Data Integration

Class #9 | GEOG 510
GIS & Spatial Analysis in Public Health

Varun Goel

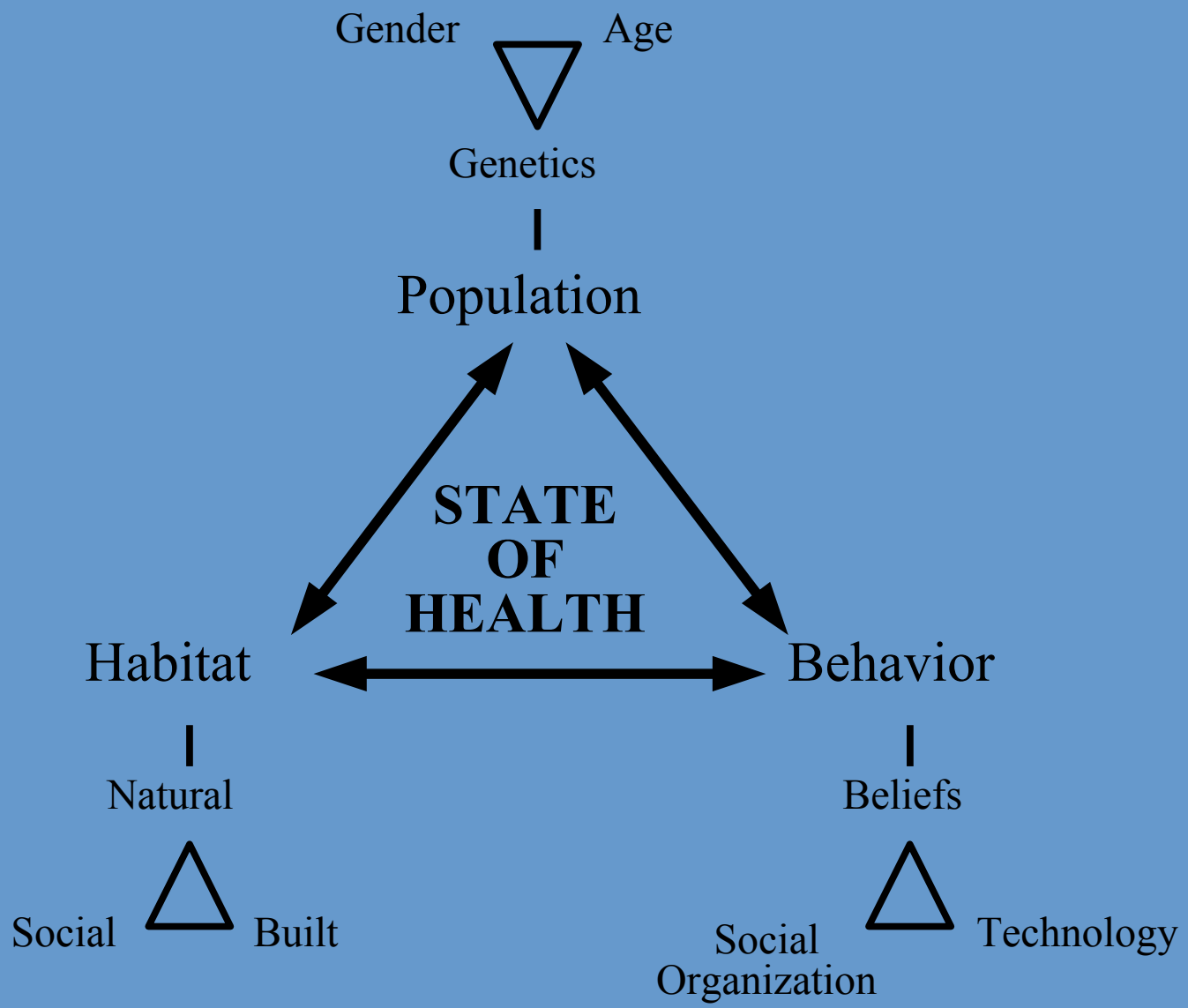
Outline

- Complexity of health-related research
- Multiple data sources
- Raster resampling
- Centroids

Health-related Research

- Whether the focus is morbidity, mortality, infectious or chronic disease, health care…
  - …the outcomes we are analyzing generally arise from a complex set of interacting factors

Triangle of Human Ecology



Health-related Research

- GIS-based analyses are generally very data dependent
  - Spatial (quantitative) data that captures both the outcomes and the factors influencing the outcomes
  - Rarely is all of this data available from the same repository, as spatial data, in the correct format, and/or ready to use

Multiple Data Sources

- When we download/acquire data from multiple sources to be used together in an analysis, what characteristics of the data should we consider?

Multiple Data Sources

- Time period
- Accuracy and precision
  - Both spatial features and attribute values!
- Logical consistency
  - Not within layers, but among

Data Integration

- In my mind, this means preparing data to be GIS analysis ready
  - For example,
    - Projection (reprojection)
    - Subset (spatial, attribute)
    - Table join, Field calculator
    - Aggregate
    - Consistency / validity
    - Resolution (raster)
    - Centroids

Data Integration

- Why discuss?
  - Your analysis will likely get messy
    - You will forget things you did last week (or last month)
  - Creating an "analysis" dataset (stored in one location) will save you time
    - Rather than having to redo steps or search for files

Projections

- Good practice: normalize your data to same coordinate system
- Operations in ArcGIS
  - Project (reproject)
    - Converts data from one coordinate system to another
  - Define Projection
    - Assigns projection information if there is none, or overwrites current projection information

Subset

- Good practice: create set of layers that do not contain extraneous information
  - Spatial extent
    - Clip to a single study area (will discuss in overlay)
  - Observations
    - Subset to only those needed (spatial or attribute query)
  - Attributes / fields
    - Remove unnecessary fields from tables

Table Join, Calculate Field

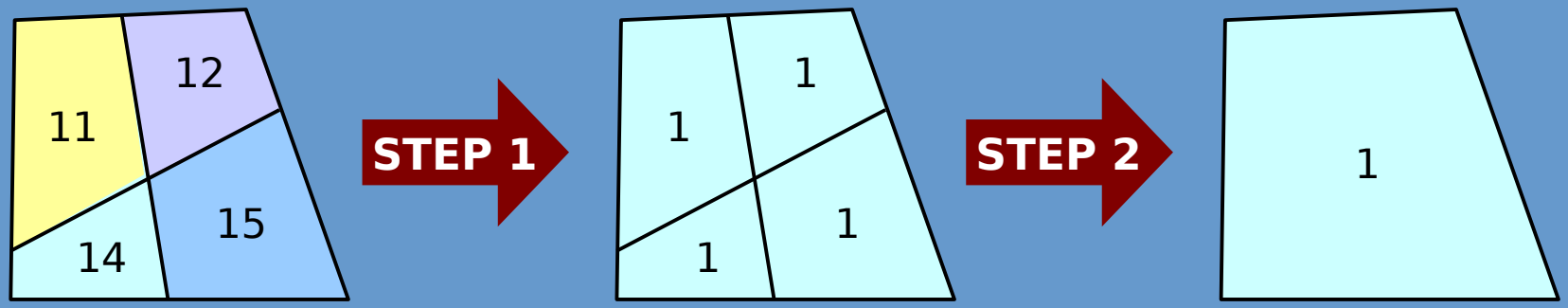
- When data are not mapping or analysis ready
  - For example, case counts by spatial (aggregated) unit
    - Not a good idea to map raw count data
    - Normalize by population size to calculate rate
      - Often requires table join, and then calculation of a new field (complete in lab, will discuss soon)

Aggregate

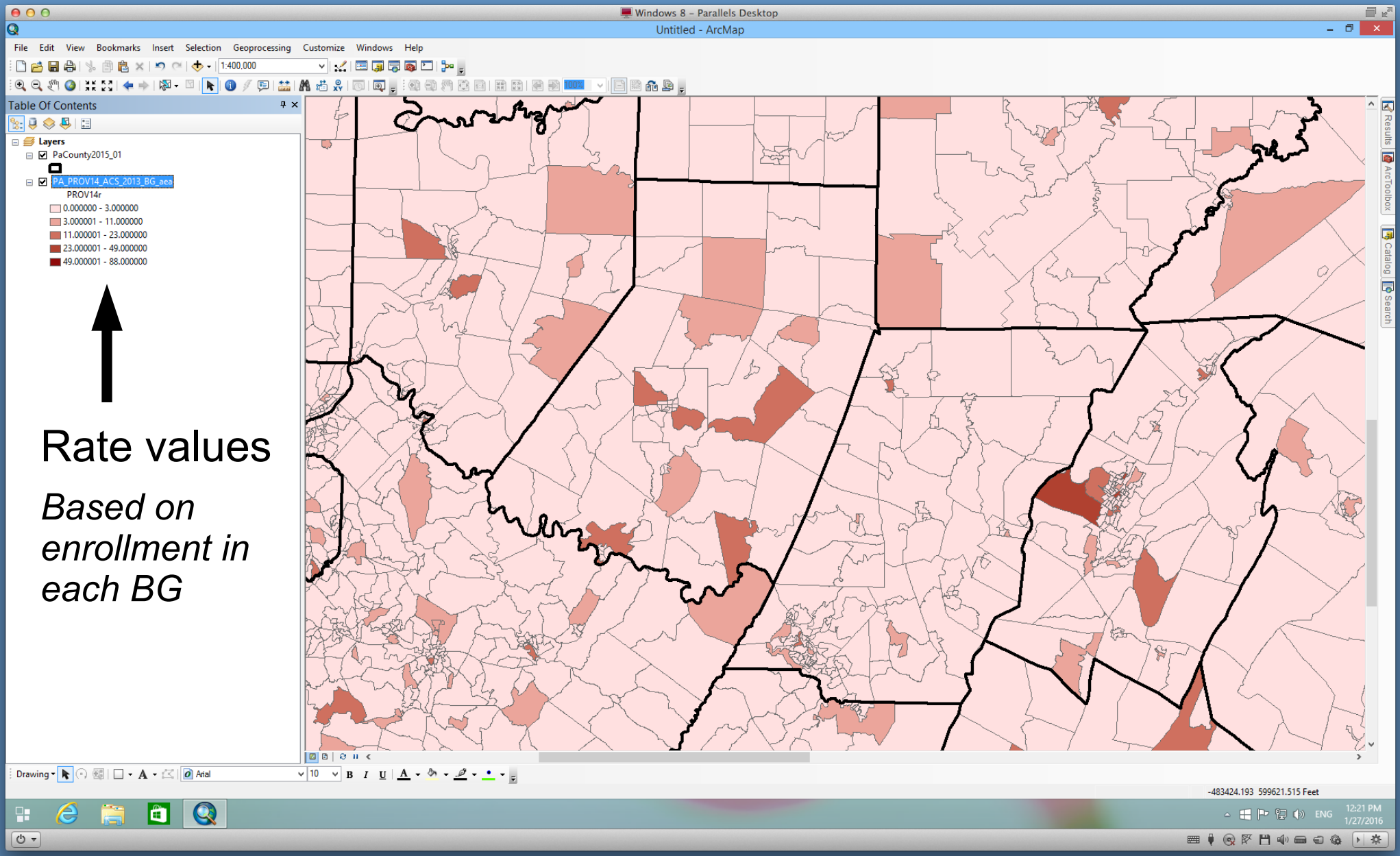
- Generally, you will be bound by the largest aggregation unit
  - e.g., if you have county, census tract, and point data, you likely have to work at the county level
    - However, there are approaches to disaggregate spatial data

Aggregate

- Used to resolve scale mismatches
  - Dissolve (by attribute) in QGIS
    - Creates new (larger) spatial features
    - Be careful!
    - Consider the values you are aggregating!



Example


Windows 8 – Parallels Desktop
Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:400,000
Table Of Contents
Layers
PaCounty2015_01
PA_PROV14_ACS_2013_BG_aea
PROV14r
0.000000 - 3.000000
3.000001 - 11.000000
11.000001 - 23.000000
23.000001 - 49.000000
49.000001 - 88.000000
Rate values
Based on enrollment in each BG
Drawing
Arial
10
-483424.193 599621.515 Feet
ENG
12:21 PM
1/27/2016

Example

BG	PROV RATE
1	10
2	0
3	0
4	12
5	20
6	5
7	0
8	3
9	7
10	11
MEAN	*6.80*

Example

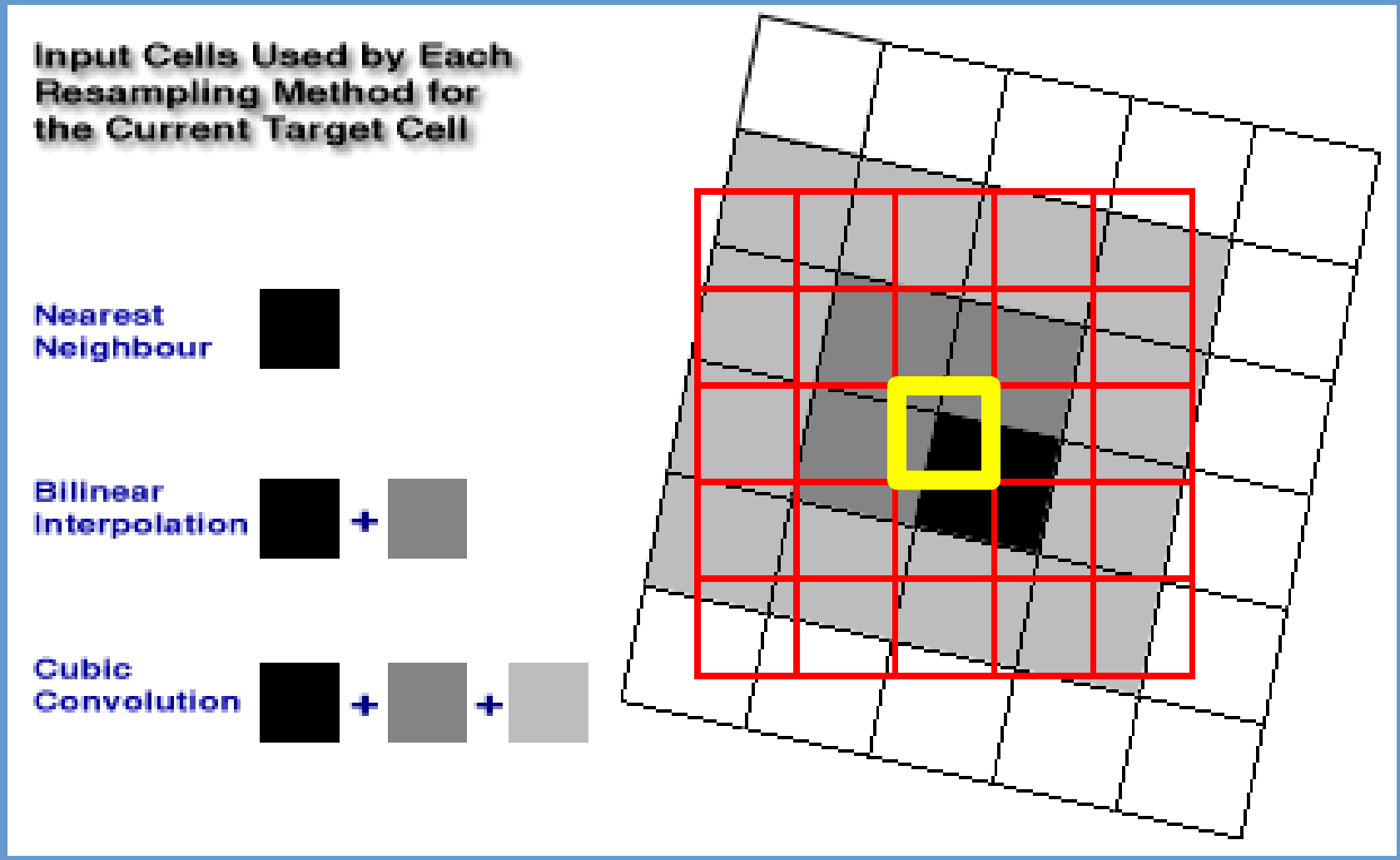
BG	PROV RATE	PROV	ENROLL	
1	10	4	40	
2	0	0	20	
3	0	0	20	
4	12	24	200	
5	20	1	5	
6	5	1	20	
7	0	0	30	
8	3	3	100	
9	7	14	200	
10	11	11	100	TRUE MEAN
MEAN	*6.80*	58	735	*7.891156463*

Raster Resampling

- Analysis of multiple raster layers generally requires that the grids have similar:
  - Coordinate systems
  - Cell size
  - Cell locations
    - Cells must align in space
- When using data from multiple sources, this can cause problems
  - One layer will have to be processed (spatially)
  - Resampling

Resampling Methods

- Nearest neighbor
  - Assigns the value of the nearest input grid cell as the value of each output cell (preserves original values, but can have a "jagged" appearance)
- Bilinear interpolation
  - Averages four nearest neighbors, weighted by distance, to calculate the output value (result is smoothest)
- Cubic convolution
  - Averages sixteen nearest neighbors, weighted with nonlinear distance function, to calculate output (distance weighting sharpens image)


Input Cells Used by Each Resampling Method for the Current Target Cell
Nearest Neighbour
Bilinear Interpolation
Cubic Convolution

Resampling

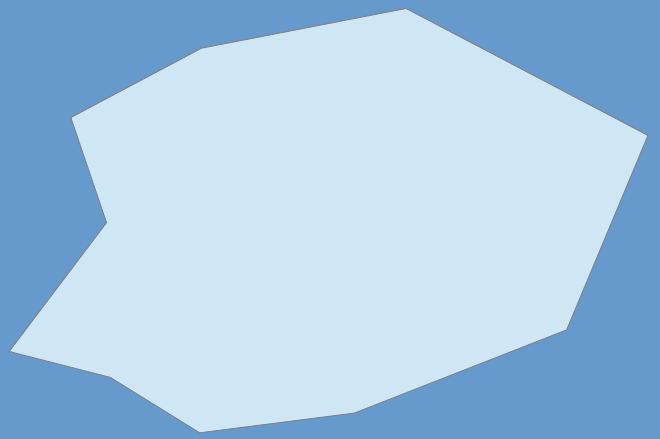
- Resampling changes the original cell values
  - Keep in mind "when" in your analysis you resample your data
  - Personal advice: Resample as late in the analysis as possible
    - Integrity of original data
    - Processing processed data
      - Input Ų Processing Ų Output Ų Resample
      - Input Ų Resample Ų Processing Ų Output

Centroids

- We often need to convert features from their original representation to a point feature
  - Polygons, lines, points
  - For example,
    - Distance calculations
    - Overlay operations
  - Can have analytical value as well

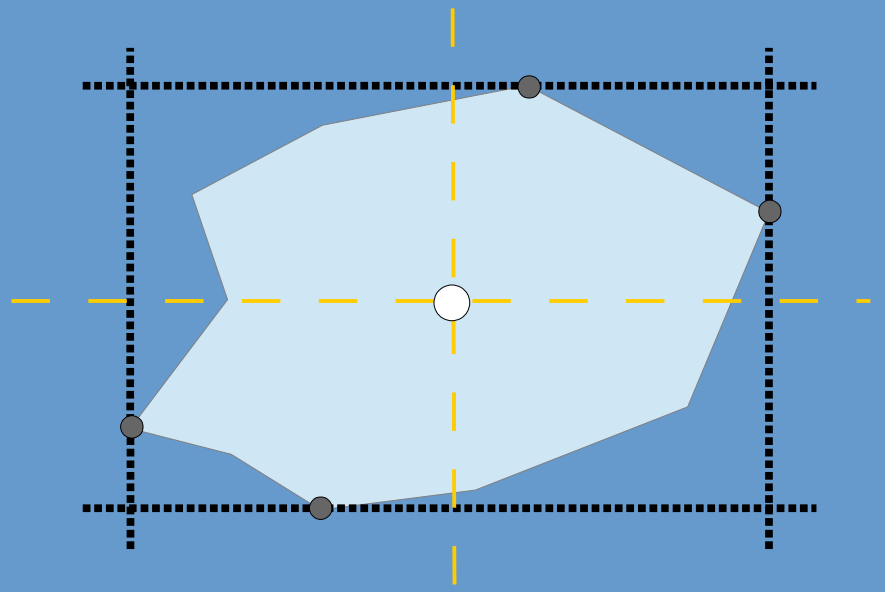
Geographic Centroid

- Simply, the central location
  - Determined by the bounding box
    - Range in X, Y
    - $X = X_{min} + (X_{max} - X_{min} / 2)$
    - $Y = Y_{min} + (Y_{max} - Y_{min} / 2)$



Geographic Centroid

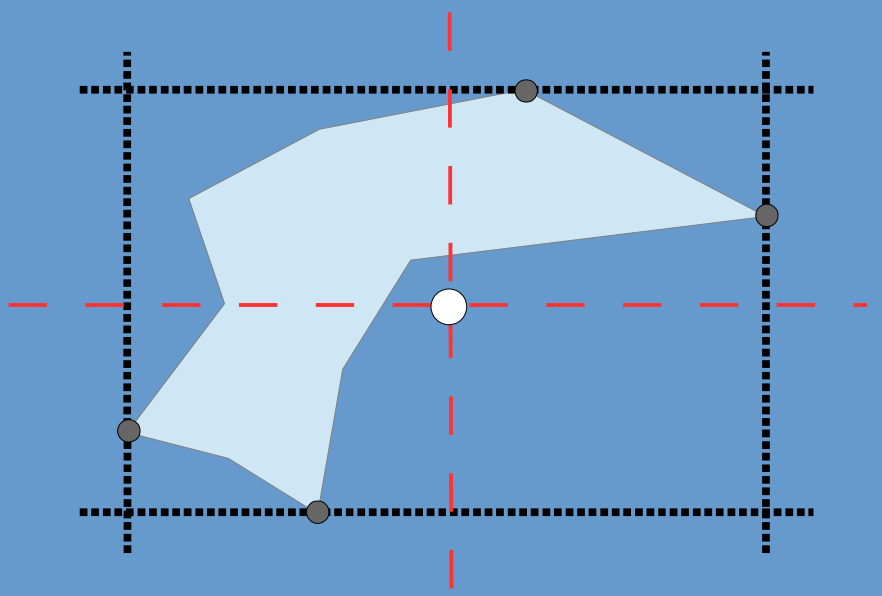
- Simply, the central location
  - Determined by the bounding box
    - Range in X, Y
    - $X = X_{min} + (X_{max} - X_{min} / 2)$
    - $Y = Y_{min} + (Y_{max} - Y_{min} / 2)$



Geographic Centroid

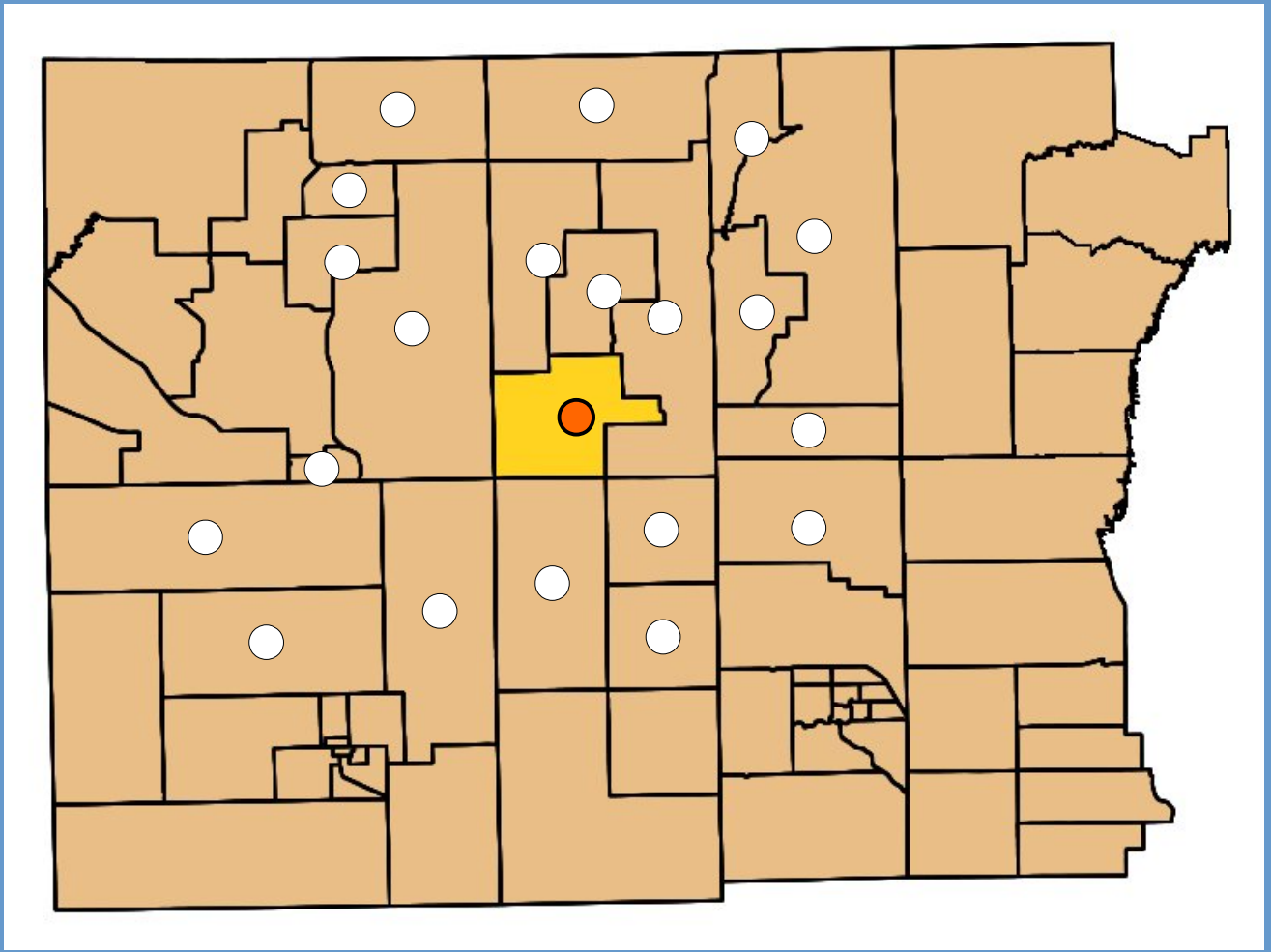
- Simply, the central location
  - Determined by the bounding box
    - Range in X, Y
    - $X = X_{min} + (X_{max} - X_{min} / 2)$
    - $Y = Y_{min} + (Y_{max} - Y_{min} / 2)$

?!?



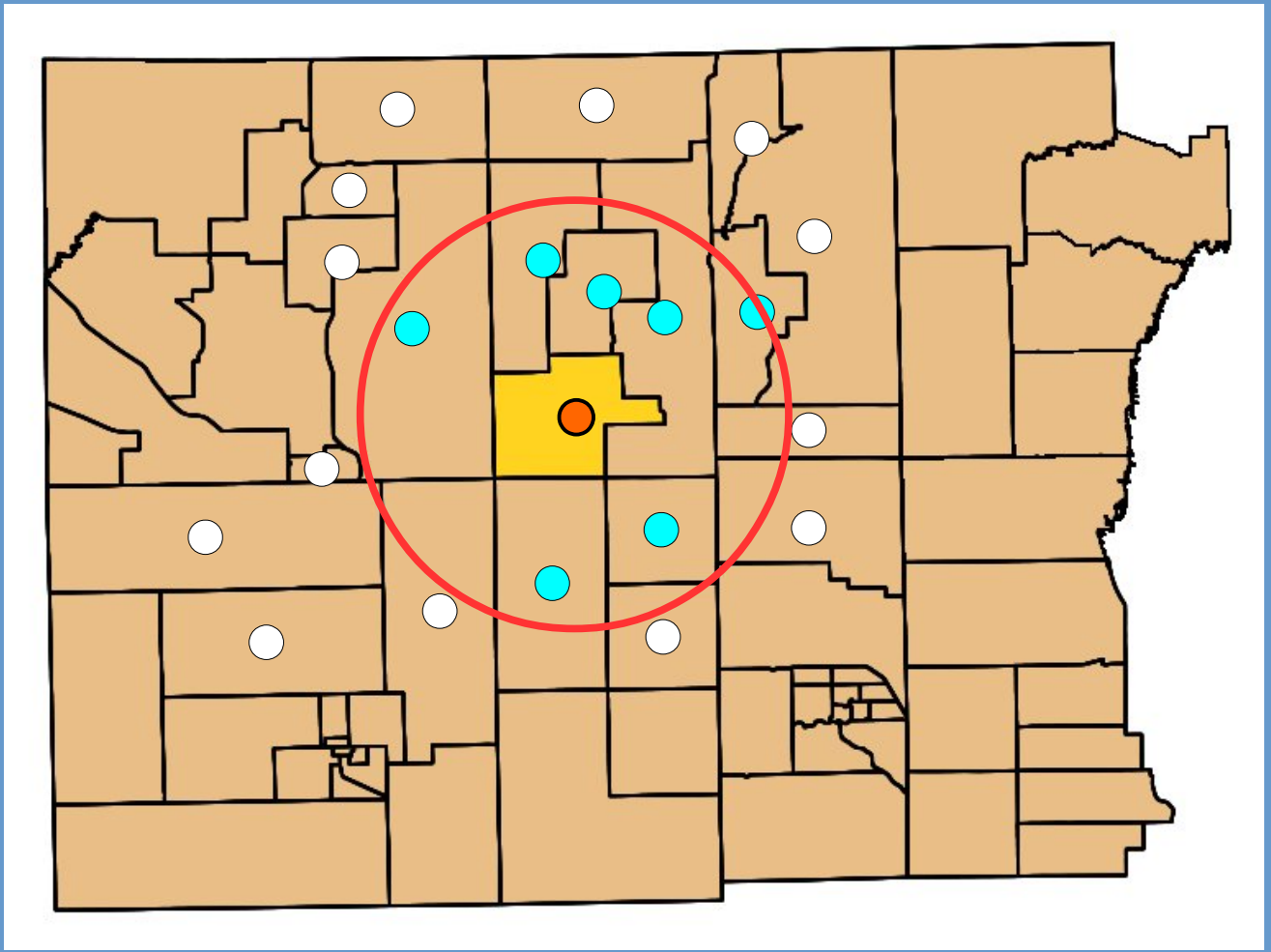
Geographic Centroid

- Measuring distance among polygon features



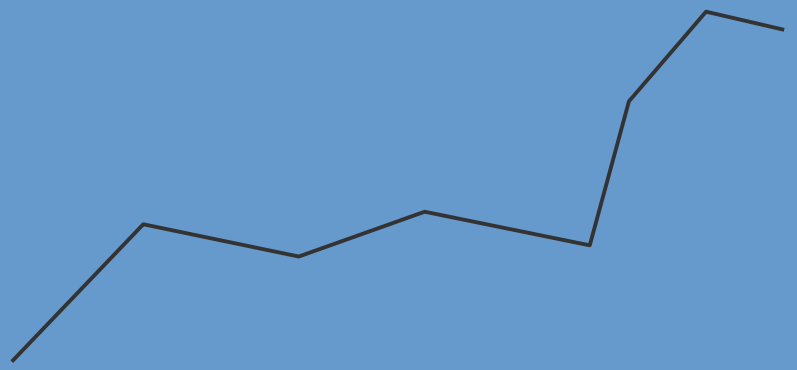
Geographic Centroid

- Measuring distance among polygon features



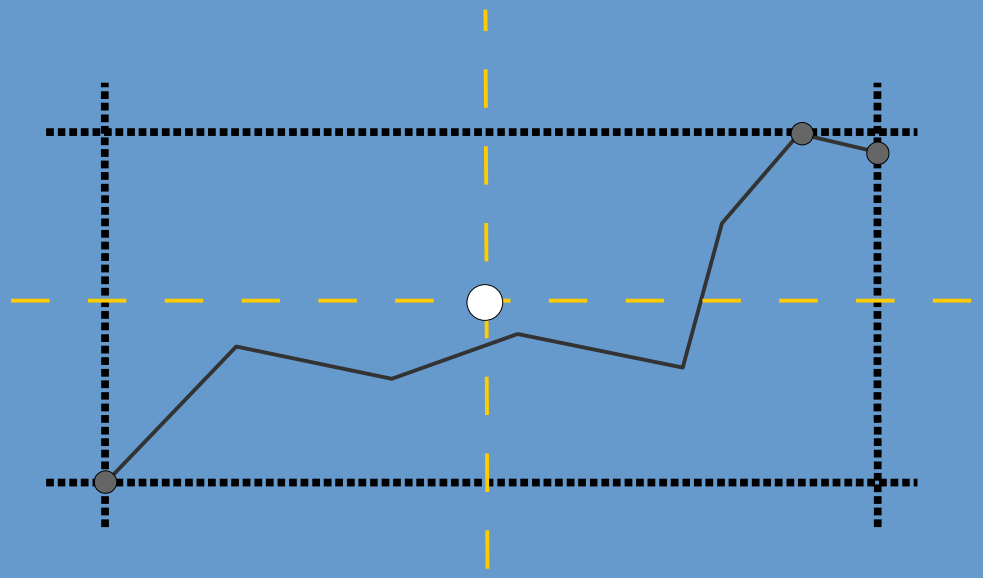
Geographic Centroid

- Simply, the central location
  - Determined by the bounding box
    - Range in X, Y
    - $X = X_{min} + (X_{max} - X_{min} / 2)$
    - $Y = Y_{min} + (Y_{max} - Y_{min} / 2)$



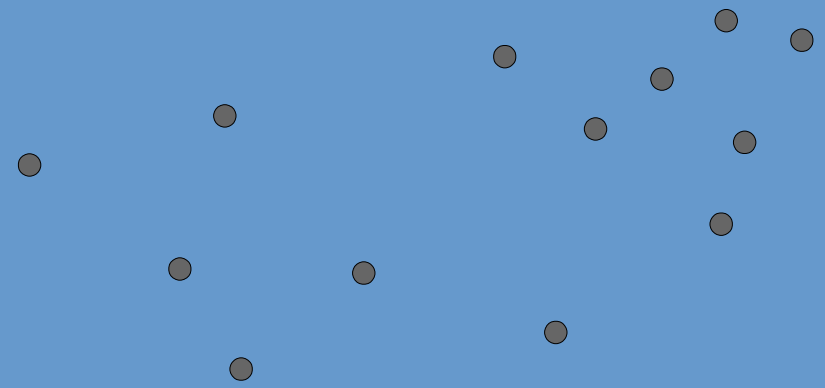
Geographic Centroid

- Simply, the central location
  - Determined by the bounding box
    - Range in X, Y
    - $X = X_{min} + (X_{max} - X_{min} / 2)$
    - $Y = Y_{min} + (Y_{max} - Y_{min} / 2)$



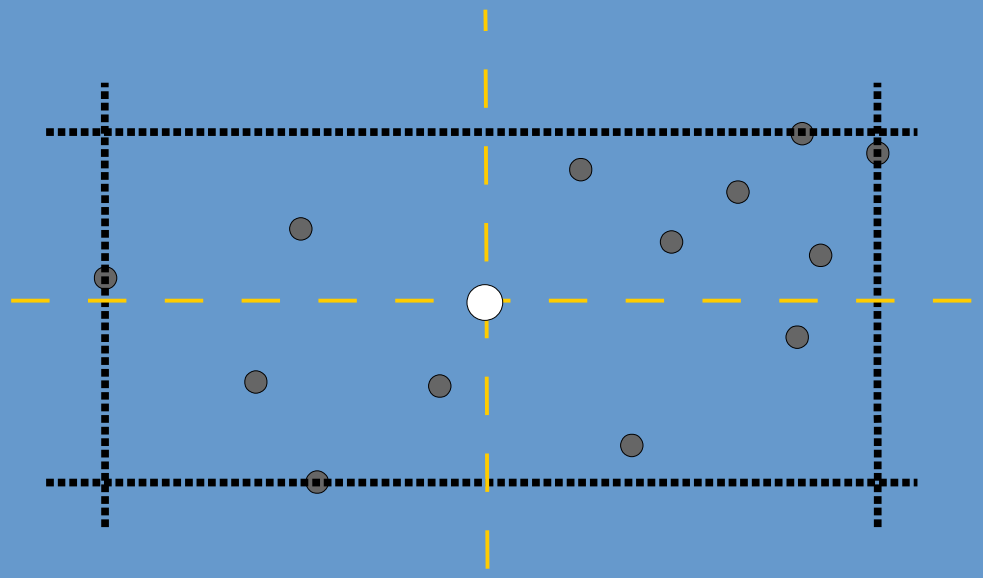
Geographic Centroid

- Simply, the central location
  - Determined by the bounding box
    - Range in X, Y
    - $X = X_{min} + (X_{max} - X_{min} / 2)$
    - $Y = Y_{min} + (Y_{max} - Y_{min} / 2)$



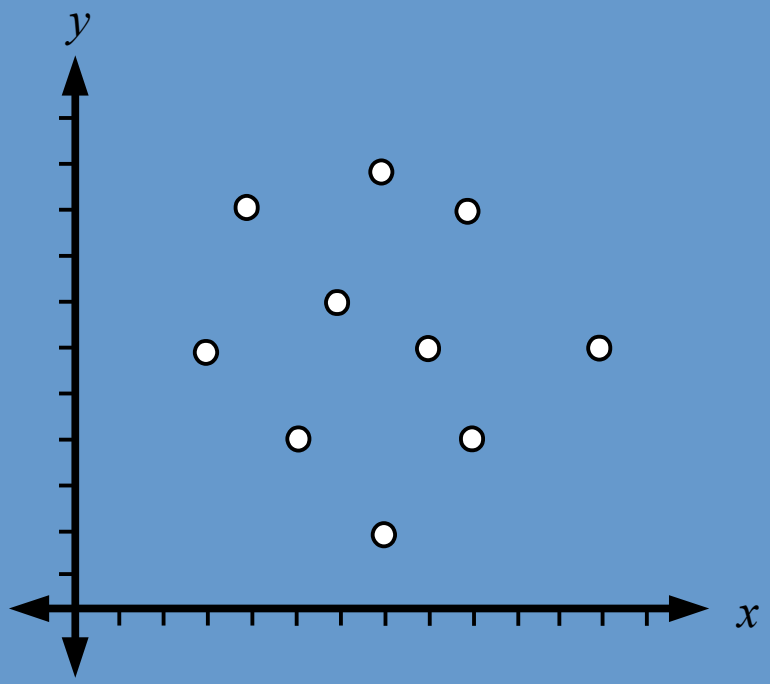
Geographic Centroid

- Simply, the central location
  - Determined by the bounding box
    - Range in X, Y
    - $X = X_{min} + (X_{max} - X_{min} / 2)$
    - $Y = Y_{min} + (Y_{max} - Y_{min} / 2)$



Central Feature

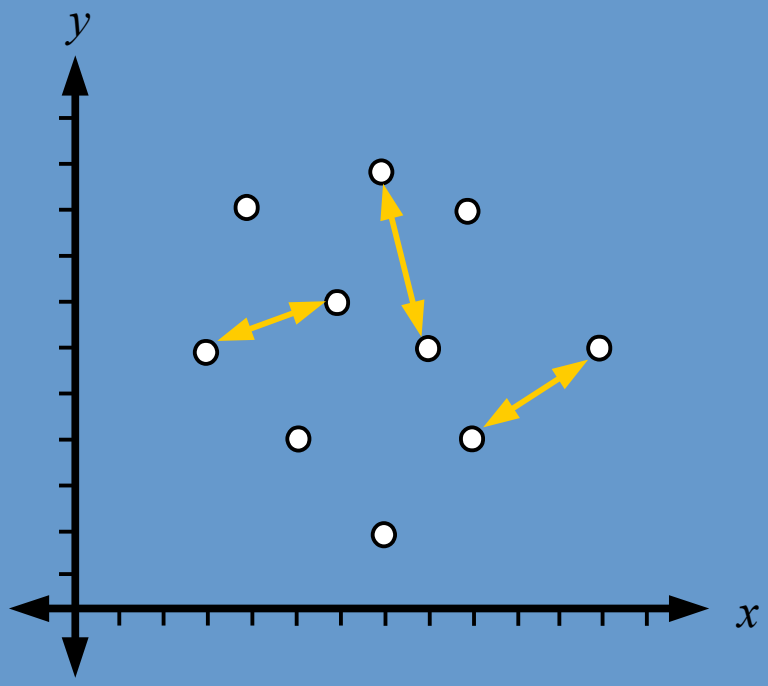
- Feature, within the observations, having the lowest total distance to the other features



Point	X	Y
1	3	6
2	4	9
3	5	4
4	6	7
5	7	2
6	7	10
7	8	6
8	9	4
9	9	9
10	12	6

Central Feature

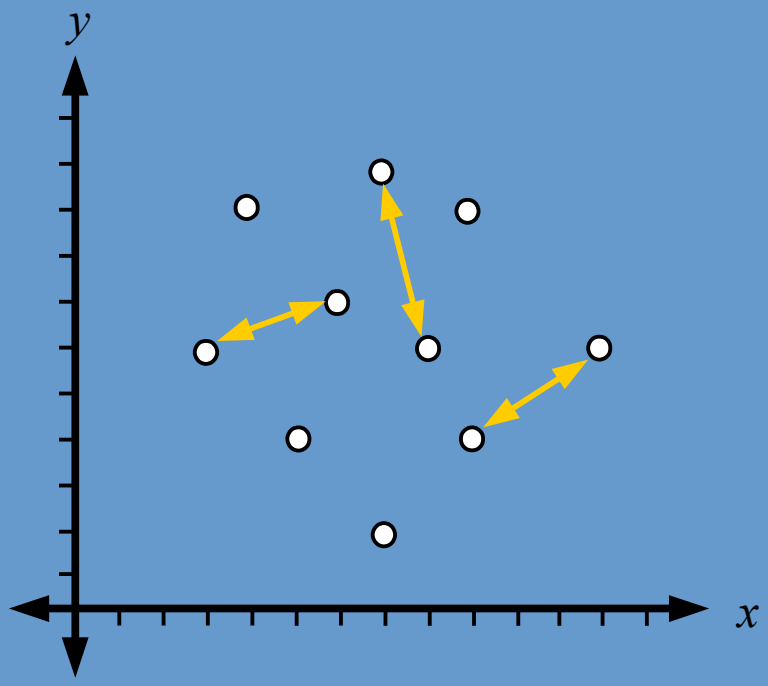
- Feature, within the observations, having the lowest total distance to the other features



Point	1	2	3	4	5	6	7	8	9	10
1	0	3.16	2.83	3.16	5.66	5.66	5.00	6.32	6.71	9.00
2	3.16	0	5.10	2.83	7.62	3.16	5.00	7.07	5.00	8.54
3	2.83	5.10	0	3.16	2.83	6.32	3.61	4.00	6.40	7.28
4	3.16	2.83	3.16	0	5.10	3.16	2.24	4.24	3.61	6.08
5	5.66	7.62	2.83	5.10	0	8.00	4.12	2.83	7.28	6.40
6	5.66	3.16	6.32	3.16	8.00	0	4.12	6.32	2.24	6.40
7	5.00	5.00	3.61	2.24	4.12	4.12	0	2.24	3.16	4.00
8	6.32	7.07	4.00	4.24	2.83	6.32	2.24	0	5.00	3.61
9	6.71	5.00	6.40	3.61	7.28	2.24	3.16	5.00	0	4.24
10	9.00	8.54	7.28	6.08	6.40	6.40	4.00	3.61	4.24	0

Central Feature

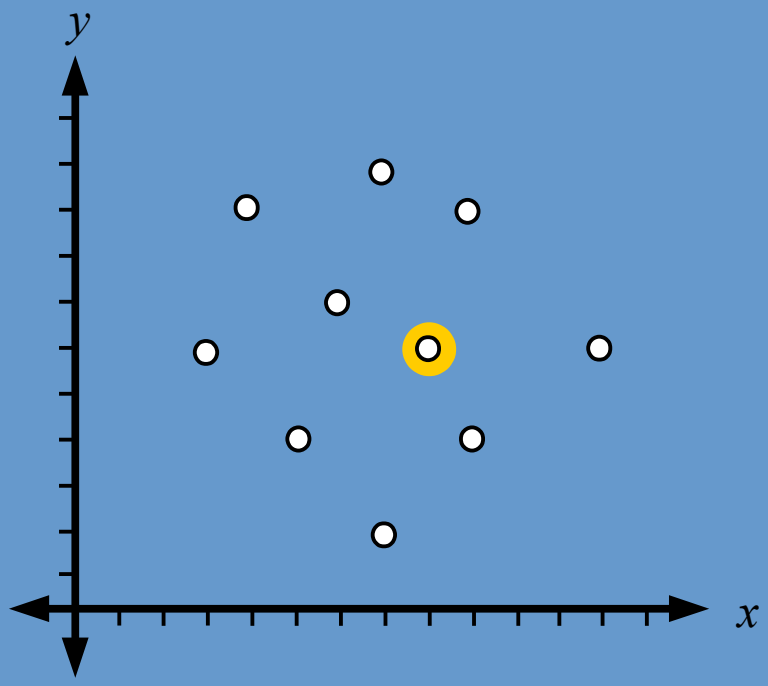
- Feature, within the observations, having the lowest total distance to the other features



Point	1	2	3	4	5	6	7	8	9	10	SUM
1	0	3.16	2.83	3.16	5.66	5.66	5.00	6.32	6.71	9.00	47.50
2	3.16	0	5.10	2.83	7.62	3.16	5.00	7.07	5.00	8.54	47.48
3	2.83	5.10	0	3.16	2.83	6.32	3.61	4.00	6.40	7.28	41.53
4	3.16	2.83	3.16	0	5.10	3.16	2.24	4.24	3.61	6.08	*33.58*
5	5.66	7.62	2.83	5.10	0	8.00	4.12	2.83	7.28	6.40	49.83
6	5.66	3.16	6.32	3.16	8.00	0	4.12	6.32	2.24	6.40	45.39
7	5.00	5.00	3.61	2.24	4.12	4.12	0	2.24	3.16	4.00	***33.49***
8	6.32	7.07	4.00	4.24	2.83	6.32	2.24	0	5.00	3.61	41.63
9	6.71	5.00	6.40	3.61	7.28	2.24	3.16	5.00	0	4.24	43.64
10	9.00	8.54	7.28	6.08	6.40	6.40	4.00	3.61	4.24	0	55.56

Central Feature

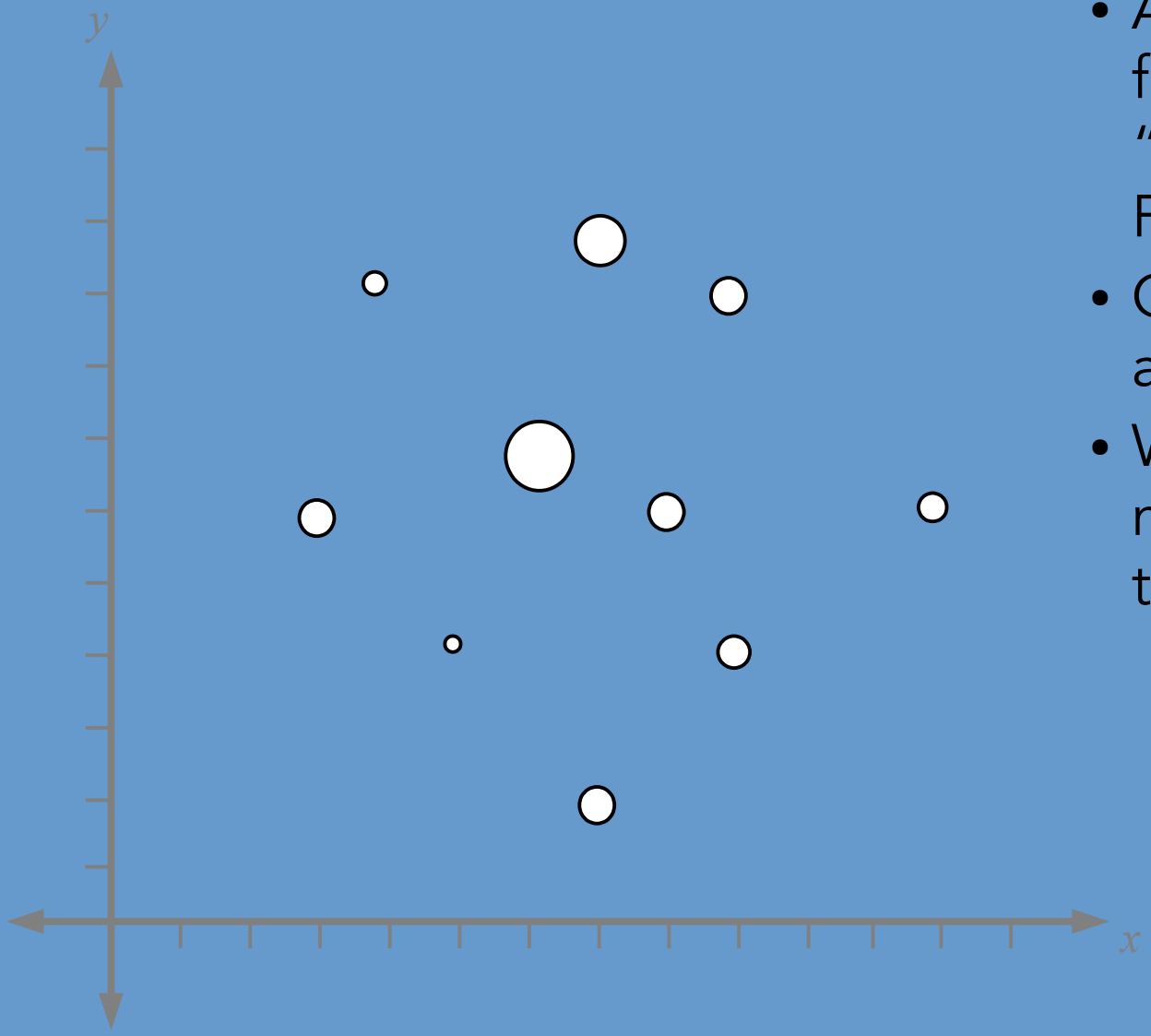
- Feature, within the observations, having the lowest total distance to the other features



Point	1	2	3	4	5	6	7	8	9	10	SUM
1	0	3.16	2.83	3.16	5.66	5.66	5.00	6.32	6.71	9.00	47.50
2	3.16	0	5.10	2.83	7.62	3.16	5.00	7.07	5.00	8.54	47.48
3	2.83	5.10	0	3.16	2.83	6.32	3.61	4.00	6.40	7.28	41.53
4	3.16	2.83	3.16					24	3.61	6.08	*33.58*
5	5.66	7.62	2.83	5.1				83	7.28	6.40	49.83
6	5.66	3.16	6.32	3.1				32	2.24	6.40	45.39
7	5.00	5.00	3.61	2.2				24	3.16	4.00	***33.49***
8	6.32	7.07	4.00	4.2				0	5.00	3.61	41.63
9	6.71	5.00	6.40	3.6				00	0	4.24	43.64
10	9.00	8.54	7.28	6.0				61	4.24	0	55.56

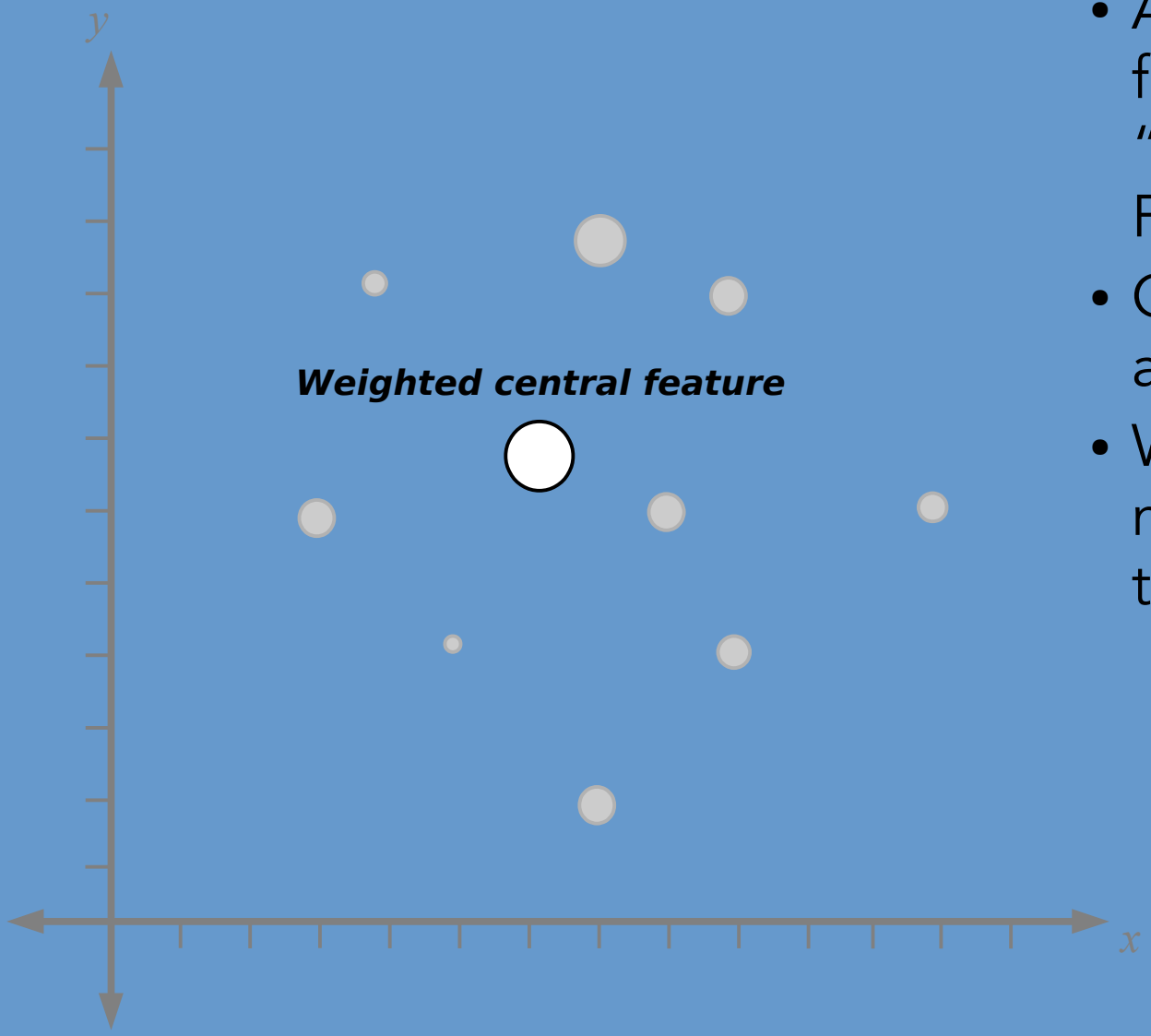
Point	*X*	*Y*
1	*3*	*6*
2	*4*	*9*
3	*5*	*4*
4	*6*	*7*
5	*7*	*2*
6	*7*	*10*
7	*8*	*6*
8	*9*	*4*
9	*9*	*9*
10	*12*	*6*

Central Feature



- Attribute value (for each feature) used as "weights" in Central Feature calculation
- Considers both distance and weight value
- Weight can be any numeric field in attribute table

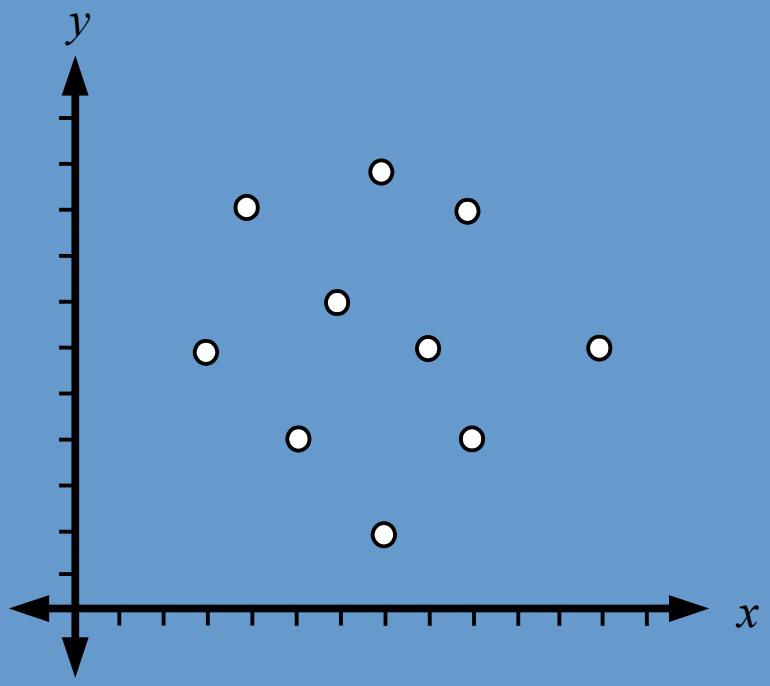
Central Feature



- Attribute value (for each feature) used as "weights" in Central Feature calculation
- Considers both distance and weight value
- Weight can be any numeric field in attribute table

Mean Center

- Simply, the mean location (in two dimensions) of the set of points



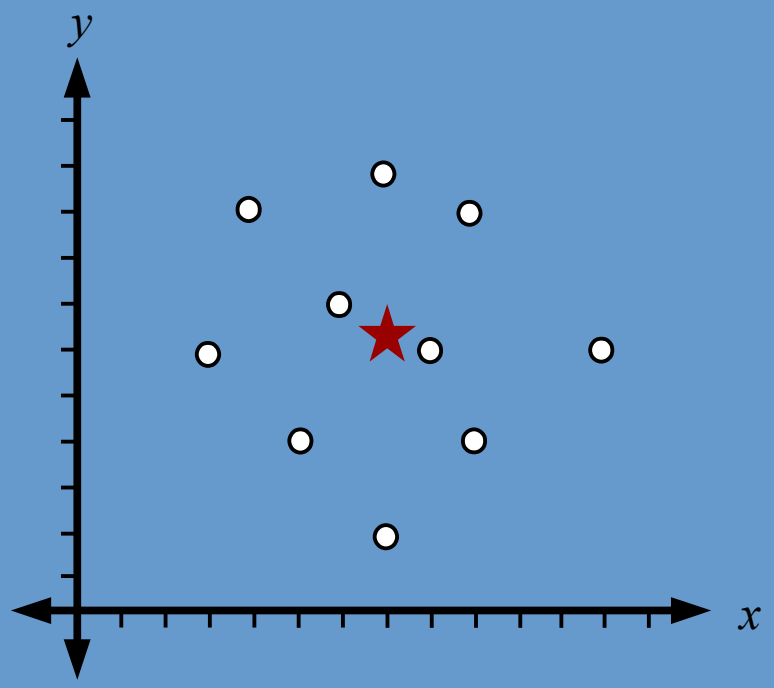
$$\overline{X}_c = \frac{\sum X_i}{n}$$

$$\overline{Y}_c = \frac{\sum Y_i}{n}$$

Point	X	Y
1	3	6
2	4	9
3	5	4
4	6	7
5	7	2
6	7	10
7	8	6
8	9	4
9	9	9
10	12	6
MEAN	**7.0**	**6.3**

Mean Center

- Simply, the mean location (in two dimensions) of the set of points

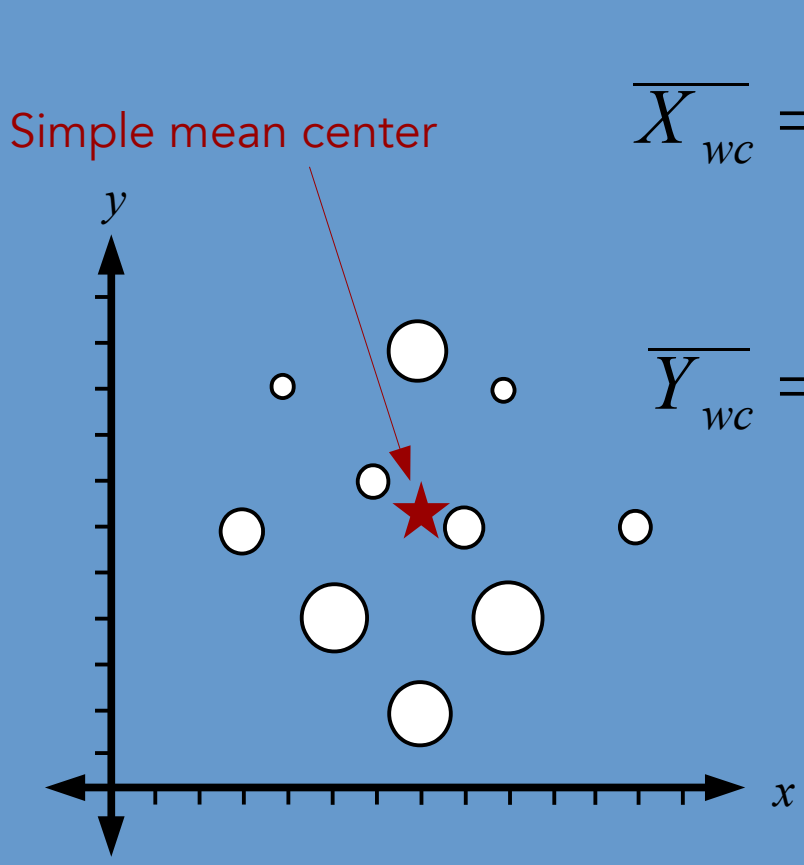


  - Output layer contains only the mean center
  - Use "Case" field for features in different classes within the same layer

Point	*X*	*Y*
1	*3*	*6*
2	*4*	*9*
3	*5*	*4*
4	*6*	*7*
5	*7*	*2*
6	*7*	*10*
7	*8*	*6*
8	*9*	*4*
9	*9*	*9*
10	*12*	*6*
MEAN	***7.0***	***6.3***

Weighted Mean Center

- Same concept as Mean Center, with weights, determined by attribute values



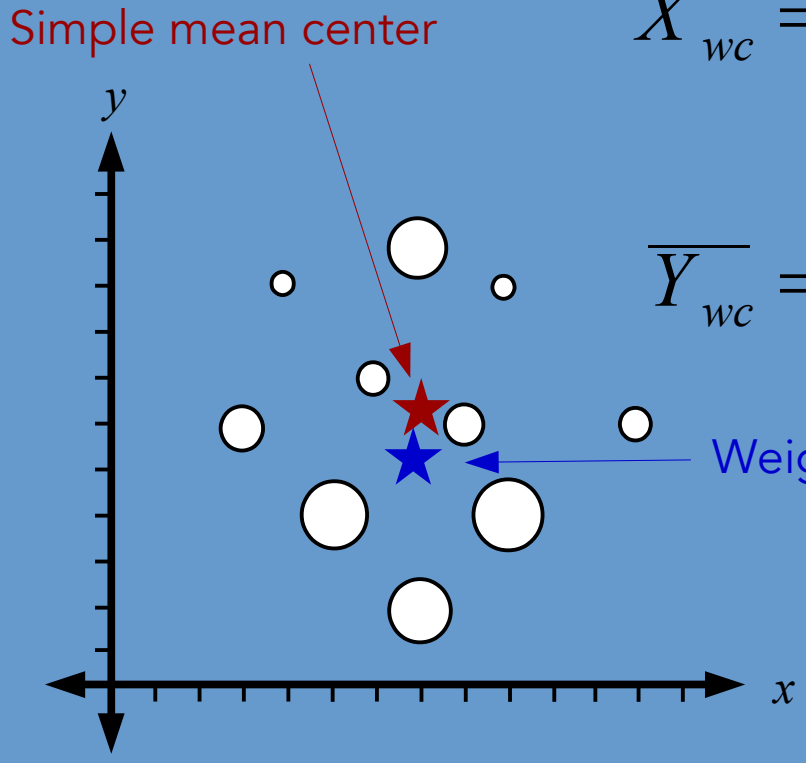
$$\overline{X}_{wc} = \frac{\sum X_i w_i}{\sum w_i}$$

$$\overline{Y}_{wc} = \frac{\sum Y_i w_i}{\sum w_i}$$

Point	X	Y	w	X * w	Y * w
1	3	6	54	162	324
2	4	9	12	48	108
3	5	4	108	540	432
4	6	7	23	138	161
5	7	2	98	686	196
6	7	10	93	651	930
7	8	6	44	352	264
8	9	4	121	1089	484
9	9	9	10	90	90
10	12	6	18	216	108
		SUM	**581**	**3972**	**3097**

Weighted Mean Center

- Same concept as Mean Center, with weights, determined by attribute values



$$\overline{X}_{wc} = \frac{\sum X_i w_i}{\sum w_i}$$

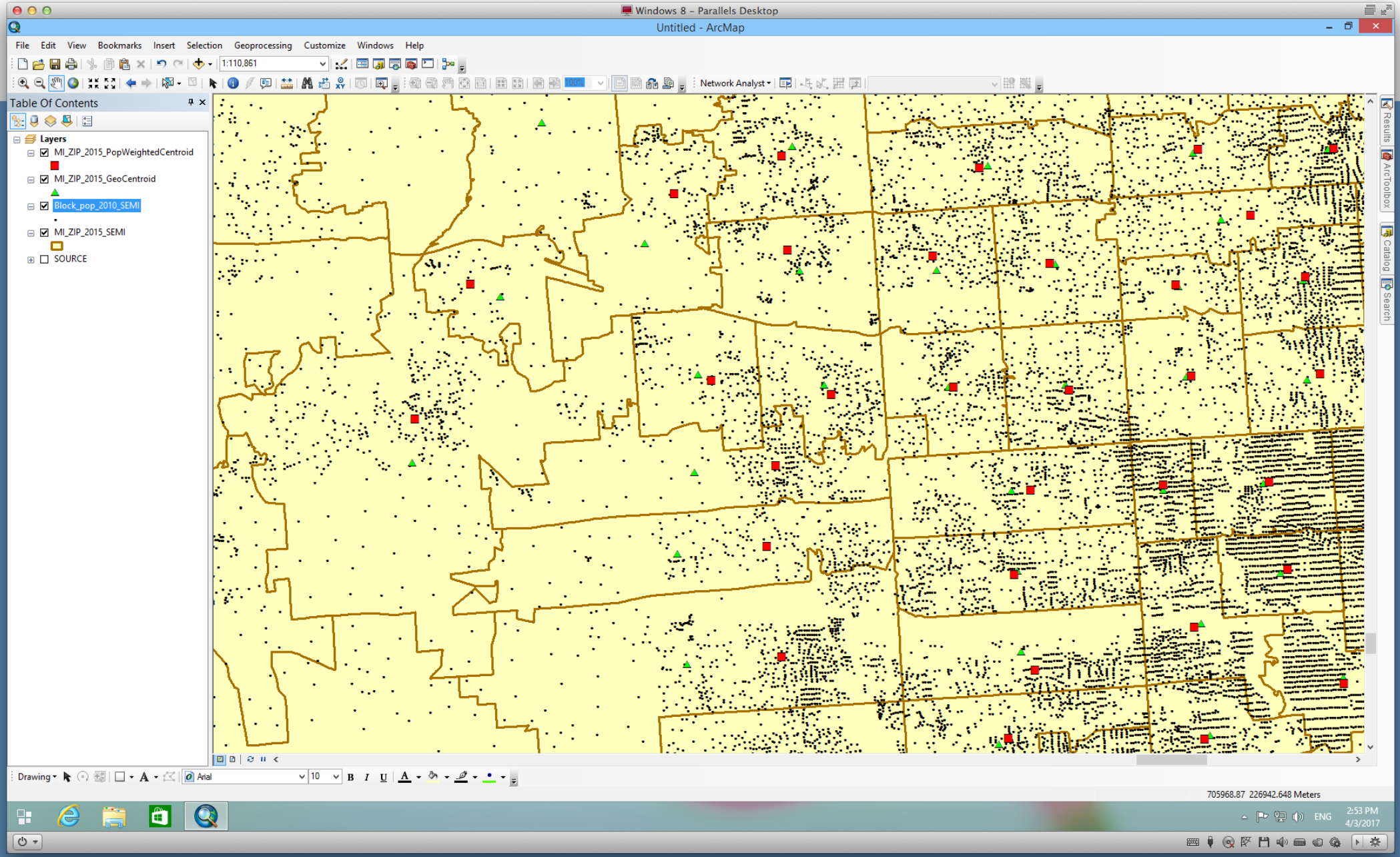
$$\overline{Y}_{wc} = \frac{\sum Y_i w_i}{\sum w_i}$$

Point	X	Y	w	X * w	Y * w
1	3	6	54	162	324
2	4	9	12	48	108
3	5	4	108	540	432
4	6	7	23	138	161
5	7	2	98	686	196
6	7	10	93	651	930
7	8	6	44	352	264
8	9	4	121	1089	484
9	9	9	10	90	90
10	12	6	18	216	108
		SUM	581	3972	3097

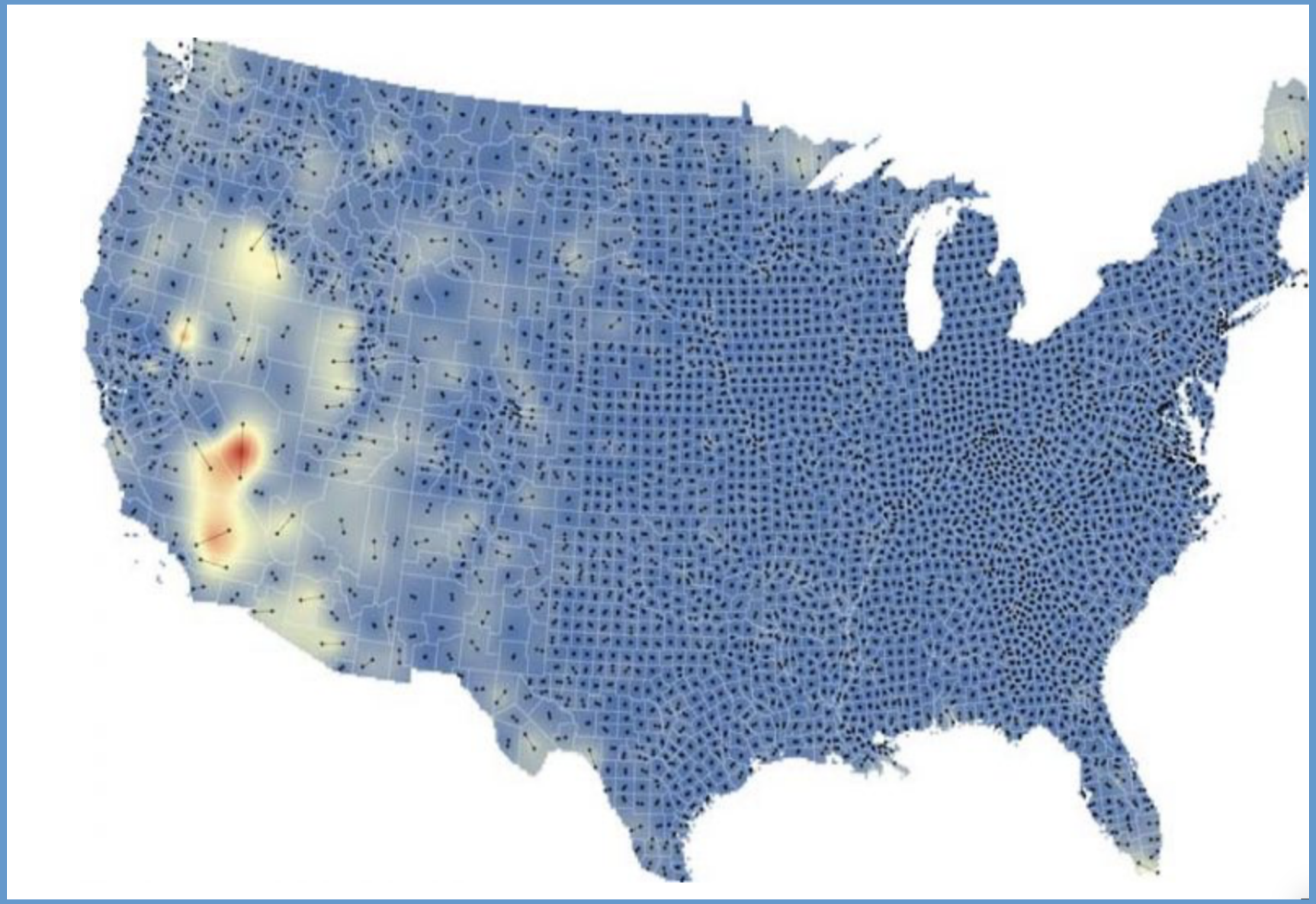
$$\overline{X}_{wc} = \frac{3972}{581} = 6.8 \qquad \overline{Y}_{wc} = \frac{3097}{581} = 5.3$$

Weighted Mean Center

- Same concept as Mean Center, with weights, determined by attribute values
  - Output layer contains only the weighted mean center
  - Use "Case" field for features in different classes within the same layer
- Extremely useful when integrating population information
  - The *Population Weighted Centroid* is the weighted mean center


Windows 8 – Parallels Desktop
Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:110,861
Network Analyst
Table Of Contents
Layers
MI_ZIP_2015_PopWeightedCentroid
MI_ZIP_2015_GeoCentroid
Block_pop_2010_SEMI
MI_ZIP_2015_SEMI
SOURCE
Results
ArcToolbox
Catalog
Search
Drawing
Arial
10
705968.87 226942.648 Meters
ENG
2:53 PM
4/3/2017

Geometric VS Population Centroid



Centroids in QGIS

- Centroid
- Mean Coordinates

Keywords

- Data Integration
- Subset
- Aggregate
  - Dissolve
- Calculate field
- Raster resampling
  - Nearest neighbor, bilinear interpolation, cubic convolution
- Centroids
  - Geographic
  - Central feature
  - Mean center
    - Population weighted centroid