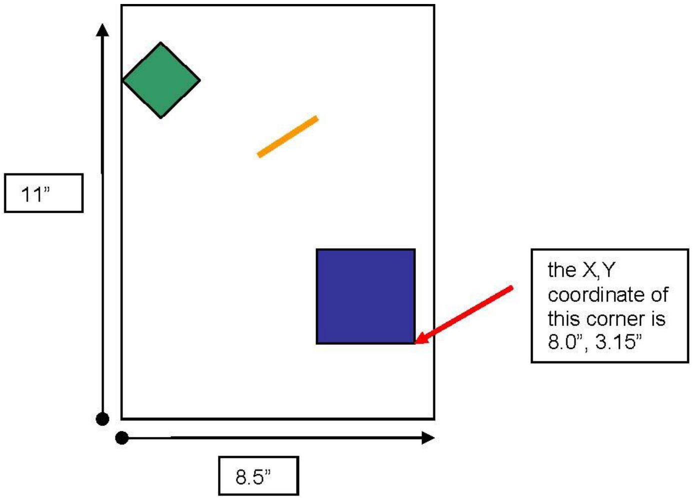
TeachGIS White Paper Number 1:
A Few Words on Locations, Coordinate Systems, and Projections

When we hand draw a map on the back of an envelope or a napkin, it usually doesn't matter that the position of the phenomena we are drawing is relative to other phenomena or to its exact location, wherever that may be. We're just trying to share basic spatial information, information that isn't designed or intended to be used for military maneuvers or exact measurements. But if we migrate that map to a computer-based map making system, such as a GIS, location can no longer relative. Drawing graphics, and mapping software requires that the depicted features be referenced or registered to *some* kind of coordinate system. Maybe that's just an arbitrary system of "page" units or digitizer units. For example, if we were to scan an 8 ½ by 11 inch piece of paper that has shapes drawn on it, we could determine the X and Y coordinates (in inches) of the four corners of the purple square, coordinates that now represent the absolute location of that square (on a piece of paper that size, measured in inches). And if we were to supply the computer with those corner coordinates again, we could draw the same shape in the same place. To draw a square, the program will use straight lines to "connect the dots" of the four corners, and voilà, you have a square. That's how the software keeps track of what goes where on the drawing.



The same thing is happening when features are located on a map. We have different coordinate systems to indicate the geographic location of those features (and for now, we'll only talk about locations on the surface of the Earth, not planets or Middle Earth, etc.).

There are many planar (X and Y, Cartesian) coordinate systems that can be used to describe the location of objects or features on the surface of the Earth. Some coordinate systems can be used to describe a location anywhere across the entire globe (such as latitude and longitude), while others have a smaller extent and are designed to be used only locally (such as a system only for a single state, or even a single city). In many places, these gridded systems may overlap each other in parts of the world.

As an example, where is the State of Liberty located? How would you know exactly where to place it on a map? Where the computer would place it on a map depends on the coordinate system we are using at that instance.

We can identify the **location** of the **crown on the Statue of Liberty** by its coordinates in several different ways:

a) **New York State Plane Coordinate System** for the **Eastern New York zone** =
X = 188468.45, Y = 205591.90 (units are meters)

b) **Universal Transverse Mercator (UTM) Coordinate System** for **Zone 18 North** =
X = 580741.73, Y = 4504121.09 (units are meters)

c) **Latitude** and **Longitude** = **X = 74° 02' 40.54" West longitude, Y = 40° 41' 20.91" North latitude** (units are degrees/minutes/seconds).

Obviously, the Statue has not moved, but that location can be *described* in many different ways through these different coordinate systems. It's like measuring an object either in inches or centimeters, or temperature in Fahrenheit or Celsius. Someone’s height may be considered static or perhaps ***absolute***, but that height can be described with reference to one of several ***relative*** and sometimes arbitrary systems (someone is 5’4” inches tall, or 162.6 cm, or 1.778 yards, or 288455054883826.56 electron diameters, or 0.00101010101 miles, or 4.671264367816092e-7 Moon diameters. You get the picture.)

In order for GIS software to display spatial data from different sources in correct alignment, with layers matching up on top of one another, the computer must be able to figure out the coordinate systems for each data set. If the coordinate systems are known, then the software is able to convert them to a common one for display purposes. That is, if the software knows that one layer has its data set measured in inches, and another one is in centimeters, it can apply a conversion factor (multiply centimeters by 2.54) and display both sets correctly, in one system or the other.

> **TIP:** One of the most common problems in GIS occurs when the software *does NOT know what the coordinate system units* are for a particular data set. A GIS may tell you that **the spatial reference is unknown**. This is easy to remedy if you know the information and can tell the computer what it is, by “defining the projection.”
>
> However, it is vastly more difficult to fix this problem when you are not aware of what spatial reference system was initially used for those data. Fortunately, there are **ways to make educated guesses.**

So far we’ve been envisioning Cartesian coordinate systems, ones whose X and Y lines of measurement are always perpendicular to each other, forming a perfect grid. This works just fine when we are measuring or mapping a small area (a town or a county), but not so well when we are mapping countries or whole continents. In those cases, the spherical surface of the Earth precludes the use of a simple, rectangular grid system.

Two coordinate systems have been designed that capitalize on the familiarity and simplicity of two-dimensional (XY) geometry yet also provide unique coordinate locations across larger areas like the United States or the world: the **U.S. State Plane System**, developed for the continental United States, and the **Universal Transverse Mercator (UTM)** System, designed for use on a global scale. To understand how these coordinate systems operate, however, requires basic familiarity with map **projections**.

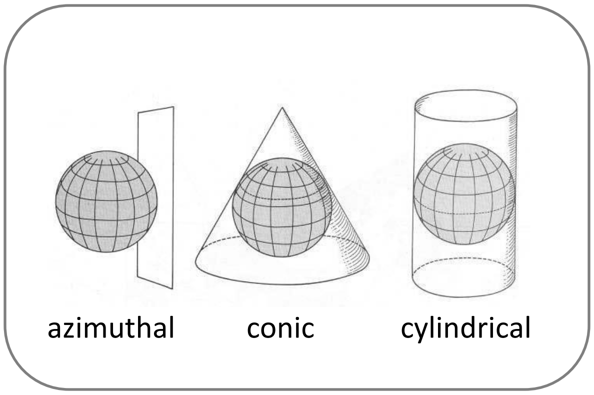
Introduction to Projections:

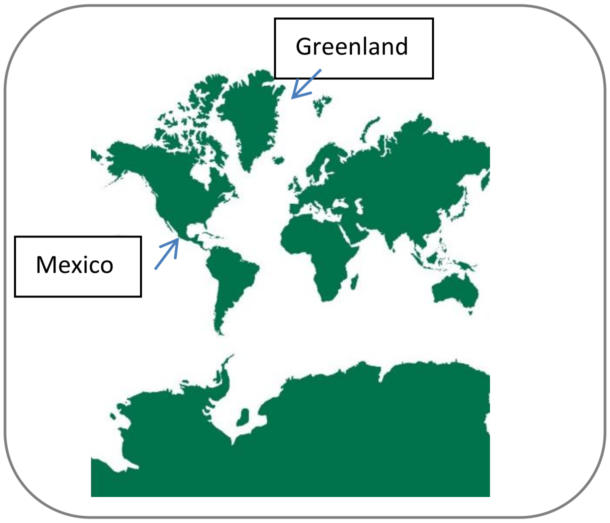
Coordinate systems are each associated with one or more map **projections**. Projections are geometric transformations that convert our 3-dimensional (round) globe onto a 2-dimensional (flat) piece of paper (or a computer screen). There are essentially an infinite number of possible projections, because each can be customized slightly with different parameters.

The projection of information from a three-dimensional sphere to a two-dimensional map is imperfect and some aspects of the three-dimensional map must be compromised, or **distorted**.

> **TIP:** Imagine that the Earth is an orange. Unpeel the orange (carefully, so that the peel remains intact in one piece), and then place the peel over a rectangular piece of paper. Flatten it out so that the peel covers as much paper as possible. Where will the peel fill most of the paper? (Answer: the middle, the equator.) Where wdill it fill the least? (Answer: the top and bottom, the poles.) Think of all of the maps of the whole world that you have seen that fill the whole rectangle. Do you usually see gaps? What has filled those "gaps"? Extra ocean? Yes, and extra land. It's a cartographic scam! Read more to learn why it has to happen.

The three main categories of projections (shown below as sketches from left to right) are **azimuthal**, **conic**, and **cylindrical**. If you imagine a light source inside of each of those globes, projecting shadows out onto the piece of paper, you can envision the types of **distortion** that different projections will create. **Distortion** (of **area**, **shape, direction**, or **distance** between true characteristics) is an inevitable result of projecting between 3D and 2D. In the small figure below in the right, you see the infamous, cylindrical **Mercator projection**, showing **distortion of area** (in reality, Greenland is about the size of Mexico; note its exaggerated size in the figure).



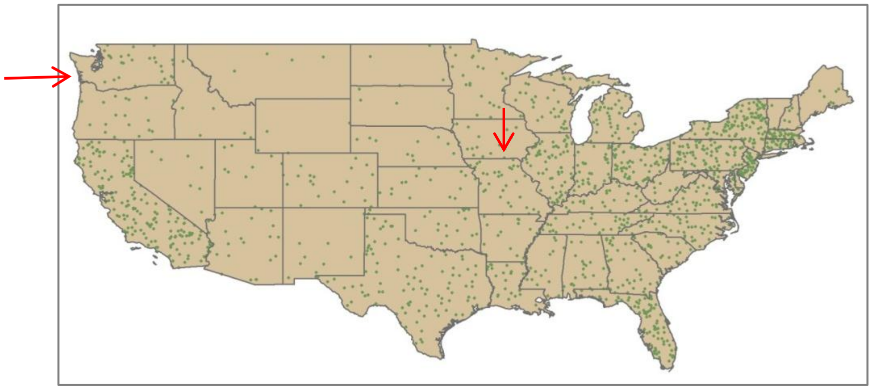


Data sets that are maintained in **latitude** and **longitude coordinates** (or a Geographic Coordinate System, **GCS**, as Esri calls it) are **NOT projected** coordinates, technically. However, we represent these in 2-D space by extending lines of longitude (meridians) directly north and south indefinitely, forming right angles at every intersection with a degree of latitude (rather than having them converge at the north and south poles, as they do in reality). The result appears similar to a cylindrical projection and is sometimes referred to as the "un-projected projection" or the Plate Carrée projection.

Why does it matter which map projection we use? Here are two reasons:

1) We expect the software to give us both accurate and precise answers when we calculate the distance between two or more points, or the area of a polygon, or the volume of a lake. When those calculations are derived from data whose projections are inappropriate for that purpose (such as measuring the surface area of polygons in a data set with a conformal projection, such as the Mercator or Plate Carrée, with which both shape and area can be grossly distorted), the answers will be wrong. Instead, areas should only be calculated from equal area projections, and the distance between two points can best be determined from a map that has a true scale throughout.

2) We should also be aware of the ways in which different projections affect the way in which a map is viewed and perceived. For example, if we are making a dot density map of population in the United States, a conformal projection will exaggerate areas in the north, creating more "space" for dots to occupy and creating the illusion that they are less densely populated with dots.



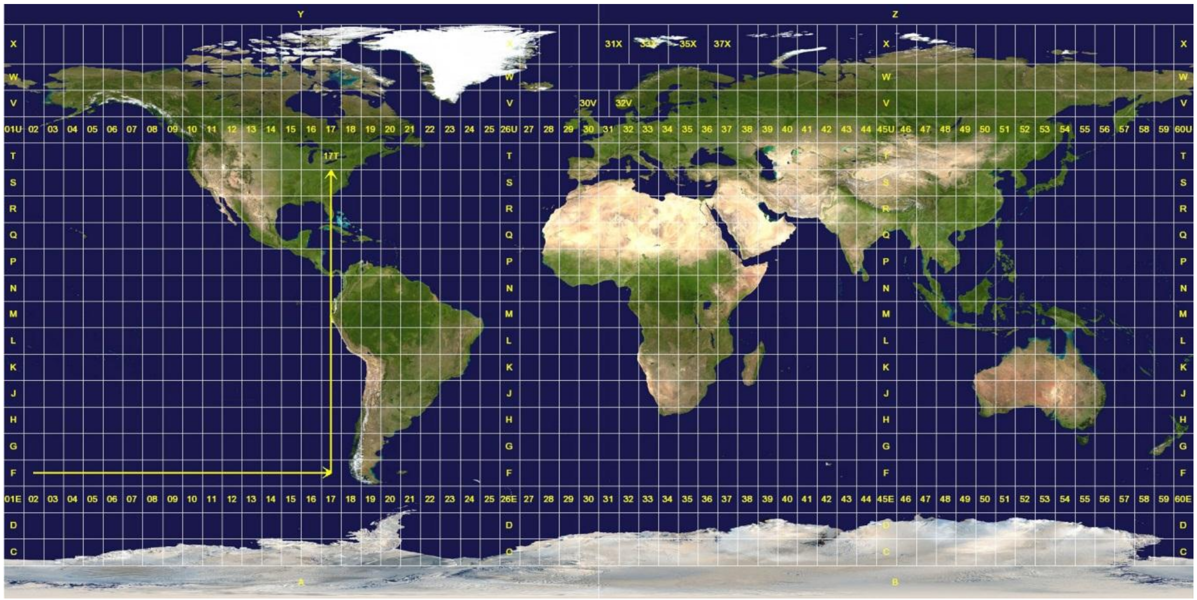
Alternatively, an equal area projection removes that source of visual error and the States can be compared on an equal basis: this map below has the same exact number of dots per state as the map above, yet the densities seem to have been altered. In the map above, Washington seems to have fewer dots than Missouri, yet they actually have an almost identical number (each dot equals 250,000 people; figures are for 1999), and the two States are approximately the same size too (though Washington is even slightly smaller than Missouri). These realities are reflected more clearly in the map below, an equal area projection.

A brief return to coordinate systems:

As mentioned earlier, there are two planar coordinate systems commonly used in the United States: the **State Plane Coordinate System (SPCS)** and the **Universal Transverse Mercator (UTM) coordinate system**. Geographical data are still often stored in longitude and latitude, but the State Plane and UTM systems offer several distinct advantages for both display and analysis.

The **State Plane** system was developed in the 1930s. At that time the country was divided into 125 zones, following state lines, each with its own projection(s) selected to minimize distortion areas within the area of each zone. Most often the projections used for each zone are a Transverse Mercator or a Lambert Conformal Conic, depending on whether the state and its zones tend to run north/south or east/west. Within each state, a planar grid is overlaid on top of each zone(s) to establish zone-specific coordinate systems.

The Universal Transverse Mercator (UTM) system is conceptually much like the State Plane system except that it is defined to provide coordinate reference for the entire world. Beginning at 180 degrees longitude and counting eastward, the globe is divided into zones that are 6 degrees of longitude in width. Zones are numbered beginning with one and proceeding sequentially eastward. A transverse Mercator projection is constructed for each zone, and top of each zone is overlaid a planar grid, just as is done with the State Plane Coordinate System. Meters are the standard unit of measurement with the UTM system.



And now onto Datums:

The surface of the Earth is far from regular and smooth. Gravitational differences affect the surface of the oceans and generate a surface that is quite irregular (a shape we call a geoid), plus there are the surface variabilities of mountain ranges and the like. Hence, the uniform coordinate system grids that we have been discussing cannot be overlaid directly on these actual shapes.

Instead, we use an **ellipsoid** – a three-dimensional ellipse – because it has a regular and predictable geometry. Ellipsoids are “fit” to geoids, and these then serve as frames of reference or **datums** for specifying position on Earth. However, ellipsoids that fit the geoid well in one area – say the continental United States – are likely to provide a very poor fit in some other area like the Saudi Arabian peninsula or the South Pacific. As a consequence, different ellipsoids and different datums are used in different parts of the world.

Moreover, as knowledge and understanding of the shape of Earth has grown, the geoid has been revised and refined. An ellipsoid that seemed to fit a particular area well at an earlier point in time may become inadequate as the ability to measure location becomes more detailed and presumably more accurate. Until recently, most locations in the United States were referenced to the **1927 North American Datum (NAD27)**, which was based on an ellipsoid that was defined in 1866. Most geographic data collected at present are referenced to the **1983 North American Datum (NAD83)**, which is based on an ellipsoid adopted in 1980. Many older data sources, however, still give locations in the 1927 datum. The distinction between NAD83 and NAD27 is especially important, because the location of the same point may shift well over 50 meters due to nothing more than a difference in datums.

To confuse the subject even further, the default datum used by most GPS receivers is known as **WGS84**, for **World Geodetic Survey of 1984**. Unlike other datums in common use, WGS84 is based not on a specific ellipsoid but on the hypothetical center of Earth, as determined through geophysics. Though they are different datums, WGS84 is almost identical to NAD83 (pdf).

Remember, data in one datum can be transformed permanently into another datum by most GIS software packages, though you may be given a warning when the software will be doing so temporarily and automatically for you for display purposes. This on-the-fly re-projecting for display purposes is often adequate for just that, display. But it is recommended that you convert your data to a single projection for analytical purposes.