2019

Unit 3

Forrest J. Bowlick
University of Massachusetts Amherst, fbowlick@umass.edu

UMassAmherst

Data Type Exploration: Vector

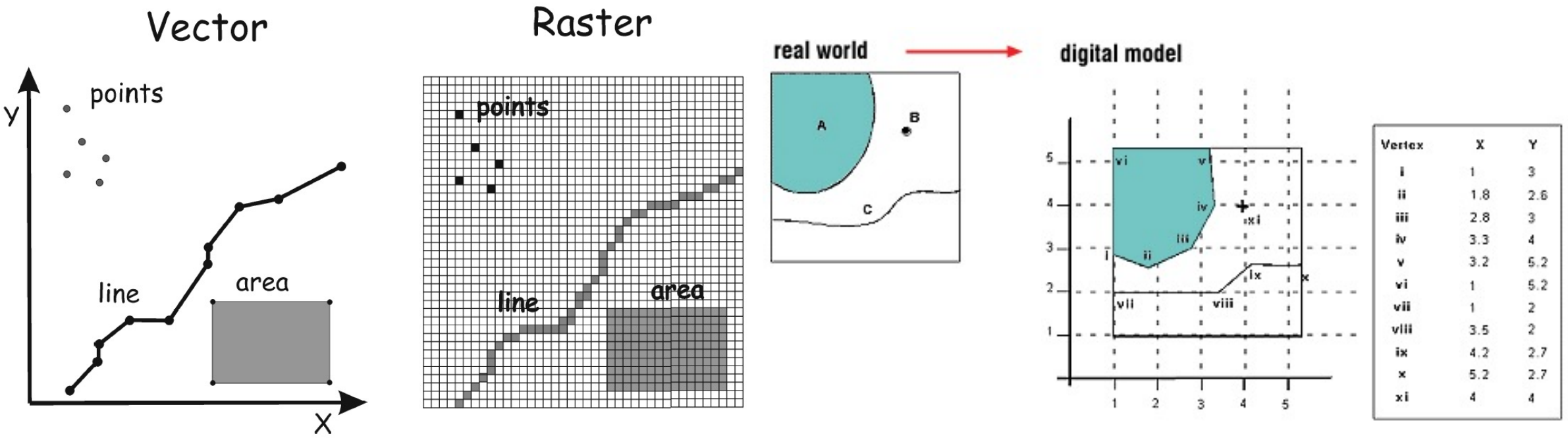


Figure 2-8: Vector and raster data models.

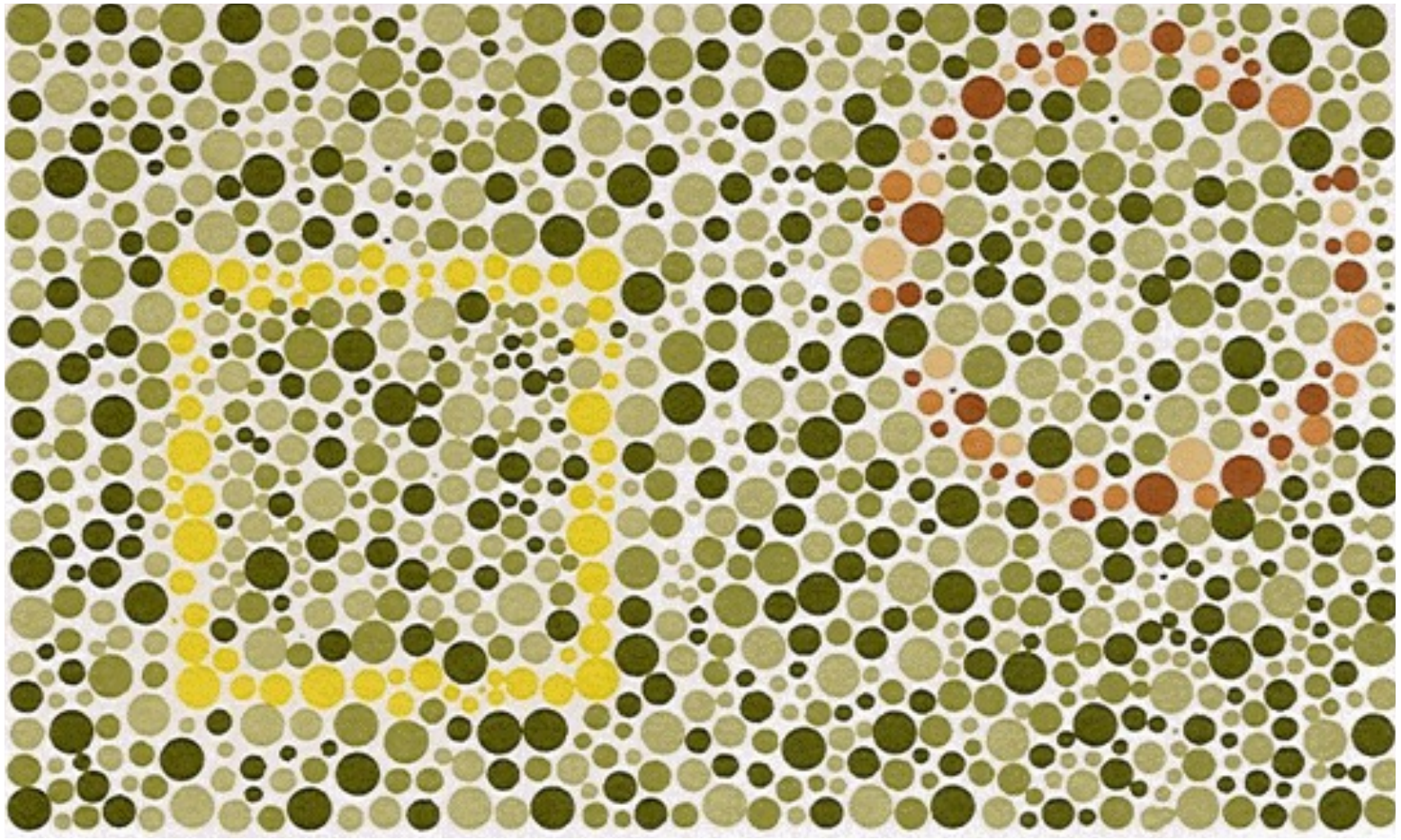
Forrest J. Bowlick, Intro GIS

UMass Amherst, Fall 2019

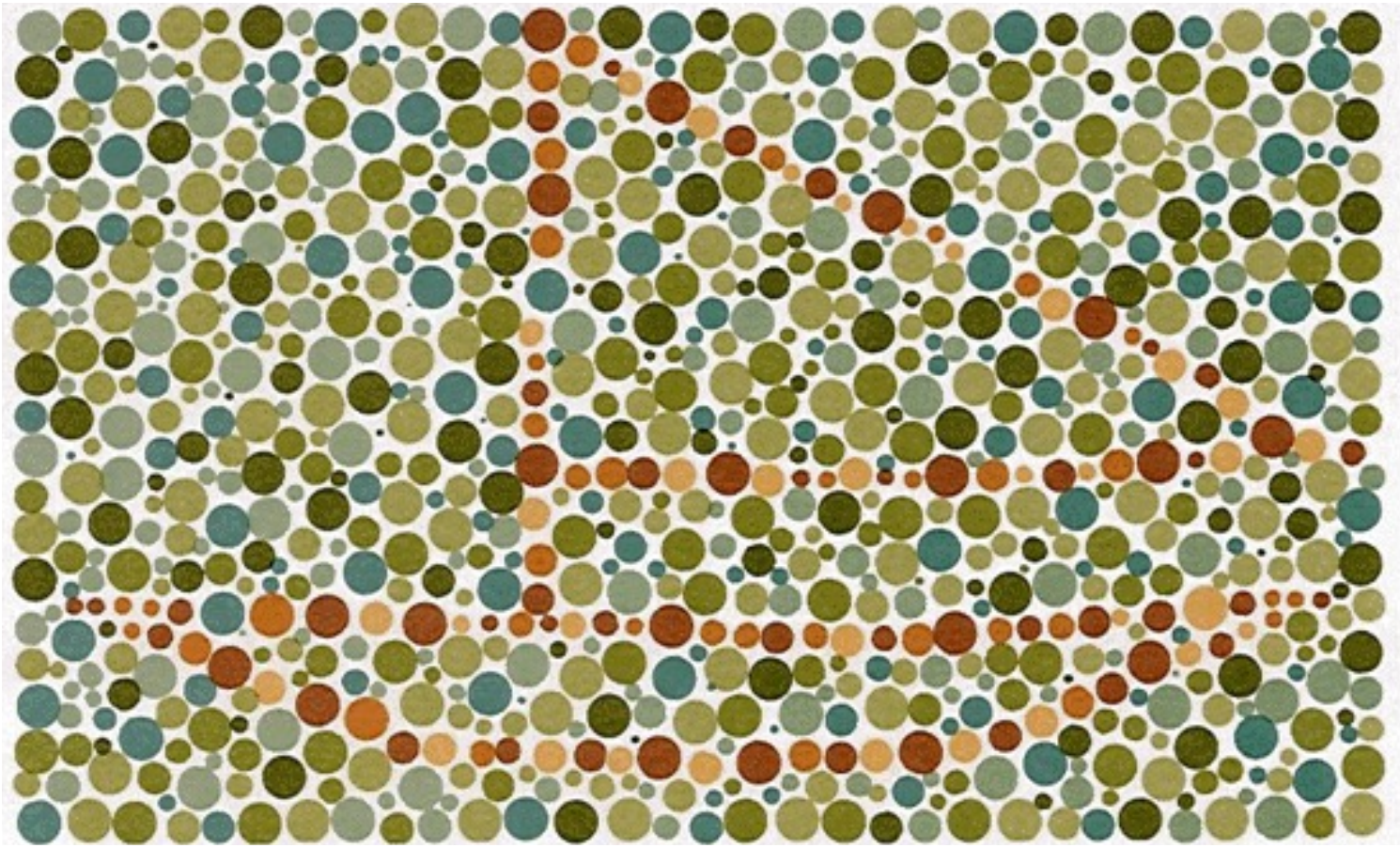
On the Agenda

- Colorblindness & Cartography
- Data Types (broadly)
- Databases
- Data Formats (broadly)
- Vector Data
- GIS Concept Practice

What shapes do you see?



What shapes do you see? (Part II)

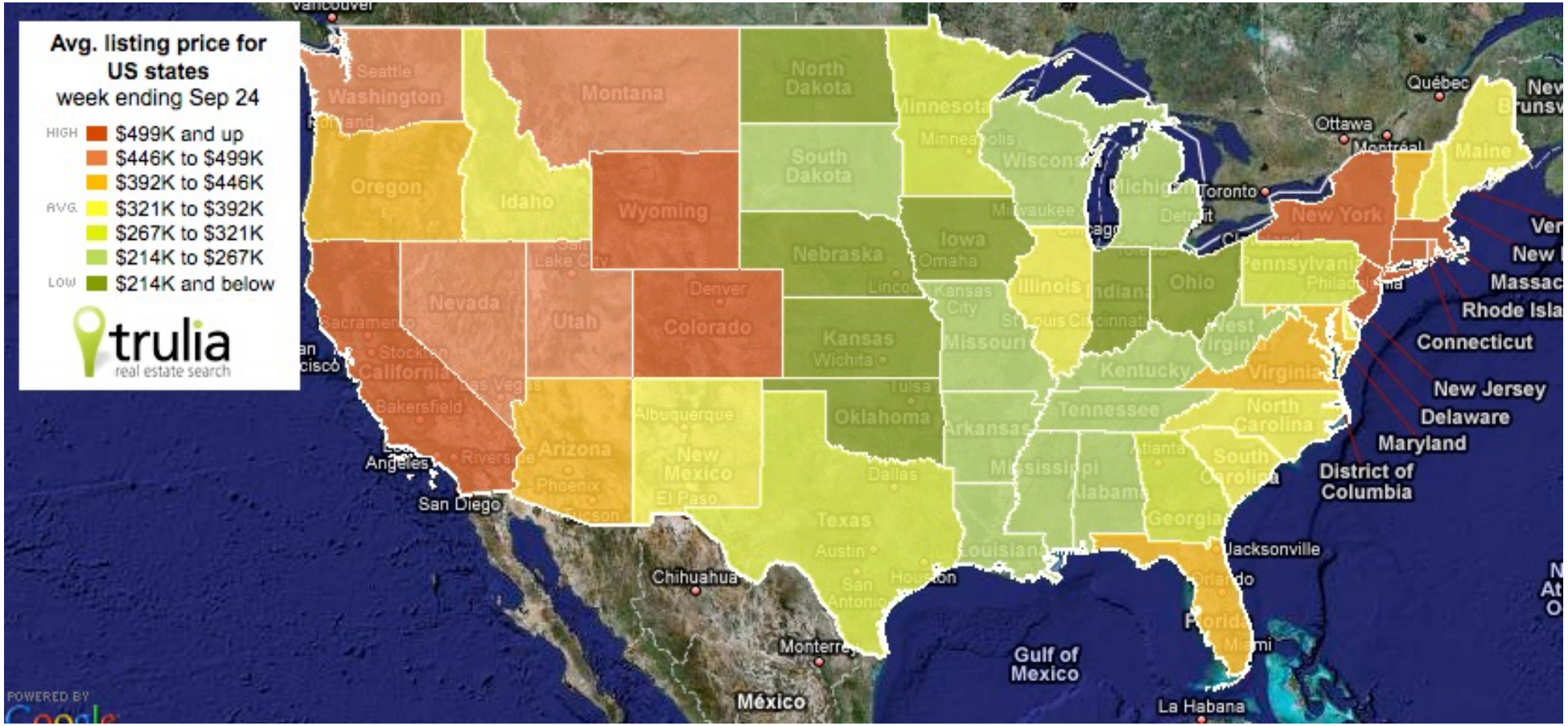


Red-green color blindness

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



Data ‘Types’

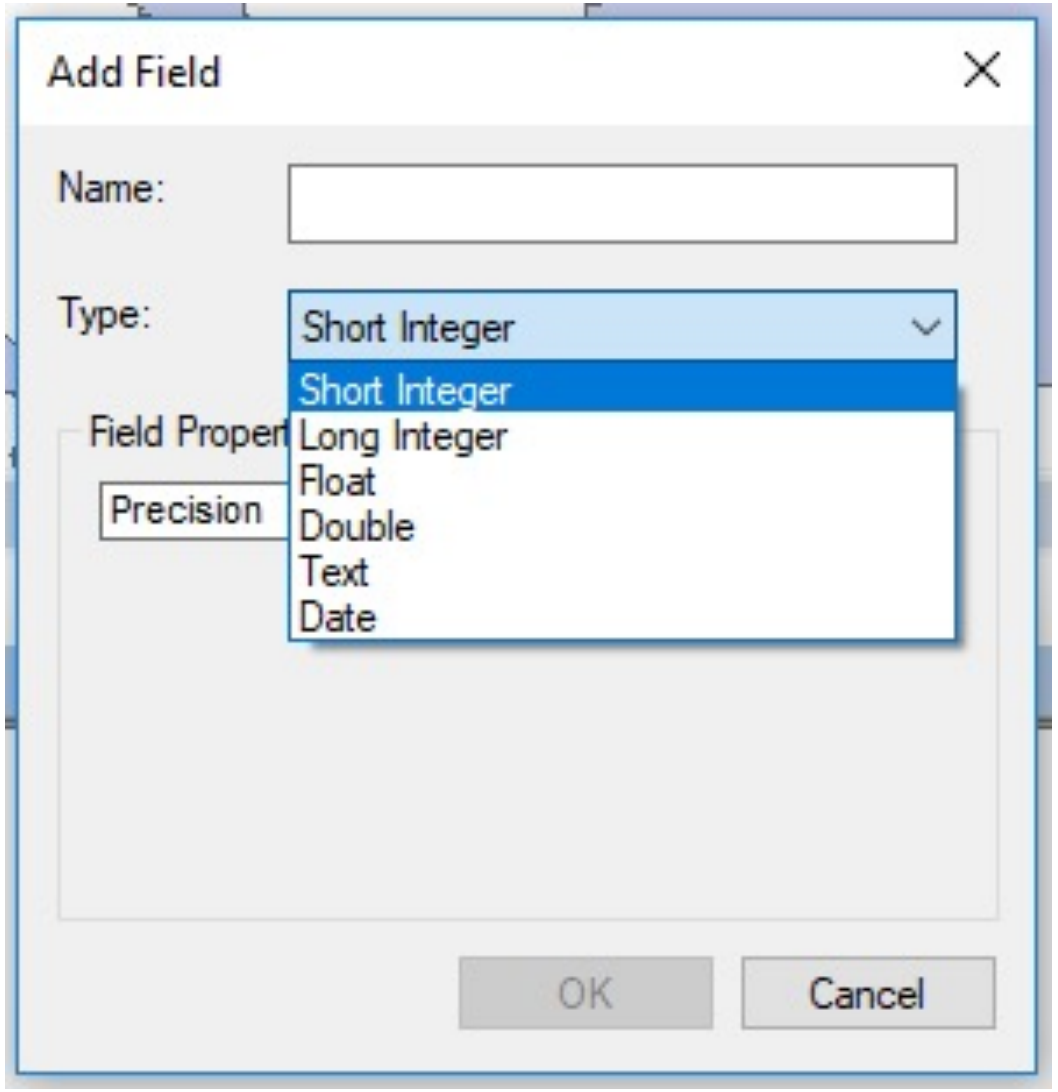
- When we reference data ‘type’, we could be referring to a few things:
  - The classification of the information stored in the file for use by a program
  - The nature of the data’s nature or distribution (better: data level or scale)
  - The file format for the data we want to use
- So data typology can be pretty complex.

Back to the Attribute Table

ArcGIS field data types

ArcMap 10.6 | Other versions

- Numbers
- Text
- Dates
- BLOBs
- Object identifiers
- Global identifiers
- Fields of type raster
- Geometry



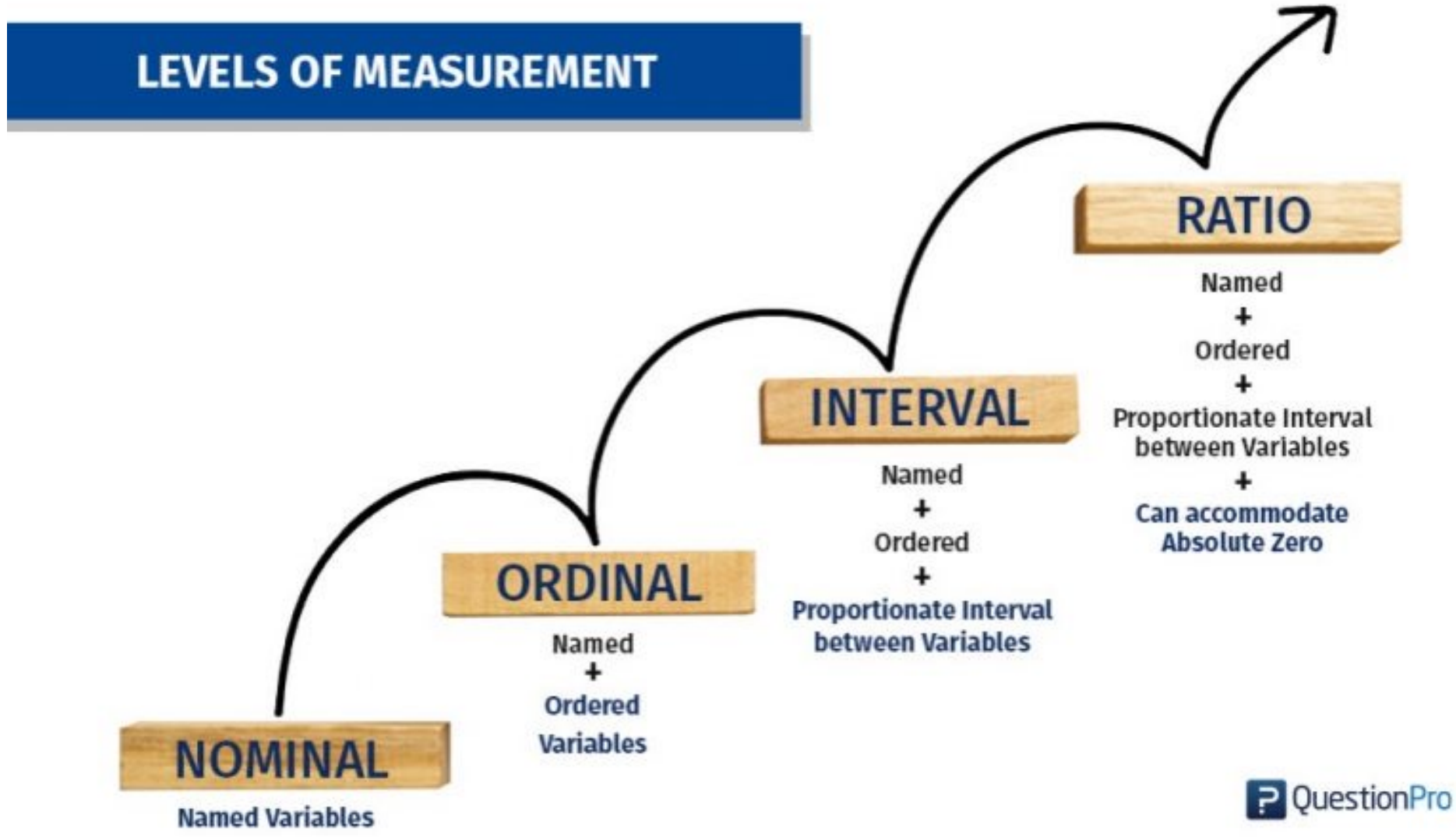
http://desktop.arcgis.com/en/arcmap/latest/manage-data/geodatabases/arcgis-field-data-types.htm

SQL is a Bit More… Complex

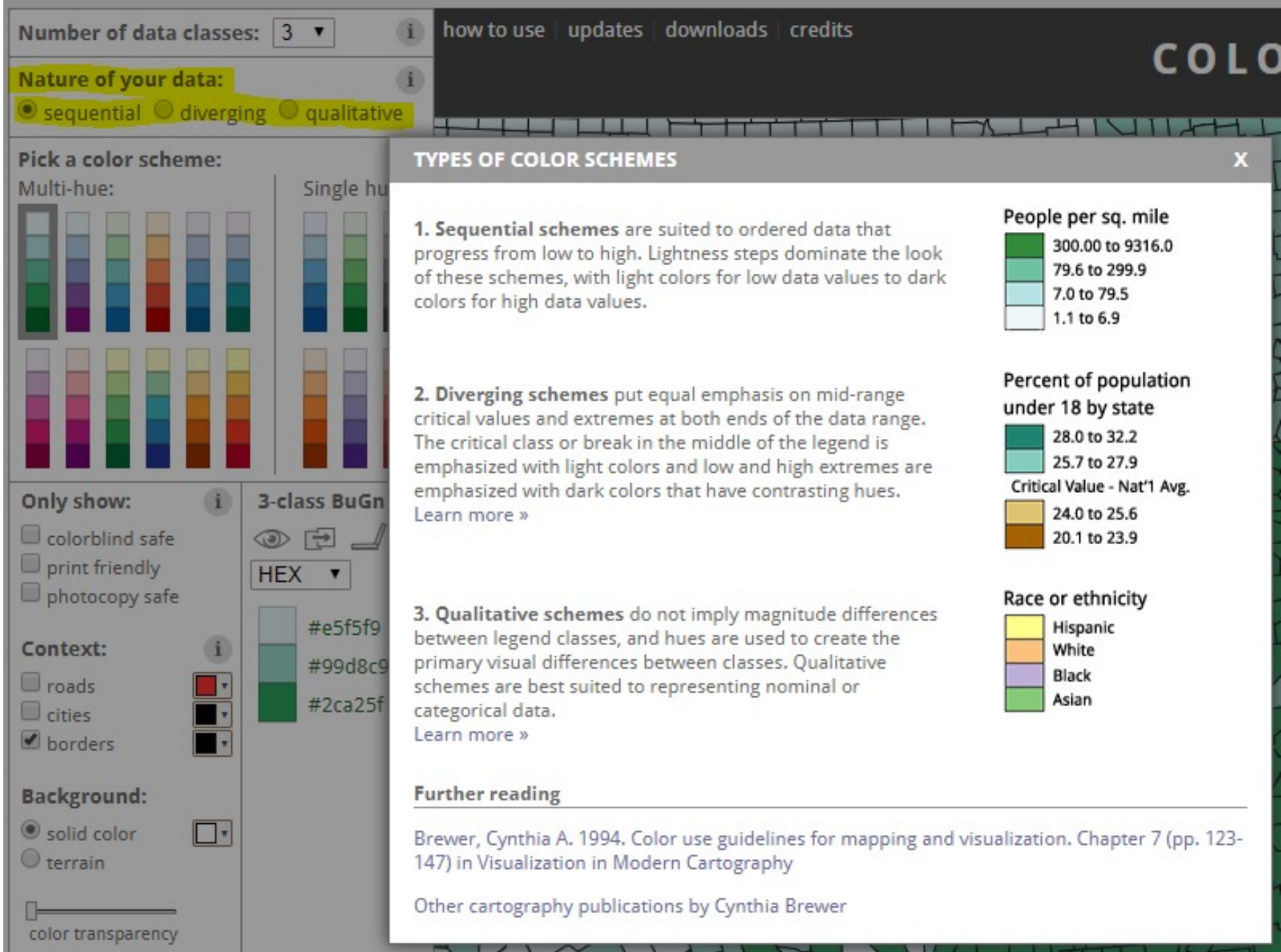
MySQL DATA TYPES

DATE TYPE	SPEC	DATA TYPE	SPEC
CHAR	String (0 - 255)	INT	Integer (-2147483648 to 214748-3647)
VARCHAR	String (0 - 255)	BIGINT	Integer (-9223372036854775808 to 9223372036854775807)
TINYTEXT	String (0 - 255)	FLOAT	Decimal (precise to 23 digits)
TEXT	String (0 - 65535)	DOUBLE	Decimal (24 to 53 digits)
BLOB	String (0 - 65535)	DECIMAL	"DOUBLE" stored as string
MEDIUMTEXT	String (0 - 16777215)	DATE	YYYY-MM-DD
MEDIUMBLOB	String (0 - 16777215)	DATETIME	YYYY-MM-DD HH:MM:SS
LONGTEXT	String (0 - 4294967295)	TIMESTAMP	YYYYMMDDHHMMSS
LONGBLOB	String (0 - 4294967295)	TIME	HH:MM:SS
TINYINT	Integer (-128 to 127)	ENUM	One of preset options
SMALLINT	Integer (-32768 to 32767)	SET	Selection of preset options
MEDIUMINT	Integer (-8388608 to 8388607)	BOOLEAN	TINYINT(1)

Data Levels to Consider



The Nature of Your Data



Let’s Try it Out!

Kahoot.it

https://play.kahoot.it/#/k/afe73d76-5e0a-4322-a22b-ac03a469c803

Database Manipulation

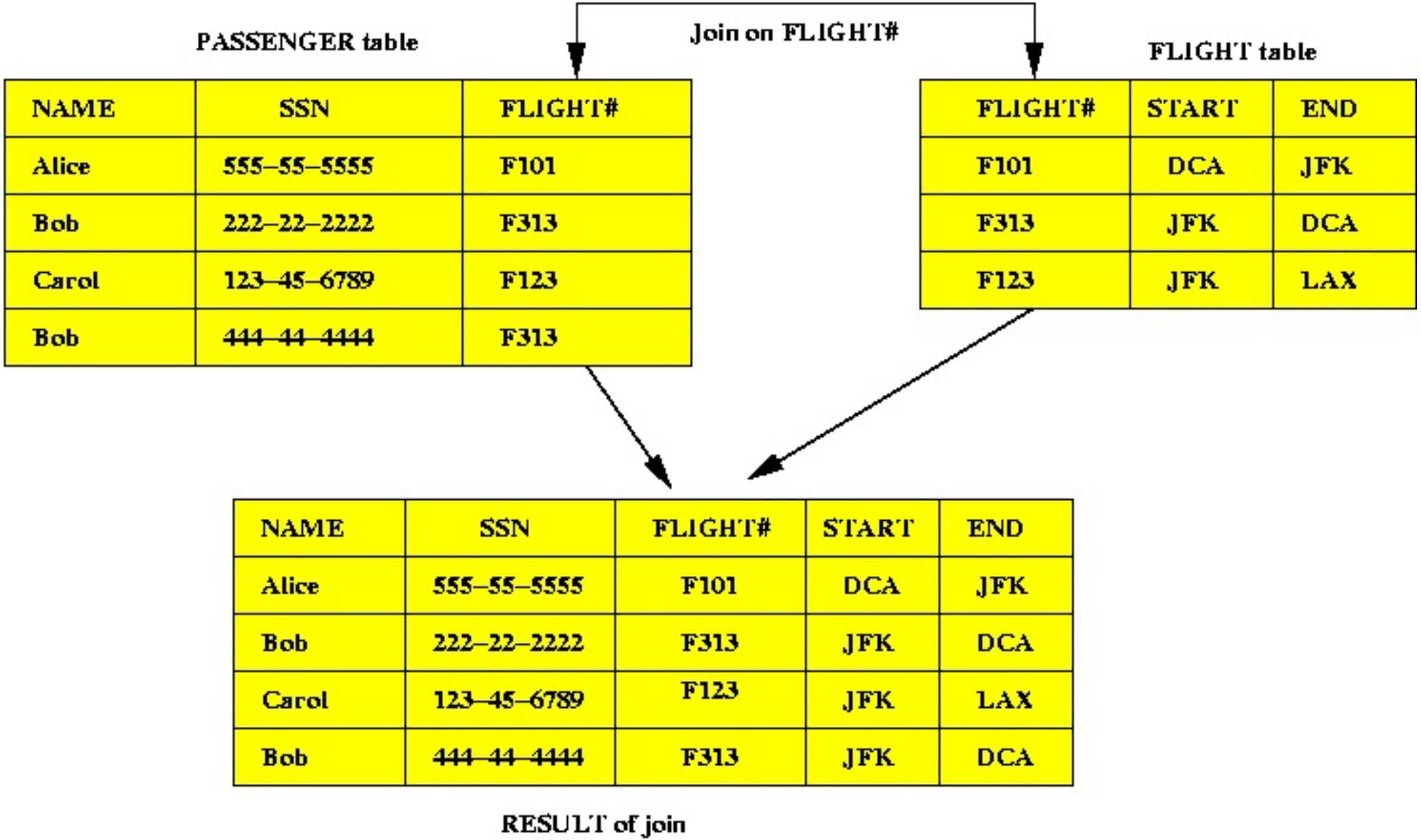
- We've worked on a small scale with databases recently, without going into much depth about the technical 'how' of data.

- Think of the 'join' operation.

‘Join’

- The join operation links two tables in a relational database by linking a key value common to each.
- ‘Relational’ means the databases can be linked.
- ‘Key’ means the values are shared and can be connected between the tables.
  - Note that the key value might not be the unique identifier!

customer_id	first_name	last_name	email	address	city	state	zip
1	George	Washington	gwashington@usa.gov	3200 Mt Vernon Hwy	Mount Vernon	VA	22121
2	John	Adams	jadams@usa.gov	1250 Hancock St	Quincy	MA	02169
3	Thomas	Jefferson	tjefferson@usa.gov	931 Thomas Jefferson Pkwy	Charlottesville	VA	22902
4	James	Madison	jmadison@usa.gov	11350 Constitution Hwy	Orange	VA	22960
5	James	Monroe	jmonroe@usa.gov	2050 James Monroe Parkway	Charlottesville	VA	22902


PASSENGER table
Join on FLIGHT#
FLIGHT table
NAME
SSN
FLIGHT#
Alice
555–55–5555
F101
Bob
222–22–2222
F313
Carol
123–45–6789
F123
Bob
444–44–4444
F313
FLIGHT#
START
END
F101
DCA
JFK
F313
JFK
DCA
F123
JFK
LAX
NAME
SSN
FLIGHT#
START
END
Alice
555–55–5555
F101
DCA
JFK
Bob
222–22–2222
F313
JFK
DCA
Carol
123–45–6789
F123
JFK
LAX
Bob
444–44–4444
F313
JFK
DCA
RESULT of join

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

Which columns can be used to relate these tables?

Which value is the ‘key’ value?

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

Which columns can be used to relate these two tables?

Which value is the ‘key’ value?

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

The JOIN operation

game

id	mdate	stadium	team1	team2
1001	8 June 2012	National Stadium, Warsaw	POL	GRE
1002	8 June 2012	Stadion Miejski (Wroclaw)	RUS	CZE
1003	12 June 2012	Stadion Miejski (Wroclaw)	GRE	CZE
1004	12 June 2012	National Stadium, Warsaw	POL	RUS
...				

goal

matchid	teamid	player	gtime
1001	POL	Robert Lewandowski	17
1001	GRE	Dimitris Salpingidis	51
1002	RUS	Alan Dzagoev	15
1002	RUS	Roman Pavlyuchenko	82
...			

eteam

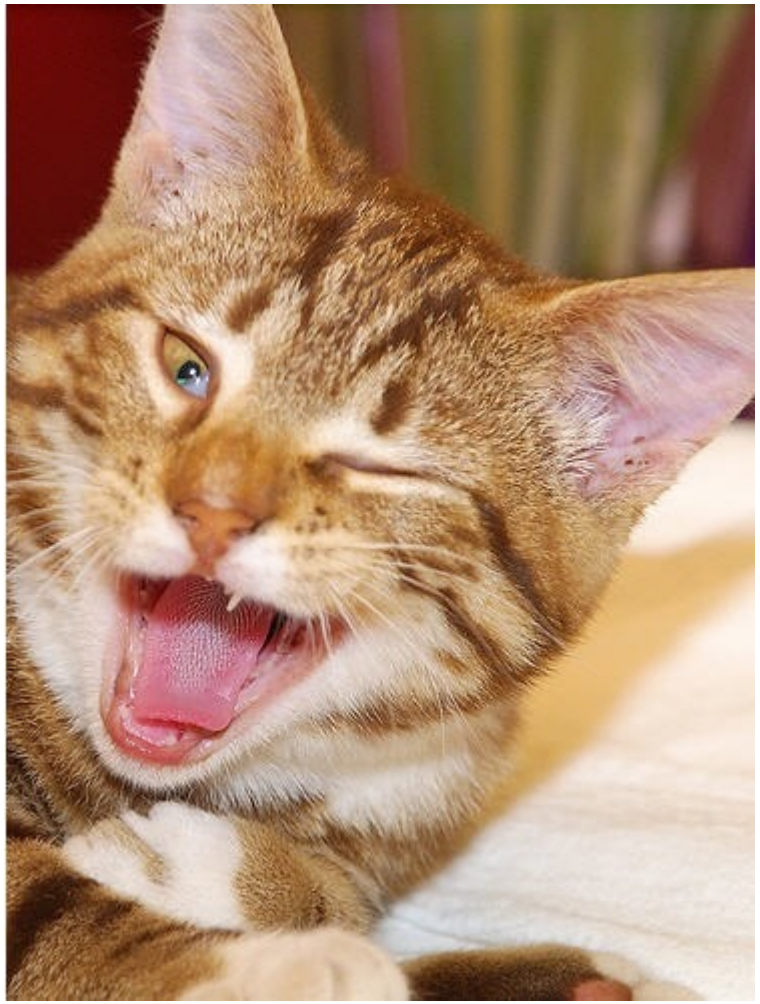
id	teamname	coach
POL	Poland	Franciszek Smuda
RUS	Russia	Dick Advocaat
CZE	Czech Republic	Michal Bilek
GRE	Greece	Fernando Santos
...		

Which columns can be used to relate these three tables?

Which value is the 'key' value?

http://sqlzoo.net/wiki/SQL_Tutorial

p.s. we’re doing all of this spatially too



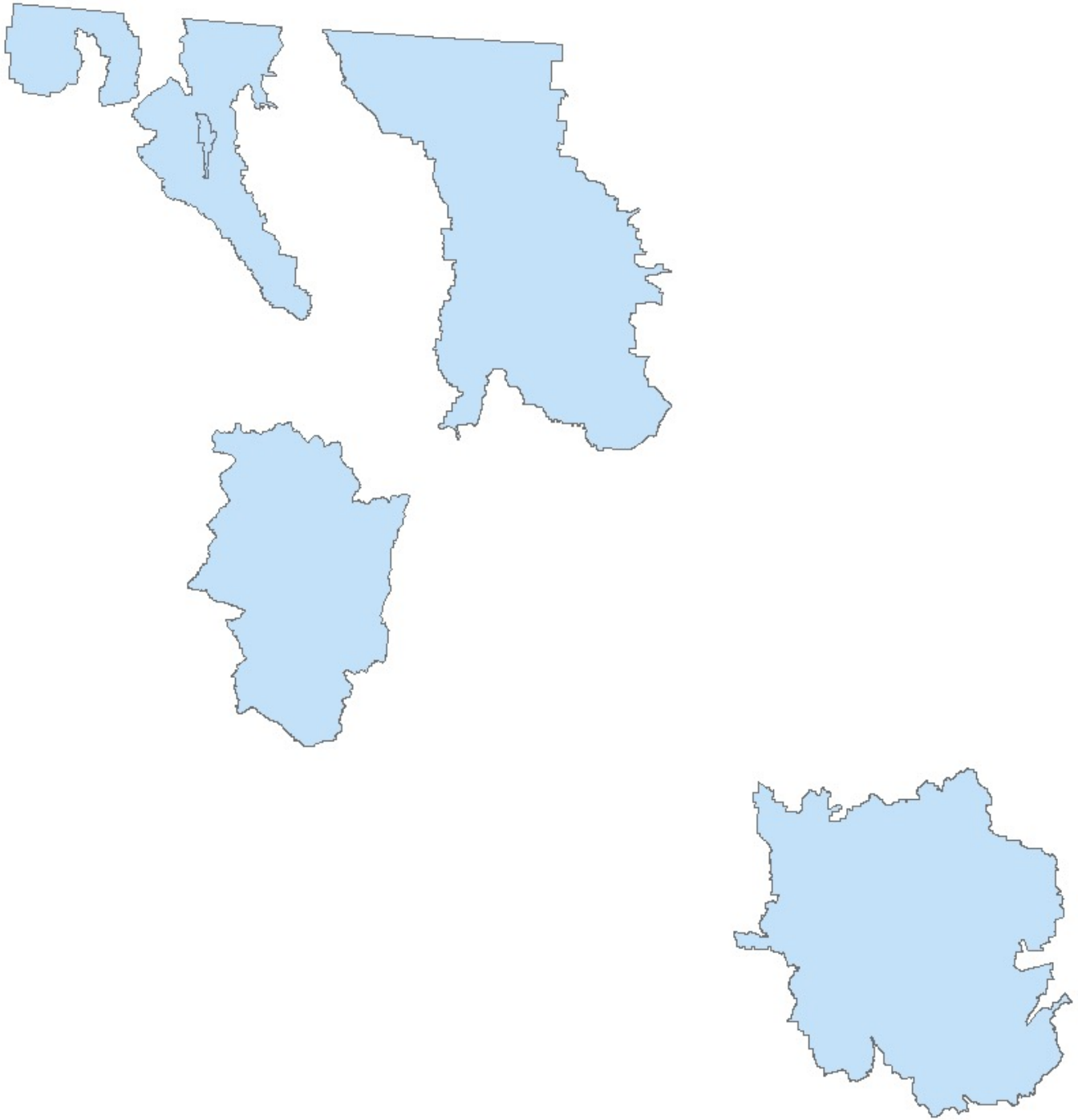
Data data data

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

- Important differences.

Raster vs. Vector vs. TIN

- Raster: Made up of cells (pixels).
- Vector: Made up of points, lines, and polygons.
- TIN: Triangular Irregular Network.



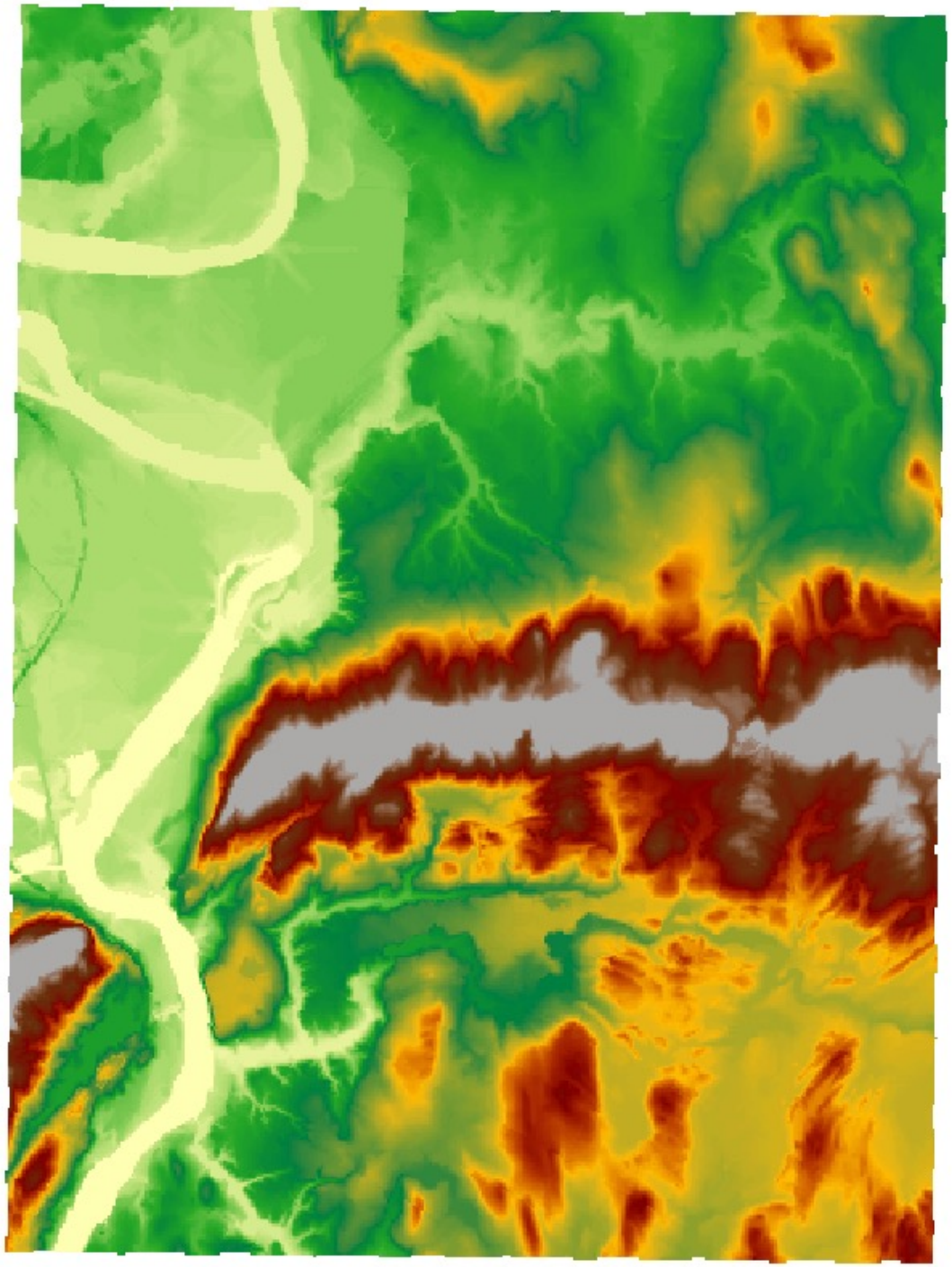


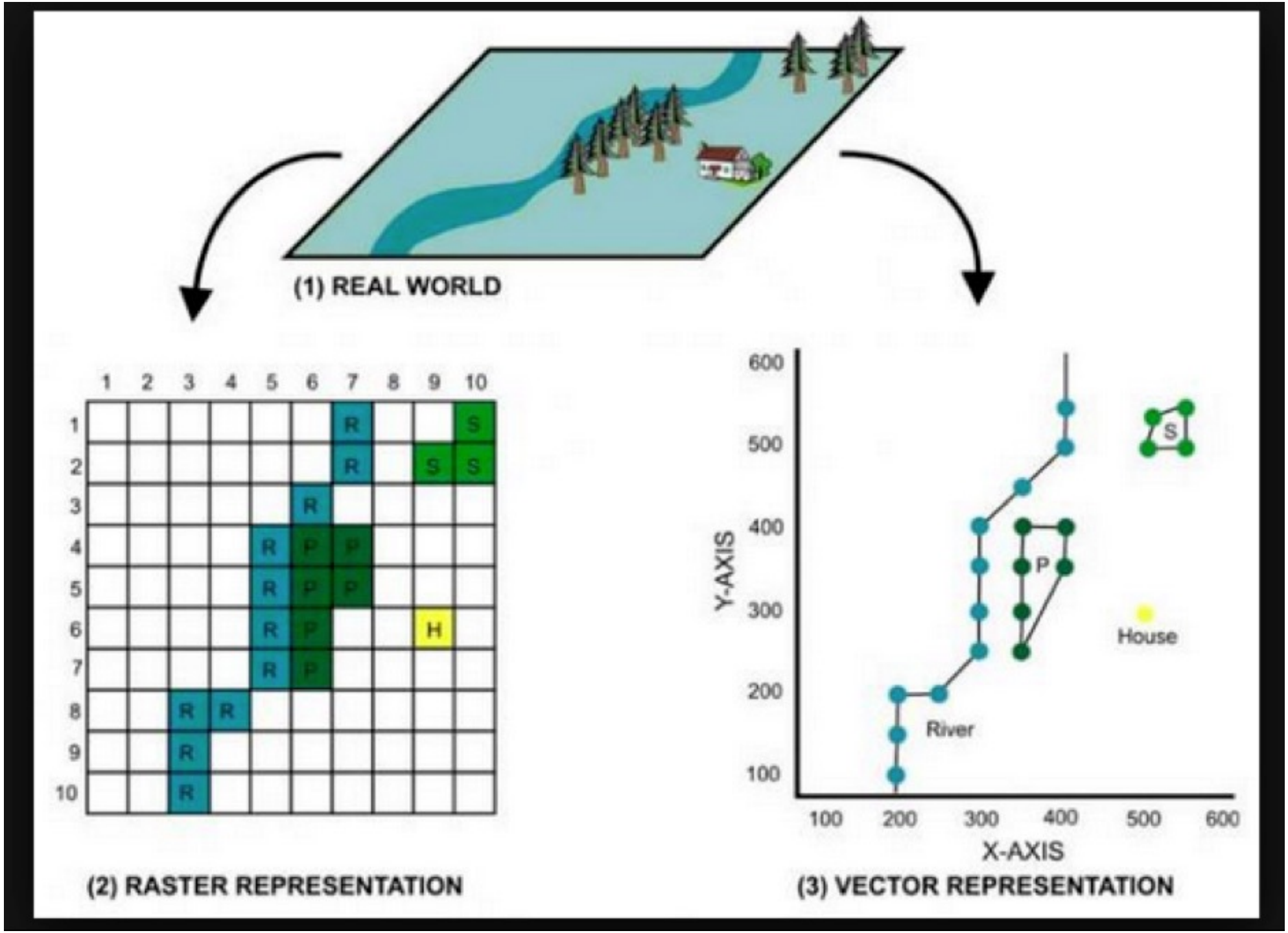
Table 2-2: A comparison of raster and vector data models.

Characteris-tic	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 net-work analyses, many other spa-tial operations more complex
positional pre-cision	floor set by cell size	limited only by positional mea-surements
accessibility	easy to modify or pro-gram, due to simple data structure	often complex
display and output	good for images, but discrete features may show "stairstep" edges	map-like, with con-tinuous curves, poor for images

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

Raster or Vector?

Tx.ag/GIS3


(1) REAL WORLD
1 2 3 4 5 6 7 8 9 10
1 R S
2 R S S
3 R
4 R P P
5 R P P
6 R P H
7 R P
8 R R
9 R
10 R
(2) RASTER REPRESENTATION
600
500
400
300
200
100
Y-AXIS
S
P
House
River
100 200 300 400 500 600
X-AXIS
(3) VECTOR REPRESENTATION

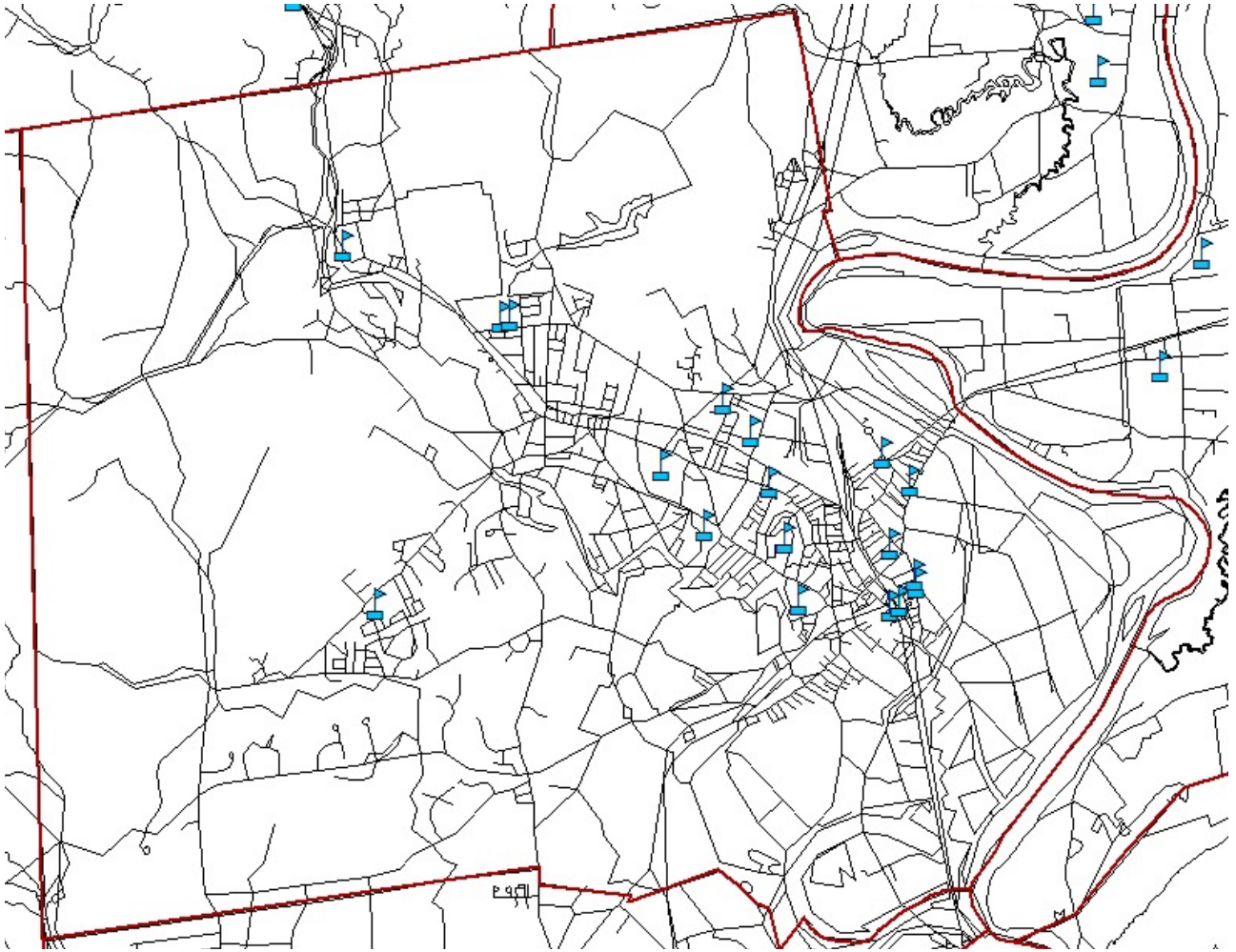
Vector data represents features as points, lines, and polygons and is best applied to discrete objects with defined shapes and boundaries.

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.

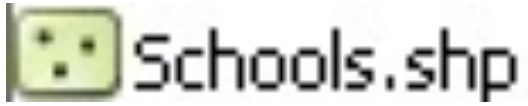
Vector (Feature) Data I

- Point
- Line
- Polygon



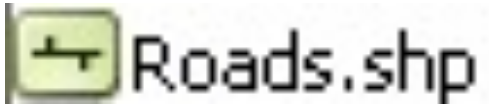
Vector (Feature) Data II

- Point
  - A specific geographic location
  - Each point may have a specific value depending on the attribute table
- Line
- Polygon



Vector (Feature) Data III

- Point
- Line
  - A segment or arc with a specific geographic location
  - Each line may have a specific value depending on the attribute table
  - Lines have direction and length
- Polygon



Vector (Feature) Data IV

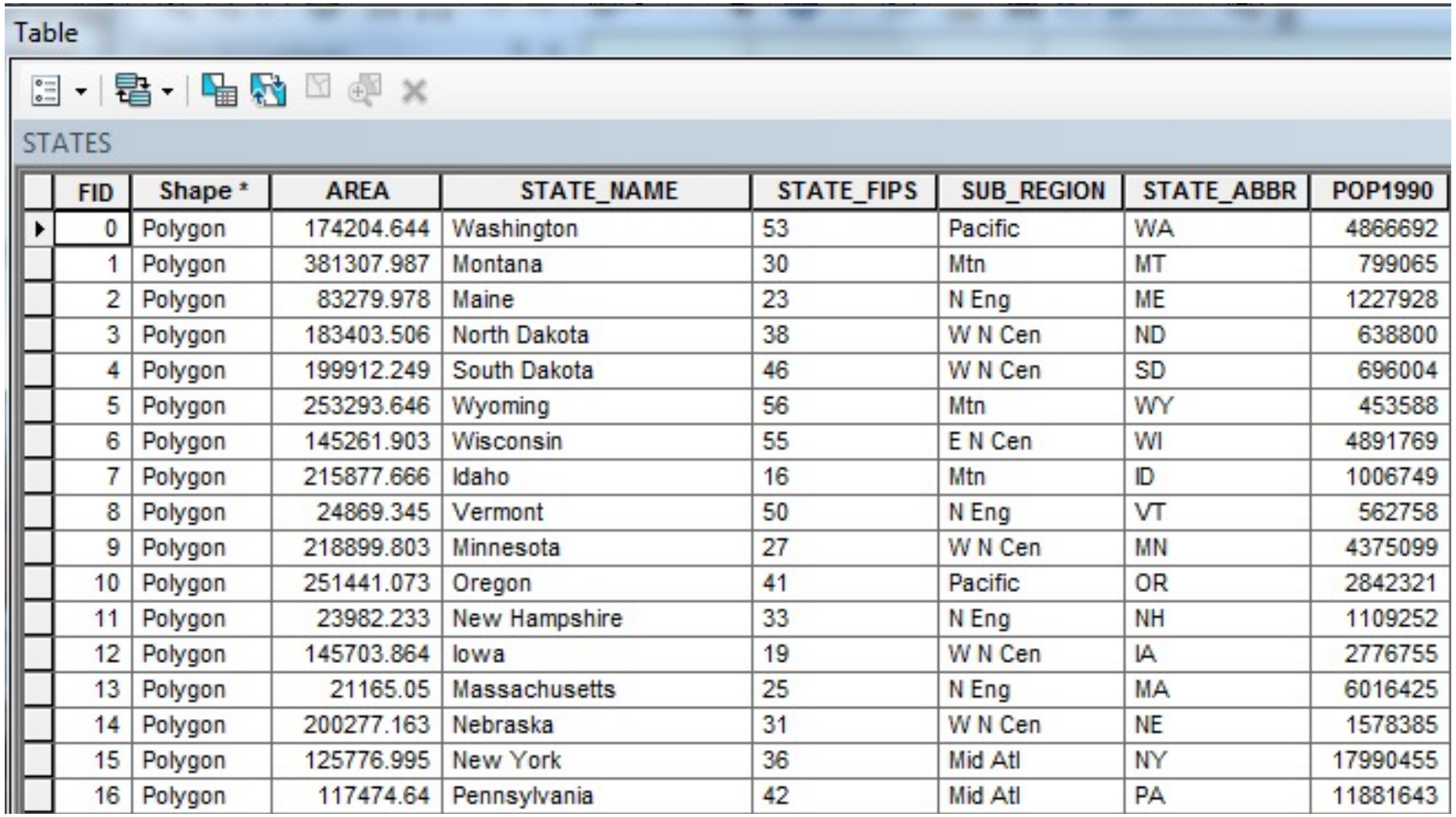
- Point
- Line
- Polygon
  - An enclosed areas with a specific geographic location
  - Has a geographically defined shape
  - Edges are straight lines with vertices

How Do You Choose?

- Kahoot (kahoot.it)

- https://play.kahoot.it/#/k/e9e2d380-4b8c-4485-afb7-272ba7a7e48f

All Vector Files Have Attributes

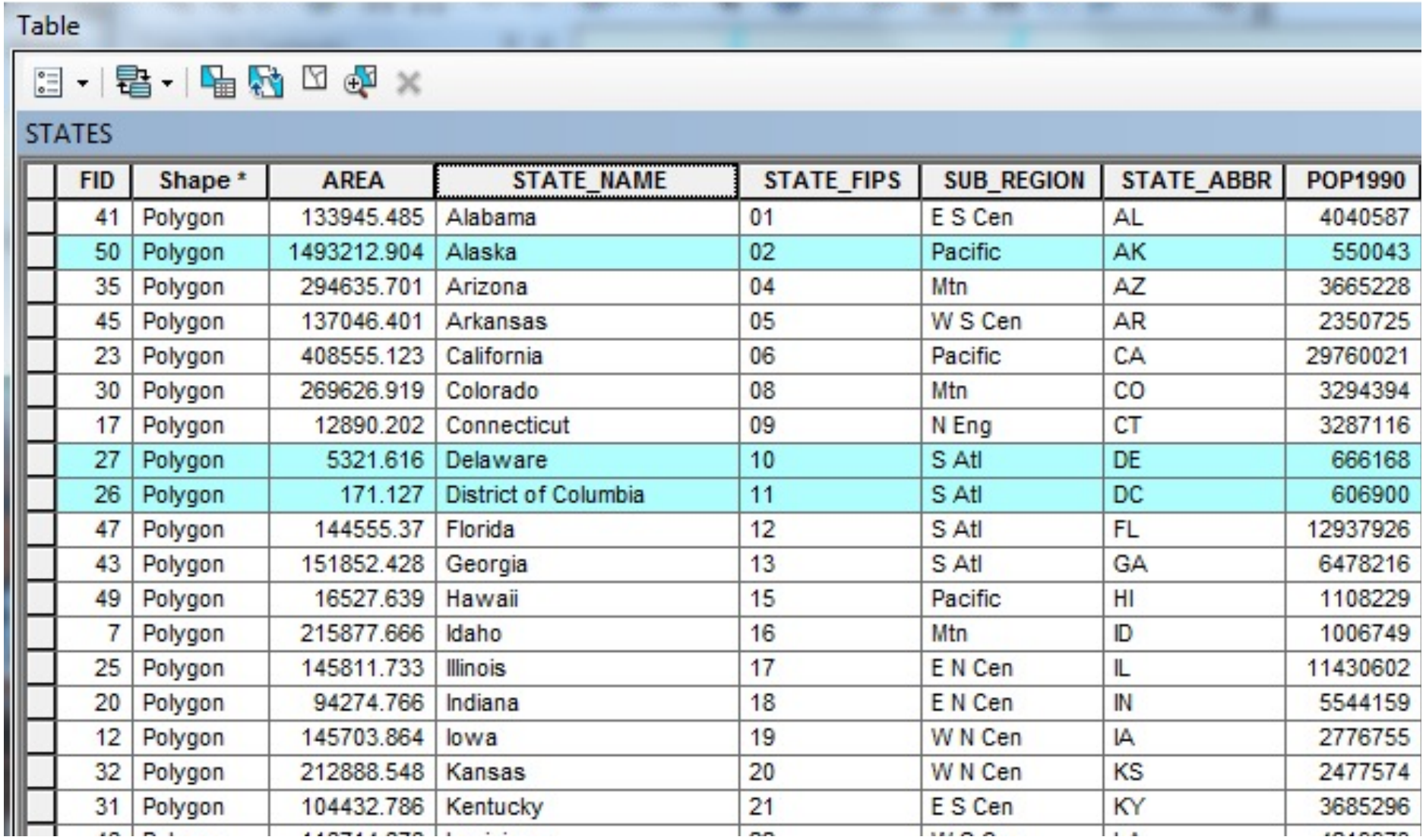


Table

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

We know that we can select features based on attributes

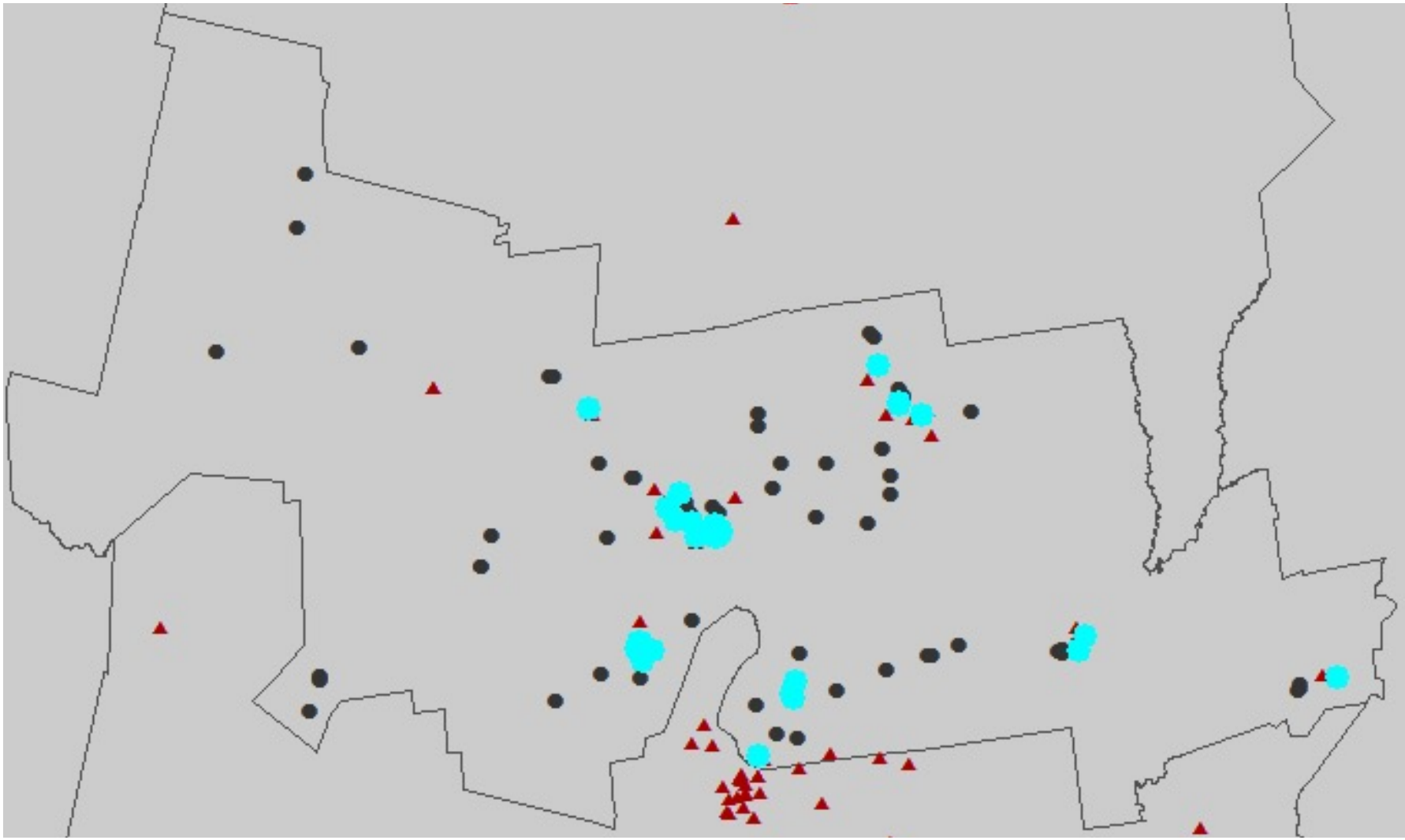


Table

STATES

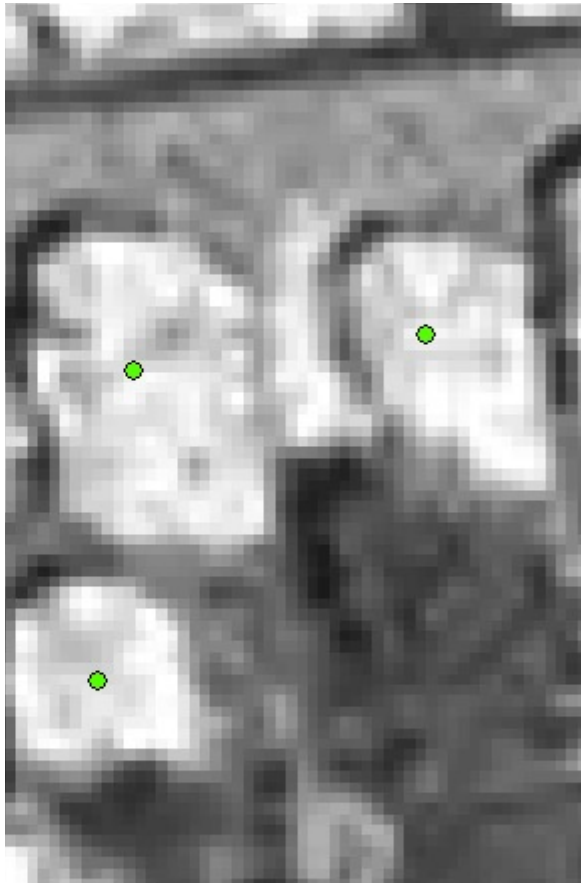
FID	Shape *	AREA	STATE_NAME	STATE_FIPS	SUB_REGION	STATE_ABBR	POP1990
41	Polygon	133945.485	Alabama	01	E S Cen	AL	4040587
50	Polygon	1493212.904	Alaska	02	Pacific	AK	550043
35	Polygon	294635.701	Arizona	04	Mtn	AZ	3665228
45	Polygon	137046.401	Arkansas	05	W S Cen	AR	2350725
23	Polygon	408555.123	California	06	Pacific	CA	29760021
30	Polygon	269626.919	Colorado	08	Mtn	CO	3294394
17	Polygon	12890.202	Connecticut	09	N Eng	CT	3287116
27	Polygon	5321.616	Delaware	10	S Atl	DE	666168
26	Polygon	171.127	District of Columbia	11	S Atl	DC	606900
47	Polygon	144555.37	Florida	12	S Atl	FL	12937926
43	Polygon	151852.428	Georgia	13	S Atl	GA	6478216
49	Polygon	16527.639	Hawaii	15	Pacific	HI	1108229
7	Polygon	215877.666	Idaho	16	Mtn	ID	1006749
25	Polygon	145811.733	Illinois	17	E N Cen	IL	11430602
20	Polygon	94274.766	Indiana	18	E N Cen	IN	5544159
12	Polygon	145703.864	Iowa	19	W N Cen	IA	2776755
32	Polygon	212888.548	Kansas	20	W N Cen	KS	2477574
31	Polygon	104432.786	Kentucky	21	E S Cen	KY	3685296

You can also select by location



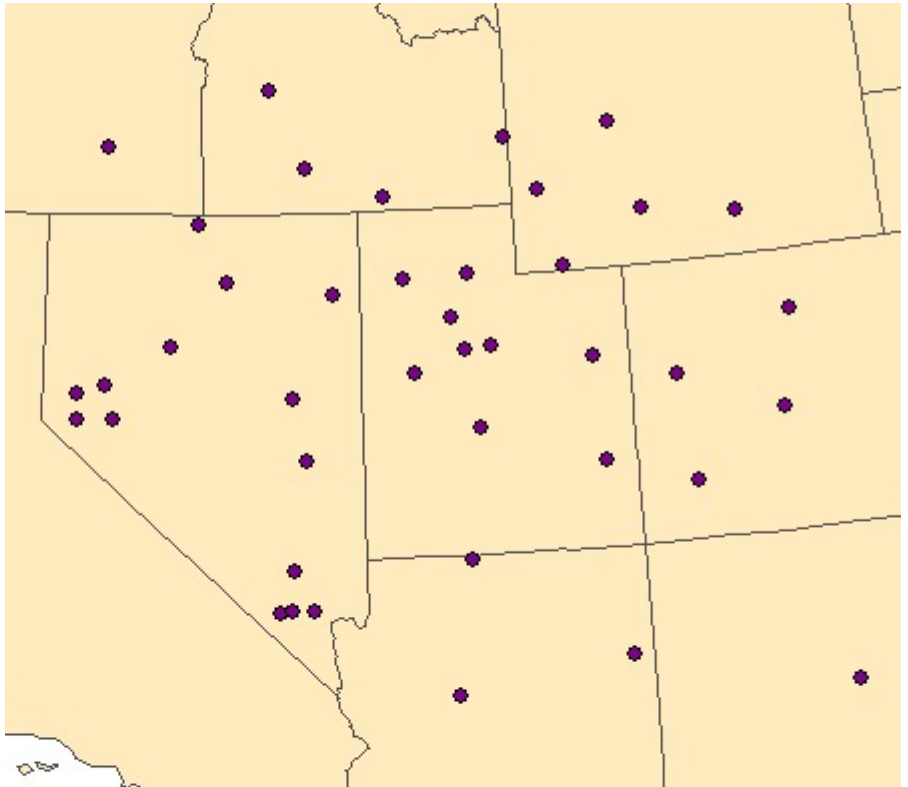
Advantages of Vector Data

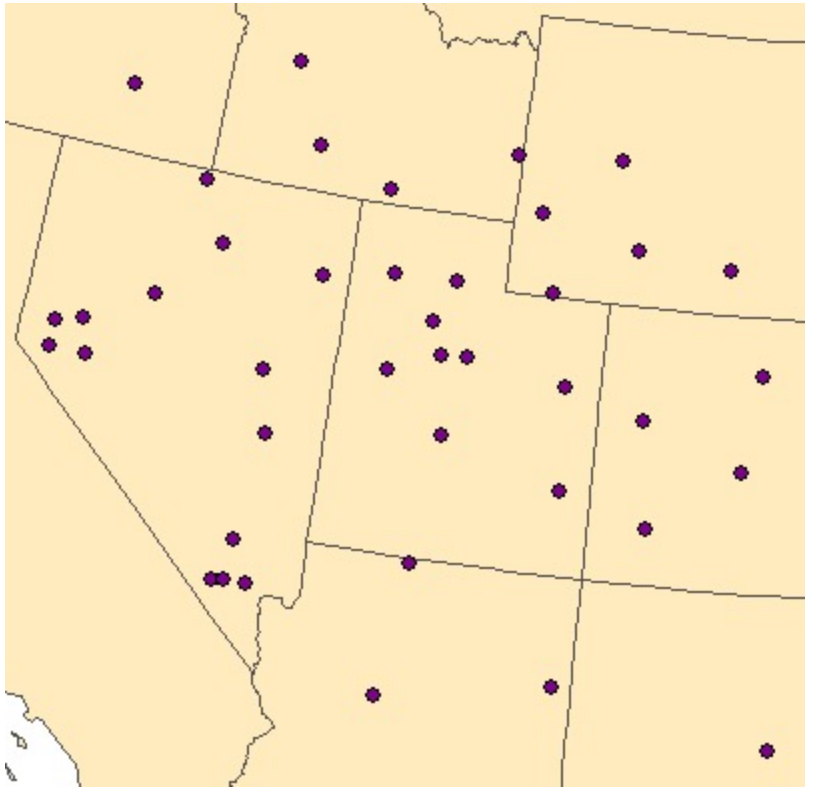
- Looks the same at any scale



Advantages of Vector Data I

- Looks the same at any scale
- Easy to re-project (change coordinate system)



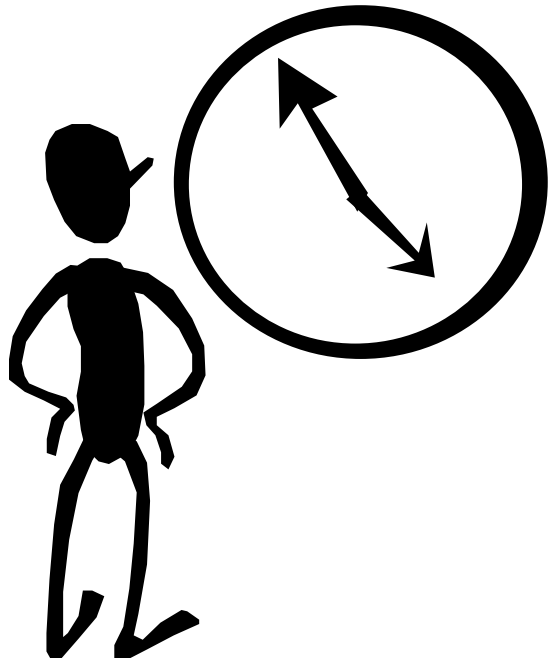


Advantages of Vector Data II

- Looks the same at any scale
- Easy to reproject (change coordinate system)
- Looks good on a map (easy to read/interpret output with nice smooth lines or points)

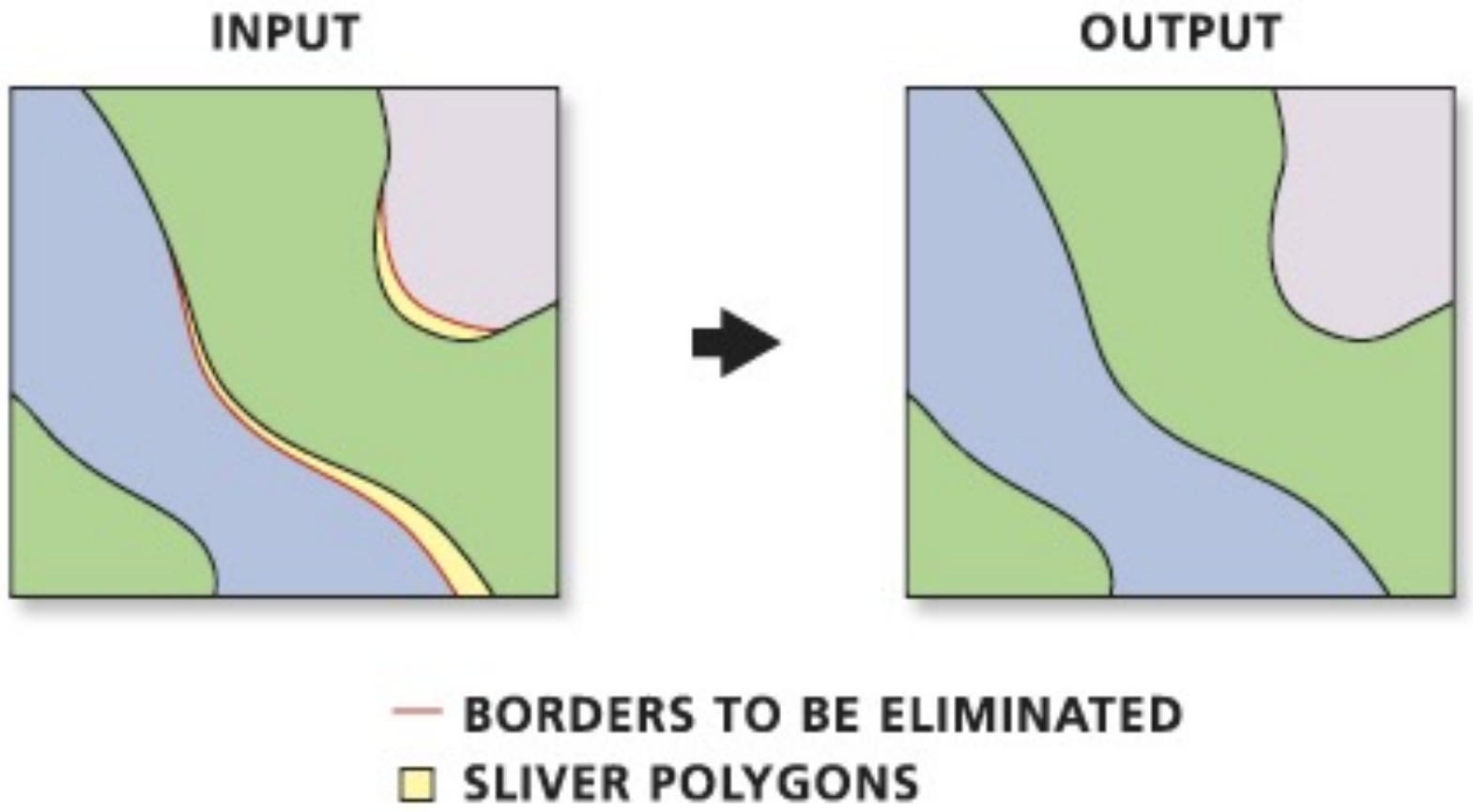
Disadvantages of Vector Data I

- It can be a memory hog! (Detailed shapefiles will take a long time to load and manipulate)

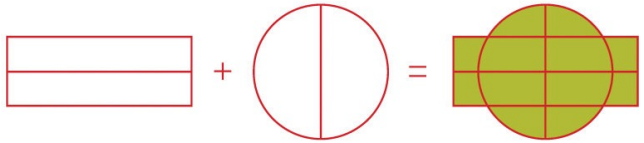


Disadvantages of Vector Data II

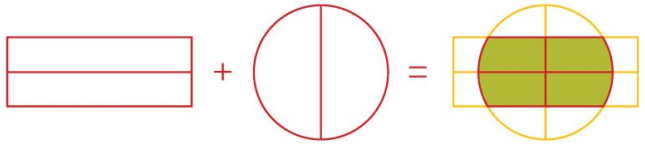
- It can be a memory hog! (Detailed shapefiles will take a long time to load and manipulate)
- Data may not be continuous spatially


INPUT
OUTPUT
BORDERS TO BE ELIMINATED
SLIVER POLYGONS

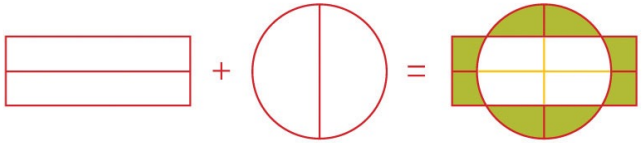
(a) Union



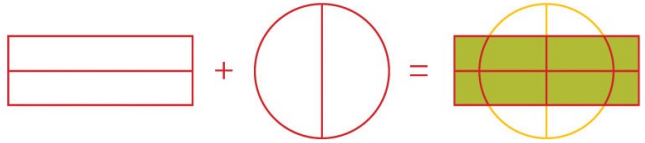
(b) Intersect



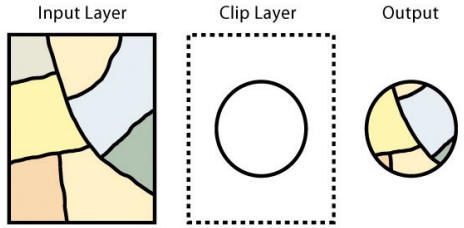
(c) Symmetrical Difference



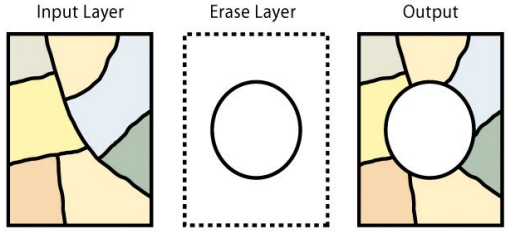
(d) Identity



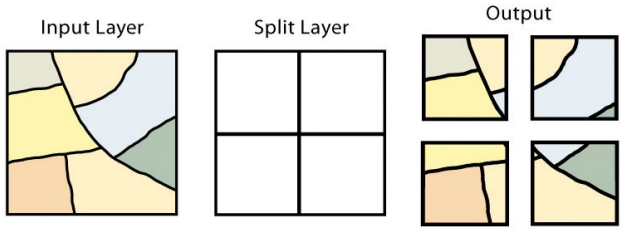
(e) Clip



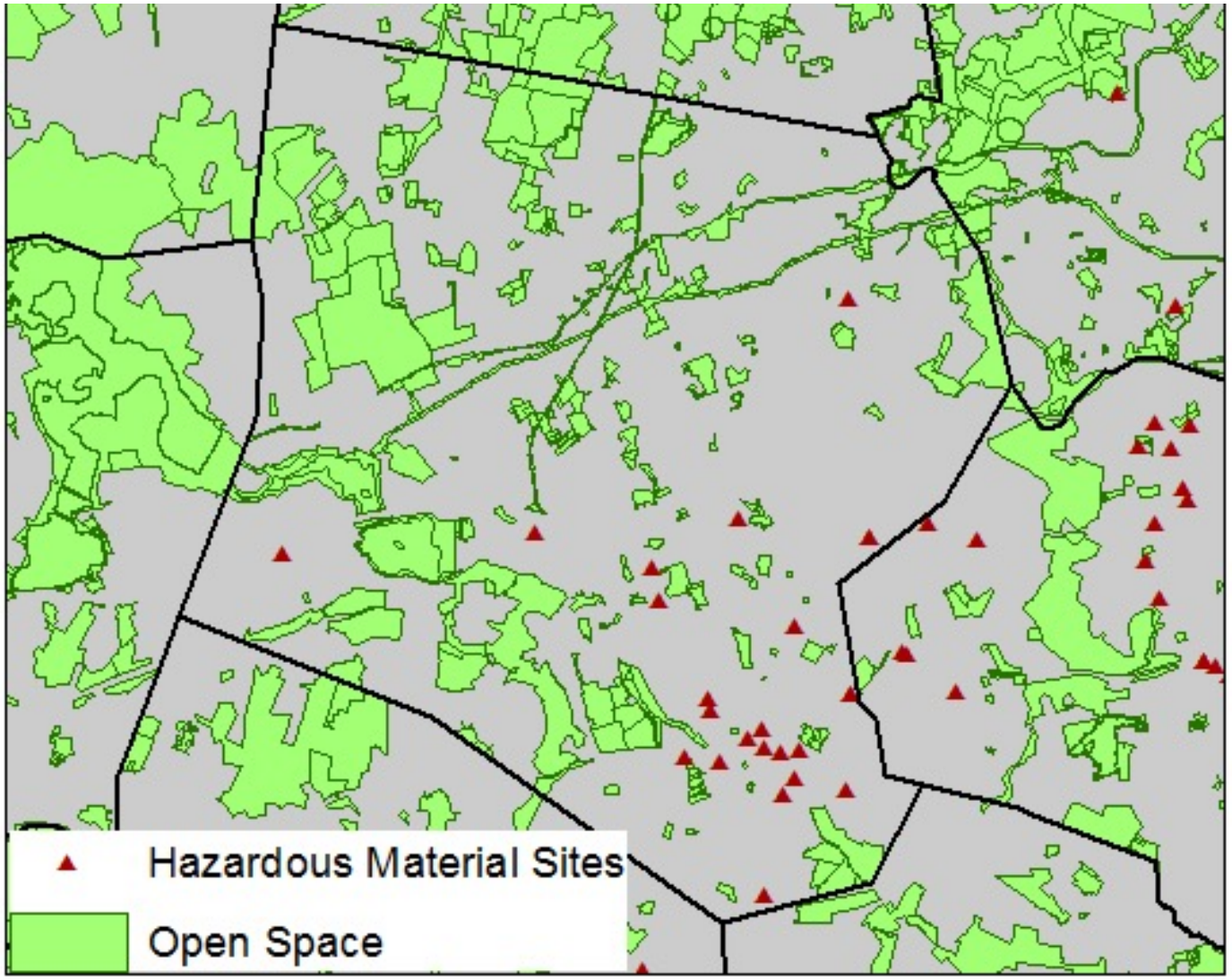
(f) Erase



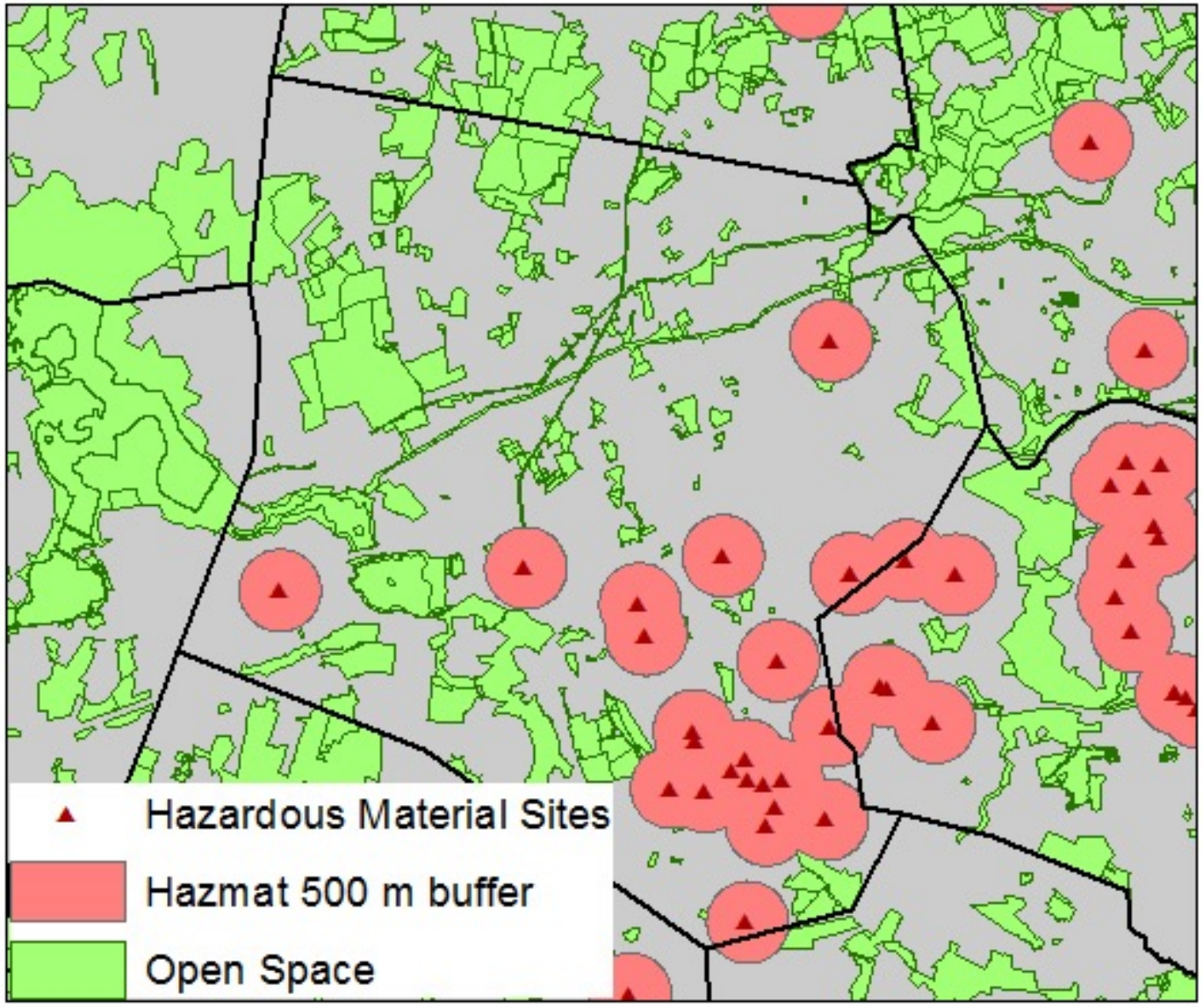
(g) Split



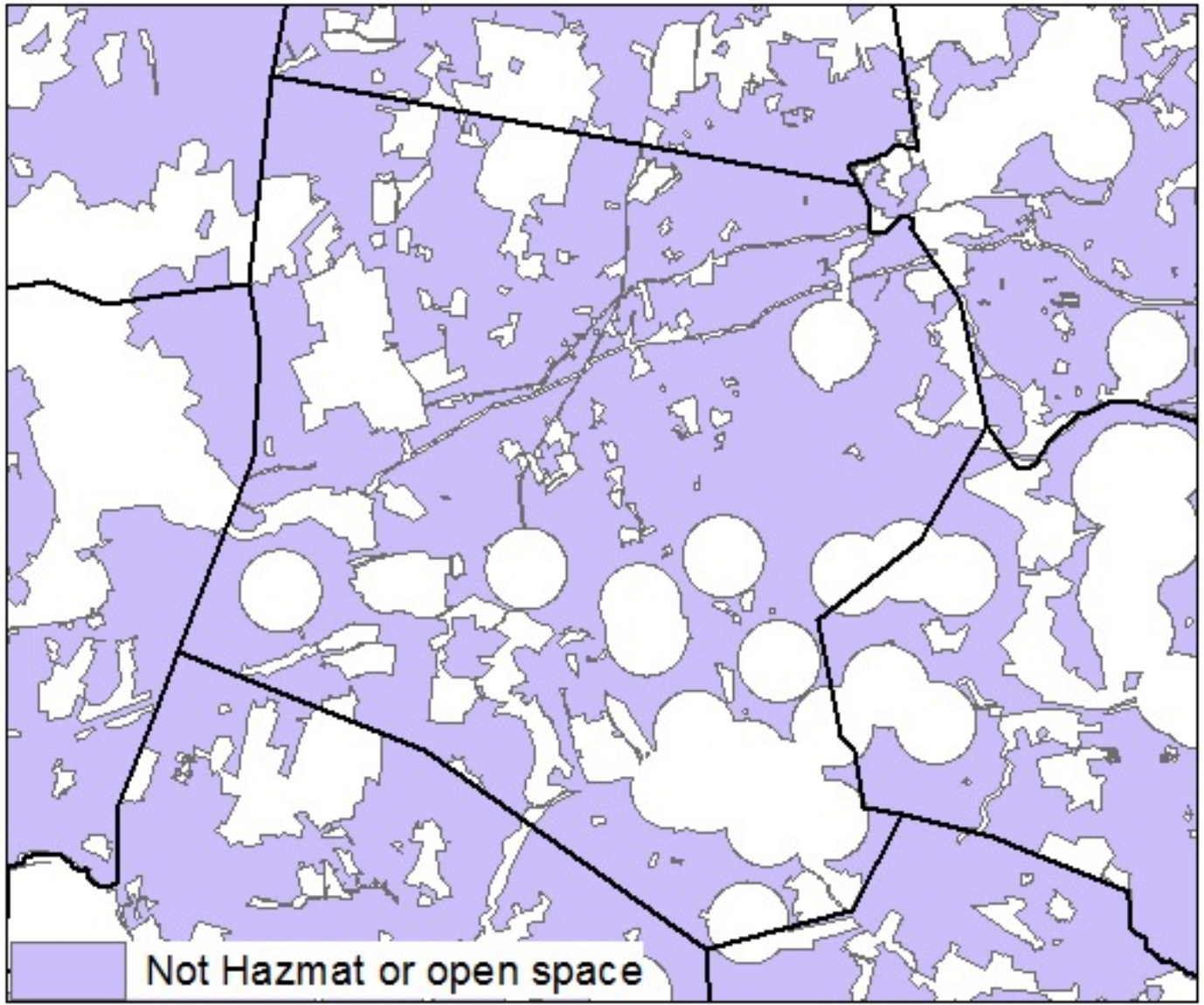
Vector data analysis: Build a school in Framingham



Far from Hazmat sites (buffer tool)



Not in open space or Hazmat buffer



Advantages of Vectors

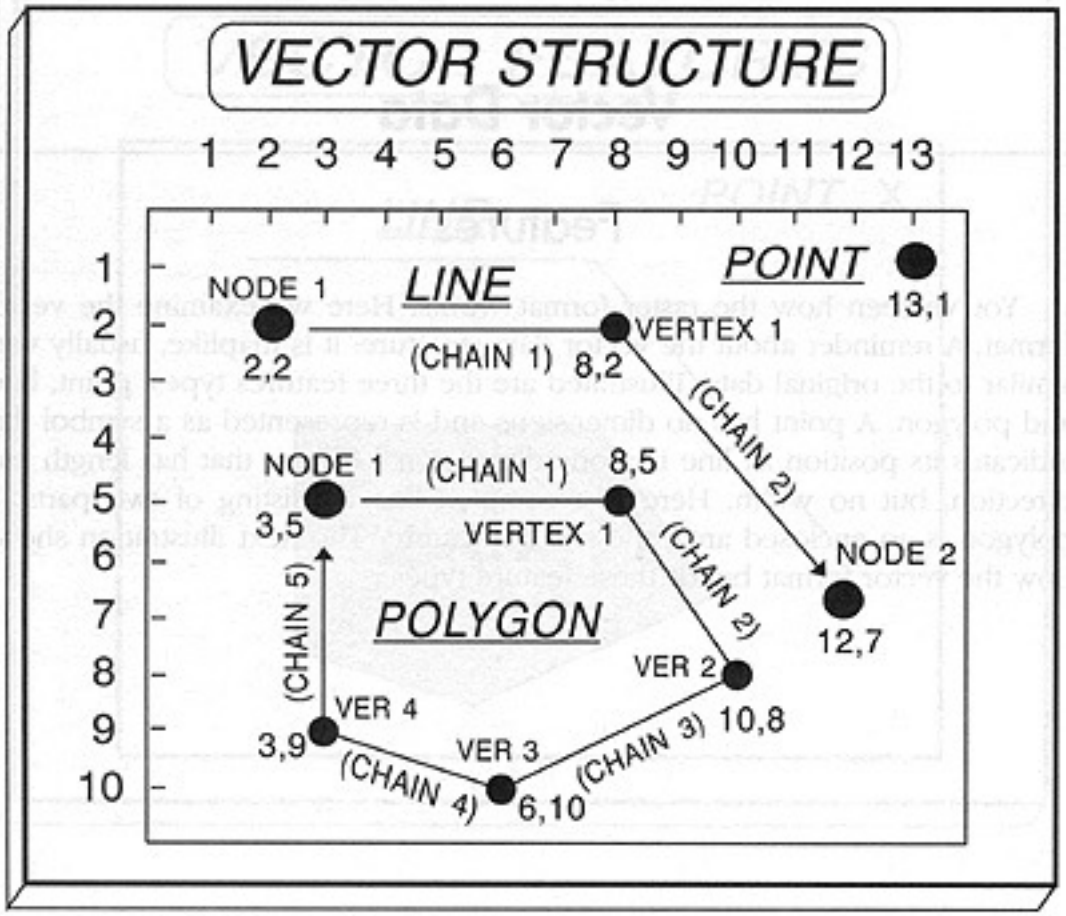
- Intuitively represent features from analog (hardcopy) maps
- Good at representing linear features
  - roads, rivers, boundaries
- Accurate: can follow features very closely
  - we can store with high precision the X,Y locations of points so we know their geographic location very precisely.

Early Vector Formats

- The first generation of vector files were simply lines, with arbitrary starting and ending points - just like a cartographer would draw a map.

- No explicit relationship between features - simply direct translation of a graphical image.

- Digitized ‘map’ files were not explicitly spatial or connected to a broader database.



http://ces.iisc.ernet.in/hpg/envis/Remote/section156.htm

Topology

Explicit information describing adjacency, connectivity, and relative spatial relationships of features

Why Topology is important…

- enables relationships to be established between features
- allows error detection to be done in a GIS
  - are all polygons completely closed
  - to the ends of arcs match up (“snapped together”)

Hierarchical Structures for Spatial Data

- Started in the 1960s and became the arc/node model.
- Based on the fact that each feature type is composed of features with one fewer dimensions.
- The problem is that it is necessary to keep track of the links between features in an arbitrary and sometimes cumbersome way.

Illustrated Hierarchical Model

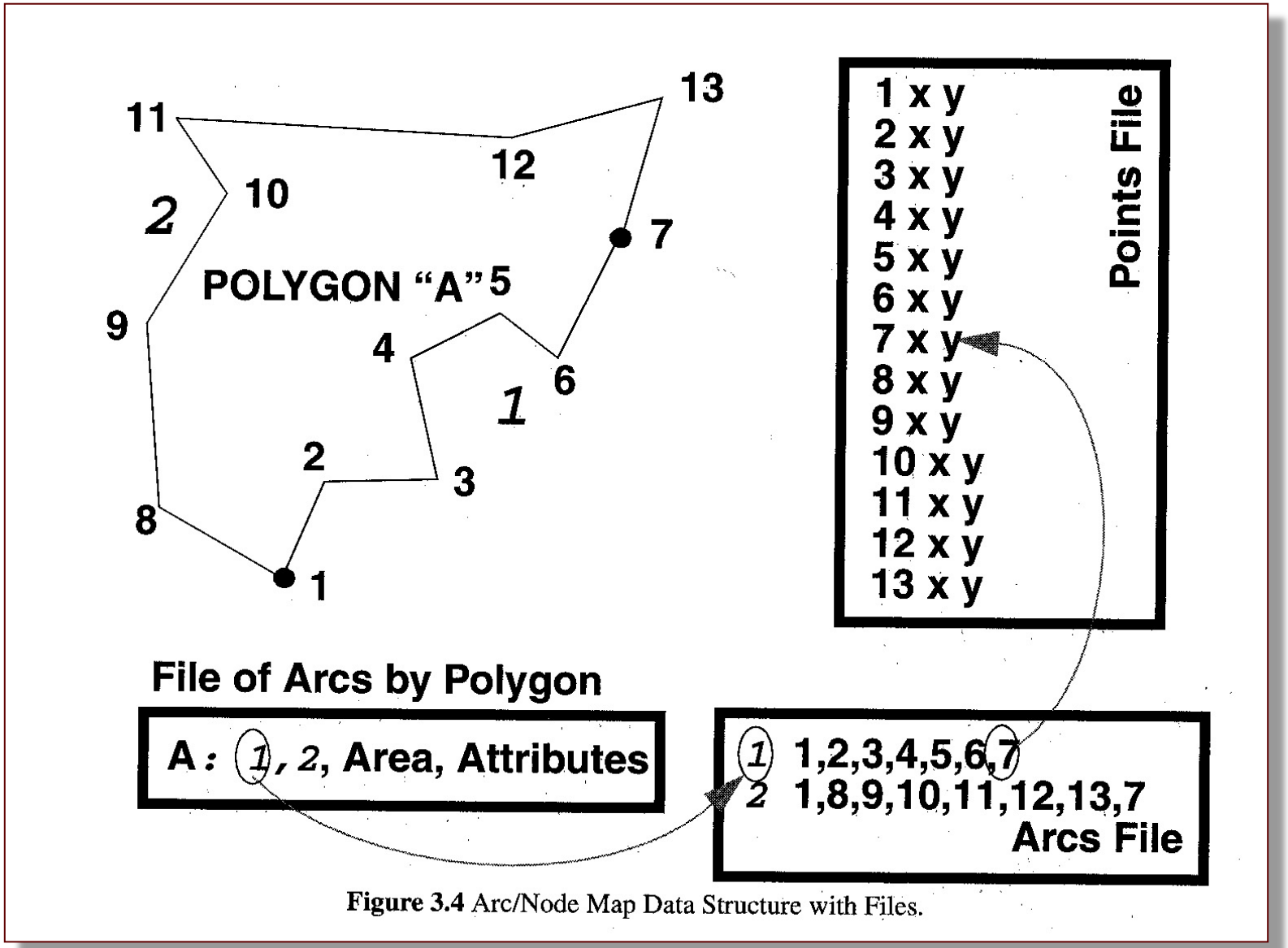


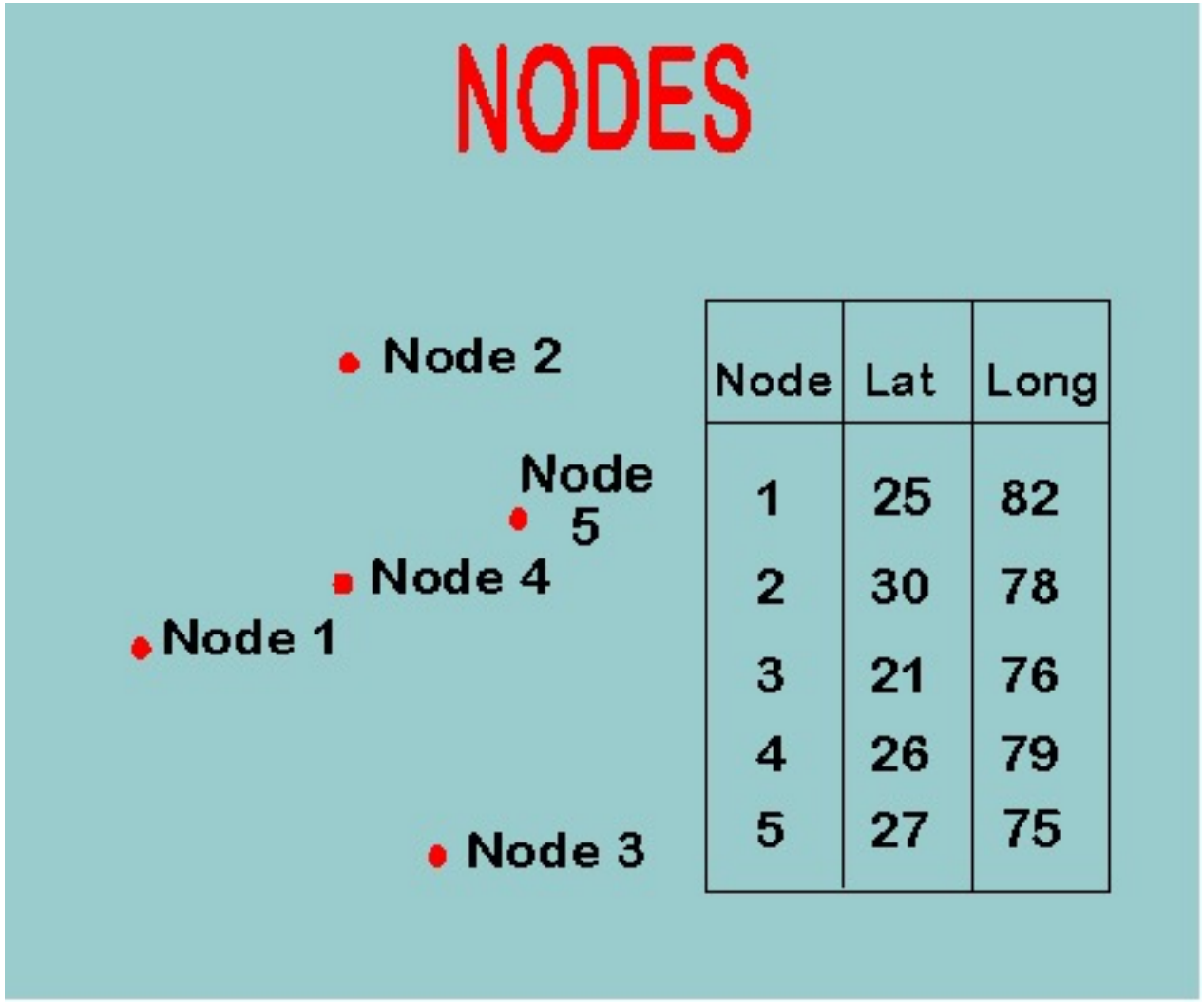
Figure 3.4 Arc/Node Map Data Structure with Files.

Figure From Clarke, K. 1999. *Getting Started with Geographic Information Systems* 2nd.

Improved Arc/Node Model

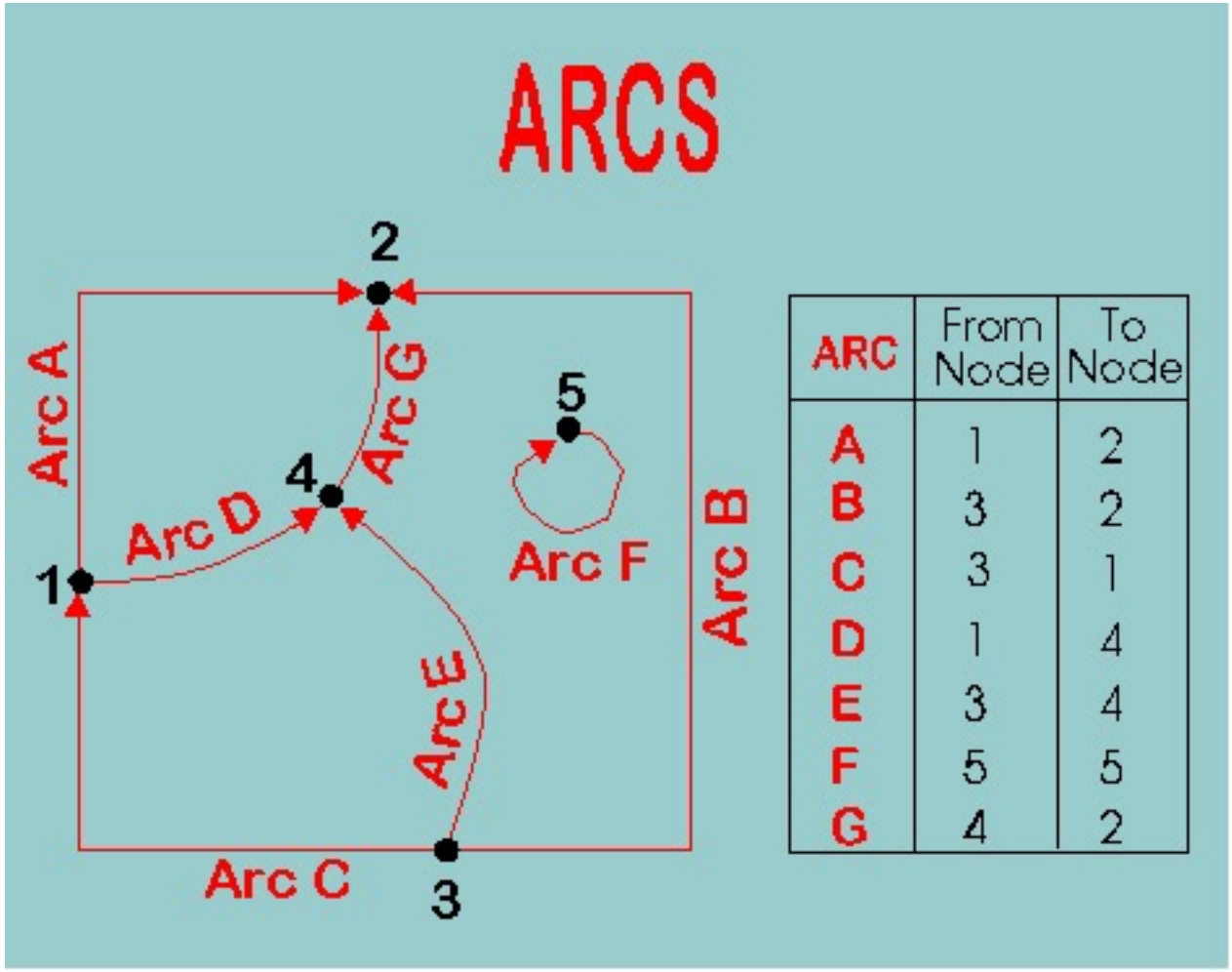
- Created after the First International Advanced Symposium on Topological Data Structures for Geographic Information Systems in 1979.

- The arc was selected as the fundamental data storage unit and polygons were reconstructed as necessary.



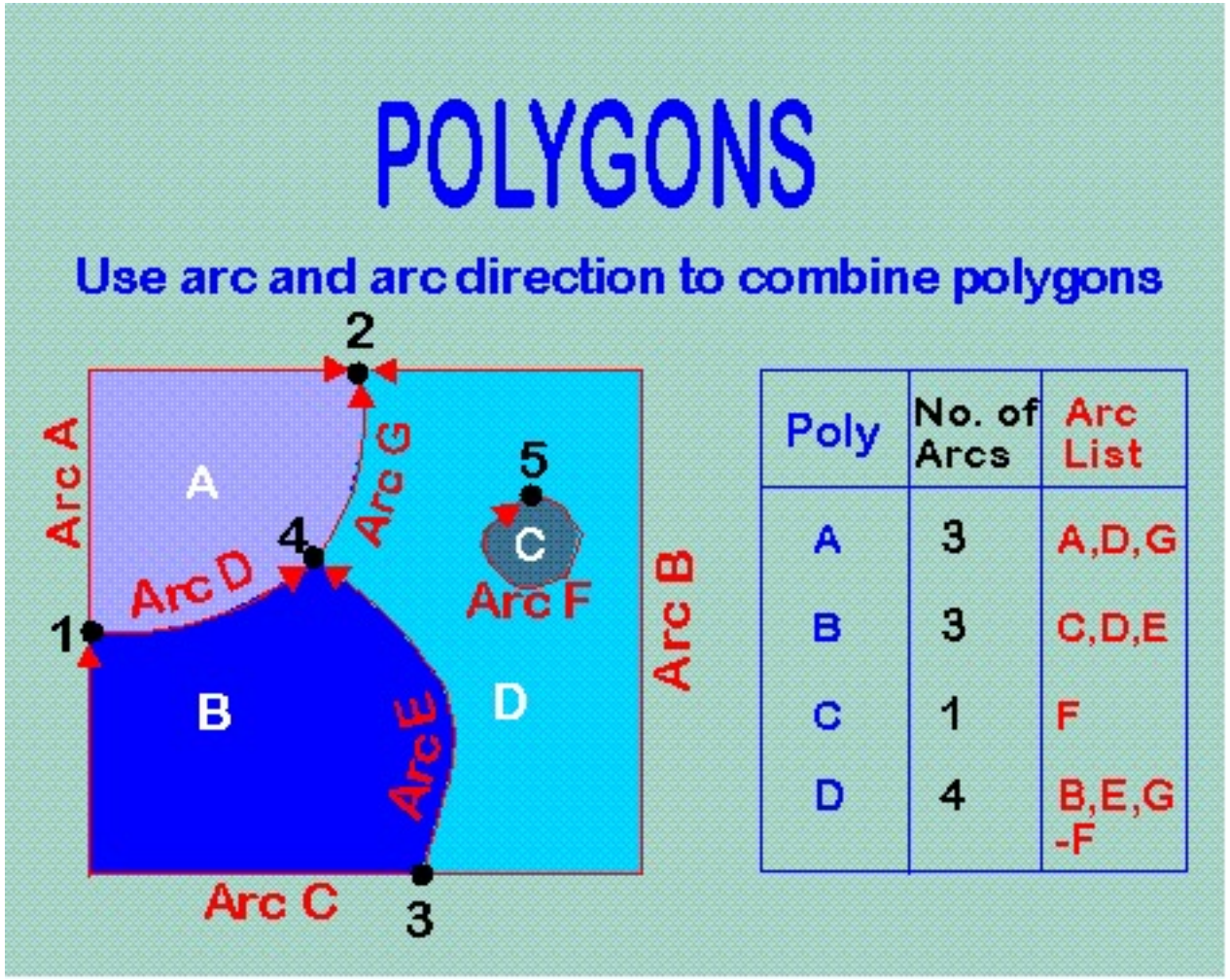
Node	Lat	Long
1	25	82
2	30	78
3	21	76
4	26	79
5	27	75

http://www.colorado.edu/geography/g

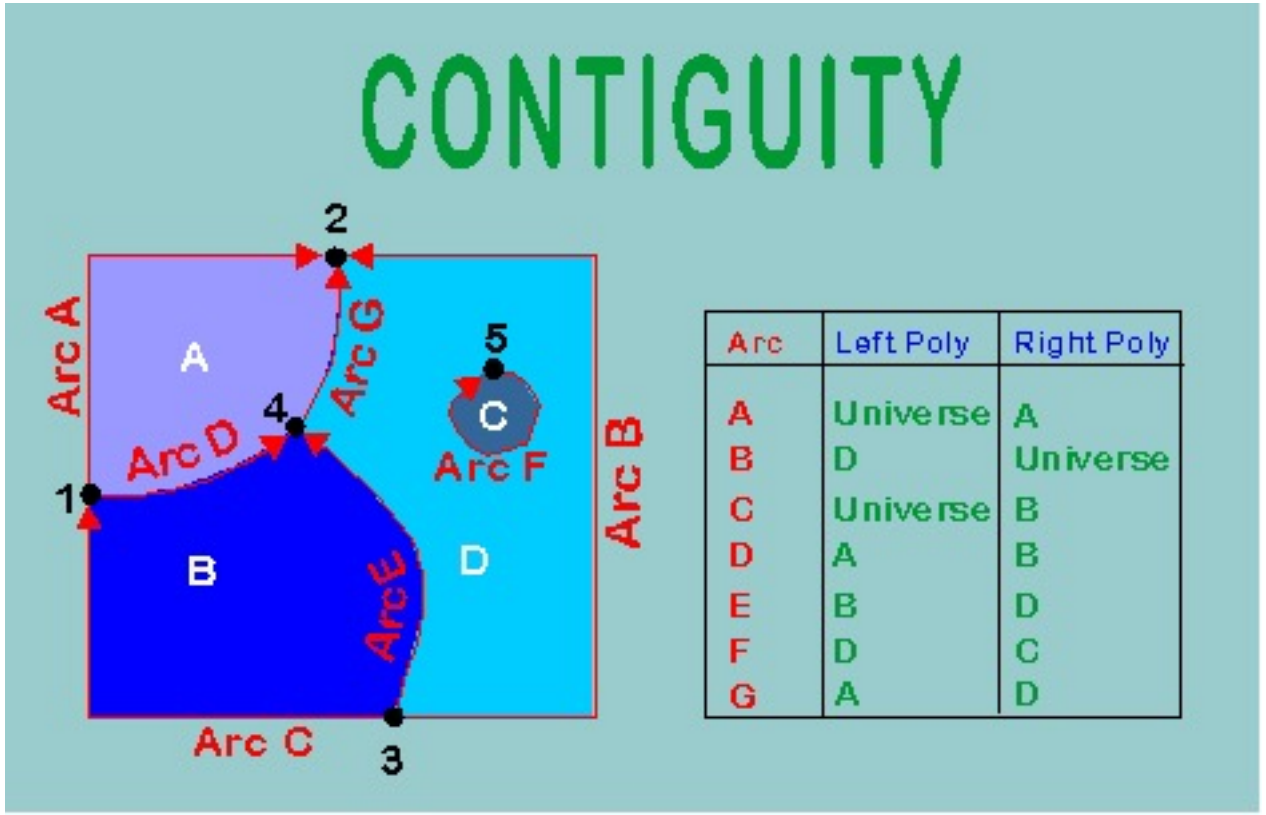


ARC	From Node	To Node
A	1	2
B	3	2
C	3	1
D	1	4
E	3	4
F	5	5
G	4	2

http://www.colorado.edu/geography/g



http://www.colorado.edu/geography/g



http://www.colorado.edu/geography/gcraft/notes/datacon/datacon.html