Deck 3: Vector Data

Spatial Vector Data – Cartography: Colors – Joining Data

Intro to GIS – UMass Amherst – Michael F. Nelson

Some notes on data types and formats

- “Data type” usage is ambiguous, it can refer to:
  - How numbers or categories are encoded in your computer:
    - Integer, short, long
      - Does this remind you of the aerial photo in lab 2?
    - Float, double: decimal or real numbers
    - Boolean: true/false
  - Spatial data formats and file types:
    - Raster: grid, image, NetCDF, etc.
    - Vector: ‘shapefile’, feature layer, lots of others

Help, I need more time to finish my lab!

- Be in touch with Ollie or Mike if you run into issues. The earlier the better!
- Windows Virtual Desktop: you can access this from a browser on any computer.
- Your computer: you can install ArcPRO on your Windows machine.

Overview

Cartography: Colors Revisited

- Accessibility: Colorblindness

Vector Spatial Data

- What is a vector?
- Types of spatial vectors.
- Vector data operations.

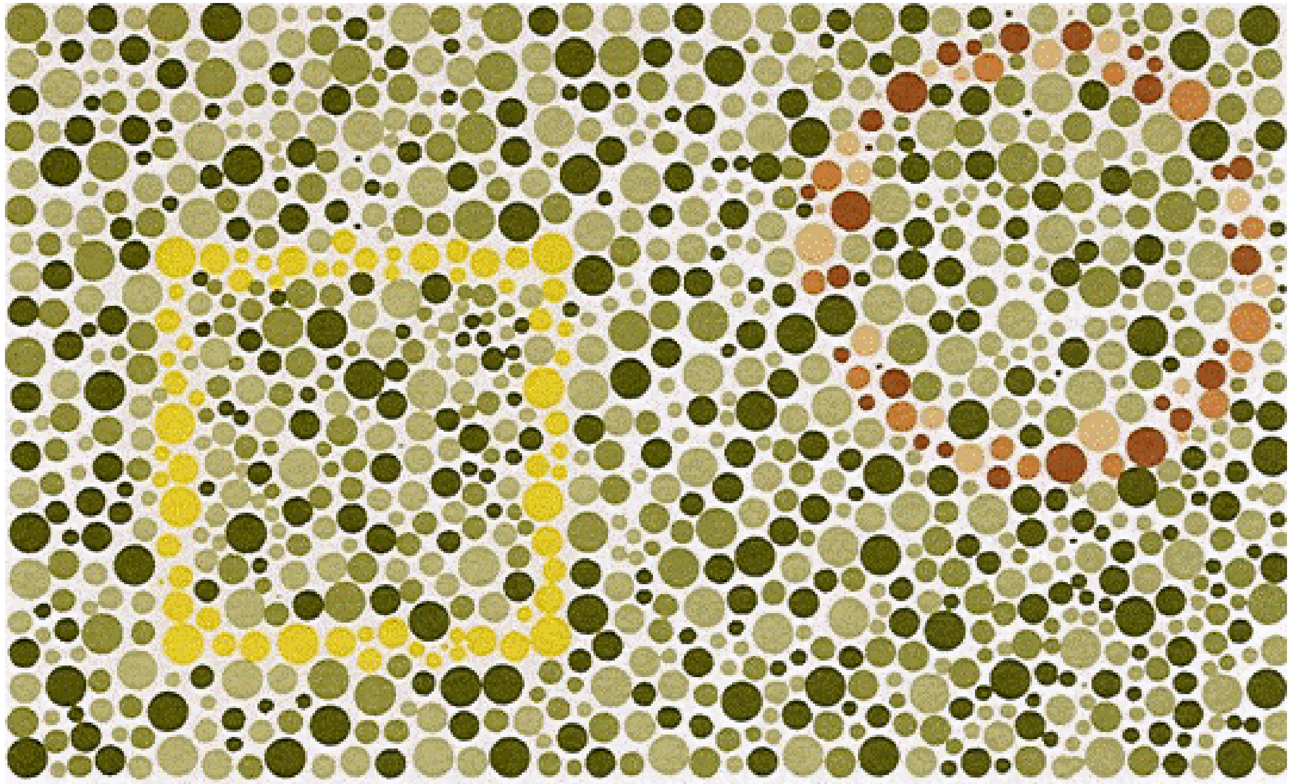
Vector Data Application

- Suitability analyses

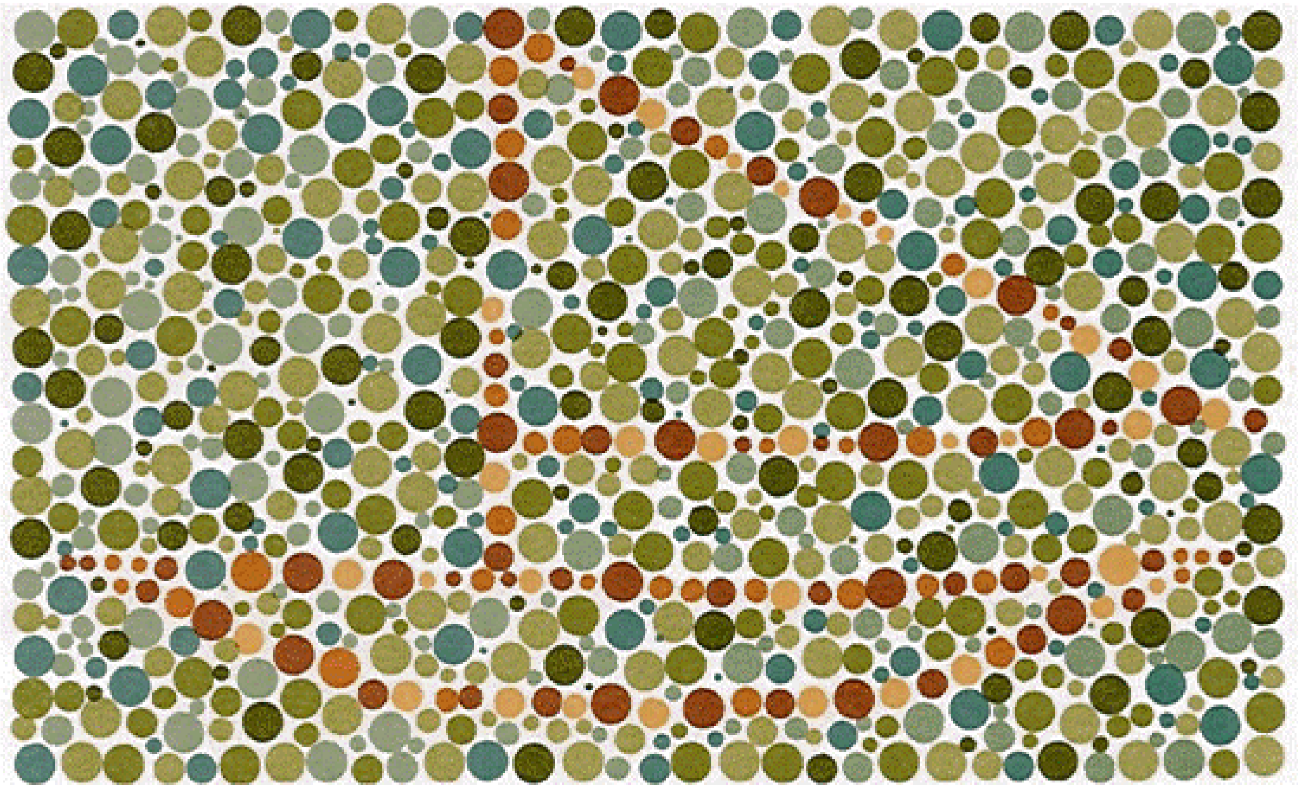
Map Design: Color

Colorblind Design

What shapes do you see?



What shapes do you see?

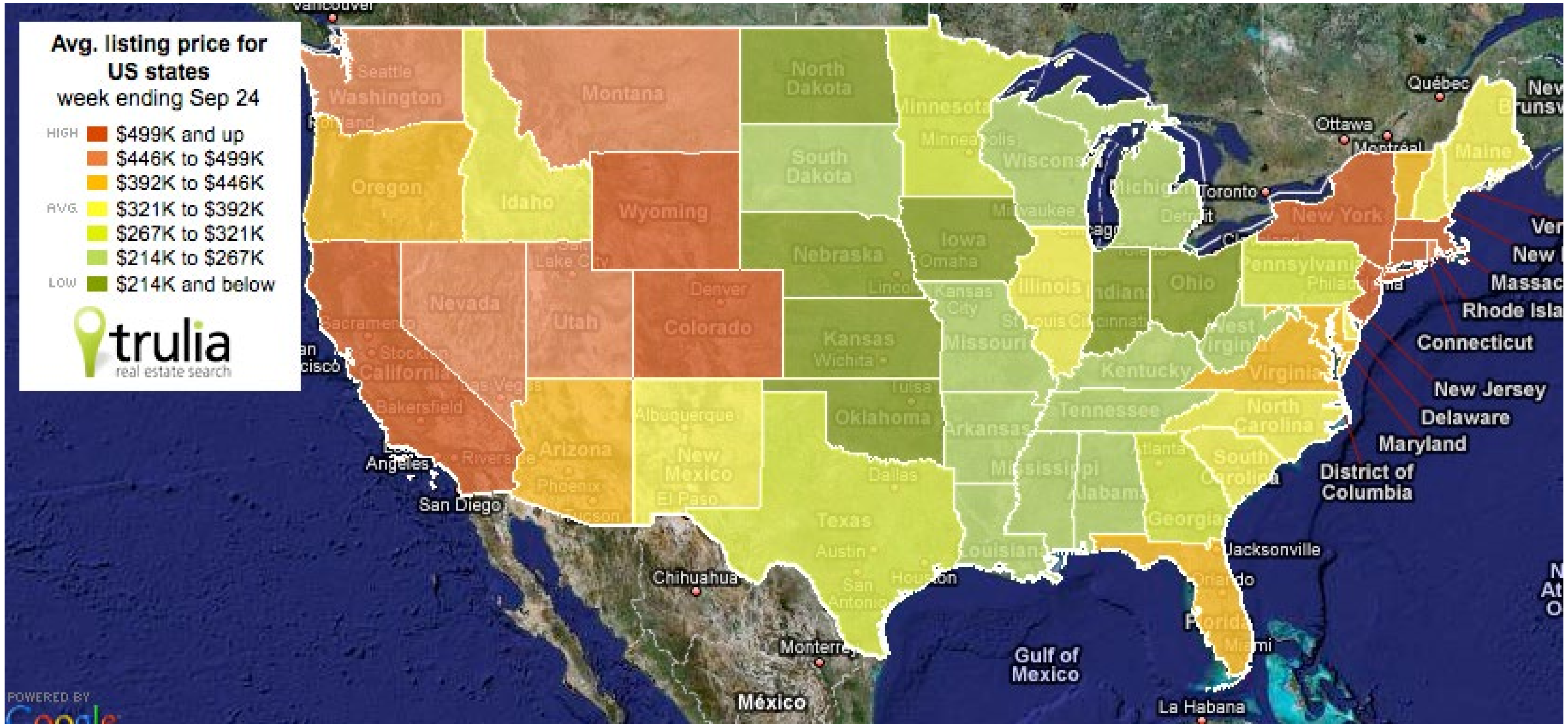


Red-green color blindness

The colors of the rainbow as viewed by a person with no color vision deficiencies.

The colors of the rainbow as viewed by a person with protanopia.

Cartography revisited: colorblind design



Color Accessibility Options

- Use a colorblindness utility:
  - Color simulator in Arc
  - Color Oracle (or other) application
  - Web tools
- Use grayscale
- Avoid red/green color contrasts.

Model Thinking

- A model is a **simplified** representation of reality.
- How can we represent spatial relationships digitally?

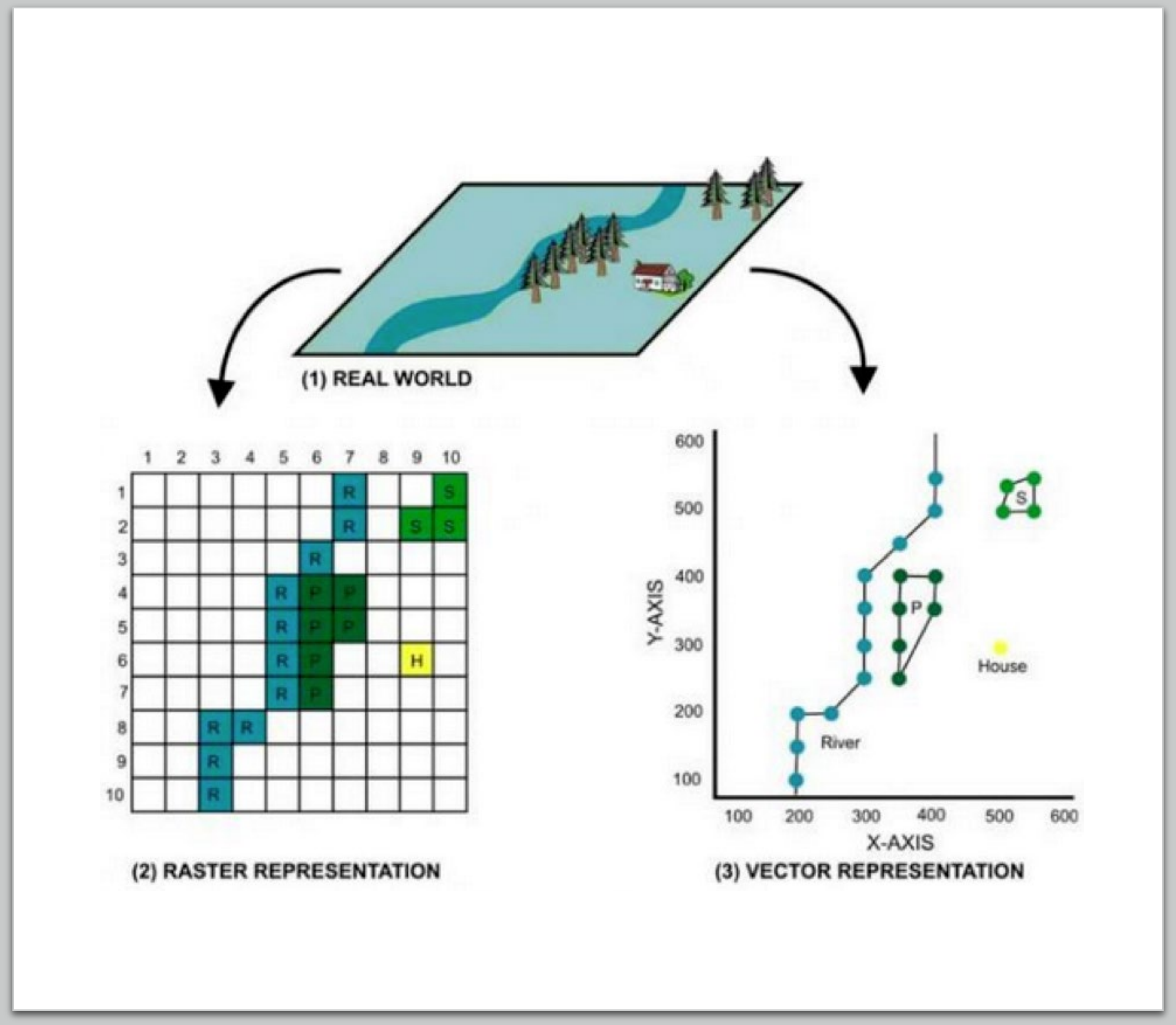


Table 2-2: A comparison of raster and vector data models.

Characteristic	Raster	Vector
data structure	usually simple	usually complex
storage requirements	large for most data sets without compression	small for most data sets
coordinate conversion	may be slow due to data volumes, and require resampling	simple
analysis	easy for continuous data, simple for many layer combinations	preferred for network analyses, many other spatial operations more complex
positional precision	floor set by cell size	limited only by positional measurements
accessibility	easy to modify or program, due to simple data structure	often complex
display and output	good for images, but discrete features may show "stairstep" edges	map-like, with continuous curves, poor for images

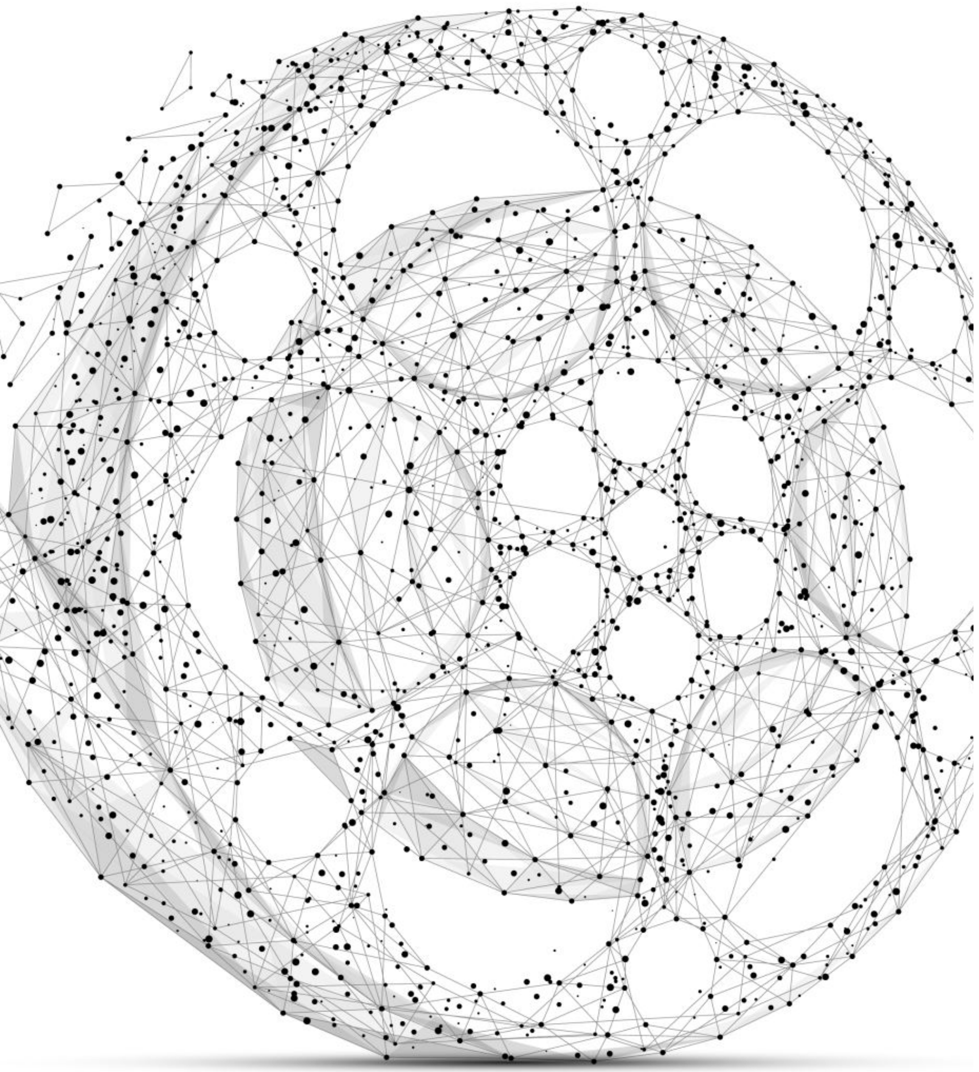
Vector and Raster are the two data models we'll focus on in this course.

Source: Paul Bolstad. 2012. GIS Fundamentals – A first text on Geographic Information Systems. 4th ed.

Data data data

- Three data ~~types~~ (formats/models) you will work with (and already have!) in GIS:
  - Raster (yes)
  - Vector (yes)
  - TIN (not yet)

- What are the key differences?



Raster vs. Vector vs. TIN

- Raster: Made up of cells (pixels).
- Vector: Made up of points, lines, and polygons.
- TIN: Triangular Irregular Network. Useful for elevation models. Very cool data model! A type of tessellation.

Vector Data

We'll start with vector data. It's often easier to work with, and it'll be on the midterm!

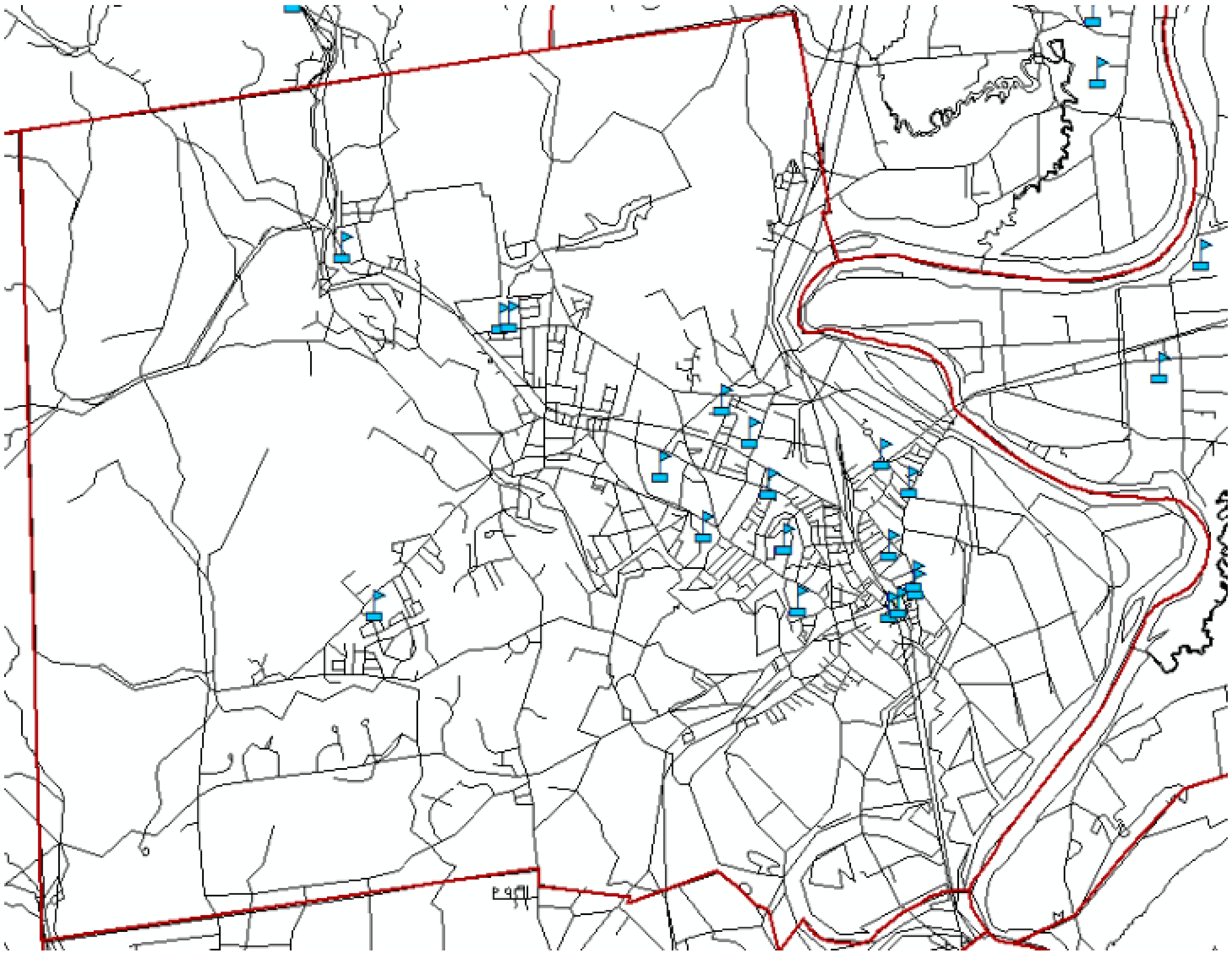
Vector data represents features as points, lines, and polygons and is best applied to discrete objects with defined shapes and boundaries.

Features have a precise shape and position, attributes and metadata, and useful behavior.

Source: Zeiler, M. 1999. *Modeling Our World: The ESRI® Guide to Geodatabase Design*. Redlands, CA: ESRI Press. 199 pp.

Vector (Feature) Data

- Vectors can represent:
  - Points
  - Lines
  - Polygons
- All vector data are built from points.
  - Each point has a coordinate



Vector Data

The vector data paradigm associates features with attributes. This sounds like the Row Data Paradigm!

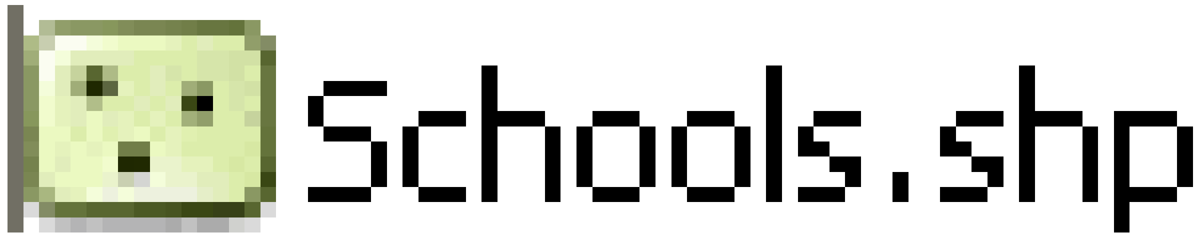
- **Feature:** stores the spatial information.
  - Each vertex in a feature has explicit x- and y- coordinates. This has important consequences!
- **Attribute table:** stores the associated data values.

Vector Data

- Key points:
  - Features and attribute tables are different data structures.
  - They're often stored in separate files.
  - The vector data paradigm associates a particular feature with a particular row in the attribute table.
  - Three main flavors: points, lines, polygons
- Raster data is a fundamentally different way of encoding spatial information.

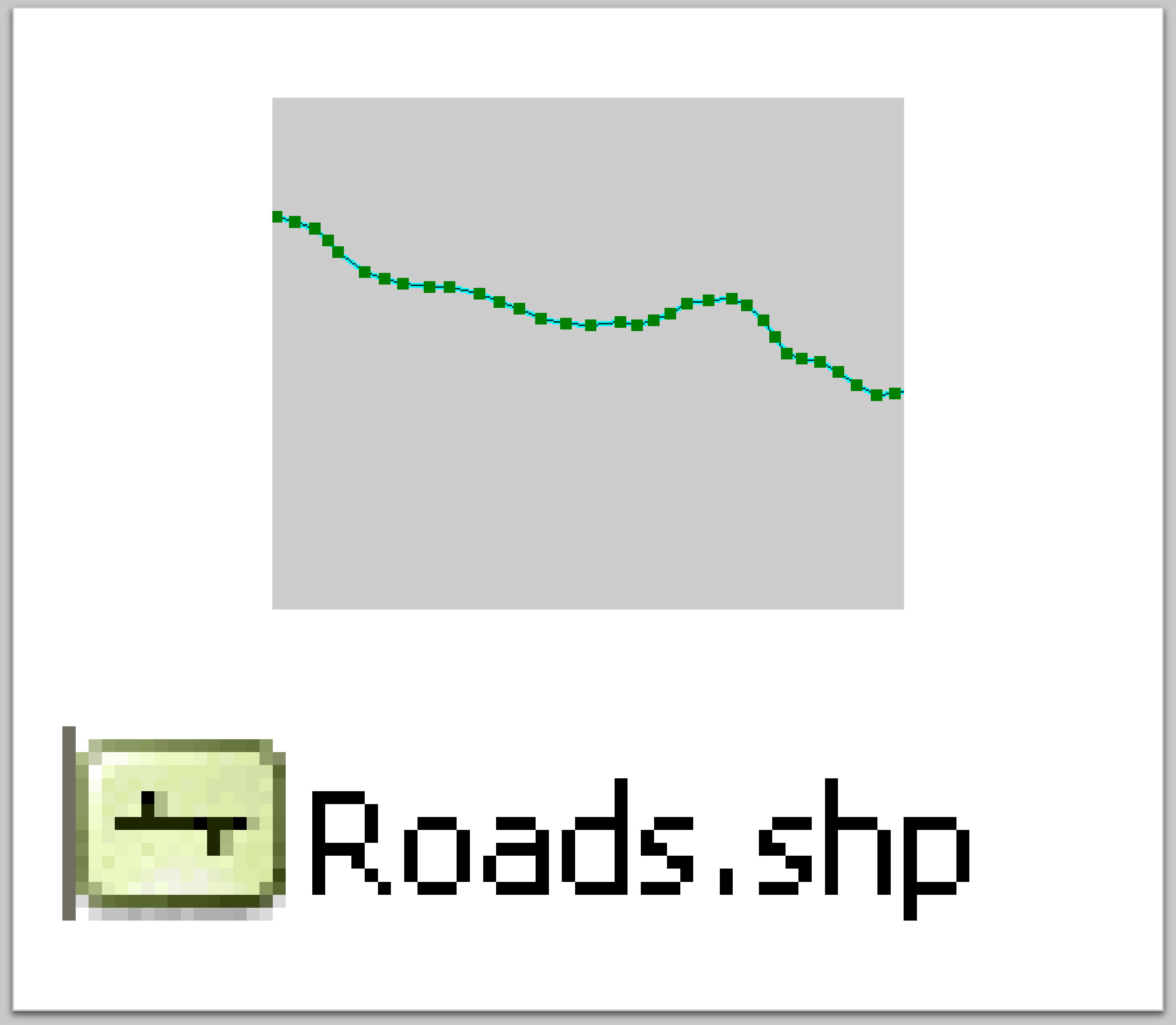
Vector (Feature) Data

- Point
  - A specific geographic location.
  - Points have no area or length.
- Line
- Polygon



Vector (Feature) Data

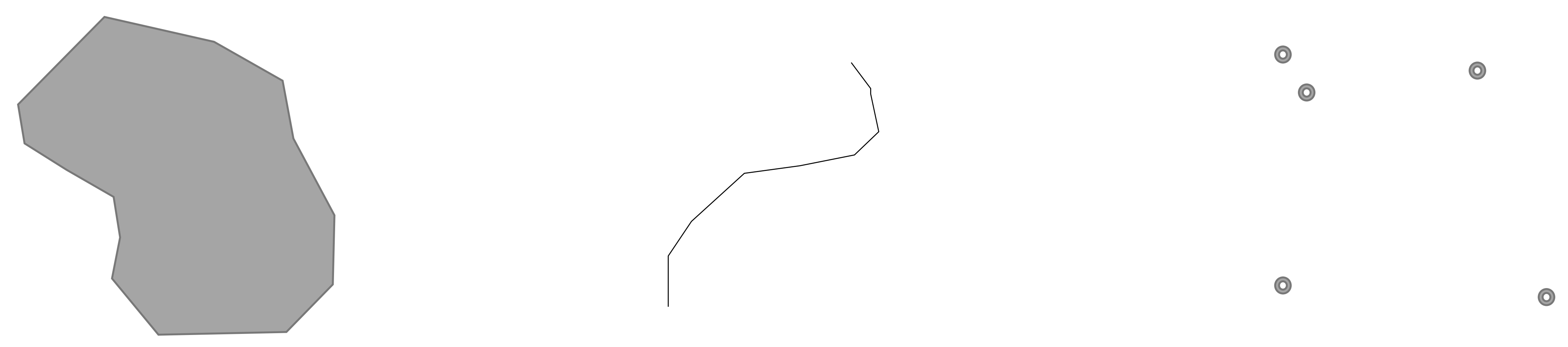
- Point
- Line
  - A segment with a specific geographic location
  - Lines have length.
  - Lines may encode direction information.
- Polygon



Vector (Feature) Data

- Point
- Line
- Polygon
    - An area enclosed within a polygon.
    - Vertices defined by points.
    - Edges are straight lines between points.

How should you represent a spatial feature?



How should you represent a spatial feature?

It depends!

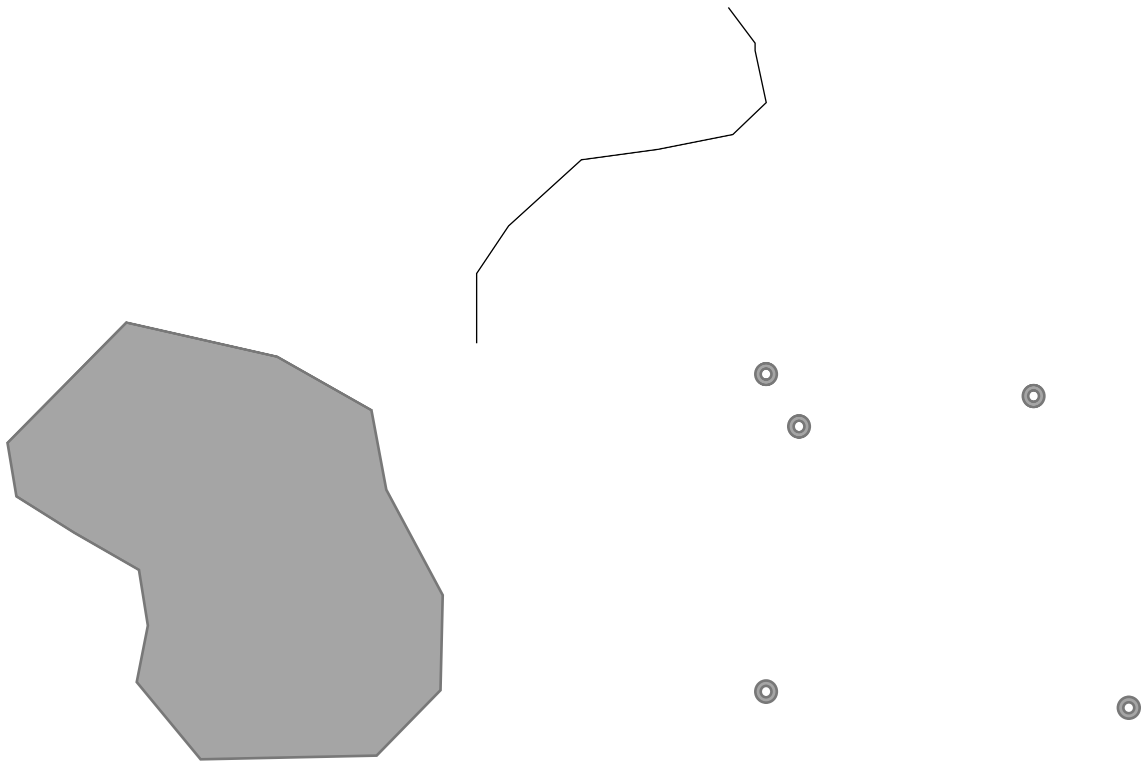
- Characteristics of the feature itself:
  - E.g. trees, buildings, rivers, roads, …
- Your mapping/analysis/research goals
  - Do I want to know about lengths, areas, locations?

Locations/addresses of properties

Area of lots

Jogging route

Vector Format: Features and Attribute



Table

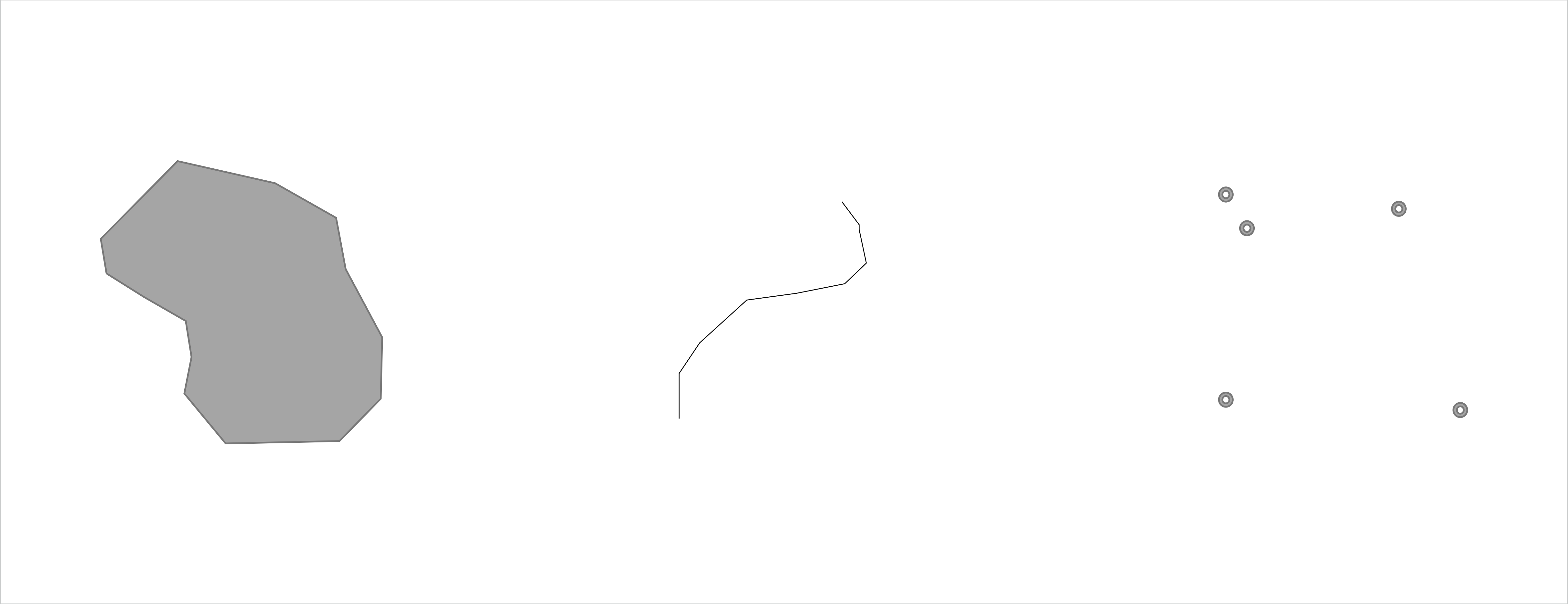
STATES

FID	Shape *	AREA	STATE_NAME
0	Polygon	174204.644	Washington
1	Polygon	381307.987	Montana
2	Polygon	83279.978	Maine
3	Polygon	183403.506	North Dakota
4	Polygon	199912.249	South Dakota
5	Polygon	253293.646	Wyoming
6	Polygon	145261.903	Wisconsin
7	Polygon	215877.666	Idaho
8	Polygon	24869.345	Vermont
9	Polygon	218899.803	Minnesota
10	Polygon	251441.073	Oregon
11	Polygon	23982.233	New Hampshire
12	Polygon	145703.864	Iowa
13	Polygon	21165.05	Massachusetts
14	Polygon	200277.163	Nebraska
15	Polygon	125776.995	New York
16	Polygon	117474.64	Pennsylvania

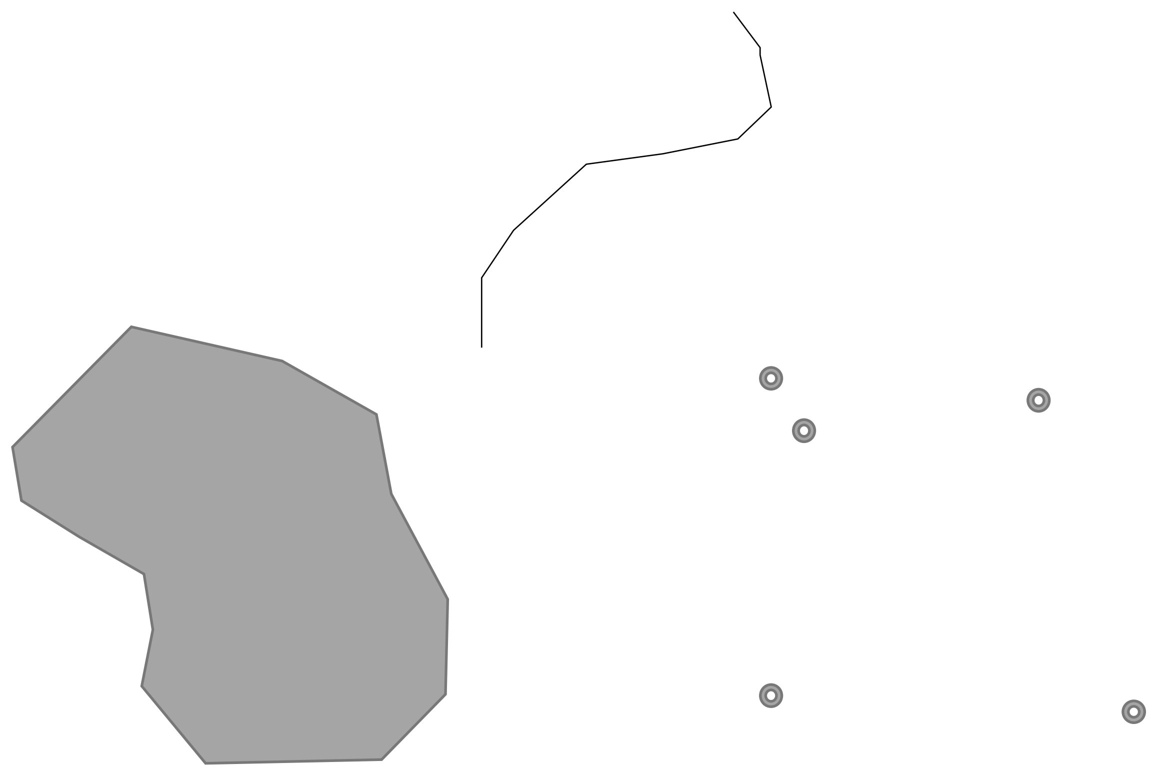
Vector Format: Features and Attributes

- The vector data model joins spatial locations (features) to attribute data.
  - Features: set of x- and y- coordinates
  - Attributes: data in the row-data format

A feature can be a complicated object: it stores the location information.



Vector Format: **Features** and Attributes



Table

STATES

FID	Shape *	AREA	STATE_NAME
0	Polygon	174204.644	Washington
1	Polygon	381307.987	Montana
2	Polygon	83279.978	Maine
3	Polygon	183403.506	North Dakota
4	Polygon	199912.249	South Dakota
5	Polygon	253293.646	Wyoming
6	Polygon	145261.903	Wisconsin
7	Polygon	215877.666	Idaho
8	Polygon	24869.345	Vermont
9	Polygon	218899.803	Minnesota
10	Polygon	251441.073	Oregon
11	Polygon	23982.233	New Hampshire
12	Polygon	145703.864	Iowa
13	Polygon	21165.05	Massachusetts
14	Polygon	200277.163	Nebraska
15	Polygon	125776.995	New York
16	Polygon	117474.64	Pennsylvania

Shapefile

Actually, a collection of many files that store:

- Feature data: coordinates of the vertices for each feature.
- Attribute data: associates each feature with a row in the data table. Often a .dbf file.
- Metadata: e.g. projection information

Table

STATES

FID	Shape *	AREA	STATE_NAME	STATE_FIPS	SUB_REGION	STATE_ABBR	POP1990
0	Polygon	174204.644	Washington	53	Pacific	WA	4866692
1	Polygon	381307.987	Montana	30	Mtn	MT	799065
2	Polygon	83279.978	Maine	23	N Eng	ME	1227928
3	Polygon	183403.506	North Dakota	38	W N Cen	ND	638800
4	Polygon	199912.249	South Dakota	46	W N Cen	SD	696004
5	Polygon	253293.646	Wyoming	56	Mtn	WY	453588
6	Polygon	145261.903	Wisconsin	55	E N Cen	WI	4891769
7	Polygon	215877.666	Idaho	16	Mtn	ID	1006749
8	Polygon	24869.345	Vermont	50	N Eng	VT	562758
9	Polygon	218899.803	Minnesota	27	W N Cen	MN	4375099
10	Polygon	251441.073	Oregon	41	Pacific	OR	2842321
11	Polygon	23982.233	New Hampshire	33	N Eng	NH	1109252
12	Polygon	145703.864	Iowa	19	W N Cen	IA	2776755
13	Polygon	21165.05	Massachusetts	25	N Eng	MA	6016425
14	Polygon	200277.163	Nebraska	31	W N Cen	NE	1578385
15	Polygon	125776.995	New York	36	Mid Atl	NY	17990455
16	Polygon	117474.64	Pennsylvania	42	Mid Atl	PA	11881643

Key Point!

Vector data format separates the spatial representation from the data. Features and attributes are associated, but *separate.*

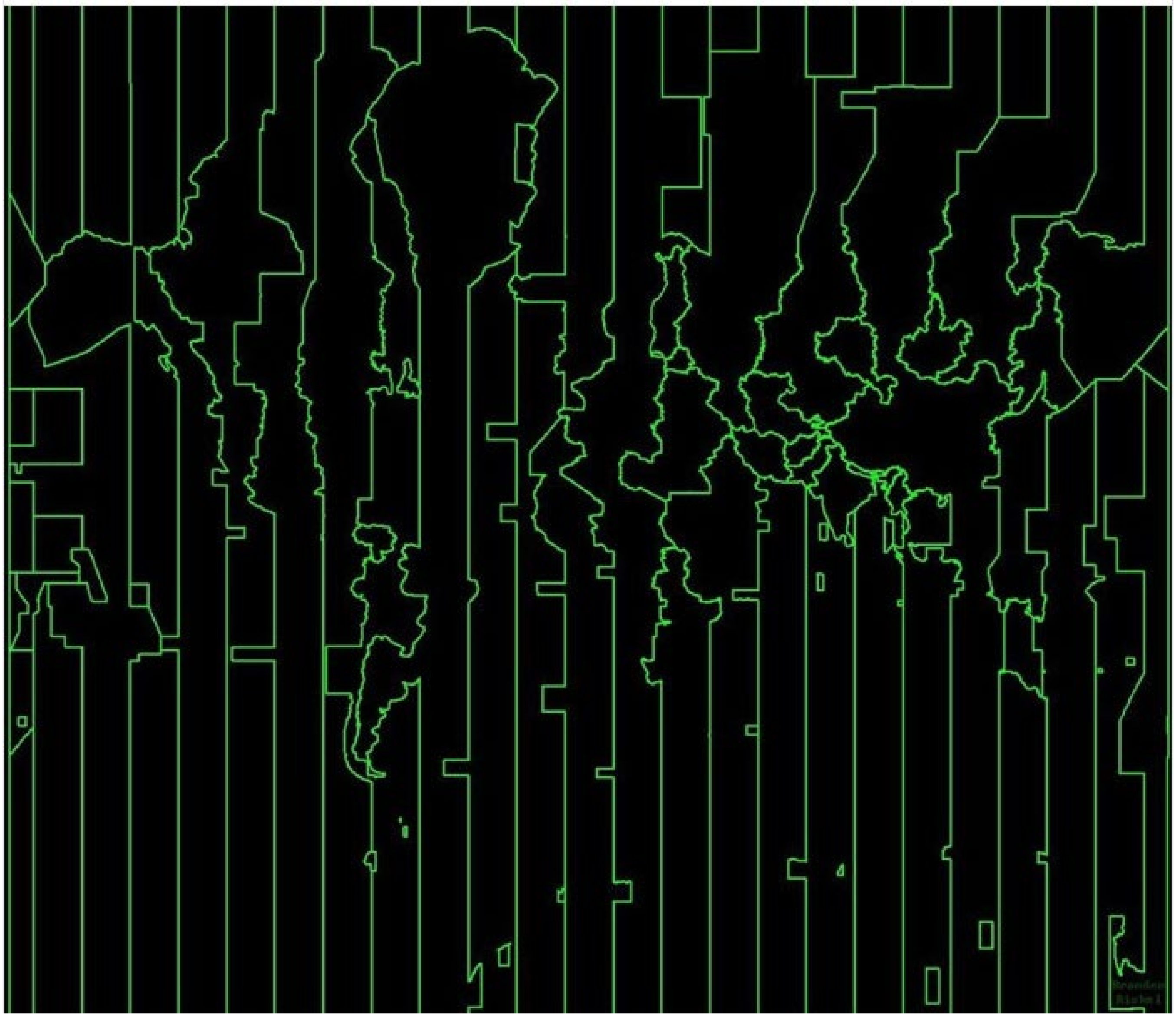
In contrast to raster data, where the spatial representation *is* the data!

Key Point!

Vector data features are represented by *explicit* coordinates: each vertex has an x- and y-coordinate

The locations of raster data elements (cells/pixels) are represented **implicitly**: corner coordinates and row/column ID.

Map puzzler: don't peek at the next slide!



Final Project Sign-Up

A reminder…

Choosing a Final Project Topic

How to select a project?

- Explore the links to data sources, project ideas, etc.
- Talk with your peers.
- Use your own ideas or data.
- What is interesting to you?
- We're only in lecture 3, but… the summer schedule is compressed so you need to start thinking about a final project.

How to do the assignment?

- Don't worry about analyses yet!
- You'll get feedback on your ideas before you worry about specific analyses to carry out!
- Think about the big-picture.
- Read the signup/description assignment for report details.

How are the labs going?

What’s good, bad, or meh???

Selecting and Joining

Features and Attributes

Select by attribute

- Selects features whose attributes match your criteria.
- Only works if your layer has the column you need.
- For example: states with low populations

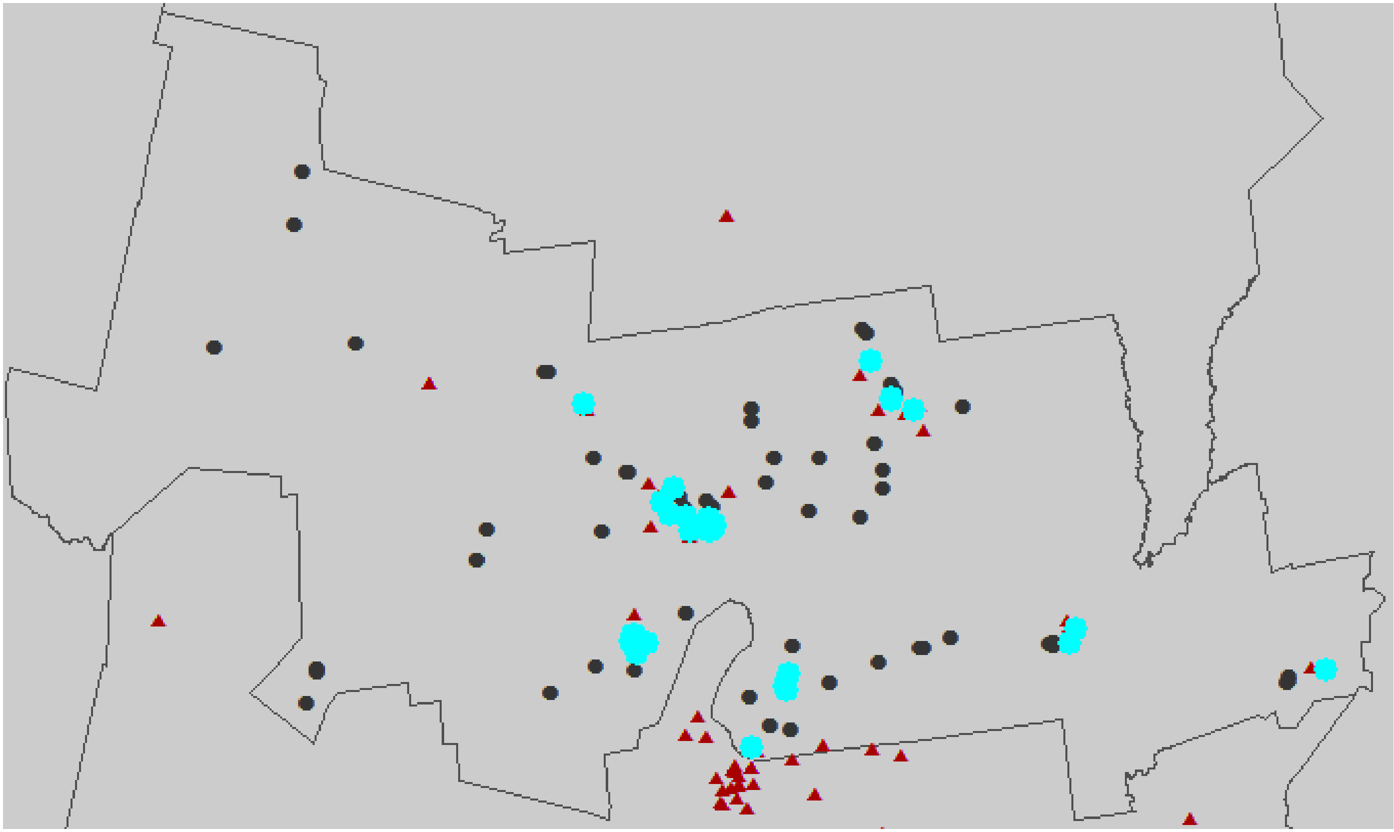
Table

STATES

FID	Shape *	AREA	STATE_NAME	STATE_FIPS	SUB_REGION
41	Polygon	133945.485	Alabama	01	E S Cen
50	Polygon	1493212.904	Alaska	02	Pacific
35	Polygon	294635.701	Arizona	04	Mtn
45	Polygon	137046.401	Arkansas	05	W S Cen
23	Polygon	408555.123	California	06	Pacific
30	Polygon	269626.919	Colorado	08	Mtn
17	Polygon	12890.202	Connecticut	09	N Eng
27	Polygon	5321.616	Delaware	10	S Atl
26	Polygon	171.127	District of Columbia	11	S Atl
47	Polygon	144555.37	Florida	12	S Atl
43	Polygon	151852.428	Georgia	13	S Atl
49	Polygon	16527.639	Hawaii	15	Pacific
7	Polygon	215877.666	Idaho	16	Mtn
25	Polygon	145811.733	Illinois	17	E N Cen
20	Polygon	94274.766	Indiana	18	E N Cen
12	Polygon	145703.864	Iowa	19	W N Cen
32	Polygon	212888.548	Kansas	20	W N Cen
31	Polygon	104432.786	Kentucky	21	E S Cen

Select by location

- Selects features in one layer that overlap one or more features in another layer.
- Can be helpful if your target layer doesn't have the column you need.
- For example: Schools that lie within a county.



Join by attribute value: common columns

ZONE_NAME	COUNT
Selkirk Mountains	515
Cabinet-Yaak	1125
Northern Continental Divide	2450
Private Inholdings-Yaak	0
Bitterroot Mountains	250
Greater Yellowstone	360

FID	Shape *	ZONE_NAME
0	Polygon	Selkirk Mountains
1	Polygon	Cabinet-Yaak
2	Polygon	Northern Continental Divide
3	Polygon	Private Inholdings-Yaak
4	Polygon	Bitterroot Mountains
5	Polygon	Greater Yellowstone

Beware: column names might not be the same!

- Column names might not match.
- Column contents might not exactly match:
  - Alternate spellings
  - Mismatch between upper/lower case.
- Tables might not be the same size

Which column could be used to join this information together?

eteam

id	teamname	coach
POL	Poland	Franciszek Smuda
RUS	Russia	Dick Advocaat
CZE	Czech Republic	Michal Bilek
GRE	Greece	Fernando Santos
...		

game

id	mdate	stadium	team1
1001	8 June 2012	National Stadium, Warsaw	POL
1002	8 June 2012	Stadion Miejski (Wroclaw)	RUS
1003	12 June 2012	Stadion Miejski (Wroclaw)	GRE
1004	12 June 2012	National Stadium, Warsaw	POL

Attribute values must match.
Column names may be different.

Student #	Major	Advisor
1110	Science	Dr. Who
1120	Art	Dr. No
1130	Math	Dr. Oz
1140	Geography	Dr. Phil

Student Name	Class	ID
Bob	Junior	1110
Jim	Junior	1120
Sally	Freshman	1130
Greta	Senior	1140

Important Spatial Data Concepts

What is Scale???

- It’s an unexpectedly complicated and deep question!
- Two important components:
  - Extent
  - Grain

Extent and Grain

Extent: How large is the area?

Grain: How much can I zoom in?

Tradeoff in file size.

Vector data advantages

Scalable: you can zoom

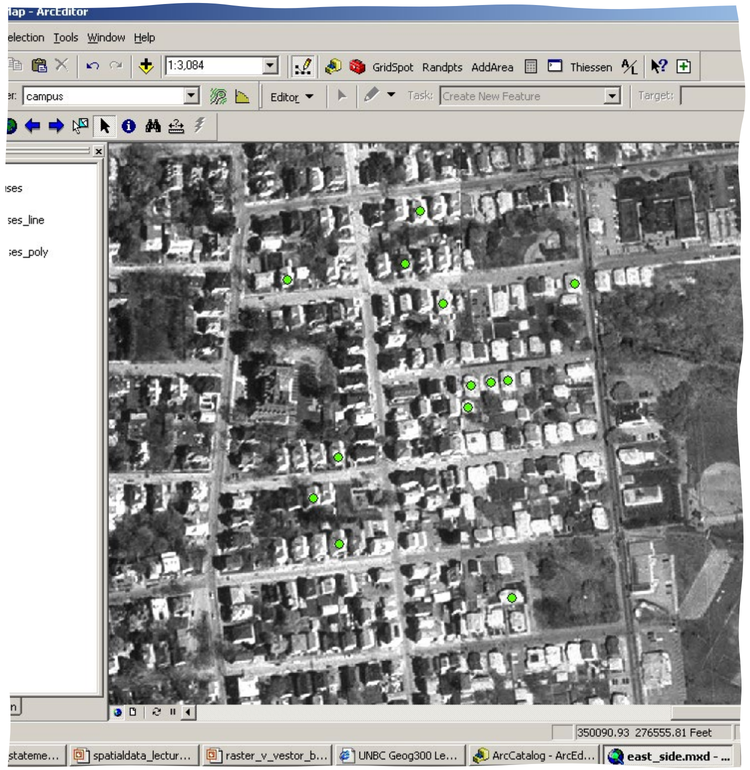
Easy to change coordinate system.

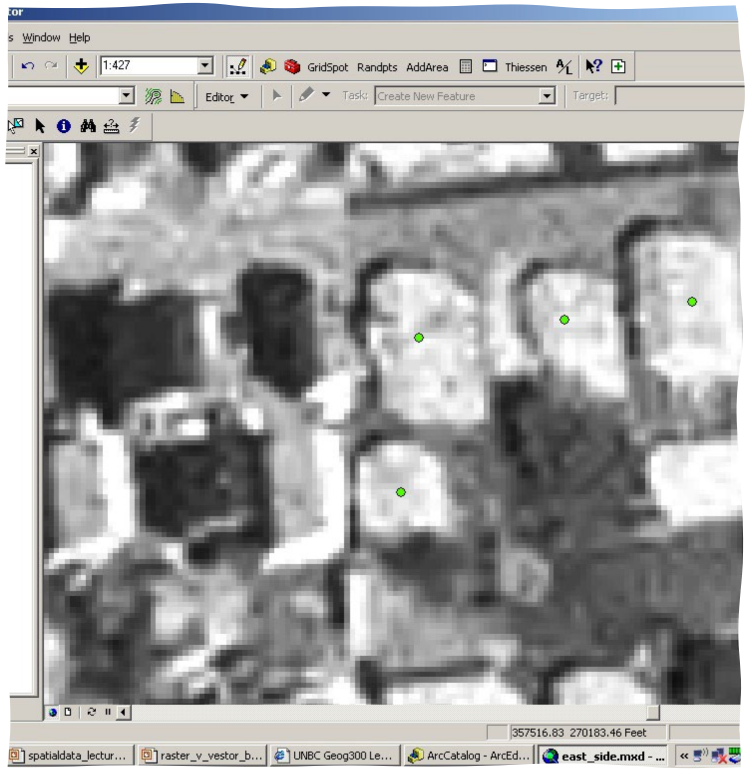
Transformations are lossless.

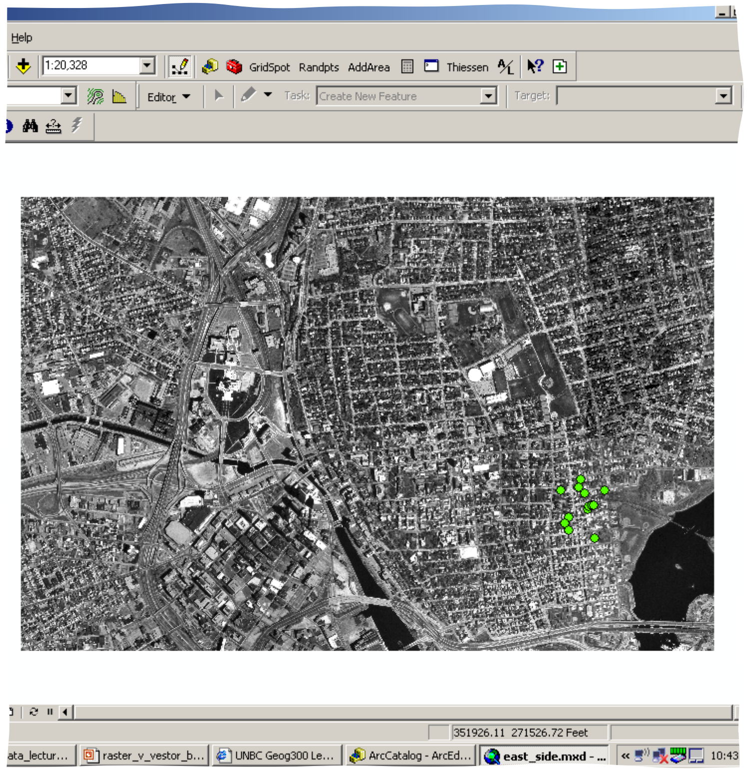
Vectors never look pixelated.*

*But they can look jagged if you zoom in too much.

Vector data never look pixelated!

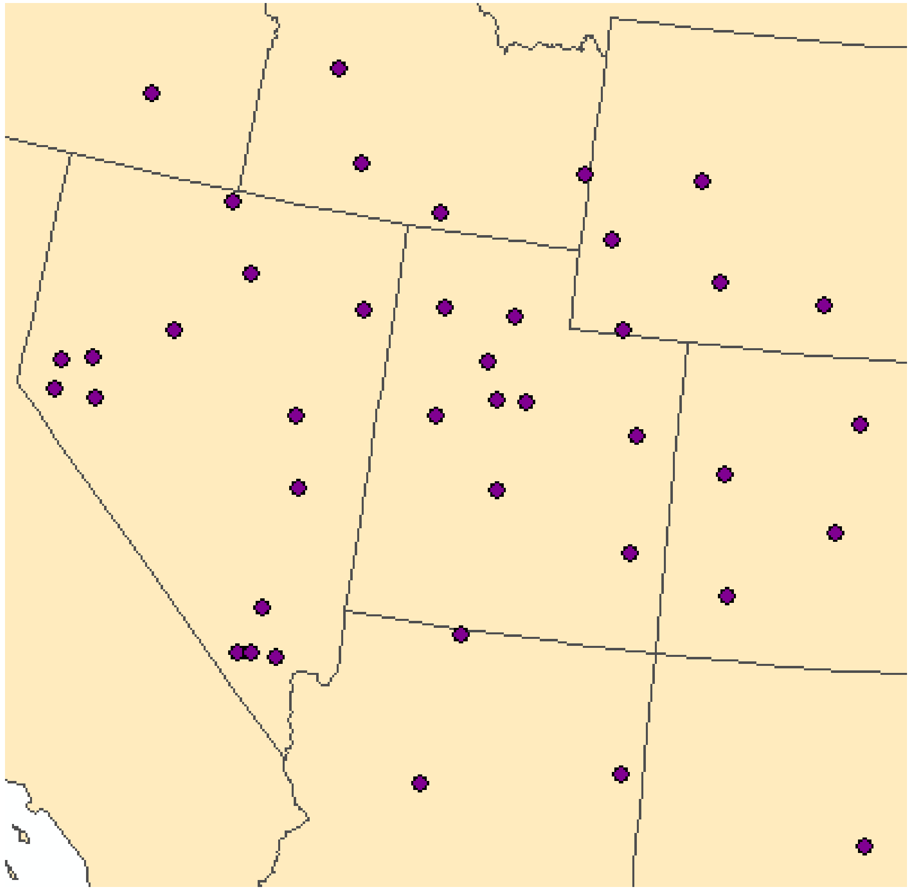


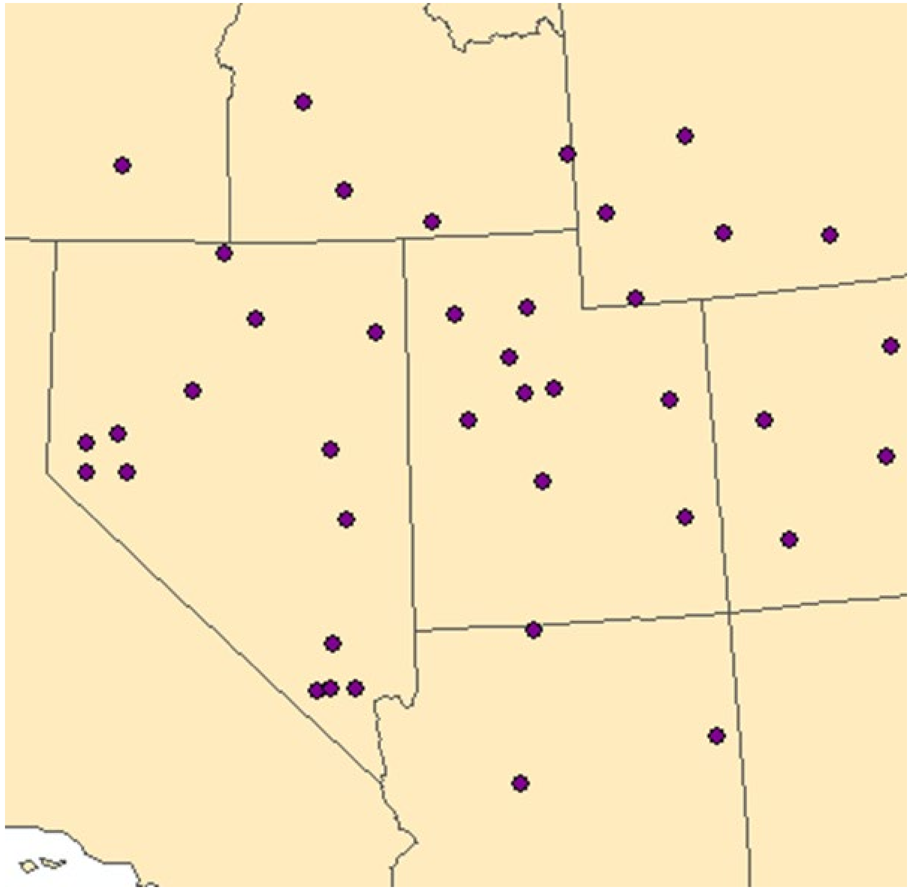




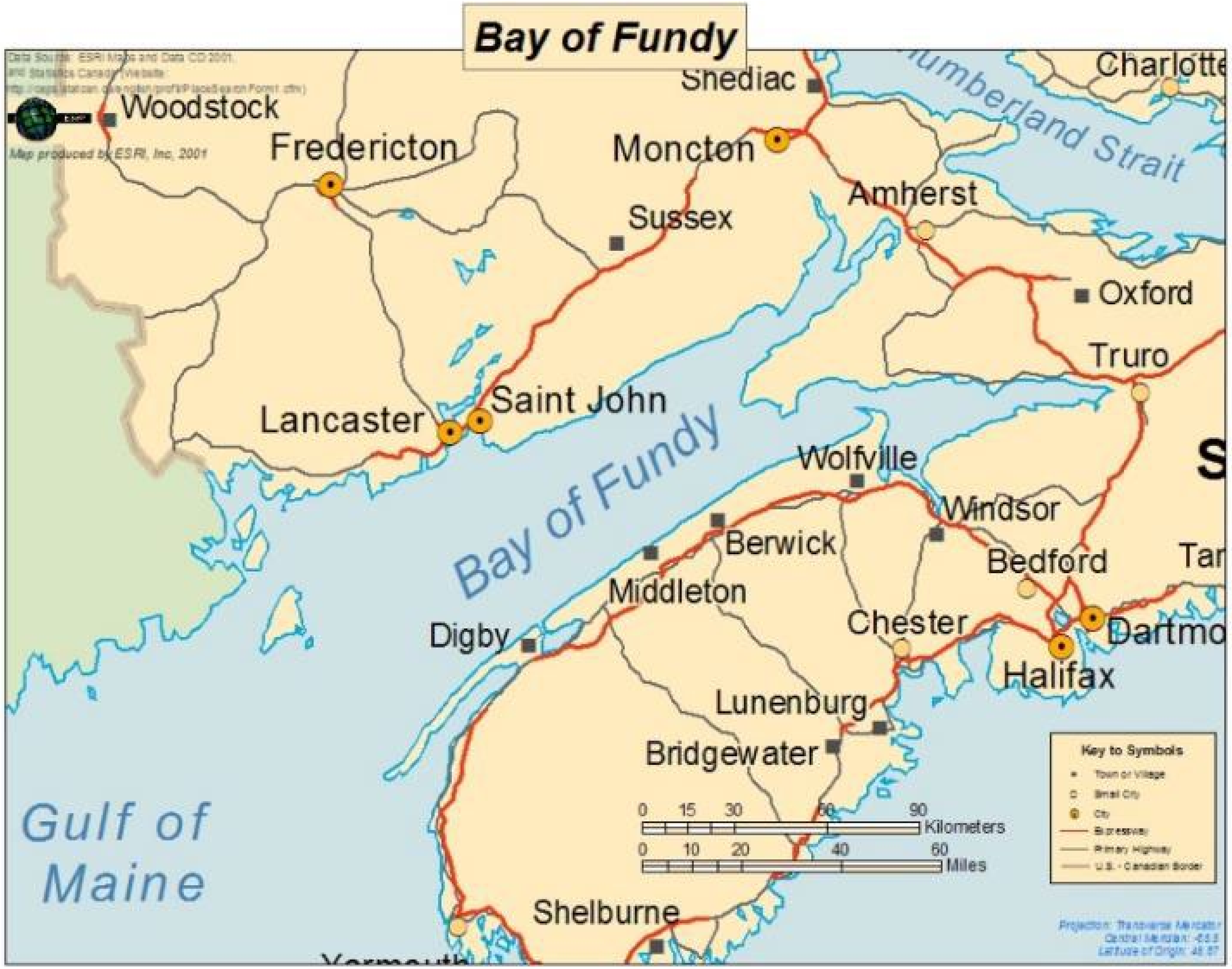
Vector data are easy to reproject.

Changes in coordinate systems are reversible.



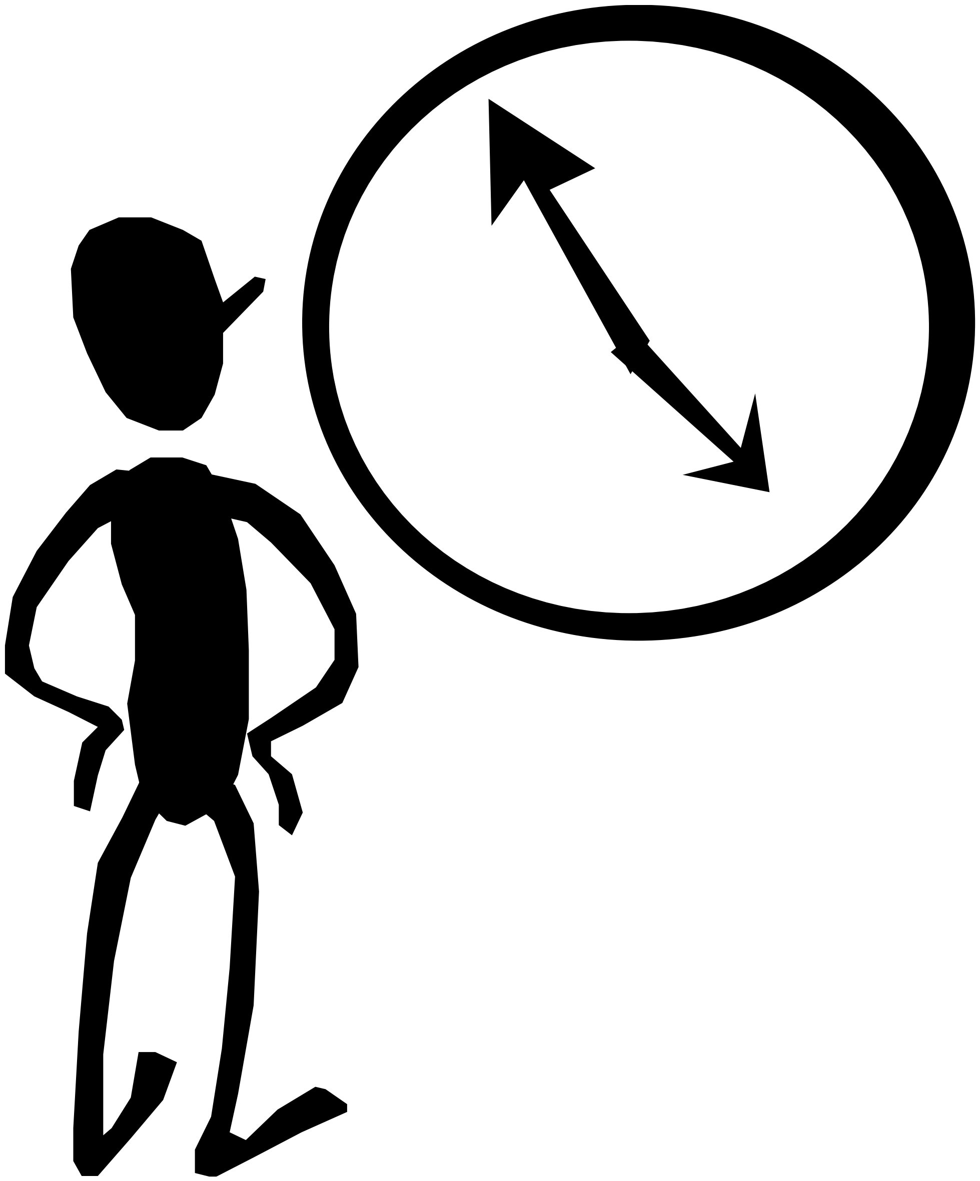


- Vector data often look good on a map
- Especially great for data aggregated at a feature level.



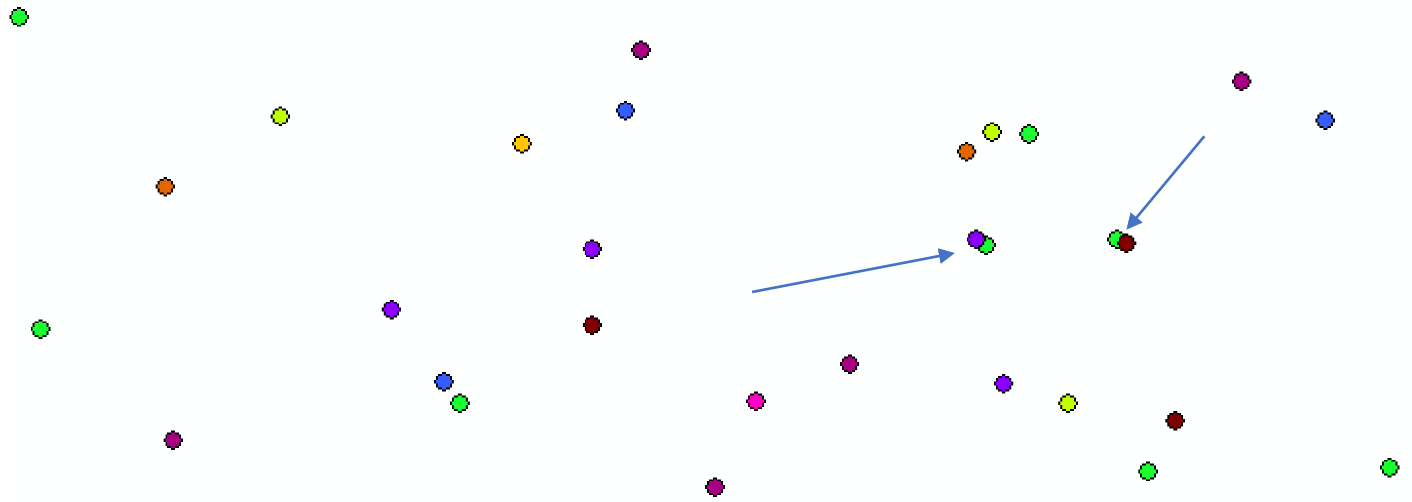
Disadvantages of Vector Data

- High resolution (small grain) data require lots of space and processing time



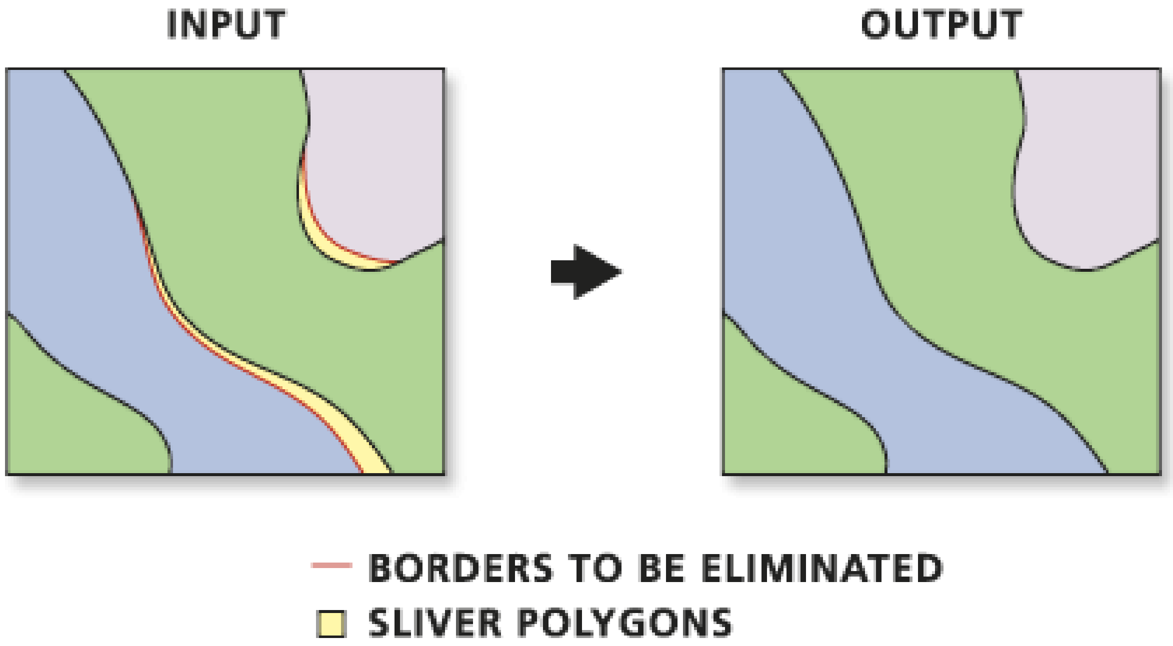
Disadvantages of Vector Data

- Data may not be continuous spatially
- Small gaps
- Round-off errors
- Misalignment



Disadvantages of Vector Data

- Data may not be continuous spatially
- Small gaps, a.k.a. ‘sliver polygons’
- Round-off errors
- Misalignment

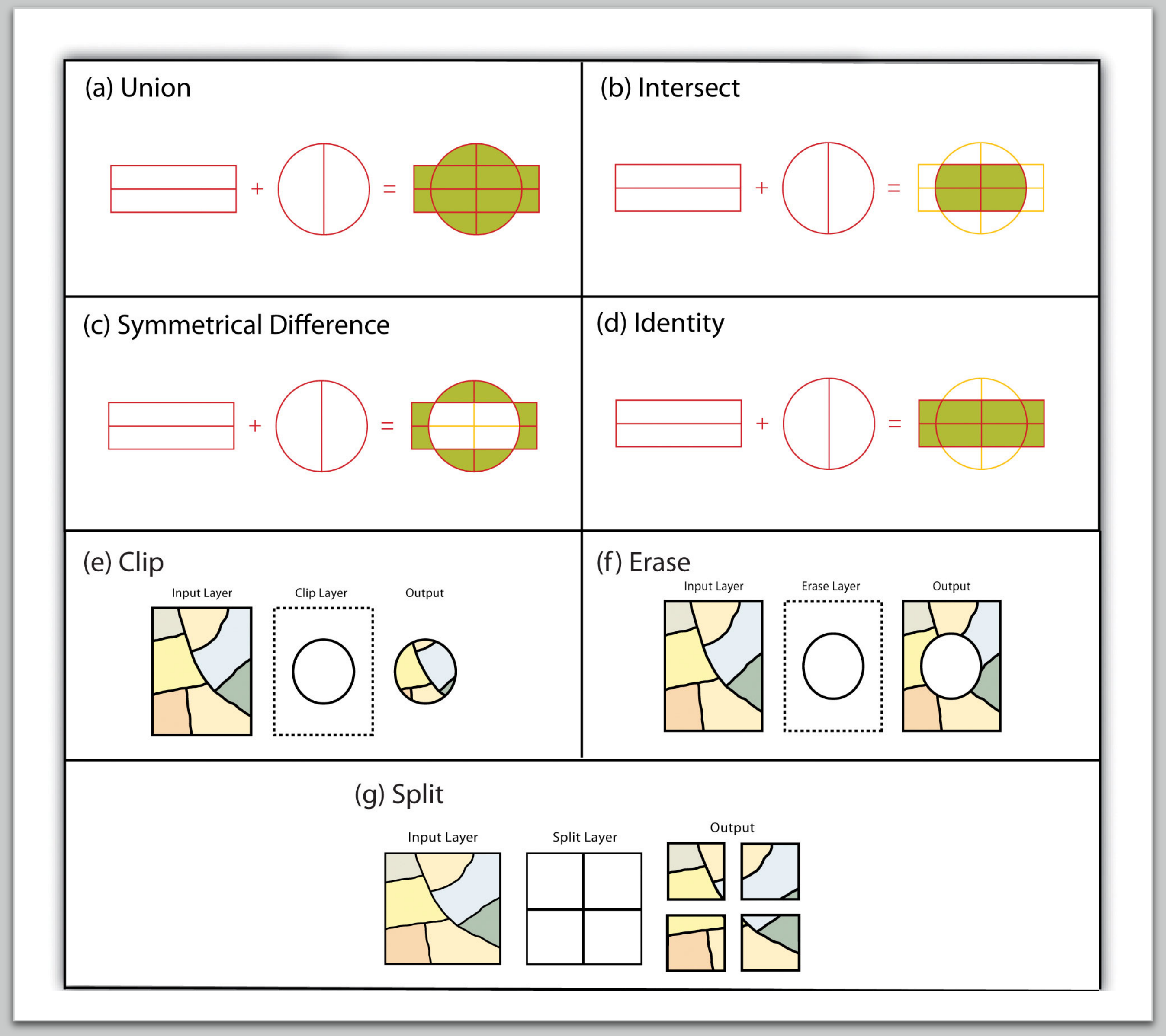


Announcements

- If you're behind on labs, reach out ASAP! Labs 3 and 4 are due Monday.
- Ollie's Thursday office hours for today: 8-ish to 9:30PM
  - Just for this week
  - I'll stick around for 30-ish minutes after class today.
- Midterm opens on June 16^{th} – you'll have 1 week to complete it.
  - Contains materials from labs 1 – 5 (that's why I'm so naggy about keeping up on labs!)
- Asynchronous lecture next Tuesday. I have a Dr. appointment that I can't reschedule during the normal class time.
  - I'll be posting the recorded lecture to Echo30 on Monday or Tuesday.

Vector Data Operations and Analysis

Vector Operations



Vector Operators, Set Theory, Boolean Algebra

Vector Operator	Set Theory	Boolean	English
Union	A∪B	OR A|B	Elements that are in at least one of A or B
Intersection	A∩B	AND A&B	Elements are on both A and B
Symmetric Difference	A∪B - A∩B	Exclusive OR, XOR A|B – A&B	Elements are in A or B, but not both

Vector Operations: Geoprocessing

Vector operations are also called 'geoprocessing'

Geoprocessing alters the topology

Geoprocessing operations are often destructive: they can't be reversed

Common Geoprocessing Operations

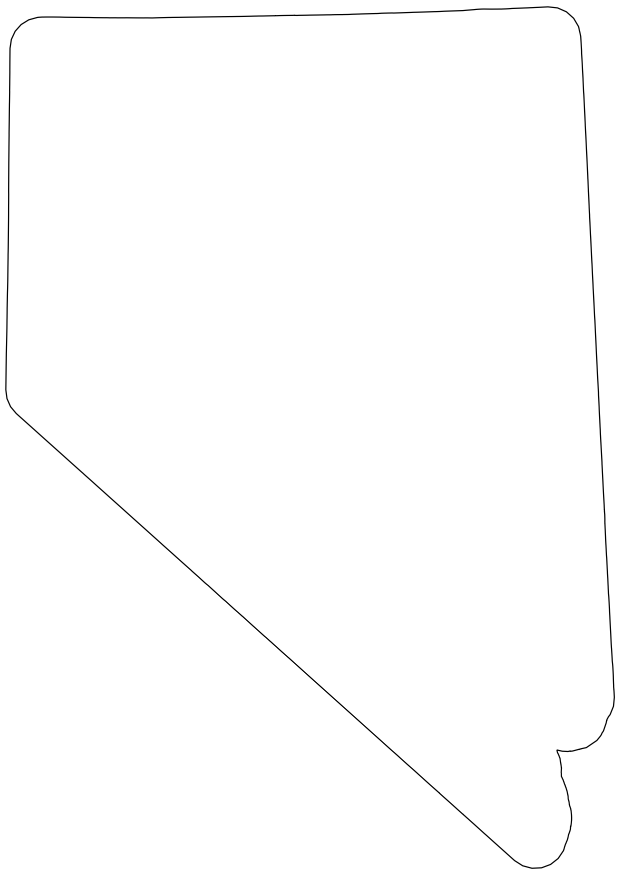
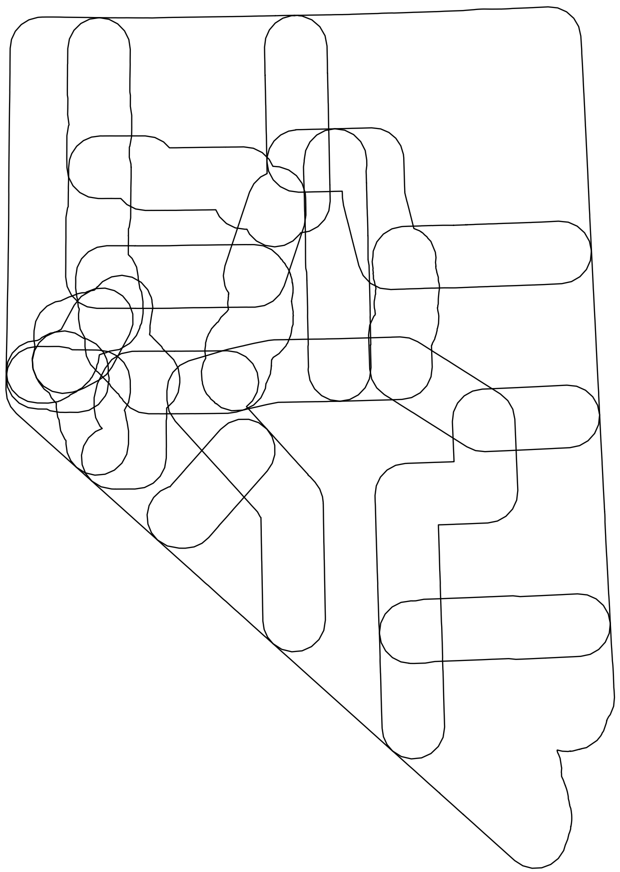
- Buffer (and dissolve)
- Clip
- Erase

The following examples were performed in R!

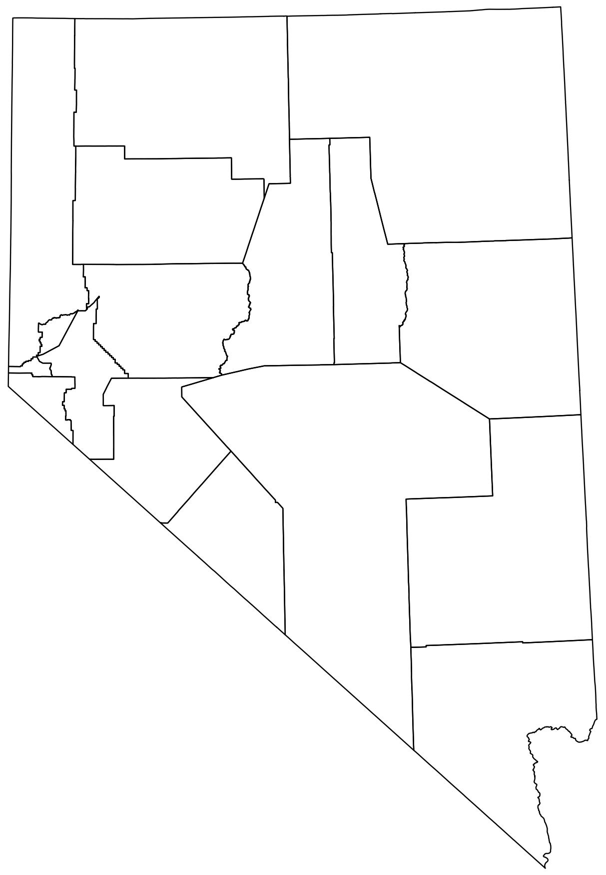
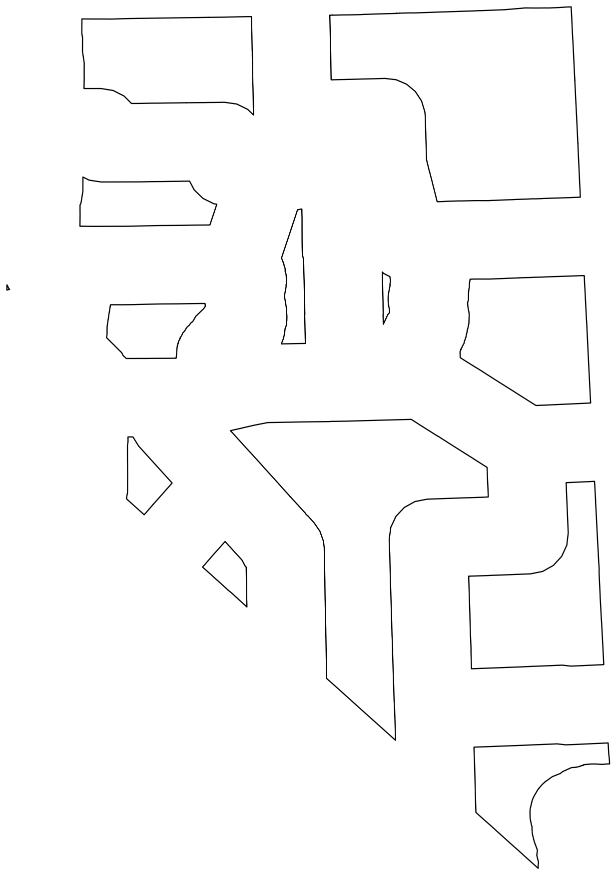
Important packages include:

- sp
- sf
- rgeos
- raster
- terra

Buffer – Buffering Nevada Counties

With Dissolve: puffy Nevada	No Dissolve: puffy Nevada with worms
	

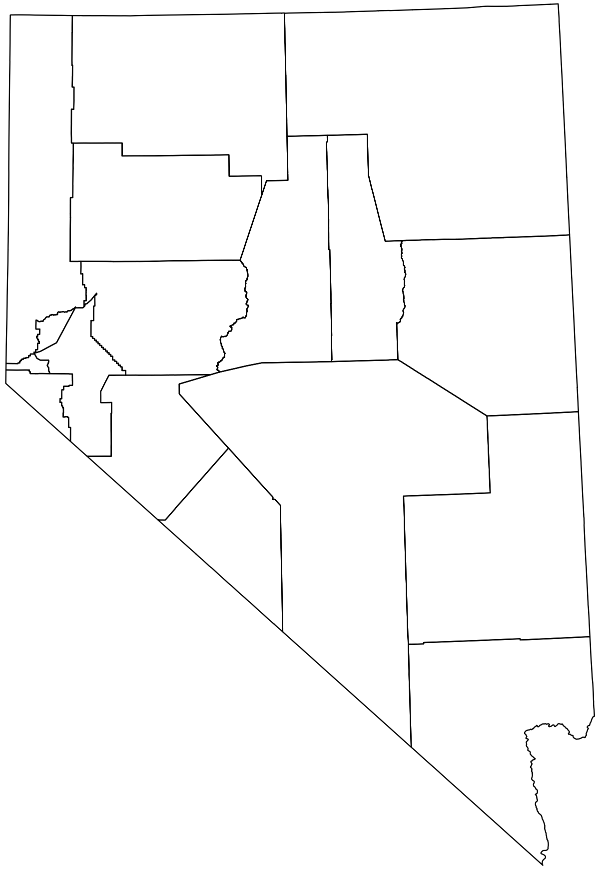
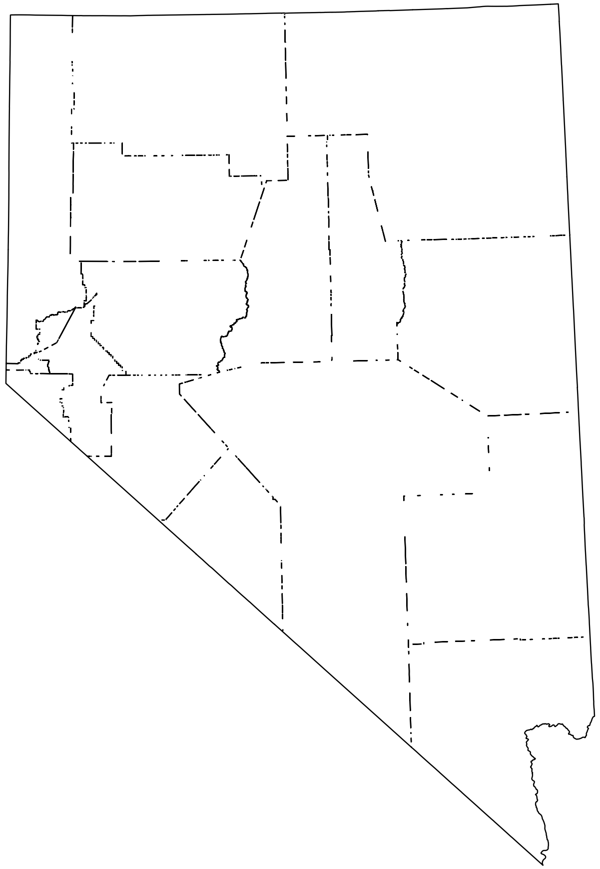
Buffer – Negative Buffer

Original	100km negative buffer
	

Dissolve

Original	Dissolved
	

Dissolve

Original	Dissolved
	

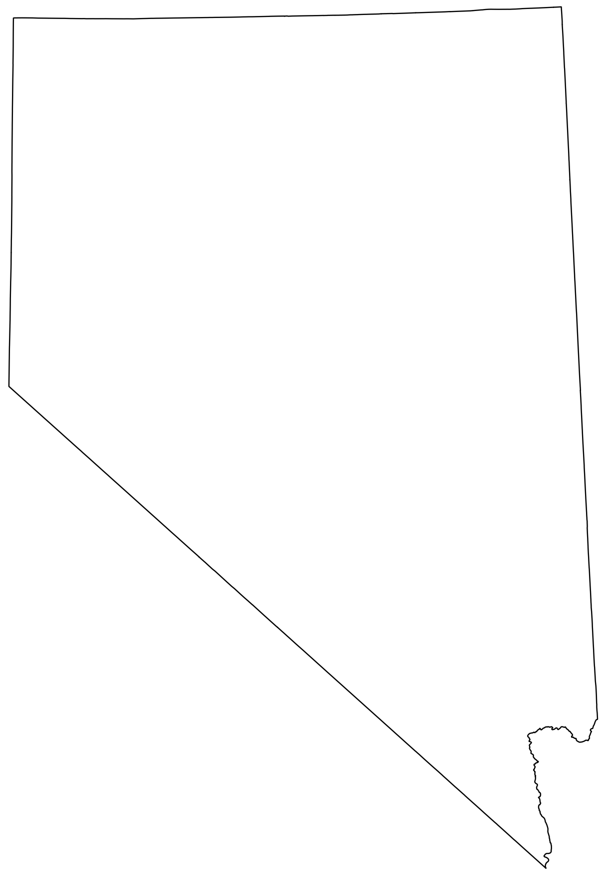
What happened???

- Nevada counties shapefile: county vertices were slightly misaligned
  - Most likely due to round-off errors: edge coordinates are stored as double or float numbers
  - Round-off errors can happen when you reproject, or if decimal values are truncated.

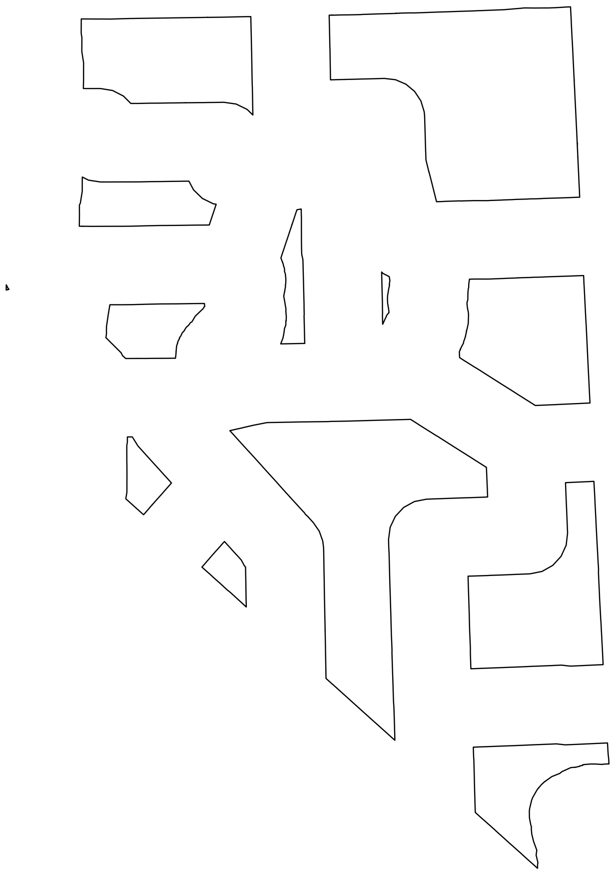
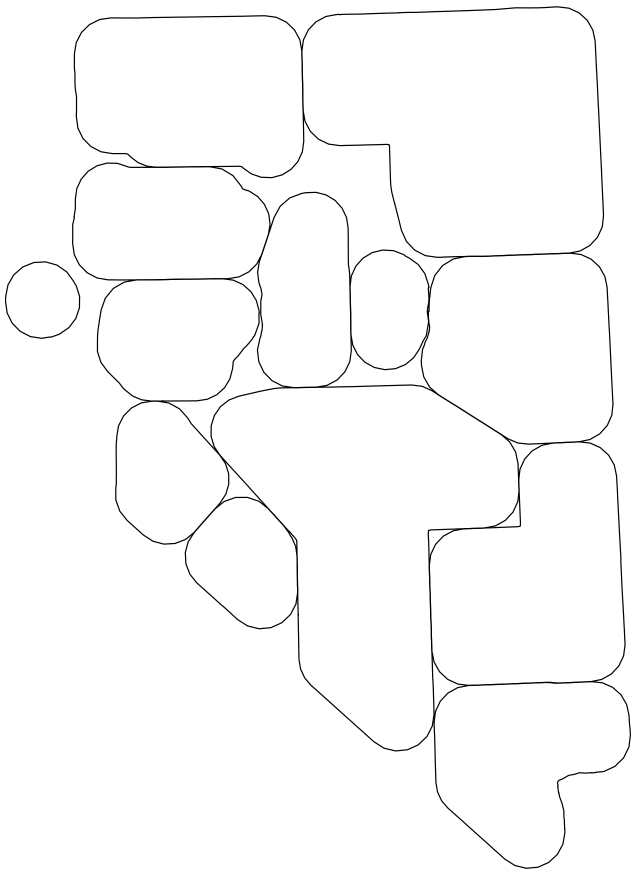
What happened???

- We're left with micropolygons: sliver polygons
  - Sliver polygons are hard to get rid of
- How to fix?
  - First buffer by a small amount, then dissolve

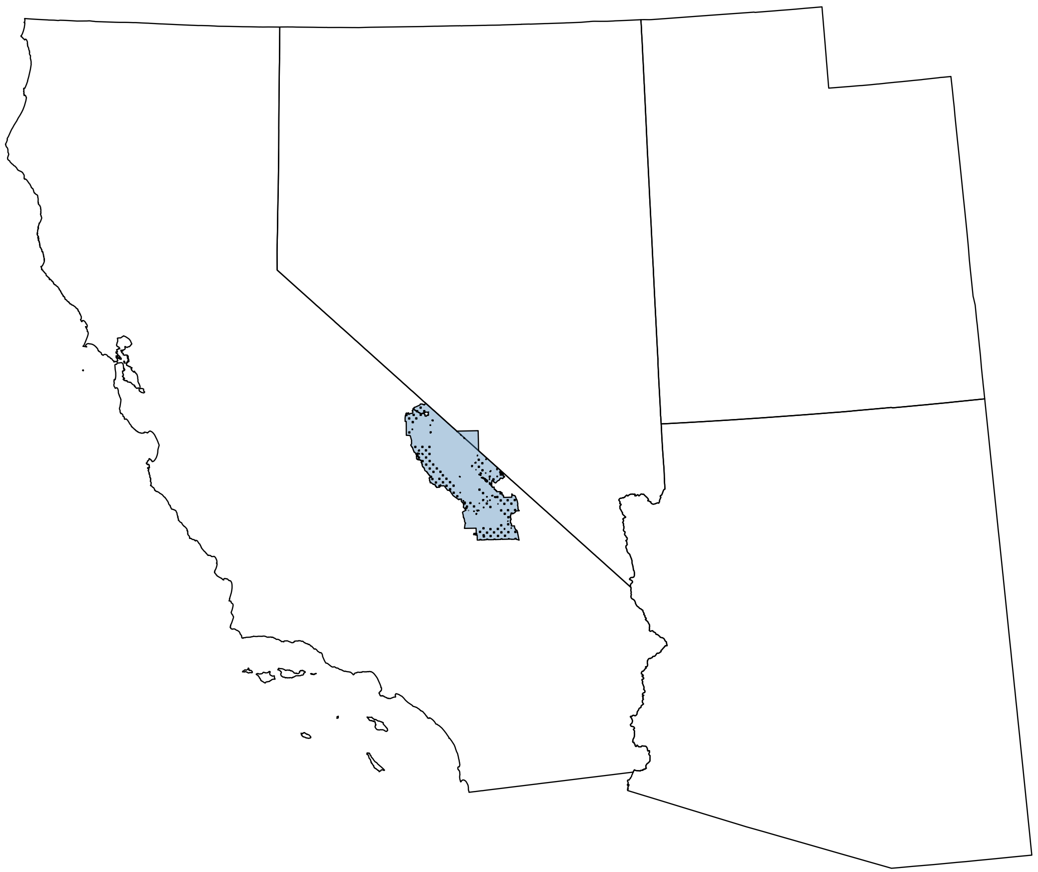
Trick to fix mis-aligned polygons

Procedure to fix it	Success!
1. First, buffer by a tiny, positive, amount. In this case, 10 meters worked. 2. Next, perform the dissolve	

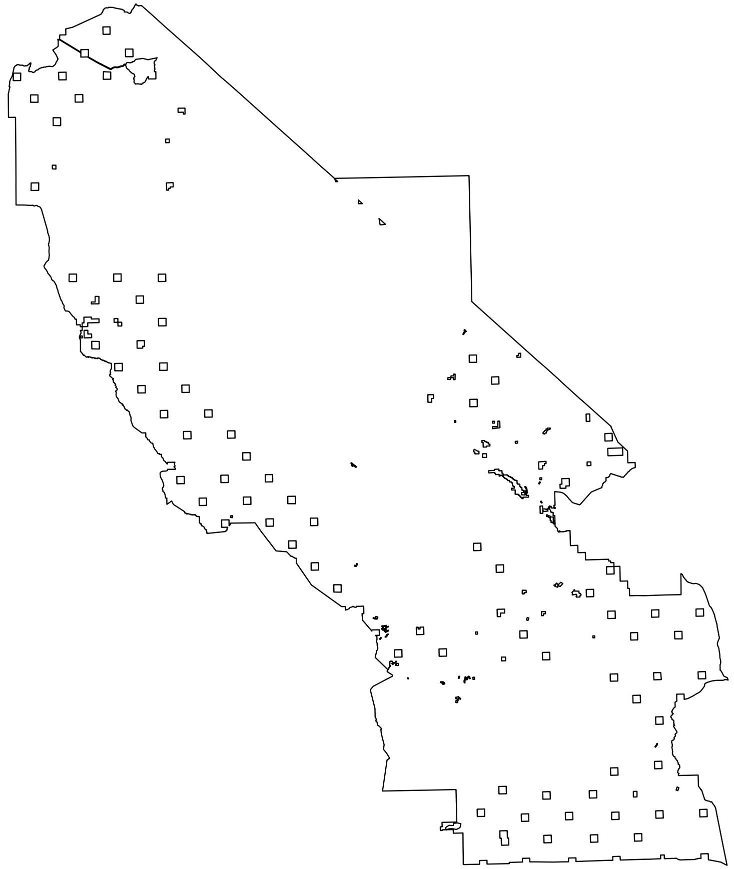
Buffering is Destructive

Negative Buffer	Negative + Positive Buffer = Marshmallow Counties
	

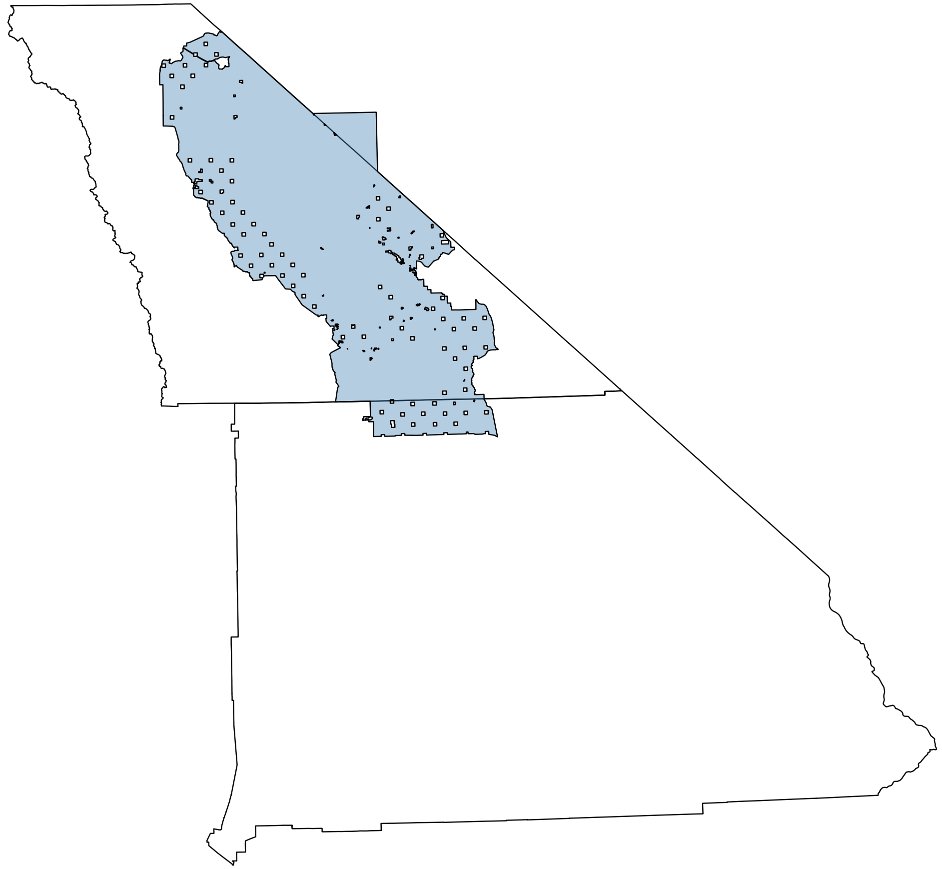
Which polygons intersect?
This is like a select by location operation in Arc GIS

Southwest States + Death Valley Which states does the park cover?	CA + NV
	

Union and Intersection

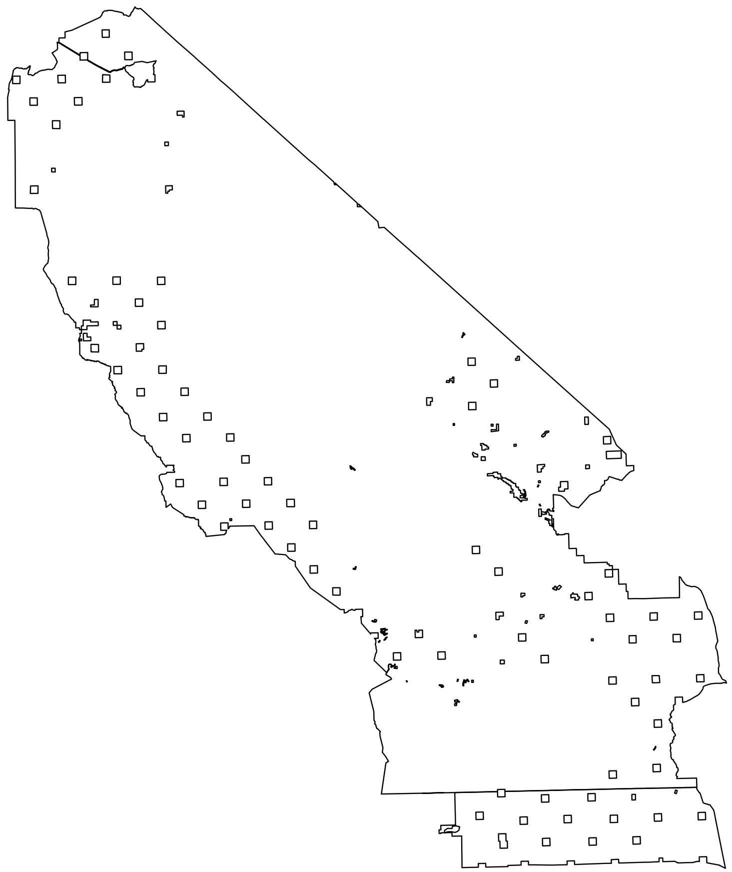
CA Counties	Death Valley National Park
	

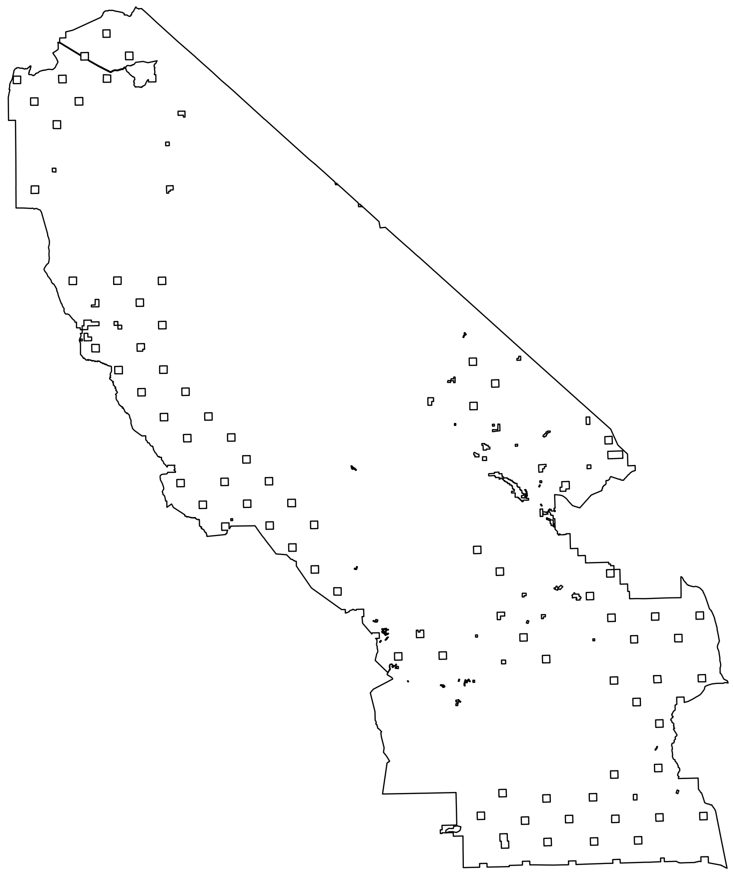
California Counties and Death Valley – Union

CA Counties + Death Valley	Zoom-in
	

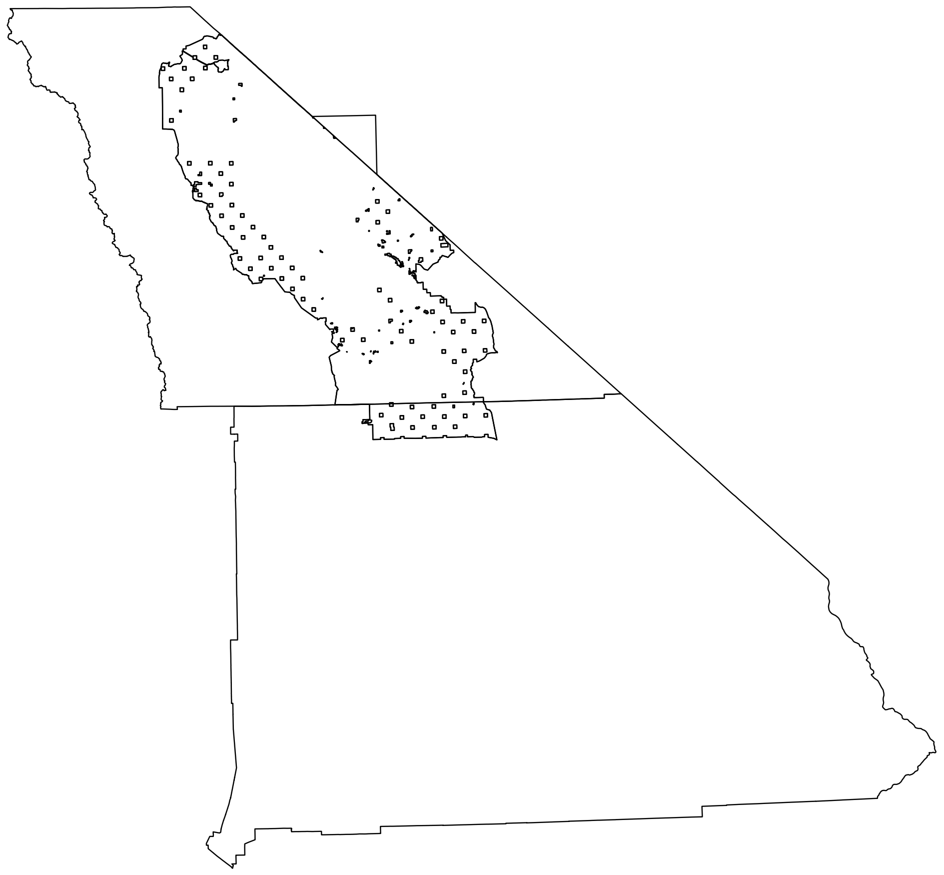
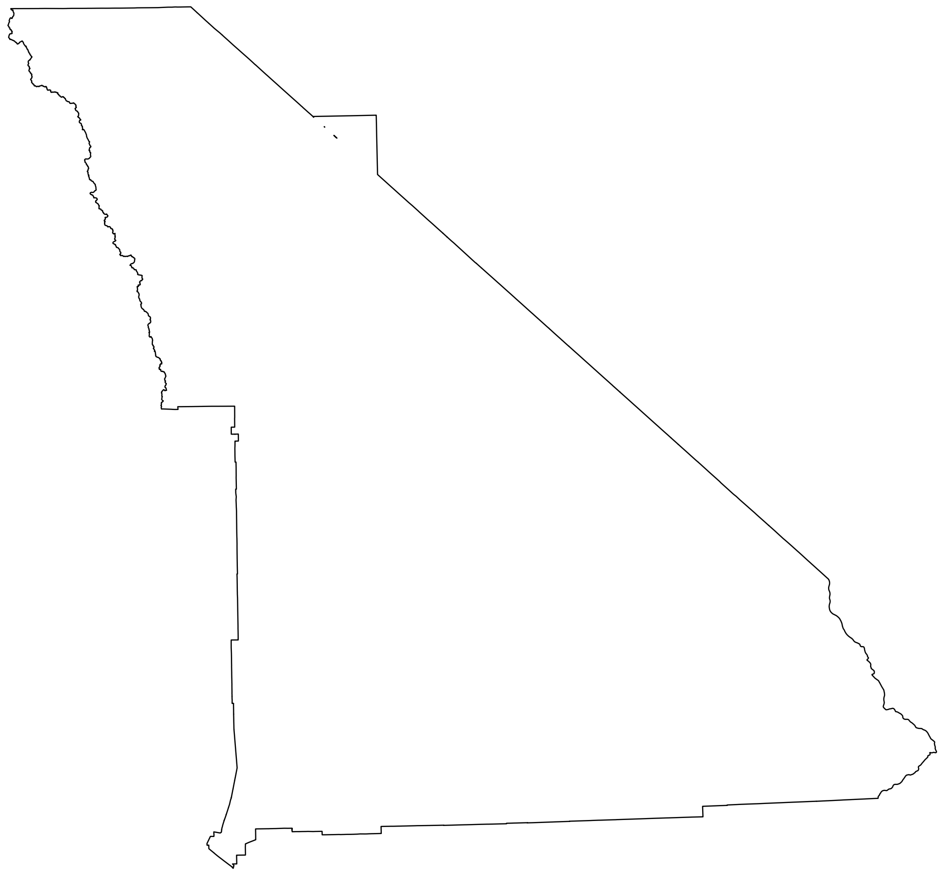
Intersection keeps overlapping parts

No Dissolve	With Dissolve
raster::intersect(dvnp, ca_cnty)	rgeos::gIntersection(dvnp, ca_cnty)





Union: keeps whole features

No Dissolve	With Dissolve
	

Key point!

The result of an intersection operation is usually smaller than either of the two input features.

The result of a union is usually larger than either of the two input features

If you perform both operations on the same two features, the result of the intersection will usually be smaller (unless there is perfect overlap).

Selections and Geoprocessing

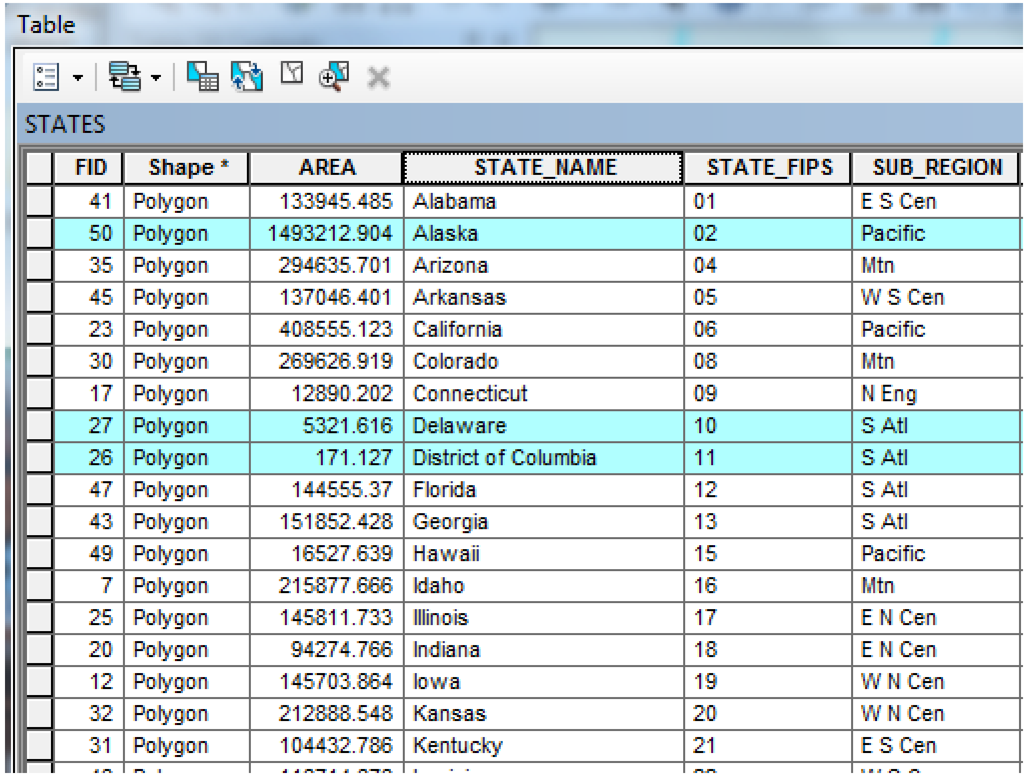
Differences are important.

Hint: you'll need to know when to use each on the midterm!

Suitability Analysis Using Vector Operations and Selections

We've looked at two main select paradigms…

By Attribute

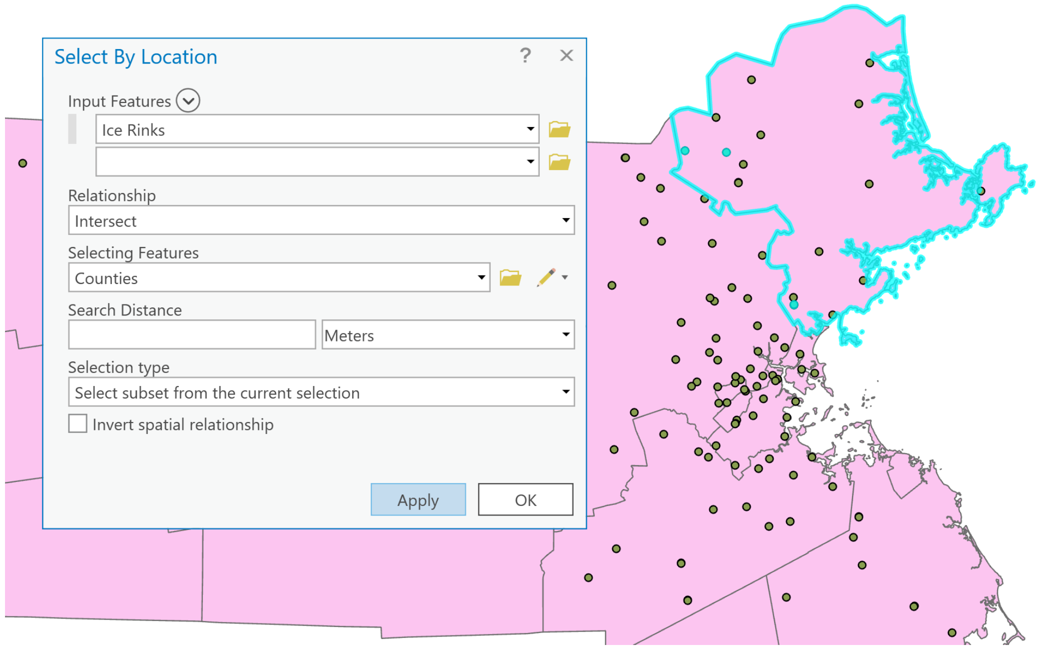


Table

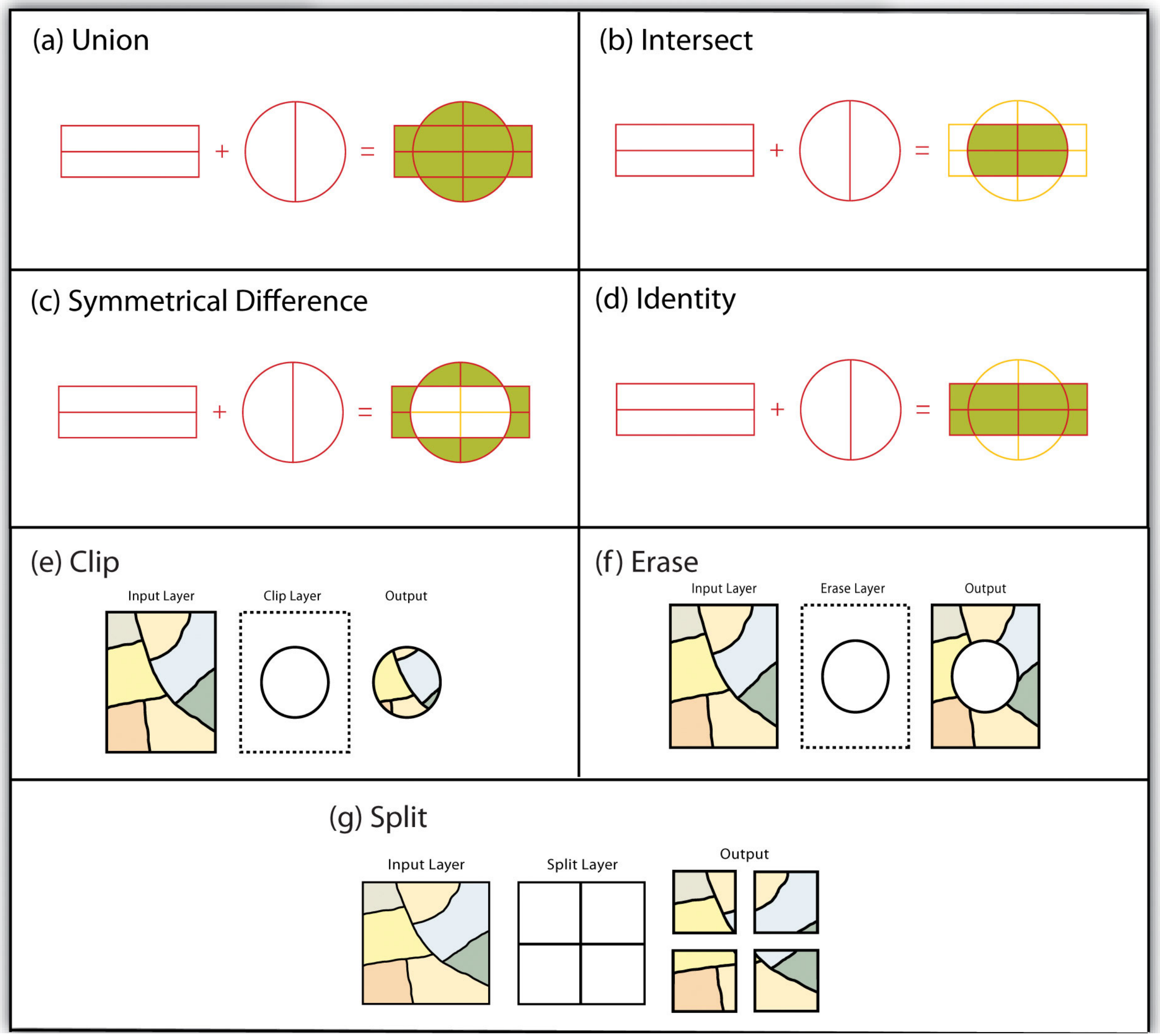
STATES

FID	Shape *	AREA	STATE_NAME	STATE_FIPS	SUB_REGION
41	Polygon	133945.485	Alabama	01	E S Cen
50	Polygon	1493212.904	Alaska	02	Pacific
35	Polygon	294635.701	Arizona	04	Mtn
45	Polygon	137046.401	Arkansas	05	W S Cen
23	Polygon	408555.123	California	06	Pacific
30	Polygon	269626.919	Colorado	08	Mtn
17	Polygon	12890.202	Connecticut	09	N Eng
27	Polygon	5321.616	Delaware	10	S Atl
26	Polygon	171.127	District of Columbia	11	S Atl
47	Polygon	144555.37	Florida	12	S Atl
43	Polygon	151852.428	Georgia	13	S Atl
49	Polygon	16527.639	Hawaii	15	Pacific
7	Polygon	215877.666	Idaho	16	Mtn
25	Polygon	145811.733	Illinois	17	E N Cen
20	Polygon	94274.766	Indiana	18	E N Cen
12	Polygon	145703.864	Iowa	19	W N Cen
32	Polygon	212888.548	Kansas	20	W N Cen
31	Polygon	104432.786	Kentucky	21	E S Cen

By Location



And several geoprocessing operations

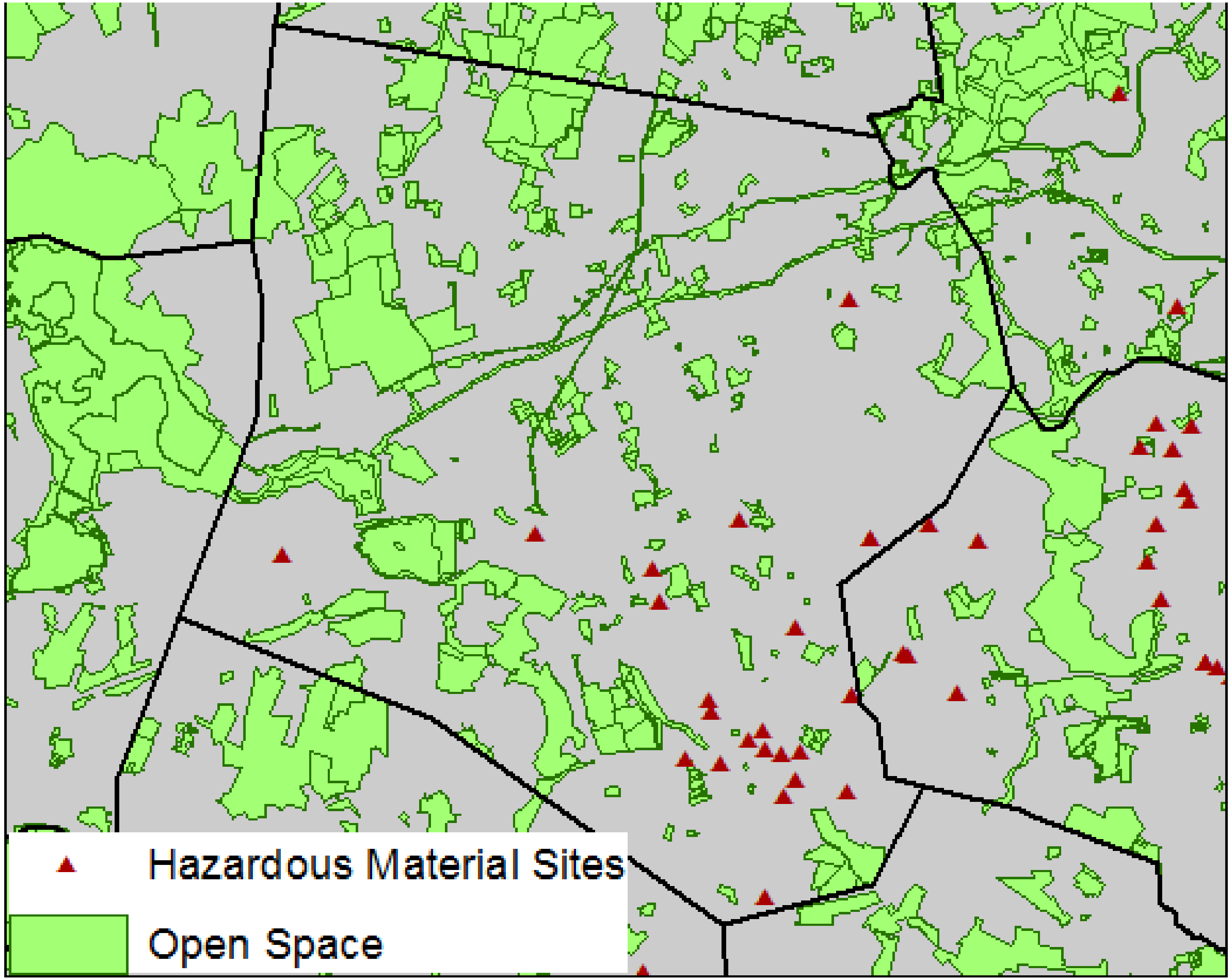


We can combine selection and geoprocessing into a powerful tool for suitability analyses!

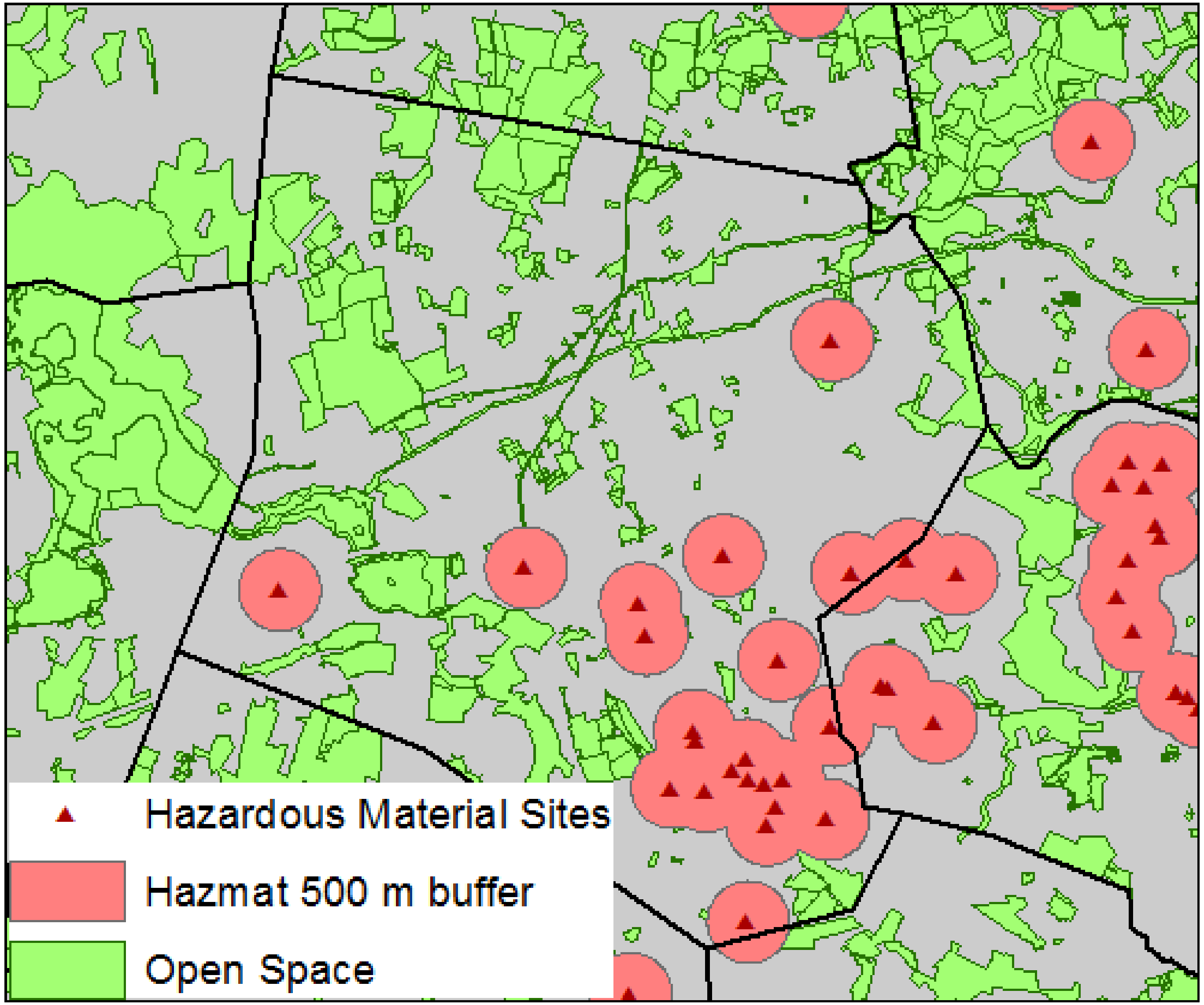
Vector data analysis: Build a school in Framingham

What factors might be important???

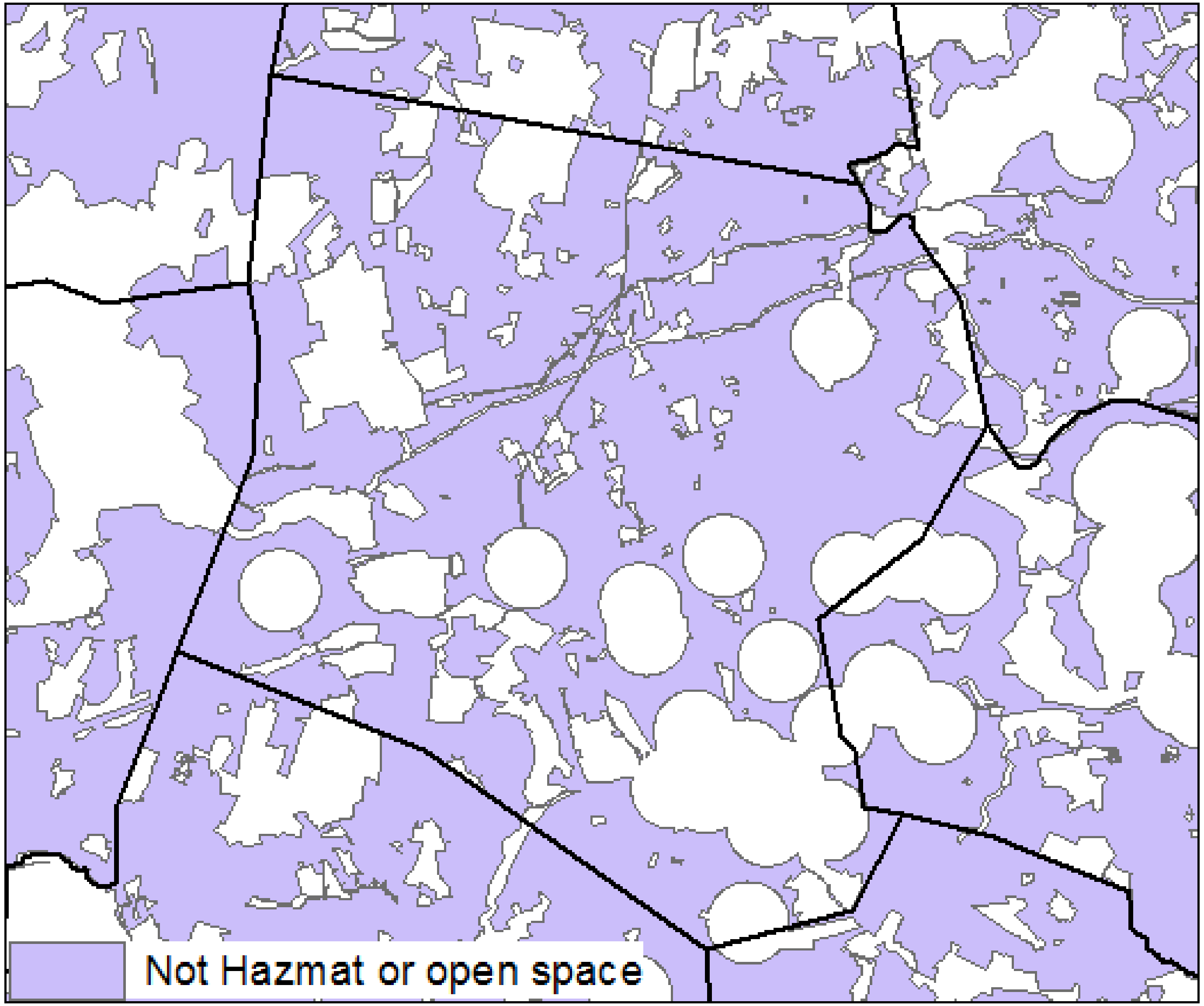
Vector data analysis: Build a school in Framingham



Far from Hazmat sites (buffer tool)



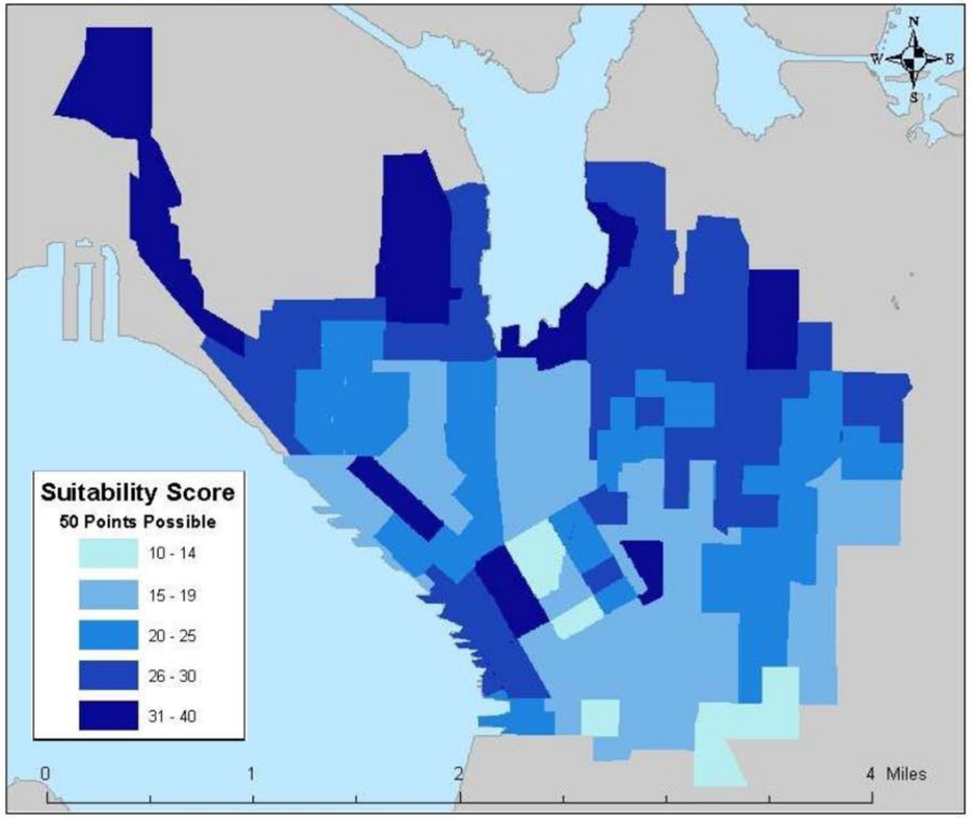
Not in open space OR hazmat buffer



Suitability based on *attributes*

Combined Criteria:

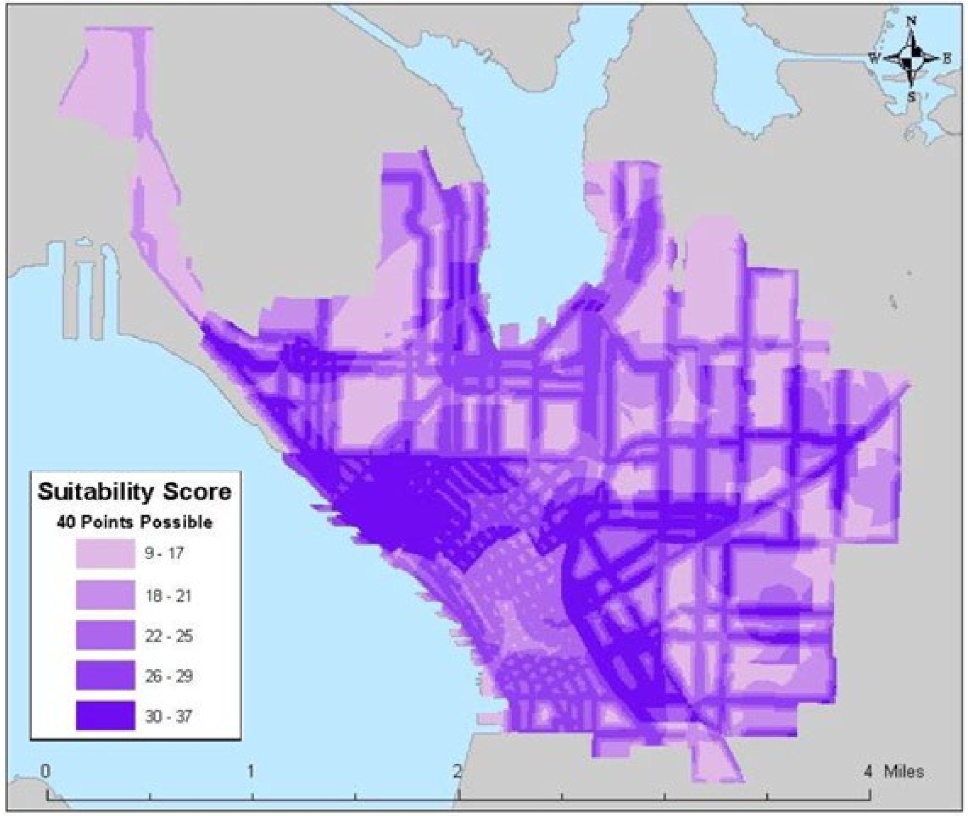
- Median Household Income
- Average Age
- Percent of Condominiums
- Percent of Professionals
- Annual Spending on Pets



Suitability based on *location*

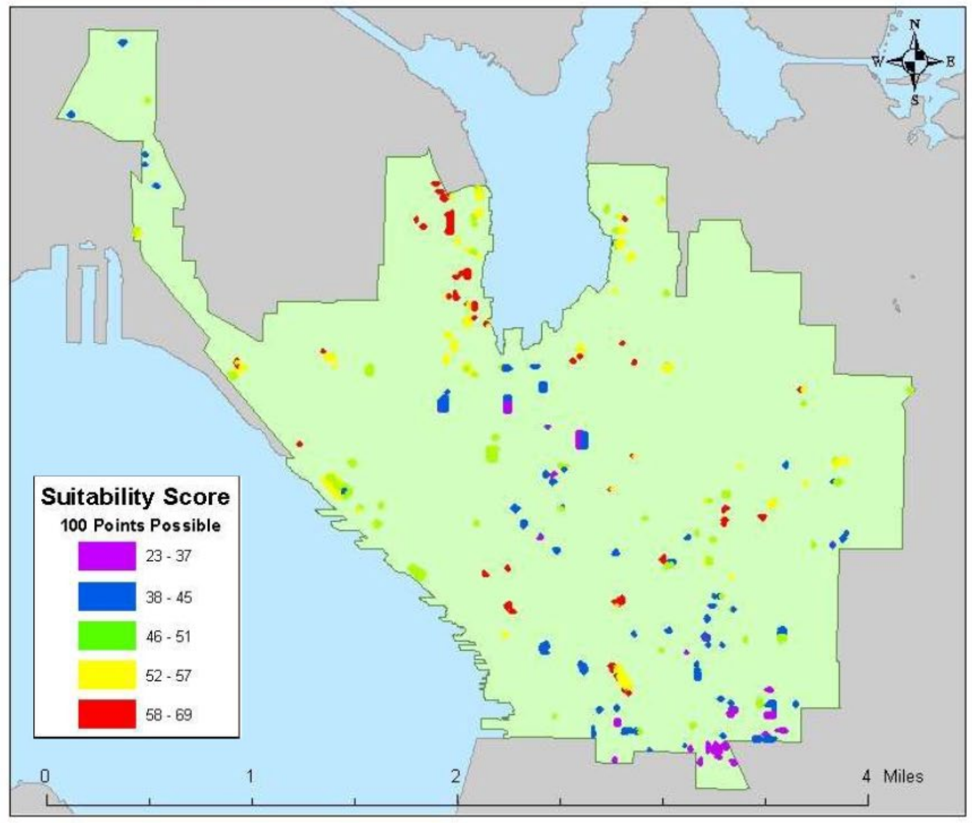
Combined Criteria

- Distance to competition
- Distance to Parks
- Distance to Arterials
- Proximity to Central Business District



Combined attribute and location

After combining Customer Suitability, Distance Suitability and parcel criteria, you end up with a map of potential properties that meet all of your requirements.



Key Points!

- Selection works with attributes
- Geoprocessing works with topology
- ‘Select by location’ uses a little bit of both