Geography 360: GIS & Mapping

Geographic Data Modeling

Data and Databases

Vaishnavi Thakar

UNIVERSITY of WASHINGTON

Review from Friday

- There are several different systems which we may use to georeference data
- Common referencing systems vary around the world
  - Place-names and points of interest
  - Postal addresses and postal codes
  - Linear referencing systems
  - Cadasters and the US Public Land Survey System
  - Measuring the Earth: latitude and longitude
- Projections

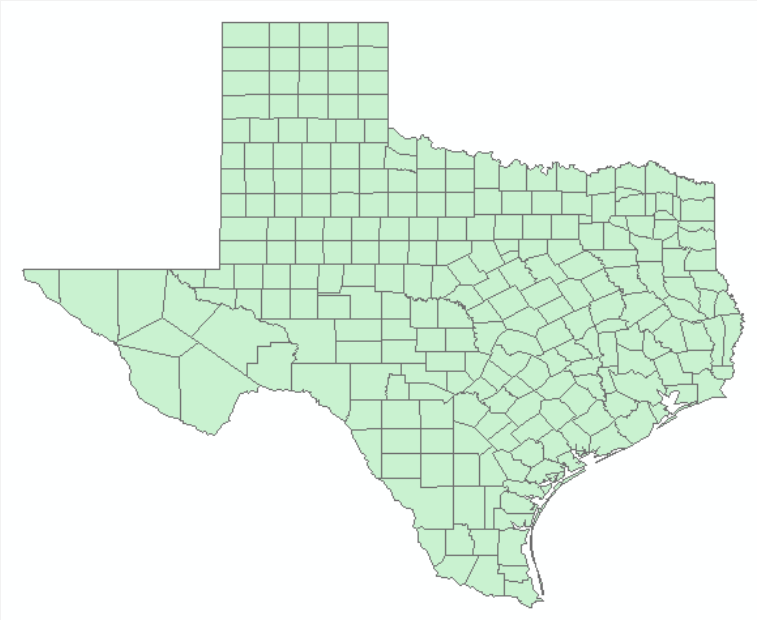
Learning Objectives

- **Define what geographic data models are and discuss their importance.**
- **Understand how to undertake GI data modeling.**
- **Outline the main geographic models used in GI systems today and their strengths and weaknesses.**
- **Understand key topology concepts and why topology is useful for data validation, analysis, and editing.**

Representing Geographic Features:
review from opening lecture

How do we describe geographical features?

- **by recognizing two *types of data*:**
  - Spatial data which describes location (where).
  - Attribute (Non-spatial) data which specifies characteristics at the location (what, how much, and when).



texas

SUMLEV	METRO	PMSA	AREANAME	PO01060D	PO01070D	PO01080D	PO01090D
1			TEXAS	9579677	11198655	14225513	16986510
5			Anderson, TX	28162	27789	38381	48024
5			Andrews, TX	13450	10372	13323	14338
5			Angelina, TX	39814	49349	64172	69884
5			Aransas, TX	7006	8902	14260	17892
3	9080		Archer, TX	6110	5759	7266	7973
5			Armstrong, TX	1966	1895	1994	2021
5			Atascosa, TX	18828	18696	25055	30533
5			Austin, TX	13777	13831	17726	19832
5			Bailey, TX	9090	8487	8168	7064
5			Bandera, TX	3892	4747	7084	10562
3	0640		Bastrop, TX	16925	17297	24726	38263
5			Baylor, TX	5893	5221	4919	4385
5			Bee, TX	23755	22737	26030	25135
3	3810		Bell, TX	94097	124483	157820	191088
3	7240		Bexar, TX	687151	830460	988971	1185394
5			Blanco, TX	3657	3567	4681	5972
5			Borden, TX	1076	888	859	799
5			Bosque, TX	10809	10966	13401	15125
3	8360		Bowie, TX	59971	68909	75301	81665
2	3362	1145	Brazoria, TX	76204	108312	169587	191707

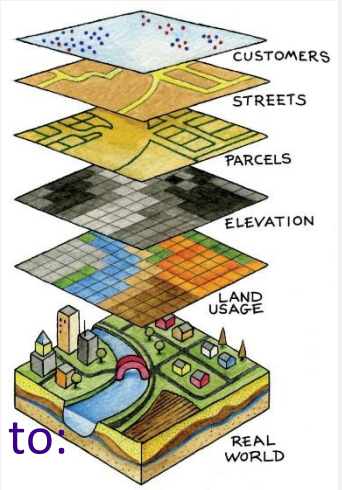
1 (0 out of 255 Selected)

Representing Geographic Features

How do we represent these geographic features digitally in a GIS?

- by grouping into layers based on similar characteristics (e.g hydrography, elevation, water lines, sewer lines, grocery sales) and using either:
  - vector data model (*shapefile* in ArcGIS)
  - raster data model (*GRID or Image* in ArcGIS)

A layer is a collection of geographic entities of same geometric type.



Layer Cake

- by selecting appropriate data properties for each layer with respect to:
  - projection, scale, accuracy, and resolution

How do we incorporate into a computer application system?

- by using a relational Data Base Management System (DBMS)

We introduced these concepts in the opening lecture. We will deal with them in more detail today (except for data properties which will be dealt with under Data Quality).

GIS Data Structures: *Topics Overview*

- **Spatial data types and Attribute data types**
- **Relational database management systems (RDBMS): basic concepts**
  - DBMS and Tables
  - Relational DBMS

- **raster data structures:** ***represents geography via grid cells***
  - **Tessellations** - A tessellation is created when a shape is repeated over and over again covering a plane without any gaps or overlaps.
  - **run length compression**
  - **BSQ/BIP/BIL**
  - **DBMS representation**
  - **File formats**

- **vector data structures:** ***represents geography via coordinates***
  - **whole polygon**
  - **point and polygon**
  - **node/arc/polygon**
  - **TINs**
  - **File formats**

- **Overview: representation of surfaces**

Spatial Data Types

- *Continuous:*
  - elevation, rainfall, ocean salinity

- *Areas:*
  - **unbounded**: landuse, market areas, soils, rock type
  - **bounded**: city/county/state boundaries, ownership parcels, zoning
  - **moving**: air masses, animal herds, schools of fish

- *Networks:*
  - roads, transmission lines, streams

- *Points:*
  - **fixed**: wells, street lamps, addresses, Brownfields
  - **moving**: cars, fish, deer

Attribute Data Types

Categorical (including letters and numbers):

- **Nominal**
  - no inherent ordering
  - landuse types, state names
  - Permissible operations
    - A=B
- **Ordinal**
  - inherent order
  - student letter grades, road class, city type
  - Permissible operations
    - A>B or A<B or A=B
  - often coded to numbers e.g. SSN but can't do arithmetic
  - may be expressed as string

Numerical (Difference between values is meaningful)

- **Interval**
  - No natural zero
  - can't say 'twice as much'
  - Permissible operations
    - A-B
  - temperature (Fahrenheit)
- **Ratio**
  - natural zero
  - ratios make sense (e.g. twice as much)
  - income, population density, age, weight
  - may be expressed as integer [whole number] or floating point [decimal fraction]

Attribute Data

- Attribute data tables can contain locational information, such as addresses or a list of X,Y coordinates.

- ArcGIS refers to these as *event tables*.

- However, these must be converted to true spatial data (shape file), for example by geocoding, before they can be displayed as a map.

Data Base Management Systems (DBMS)

- **Rows:**

  entities, records, observations, features

  - Each row corresponding to a different discrete object
  - All information about one occurrence of a feature

- **Columns:**

  attributes, fields, variables

  - Each column corresponding to an attribute of the object
  - One type of information for all features

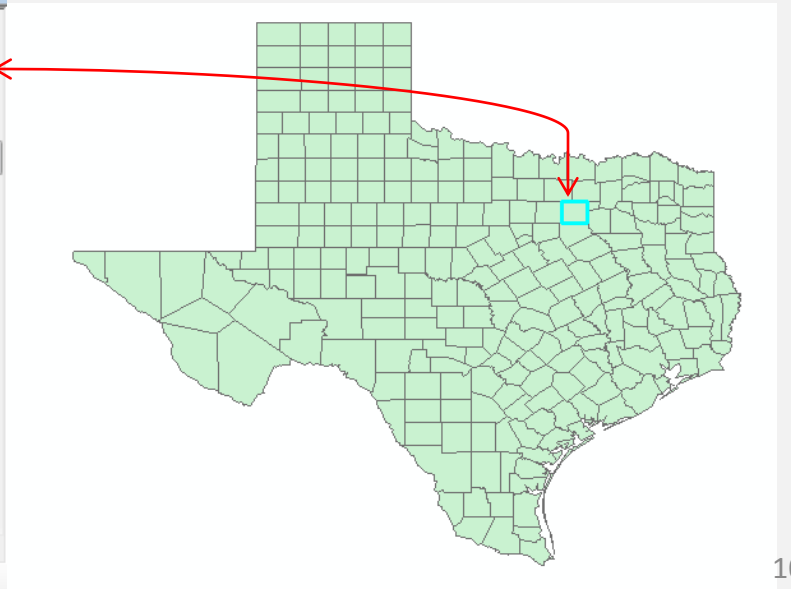
- **The key field is an attribute whose values uniquely identify each row**

texas

OID	ST	COU	SUMLEV	METRO	PMSA	AREANAME	PO01060D	PO01070D	PO01080D	PO01090D
56	48	111	5			Dallam, TX	6302	6012	6531	5461
57	48	113	2	1922	1920	Dallas, TX	951527	1327695	1556419	1852810
58	48	115	5			Dawson, TX	19185	16604	16184	14349
59	48	117	5			Deaf Smith, TX	13187	18999	21165	19153
60	48	119	5			Delta, TX	5860	4927	4839	4857
61	48	121	2	1922	1920	Denton, TX	47432	75633	143126	273525
62	48	123	5			De Witt, TX	20683	18660	18903	18840
63	48	125	5			Dickens, TX	4963	3737	3539	2571
64	48	127	5			Dimmit, TX	10095	9039	11367	10433
65	48	129	5			Donley, TX	4449	3641	4075	3696
66	48	131	5			Duval, TX	13398	11722	12517	12918
67	48	133	5			Eastland, TX	19526	18092	19480	18488
68	48	135	3	5800		Ector, TX	90995	92660	115374	118934
69	48	137	5			Edwards, TX	2317	2107	2033	2266
70	48	139	2	1922	1920	Ellis, TX	43395	46638	59743	85167
71	48	141	3	2320		El Paso, TX	314070	359291	479899	591610
72	48	143	5			Erath, TX	16236	18141	22560	27991
73	48	145	5			Falls, TX	21263	17300	17946	17712
74	48	147	5			Fannin, TX	23880	22705	24285	24804
75	48	149	5			Fayette, TX	20384	17650	18832	20095
76	48	151	5			Fisher, TX	7865			

0 (1 out of 255 Selected)

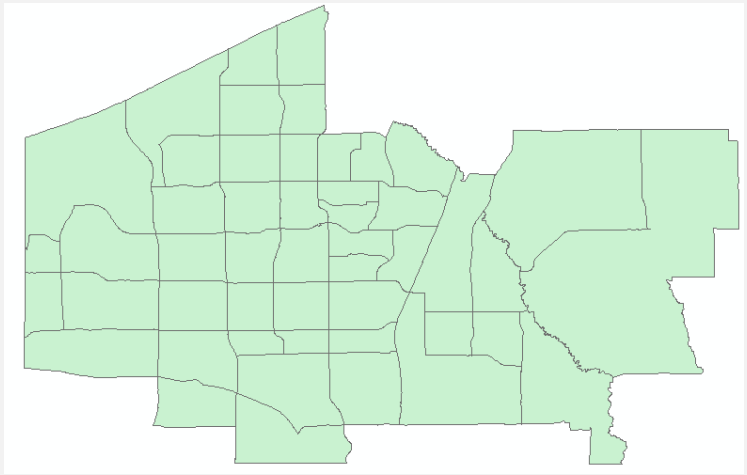
Attribute Table



Relational DBMS

- Tables are related, or *joined*, using a common record identifier (column variable), present in both tables, called a *secondary (or foreign) key*, which may or may not be the same as the key field.

Goal: Produce map of values by kid's density
Problem: no kid's density available in Tract table



TRACTS

FID	Shape *	AREA	PERIMET	TRACT	COUNT	AREA_SQM
0	Polygon	40997916	26635.33	0316.40	Collin	1.47
1	Polygon	29882178	23446.91	0316.39	Collin	1.07
2	Polygon	79922680	46537.53	0316.50	Collin	2.87
3	Polygon	26295940	20917.97	0316.41	Collin	0.94
4	Polygon	37443500	24260.71	0316.38	Collin	1.34

0 (0 out of 49 Selected)

Secondary or foreign key

Solution: join Planocen table, containing density values, with Tract table, containing location information, using TRACT as a key field

PLANOCEN

OID	TRACT *	COUNTY	AREA	AREA_SQM	POP	WHITE	BLACK	AMIND	ASIAN
0	0316.40	Collin	40997916	1.47	3506	2989	114	21	195
1	0316.39	Collin	29882178	1.07	1753	1354	148	40	172
2	0316.50	Collin	79922680	2.87	4680	3461	247	32	794
3	0316.41	Collin	26295940	0.94	3780	2561	162	14	814
4	0316.38	Collin	37443500	1.34	5640	4254	154	0	1046

0 (0 out of 49 Selected)

Spatial Data

- **There are two fundamental ways of representing spatial data:**
  - Discrete Objects
    - The discrete object view represents the geographic world as objects with **well-defined boundaries in otherwise empty space**.
    - Objects can be **counted**.
    - Objects have dimensionality: 0-dimension (points), 1-dimension (lines), 2-dimensions (areas, polygons).
    - E.g. Animals (bears), manufactured objects (cars), buildings
  - Continuous Fields
    - The continuous field view represents the real world as a finite number of variables, each one defined at every possible position.
    - E.g. Elevation, population density, land use, soil type

Spatial Data

Monitoring Bear Population

Bears are easily conceived as discrete objects (points), maintaining their identity as objects through time and surrounded by empty space.

Spatial Data

(Oliviero Olivieri/Getty Images, Inc.)

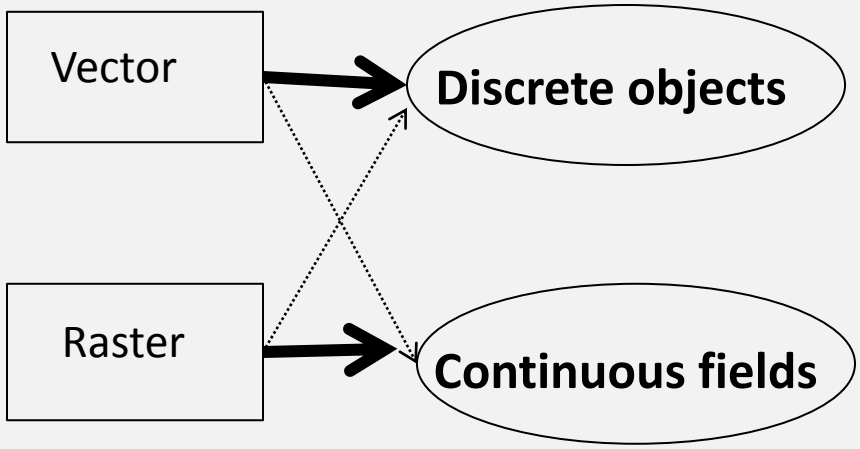
Lakes are difficult to conceptualize as discrete objects because it is often difficult to tell where a lake begins and ends, or to distinguish a wide river from a lake.

If discrete object view : Can count number of lakes

If continuous filed view: All points are either lake or nonlake

Spatial Data Structures

- Raster and vector are two methods that are used to reduce geographic phenomena to forms that can be coded in computer databases.
- In principle, each can be used to code both discrete objects and continuous fields, but in practice there is a strong association between vector and discrete objects, and between raster and continuous fields.



Representing trees/forest ?

Depends on scale of mapping

Representing Data

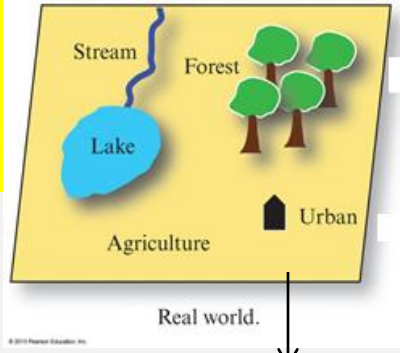
Different representational models of the same area in Colorado, USA

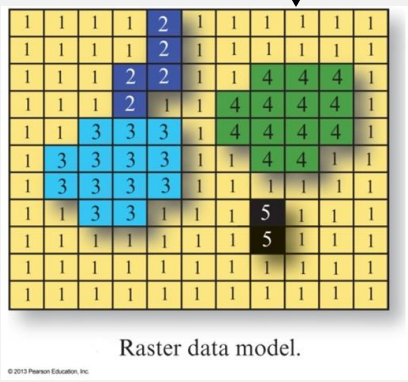
Roads, buildings, cars, other features on the surface

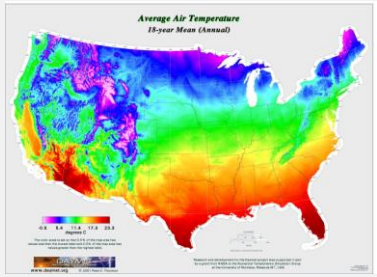
Roads, buildings, lot/property boundaries, water mains and valves

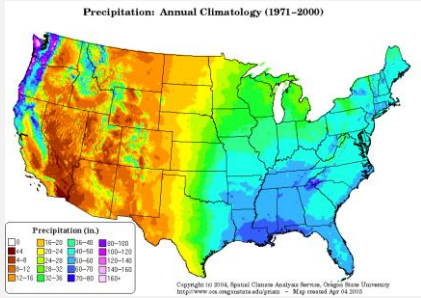
Representing Data using *Raster* Model

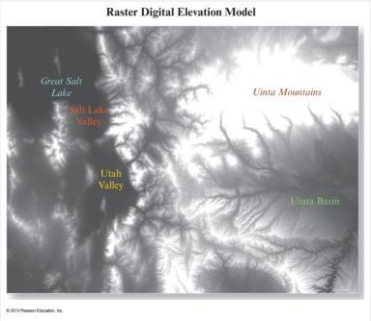
- Area is covered by grid with (usually) equal-sized cells.
- **location** of each cell calculated from origin of grid:
  - "two down, three over"
- cells often called *pixels* (picture elements); raster data often called *image* data
- **attributes** are recorded by assigning each cell a single value based on the majority feature (attribute) in the cell, such as land use type.
- There are no gaps in the coverage
- easy to do overlays/analyses, just by 'combining' corresponding cell values: "*yield= rainfall + fertilizer*" (why raster is faster, at least for some things)









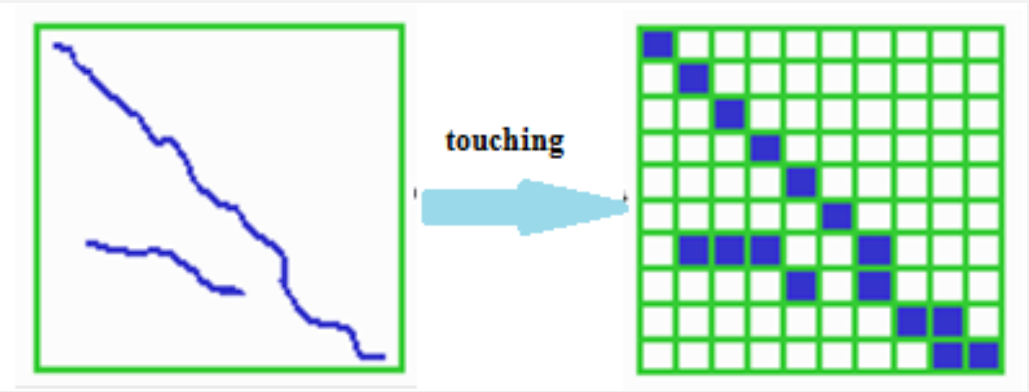


Raster Data Structures

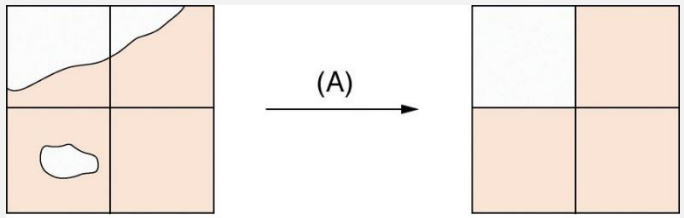
- **Grid often has its origin in the upper left but note:**
  - State Plane and UTM, lower left
  - lat/long & cartesian, center

- **Single values associated with each cell**
  - typically 8 bits assigned to values therefore 256 possible values (0-255)

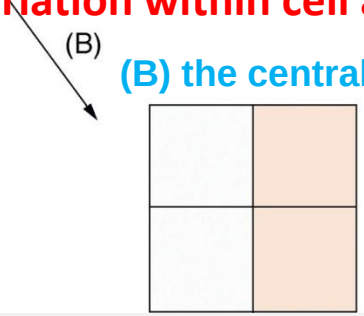
- **Rules needed to assign value to cell if object does not cover entire cell**
  - the largest share rule
  - the central point rule
  - the touching rule for linear feature such as road



(A)the largest share rule



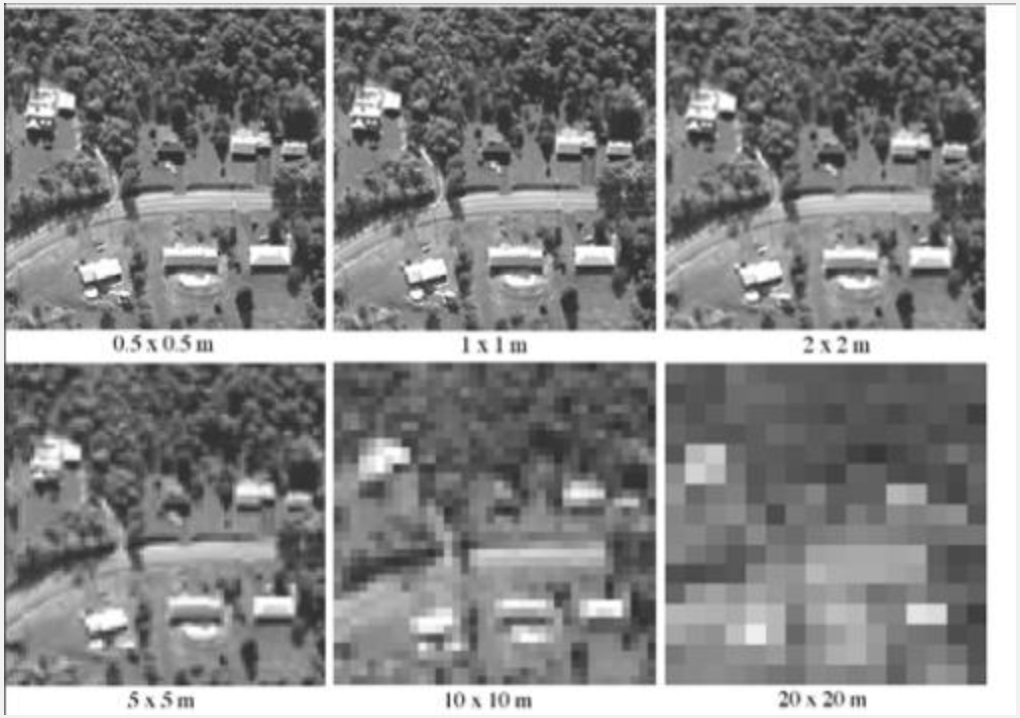
Details about variation within cell are lost!

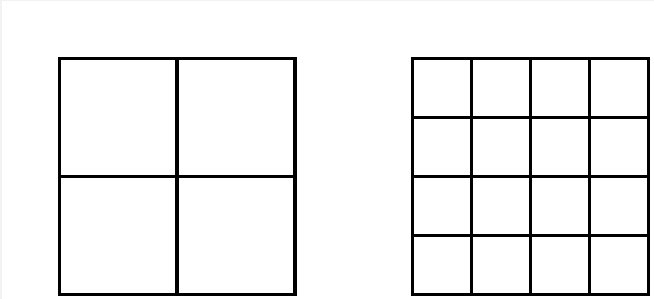


(B) the central point rule

Raster

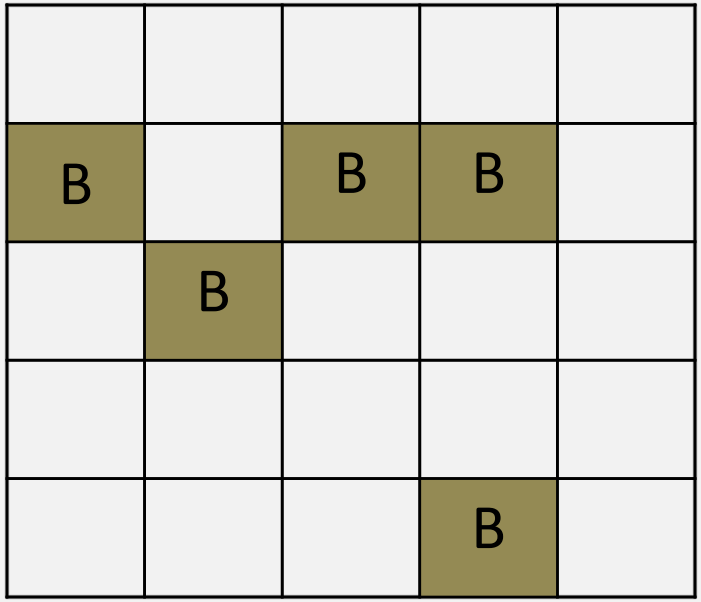
- ***Resolution*** **of a raster is the distance that one side of a grid cell represents on the ground.**
- **The higher the resolution (smaller the grid cell), the higher the precision, but the greater the cost in data storage.**



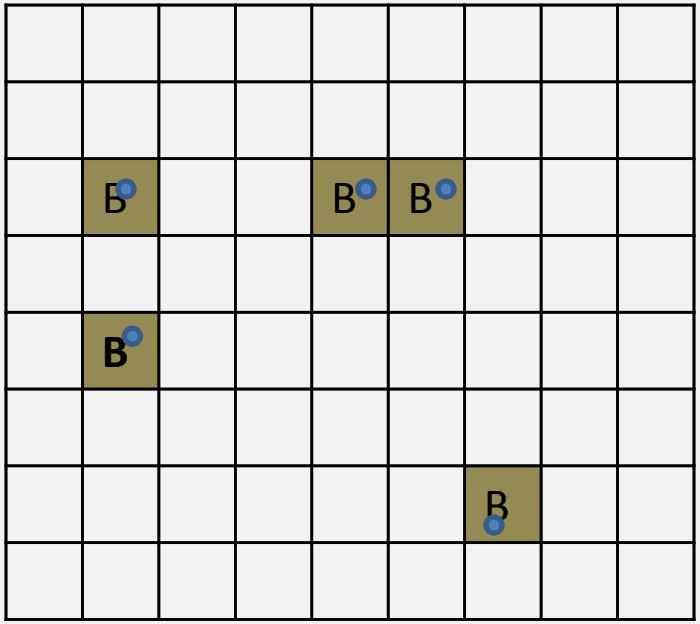


Raster

30 x 30 m



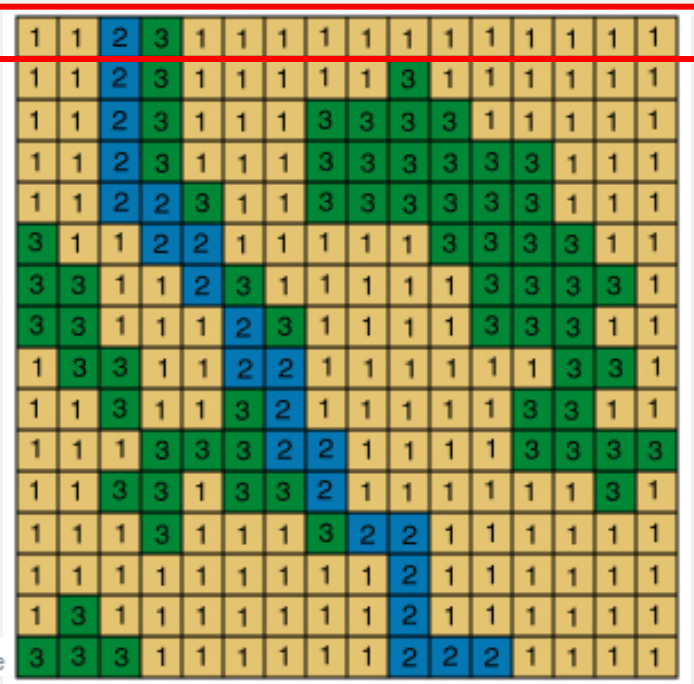
0.5 x 0.5 m



Accuracy of bear locations in Raster?

Raster Data Structures

- ***Runlength Compression (for <u>single</u> layer)***
  - A way of compressing raster data based on eliminating redundancy for attributes along rows of grids.
  - Improve storage efficiency.
  - *"Value through column" coding: 1st number is value, 2nd is last column with that value.*
  - This is a "*lossless*" compression, as opposed to "*lossy,"* since the original data can be exactly reproduced (no loss/degradation of information).



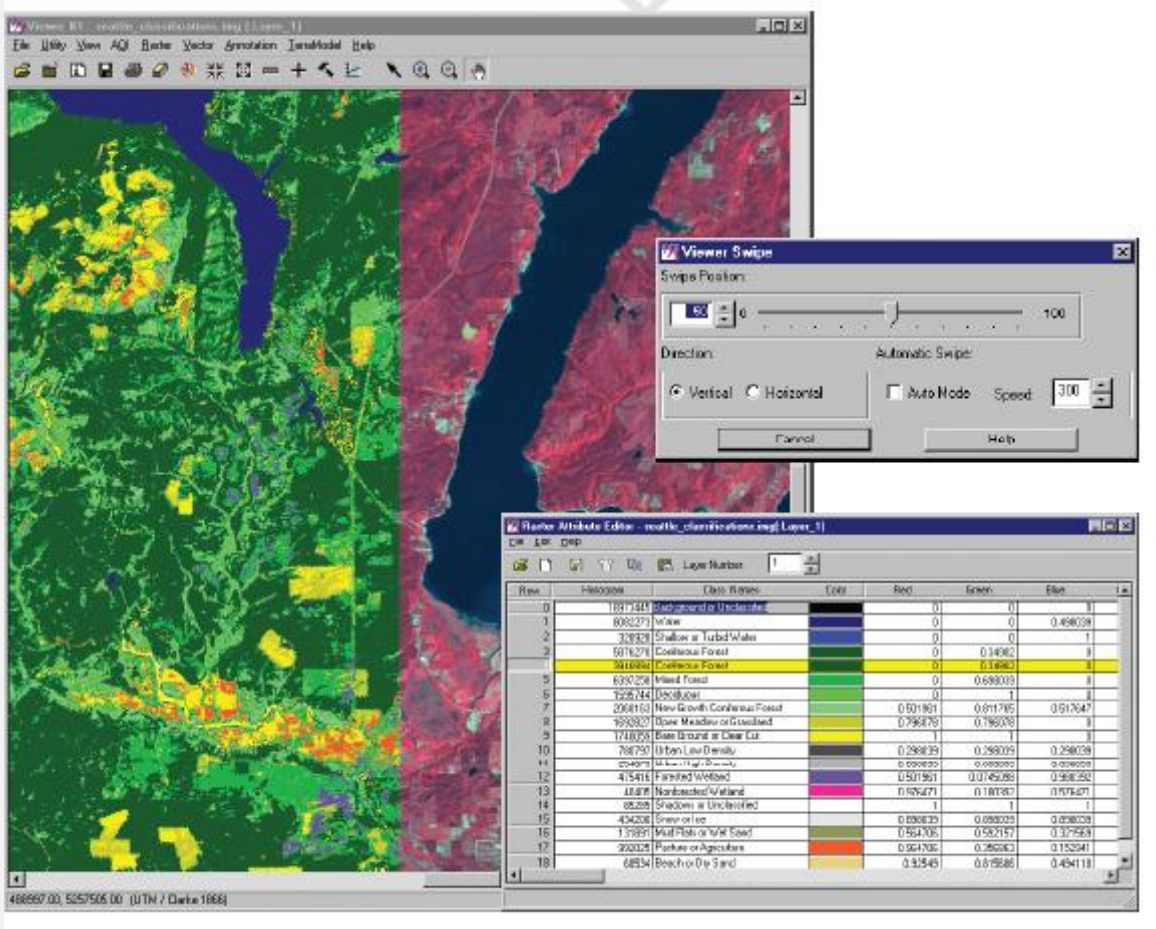
1 2 2 3 3 4 1 16
1 2 2 3 3 4 1 9 3 10 1 16
1 2 2 3 3 4 1 7 3 11 1 16
1 2 2 3 3 4 1 7 3 13 1 16
1 2 2 4 3 5 1 7 3 13 1 16
3 1 1 3 2 5 1 10 3 14 1 16
3 2 1 4 2 5 3 6 1 11 3 15 1 16
3 2 1 5 2 6 3 7 2 11 3 14 1 16
1 1 3 3 1 5 2 7 1 13 3 15 1 16
1 2 3 3 1 5 3 6 2 7 1 12 3 14 1 16
1 3 3 6 2 8 1 12 3 16
1 2 3 4 1 5 3 7 2 8 1 14 3 15 1 16
1 3 3 4 1 7 3 8 2 10 1 16
1 9 2 10 1 16
1 1 3 2 1 9 2 10 1 16
3 3 1 9 2 12 1 16

Credit: Photo Source

Full Matrix--256 bytes **Run Length (row)--188 bytes**

Raster Data Structures

In some systems, multiple attributes can be stored for each cell in a type of value attribute table where each column is an attribute and each row either a pixel or a pixel class



Raster data of the Olympic Peninsula, Washington State, with associated value attribute table. Bands 4,3,2 from Landsat 5 satellite with land cover classification overlaid

File Formats for Raster Data

The generic raster data model is actually implemented in several different computer file formats:

- **GRID is ESRI's proprietary format for storing and processing raster data.**
- **Standard industry formats for image data such as JPEG, TIFF, DAT, IMG and MrSid formats can also be imported into ArcGIS**

Conclusions

- Discrete objects and continuous fields
  - Two fundamental ways of representing geography

- Data model: describing and representing parts of the real world in a digital computer system

- Raster and vector
  - two methods of representing geographic data in digital computers

Questions ?

https://www.google.com/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwihuqvgkpjcAhU3IDQIHZbqB8QQjRx6BAgBEAU&url=http%3A%2F%2Fwww.cityofrockhill.com%2Fdepartments%2Finformation-technology-services%2Fmore%2Finformation-technology-services%2Fgeographic-information-systems-gis-%2Fgis-frequently-asked-questions&psig=AOvVaw2fEUXAIjbJY26W-bnjS0wY&ust=1553136220322311

Upcoming

- Wednesday (Lecture) : Data and Databases II
- Submit Assignments
- Readings updated on canvas.
- Week 4, GIS Lab 04 will be assigned.
- DRS Requests: Contact DRS for exam 1.