Course Title: Environmental Data Analytics.
Degree Program: Masters in data science.
Instructor: Mohammad Mahdi Rajabi

UNIVERSITÉ DU LUXEMBOURG

Geospatial Data I

This section provides an overview of Definition and Types of Geospatial Data; Coordinate Systems and Projections; Common Data Formats; Vector vs. Raster Models; and Time in Geospatial Data.

Definition of Geo-spatial Data

Geospatial data (or geographic information) refers to information that has a geographic aspect, meaning it is associated with a specific location on Earth, through coordinates, addresses, or other geographical identifiers. Geospatial data may have **Attributes**, which are additional details about the geographical feature. For example, for a set of coordinates representing a city, attributes might include population, climate data, or economic indicators.

Importance of Geo-spatial Data in EDA

Geospatial data is a fundamental component of environmental data analytics because environmental phenomena are inherently spatial—they vary across different geographic regions and are shaped by location-specific factors like altitude, proximity to water bodies, levels of urbanization etc.

By associating environmental data with specific geographic coordinates or other identifiers, geo-spatial data facilitates the ***analysis of spatial relationships and patterns***, which is crucial for tasks such as mapping environmental changes over time, identifying areas at risk, and optimizing resource management. Additionally, geo-spatial data enables the ***integration of diverse environmental datasets***—such as satellite imagery, sensor readings, and field observations—by leveraging location overlap and proximity, thereby offering a holistic view of environmental conditions.

Geospatial data is a crucial aspect of data science due to its extensive applications across various domains, including supply chain and logistics, transportation, real estate, land use and urban planning, environmental science, mining, retail and consumer analytics, insurance, and more. However, despite its significance, geo-spatial data is often underrepresented in standard data science curricula. Several reasons contribute to this:

- **Specialization Required**: Geospatial data science demands specific expertise in areas like Geographic Information Systems (GIS), spatial analysis, and mapping techniques. These are typically regarded as specialized topics, often covered in detail within dedicated programs or courses. In contrast, general data science curricula tend to focus on more universally applicable topics that can be used across a broad range of domains, leading to less emphasis on niche areas like geo-spatial data.

- **Complexity**: Geospatial data can be more complex to work with compared to other types of data. It requires an understanding of coordinate systems, spatial relationships, and spatial data structures, which are considered advanced topics. This complexity can make it less suitable for inclusion in introductory data science courses, where foundational concepts are prioritized.
- **Software and Tools**: The analysis of geospatial data often depends on specialized software and libraries. These tools are not always part of the standard data science toolkit and learning them requires additional time and resources.
- **Cross-Disciplinary Nature**: Geospatial data science is inherently interdisciplinary, combining elements of geography, computer science, and statistics. This cross-disciplinary aspect may result in geospatial data science being taught in more focused or specialized programs rather than in broader data science programs.

Geospatial Reference

There are several ways to locate or reference positions on Earth:

- **Coordinates**: Geospatial data often relies on coordinates to represent precise locations. A **coordinate system** is essential, as the same numerical coordinates can point to different locations depending on the system used. Without specifying a coordinate system, the data becomes ambiguous or meaningless.
- **Addresses**: Addresses are another common geospatial identifier. Unlike coordinates, they do not require a coordinate system. Instead, an address refers directly to a specific location, which can be converted into coordinates through a process called **geocoding**.
- **Other Geographical Identifiers**: These include postal codes, place names, or administrative boundaries. Like addresses, they don't need a coordinate system but can be translated into coordinates for mapping or spatial analysis when necessary.

Geocoding is the process of converting textual geographical identifiers, such as addresses or place names, into precise geographic coordinates within a specified coordinate reference system.

Note that for data analytics and machine learning, geospatial references typically need to be represented using coordinates (as quantitative measures). Therefore, geocoding is often required to convert addresses or other geographical identifiers into coordinates.

Geo-spatial Coordinates

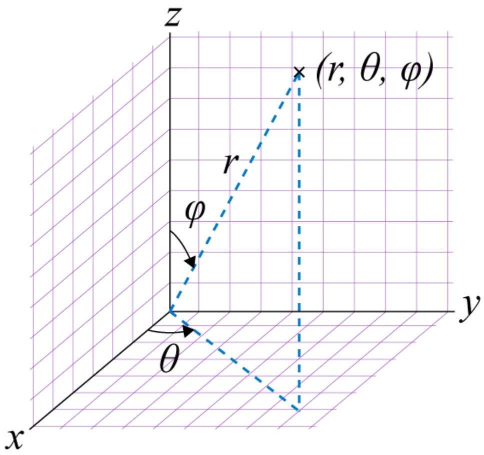
Think of Earth as a 3D space, and if we ignore elevation, as a 2D space. To describe a specific location on Earth, much like any point in space, a set of coordinates is required. These coordinates are defined by an agreed-upon system that includes a reference point (or origin), a method for describing the location, and a unit of measurement. Over the years, many such coordinate systems have been developed by different people and organizations for various purposes. Each system has its own way of specifying locations, depending on the context in which it is used.

The most popular coordinate systems that cover a broad spectrum of applications, are the following two.

1) Latitude and Longitude:

The latitude and longitude system is a global, ***spherical coordinate system*** used to describe locations on Earth's surface using angular measurements. It is the most widely used coordinate system and is based on the concept of the ***Equator*** and ***Prime Meridian*** as reference lines.

- ***Spherical Coordinate System***: A three-dimensional coordinate system where a point's position is a coordinate system defined by three values: the radial distance from a central point (often called the origin), the polar angle (the angle measured from a reference axis, usually the z-axis), and the azimuthal angle (the angle measured in the plane perpendicular to the reference axis, typically from a reference direction in that plane, such as the x-axis).
- ***Equator***: The imaginary line around the middle of the Earth, equidistant from the North and South Poles, dividing the Earth into the Northern and Southern Hemispheres.
- ***Prime Meridian***: The imaginary line running from the North Pole to the South Pole through Greenwich, England. It divides the Earth into the Eastern and Western Hemispheres.



Spherical Coordinate System.

In the Latitude and Longitude coordinate system, any point on Earth is described using two numbers—latitude and longitude. **Latitude** corresponds to the **polar angle**, which measures how far north or south a point is from the Equator (which is 0° latitude), with values ranging from -90° at the South Pole to +90° at the North Pole. **Longitude** corresponds to the **azimuthal angle**, which measures how far east or west a point is from the Prime Meridian (which is 0° longitude), with values ranging from -180° to +180°.

These values can also be expressed using directional indicators. So instead of using positive/negative signs, you might see latitude and longitude expressed with directional indicators (N, S for latitude, and E, W for longitude).

Unlike a full spherical coordinate system, the Latitude and Longitude system does not include the **radial distance**, because it specifically describes locations on the Earth's surface, assuming a constant radius (i.e., the distance from the Earth's center to the surface). Together, latitude and longitude uniquely identify any location on Earth.

Here's an example of how latitude and longitude are used to describe a location. Let's take the Eiffel Tower in Paris, France:

- **Latitude**: 48.8584° N
- **Longitude**: 2.2945° E

In standard databases, these coordinates are often stored in a more compact numerical form, known as decimal degrees, without the directional indicators:

- **Latitude**: 48.8584
- **Longitude**: 2.2945

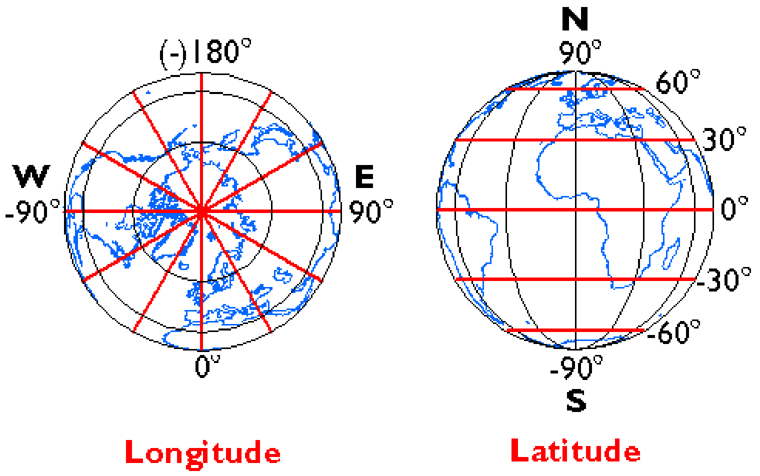
Here's how you can interpret these numbers:

- The latitude of 48.8584° N means the location is 48.8584 degrees north of the Equator.
- The longitude of 2.2945° E means the location is 2.2945 degrees east of the Prime Meridian.

The fractional part of the degree (the digits after the decimal point) is based on the standard decimal system (base-10). In this system:

- The first digit after the decimal represents tenths of a degree.
- The second digit represents hundredths of a degree, and so on.

This format is commonly used in GPS devices, mapping software, and geospatial databases to accurately pinpoint locations. In many standard databases, the coordinates might be stored in separate fields—one for latitude and one for longitude—often with additional metadata describing the coordinate system used (e.g., WGS84, which is the standard for GPS).



Latitude vs. longitude.

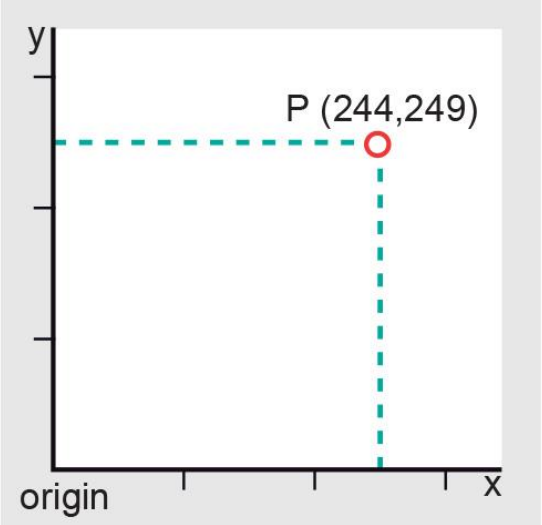
The WGS84 System

While the latitude/longitude system provides a framework for describing locations on Earth's surface, WGS84 ensures the accuracy of these descriptions by precisely defining the Earth's shape and reference points. The Earth is not a perfect sphere but an oblate spheroid, slightly flattened at the poles and bulging at the equator. WGS84 provides a mathematical model (an ellipsoid) that closely approximates this shape, as well as the Earth's center of mass and the orientation of its rotational axis. Additionally, WGS84 extends the system to include elevation, enabling accurate 3D positioning. As a result, WGS84 serves as the global reference system that makes latitude and longitude coordinates accurate and meaningful on a worldwide scale.

2) Universal Transverse Mercator (UTM):

The Universal Transverse Mercator (UTM) system is a global, ***planar coordinate system*** used to describe locations on Earth's surface by dividing it into a series of zones, each of which is mapped onto a flat grid using a transverse Mercator projection. It is widely used for its accuracy over localized areas, particularly in mapping, and surveying.

Planar Coordinate System: Unlike spherical systems, the UTM system projects the Earth's curved surface onto a series of two-dimensional flat grids. This is done by dividing the globe into 60 longitudinal zones, each covering 6 degrees of longitude, and applying a **transverse Mercator projection** to each zone, which maps a cylindrical surface onto a plane.



A Planar Coordinate System.

Transverse Mercator Projection: A map projection where a cylinder is wrapped around the Earth along a meridian (a line of longitude), rather than the equator. This projection minimizes distortion within the UTM zones, particularly along the central meridian of each zone.

Transverse Mercator Projection: A map projection where a cylinder is wrapped around the Earth along a meridian (a line of longitude), rather than the equator. This projection minimizes distortion within the UTM zones, particularly along the central meridian of each zone.

Mercator Map Projection.

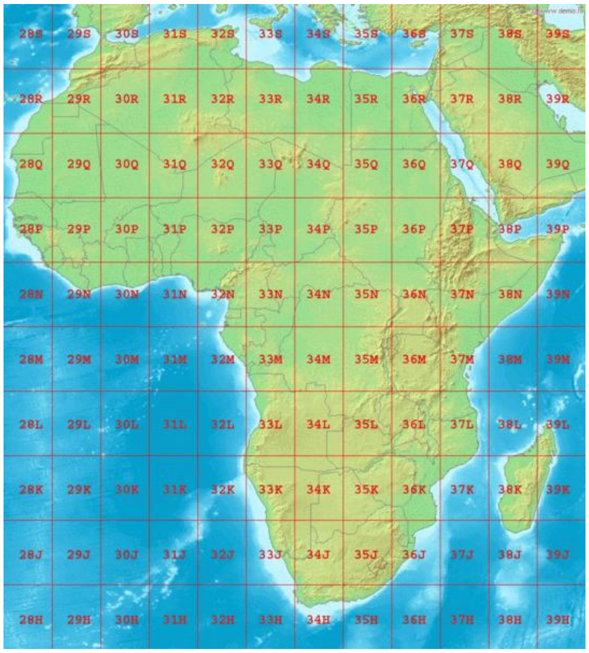
UTM Zones: Each zone in the UTM system has a unique grid and covers a specific region of the Earth, running from 84°N to 80°S latitude. A location's position is described by two coordinates: **easting** (the horizontal component) and **northing** (the vertical component). Easting measures the distance eastward from the central meridian of the zone (with a false easting added to avoid negative numbers), while northing measures the distance from the equator in the northern hemisphere, or from a point 10,000,000 meters south of the equator in the southern hemisphere.

In the UTM system, locations are uniquely described by three components:

- The **UTM zone number**, which identifies the longitudinal slice of the Earth where the location lies.
- The **easting**, which is the distance from the central meridian of the zone (in meters).
- The **northing**, which is the distance from the equator (or false origin in the southern hemisphere, also in meters).

For example, the Eiffel Tower in Paris is located in:

- **UTM Zone**: 31T
- **Easting**: 448251 meters
- **Northing**: 5411932 meters



UTM zones for the African Continent.

Accuracy and Usage: UTM is especially useful in regional mapping and engineering projects because it maintains accurate distance and area measurements within individual zones.

WGS84 in UTM: Like with latitude and longitude, the **WGS84 datum** is often used with the UTM system to ensure accurate geographic positioning by defining the Earth's shape and reference points. Together, UTM and WGS84 provide a highly precise and standardized way of locating any position on Earth within a local zone using metric units.

Common Geo-spatial Data Formats

Apart from point-wise geospatial data, which are commonly used to represent sensor or measurement locations, well locations, GPS points, and more, there are several other types of geospatial data tailored to specific use cases:

Line (Polyline) Data

Polylines represent **linear features or paths** by connecting multiple points and are used to model things like river networks, pollutant dispersion paths, or migration routes in environmental studies. Each line segment in a polyline is defined by at least two geographic coordinates (e.g., latitude/longitude pairs or other coordinate systems). A polyline is typically represented as a sequence of points, with each point having its own set of coordinates, and the lines connect these points in the defined order, creating the overall path.

Example: A river's flow path connecting three monitoring stations tracking water quality:

- Station 1: (Lat: 45.5128, Long: -122.6587) – Upstream water quality sensor.
- Station 2: (Lat: 45.5017, Long: -122.6750) – Midstream water quality sensor.
- Station 3: (Lat: 45.4815, Long: -122.6548) – Downstream water quality sensor.

This polyline represents the river's flow path between the three monitoring stations, helping environmental analysts track water quality changes along the river.

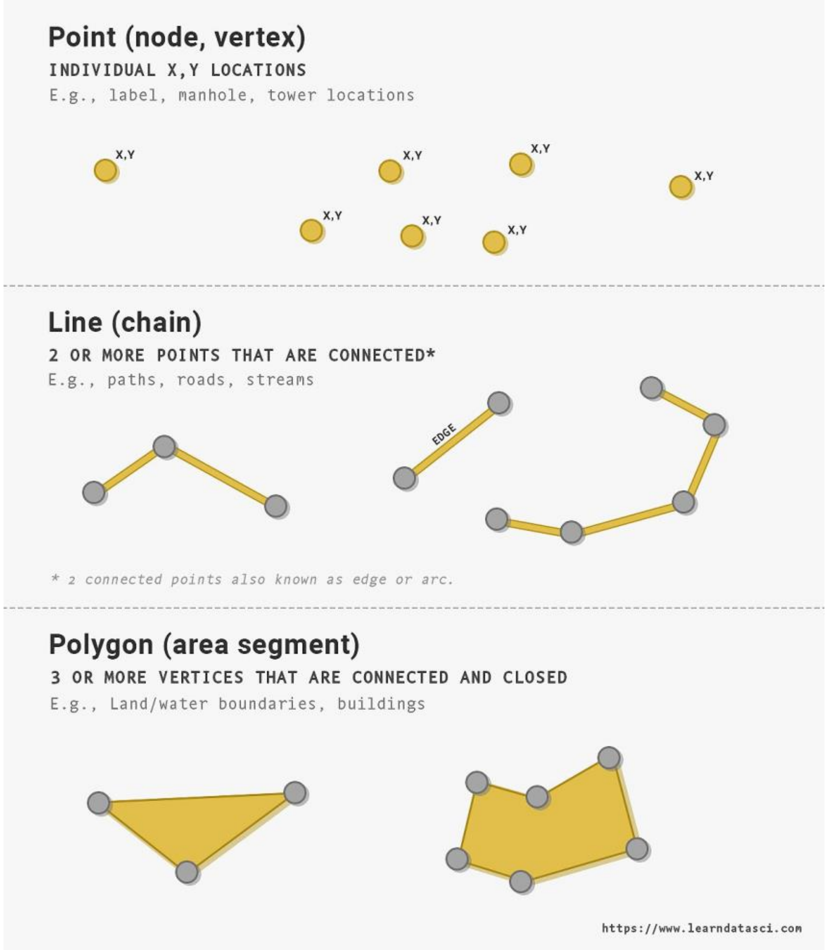
Polygon Data

Polygons represent areas or boundaries by connecting multiple points to form a **closed shape**, commonly used to model regions such as lakes, forests, land parcels, or protected areas in environmental studies. Each polygon is defined by a series of geographic coordinates (e.g., latitude/longitude pairs) that represent its boundary. The first and last points must be the same to form a closed loop, enclosing an area.

Example: A protected forest area defined by four boundary points:

- Point 1: (Lat: 42.0000, Long: -123.0000)
- Point 2: (Lat: 42.5000, Long: -123.0000)
- Point 3: (Lat: 42.5000, Long: -122.5000)
- Point 4: (Lat: 42.0000, Long: -122.5000)
- Back to Point 1 to close the polygon.

This polygon represents a forest area enclosed within the specified boundary points, which can be used to analyze land cover, biodiversity, or environmental protection efforts within the region.


Point (node, vertex)
INDIVIDUAL X,Y LOCATIONS
E.g., label, manhole, tower locations
X,Y
X,Y
X,Y
X,Y
X,Y
X,Y
X,Y
X,Y
Line (chain)
2 OR MORE POINTS THAT ARE CONNECTED*
E.g., paths, roads, streams
EDGE
* 2 connected points also known as edge or arc.
Polygon (area segment)
3 OR MORE VERTICES THAT ARE CONNECTED AND CLOSED
E.g., Land/water boundaries, buildings
https://www.learndatasci.com

Comparison of point-wise, polyline and polygon data.

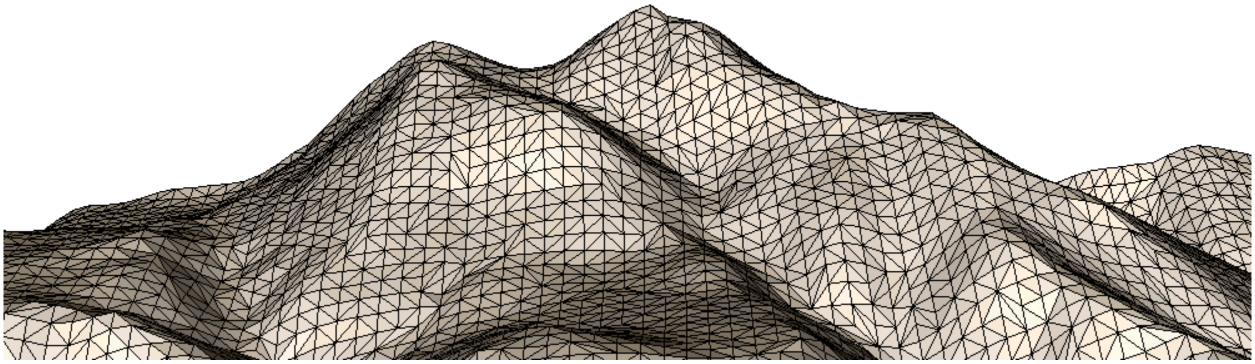
Triangular Irregular Network (TIN)

A **Triangular Irregular Network (TIN)** is a vector-based representation used to model surfaces, typically used in environmental studies to represent terrain, elevation, or other continuous surfaces. A TIN consists of a series of non-overlapping triangles that are formed by connecting points with known elevation values (e.g., latitude, longitude, and elevation). These triangles collectively approximate the surface, making TIN useful for analyzing terrain features like slopes, watersheds, or erosion patterns.

Example: A TIN model of a hilly landscape, defined by three points:

- Point 1: (Lat: 37.7749, Long: -122.4194, Elevation: 30m)
- Point 2: (Lat: 37.7789, Long: -122.4185, Elevation: 50m)
- Point 3: (Lat: 37.7750, Long: -122.4170, Elevation: 40m)

These points form a triangle that represents a small portion of the landscape's surface. Multiple triangles are combined to form a TIN that models the entire terrain. This TIN can be used for environmental applications such as watershed delineation or calculating slope gradients for erosion analysis.



Representation of topography using Triangular Irregular Networks (TINs)

Vector Data

Point, **polyline (line)**, **polygon** and **Triangular Irregular Network (TIN)** are collectively referred to as **vector data** because they represent geographic features as ***discrete*** geometric shapes.

- Vector data is generally more storage-efficient for discrete features compared to raster data, which can be more data-intensive.
- Vector data typically produces cleaner and more aesthetically pleasing maps, especially for features like boundaries and road networks.

Raster Data

Raster Data is a **grid-based representation** of continuous spatial phenomena, commonly used in environmental studies to model data such as temperature, elevation, or land cover. Each cell (or pixel) in the grid represents a specific geographic area and is assigned a value corresponding to the attribute being measured (e.g., temperature in degrees or elevation in meters). Raster data is particularly suited for representing continuous surfaces where values change gradually across space.

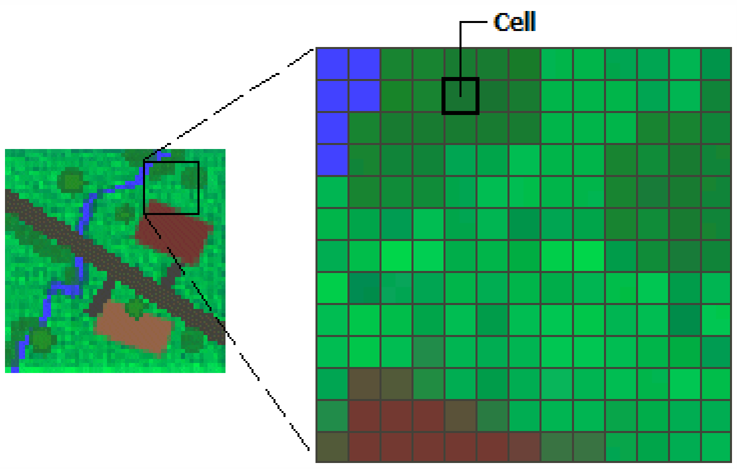
Example: A raster representing soil moisture across a region:

- Each cell in the grid covers a 1 km² area.
- The value in each cell represents the soil moisture percentage at that location, ranging from 0% (dry) to 100% (fully saturated).

This raster data allows environmental analysts to assess spatial patterns in soil moisture across the landscape, helping to identify areas of drought or predict agricultural productivity. Raster data is commonly used for satellite imagery, climate models, and surface analysis.

Raster data is geo-referenced by providing the coordinates of a reference point in the raster image, commonly the upper-left corner of the raster grid. This, along with information about the pixel size (resolution) and the coordinate system, allows the raster to be accurately placed within a geographic space.

Raster data differs from vector data in that it provides a **continuous representation** of spatial information, while vector data represents **discrete features** like points, lines, and polygons.



Raster Data.

Trajectory Data

Trajectory data differs from both typical vector and raster data as it represents the **movement of objects over time**, capturing both spatial and temporal dimensions. While vector data focuses on static, discrete features like points, lines, or polygons, trajectory data tracks the **dynamic movement** of objects, such as animals, vehicles, or weather patterns. Each trajectory is composed of a sequence of time-stamped geographic points, representing the object's changing position as it moves through space over time.

Example: In environmental data analytics, trajectory data can be used to track the migration patterns of wildlife. For example, the movement of a tagged bird might be recorded at various intervals:

- Time 1: (Lat: 45.5017, Long: -122.6750) – 8:00 AM
- Time 2: (Lat: 45.6098, Long: -122.7565) – 10:00 AM

- Time 3: (Lat: 45.7320, Long: -122.8345) – 12:00 PM

By analyzing the bird's trajectory, researchers can study its migratory behavior, travel speed, and preferred habitats over time

Time in Geospatial Data

Apart from trajectory data, which is inherently time-stamped, both raster and vector data can also be time-stamped to represent dynamic, time-varying features.

1. ***Time-Stamped Raster Data***:

- **Dynamic raster data** represents continuous variables (e.g., temperature, precipitation, or air quality) that change over time.
- Each raster "snapshot" corresponds to a specific moment or time period, and by time-stamping each snapshot, you can track how the variable changes over time.
- **Example**: Satellite imagery showing vegetation cover could be time-stamped to observe seasonal changes in forest cover over a year.

2. ***Time-Stamped Vector Data***:

- **Dynamic vector data** involves time-stamping features like points, lines, or polygons to represent changes in their position or attributes over time.
- This is commonly used in scenarios where spatial features evolve, such as shifting boundaries, moving objects, or fluctuating attributes.
- **Example**: Time-stamped polygons could represent the changing boundaries of a floodplain over the course of a storm.

Common File Types for Geo-spatial data

There are many file formats, some specifically developed for **geospatial data** and others designed for storing general **geometric data** that can also be applied to geospatial contexts. Here, we review some of the common formats, though this is not an exhaustive list. It's important to note that none of these formats were developed exclusively for **environmental data analytics**, but they are widely used for geospatial data and are often applied in environmental analyses as well.

Shapefile (.shp extension)

A widely used format for geospatial vector data, developed by Esri. A Shapefile can store basic vector data, including points, polylines (lines), and polygons, along with associated attribute information. **Shapefiles do not support TINs or raster data**. They are typically used for **static data** or **static time-stamped snapshots**, as they do not natively support dynamic or time-series data. A shapefile requires additional files such as:

- **.shx**: Index file to link geometry and attribute data.
- **.dbf**: Attribute data file containing tabular information related to the features.

A **shapefile** is **not human-readable**. The main **.shp** file and its associated files (e.g., **.shx** and **.dbf**) are binary formats, which means they are meant to be read and processed by GIS software, not by humans directly.

GeoJSON (.geojson extension)

A widely used format for encoding a variety of geospatial data structures in **JSON (JavaScript Object Notation)**. GeoJSON can store basic vector data, including points, polylines (LineString), and polygons, along with associated attribute information. Unlike shapefiles, **GeoJSON natively supports time-stamped and dynamic data**, making it suitable for tracking changes over time or representing movement.

GeoJSON files are human-readable and lightweight, making them ideal for **web applications** and **API-based data sharing**. They use a coordinate system based on **WGS84** (World Geodetic System 1984).

GeoJSON files are self-contained and do not require additional index or attribute files like shapefiles.

GeoJSON is not suitable for representing **raster** or **TIN** data.

KML (Keyhole Markup Language, .kml extension)

A widely used XML-based format for representing geospatial data, originally developed for **Google Earth**. KML can store basic **vector data**, including points, polylines, and polygons. KML is also capable of storing **3D geometries** and can include additional styling information (such as colors and icons for points) for better visualization.

KML supports time-stamped and dynamic data, making it suitable for visualizing **moving objects** or changes over time (e.g., animated paths or temporal changes in geographic features). KML can also handle **placemarks**, **network links** (for fetching data dynamically), and **overlays** (like images or raster data, though the raster data itself is not stored directly in KML).

Limitations:

1. **Raster Data**: KML does not directly store raster data, but it can reference external images or overlays that are draped over geographic features.

2. **TIN Data**: KML does not natively support **TINs** or topological data structures like triangulated surfaces, making it unsuitable for detailed 3D surface modeling as required for TIN data.

KML is widely used for **web-based applications** and **virtual globes** like Google Earth, where visual representation and ease of sharing geospatial data are priorities.

GeoTIFF (.tif or .tiff extension)

GeoTIFF is a widely used format for storing **raster data** that includes geographic or spatial information. It is an extension of the standard TIFF format (Tagged Image File Format), commonly used for images, with additional metadata that allows the raster data to be **geo-referenced**. GeoTIFF can store data like satellite imagery, digital elevation models (DEMs), or any grid-based data, with each pixel corresponding to a specific geographic location.

Key Features:

- **Geo-referencing**: GeoTIFF includes metadata such as the coordinate reference system (CRS), origin (e.g., upper-left corner coordinates), pixel size (resolution), and transformation parameters, which allow the raster image to be accurately placed on a map.
- **Continuous or Categorical Data**: GeoTIFF can represent both **continuous data** (e.g., elevation, temperature, or rainfall) and **categorical data** (e.g., land cover types, soil classifications).
- **Multiple Bands**: GeoTIFF supports storing multiple bands of data within a single file, which is useful for applications like remote sensing (e.g., RGB bands, infrared bands).

Limitations:

1. **Vector Data**: GeoTIFF is designed for **raster data** and cannot store vector data (points, lines, polygons).
2. **TIN Data**: GeoTIFF cannot represent **TIN** data, as it is a grid-based format and does not handle topological relationships between points, which are necessary for TINs.

OBJ (Wavefront OBJ, .obj extension)

OBJ is a widely used format for representing **3D geometric data**, including meshes, surfaces, and **triangular irregular networks (TINs)**. The **OBJ** format is simple and human-readable, using plain text to define the vertices, edges, and faces that make up a 3D model. It is widely supported in both 3D modeling software and programming environments (like Python), making it a common choice for applications involving 3D geometry.

Limitations:

1. **No Georeferencing**: Unlike formats such as **GeoTIFF** or **Shapefile**, OBJ does not inherently store geographic information (coordinate systems or spatial references). However, it can be used in conjunction with other files or systems to provide spatial context.
2. **No Metadata for Time or Attributes**: OBJ focuses purely on 3D geometry and does not support time-stamping or attribute data. For dynamic data or additional attributes, other formats (like **GeoJSON** or **KML**) may be more appropriate.

Example Use Cases:

- **3D Terrain Models**: OBJ is frequently used to represent **terrain surfaces** as TINs in 3D, where each face of the triangle defines part of the surface.