**EpiDataFuse Quickstart**

EpiDataFuse allows researchers to easily build pipelines for **collection, fusion and integration** of **spatio-temporal** (collected across both space and time) data for an ample of use cases such as online machine learning and for continuous prediction using computational models.

## A. Introduction to Spatio-temporal data and terminology

### 1. Spatio-temporal data

Spatio-temporal data is collected across both **space** and **time**. Spatial refers to space and temporal refers to time. A spatio-temporal data point is only valid in a particular spatial context and a temporal context. For instance, consider the following data frame which depicts the new Coivd-19 cases in each **WHO region** from **October 26th to November 9th** of 2020.

| | WHO region | New covid cases | Date |
|---|---|---|---|
| 0 | Americas | 1031573 | Nov 2 |
| 1 | Europe | 2043493 | Nov 2 |
| 2 | South East Asia | 381442 | Nov 2 |
| 3 | Eastern Mediterranean | 214072 | Nov 2 |
| 4 | Africa | 33687 | Nov 2 |
| 5 | Western Pacific | 31370 | Nov 2 |
| 6 | Americas | 1450998 | Nov 9 |
| 7 | Europe | 1892039 | Nov 9 |
| 8 | South East Asia | 373786 | Nov 9 |
| 9 | Eastern Mediterranean | 238492 | Nov 9 |
| 10 | Africa | 40990 | Nov 9 |
| 11 | Western Pacific | 32973 | Nov 9 |

Figure 1.1: Data frame for covid cases
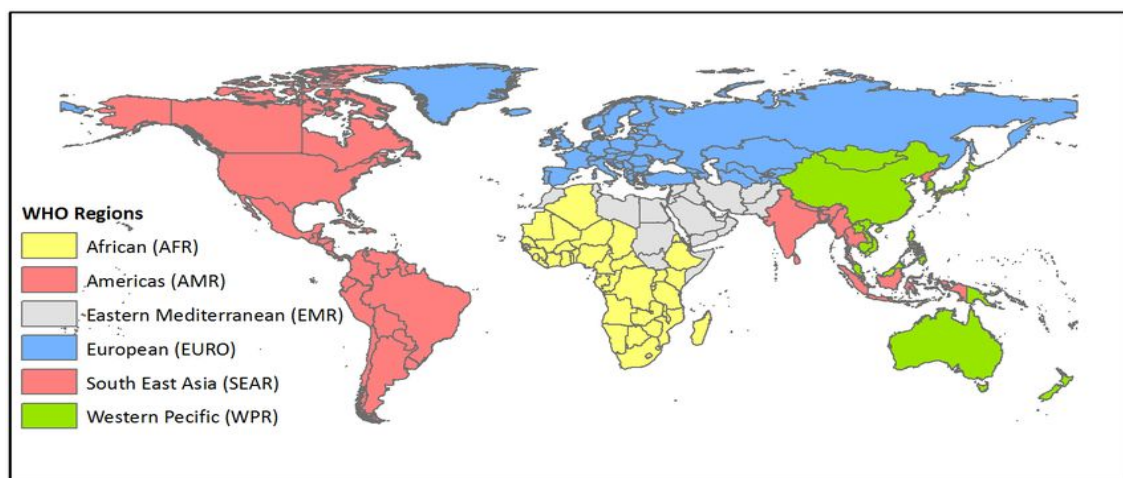


Figure 1.2: WHO Regions

As depicted above record 0 is only valid in the **Americas WHO region** and record 1 is valid only in **Europe WHO region**. However, record 0 and 1 are both valid in the **week ending on Nov 2nd**. In contrast, record 0 and record 6 is valid in the **Americas WHO region**, but have different temporal contexts. The spatial and temporal context above is technically termed as granularity. In the above scenario, the number of covid-19 cases is the feature whereas the WHO region and the cumulative week(from Monday to Sunday) are respectively spatial and temporal granularities of the feature.

In the above instance, Americas, Europe, South East Asia, Eastern Mediterranean Africa and Western Pacific are granules of **WHO region** spatial granularity, whereas **Week 1(Nov2)**, **Week 2(Nov 9)** are granules of **Week** temporal granularity.

Above example depicts is only one type of spatio-temporal data where spatial locations/regions are treated as objects and measurements are collected over time to form features. This type of data is also known as **Raster data**. Event data, Trajectory data and Point reference data are the other types of data which are summarized in a survey done by Atluri et al. EpiDataFuse currently provide fusion mechanism for only raster data.

## 2. Spatio-temporal granularity

As described above, spatio-temporal granularity is the spatial and temporal granularities(context of validity) of a particular data point. Further, multi-granular data are data which can be represented in multiple granularities. For instance, in the above example, the number of covid-19 cases is represented in **WHO region spatial granularity** which is an aggregation of covid cases of **countries**, where **WHO region** is the coarser granularity and **country** is the finer granularity. Further, **Country** is a granularity which is an aggregation of **Districts or States** which can be further divided into locations of hospitals, labs, etc. As such, coarser granularities can be formed by aggregating finer granularities and finer granularities can be formed by dividing coarser granularities. The same applies to the temporal domain as well, which can easily be understood by the domain hierarchy of the temporal domain which is nanoseconds, microseconds, milliseconds, seconds, minutes, hours, days, etc.

## 3. Spatial granularity
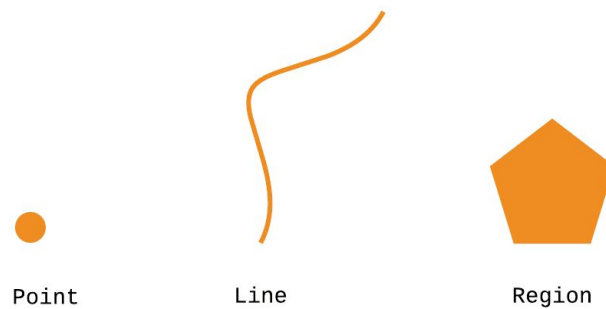


In the spatial domain, granularity can be defined in terms of geometry using a spatial object type. The following table summarizes the primitive types of spatial objects with corresponding example spatial granularities.

| Spatial object type | Spatial granularity |
|---|---|
| **Point** | 1. Airports<br>2. Weather stations<br>3. Train Stations<br>4. Etc. |
| **Line** | 1. Roads<br>2. Flight routes<br>3. Train tracks<br>4. Migration patterns of Animals<br>5. Etc. |
| **Polygon** | Geopolitical regions and Administrative boundaries<br>1. Continent<br>2. Country<br>3. WHO Region<br>4. MOH Area<br>5. Etc. |

Multi-point, Multi-Line, Multi-Polygon are extensions of these primitive spatial objects that also used to define spatial granularities. For each granularity, there is a set of **granules**. For instance, if we consider **Airports in the world** as a granularity the set of airports around the globe forms the granules set for that particular granularity.

- Los Angeles International Airport
- Denver International Airport
- John F. Kennedy International Airport
- Etc.

One other example of granules is the Americas, Europe, South East Asia, Eastern Mediterranean Africa and Western Pacific regions depicted above in figure 1.2 to form MOH Region granularity.

## 4. Temporal granularity

In the temporal domain, the granularities can be considered the standardized units of measuring time which is known as the concept hierarchy of the temporal domain. Further, custom temporal granularities can be formed using standard units of measurement for the requirements of the application. Following table depicts some of the temporal granularities and granules of each granularity.

| Temporal granularity | Granules |
|---|---|
| Second | 1, 2, 3, … , 60 |
| Minute | 1,2,3, … ,60 |
| Hour | 1,2,3, … , 24 |
| Day | 1,2,3, … , 31 |
| Week | Mon, Tue, … , Sun |
| Month | Jan, Feb, … , Dec |
| Quater | 1,2,3,4 |
| Year | …, 2000, 2001, 2002, … , 2020, ... |
| Season | Spring, Summer, Fall, Winter |

When spatial and temporal granularities are combined it creates a spatio-temporal granularity which makes spatio-temporal data comparatively complex. Moreover, the spatio-temporal granularity is heterogeneous from feature to feature. Due to this heterogeneity, it very difficult to integrate data points of different features to generate new combined or summarized data points. The following section describes the process of integration of spatio-temporal data.

### 6. Fusing and Integrating spatio-temporal data points

Fusion and integration of data points are performed to generate consumable data points. For instance, consider two data points **a1** and **b1** belonging to two different features A and B. We can generate a new data point **c1** by Integrating a1 and b1. Thereafter, the generated data point **c1** may be consumed by a computational model. However, integrating two or more spatio-temporal data points is a complex process and which consumes a lot of time and effort. Consider the following two spatio-temporal data points.