Lesson 2: Understanding Geospatial Data Models

## INTRODUCTION

This lesson focuses on geospatial data models. The information and images explain the elements that are used in making data models. You will learn about discrete and continuous spatial phenomena and how these concepts are applied in mapped examples. The lesson also provides information on attributes and why it is important to understand attributes when creating geospatial data models.

## LESSON OBJECTIVES

By the end of this lesson, you will be able to:

1. Identify different types of data and data models used to store and analyze information within a Geographic Information System (GIS).

2. Use a GIS application to add and view data.

**LEARNING SEQUENCE**

|  |  |
| --- | --- |
|  | |
| Required Reading | Read the following:  Understanding Geospatial Data Models   * Geospatial Data Models * Attributes * Spatial Data Models * Raster Data Models * Triangulated Irregular Network (TIN) |
| Assignments | Complete the following assignments:   * Lab: Exploring Geospatial Data Models * Quiz: Understanding Geospatial Data Models |

## INSTRUCTION

Geospatial Data Models

## What are Geospatial Data Models?

A geospatial data model is a formal means of representing spatially referenced information. It is a simplified view of physical entities and a conceptualization of reality. What I mean by "simplified conceptualization of reality" is that while we would like to store a complete model of reality, we really can't do that as it is not possible to represent every possible variable that exists in reality. For instance, if we want to represent a road in a geospatial data model, we would probably store the pavement type, number of lanes, and speed limit. We would not be able to store the amount of wear on the road based on the friction of all the tires that have traveled the road, nor would we necessarily want to. For our conceptualization of reality, we only store the information that is important for our task.

|  |
| --- |
| **Key Facts**  Geospatial Data Models   * Formal means of representing spatially-referenced information * Simplified view of physical entities * A conceptualization of reality |

## Geospatial Data Models: Object and Attributes

A geospatial data model is composed of two parts: object and attributes. The object and attributes are combined together to form a simplified concept of reality. The object stores the geometry of the spatial phenomenon that we are modeling. The attribute stores the descriptive information related to the object part of the geospatial data model.

|  |  |
| --- | --- |
|  | Geospatial Data Models are composed of two parts, object and attribute to form a concept of reality. |

## Satellite View Example

As an example, here's a satellite view of the lower 48 States of the United States of America. Let's assume that this satellite picture represents reality. So the question is: "How can we simplify and conceptualize this picture of reality for our purpose?" What if we want to conceptualize the USA as a basic reference map that only shows the boundary of the United States of America, state boundaries, water features, and major roads?

We are going to represent our chosen model of reality using a geospatial data model. Once we have the data in the geospatial data model, we are then going to apply some good map design principles to arrive at a map that looks like what we would expect.



## General Reference Map

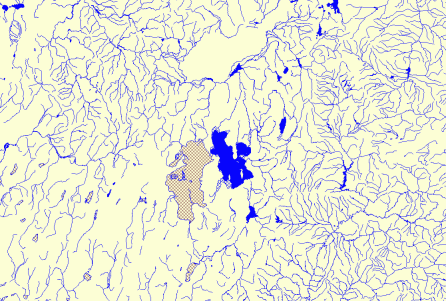
|  |  |
| --- | --- |
|  | This is a general reference map of the United States of America. What we see on this map is our simplified view of reality. We see the national outline, state outlines, water features, roads, and even elevation information. Note that while we are representing roads, we chose only to represent major roads and are leaving off all of the other roads for the purpose of readability.  By only showing major roads, we have simplified reality for our purposes. In fact, every facet of reality that we are displaying on this map has been simplified in at least one way. |

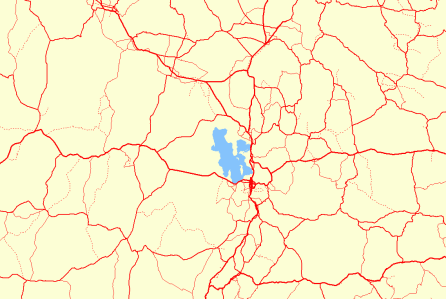
## Types of Spatial Phenomenon

Now, if we're going to simplify reality and are going to store the concept of reality in a geospatial data model, we need to understand the way in which spatial phenomenon are structured. There are two types of spatial phenomenon that we are going to discuss: discrete and continuous.

### Discrete Spatial Phenomenon

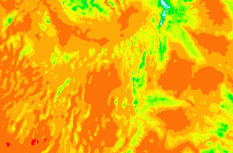
A discrete spatial phenomenon is anything that exists that is individually distinguishable. It has well-defined boundaries and it is easy to see where it begins and ends. It does not exist between where we observe it to be. Here are two examples.

It is very easy to tell where a stream or lake begins, where they end, and they are both individually distinguishable on the Earth.

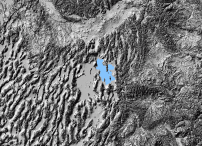
Roads have well-defined boundaries and are each easily distinguishable as individual roads.

### Continuous Spatial Phenomenon

The second type of spatial phenomenon is continuous. A continuous spatial phenomenon is something that exists between our observations. It is data of a continuous nature that cannot be isolated as an individual. It is perhaps best explained through examples.



For instance, temperature is a continuous phenomenon that gently varies throughout space. A temperature reading at a single location does not represent a well-defined location where it is exactly that temperature. Instead, the temperature reading must be put in the context of the surrounding area as part of the larger "surface" of temperature readings; Temperatures should be considered as part of a larger continuous spatial phenomenon.



Elevation is another example of a continuous spatial phenomenon. An elevation reading is only a single point in a larger surface.

## Map Representation of Discrete and Spatial Phenomenon

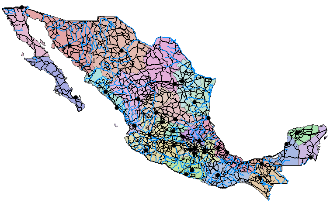
If we consider continuous and discrete spatial phenomena, we can simplify them to a map that looks like this.

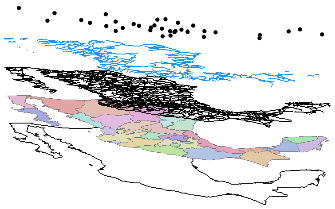
We would consider this map to be a reasonable representation of the United States of America and the spatial phenomenon that falls within. The discrete spatial phenomenon would be the roads, the national boundary, the state boundary, and the water features. The continuous spatial phenomenon would be the elevation information on this map.

## Organizing Geospatial Data Models

Now that we know that we can store representations of discrete and continuous spatial phenomena in geospatial data models, we should consider how we organize them. Typically, spatial phenomena are organized into separate geospatial data models by theme. Separating our models of reality will provide us with many benefits when it comes to querying and analysis.

Additionally, we tend to group things that are similar together in our minds, so we mirror that tendency in our geospatial data models. In the example on the screen, we have modeled five facets of Mexico: Cities (the top layer), water features, roads, state boundaries, and the national boundary. While we do separate the spatial phenomena into separate geospatial data models in the computer. If we lay these layers on top of each other, and flatten them, we can then produce a map that looks like this.

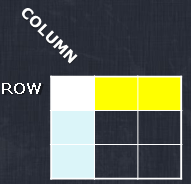
This map is composed of five geospatial data models, one for each theme of information that we have modeled from reality. By placing these "layers" of information on top of each other, we can produce useful visualizations, conduct analysis, and execute queries.

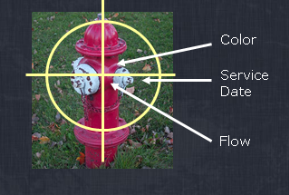


**Attributes**

## What are Attributes?

Attributes are the non-spatial characteristics that describe spatial entities. Attributes are commonly arranged in tables where a row is equivalent to one entity and a column is equivalent to one attribute, or descriptor, of that entity. Typically, each row relates to a single object and a spatial data model. Additionally it is typical for each object to have multiple attributes that describe the object. All attributes are often displayed in a table format; on the computer it can be stored in many ways, such as a flat file format, or in a database management system.

* + **Row** = 1 entity
  + **Column** = 1 attribute 

Let’s look at an example. Let’s say we have a spatial data model that stores the location of fire hydrants. For each fire hydrant, to represent the object, we would store the position. In addition to the positional information, we will also store attributes that will describe those fire hydrants. In this example, we are storing color, service state, and flow as three attributes that describe this particular fire hydrant at this particular position on Earth. The position, color, service state, and flow will be stored as one row in an attribute table that will contain four columns because there are four descriptors for this fire hydrant.

## Storing Descriptive Information

Attributes can store all kinds of different descriptive information, however, depending on the descriptive information stored in each attribute, some properties are implied, so operations are permitted, and other operations and properties are not allowed and not applied respectively. We can break attributes down into four different categories: nominal, ordinal, interval, and ratio.

### Nominal Attribute

The first attribute category is the nominal attribute category. A nominal attribute provides descriptive information about the object such as the color of the object, or the name of an object for instance a city name, or the type of an object. What’s important here is that this descriptive information does not imply any order, size, or any other quantitative information. That means that you cannot state that one attribute is greater than or less than another attribute or you cannot multiply attributes together, so for instance, it does not make sense to multiply the color blue by the color red. The only comparisons you can do with nominal attributes are to check whether two attributes are equal or not equal.

In addition to text descriptions, the nominal attribute category includes descriptive information such as images, movies, and sounds.

|  |
| --- |
| **Key Facts**  **Nominal Attribute**   * + Provides descriptive information * Example: color, city name, plant type   + No implied order, size, or other quantitative info   + May also be images, movies, sounds |

### Ordinal Attributes

The next attribute category is the ordinal attribute category. Ordinal attributes imply a ranking or order based on their values. These values can be descriptive text, or numerical. For example, I can describe an object as having a high/medium/low ranking, or a ranking of 100/50/1. In either case, these ordinal attributes allow us to specify rank only, and not scale. So for instance, we can state that high is ordered higher than low, and high is ordered higher than medium, and low is ordered lower than high, but we cannot say that high is twice as high as medium, and medium as twice as high as low. Additionally, if the numerical attributes are of the ordinal attribute category, again we can say that 50 is ordered higher than 20 and 20 is ordered higher than 10 but we cannot say that 50 is twice as high as 25 and 25 is twice as high as 12 ½. Even though we are using numbers to describe a rank, do not let that confuse you into thinking that a scale is implied.

|  |
| --- |
| **Key Facts**  **Ordinal Attribute**   * + Imply a ranking or order by their values   + Uses descriptive or numerical attributes * Example: High/Medium/Low or 100/50/1   + Specifies rank only, does not specify scale * Example: High/50 is ordered higher than Low/1, but we do not know by how much |

### Interval Attributes

The third category is the interval attributes category. Interval attributes imply a rank order and magnitude or scale. Interval attributes use numbers, however those numbers do not have a natural zero, and use an arbitrary zero point instead. For instance if we look at temperature on the Fahrenheit scale, 0°F is not a natural zero point for temperature, it is a human defined zero point. Therefore, while we can say that 50°F is 10°F more than 40°F, we cannot say that 50°F is twice as hot as 25°F, again, because 0°F is a human created zero, and not a natural phenomenon. With an interval attribute, addition and subtraction to make sense but not multiplication since values are relative from that arbitrary zero.

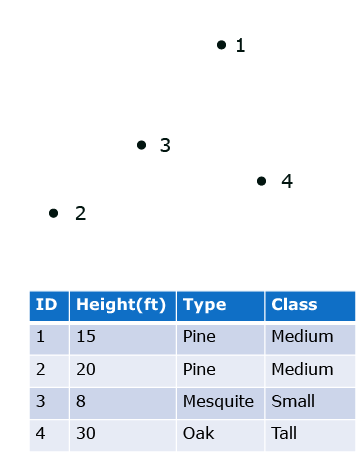
|  |
| --- |
| **Key Facts**  **Interval Attribute**   * + Rank order and magnitude is implied   + Does not have a natural zero     - Uses an arbitrary zero point instead   + Uses numerical attributes.     - Example: 50°F is 10°F more than 40°F   + Addition and subtraction make sense, but *not* multiplication since values are *relative* from an arbitrary zero. |

### Ratio Attribute

The fourth and final category is the ratio attribute category. A ratio attribute implies both rank order and magnitude about a natural zero. The ratio attribute category uses numerical attributes that allow addition, subtraction, multiplication, and division. These mathematical operations make sense, as the values are based around a natural zero. So for example, if we are measuring speed in miles per hour, an item rest is at a natural 0 miles per hour, as it is not moving. Therefore, not only can we say that 50 mi./h is twice as fast than 25 mi./h we can also say that is greater than 25 mi./h, therefore we are implying both rank and magnitude.

|  |
| --- |
| **Key Facts**  **Ratio Attribute**   * + Rank order and magnitude is implied   + Has a natural zero   + Uses numerical attributes     - Example: 50MPH is twice as fast and greater than 25MPH   + Addition, subtraction, multiplication, and division make sense, as the values are absolute from a natural zero. |

## Attribute Table: Example Data Set

Now you know the four different attribute categories, let’s take a look at an example data set and its related attribute table. Try to identify each column as holding nominal, ordinal, interval, or ratio data.

The data set we are looking at contains four objects, and each one of those objects represents a tree. Each object has four attributes showing in the attribute table: ID, height in feet, type, and class.

### ID Field

In this case, the ID was automatically generated by the GPS unit used to record the location of the trees. It does not represent any meaningful measure of the tree, other than to differentiate one tree from another. Note that the ID is a number. Since it is a number, this does not rule out any of the four attribute categories, as all four attribute categories can hold numbers. Therefore, we need to think about the following questions:

* What does an ID mean?

Let’s think about a driver’s license number. Your driver’s license number is a number that uniquely identifies you as a driver in that state.

* Would it be logical for me to ask you to multiply your driver’s license number by my driver’s license number?

No, that would not produce meaningful results.

* Is it meaningful for me to state that my driver’s license number is greater and better than your driver’s license number?

No, it only makes sense to state that my driver’s license number is different than your driver’s license number. Therefore, if we cannot do multiplication, subtraction, addition, or rank driver’s license numbers, then it just leaves us with the nominal attribute category.

In the case of our trees data set, the ID is equivalent to a driver’s license number. Even though it is a number, it is simply a unique identifier that we used to separate trees from each other, just as we would separate individual drivers using a driver’s license number.

### Height Attribute

If we think about height, it is a natural phenomenon that has a natural zero point: 0 feet. Therefore, as we know we are modeling height, and heights have a natural zero point, we can infer that the height attribute contains attributes of the ratio category.

### Type Attribute

The type attribute contains a description of the type of tree. Since these attributes do not use numbers, we can immediately remove interval and ratio from the list of possibilities. That leaves nominal, and ordinal. We think about what a type represents, it is a description of the type of tree but the question becomes, is there any rank implied when we describe the type of tree. The answer is no, we cannot state that a pine tree is greater than or less than a mesquite tree, all we can compare is whether a pine tree is equal to or not equal to a mesquite tree, therefore, the type attribute holds nominal attributes.

### Class Attribute

It represents a relative measure of the height of the tree. Again, as we’re not seeing numbers in this attribute column, we can rule out interval and ratio attribute categories. This leaves nominal and ordinal. Let’s first consider ordinal for this attribute column.

Question: Logically, can we order tall, medium, and small?

Answer: Yes, we can state, that tall is greater than medium, and medium is greater than small, and small is less than tall. Therefore, this data represents ordinal information.

### Attribute Data Types

Let’s finish this section by discussing attribute data types. Computers fundamentally “think” differently than humans. While humans see numbers, letters, pictures, and sounds, a computer only sees zeros and ones, or binary data. Therefore, we need a way to translate the numbers, sounds, and videos, as humans know it, to a form in which a computer can understand, and store the information. Therefore, computer scientists have created data structures that can be used by us to translate our information into a format which the computer can store in its memory. This data structure is known as a data type.

## Data Types

There are four typical data types that we use in GIS: integer, float/real, text/string, and date. It is important that we specify which data type we are going to use to store information in the computer’s memory, so that we may use the computer’s memory in the most efficient manner. It is important to let the computer know which operations are allowed for each data point stored in that memory location using a specific data type.

### Integers

The first data type is the integer. An integer is a whole number, such as the number one, the number 2458, and the number -54. Integers can be used for mathematical calculations; however, any resulting fraction of a whole number will be rounded, or truncated.

### Float or Real

The float, or real, data type holds a decimal number such as the number 1.452, the number 254,783.1, or -845.157. Like the integer data type, the float or real data type can be used for mathematical calculations. No rounding or truncation will take place when using float or real numbers, depending on the number of significant digits you have specified.

### Text or String

The text, or string, data type contains characters such as character “A”, the characters “GIS”, the characters “125 Main St.”, or the number “9”. Even though the text may contain numbers, it is important to note that they cannot be used for mathematical calculations. However, strings can be manipulated to find substrings, or to cut strings and locations.

### Date

The last common data type is date. The date data type holds time and date information such as 12/10/2010, or 10/12/10, or December 10, 2010. The date data type cannot be used for mathematical calculations however, it can be used to determine and calculate lengths of time between two different dates or times. Additionally, the computer stores the date information in its own internal data structure, but can be formatted to output the date in many different ways, as shown in these examples.

Spatial Data Models

## Types of Spatial Data Models

There are three common spatial data models being used in GIS today: vector, raster, and triangulated irregular network (TIN).

**Vector Data Model**

A vector data model defines discrete objects. Examples of discrete objects are fire hydrants, roads, ponds, or a cadastral. A vector data model is broken down into three basic types: points, lines, and polygons. All three types of vector data are composed of coordinates, and attributes.

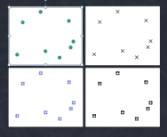
### Points

A point uses a single coordinate pair to define its location. Points are considered to have no dimension even though they may have a real world dimension. For the purposes of a GIS, no dimension is assumed. Each point has associated attribute information, and

the information is attached to the center of the point. Examples

of spatial phenomenon that would be modeled well as points are

light poles, manhole covers, and crime locations.

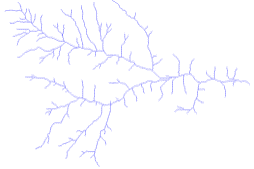
What you see in this image are the different ways to represent airports using a point vector type. Even though the symbols are different, they all represent airports, and, all of the attributes for each airport are linked to the center of each one of these symbols.

### A Line Vector

A line vector type is defined by an ordered set of coordinates. Each line, and curve, is made up of multiple line segments, however, on occasion, curved lines are represented mathematically.

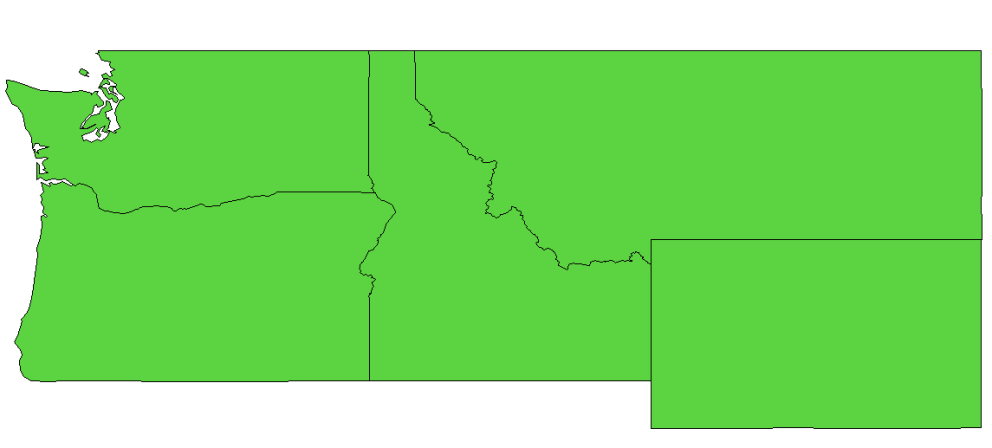
There are two words that we need to define when discussing lines: the node, and a vertex. A node is where a line begins or ends. A vertex is where a line changes direction. The smallest possible line will have two nodes, a start node, and an end node. Longer lines will have at least two nodes, and many vertices in between where the line changes direction. Attributes may be attached to the entire line, individual node, or individual vertices, therefore, each line may have multiple rows of attributes in the attribute table.

For example, if a line represents a road, each road segment between two intersections, may have its own address information, such as the start address, and the end address for that block. An intersection may have an attribute that describes where the intersection has a stop sign, or stoplight. The other option, is for the entire line to have one row of attributes no matter how complex the line. Examples of spatial phenomenon that are modeled well by lines are roads, pipelines, outlines of objects, and power lines.

This image shows a stream network represented by the line vector type.

### Polygon

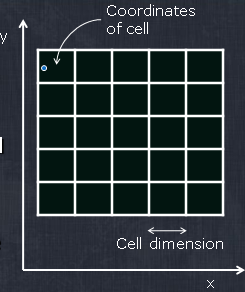
The last vector data type is the polygon. A polygon is formed by a set of connected lines where the start and end point have the same coordinate. Because the start and end point have the same coordinate, the polygon will close and will have an interior region. Attribute information is attached to the center of the polygon no matter how complex the polygon. Examples of spatial phenomenon modeled well by polygons are lakes, cities, tree stands, and political boundaries.



**Raster Data Models**

## What is a Raster Data Model?

The raster data model best represents continuous objects such as temperature or elevation. A raster is a regular set of cells, or pixels in the case of a photograph, or in a grid pattern, which is also referred to as a matrix. Each cell in the raster contains a single value, and the coordinate of each cell of the raster refers to the center of the cell. Therefore, the single value stored in each cell of the raster, applies to the entire cell in the raster matrix. Each cell can be defined by a cell dimension such as the cell width and height. Often, cells in a raster are square, so the cell width and cell height will be the same.



This is an example of a raster. In this case the raster is an aerial photograph showing a portion of the port of Corpus Christi, Texas. Photographs are the most common data sets that are represented by rasters. In fact, the photos that you take with your camera, or camera phone, are stored in a raster format. So for instance, a JPEG, is a raster file stored in the JPEG format. In these pictures, each cell contains a single value that represents the color reflected from the objects.

## Raster Resolution

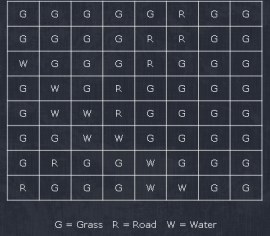
In GIS, it is important that we know the resolution of the raster. The raster resolution is the cell size of each cell of the raster. Unlike how photographers represent resolution, as the number of megapixels their camera uses, in GIS, we are not as concerned about the number of cells, but of how much area on the ground each cell covers. There is a direct trade-off between resolution and file size, the cell coordinate is a center of the point cell, and the coordinate applies to the entire cell area.

Again, it is important to reiterate that each raster cell represents a given area and the value assigned applies to the entire cell area. If there is more than one value, they can fall inside the raster cell, then the raster cell may contain the average, central, most common, or only value covered by the cell.

For instance, let’s look at this picture of the port of Corpus Christi, Texas. If we zoom in to the ship, we will be able to see each individual cell.

You can now clearly see that each cell only contains a single color. Here zoomed into the raster, you can see that the resolution is not infinite, and each cell covers a large location on the ground. Additionally, we know that each cell can only contain a single value, but on the ground multiple colors were reflected. So in this case, the camera has the choice of taking the average color, most common color, or a single color for the cell. When this happens, data is lost through aggregation, and it is important to keep that limitation in mind. While we would like to have a much higher resolution photo of the ship, this would be expensive, and would produce an outrageously large file size, so the trade-off was coverage of a larger area, but at a lower resolution.

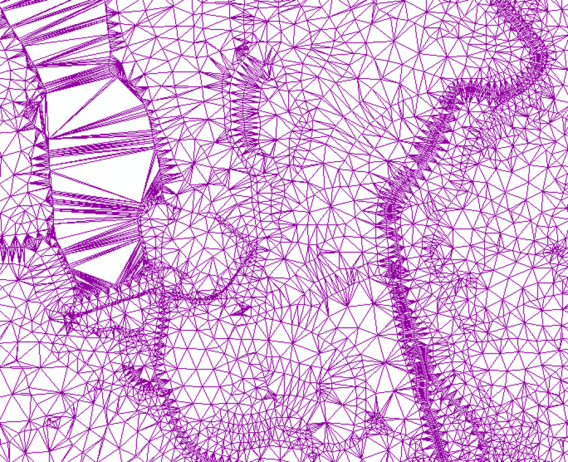
In addition to colors, a raster can store other data values. For instance, this raster contains a single character and each cell, where each character represents grass, road, or water. Again, we can guess that each cell doesn’t contain exactly water, a road, or grass, so we assume that some aggregation has taken place again. But in either case, this raster can store information about the continuous phenomenon, in this case land cover, in a meaningful way.

If we apply colors to the previous Grass, Road, and Water values, we can visualize the data like this. This visualization shows what is covering the earth in each cell of the raster and allows us to simplify our vision of reality to see patterns.

|  |
| --- |
| **Key Facts**  Raster Resolution   * The cell size is the resolution * There is a trade-off between resolution and raster file size * The cell coordinate is the center point of the cell * The coordinate applies to the entire cell area * Each raster cell represents a given area and the value assigned applies to the entire cell * The raster cell value represents the average, central, most common, or only value covered by the cell |

**Triangulated Irregular Network (TIN)**

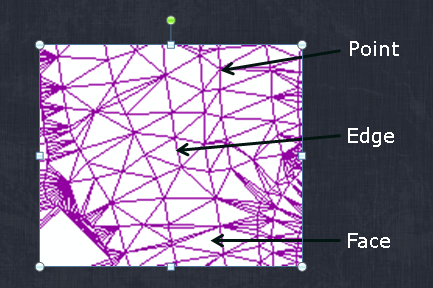
## What is a Triangulated Irregular Network (TIN)?

****A triangulated irregular network, most commonly referred to as a TIN, is a network of triangles connected together to create a 3-D surface where the triangles in this network are not crossing. Triangulated irregular networks are more complex than rasters, however they are more efficient space wise. Additionally, triangular irregular networks can easily accommodate different sampling densities, where rasters cannot. Lastly, triangular irregular networks preserve each input measurement point, again, where rasters cannot.

Triangulated irregular networks can be quite large, and look quite complex. However, if you think about what each point of the triangle represents, in this case, an elevation point, and you know that faces are the flat face of a triangle, you can decipher some of the features in a TIN without seeing it colorized. For instance, the very large triangles in the top left corner of this triangulated irregular network represent a lake behind a dam. The very dense triangles flowing from the top right of the triangulated irregular network to the bottom right of triangular irregular network represents a river.

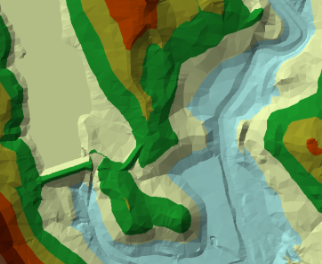
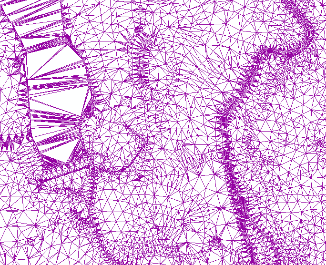
### Anatomy of a TIN

If we look at the anatomy of a TIN, is composed of points, edges, and faces. A point represents an input data value that is preserved, and defines an endpoint for a triangle. The edge is the line drawn between each point which creates the outline of the triangles. The face is the area, or surface, inside each of the triangles.



### Applying Colors

When we apply colors to this triangular irregular network based on the elevation value, it becomes clearer what the triangulated irregular network is representing. It is representing the continuous spatial phenomenon known as elevation. And in this particular case, it shows a lake behind the dam, the surrounding terrain, and the river leading away from the dam.

 ****

## SUMMARY

Geospatial data models represent spatially referenced information in a simplified view. In this lesson you learned about the elements of these models to further your understanding on what is considered, how information is placed, and why. You learned that spatial phenomenon is part of the mapping process, and that attributes are important in understanding geospatial data models. The various types of spatial data models section of the lesson provided insight as to what types of models are used, when they should be used, and why.

## ASSIGNMENTS

1. Lab: Exploring Geospatial Data Models

2. Quiz: Understanding Geospatial Data Models