Lecture 1: What is a GIS?

1.1 Getting Started

1.2 Some Definitions of GIS

1.3 A Brief History of GIS

1.4 Sources of Information on GIS

Lecture 1: What is a GIS?

- GIS (usually) stands for **Geographic Information System**.
- It is **comprised** of hardware, software, network, data, and trained personnel to support the capture, management, manipulation, analysis, and display of geographically referenced data for solving complex municipal management and planning problems, and for serving the public better and more efficiently.

Defining GIS

- **Different definitions** of a GIS have evolved in different areas and disciplines
  - a toolbox
  - an information system
  - an approach to science
  - an multi-billion dollar business
  - plays an important role in society
- All GIS definitions recognize that **spatial data are unique** because they are linked to maps (Space matters!)
- A GIS **at least consists** of a database, map information, and a computer-based link between them

Definition 1: A GIS is a toolbox

- "*a powerful set of tools for storing and retrieving at will, transforming and displaying spatial data from the real world for a particular set of purposes*"

  (Burrough, 1986, p. 6).

- "*automated systems for the capture, storage, retrieval, analysis, and display of spatial data*." (Clarke, 1995, p. 13).

Definition 2: A GIS is an information system

"*An information system that is designed to work with data referenced by spatial or geographic coordinates. In other words, a GIS is both a database system with specific capabilities for spatially-referenced data, as well as a set of operations for working with the data*" (Star and Estes, 1990, p. 2).

Dueker's 1979 definition (p. 20) has survived the test of time.

"*A geographic information system is a special case of information systems where the database consists of observations on spatially distributed features, activities or events, which are definable in space as points, lines, or areas. A geographic information system manipulates data about these points, lines, and areas to retrieve data for ad hoc queries and analyses*" (Dueker, 1979, p 106).

Definition 3: GIS is an approach to science

- **Geographic Information Science** is research both *on* and *with* GIS.
- The technology of GIS has become **much simpler, more distributed, cheaper** and has crossed the boundary into disciplines such as anthropology, epidemiology, facilities management, forestry, geology, and business.
- GIS is used as **a new approach** to science.

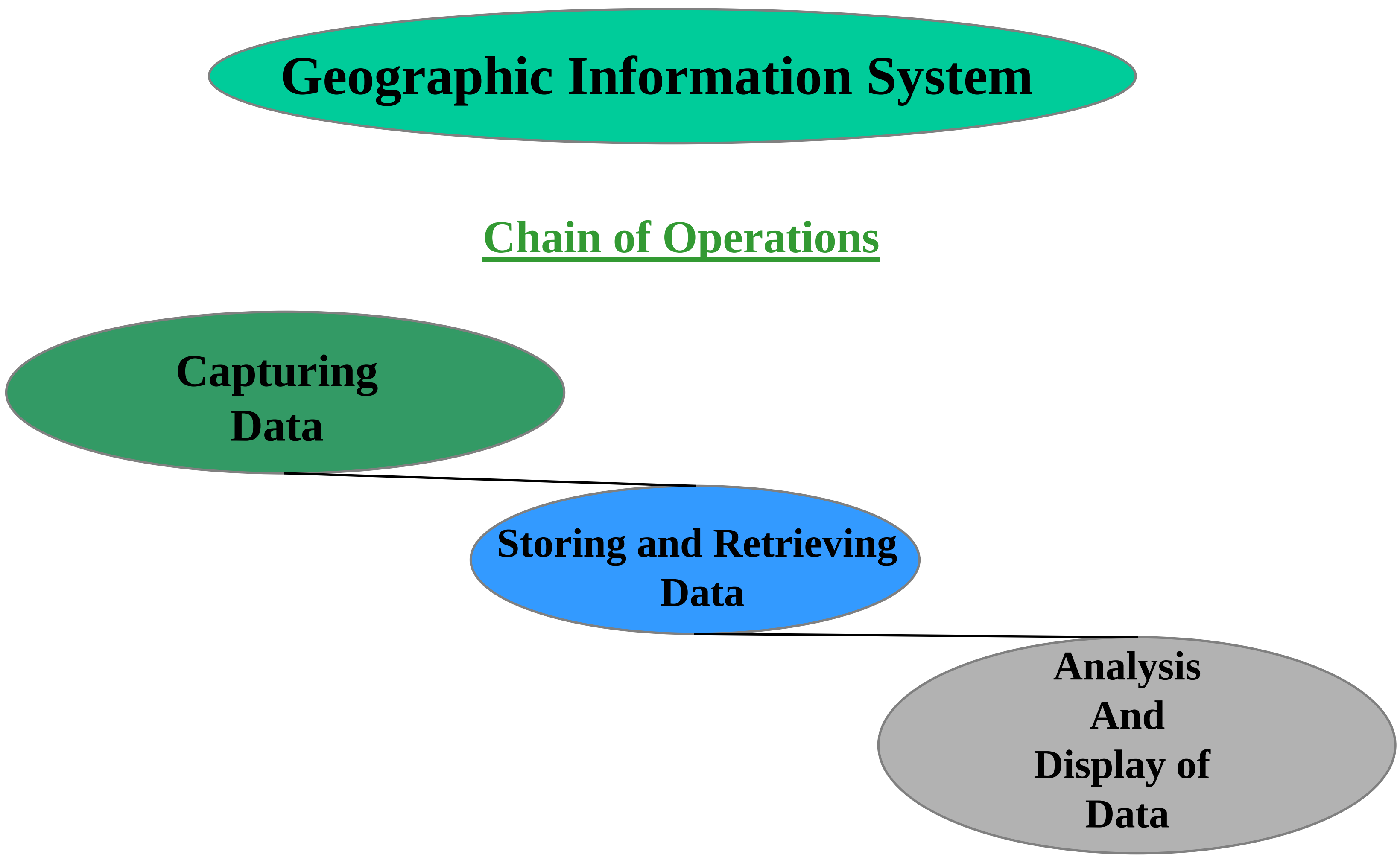
Definition 4: GIS is a multi-billion dollar business

"*The growth of GIS has been a marketing phenomenon of amazing breadth and depth and will remain so for many years to come. Clearly, GIS will integrate its way into our everyday life to such an extent that it will soon be impossible to imagine how we functioned before*"

Definition 5:
GIS plays a role in society.

Nick Chrisman (1999) has defined GIS as "organized activity by which people measure and represent geographic phenomena, and then transform these representations into other forms while interacting with social structures."

How Does GIS Work?



GIS's Roots in Cartography

- Earth models
- Datum
- Geographic coordinates
- Map projections
- Coordinate systems
- Basic properties of geographic features

The Elements of GIS

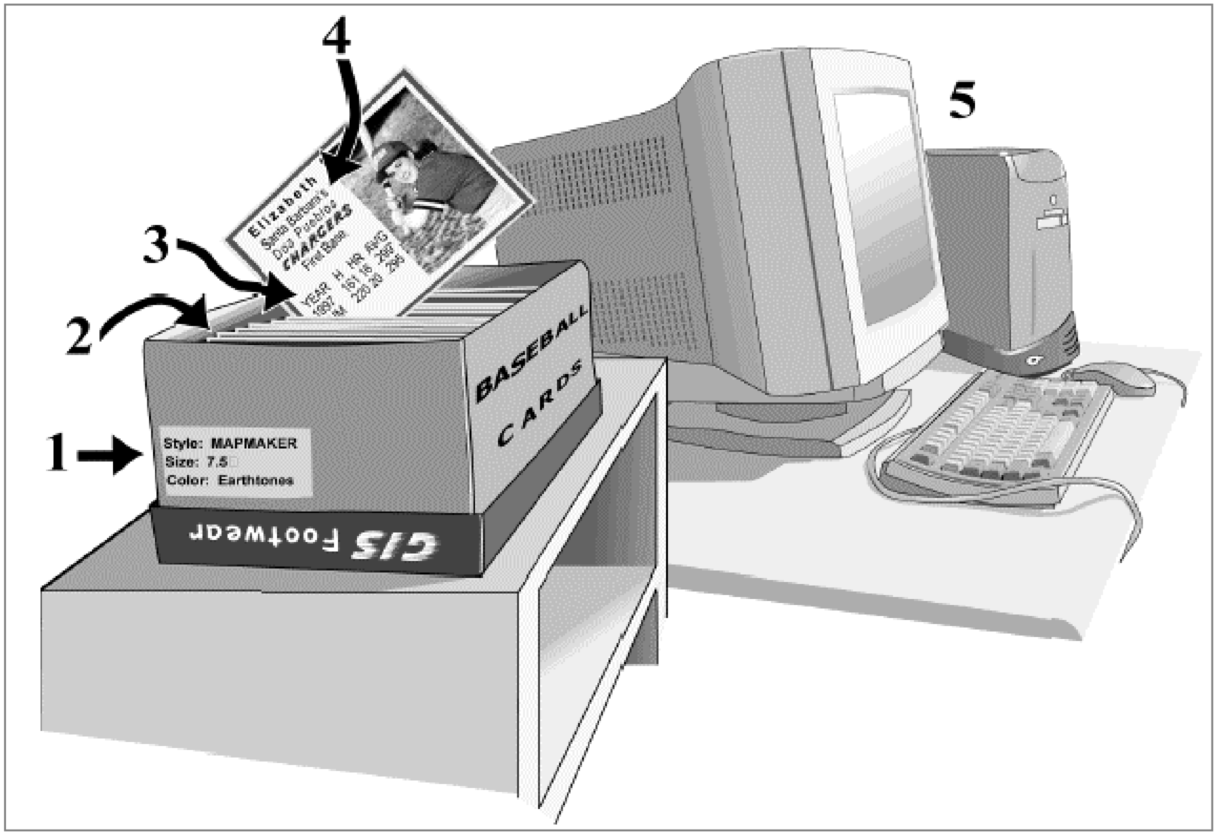


Figure 2.1 The elements of a GIS. (1) The database (shoebox); (2) the records (baseball cards); (3) the attributes (the categories on the cards, such as batting average, (4) the geographic information (locations of the team's stadium in latitude and longitude); (5) a means to use the information (the computer).

The GIS Database

- In a database, we store **attributes as column headers** and **records as rows**.
- The contents of an attribute for one record is a **value**.
- A value can be **numerical or text**.

The GIS Database (continued)

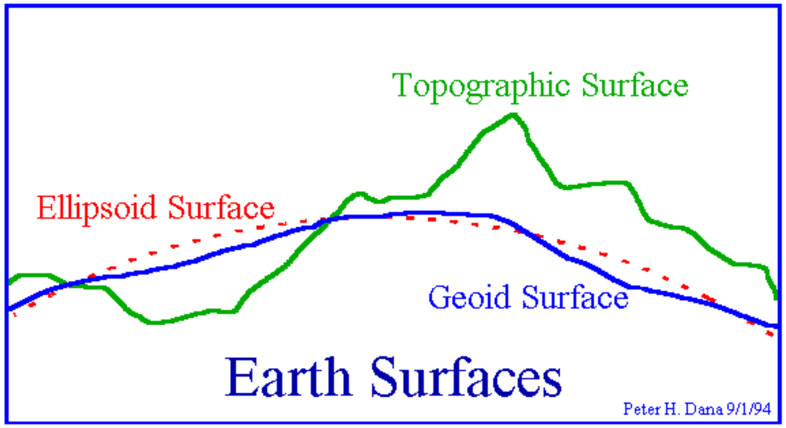
- Data in a GIS must contain a **geographic reference** to a map, such as latitude and longitude.
- The GIS cross-references **the attribute data with the map data**, allowing searches based on either or both.
- The **cross-reference is a link**.

Models of the Earth

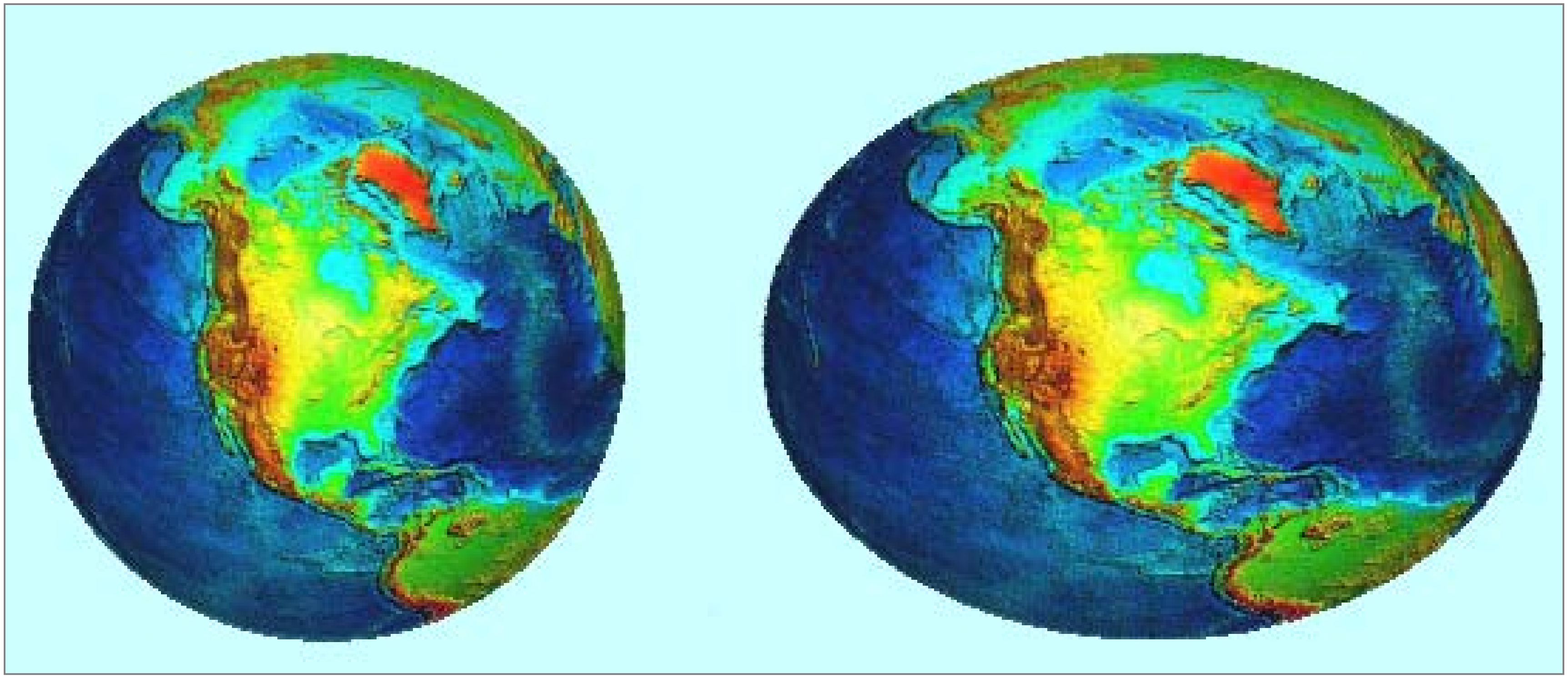
A Sphere

An Ellipsoid

A Geoid



Earth Shape: Sphere and Ellipsoid



Pole to pole distance: 39,939,593.9 meters
Around the Equator distance: 40,075,452.7 meters

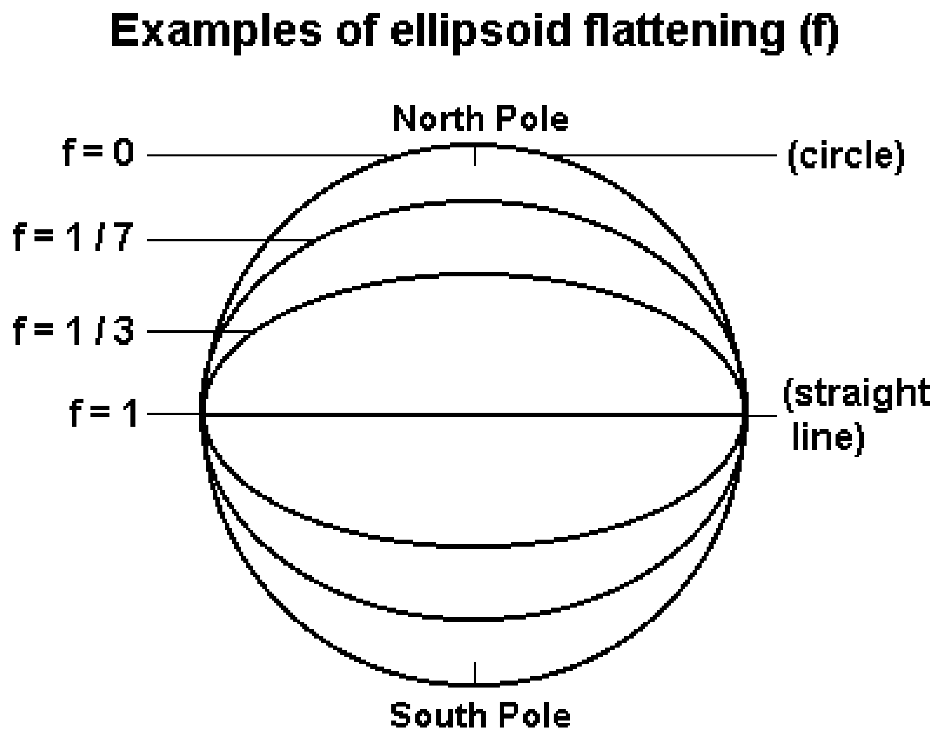
Ellipticity of the Earth

•**Newton** estimated the Earth's ellipticity to be about f = 1/300

•**Modern satellite technology** gives an f = 1/298 (~0.003357)

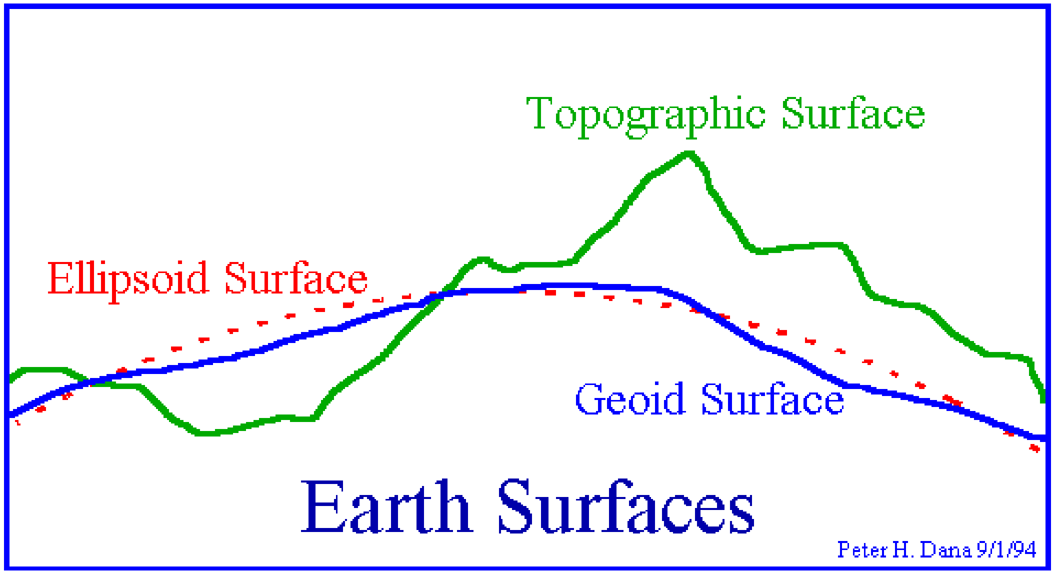
↑

These small values of f tell us that the Earth is very close to being a sphere, but not close enough to ignore its ellipticity if we want to accurately locate features on the Earth



The Earth as Geoid

- **Geoid** → The surface on which gravity is the same as its strength at mean sea level
- **Geodesy** is the science of measuring the size and shape of the earth and its gravitational and magnetic fields.



Earth Models and Datums

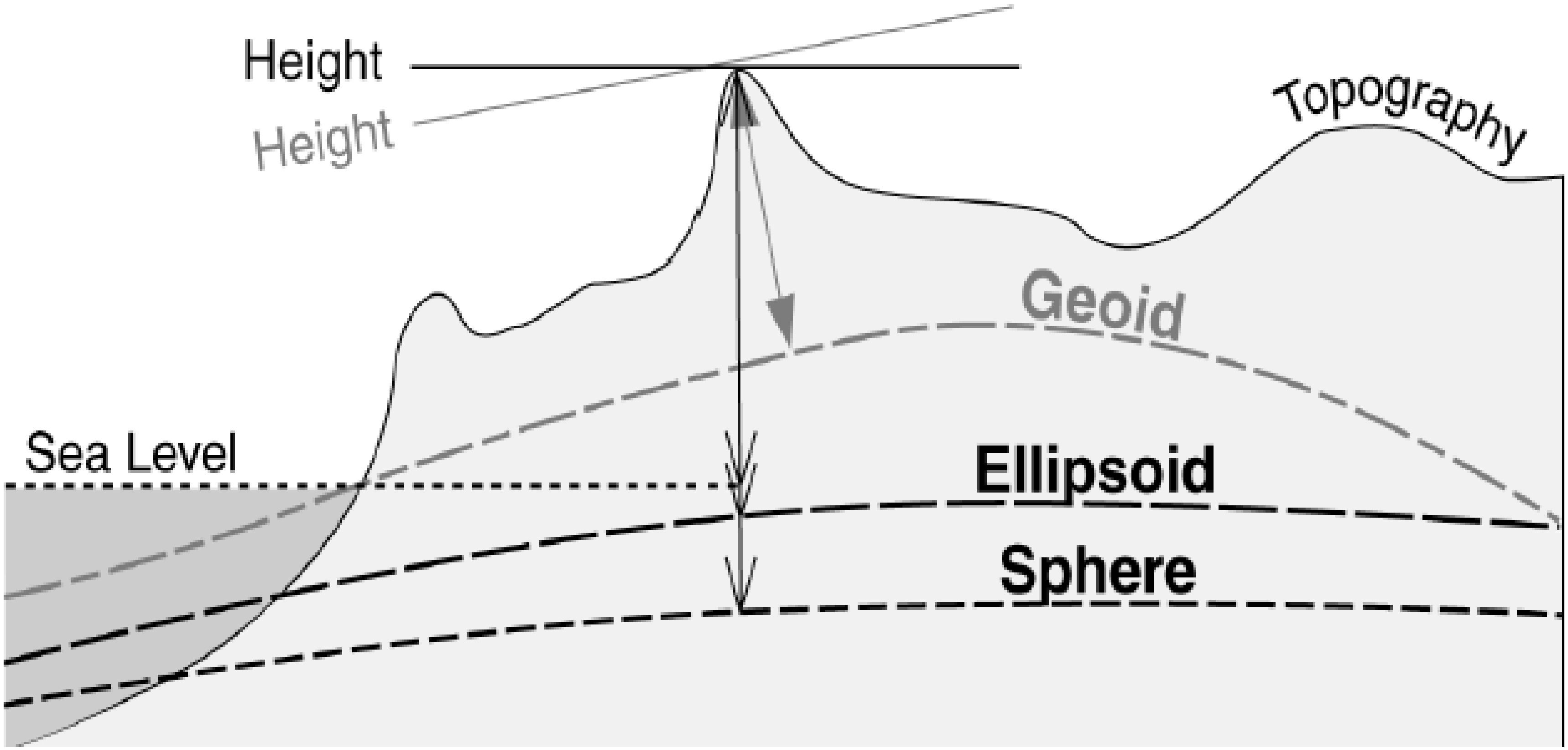


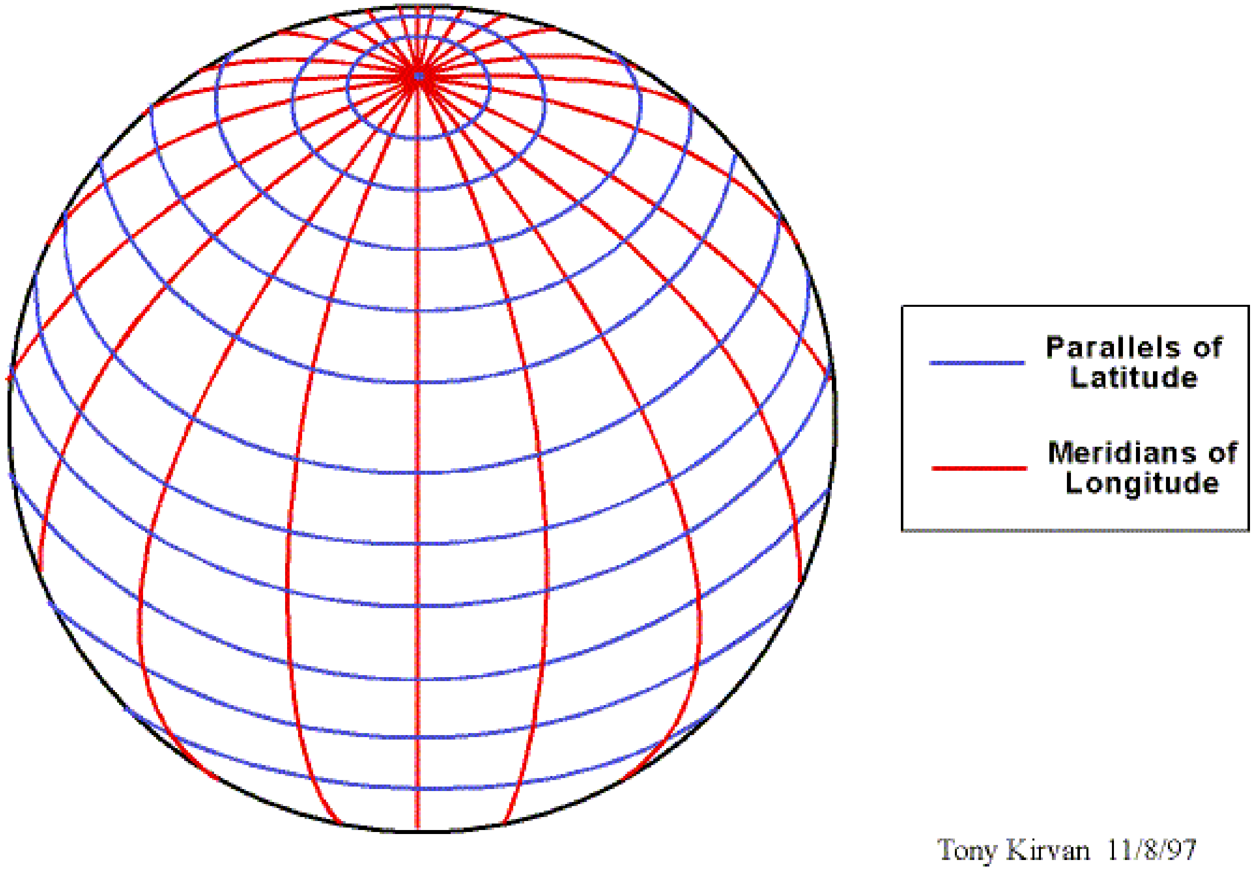
Figure 2.4 Elevations defined with reference to a sphere, ellipsoid, geoid, or local sea level will all be different. Even locations as latitude and longitude will vary somewhat. When linking field data such as GPS with a GIS, the user must know what base to use.

Map Scale

- Map scale is based on the **representative fraction**, the **ratio** of a **distance on the map** to the **same distance on the ground**.
- Most maps in GIS fall between **1:1 million and 1:1000**.
- A GIS is **scaleless** because maps can be enlarged and reduced and plotted at many scales other than that of the original data.
- To **compare** maps in a GIS, **both** maps **<u>MUST</u>** be at the **same scale** and have the **same extent**.

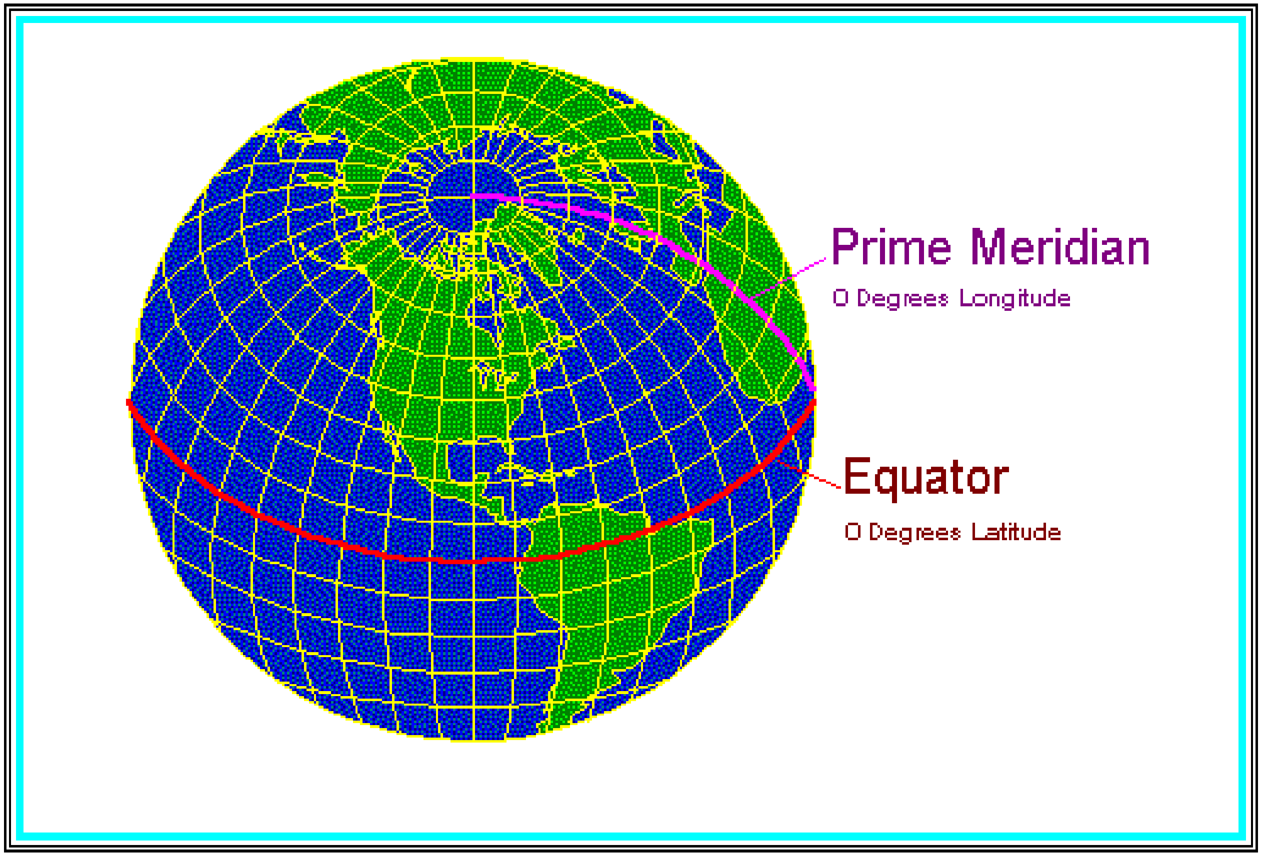
Geographic Coordinates

- Lines of latitude are called **parallels**
- Lines of longitude are called **meridians**



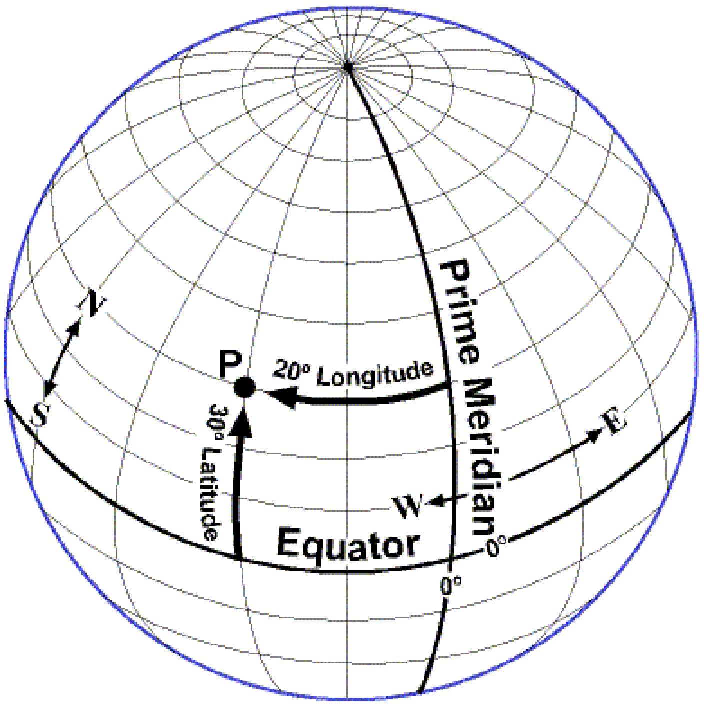
Geographic Coordinates

- The **Prime Meridian** and the **Equator** are the origin lines used to define latitude and longitude



Geographic Coordinates

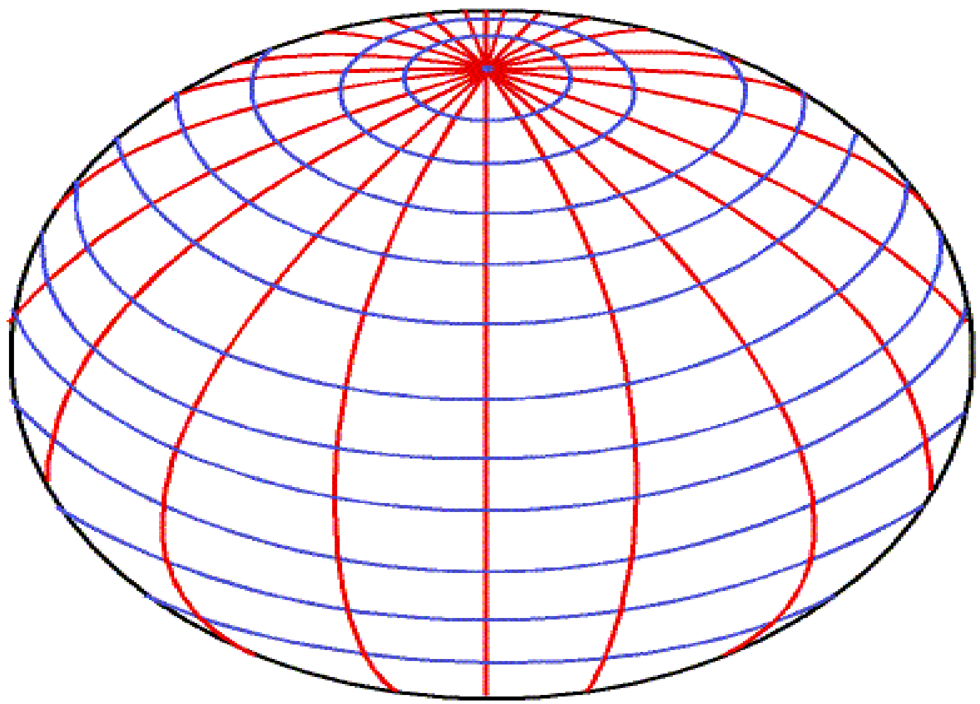
- Geographic coordinates are calculated using **angles**
- **Units** are in degrees, minutes, and seconds
- **Any location** on the planet can be specified with a **unique pair** of geographic coordinates



Tony Kirvan 11-8-97

Latitude & Longitude on an Ellipsoid

- On a sphere, lines of latitude (parallels) are an **equal distance** apart everywhere
- On an ellipsoid, the distance between parallels **increases slightly** as the latitude increases



Using Projections to Map the Earth

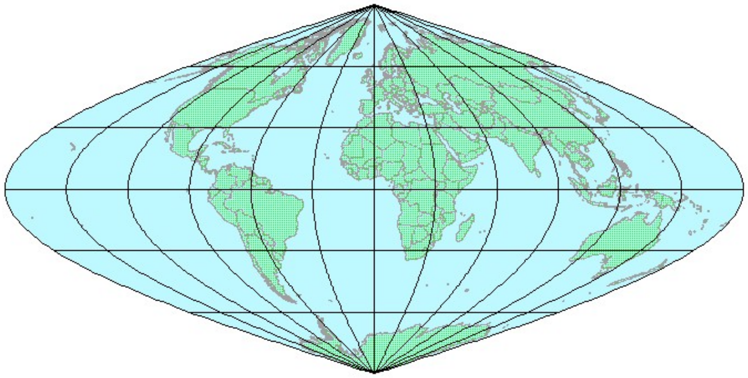
•We have discussed **geodesy**, and we now know about modeling the shape of Earth as an ellipsoid and geoid

•We are ready to tackle the problem of transforming the **3-dimensional Earth → 2-dimensional representation** that suits our purposes:

Earth surface

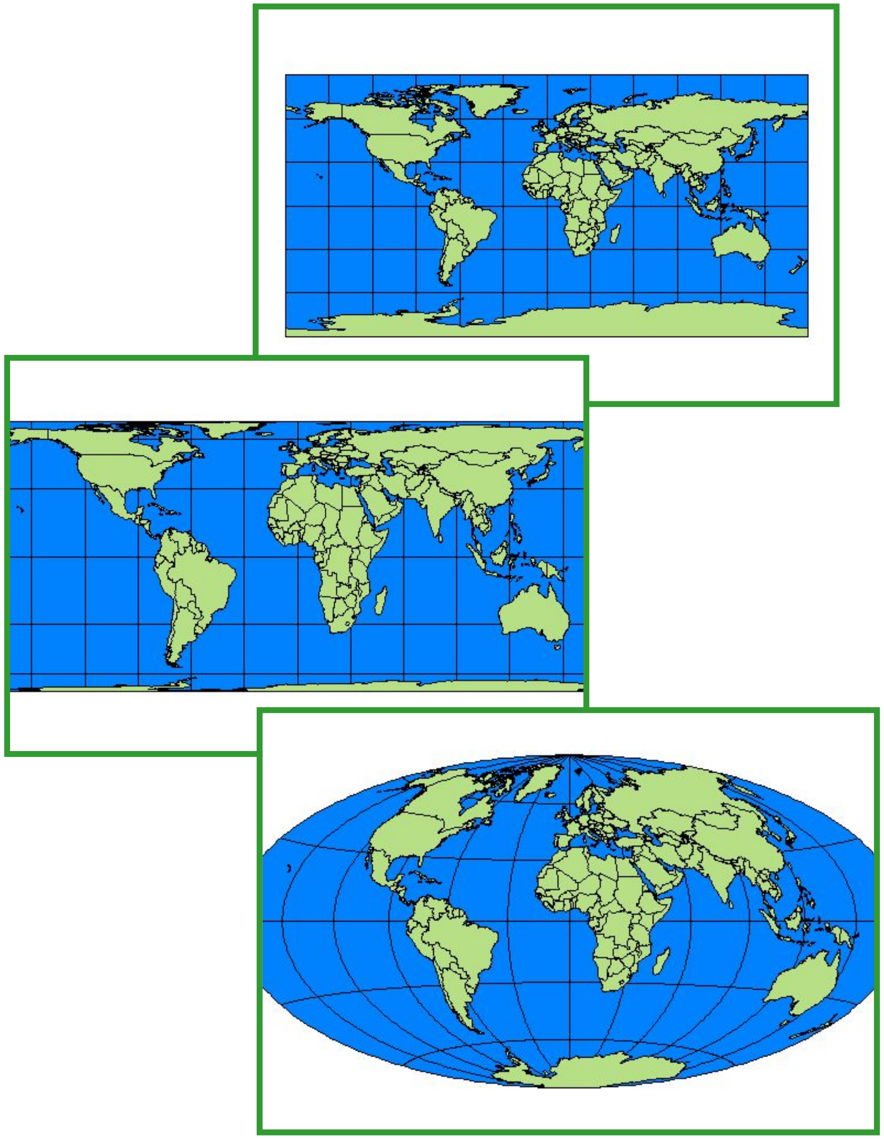
map

Paper map or GIS



Projections Distort

- Because we are going from the 3D Earth → 2D planar surface, projections **always** introduce some type of **distortion**
- When we select a map projection, we choose a particular projection to **minimize the distortions** that are important to a **particular application**



Three Families of Projections

- There are **three major families** of projections, each tends to introduce **certain kinds of distortions**, or conversely each has certain **properties** that it used to **preserve** (i.e. spatial characteristics that it does not distort):
- **Three families:**
    1. Cylindrical projections
    2. Conical projections
    3. Planar projections

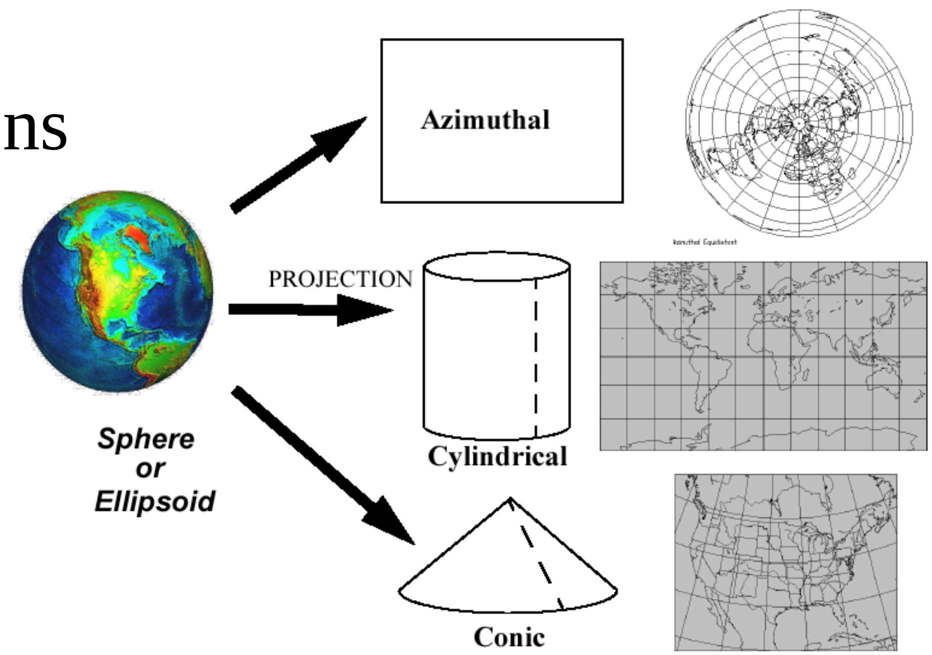
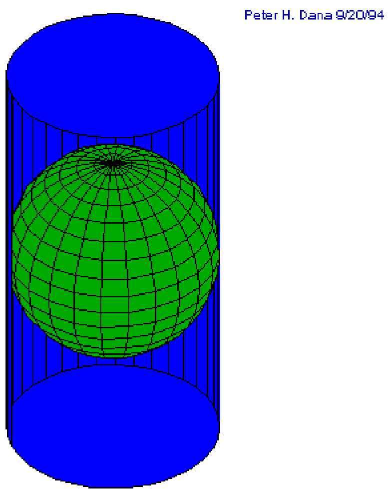
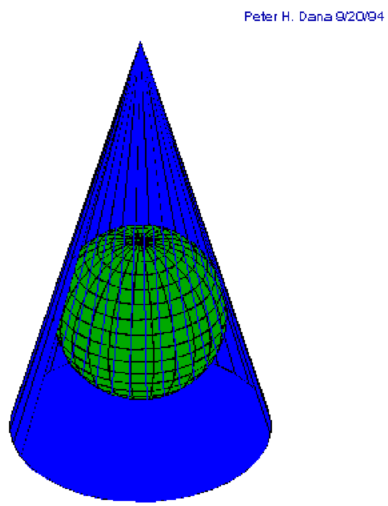


Figure 2.7 The earth can be projected in many ways, but basically onto three shapes that can be unrolled into a flat map: a flat plane, a cylinder, and a cone.

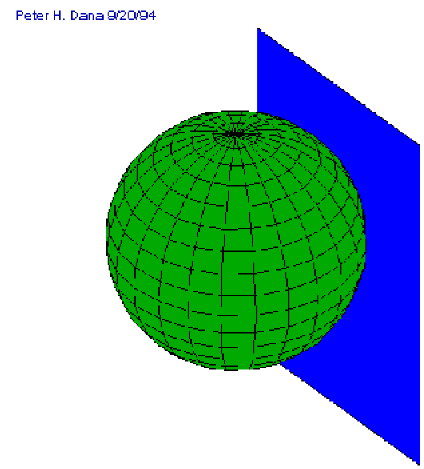
Tangent Projections



Cylindrical Projection Surface



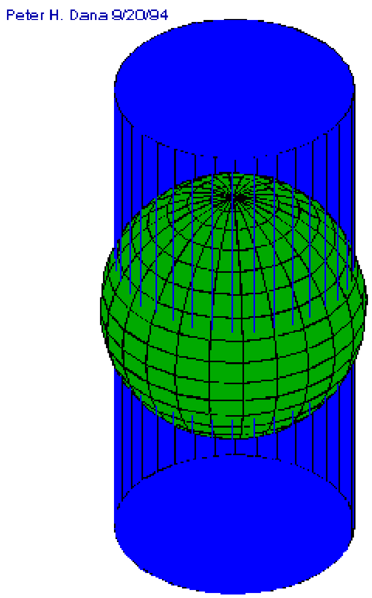
Conical Projection Surface



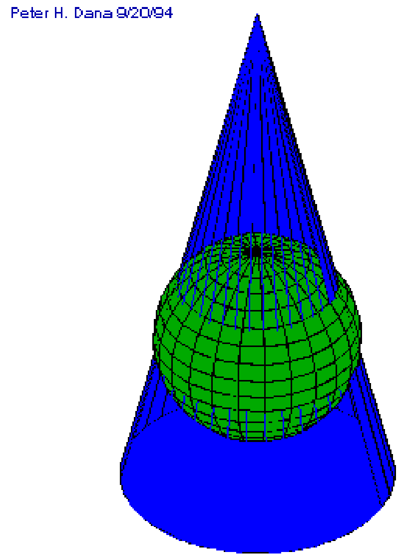
Planar Projection Surface

•Tangent projections have a **single standard point** (in the case of planar projection surfaces) or a **standard line** (for conical and cylindrical projection surfaces) of contact between the developable surface and globe

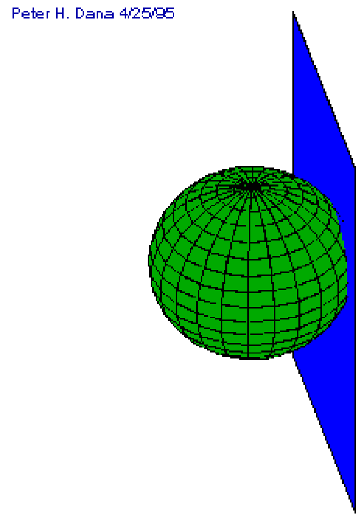
Secant Projections



Secant Cylindrical Projection



Secant Conic Projection



Secant Planar Projection

- Secant projections have a **single standard line** (in the case of planar projection surfaces) or **multiple standard lines** (for conical and cylindrical projection surfaces) of contact between the developable surface and the globe

Preservation of Properties

- Every map projection introduces some sort of **distortion** because there is always distortion when reducing our 3-dimensional reality to a 2-dimensional representation

- **Q:** How should we **choose** which projections to use?

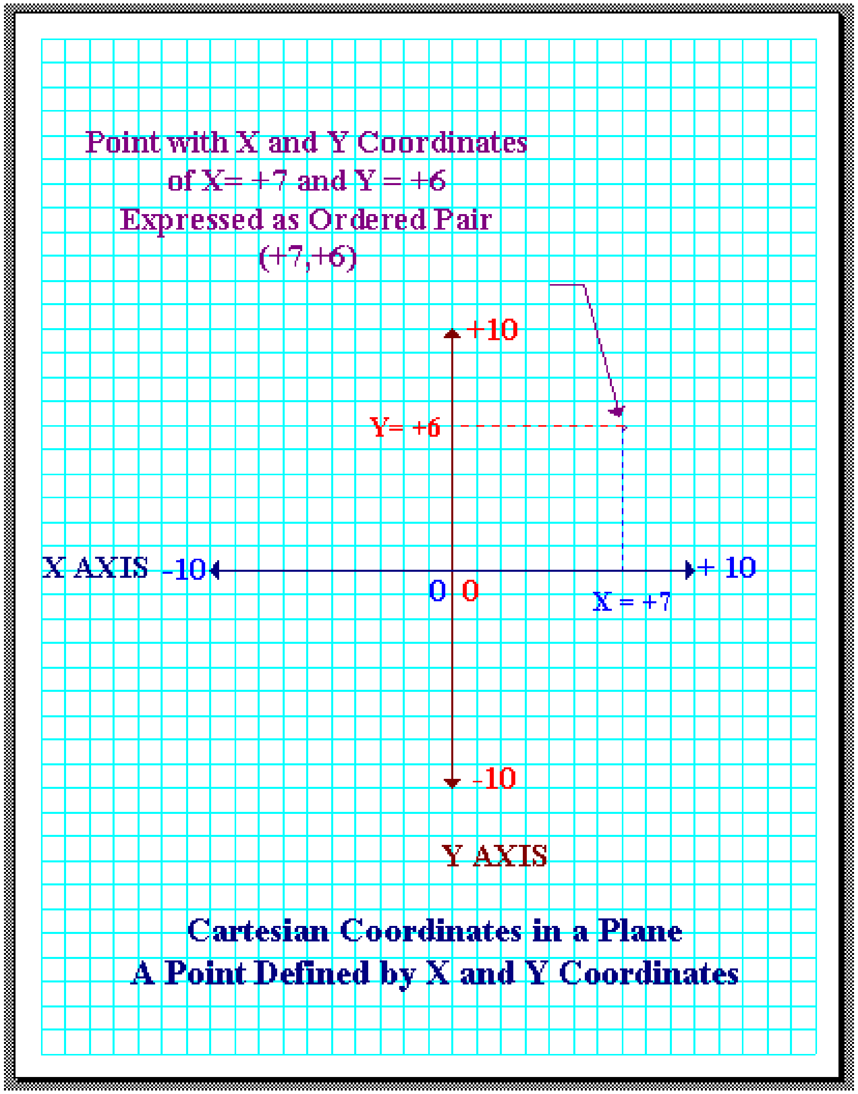
  A: We should choose a map projection that **preserves the properties** appropriate for the application, choosing from the following properties:

  1. Shape
  2. Area
  3. Distance
  4. Direction

Note: It may be more useful to classify map projections by the **properties they preserve**, rather than by the shape of their surfaces

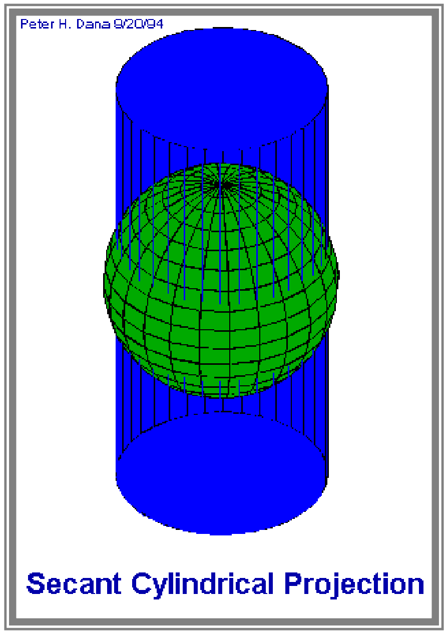
Planar Coordinate Systems

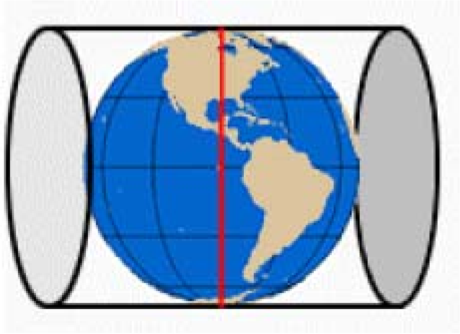
- Once we start working with **projected** spatial information, using latitude and longitude becomes **less convenient**
- We can instead use a **planar coordinate system** that has *x* and *y* axes, an arbitrary origin (a Cartesian plane), and some convenient units (e.g. ft. or m.)
- When applied in a geographic context:
  - **Eastings** are ***x*** values
  - **Northings** are ***y*** values



Universal Transverse Mercator

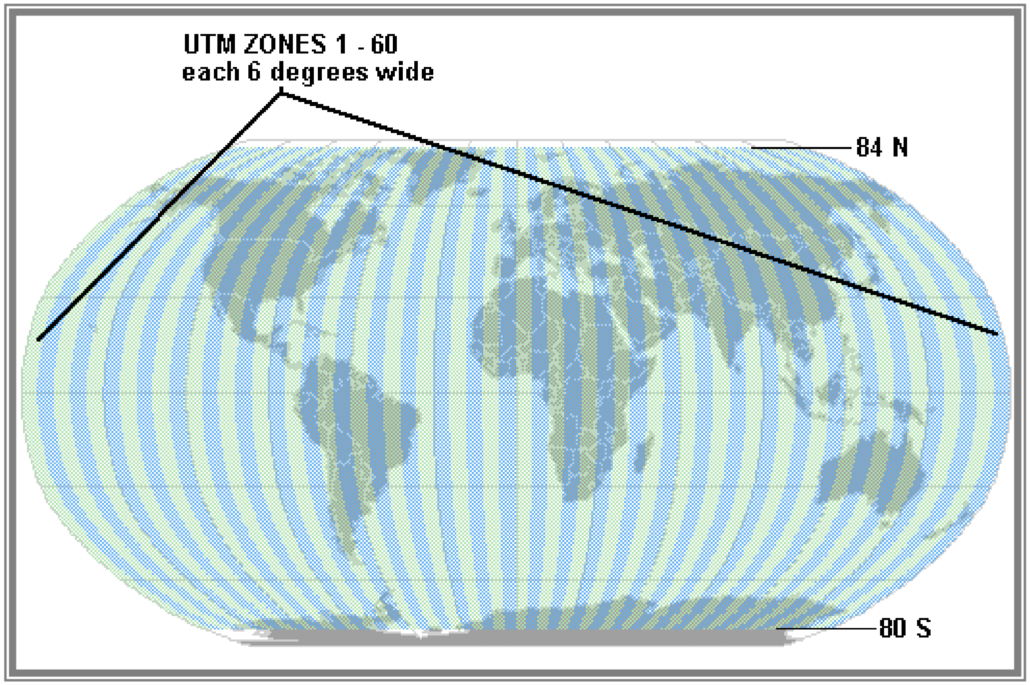
- Earlier, you were introduced to the Transverse Mercator **projection**
- That projection is used as the **basis** of the UTM **coordinate system**, which is widely used for topographical maps, satellite images, and many other uses
- The projection is based on a **secant transverse cylindrical projection**
- Recall that this projection uses a **transverse cylinder** that has **standard lines** that run **north-south**, and **distortion** increases as we move further **east or west**



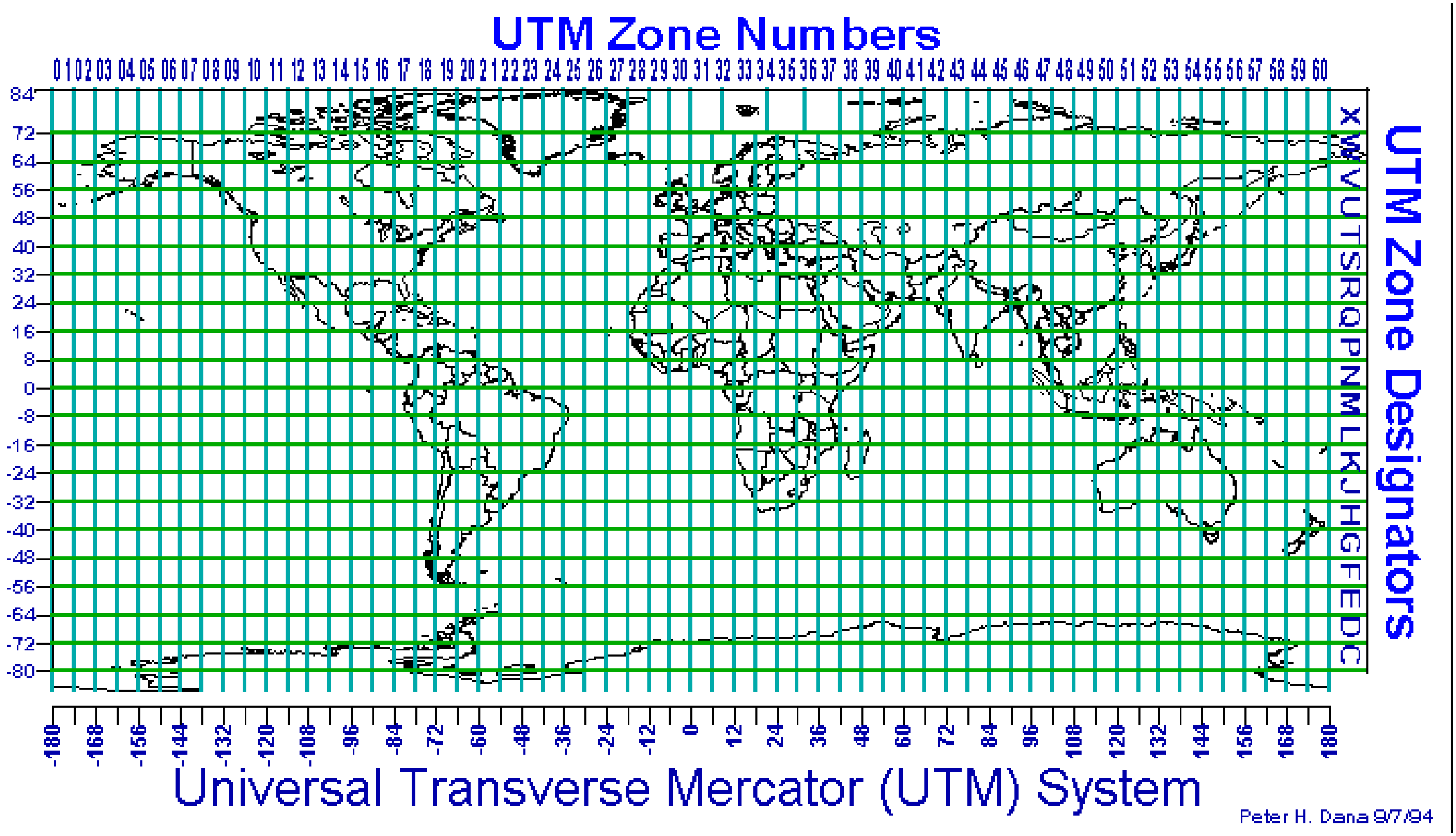


Universal Transverse Mercator

- In order to **minimize** the **distortion** associated with the projection, the UTM coordinate system uses a **separate Transverse Mercator projection** for every **6 degrees** of longitude → the world is divided into **60 zones**, each 6 degrees of longitude in width, each with its own UTM projection:

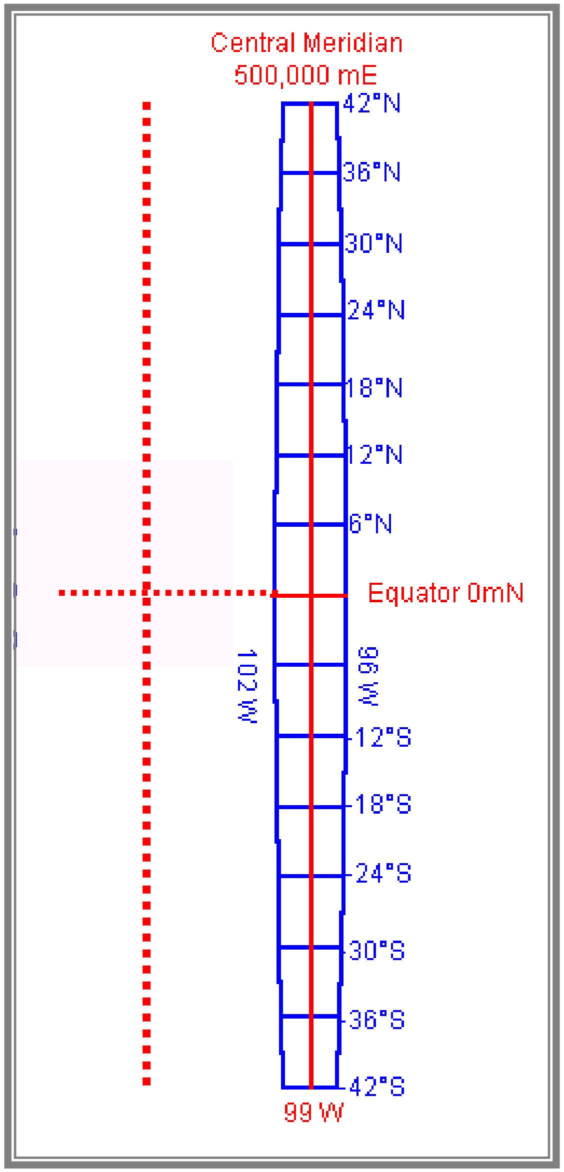


Universal Transverse Mercator



Universal Transverse Mercator

- The **central meridian**, which runs down the middle of the zone, is used to define the position of the origin
- **Distance units** in UTM are defined to be in **meters**, and distance from the origin is measured as an **Easting** (in the x-direction) and a **Northing** (in the y-direction)
- The x-origin is west of the zone (a false easting), and is placed such that the central meridian has an Easting of **500,000 meters**



Chapter 3: Maps as Numbers

- 3.1 Representing Maps as Numbers
- 3.2 Structuring Attributes
- 3.3 Structuring Maps
- 3.4 Why Topology Matters
- 3.5 Formats for GIS Data
- 3.6 Exchanging Data

Maps as Numbers

- GIS requires that both data and maps be **represented as numbers**.
- The GIS places data into the computer's memory in a **physical data structure** (i.e. files and directories).
- Files can be written in **binary** or as **ASCII text**.
- Binary is **faster to read and smaller**, ASCII can be **read by humans and edited** but uses more space.

Binary Notation

•Everything is represented as 0s and 1s in a computer. These two-state forms correspond to yes/no, on/off, open/closed

	Binary		Decimal	One to one correspondence	
				Decimal	Binary
1 digit	0, 1	1 bit	0,1,2,…9		
				0	0
2 digits	00, 01	2 bits	00, 01,…	1	1
	10, 11		97, 99	2	10
				3	11
3 digits	000, 001	3 bits	000, 001,	4	100
	010, 011		002, 003,	5	101
	100, 101		…	6	?
	110, 111		998, 999		

ASCII Encoding

•If computers store everything using 0s and 1s, then how are **characters** represented?

•The **ASCII** (American Standard Code for Information Interchange) code assigns the numbers 0 through 127 to 128 characters, including upper and lower case alphabets plus various special characters, such as white space etc.

•e.g. decimal 85 is assigned to represent upper case U. In binary, 01010101 = 85. Thus the computer represents U using 01010101.

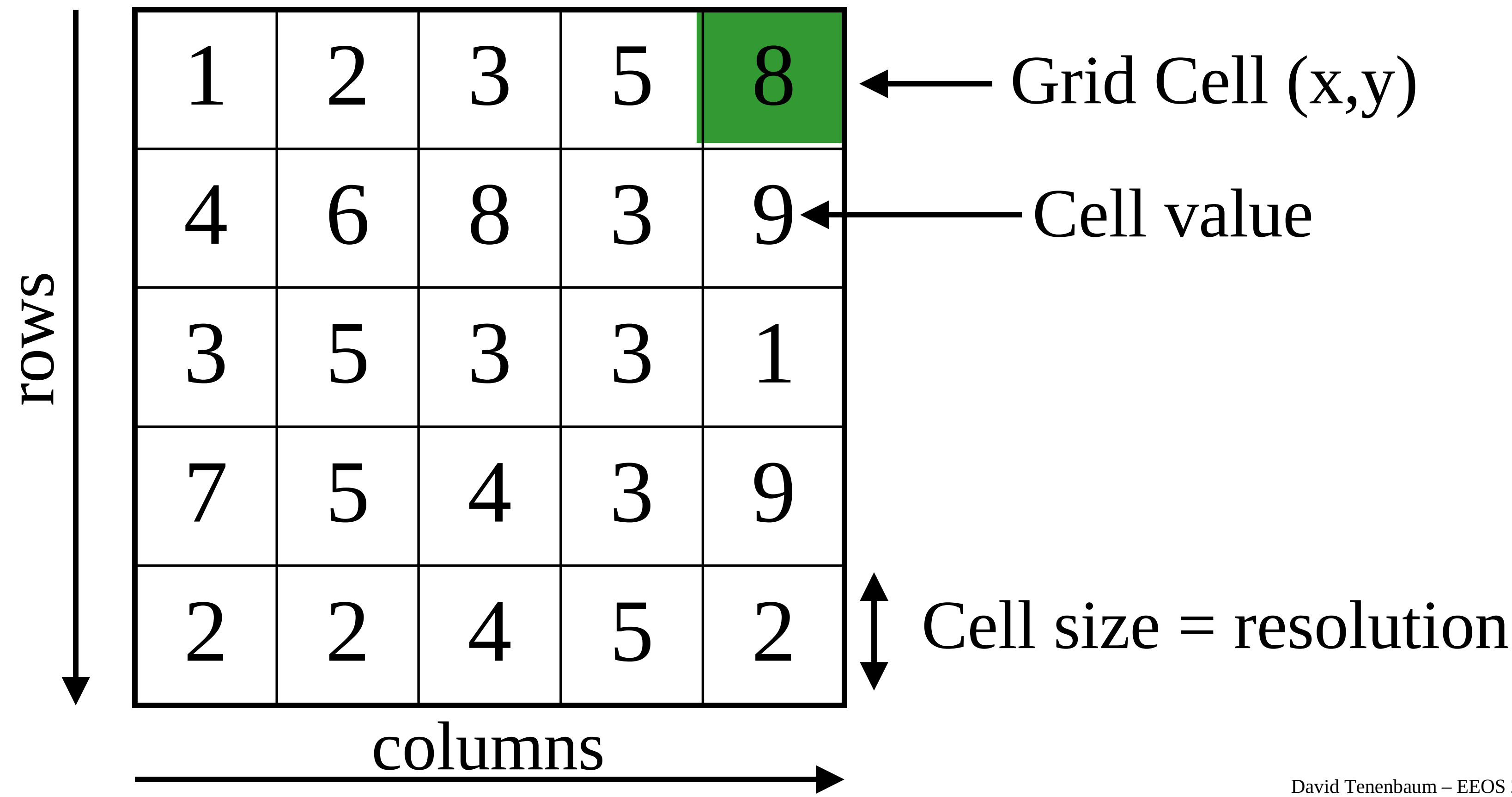
•Files which contain information encoded in ASCII are **easily transferred** and processed by different computers and programs. These are called “ASCII” or “text” files.

GIS Data Models

- A GIS map is a **scaled-down** digital representation of **point, line, area, and volume** features.
- A **logical data model** is how data are organized for use by the GIS.
- Traditionally there are **two GIS data models** used:
  - Raster
  - Vector

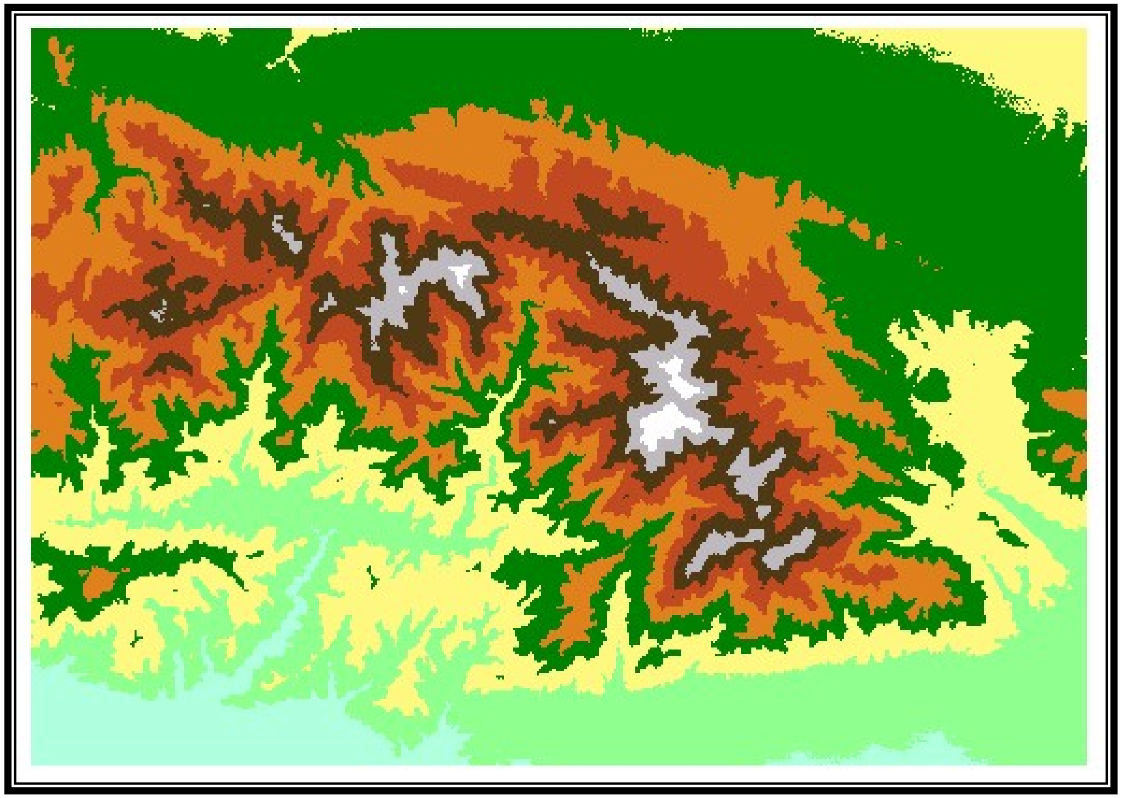
Generic structure for a grid

•The raster data model represents the Earth's surface as an **array** of two-dimensional grid cells, with each cell having an associated value:



Cells - Absolute Values

•In this instance, the **value** of the cell is actually the value of the phenomenon of interest, e.g. elevation data (whether floating point or integer):



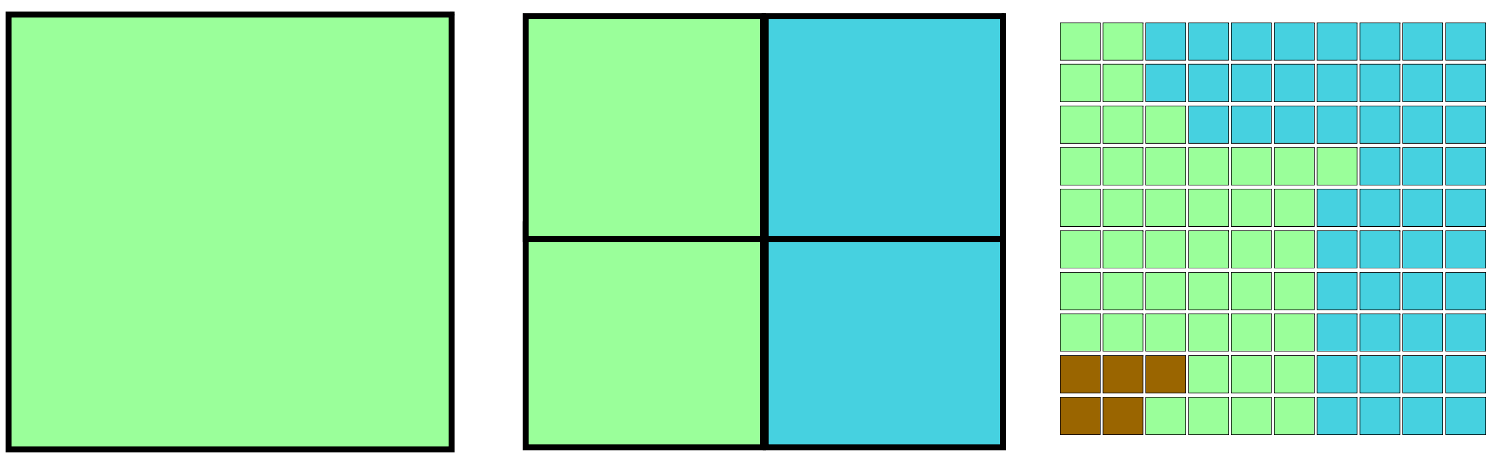
Cells - Coded Values

•Here, the values stored in each cell are used as **substitutes** for some **nominal** or **categorical** data, e.g. land cover classes:

ID	Land Cover Type	Ownership
1	Grass	Smith
…		
8	Water	Smith
13	Sand	Smith

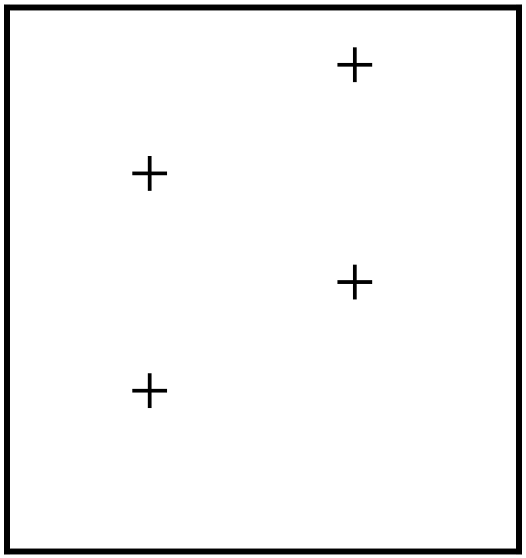
Cell Size & Resolution

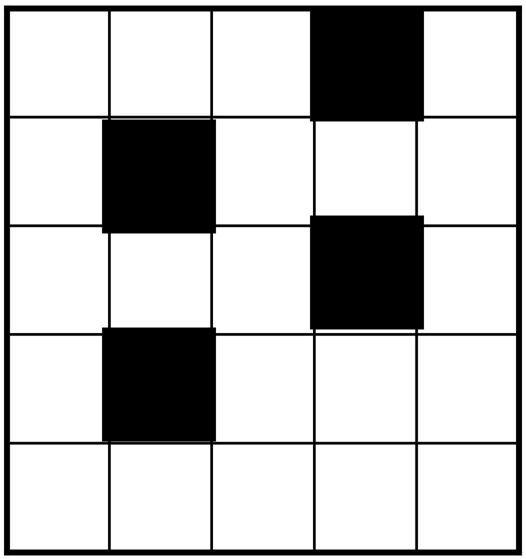
- The **size** of the cells in the raster data model determines the **resolution** at which features can be represented
- The selected **resolution** can have an **effect** on how features are represented:



10 m Resolution   5 m Resolution   1 m Resolution

Raster Data Model - Points

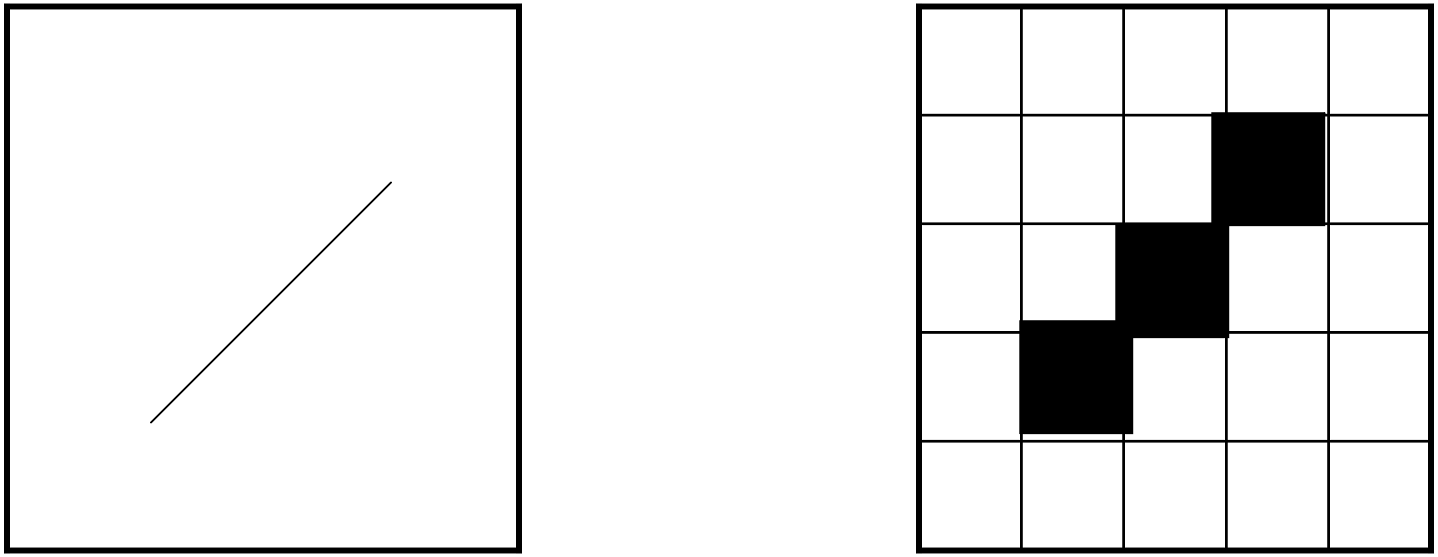




1 point = 1 cell

What problem do we have here? How can we solve it?

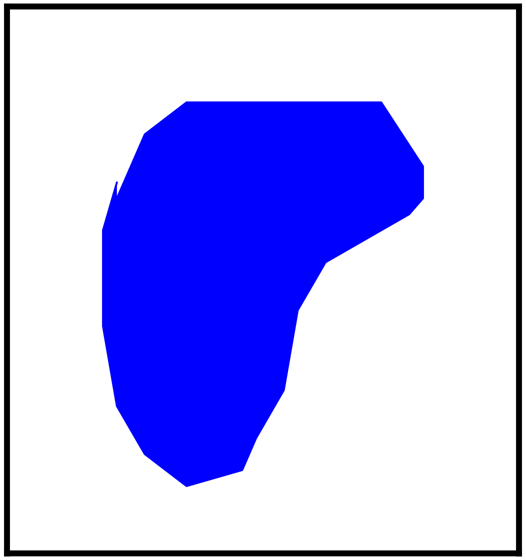
Raster Data Model - Lines

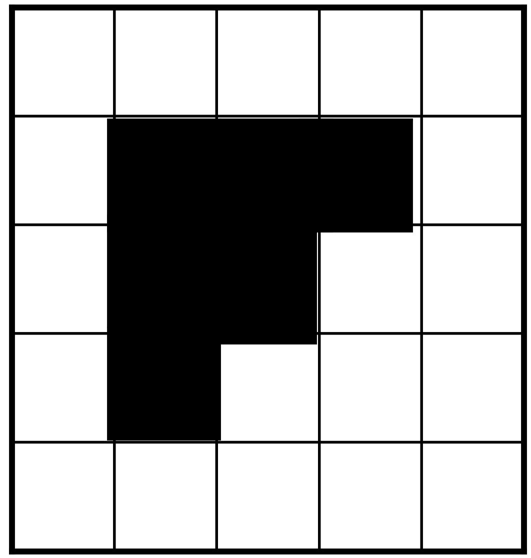


A line = a series of connected cells that portray length

Is there a problem with this representation?

Raster Data Model - Areas



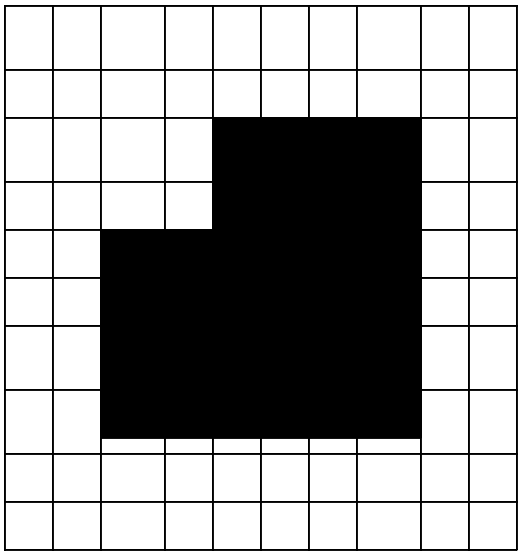


Area = a group of connected cells that portray a shape

What problems could we have with this representation?

Raster Data Storage – No Compaction

This approach represents each cell **individually** in the file:



0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	0	0
0	0	0	0	1	1	1	1	0	0
0	0	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

max. cell value

rows

columns

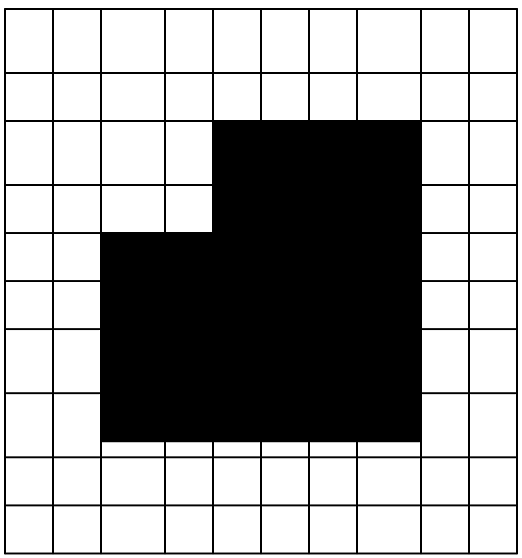
10, 10, 1
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 1 1 1 0 0
0 0 0 0 1 1 1 1 0 0
0 0 1 1 1 1 1 1 0 0
0 0 1 1 1 1 1 1 0 0
0 0 1 1 1 1 1 1 0 0
0 0 1 1 1 1 1 1 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0

103 values

Problem: too much **redundancy**

Raster Data Storage – Run Length Encoding

This approach takes advantage of **patterns** in the data, taking advantage of the **repetition** of values in a row:



0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	0	0
0	0	0	0	1	1	1	1	0	0
0	0	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

header → 10,10,1

row by row:

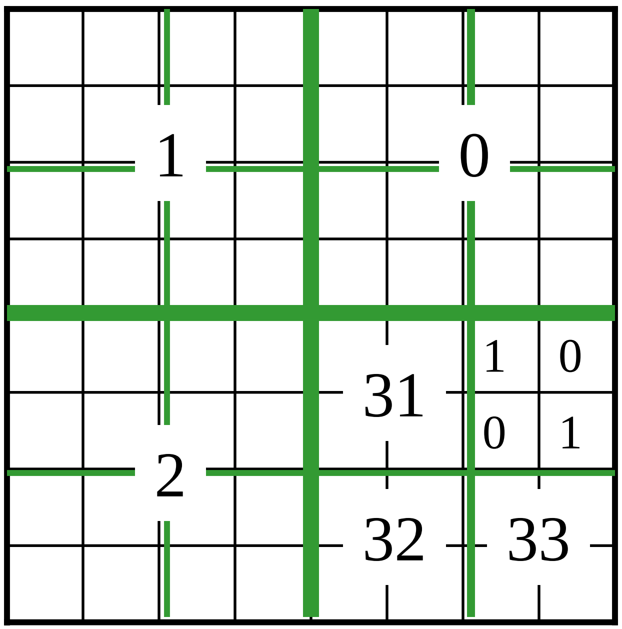
0, 10
0, 10
0, 4, 1, 4, 0,2
0, 4, 1, 4, 0,2
0, 2, 1, 6, 0,2
0, 2, 1, 6, 0,2
0, 2, 1, 6, 0,2
0, 2, 1, 6, 0,2
0, 10
0, 10

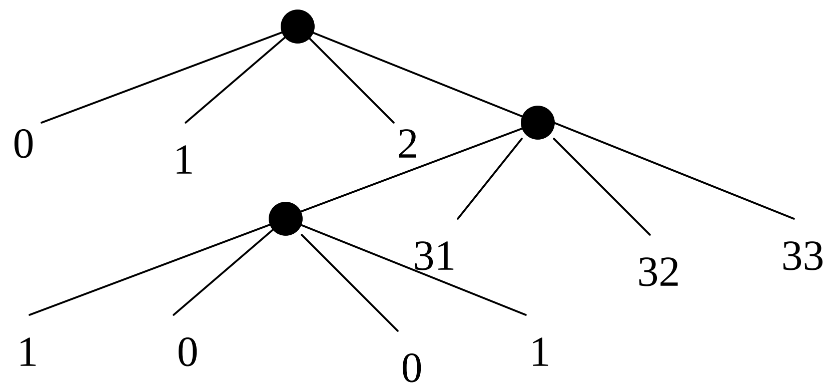
45 values

There is a tendency towards **spatial autocorrelation**; for **nearby cells** to have **similar values** - values often occur in runs across several cells

Raster Data Storage - Quadtrees

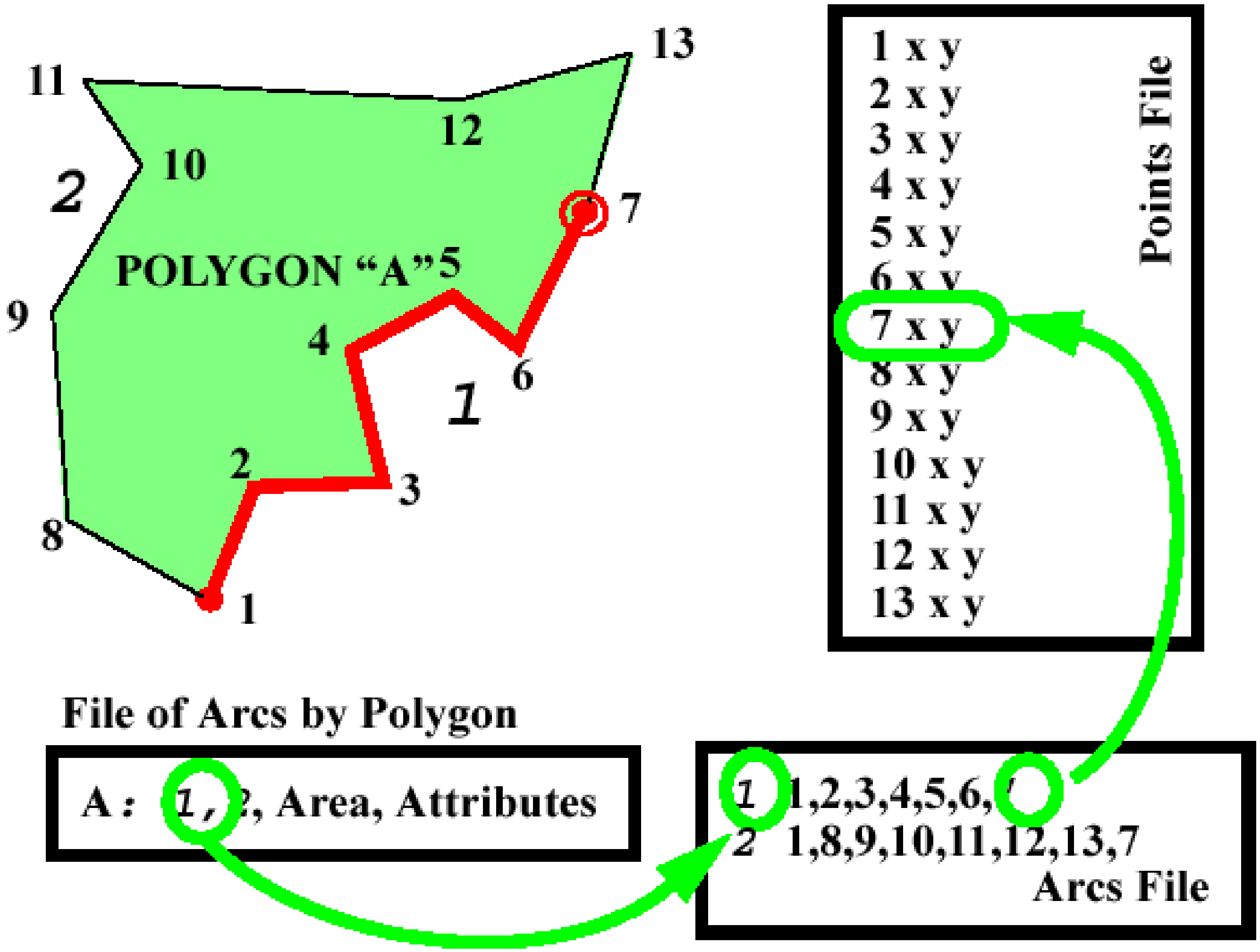
The quadtree method **recursively subdivides** the cells of a raster grid into quads (quarters) until **each quad** can be represented by a **unique cell value**:





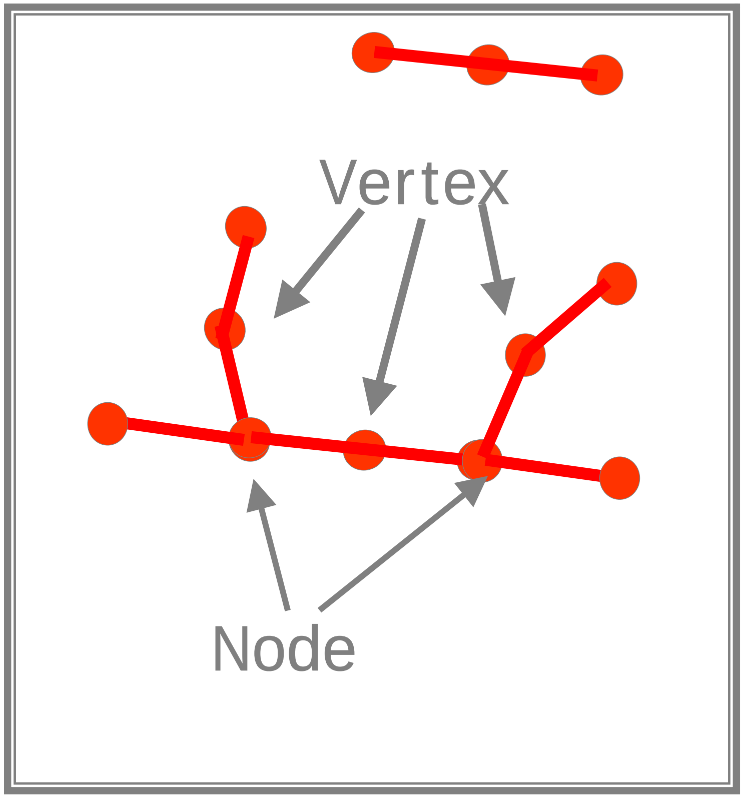
The number of subdivisions depends on the **complexity** of features and stores **more detail** in areas of greater complexity

Arc/node map data structure with files



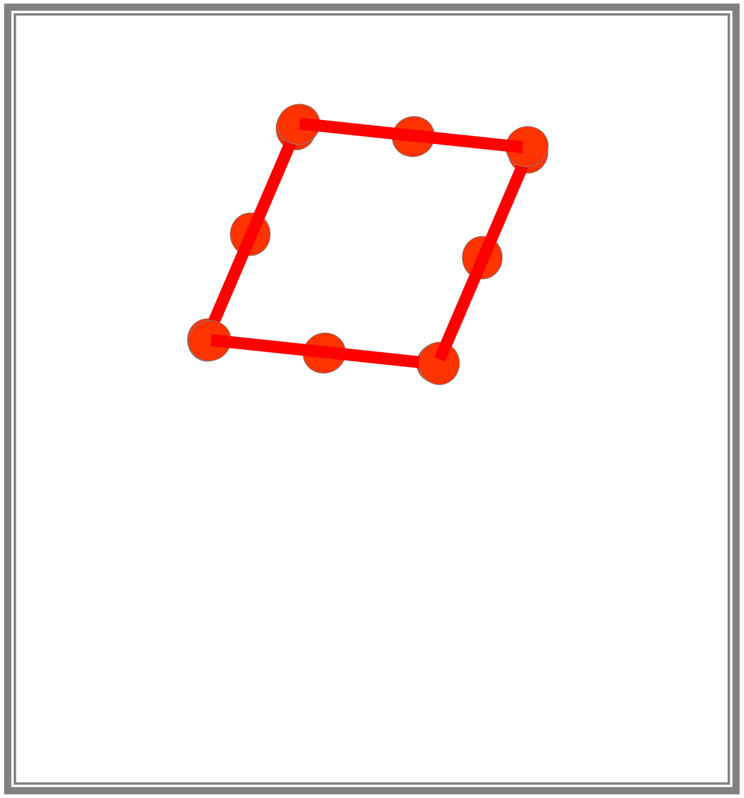
Vector Data Model - Objects

- Lines/Arcs
  - these are formed by **joining** multiple points
  - points at the junctions of lines are called **nodes**

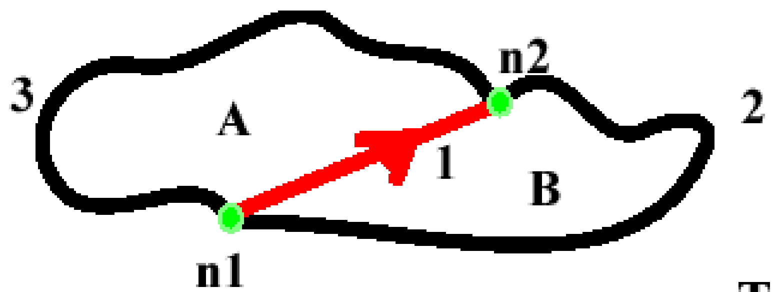


Vector Data Model - Objects

- Polygons
  - these are composed of **multiple lines or arcs**
  - They are required to have the property of **closure**, meaning that the multiple lines/arcs must form a **closed shape** for it to be a polygon



Basic arc topology



Topological Arcs File

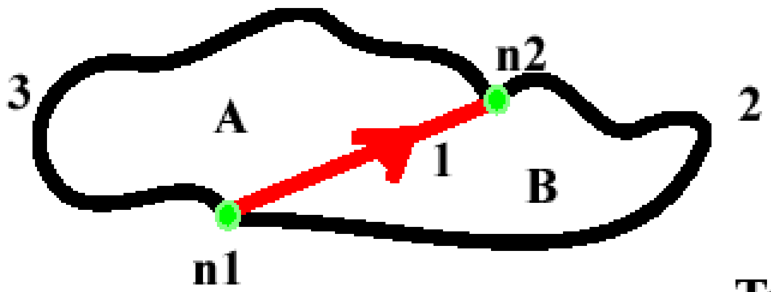
Arc	*From*	*To*	*PL*	*PR*	*n1x*	*n1y*	*n2x*	*n2y*
1	n1	n2	A	B	x	y	x	y

Vector Data Model - Topology

- Topology defines spatial relationships. The arc-node data structure supports the following topological concepts:
- **Area definition:** Arcs connect to surround an area, defining a polygon
- **Containment:** Nodes (or arcs) can be found within a polygon
- **Connectivity:** Arcs connect to each other at shared nodes
- **Contiguity:** Arcs have a defined direction, and left and right sides

Topology

- A **spatial data structure** used primarily to ensure that the associated data forms a **consistent and clean topological fabric**. For instance, the **arc-node topology**: Typically the arc is stored as the base unit, storing with it the polygon left and right, the forward and reverse arc linkages and the arc end nodes.



Topological Arcs File

Arc	*From*	*To*	*PL*	*PR*	*n1x*	*n1y*	*n2x*	*n2y*
1	n1	n2	A	B	x	y	x	y

Getting the Map into the Computer: GIS Data Development

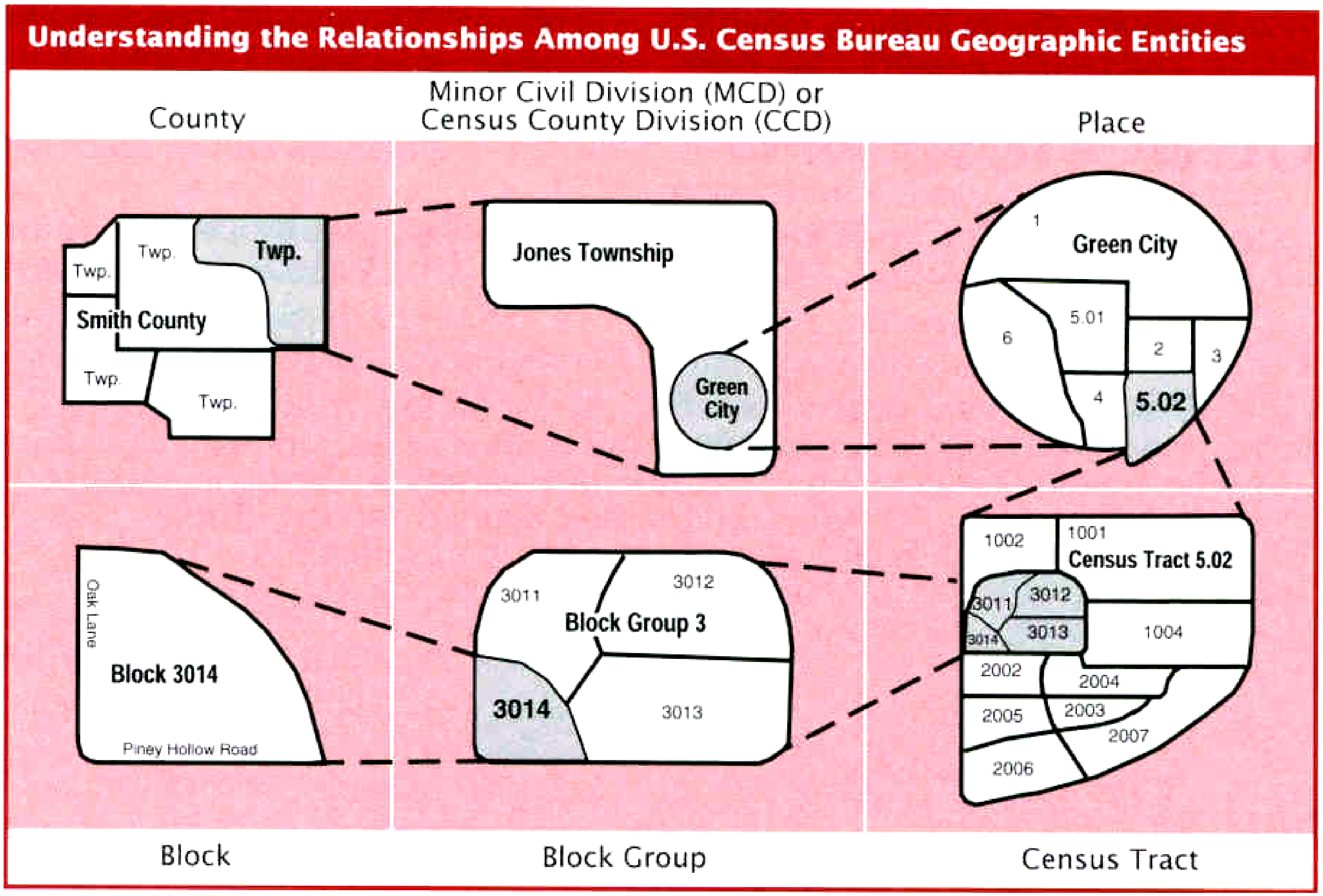
- 4.1 Analog-to-Digital Maps
- 4.2 Finding Existing Map Data
- 4.3 Digitizing and Scanning
- 4.4 Field and Image Data
- 4.5 Data Entry
- 4.6 Editing and Validation

GIS maps are digital not analog

- Maps have a **communications function** but...
- A map has a **storage function** for spatial data
- Somehow, the visually "stored" data must **get digital**
- **Real and Virtual** maps

Small-Area Geography Overview

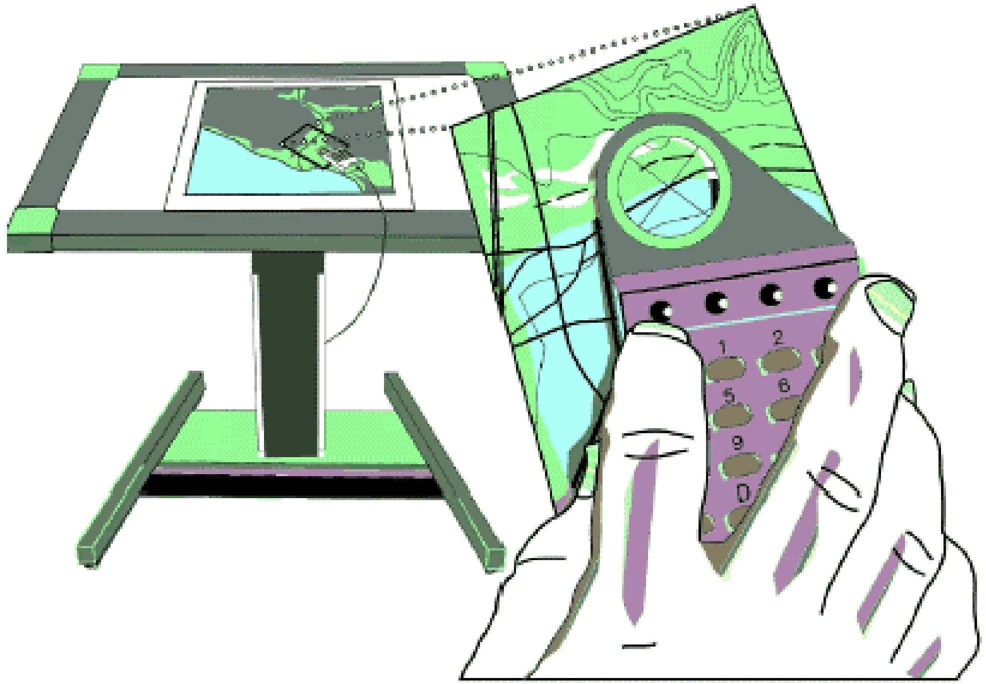
Census Small-Area Geography



GEOCODING

- Geocoding is the **conversion of spatial information** into **digital form**
- Geocoding involves **capturing the map**, and sometimes also **capturing the attributes**
- Often involves **address matching**

The Digitizing Tablet



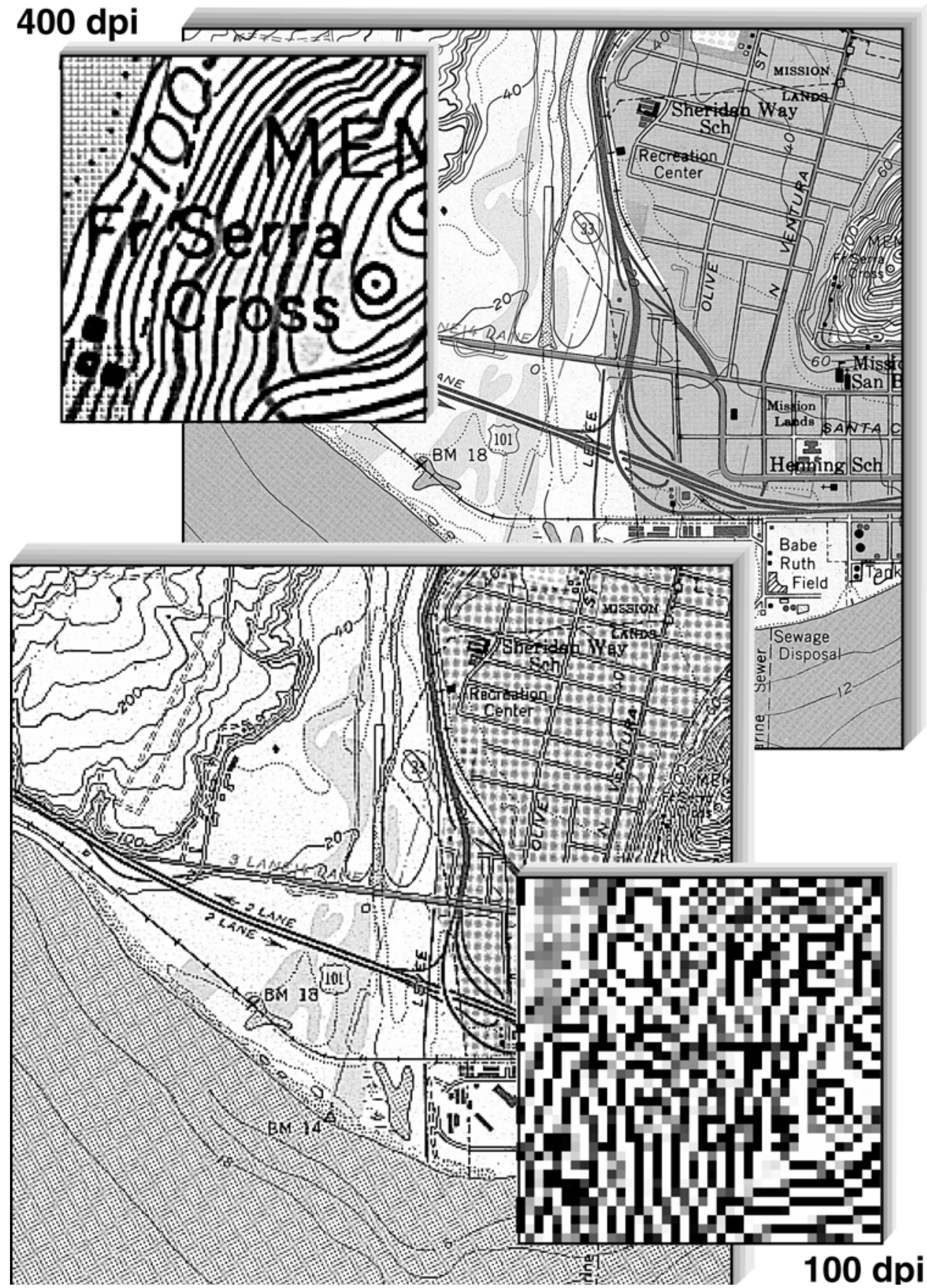
1. Digitizer cursor transmits a pulse from an electromagnetic coil under the view lens.
2. Pulse is picked up by nearest grid wires under tablet surface.
3. Result is sent to computer after conversion to x and y units.

Scanning

- Places a **map on a glass plate**, and passes a **light beam** over it
- Measures the **reflected light** intensity
- Result is a **grid of pixels**
- **Image size** and **resolution** are important
- Features can **"drop out"**

Scanning example

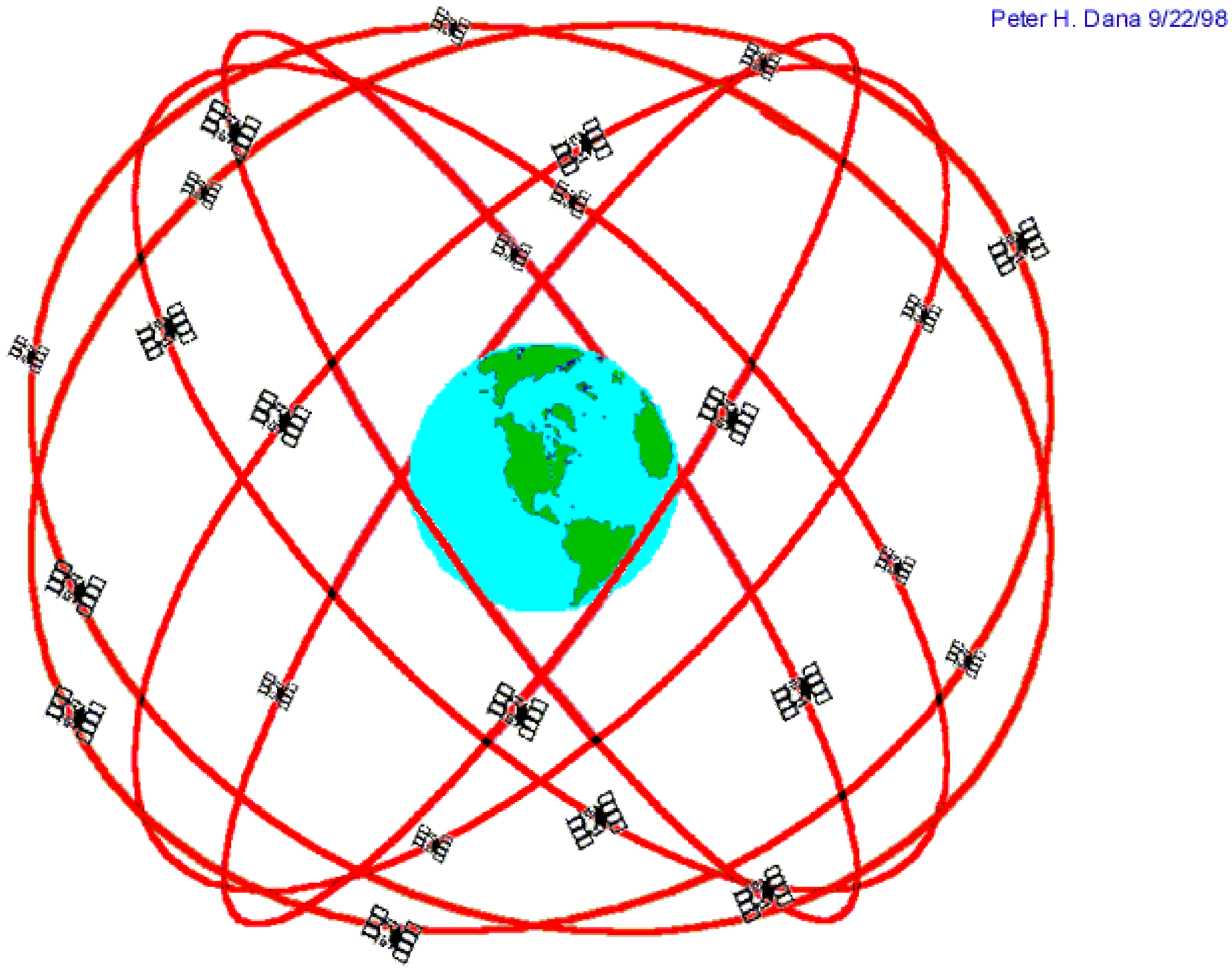
- 15 x 15 cm (3.6 x 3.6 km)
- grid is 0.25 mm
- ground equivalent is 6 m
- 600 x 600 pixels
- one byte per color (0-255)
- 1.08 MB



Field data collection

Global Positioning System (GPS)

- A space-based 3-dimensional **measurement and positioning system** that operates using radio signals from satellites orbiting the earth
- Created and maintained by the US Dept. of Defense and the US Air Force
- The system as a whole consists of three **segments**:
  - satellites (space segment)
  - receivers (user segment)
  - ground stations (control segment)
- Note: Russia and a European consortium are implementing similar systems.

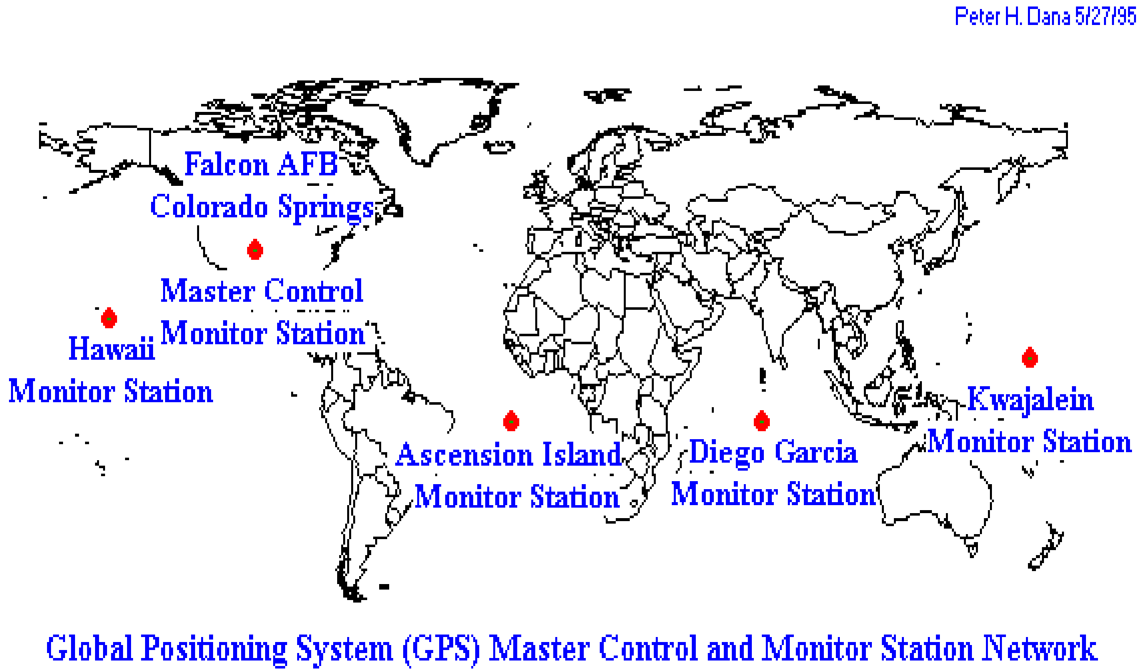


GPS Nominal Constellation
24 Satellites in 6 Orbital Planes
4 Satellites in each Plane
20,200 km Altitudes, 55 Degree Inclination

GPS – User Segment (Receivers)

- Ground-based devices that can read and **interpret the radio signals** from several of the NAVSTAR satellites at once
- Use **timing** of radio signals to calculate the receiver's **position** on the Earth's surface
- Calculations result in varying **degrees of accuracy** that depend on:
  - quality of the receiver
  - user operation of the receiver
  - local & atmospheric conditions
  - current status of system

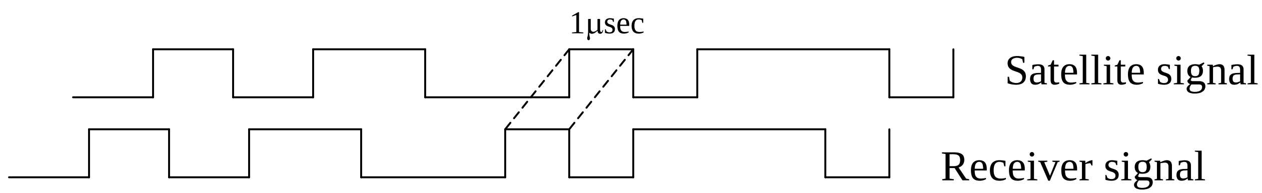
GPS – Control Segment (Ground Stations)



- Five control stations
  - master station at Falcon (Schriever) AFB, Colorado
  - **monitor** satellite orbits & clocks
  - **broadcast** orbital data and clock **corrections** to satellites

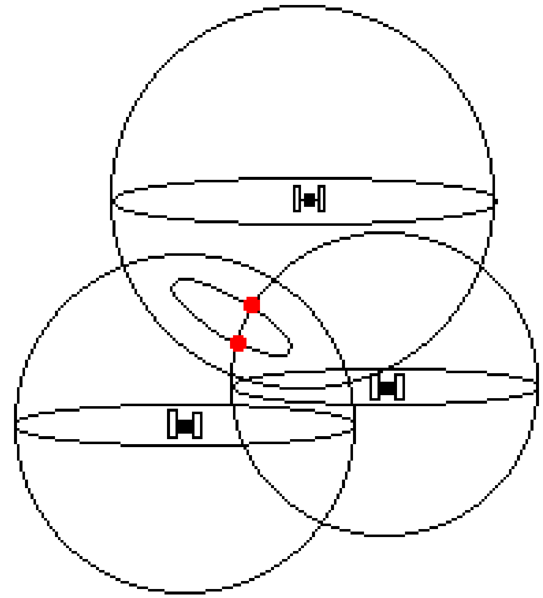
GPS - How Does it Work?

- GPS allows us to determine a position by calculating the **distance** between a receiver and multiple satellites
  - Distance is determined by **timing** how long it takes the signal to travel from satellite to receiver
  - Radio signals travel at **speed of light**: 186,000 mi / sec
  - Satellites and receivers generate **exactly the same signal** at exactly the same time
  - Signal travel time = delay of satellite signal relative to the receiver signal

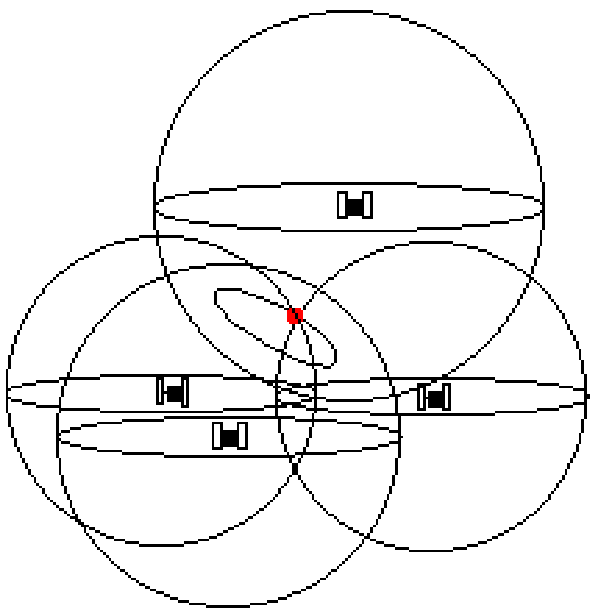


GPS - Trilateration Cont.

Adding a third satellite narrows down the position to **two points** where the three sphere intersect, and usually only one point is a **‘reasonable’** answer



3 satellites = 2 points

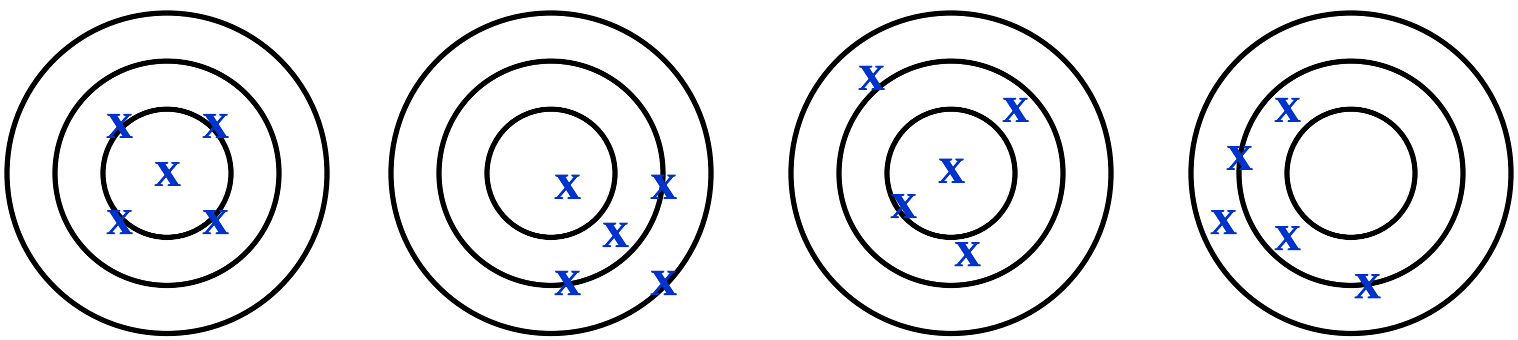


4 satellites = 1 point

The intersection of four spheres occurs at **one point**, but the 4th measurement is not needed, and is used for **timing** purposes instead

Precision and Accuracy

- These related concepts are often confused:
  - **Precision** refers to the exactness associated with a measurement (i.e. closely clustered)
  - **Accuracy** refers to the extent of systematic bias in the measurement process (i.e. centered on the middle)



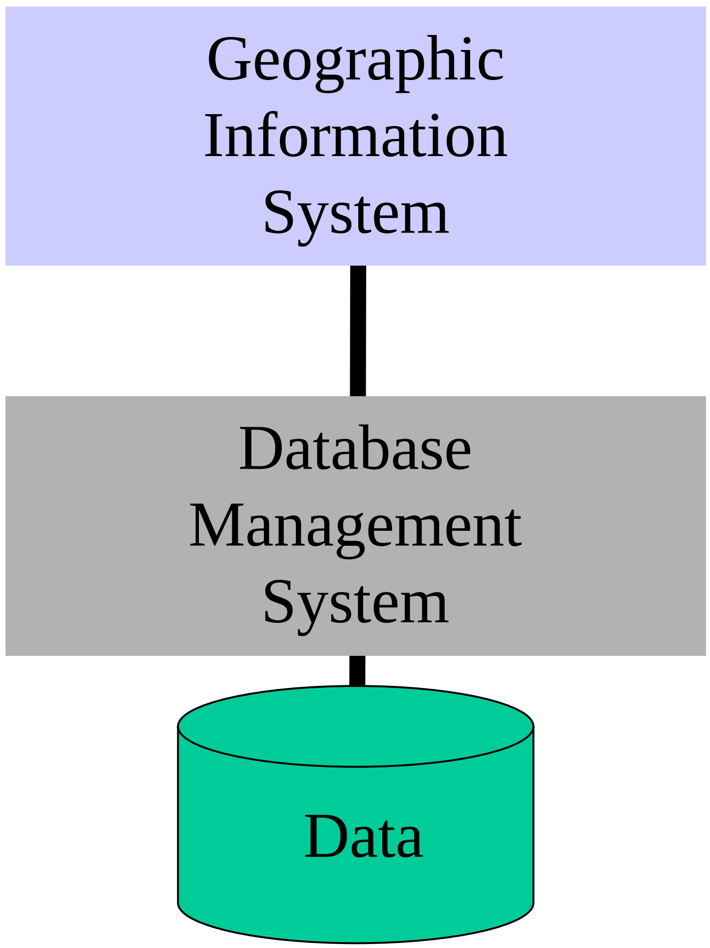
Precise & Accurate | Precise & Inaccurate | Imprecise & Accurate | Imprecise & Inaccurate

Chapter 5: What is Where?

- 5.1 Basic Database Management
- 5.2 Searches By Attribute
- 5.3 Searches By Geography
- 5.4 The Query Interface

The Role of DBMS in GIS

System



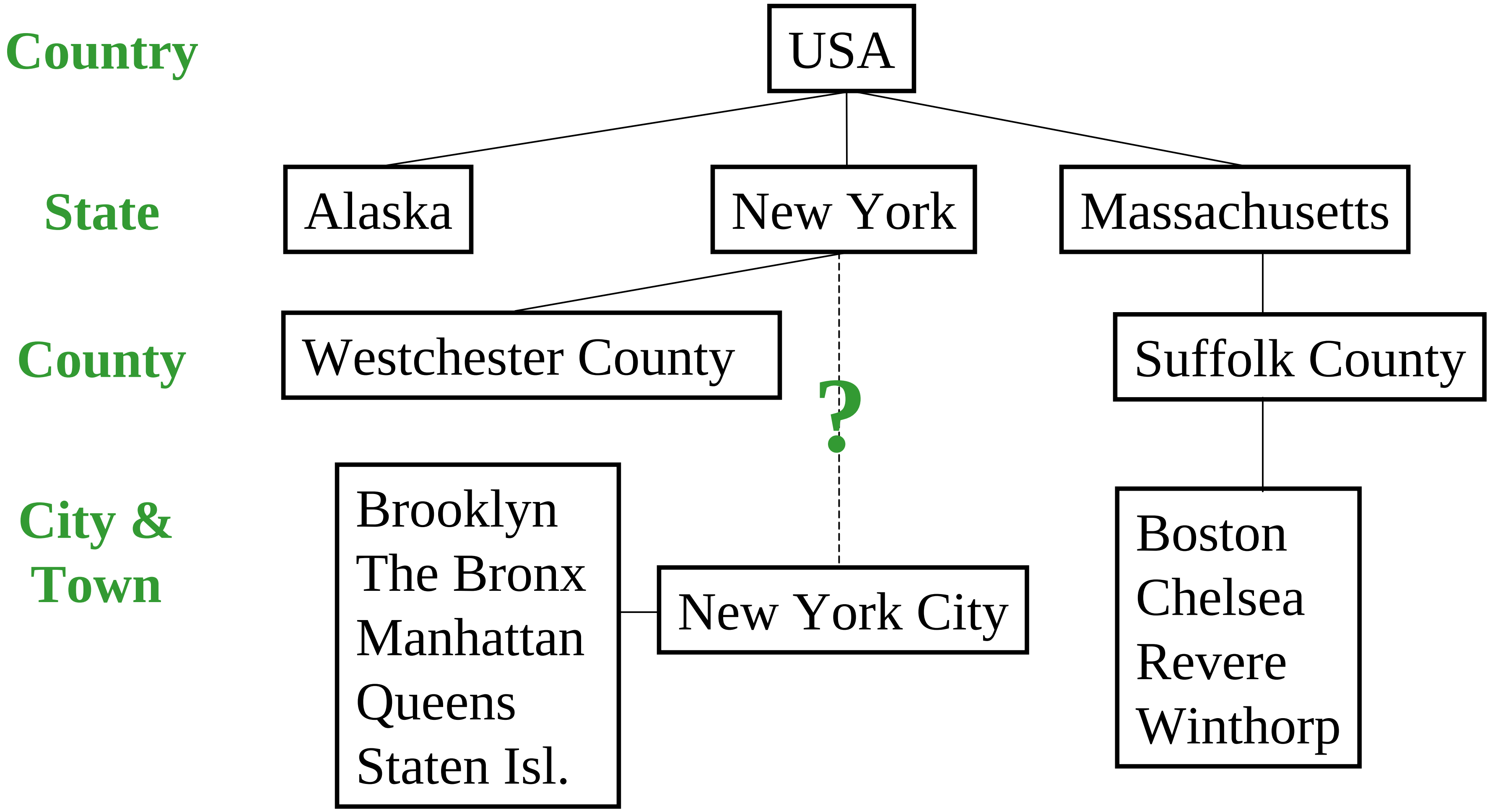
Task

- **Data load**
- **Editing**
- **Visualization**
- **Mapping**
- **Analysis**

- **Storage**
- **Indexing**
- **Security**
- **Query**

Hierarchical Data Model

- Now, suppose we are creating a model for places in the USA



Types of DBMS Models

- Hierarchical
- Network
- Relational - RDBMS ← **Our focus**
- Object-oriented - OODBMS
- Object-relational - ORDBMS

Most current GIS DBM is by relational databases.

- Based on **multiple flat files** for records
- Connected by a **common key attribute**.
- Key is a **UNIQUE** identifier at the "atomic" level for every record (Primary Key)

Relational Data Model

The **relational model** organizes data in a series of two-dimensional tables, each of which contains records for one kind of entity

Fields →

records ↓

PID #	Name	Major	Phone #	…
1010789	John	EEOS	555-4321	…
1021384	David	Comm.	555-6789	…

This model is a **revolution** in database management → It replaced almost all other approaches in database management because it allows more **flexible relations** between kinds of entities

Relation Rules (Codd, 1970)

- Only one value in **each cell** (intersection of row and column)
- All values in a column are about the **same subject**
- Each row is **unique**
- No significance in **column** sequence
- No significance in **row** sequence

Relational Join

- Take two tables full of different data (e.g. registrar info & parking data), and **join** them:

PID #	Name	Major	Phone #	…
1010789	John D.	EEOS	555-4321	…
1021384	David Q.	Comm.	555-6789	…

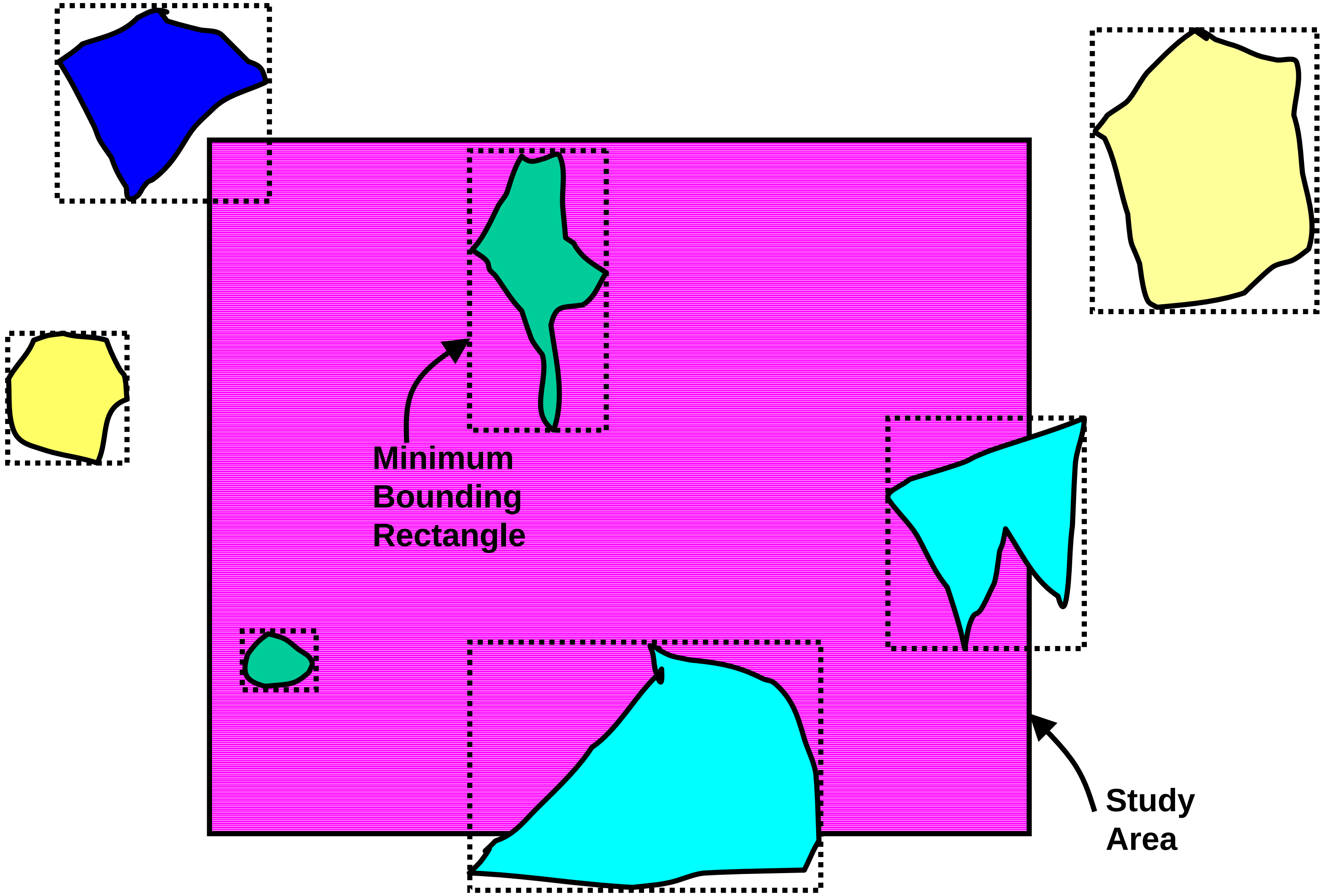
PID #	Name	Parking Lot	License Plate	…
1010789	John D.	North Lot	PNT3465	…
1021384	David Q.	Lot 150	JRS4089	…

The tables are joined through a **common key** which has a unique value for each record

Indexing

- Used to locate rows **quickly**, speed up access
- RDBMS use **simple 1-d indexing**
- Spatial DBMS need **2-d, hierarchical indexing** to allow features in a given vicinity to be found quickly, using a variety of methods:
  - Grid
  - Quadtree
  - R-tree
  - Others
- Hierarchical in the sense that **multi-level queries** are often used for better performance

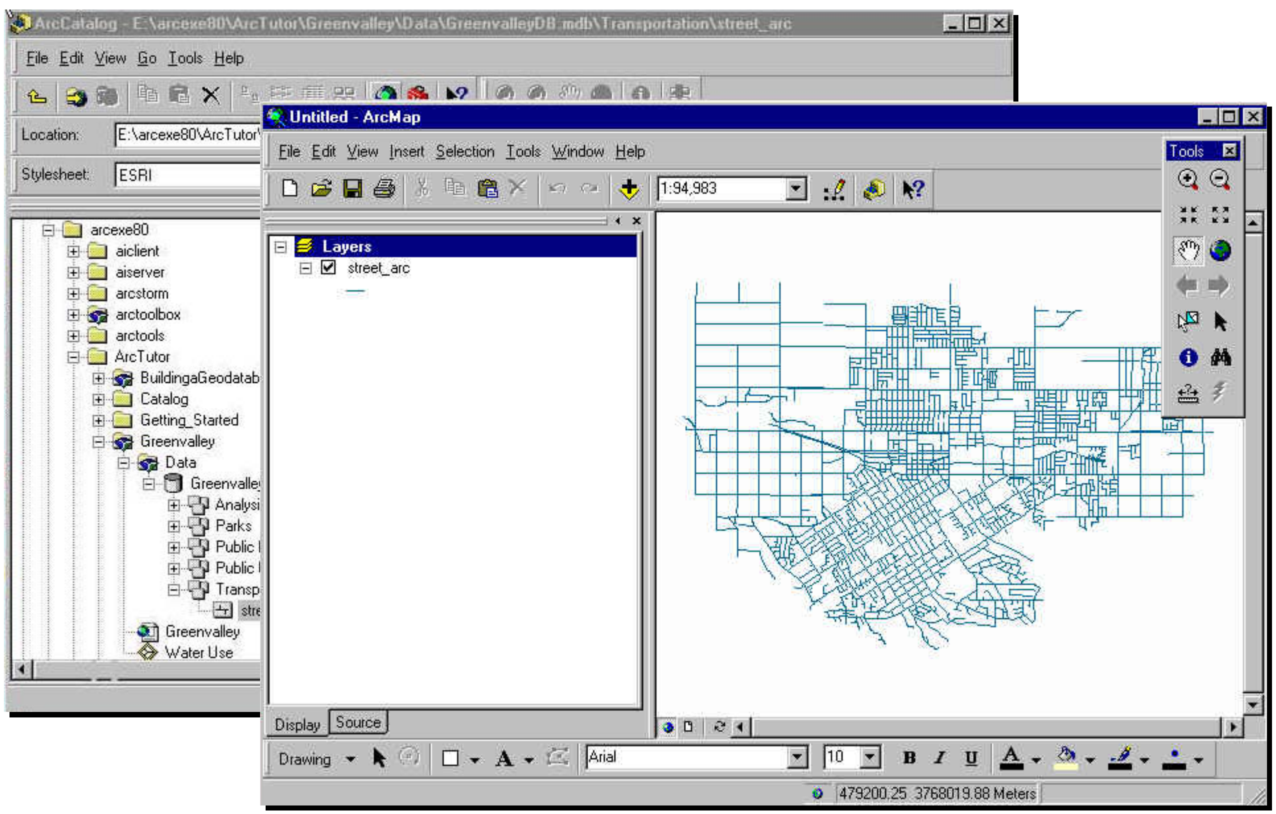
Minimum Bounding Rectangle



The Retrieval User Interface

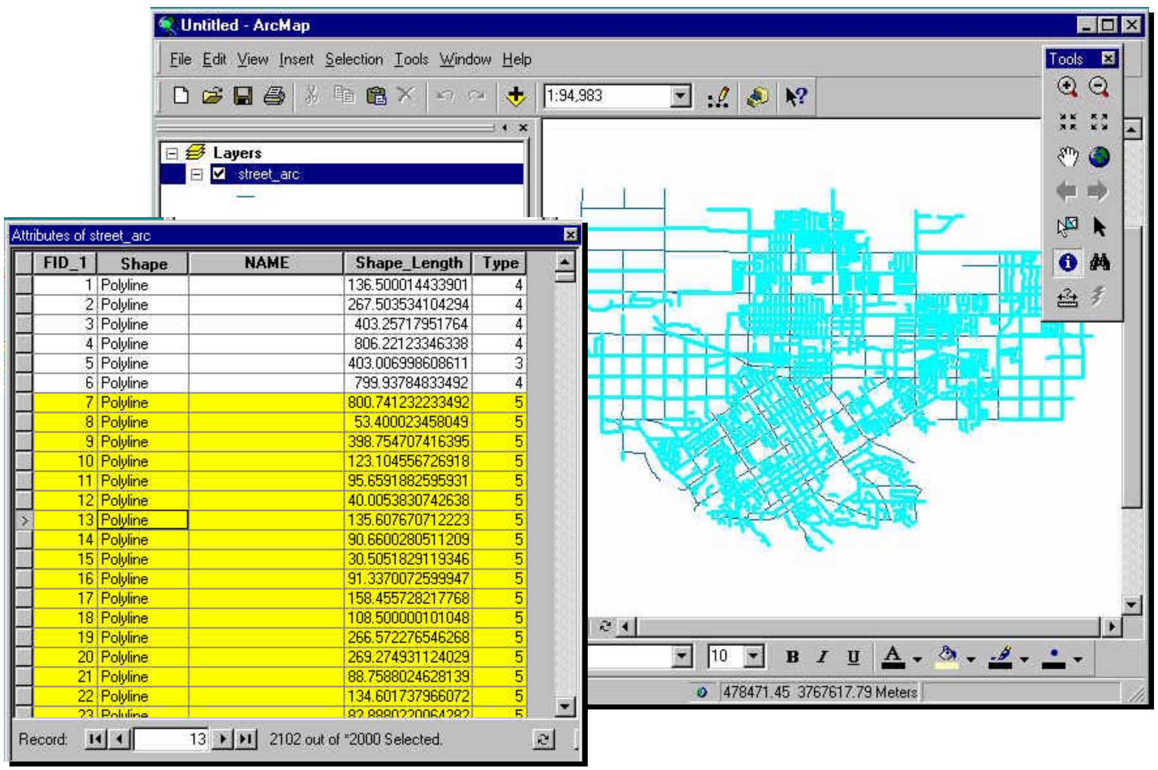
- **GIS query** is usually by command line, batch, menu (GUI) or macro.
- Most GIS packages use the GUI of the computer's operating system to support both a **menu-type query interface** and a **macro or programming language**.
- **SQL** is a standard interface to relational databases and is supported by many GISs.

The Map View (~ Data View)



•A user can interact with a **map view** to identify objects and query their attributes, to search for objects meeting specified criteria, or to find the coordinates of objects. This illustration uses ESRI’s ArcMap.

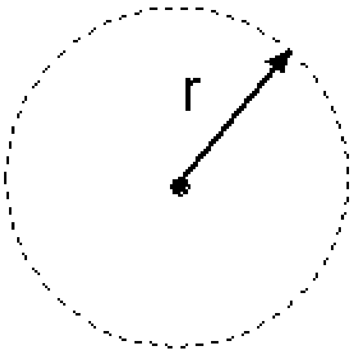
The Table View (~ a Table)

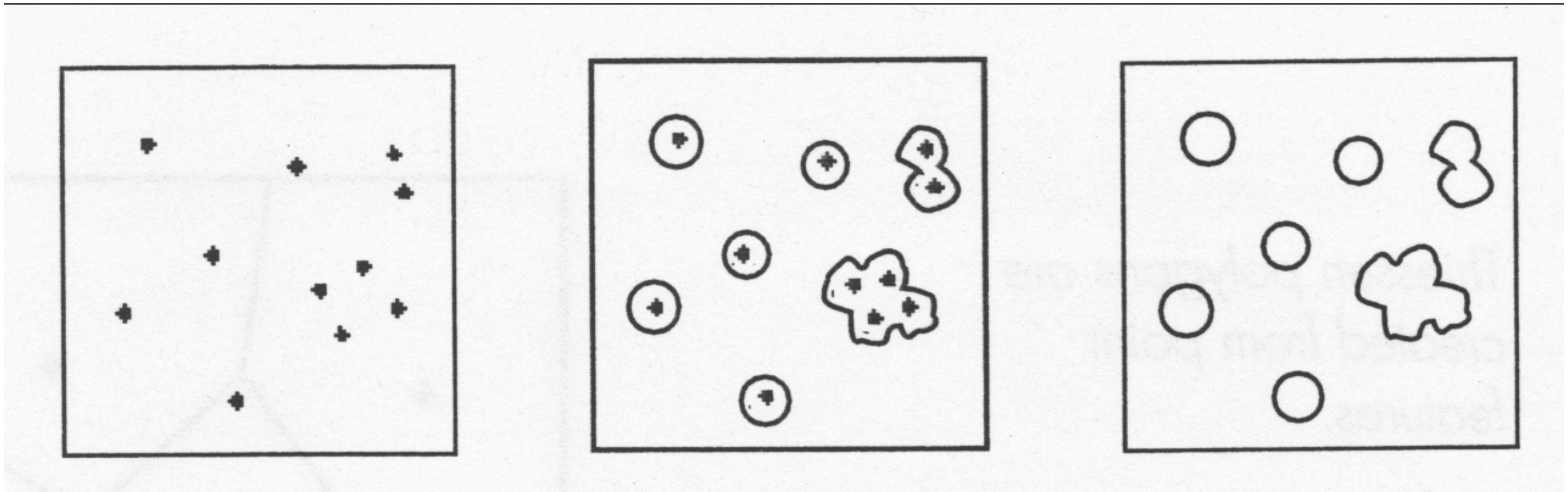


•Here attributes are displayed in the form of **a table**, linked to a map view. When objects are selected in the table, they are automatically highlighted in the map view, and vice versa. The table view can be used to answer simple queries about objects and their attributes.

Buffering (Proximity Analysis)

Buffering: The delineation of a zone around the feature of interest **within a given distance**. For a point feature, it is simply a circle with its radius equal to the buffer distance:

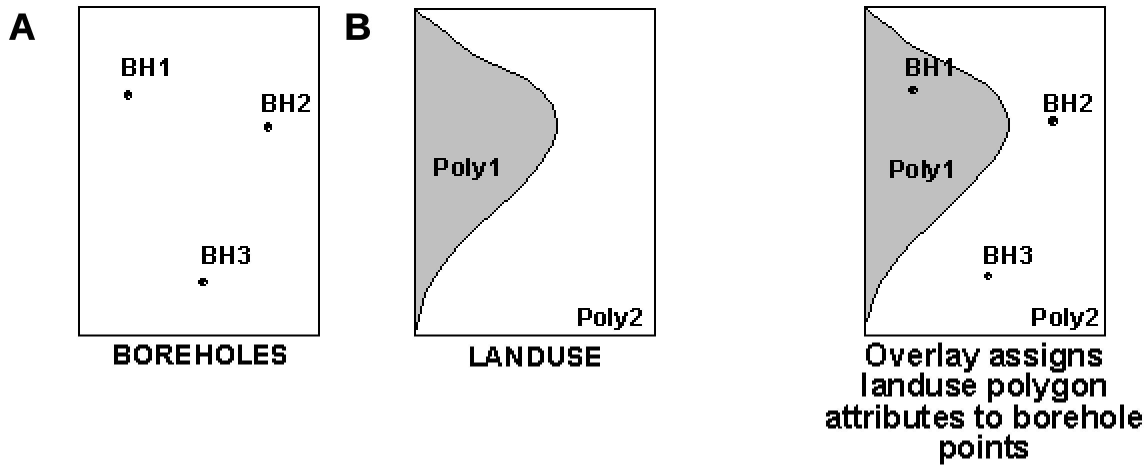




Point in Polygon Analysis

- Overlay point layer (A) with polygon layer (B)
  - **In which** B polygon are A points located?
    - » **Assign polygon attributes** from B to points in A

Example: Comparing soil mineral content at sample borehole locations (points) with land use (polygons)...



Point	Zn	Pb
BH1	140	65
BH2	178	54
BH3	101	87

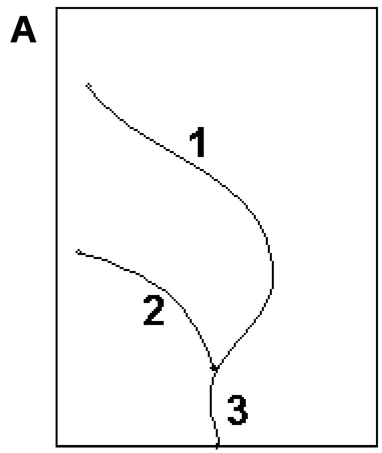
Poly	Landuse
1	Agriculture
2	Urban

Point	Zn	Pb	Landuse
BH1	140	65	Agriculture
BH2	178	54	Urban
BH3	101	87	Urban

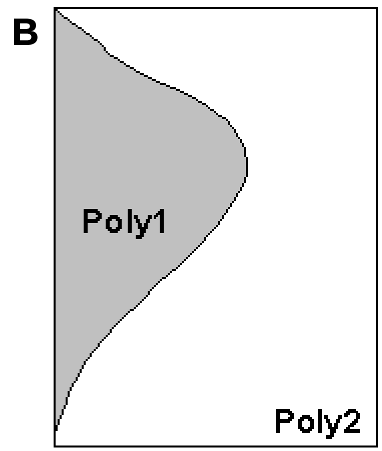
Line in Polygon Analysis

- Overlay line layer (A) with polygon layer (B)
  - **In which** B polygons are A lines located?
    - » **Assign polygon attributes** from B to lines in A

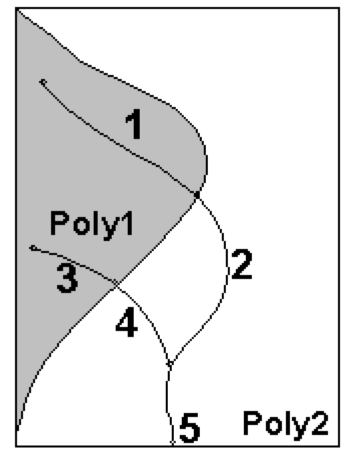
Example: Assign land use attributes (polygons) to streams (lines):



STREAMS



LANDUSE



Overlay splits stream line entities and assigns landuse poly attributes to streams

Line	Length
1	780
2	520
3	225

Poly	Landuse
1	Agriculture
2	Urban

Line	Length	Landuse
1	440	Agriculture
2	340	Urban
3	220	Agriculture
4	300	Urban
5	225	Urban

Polygon Overlay Analysis

UNION

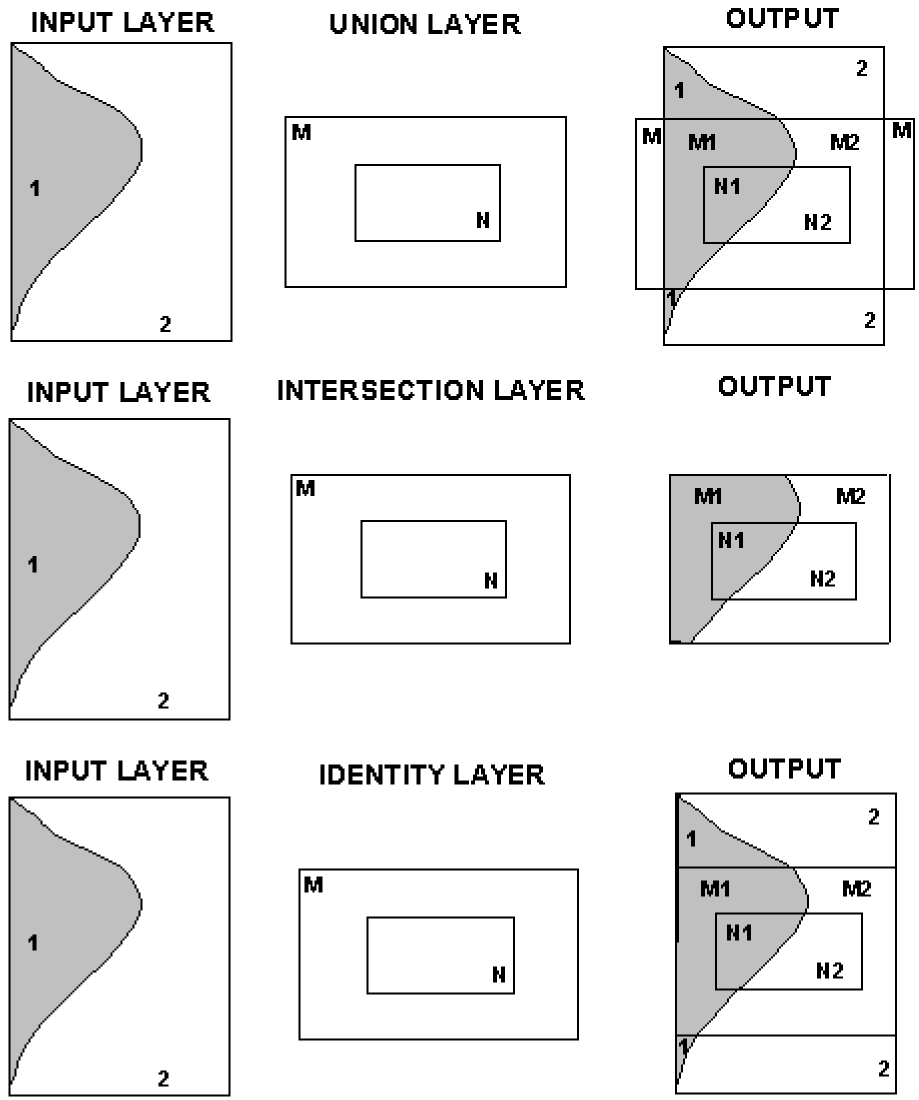
- overlay polygons and **keep areas from both layers**

INTERSECTION

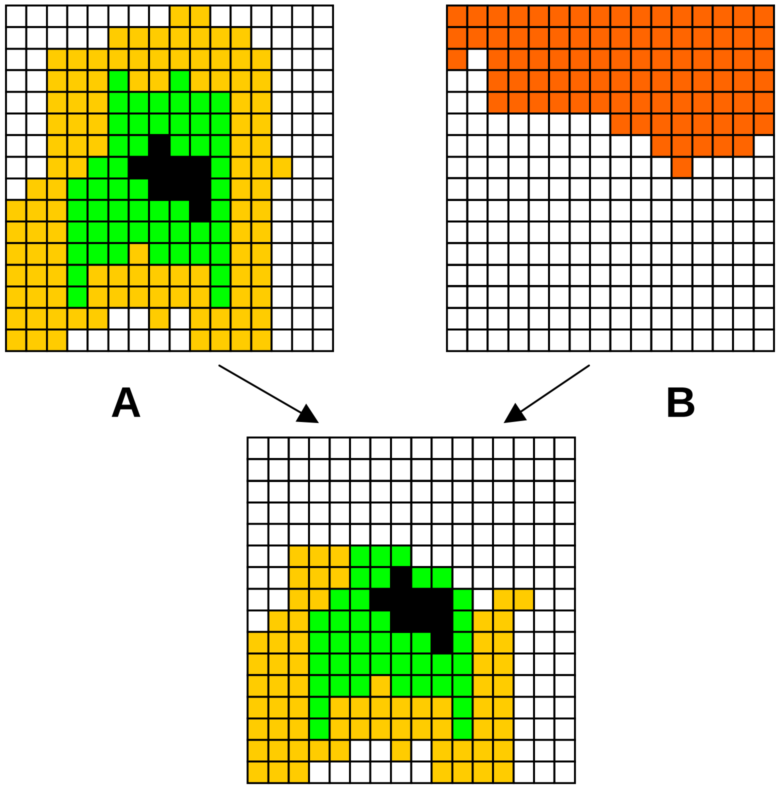
- overlay polygons and **keep only areas** in the input layer that fall **within the intersection** layer

IDENTITY

- overlay polygons and **keep areas from input layer**



Overlay of Fields Represented as Rasters



The two input data sets are maps of (A) travel time from the urban area shown in black, and (B) county (red indicates County X, white indicates County Y). The output map identifies **travel time to areas in County Y only**, and might be used to compute average travel time to points in that county in a subsequent step

Simple Arithmetic Operations

Summation

0	1	1
0	0	1
1	0	1

+

0	0	0
1	1	1
0	0	1

=

0	1	1
1	1	2
1	0	2

Multiplication

0	1	1
0	0	1
1	0	1

×

0	0	0
1	1	1
0	0	1

=

0	0	0
0	0	1
0	0	1

Summation of more than two layers

0	1	1
0	0	1
1	0	1

+

0	0	0
1	1	1
0	0	1

+

0	0	0
1	1	1
0	0	1

=

0	1	1
2	2	3
1	0	3

Chapter 6: Why is it There?

- 6.1 Describing Attributes
- 6.2 Statistical Analysis
- 6.3 Spatial Description
- 6.4 Spatial Analysis

Scales of Measurement

- **Attribute data** can be divided into four types

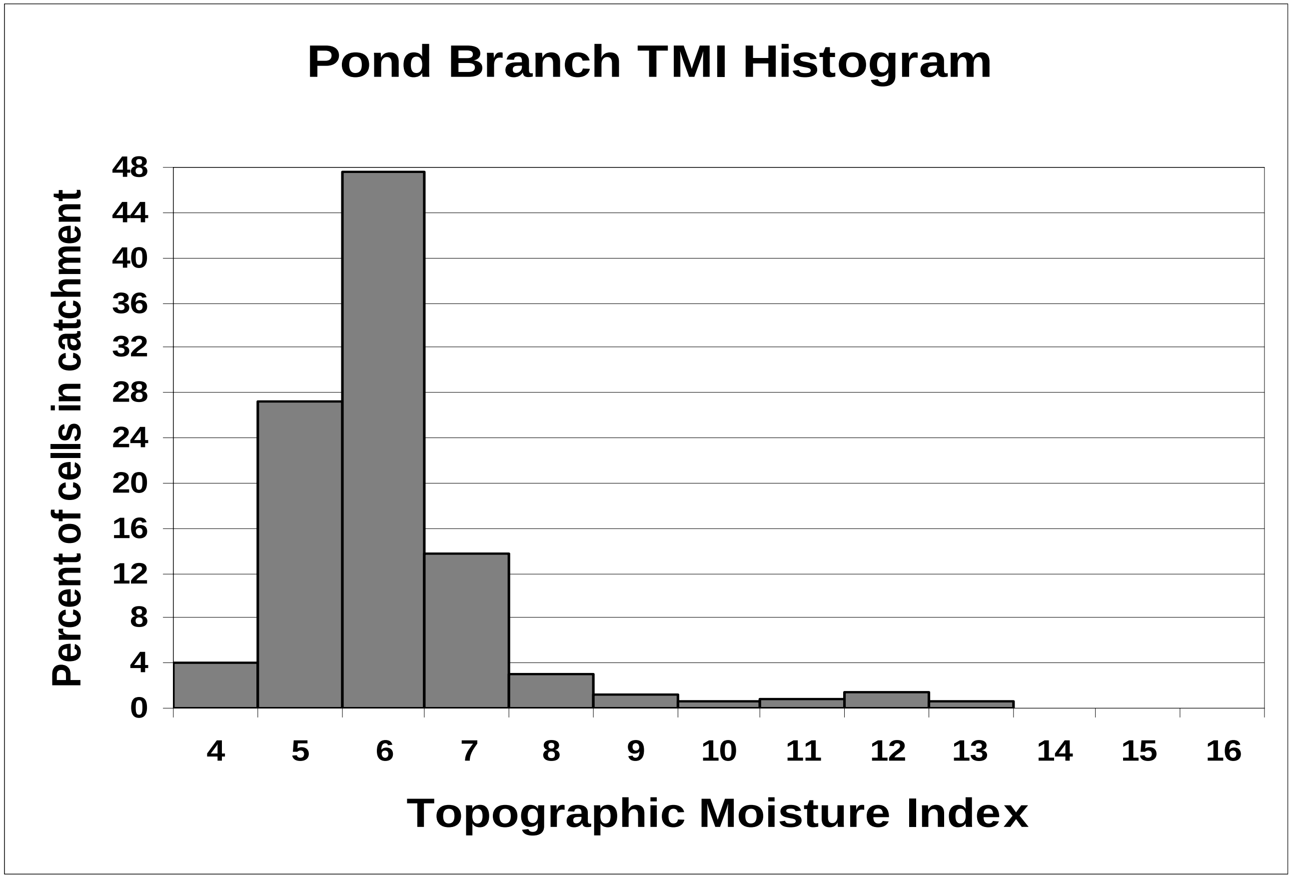
1. The Nominal Scale
2. The Ordinal Scale
3. The Interval Scale
4. The Ratio Scale

As we progress through these scales, the types of data they describe have increasing information content

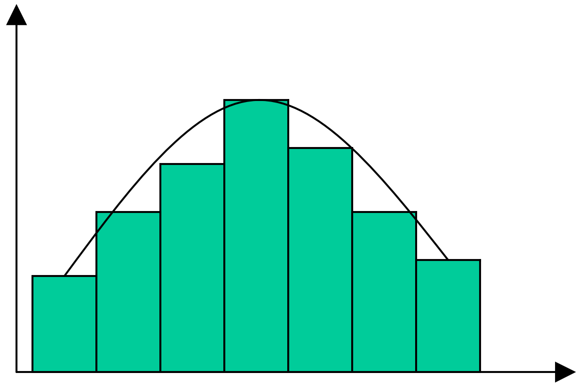
Building a Histogram

4. Plot the frequencies of each class

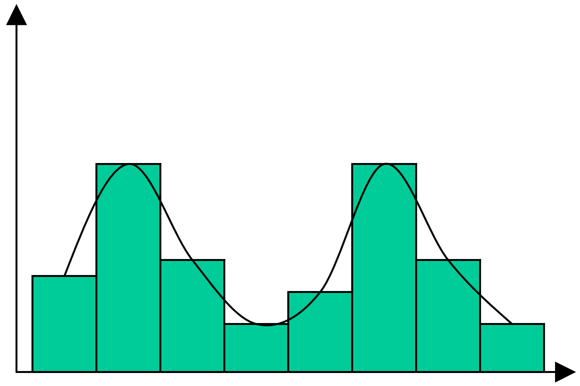
- All that remains is to create the plot:



Shapes of Histograms

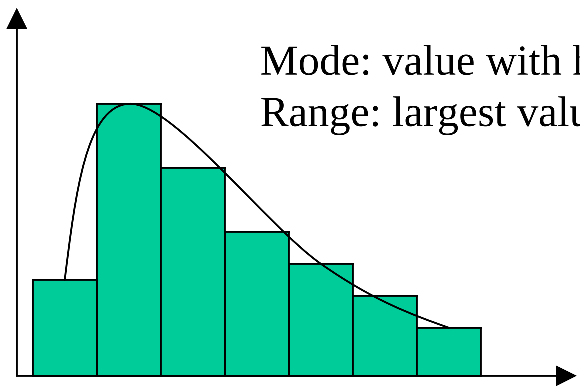


Bell Shaped

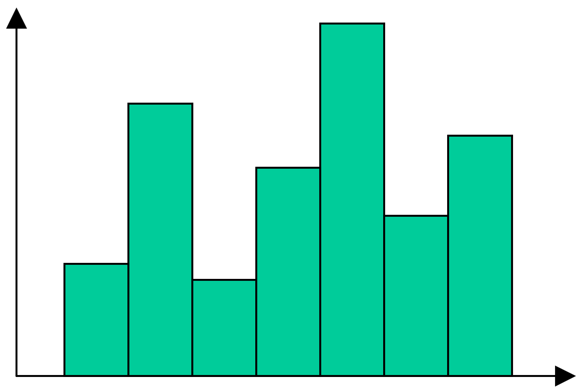


Bimodal

Mode: value with highest frequency
Range: largest value-smallest value



Skewed



Random

•Developing a histogram from attribute data is one level of data reduction; we can describe bell shaped distributions using parameters that provide a more concise summary

Measures of Central Tendency

- Think of this from the following point of view: We have some distribution in which we want to **locate the center**, and we need to choose an appropriate measure of central tendency. We can choose from:

  1. Mode
  2. Median
  3. Mean

- Each of these measures is appropriate to different distributions / under different circumstances

Measures of Central Tendency - Mode

1. **Mode** – This is the most frequently occurring value in the distribution

- In the event that **multiple values** tie for the **highest frequency**, we have a **problem** …
- A potential **solution** in this situation involves constructing **frequency classes** and identify the **most frequently occurring** class
- This is the only measure of central tendency that can be used with **nominal data**
- The mode allows the distribution's peak to be located quickly

Measures of Central Tendency - Median

2. **Median** – This is the value of a variable such that **half** of the observations **are above** and **half are below** this value i.e. this value divides the distribution into two groups of equal size

- Note: When the distribution has an **even number** of observations, finding the median requires **averaging two numbers**
- The **key advantage** of the median is that its value is **unaffected** by extreme values at the end of a distribution (which potentially are **outliers**)

Measures of Central Tendency - Mean

3. Mean – a.k.a. average, the most commonly used measure of central tendency

$$\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n}$$

Sample mean

- When we compute a mean using these basic formulae, we are assuming that each observation is **equally significant**

Measures of Central Tendency - Mean

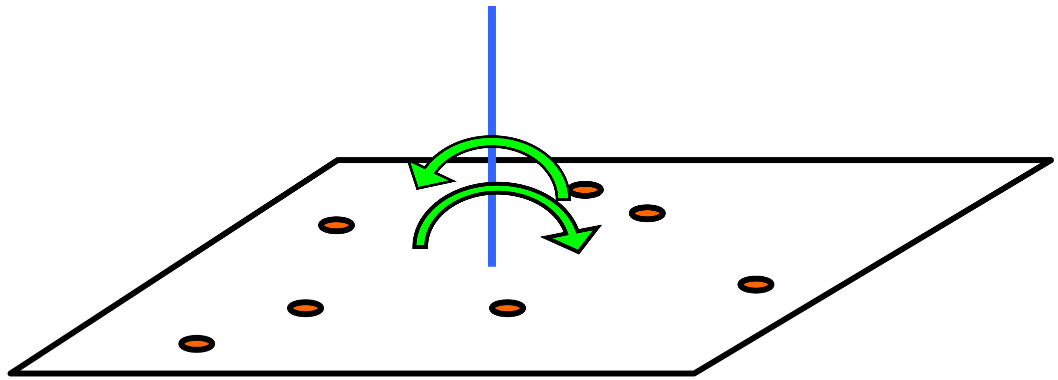
3. **Mean cont.** – A standard geographic application of the mean is to locate the center (a.k.a. **centroid**) of a **spatial distribution** by assigning to each member of the spatial distribution a gridded coordinate and calculating the mean value in each coordinate direction → **Bivariate mean** or **mean center**

For a set of (x,y) coordinates, the **mean center** $(\bar{x},\bar{y})$ is computed using:

$$\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n} \qquad \bar{y} = \frac{\sum_{i=1}^{i=n} y_i}{n}$$

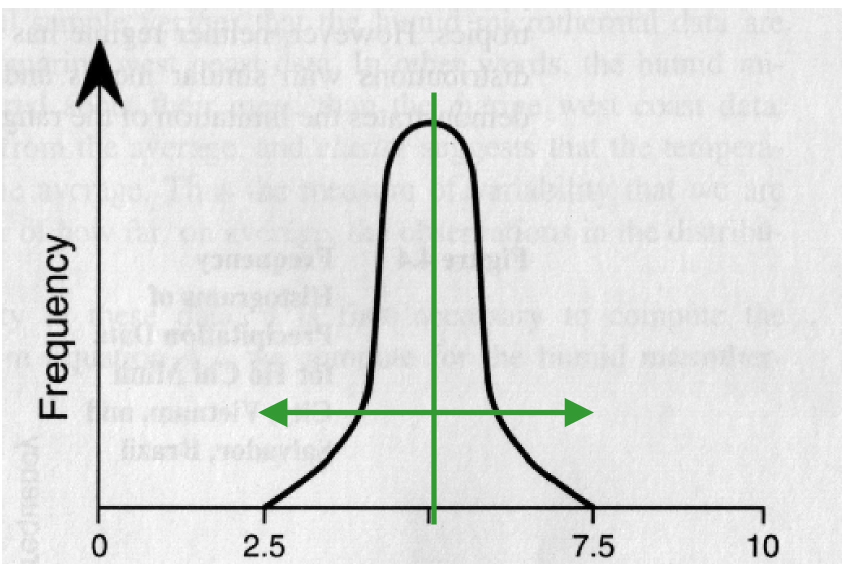
The Centroid

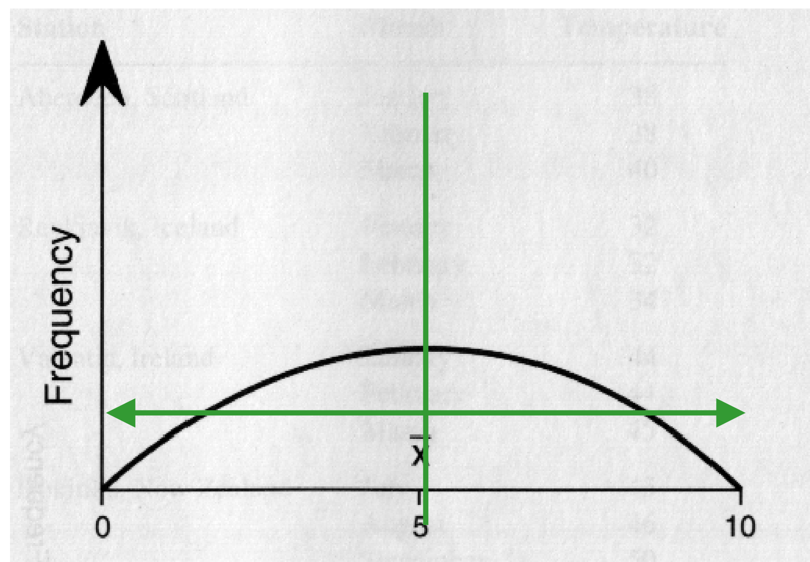
- The **mean center** can be found for a set of points by taking an **average of coordinates**, and this is also known as a **centroid**
- The centroid can also be thought of as the **balance point** of a set of points, as it **minimizes** the sum of the distances squared



Why Do We Need Measures of Dispersion at all?

•Measures of central tendency **tell us nothing** about the variability / dispersion / deviation / range of values about the central value. Consider the following two unimodal symmetric distributions:





Source: Earickson, RJ, and Harlin, JM. 1994. Geographic Measurement and Quantitative Analysis. USA: Macmillan College Publishing Co., p. 91.

Measures of Dispersion - Range

1. **Range** – this is the most **simply formulated** of all measures of dispersion

- Given a set of measurements $x_1, x_2, x_3, \ldots, x_{n-1}, x_n$, the range is defined as the **difference** between the largest and smallest values:

$$\text{Range} = x_{max} - x_{min}$$

- This is another descriptive measure that is **vulnerable** to the influence of **outliers** in a data set, which result in a range that is not really descriptive of most of the data

Measures of Dispersion – Variance, Standard Deviation, Z-scores

2. **Variance etc.** – As an alternative to taking the absolute values of the statistical distances, we can square each deviation before taking their sum, which yields the **sum of squares**:

$$\text{Sum of Squares} = \sum_{i=1}^{i=n} (x_i - \bar{x})^2$$

- The sum of squares expresses the **total square variation about the mean**, and using this value we can calculate variances and standard deviations for both populations and samples

Measures of Dispersion – Variance, Standard Deviation, Z-scores

2. **Variance etc. cont.** – Variance is formulated as the sum of squares divided by the population size or the sample size minus one:

$$S^2 = \frac{\sum_{i=1}^{i=N} (x_i - \overline{x})^2}{n - 1}$$

Sample variance

Measures of Dispersion – Variance, Standard Deviation, Z-scores

2. **Variance etc. cont.** – Standard deviation is calculated by taking the square root of variance:

$$S = \sqrt{\frac{\sum_{i=1}^{i=N} (x_i - \overline{x})^2}{n - 1}}$$

Sample standard deviation

- Why do we **prefer** standard deviation over variance as a measure of dispersion? Magnitude of values and units match means

Measures of Dispersion – Variance, Standard Deviation, Z-scores

2. **Variance etc. cont.** – Just as the mean can be applied to spatial distributions through the bivariate mean center and weighted mean center formulae (computed by considering the (x,y) coordinates of a set of spatial objects), standard deviation can be applied to examining the dispersion of a spatial distribution. This is called **standard distance** (SD):

$$SD = \sqrt{\frac{\sum_{i=1}^{i=n} (x_i - \overline{x})^2}{n - 1} + \frac{\sum_{i=1}^{i=n} (y_i - \overline{y})^2}{n - 1}}$$

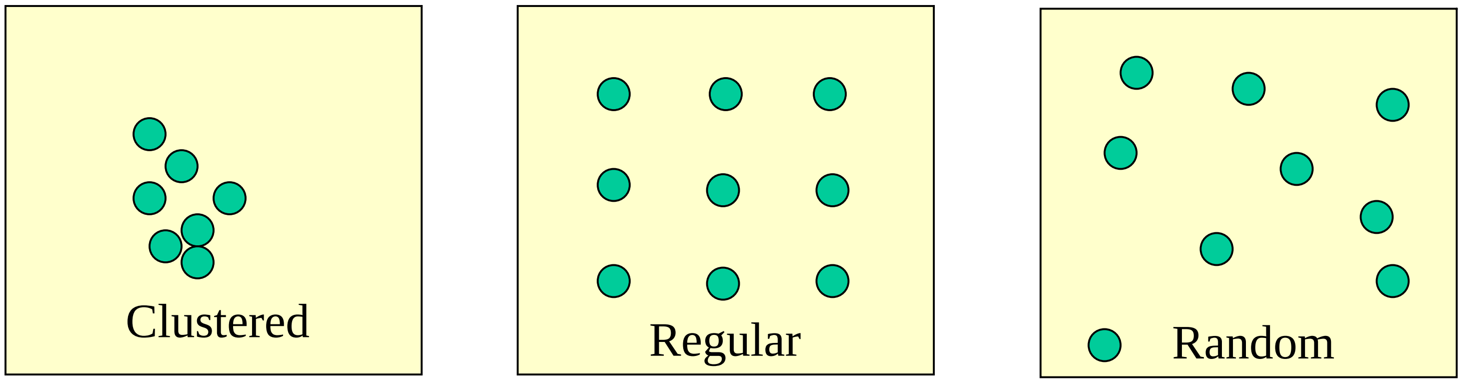
Measures of Dispersion – Variance, Standard Deviation, Z-scores

2. **Variance etc. cont.** – Sometimes, we want to **compare** data from different distributions, which in turn have different means and variances

- In these circumstances, it's convenient to have a **standardized** measure of dispersion that can be calculated for an individual observation. The **z-score** (a.k.a. standard normal variate, standard normal deviate, or just the standard score) is calculated by subtracting the sample mean from the observation, and then dividing that difference by the sample standard deviation:

$$\text{Z-score} = \frac{x - \bar{x}}{S}$$

Average Nearest Neighbor



- Point patterns can be characterized by the **distance between neighboring points**. If we define d_i as the distance between a point and its nearest neighbor, the **average distance between neighboring points** can be written as:

$$D_A = \frac{\sum_{i=1}^{n} d_i}{n}$$

The Nearest Neighbor Index

- We can calculate the **expected distance** (D_E) between randomly distributed points using:

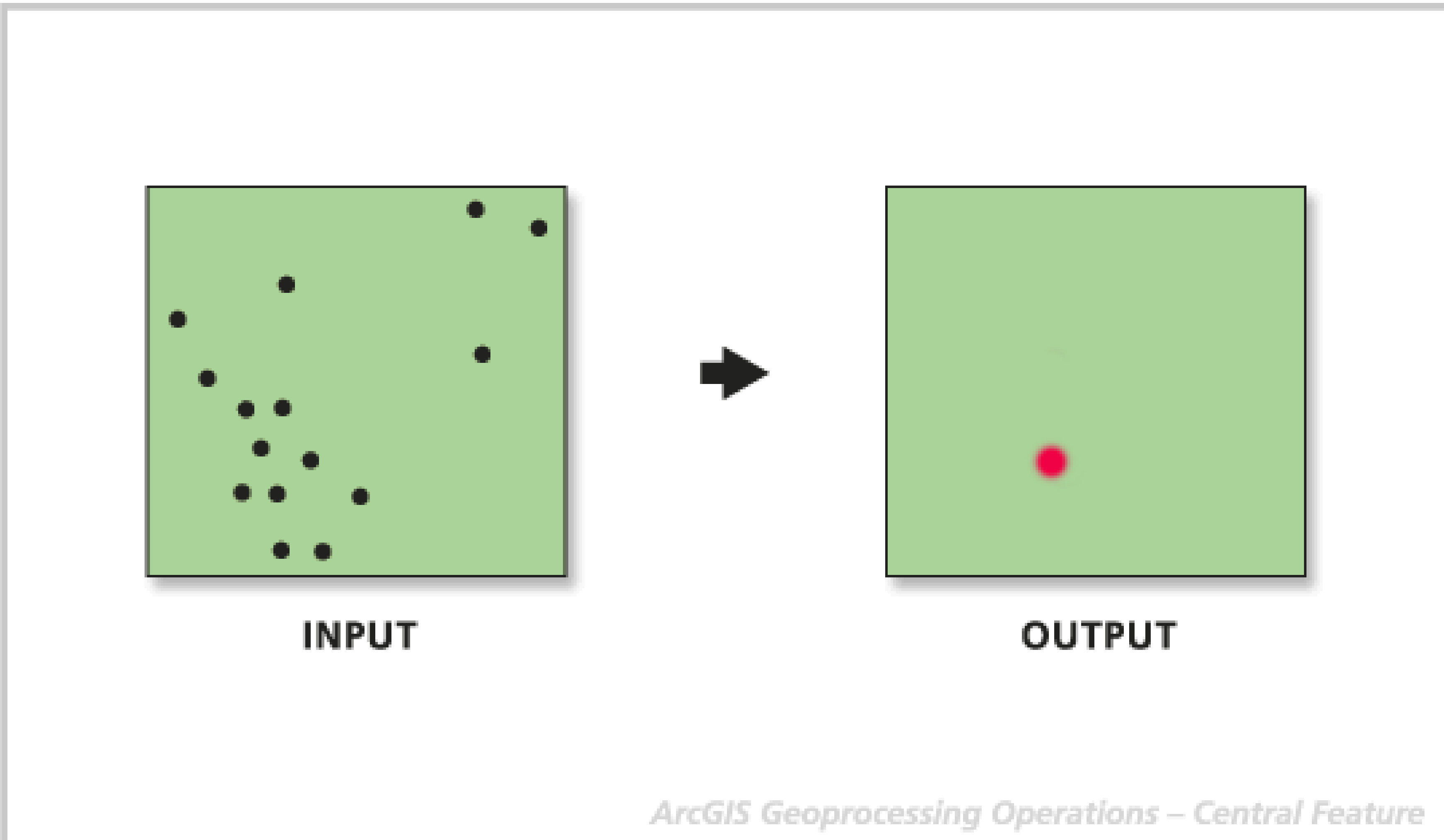
  $$D_E = \frac{1}{2}\sqrt{\frac{A}{n}}$$ where A is the area and n is the # of points

- We can determine the degree to which a set of points is randomly distributed by **comparing** the **actual distance** between the points (D_A) with the **expected distance** (D_E), taking the **ratio** between the two, known as the **nearest neighbor index** (NNI):

  $$NNI = \frac{D_A}{D_E}$$

- **Random** points: $D_A \sim D_E$, $\therefore$ NNI ~ 1
- **Clustered** points: $D_A \sim 0$, $\therefore$ NNI ~ 0
- **Dispersed** points: D_A larger up to max. NNI = 2.1491

Spatial Statistics Tools – Central Feature

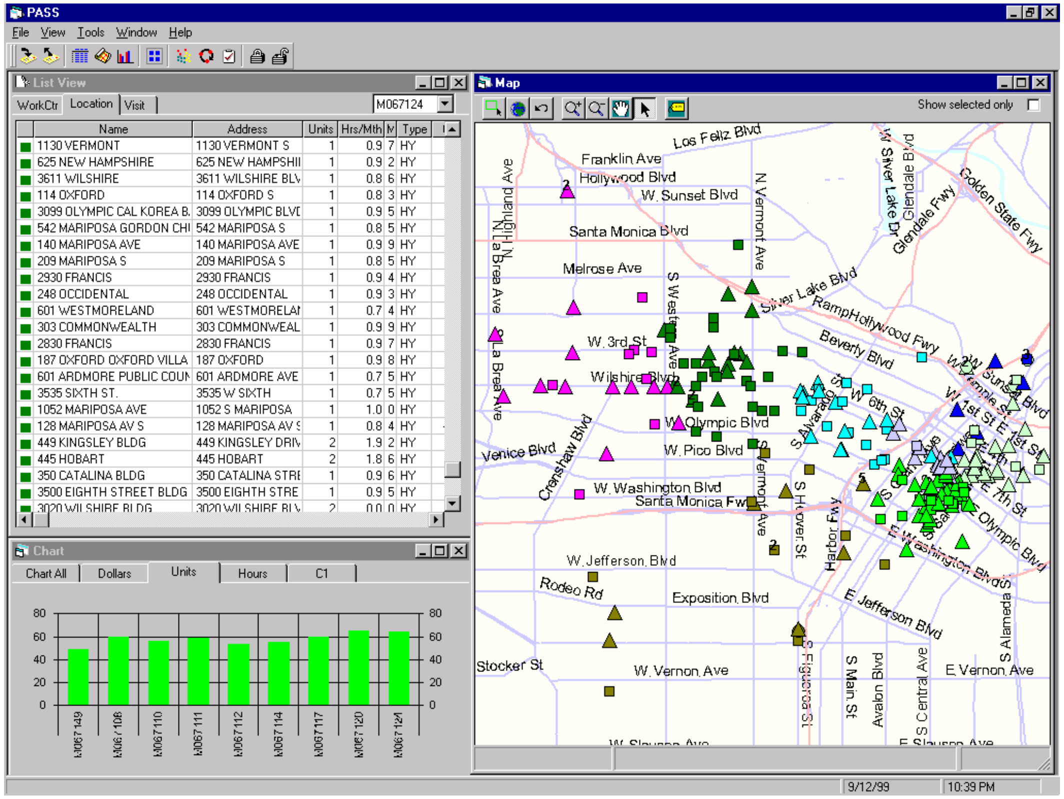


Central Feature illustration

Optimization

- Spatial analysis can be used to solve many **problems of design**, such as "where is the best place to build a new x"
- The decision as to where to build a new facility is often approached from the point of view of **maximizing access**, or **minimizing travel time** from a certain catchment or service area,
  - e.g. if we identify a developing area where the nearest hospital is an unacceptably long drive away, we may know we want to locate a hospital in that area … but **where should we put it** to best serve the residents in the area and minimize overall travel time for the area?
- To do, we can identify the **point of minimum aggregate travel** (MAT)

Routing Problems – The Traveling Salesman



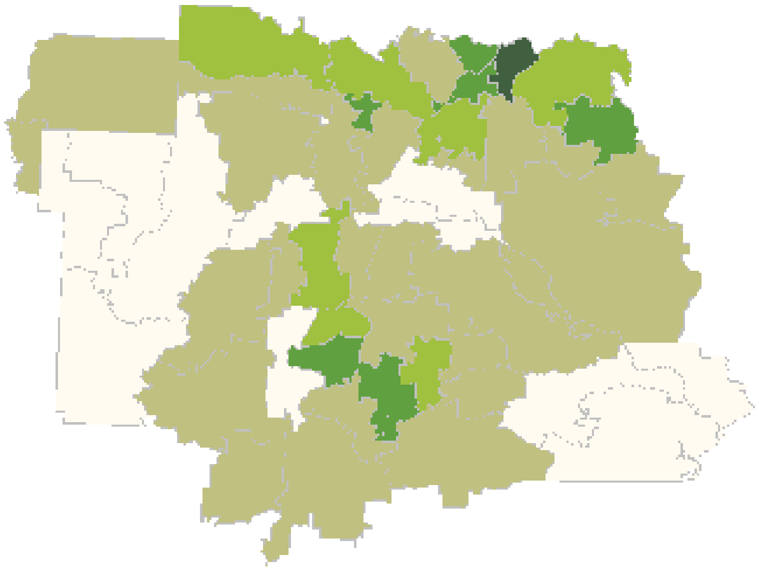
•Routing service technicians for Schindler Elevator: Every day this company's service crews must **visit a different set of locations** in Los Angeles. GIS is used to partition the day's workload among the crews and trucks (color coding) and to **optimize the route to minimize time and cost**

Least-Cost Path Example

- The figure to the left shows the solution of a **least-cost path problem**:
- The white line represents the **optimum solution**, or path of least total cost, across a friction surface represented as a raster layer
- The area is dominated by a mountain range, and **cost** in this example is determined by **elevation and slope**
- The best route uses a **narrow pass** through the range. The blue line results from solving the same problem using a coarser raster

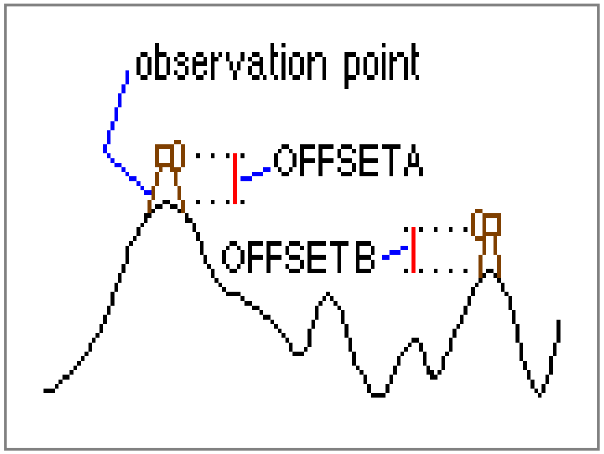
Hot Spot Analysis

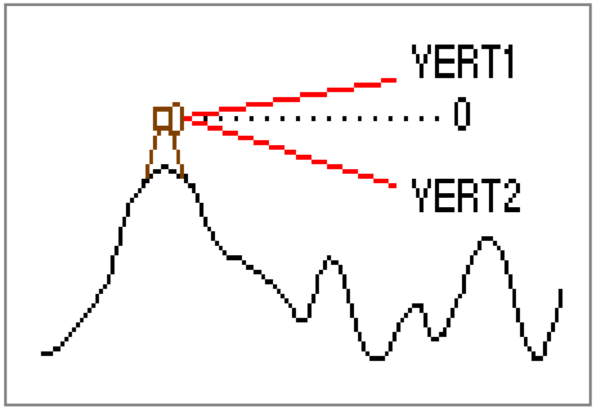
- The hot spot analysis tool creates a **new feature class** that duplicates the input feature class then **adds a new results column** for the hot spot (Gi*) Z score values.



Spatial Analysis Tools – ViewShade Analysis

- **Which areas can be seen** from a fire lookout tower that is 15m high?
- **How frequently** can a proposed disposal site **be seen** from an existing highway?
- **Where** should the next communications repeater tower in a series be located?





Spatial Analysis Tools – Line of Sight

- Uses an **input 3D line** feature class to **determine visibility along its lines**.
- Produces an **output line** feature class that contains **line and target visibility information**.
- If the **target is not visible**, Line of Sight produces an output point feature class that shows the **first obstruction points** along the lines..

Spatial Analysis Tools – Surface Volume

- Calculates the **area and volume** of a functional surface above or below a **given reference plane**.