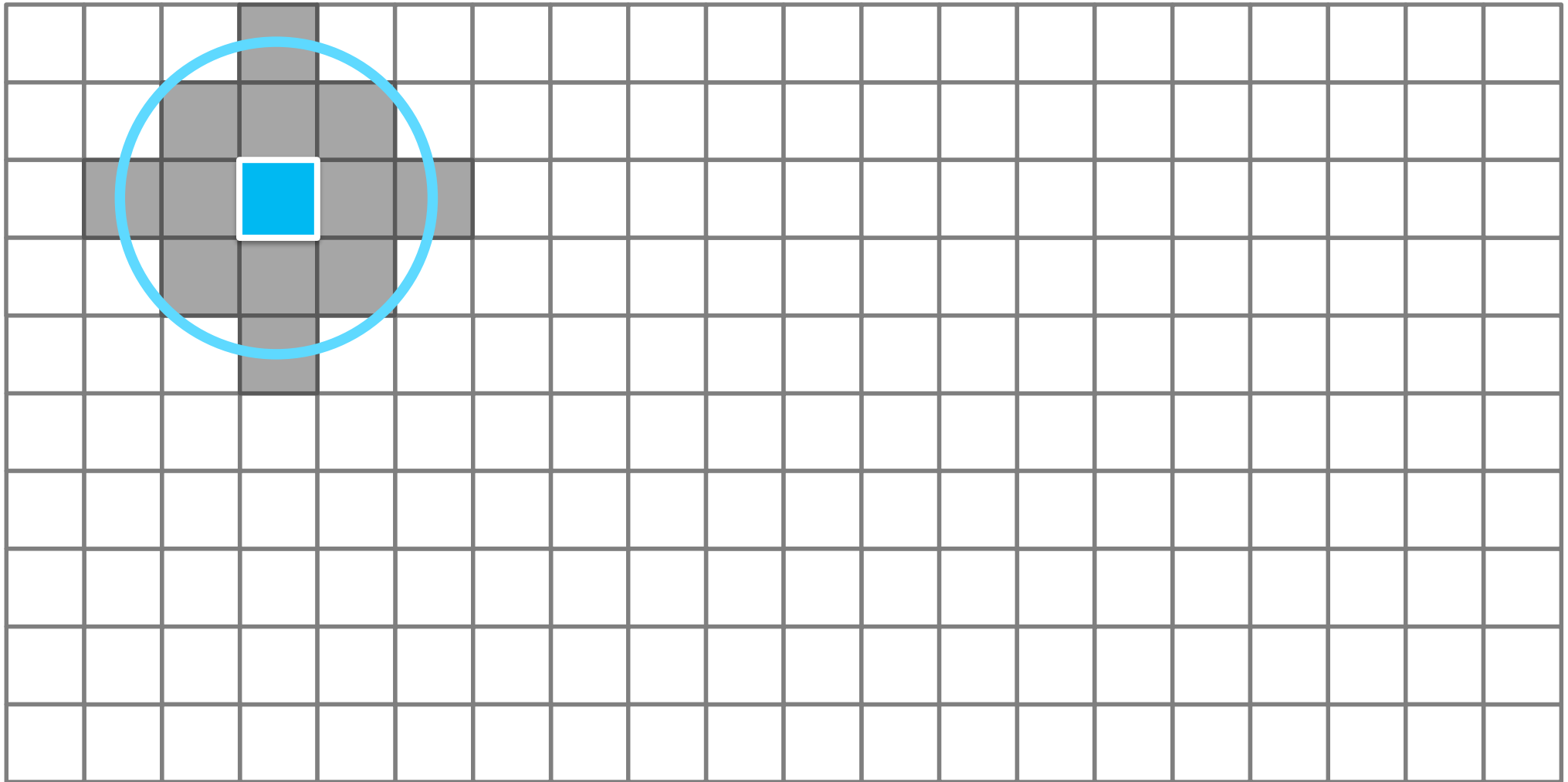


How are
neighborhood
sizes determined?

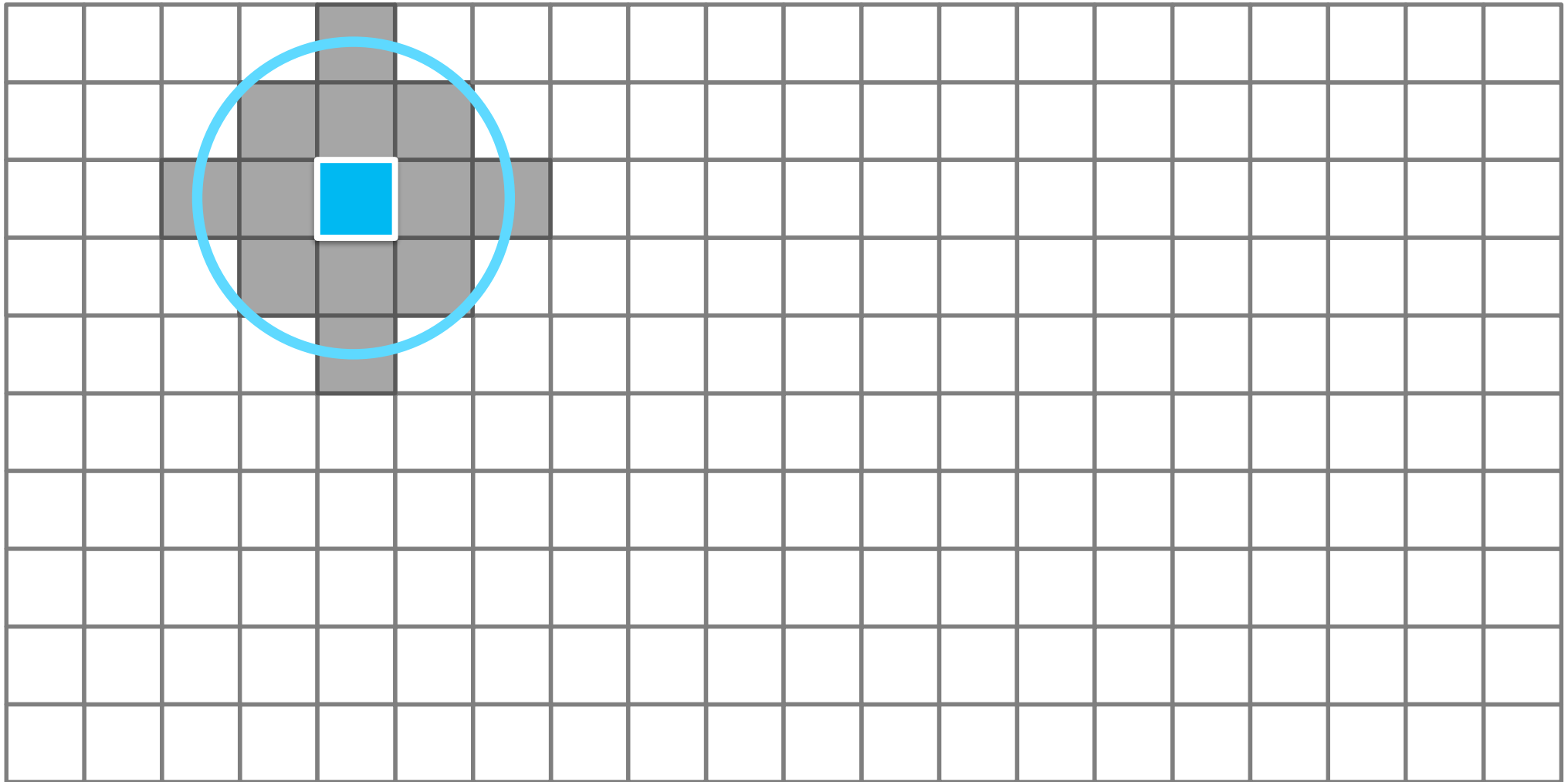
Fixed Distance Band



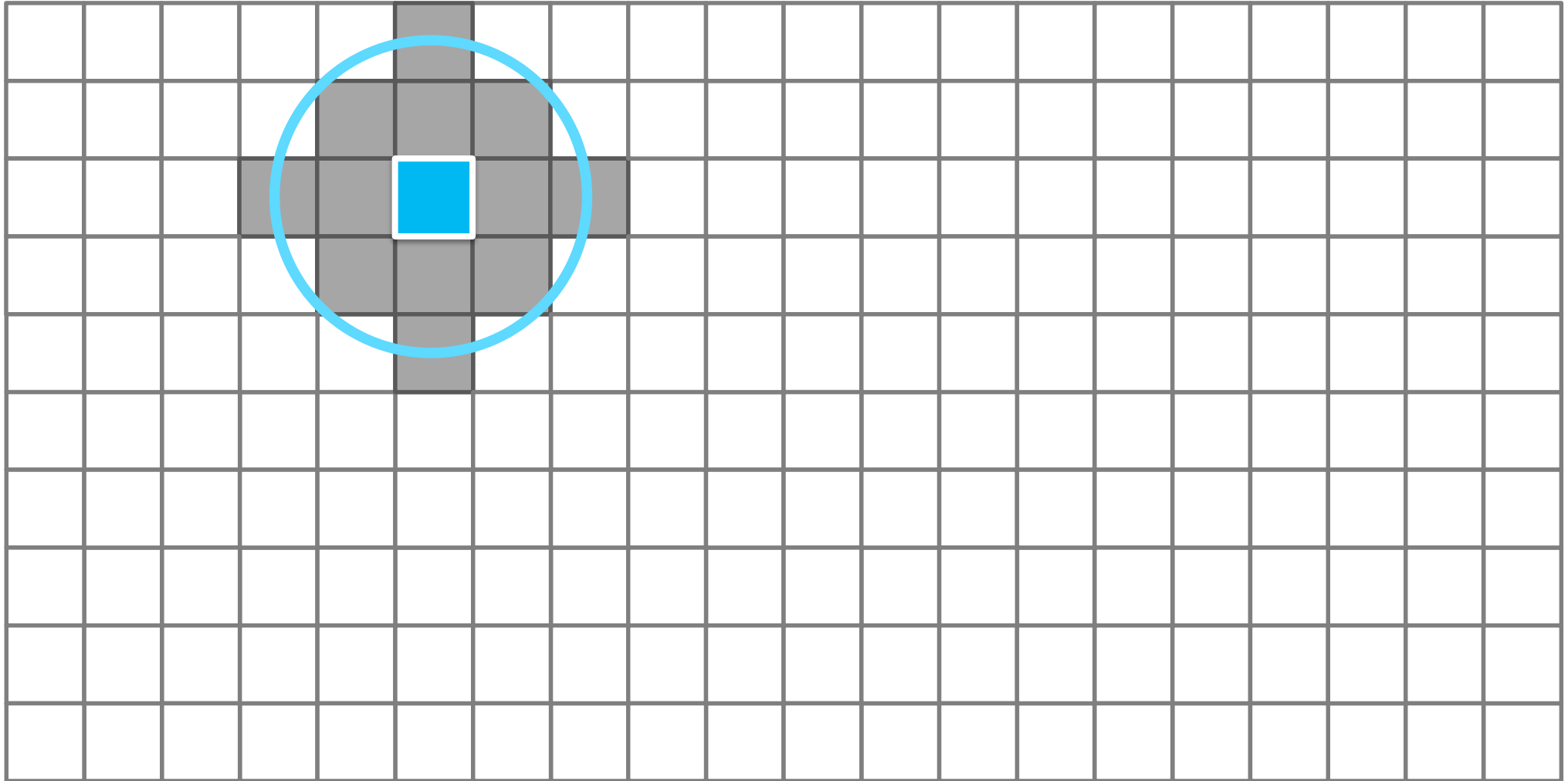
Fixed Distance Band



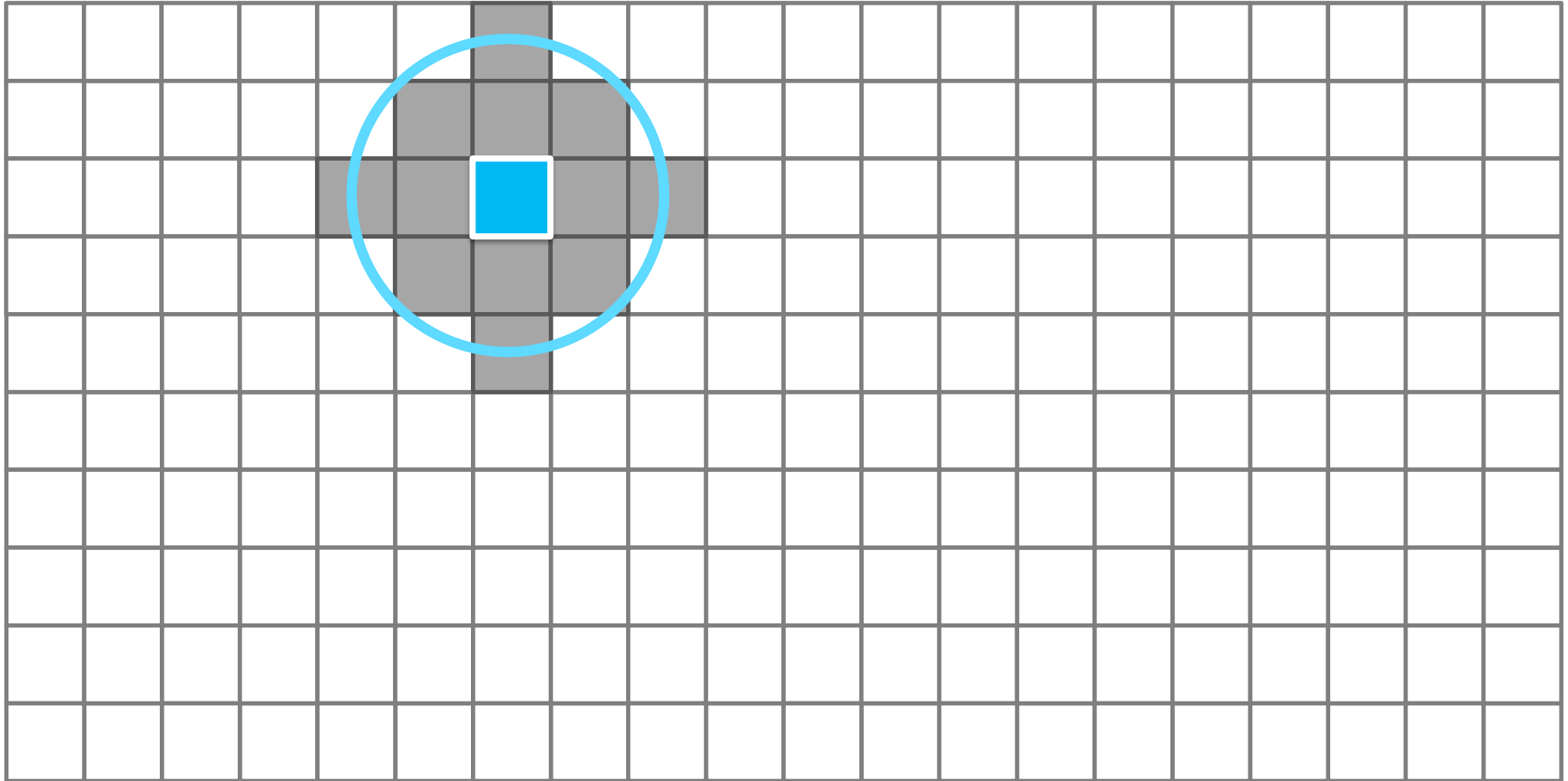
Fixed Distance Band



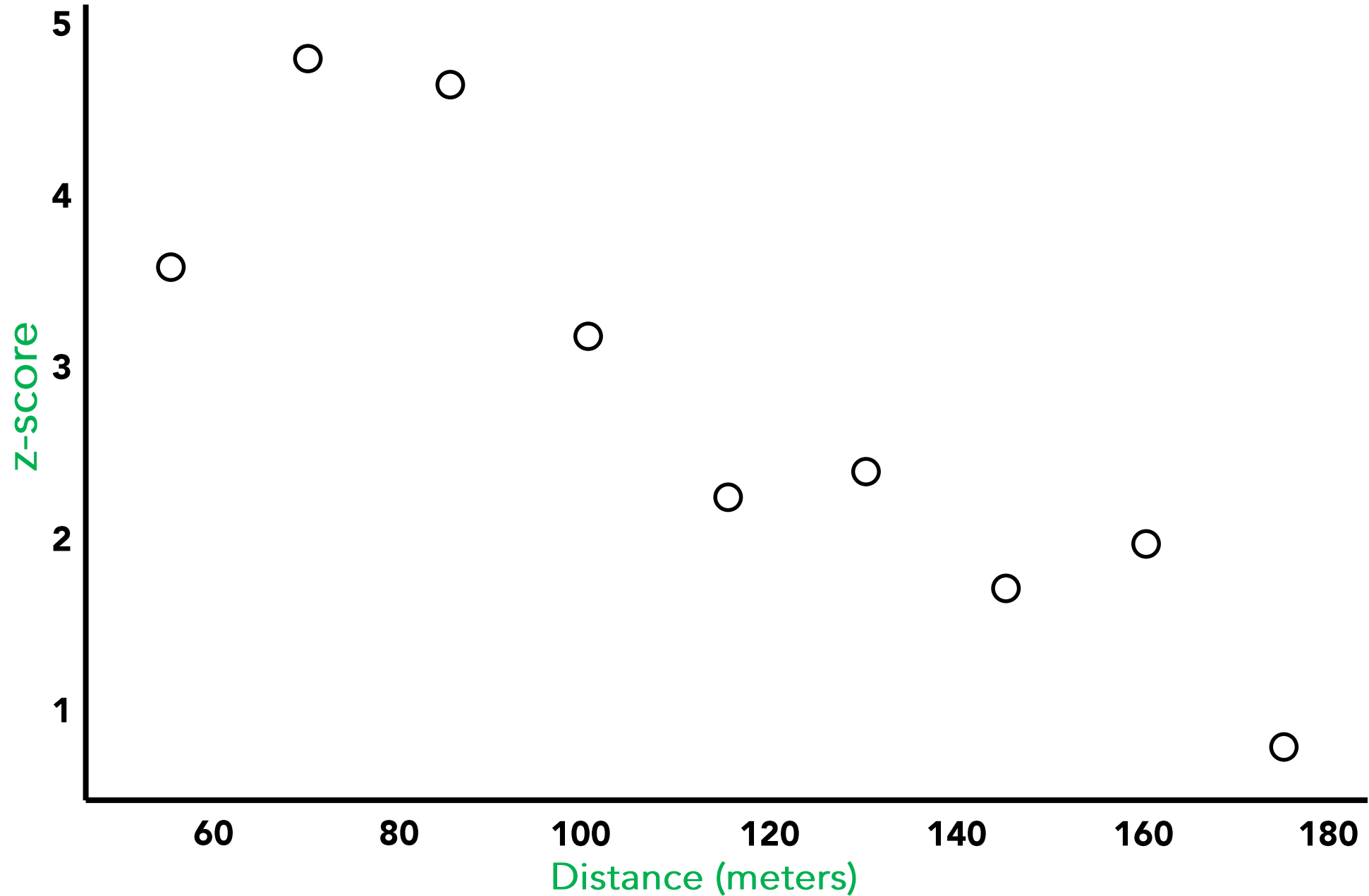
Fixed Distance Band



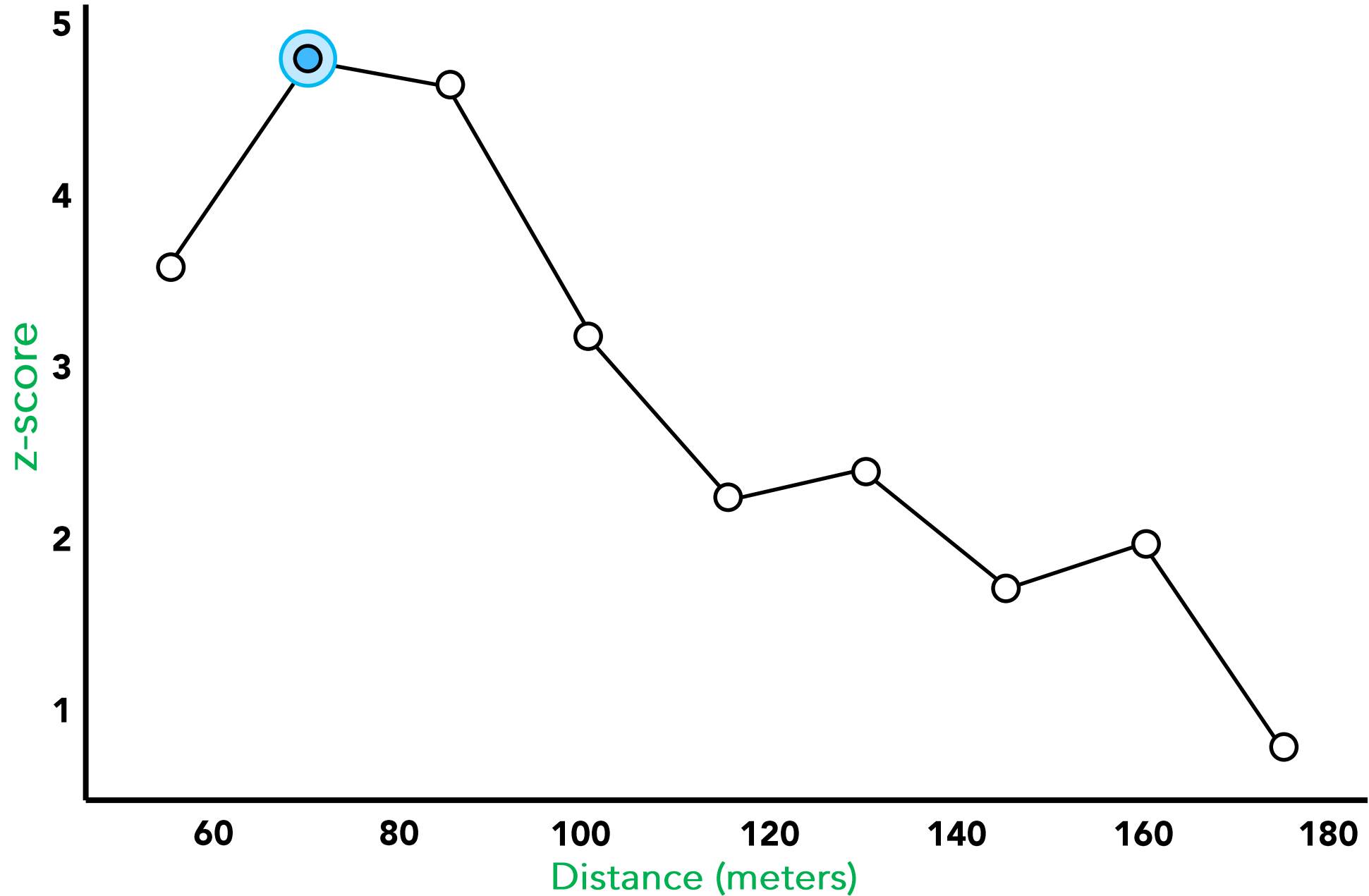
Fixed Distance Band



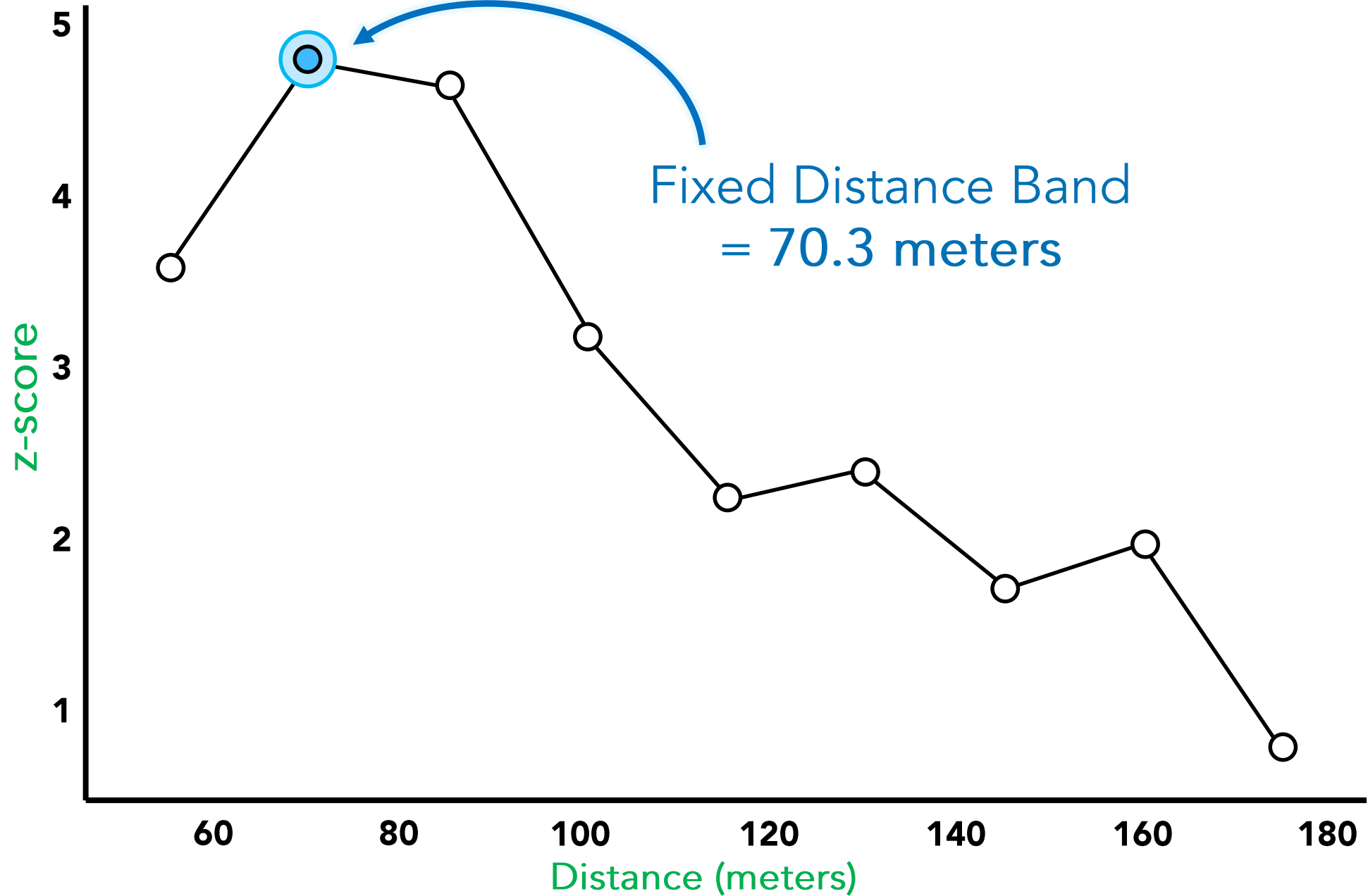
Spatial Autocorrelation by Distance

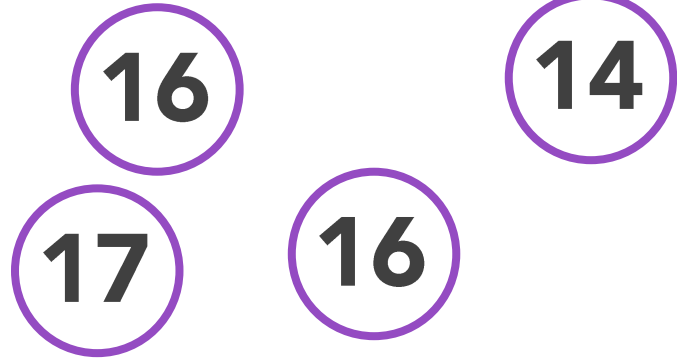


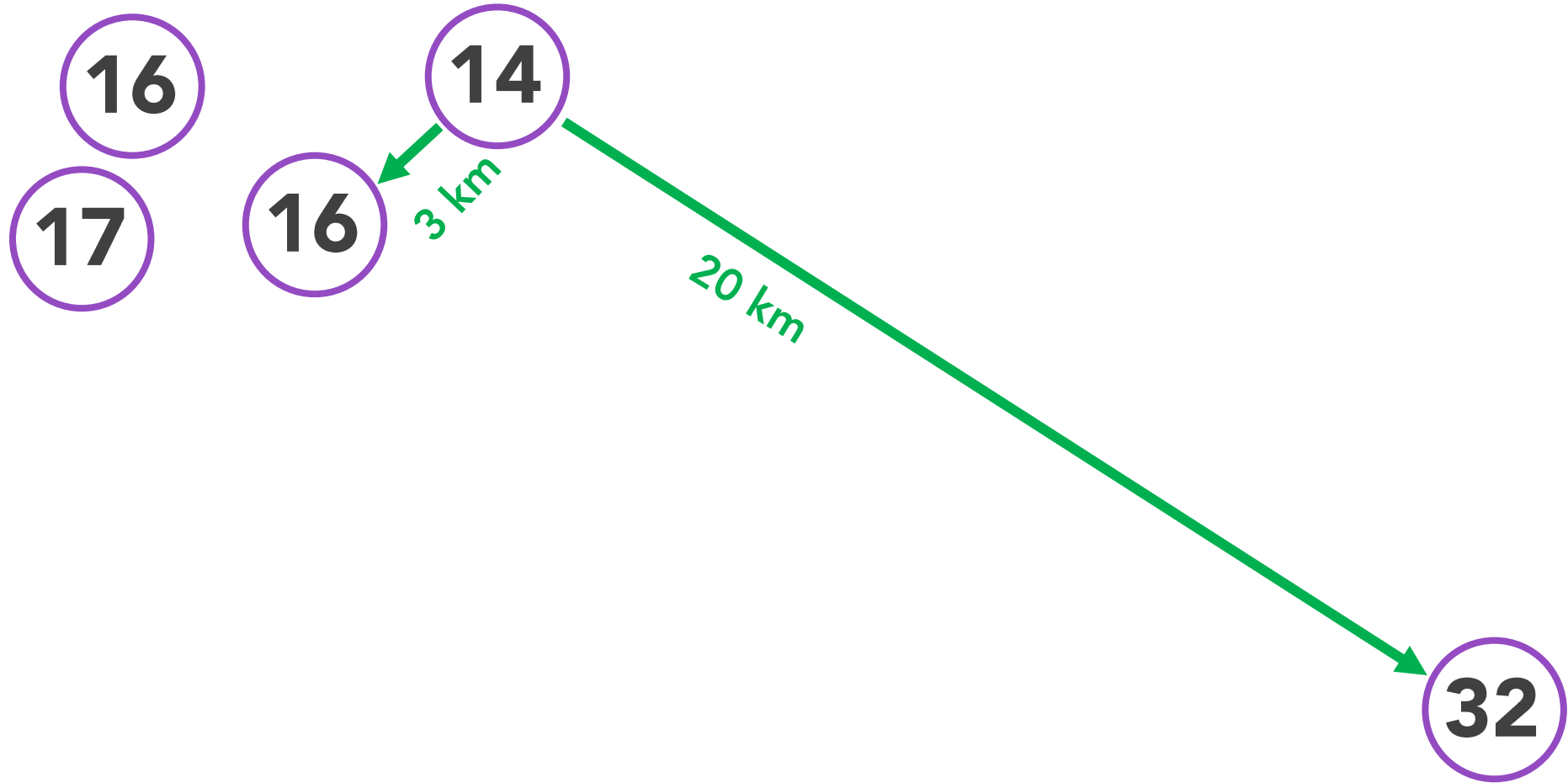
Spatial Autocorrelation by Distance

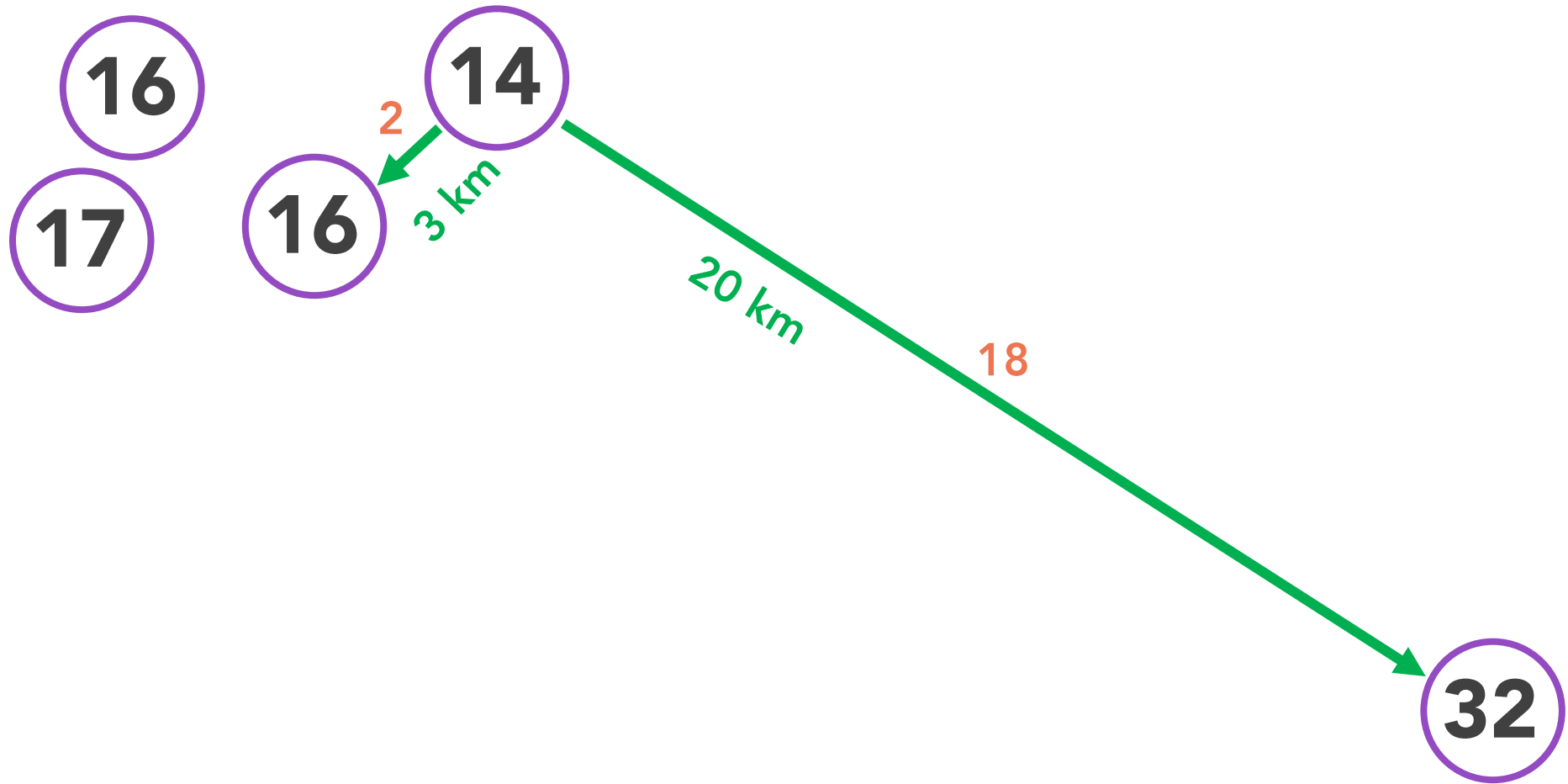


Spatial Autocorrelation by Distance

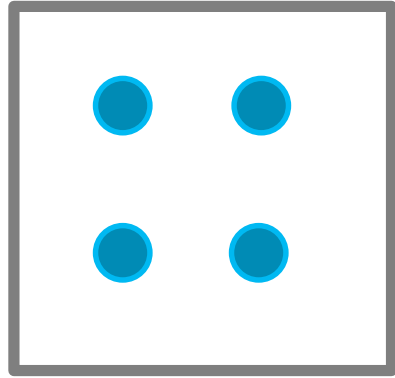




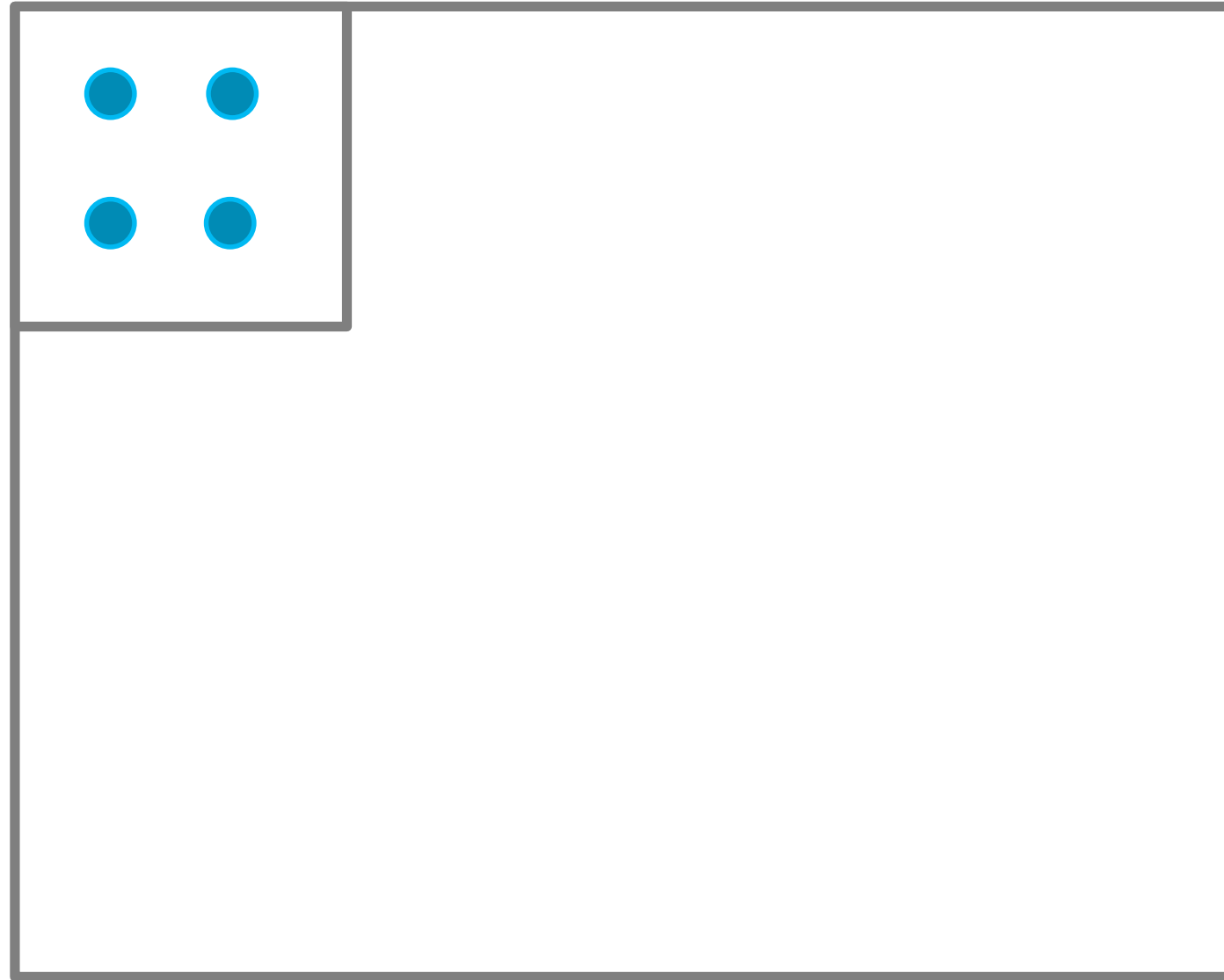




dispersed

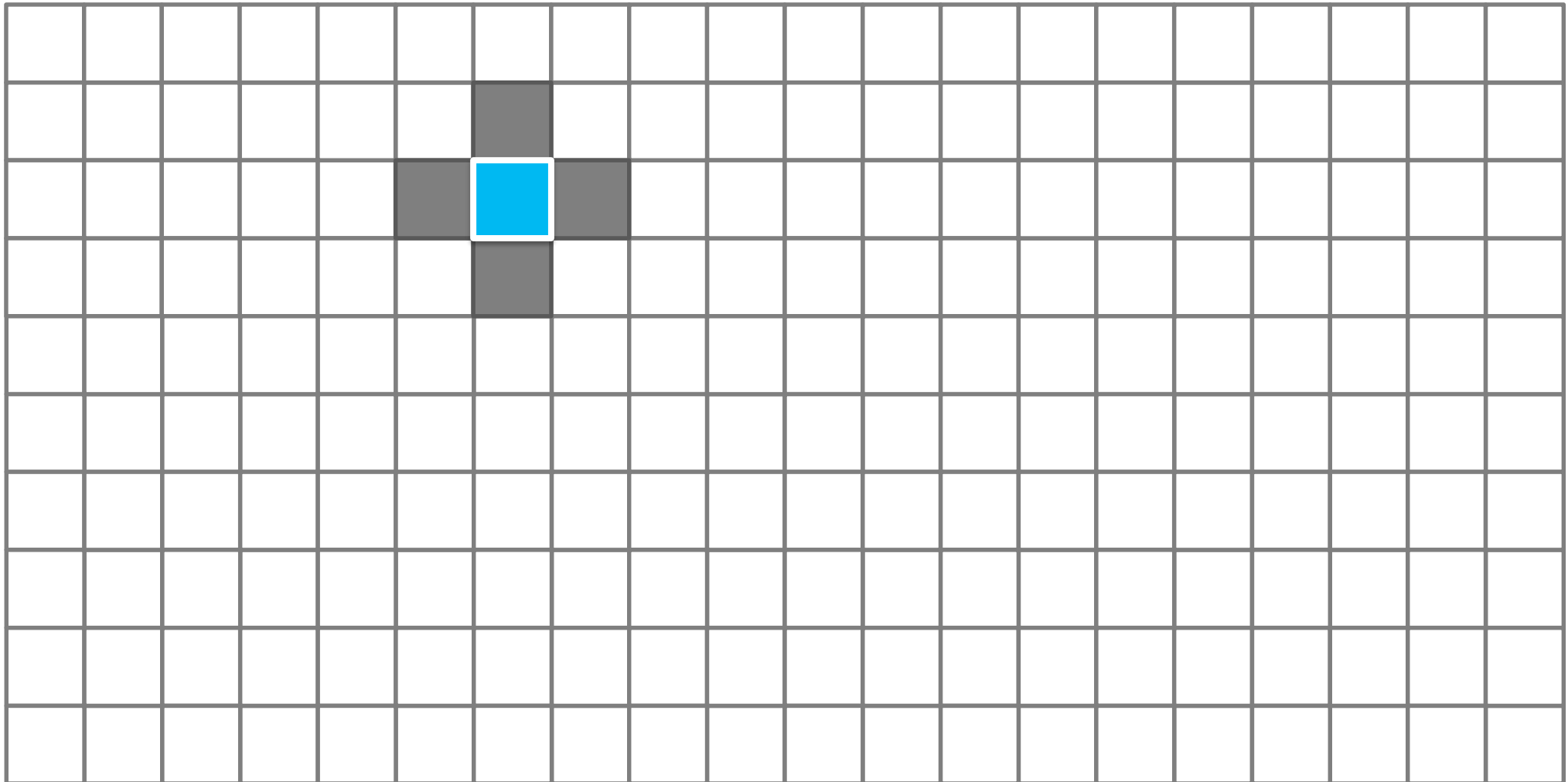


clustered



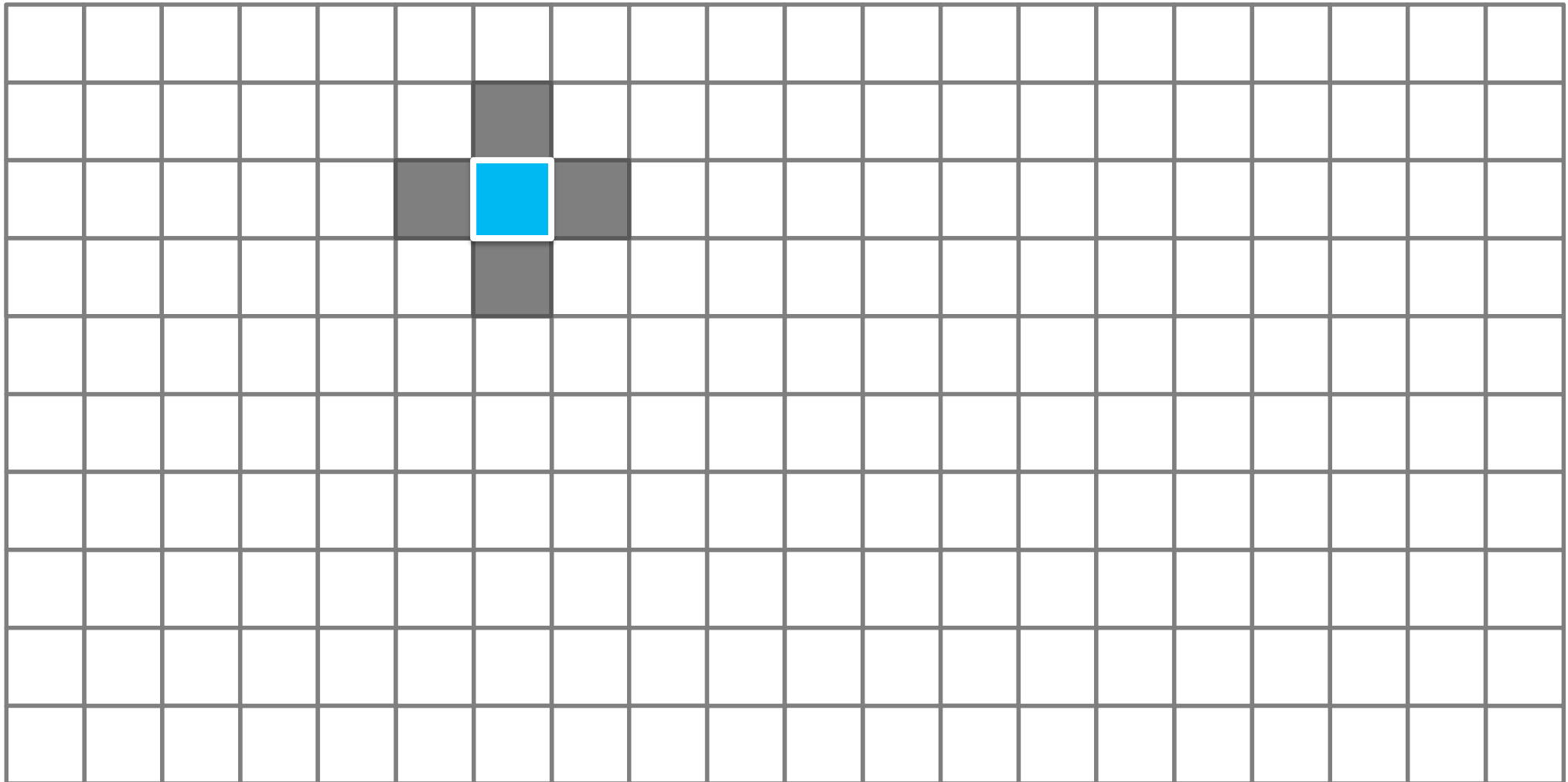
Contiguity

Edges



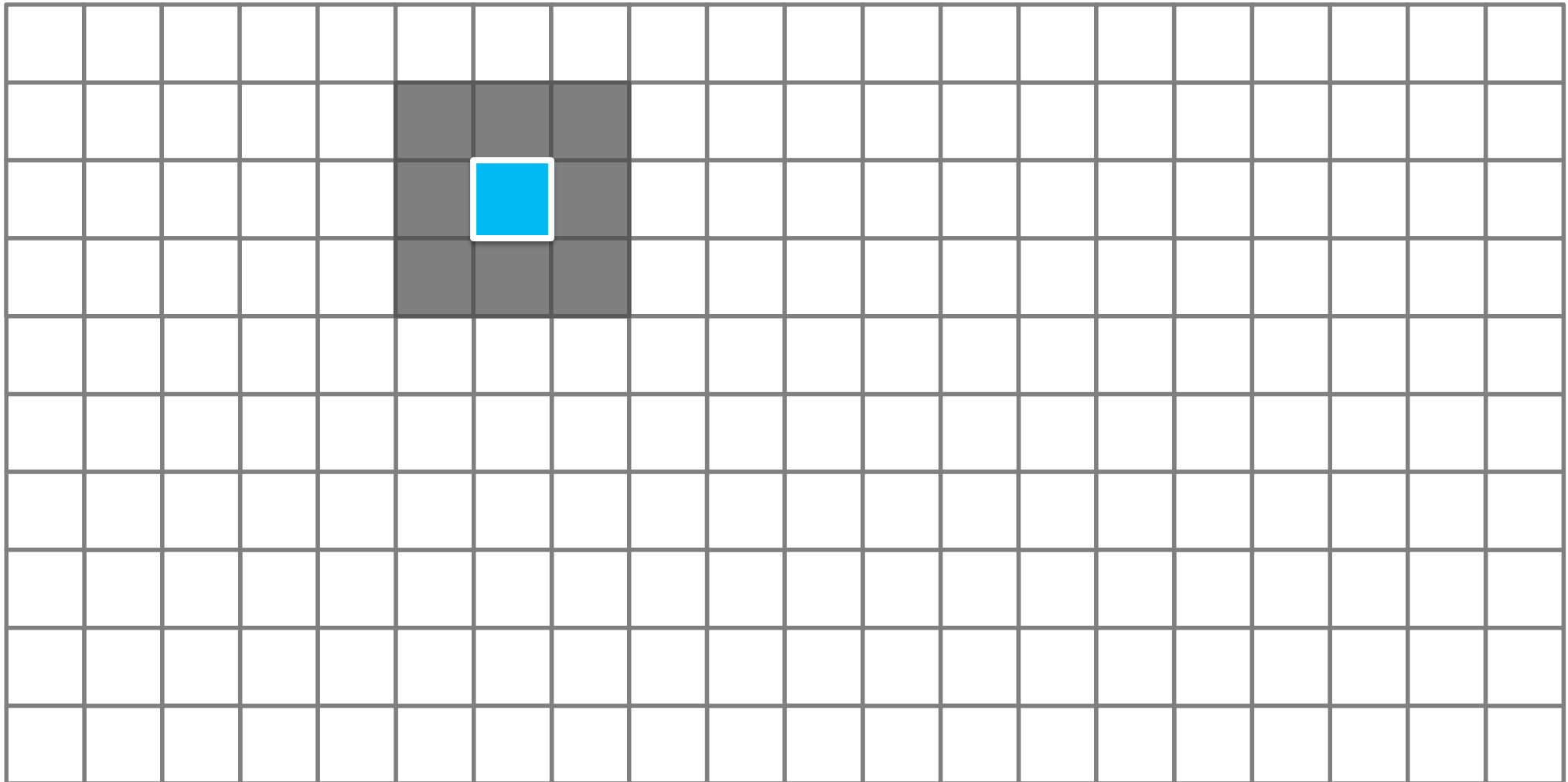
Contiguity

Rook's Case



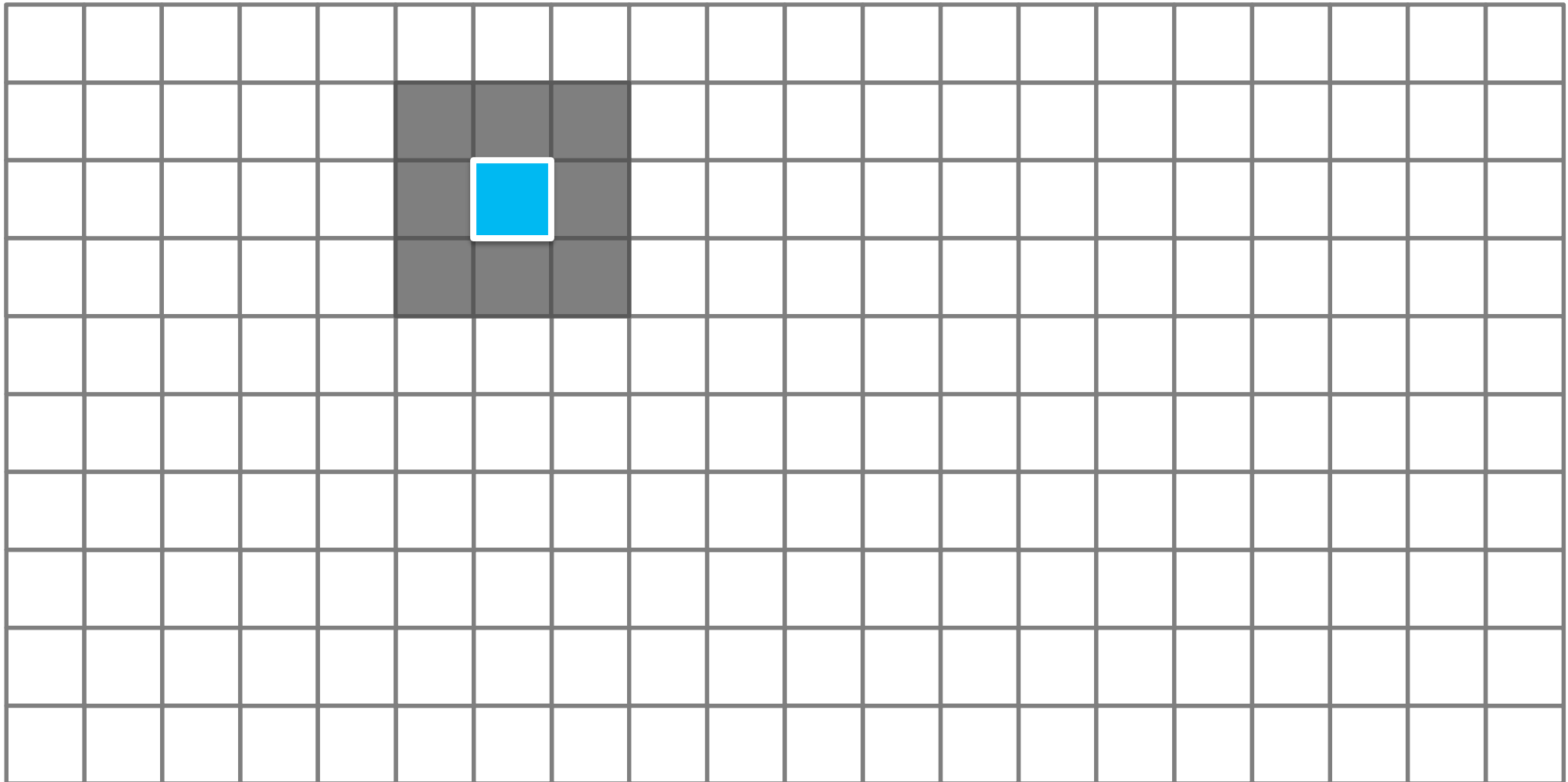
Contiguity

Edges/Corners

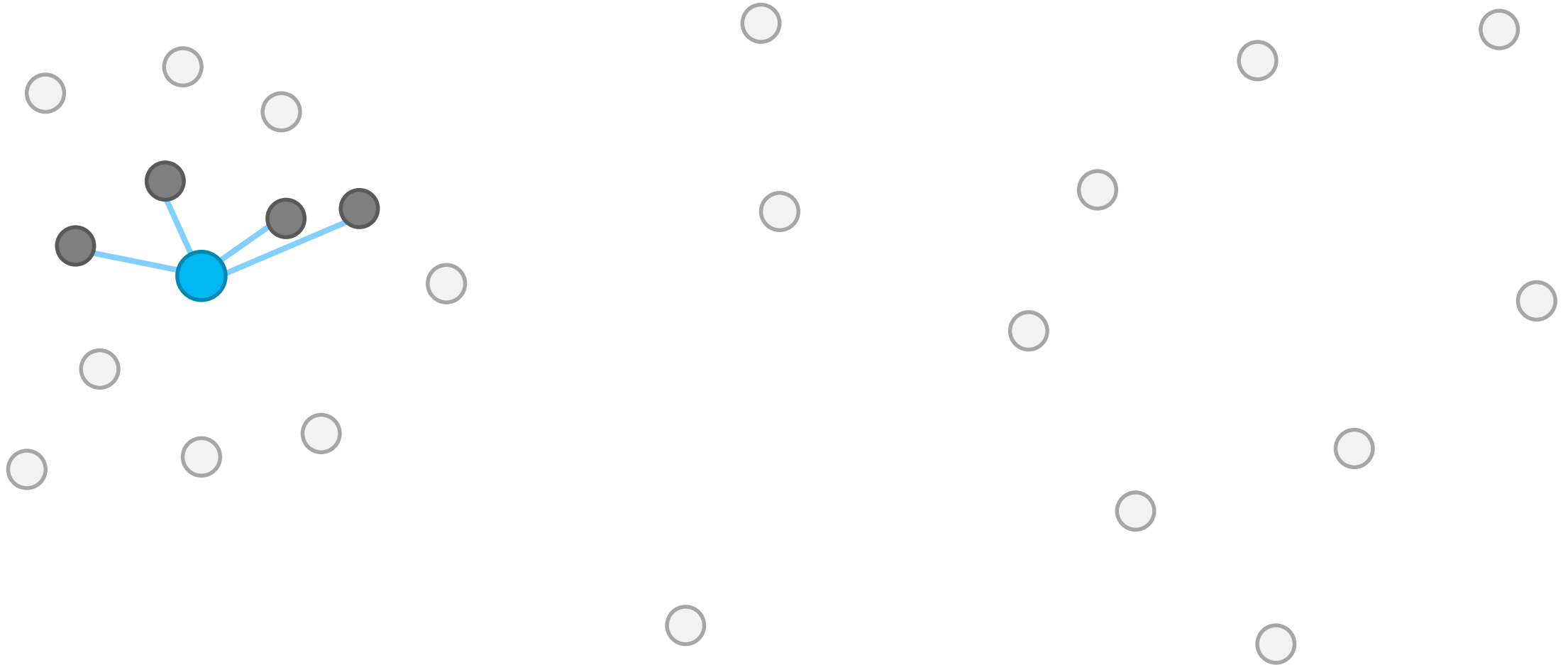


Contiguity

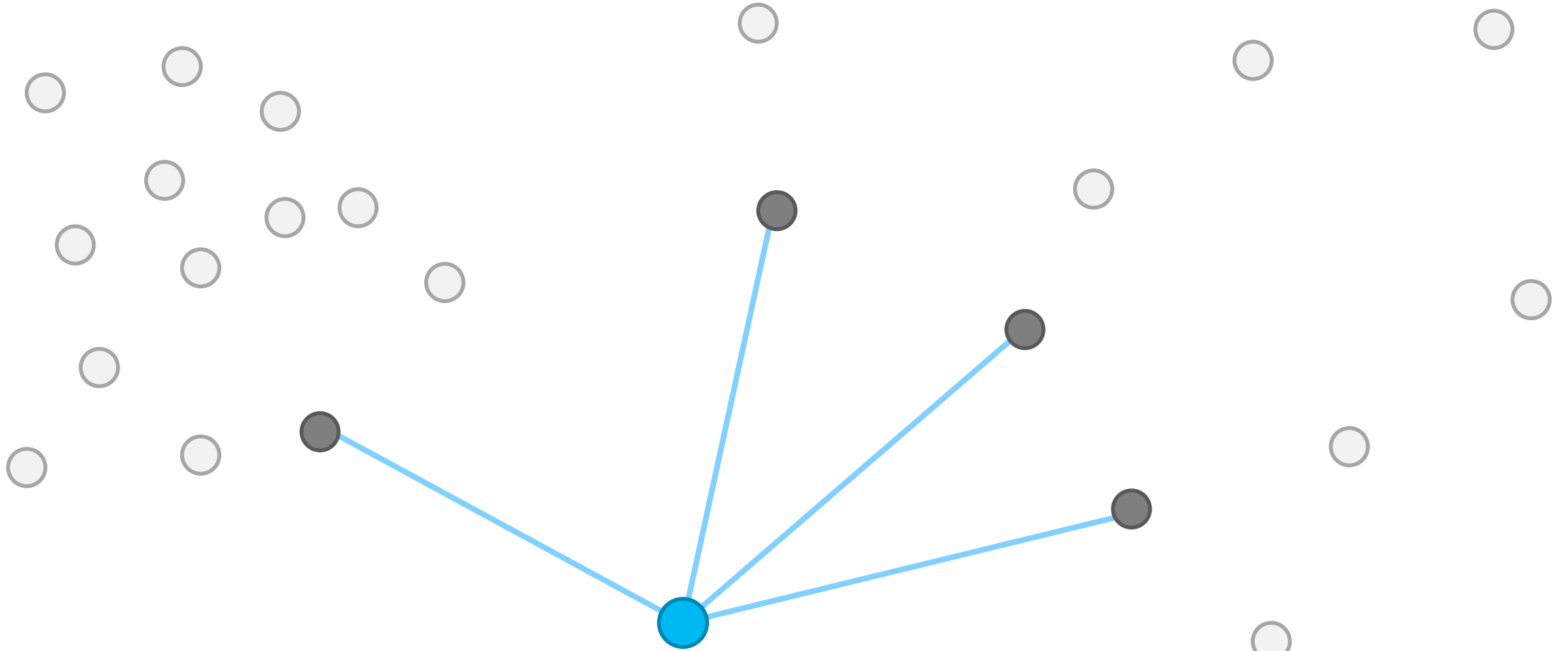
Queen's Case



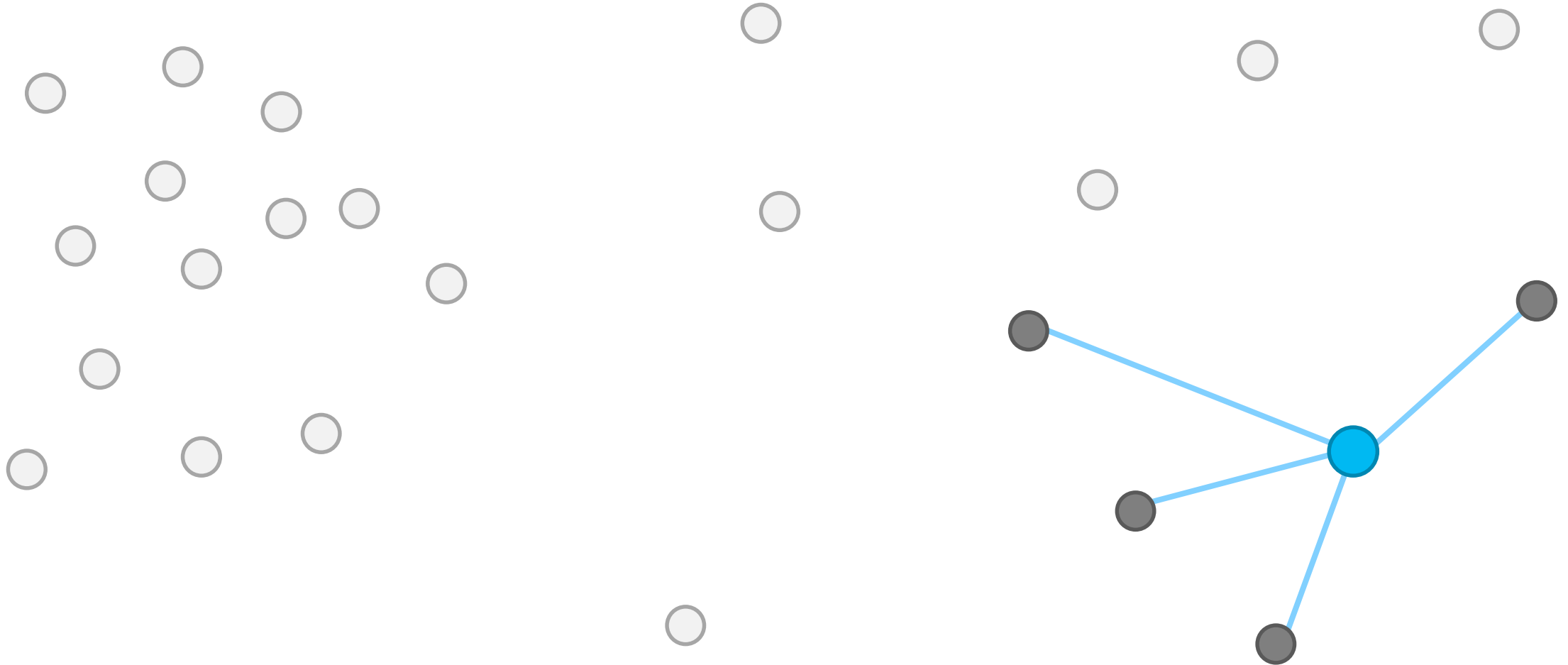
K Nearest Neighbors



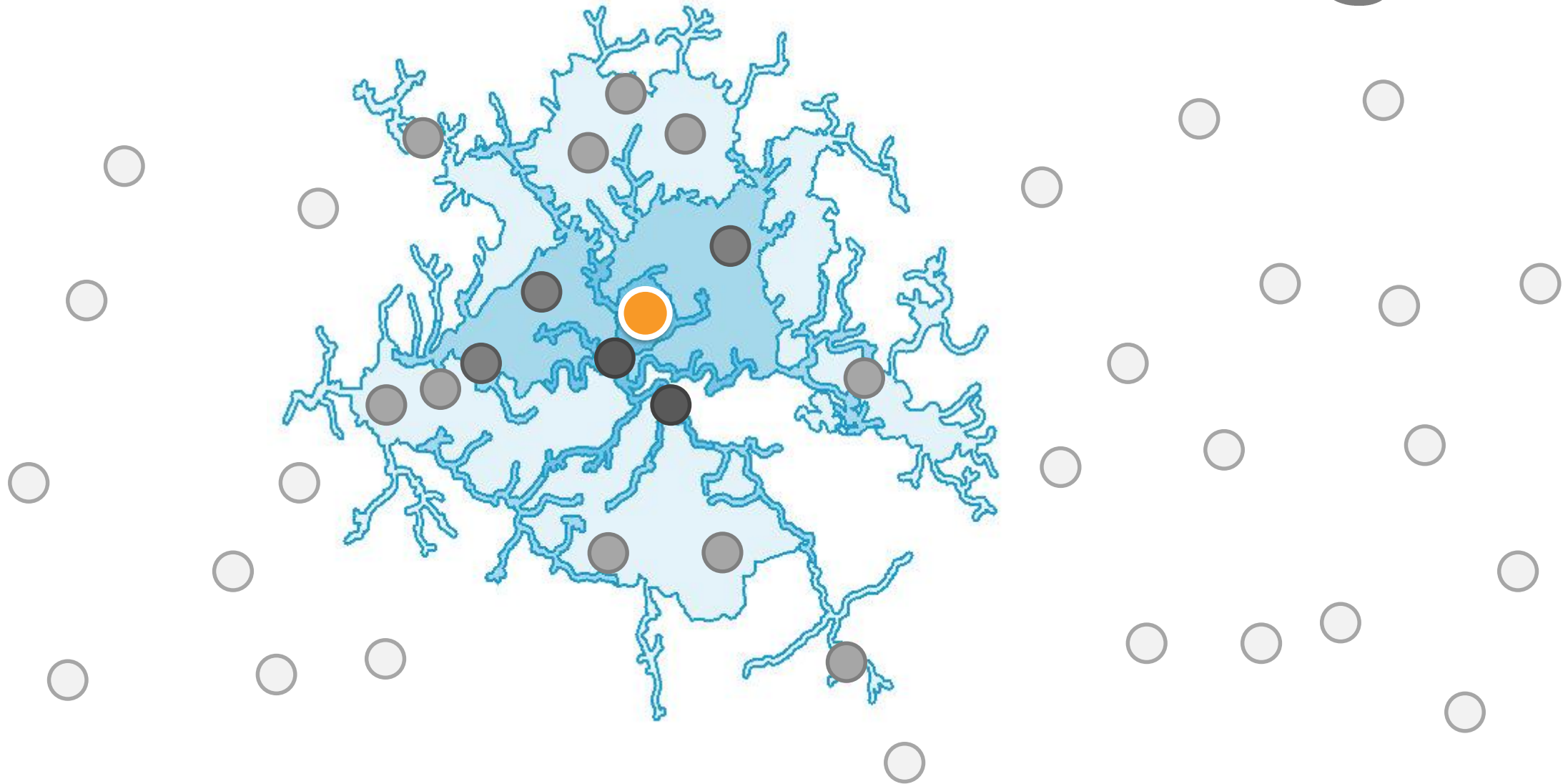
K Nearest Neighbors



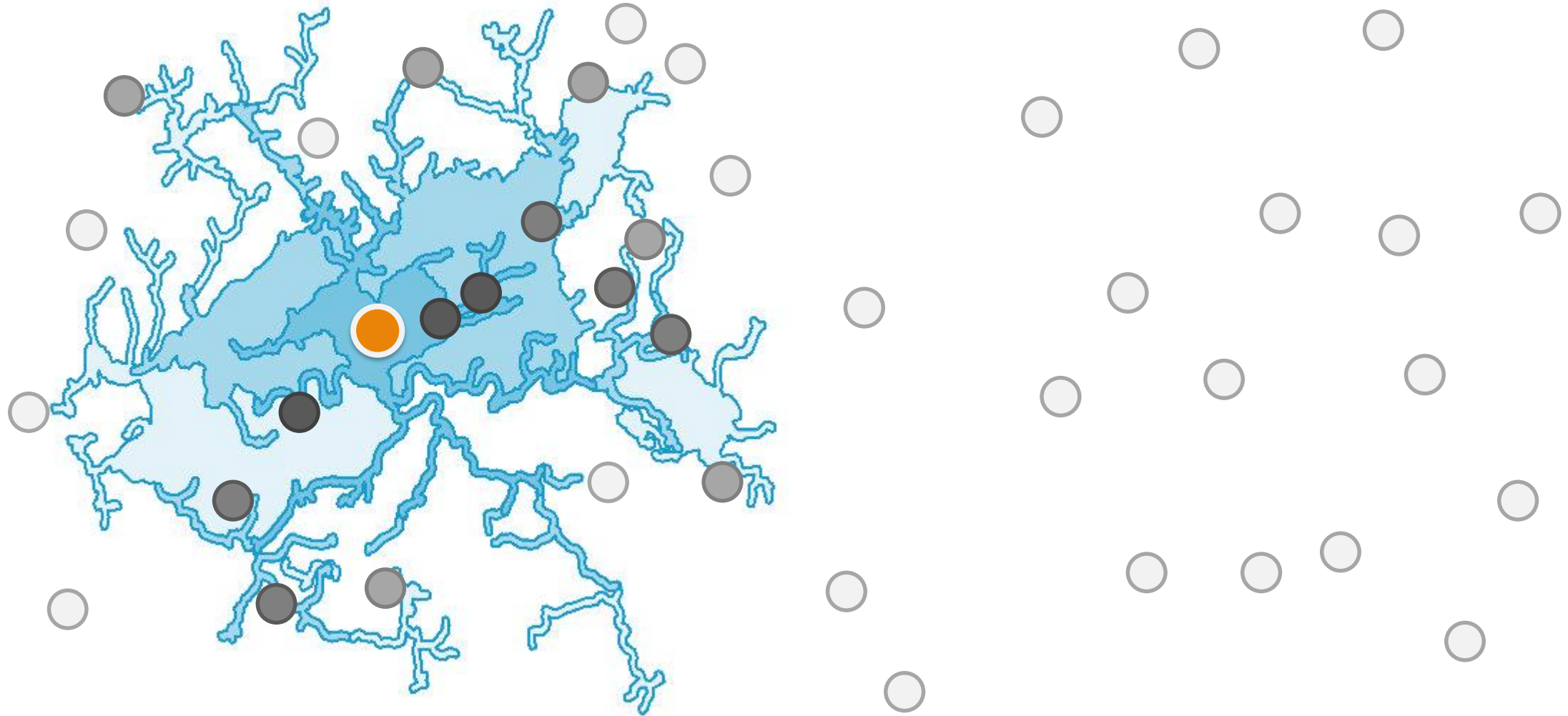
K Nearest Neighbors



Network Spatial Weights



Network Spatial Weights

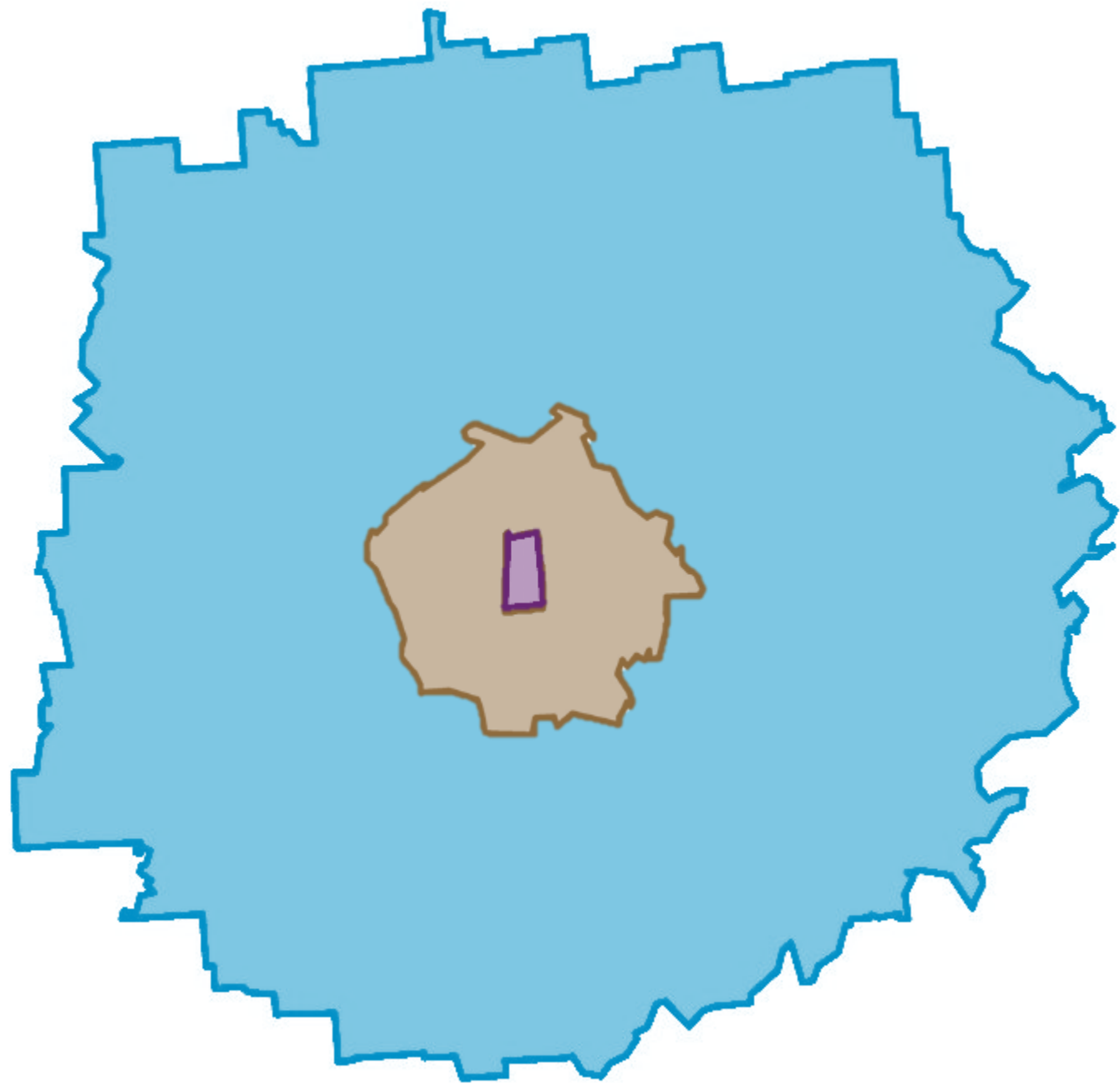


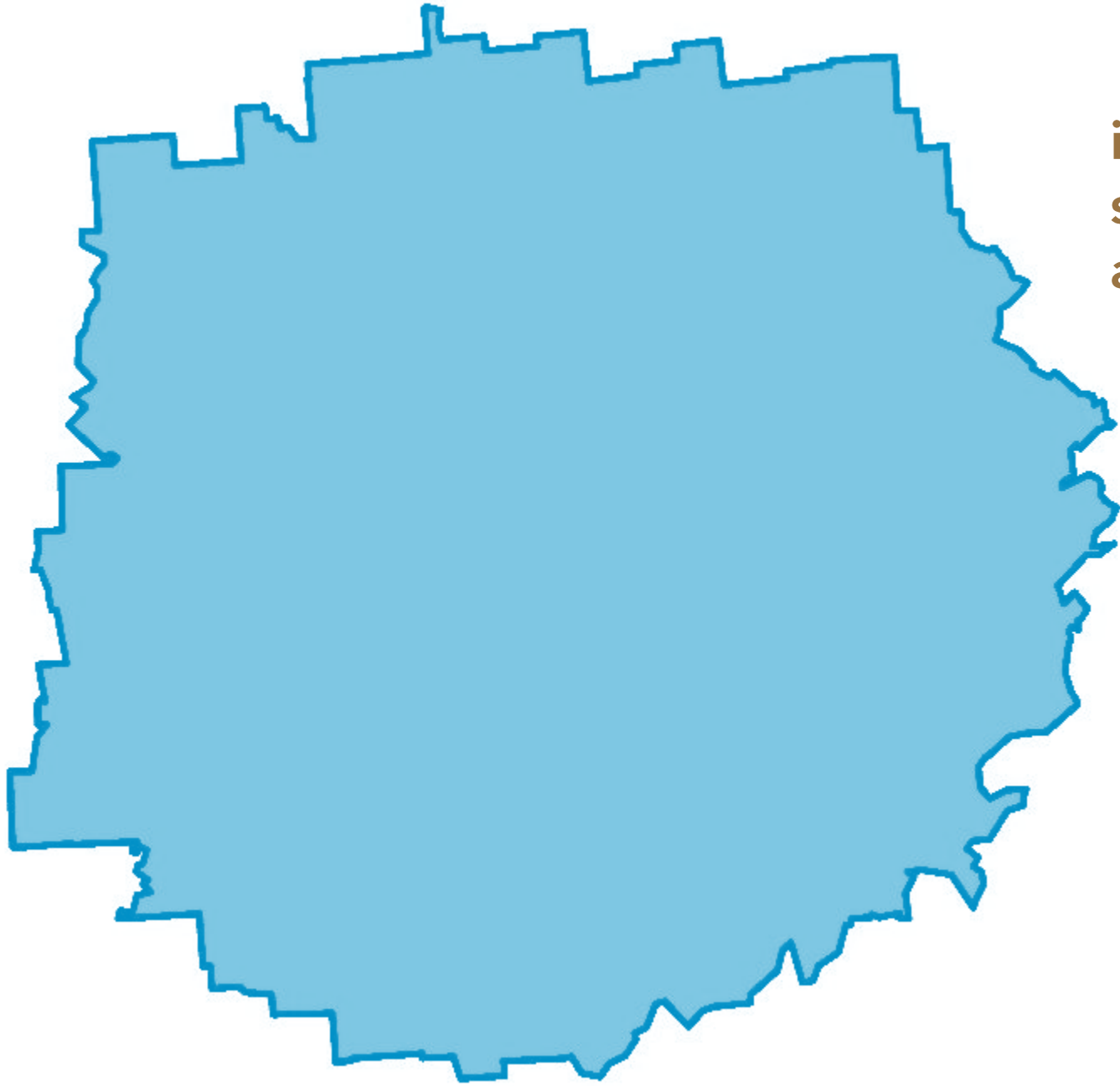
demo



Cluster and Outlier Analysis

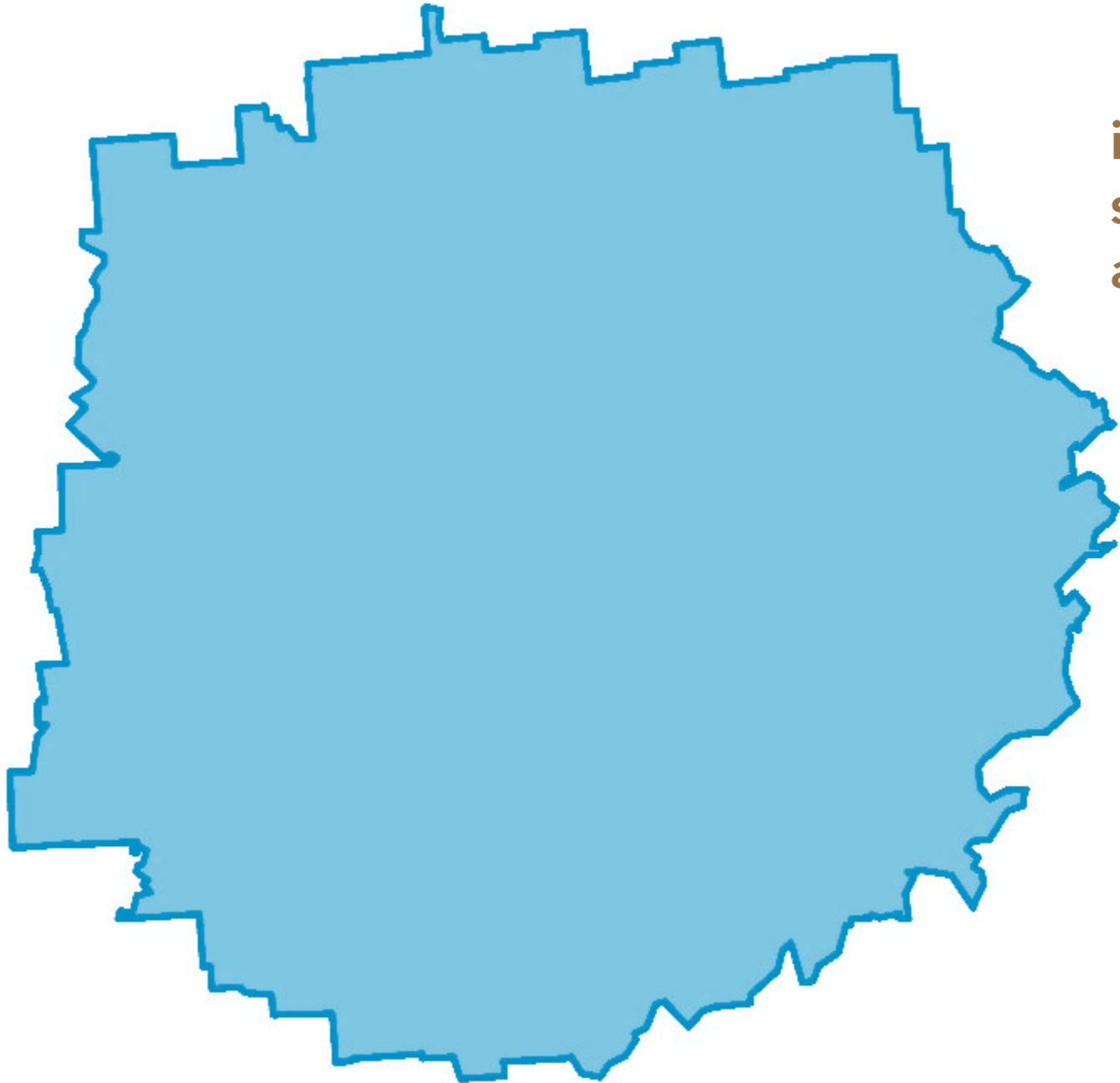
given a set of weighted features, identifies statistically significant hot spots, cold spots, and spatial outliers using the Anselin Local Moran's I statistic



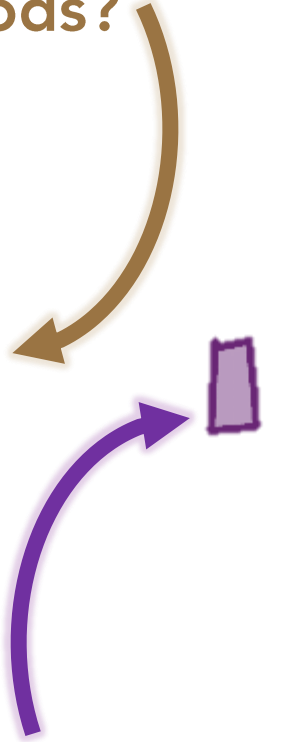


is the neighborhood significantly different from all other neighborhoods?

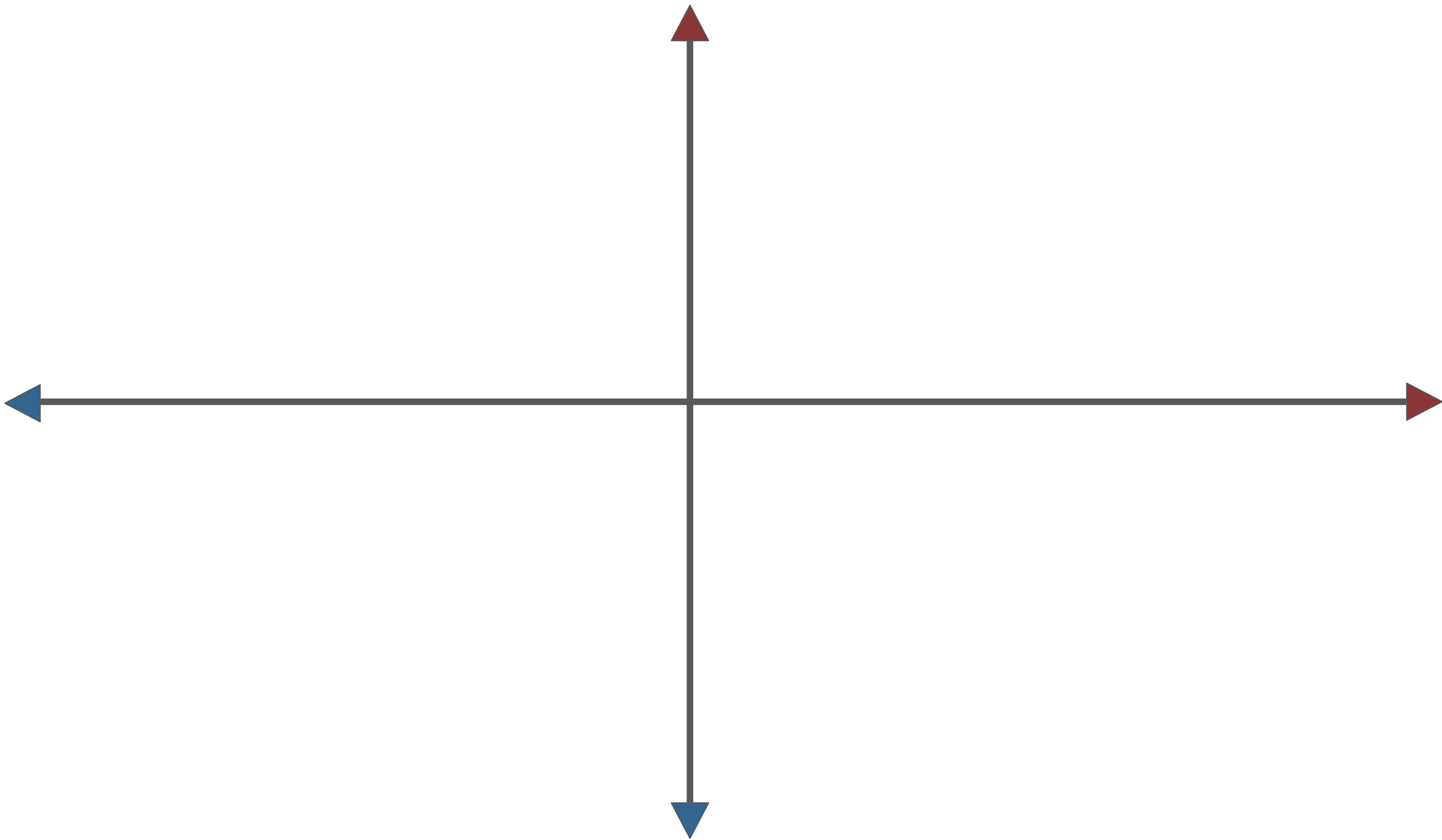


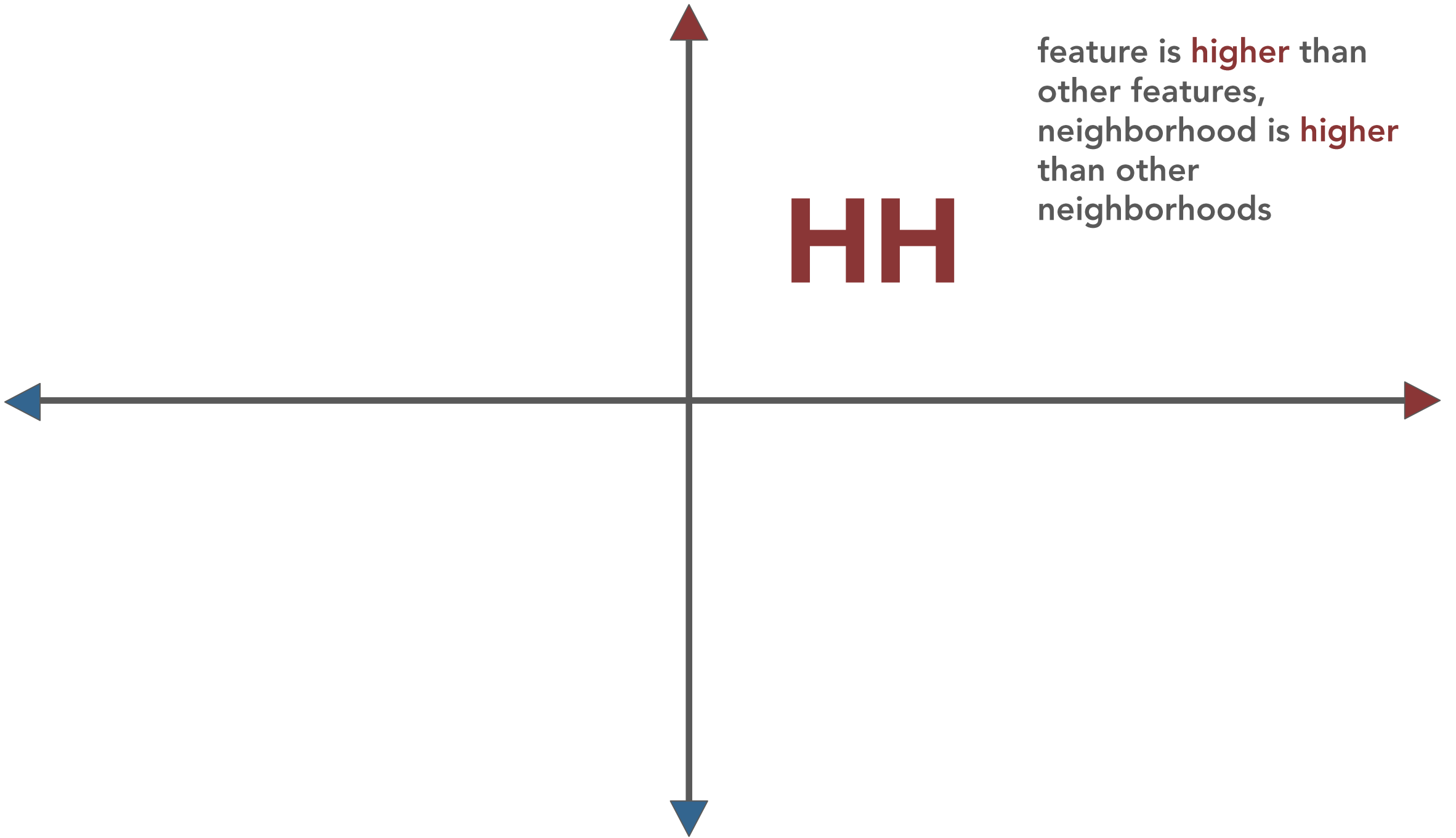


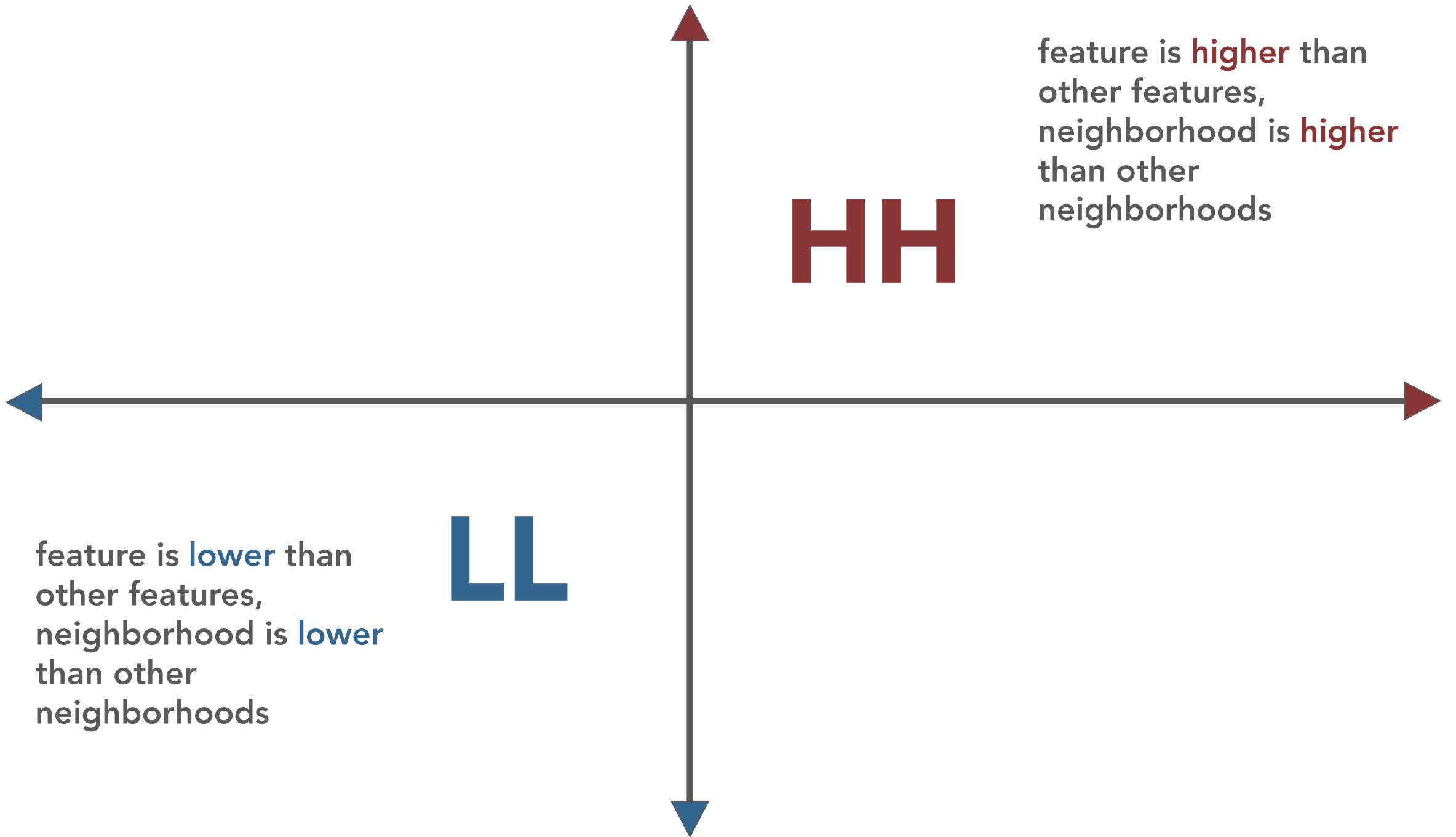
is the neighborhood significantly different from all other neighborhoods?



is the feature significantly different from all other features?







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

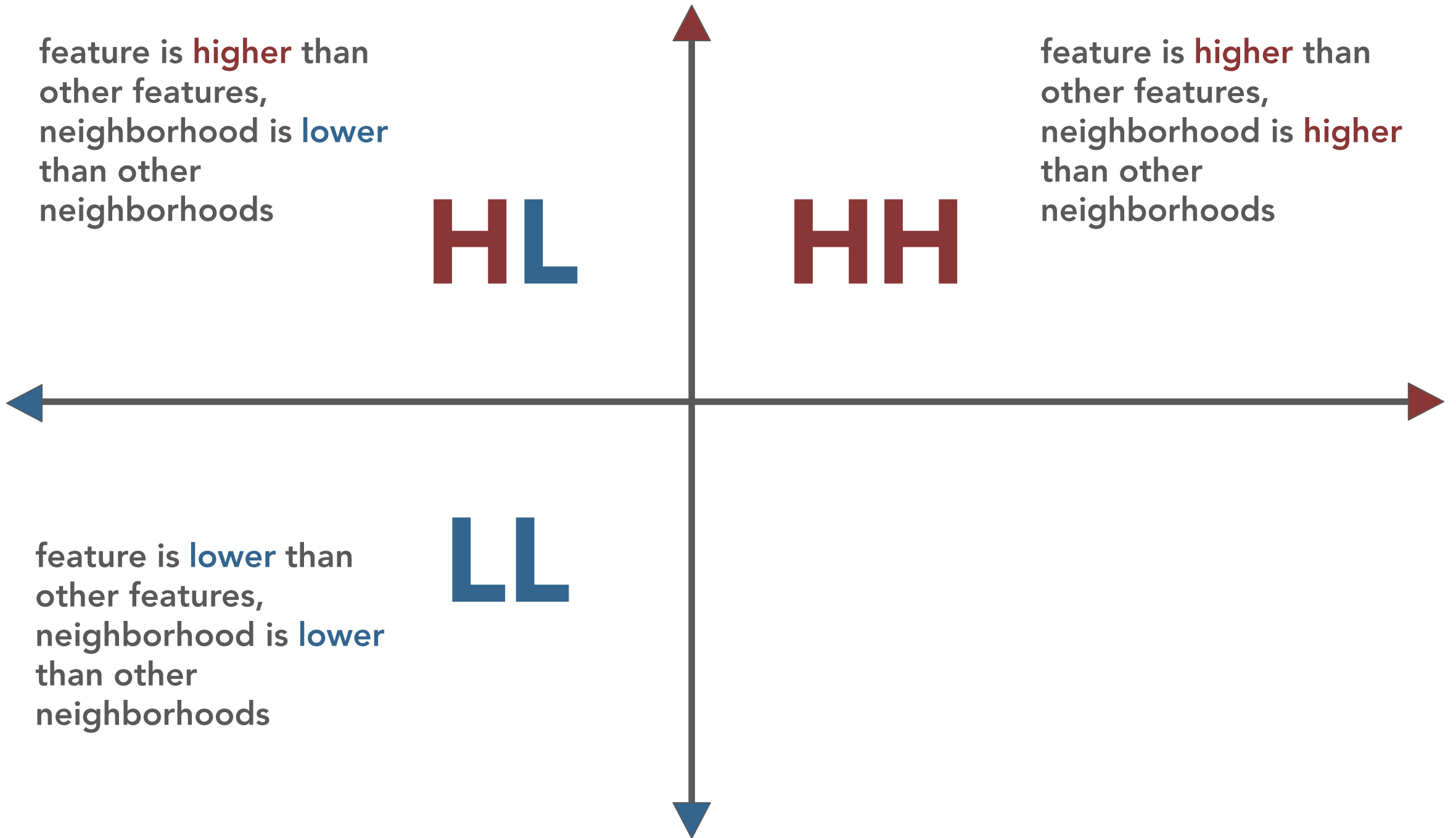
H**L**

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

H**H**

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

L**L**



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

H**L**

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

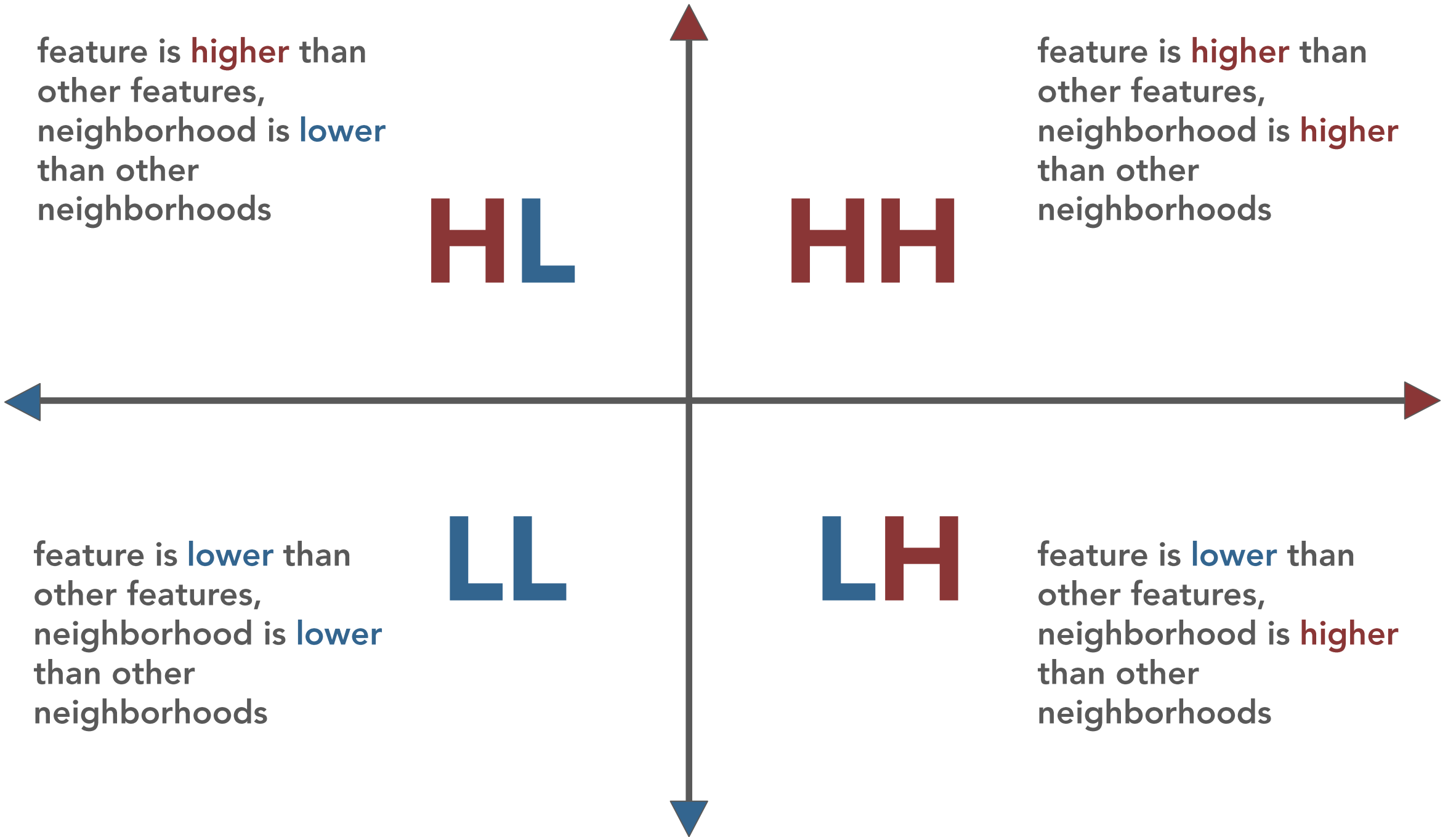
H**H**

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

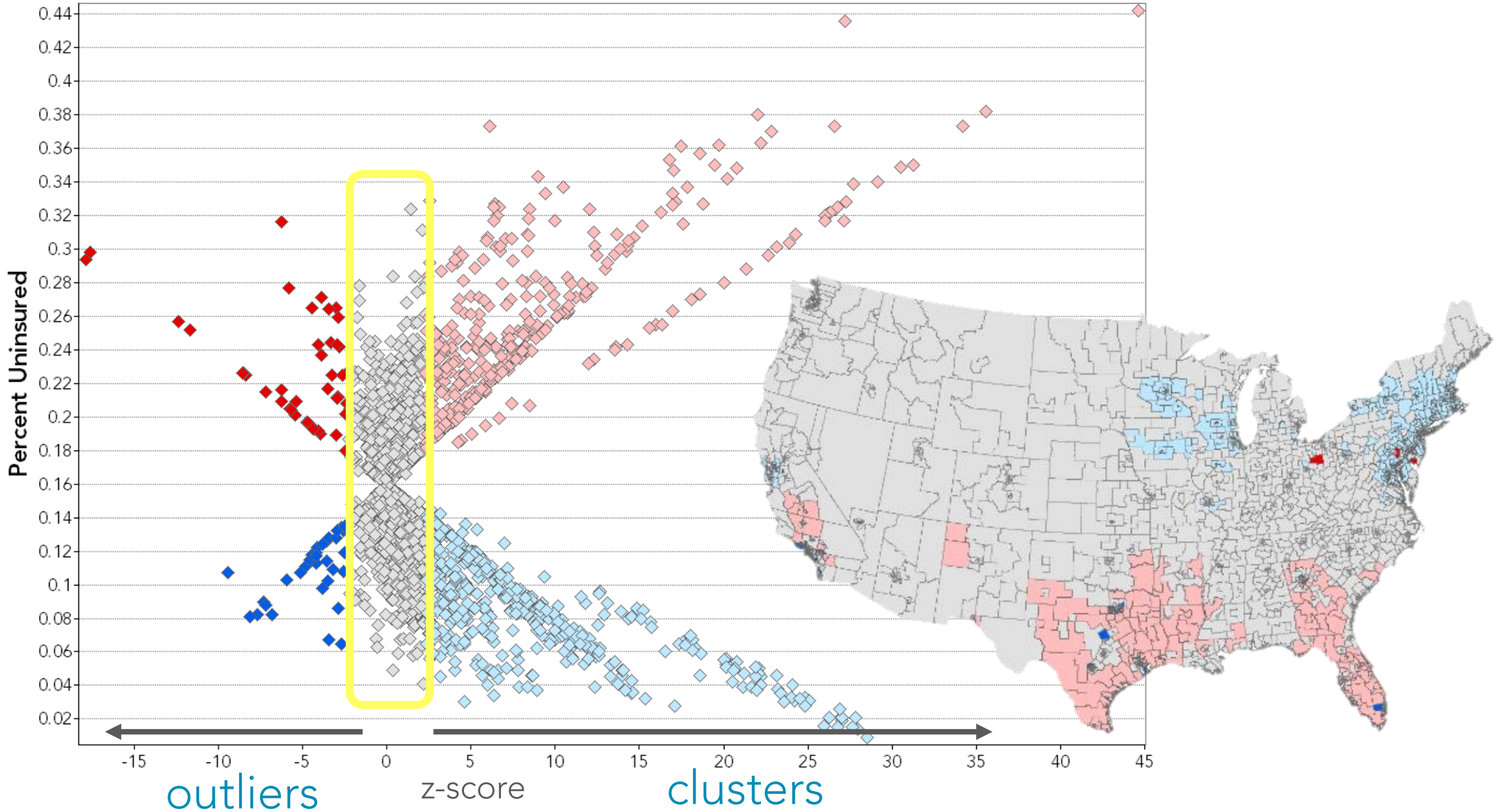
L**L**

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

L**H**



Cluster and Outlier Analysis



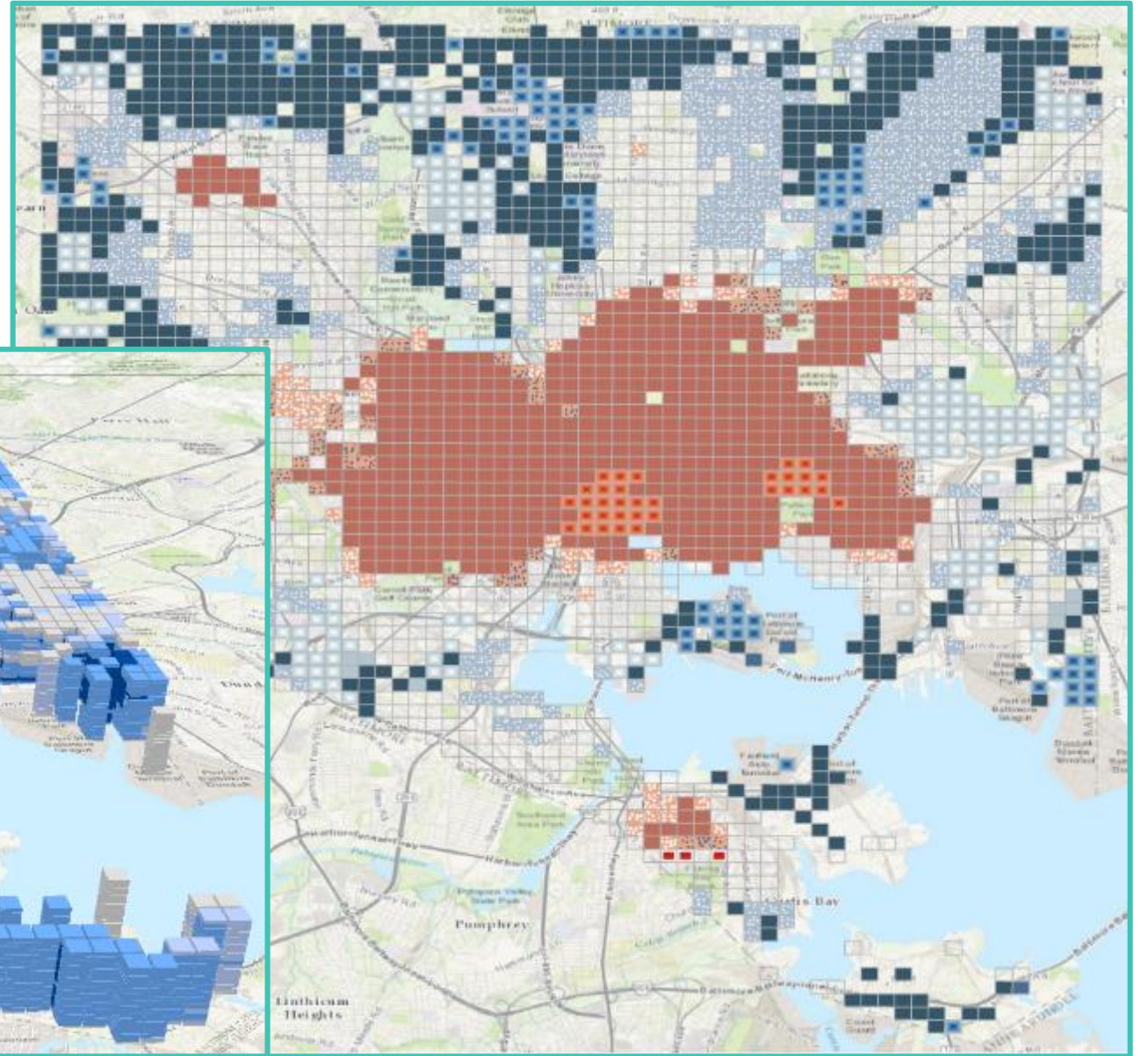
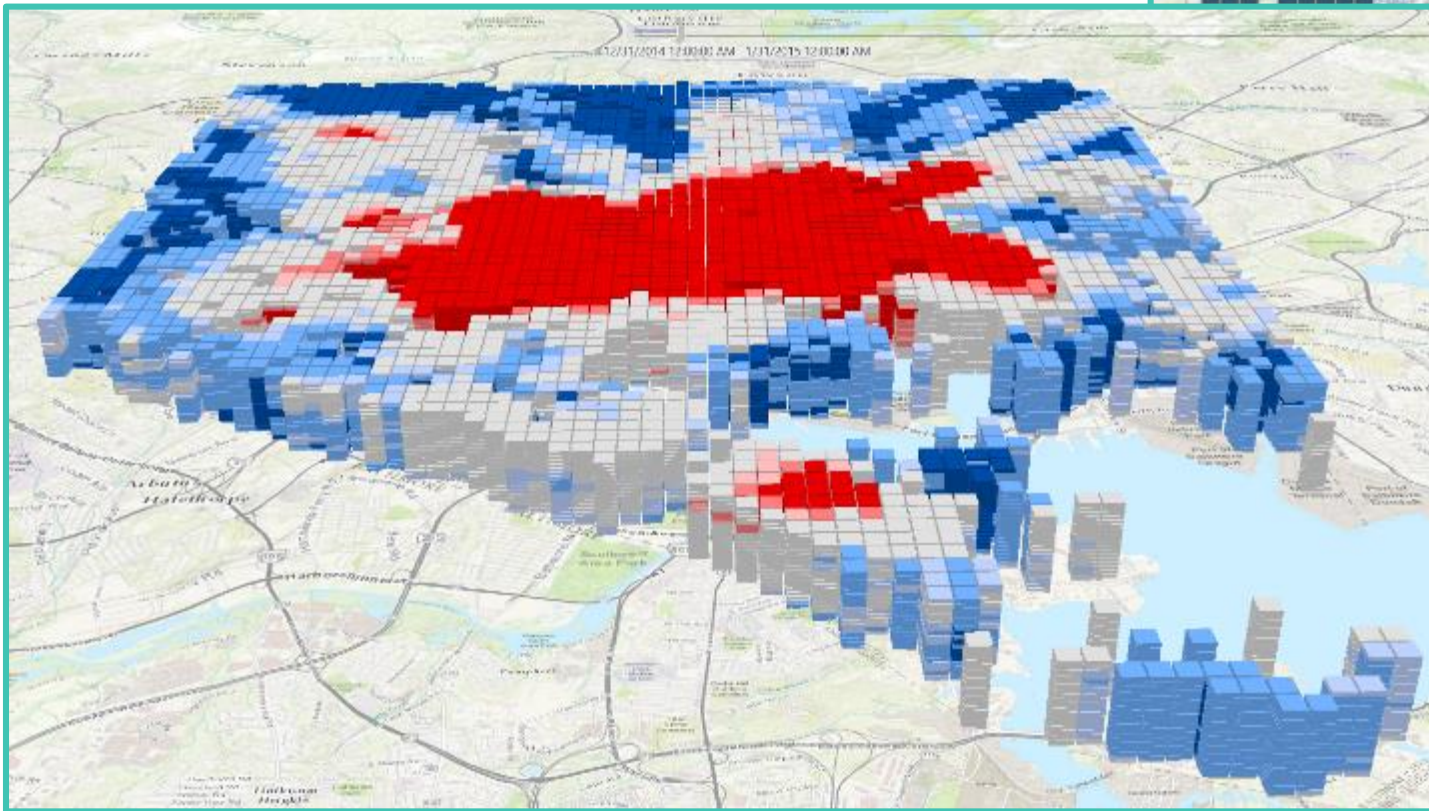
demo



Space Time Pattern Mining

Defining neighbors in space and time

911 Calls





esriurl.com/spatialstats