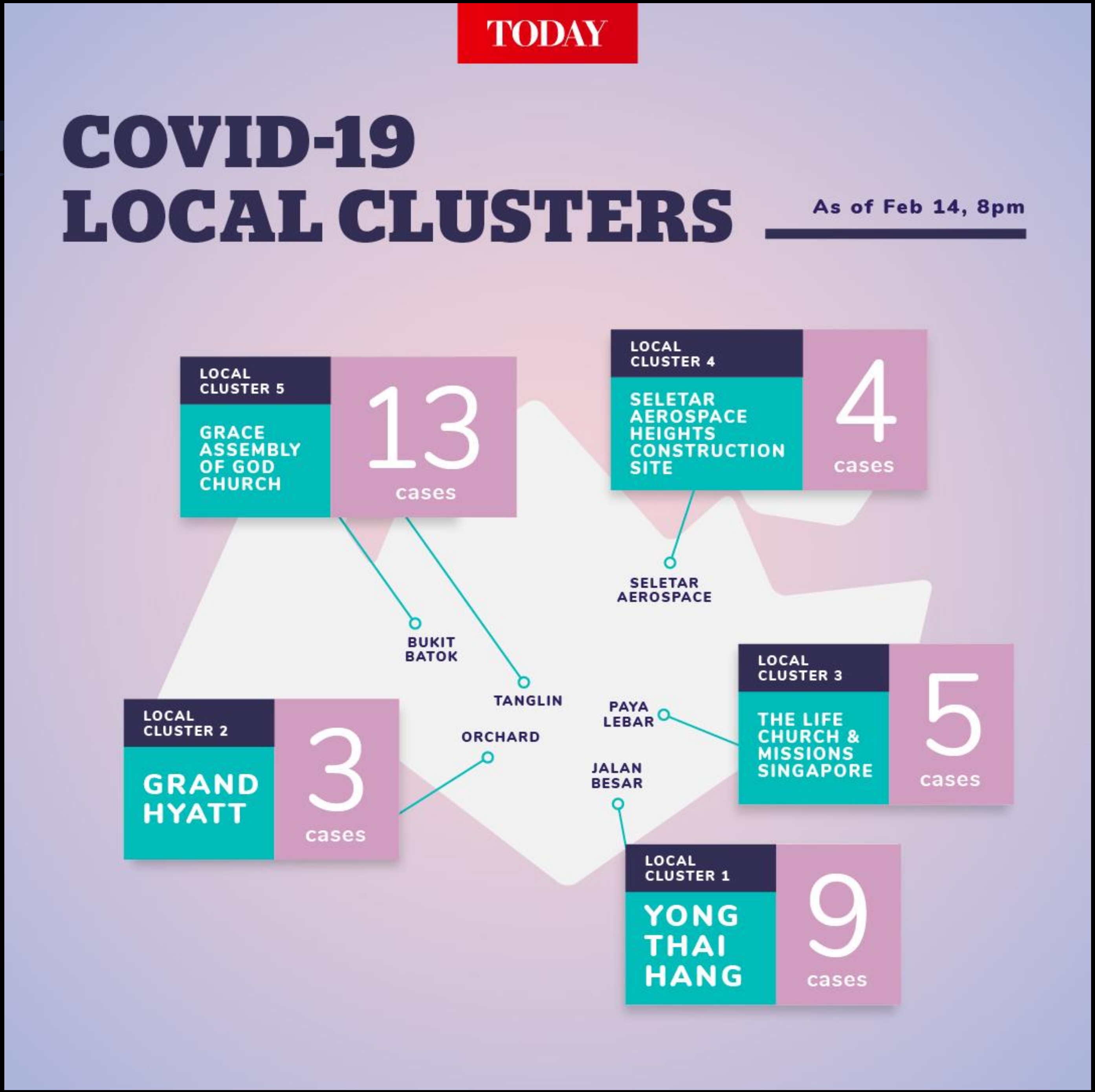
ECON6027 1a

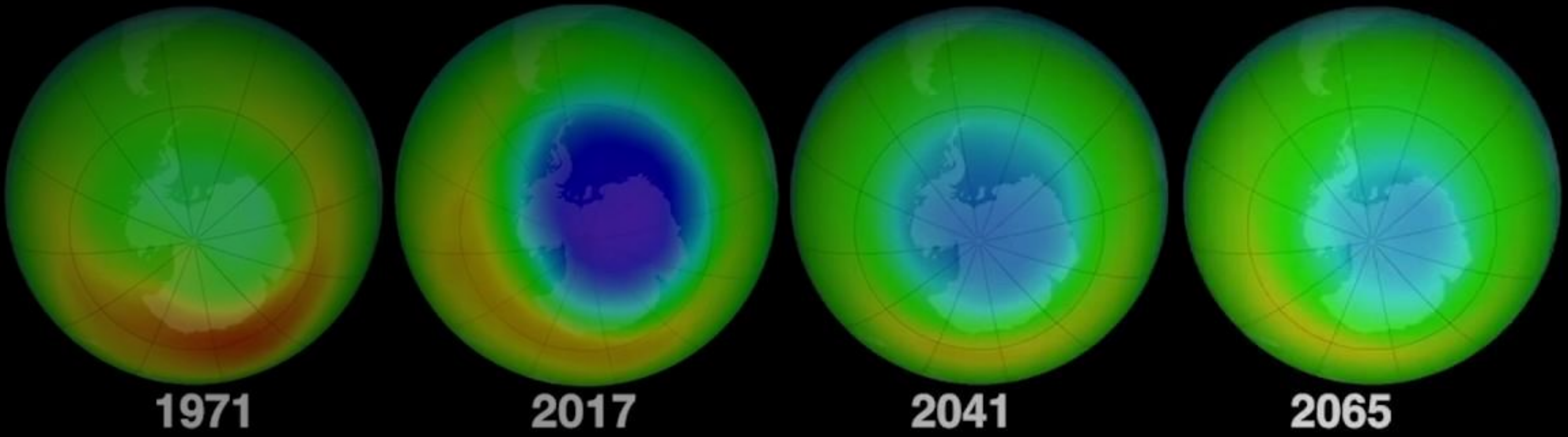
INTRODUCTION

Why Spatial Data?

Tobler's first law of geography (1970):

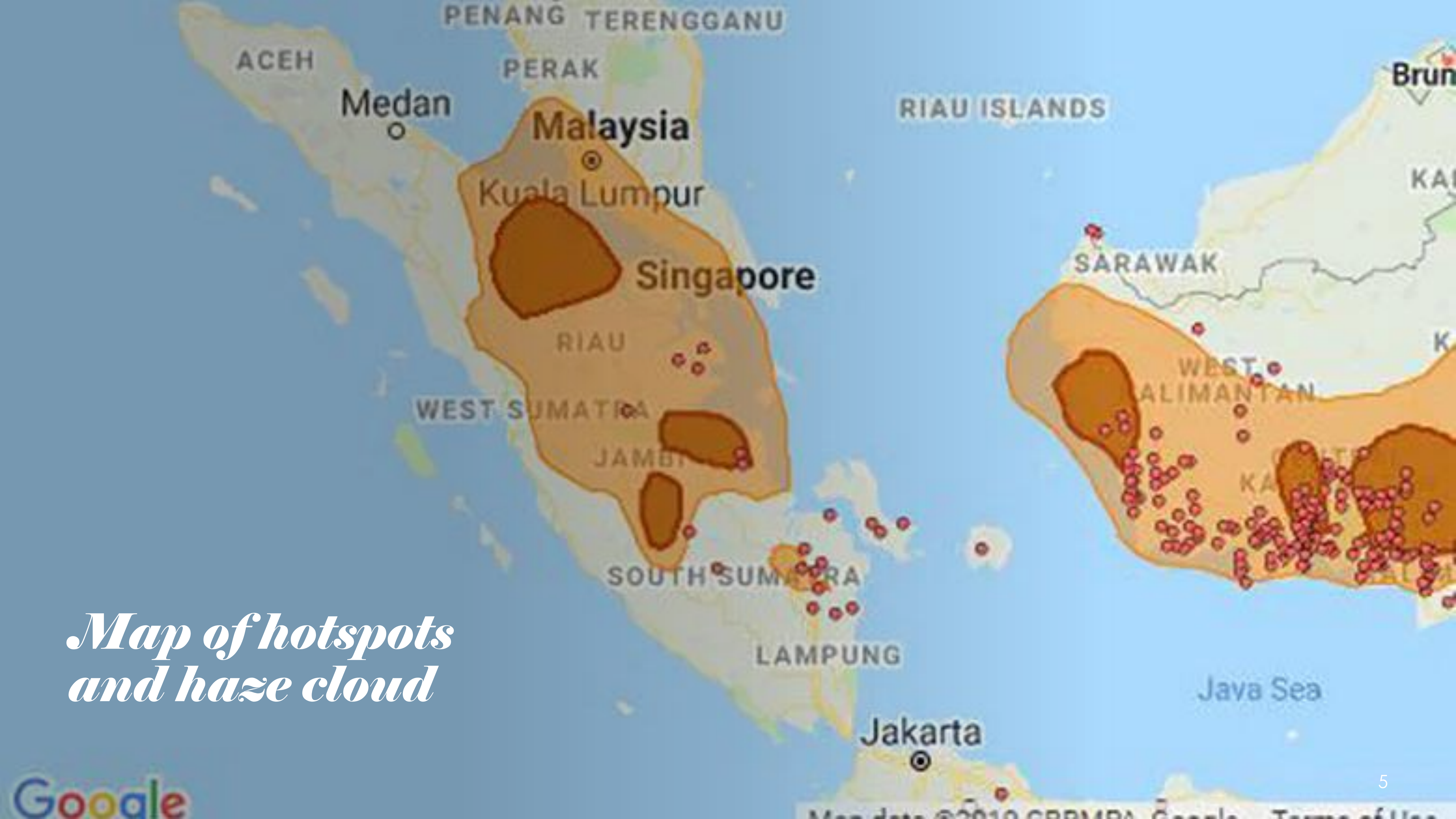
"everything is related to everything else, but near things are more related than distant things"


TODAY
COVID-19
LOCAL CLUSTERS
As of Feb 14, 8pm
LOCAL CLUSTER 5
GRACE ASSEMBLY OF GOD CHURCH
13 cases
LOCAL CLUSTER 4
SELETAR AEROSPACE HEIGHTS CONSTRUCTION SITE
4 cases
SELETAR AEROSPACE
BUKIT BATOK
TANGLIN
ORCHARD
PAYA LEBAR
JALAN BESAR
LOCAL CLUSTER 3
THE LIFE CHURCH & MISSIONS SINGAPORE
5 cases
LOCAL CLUSTER 2
GRAND HYATT
3 cases
LOCAL CLUSTER 1
YONG THAI HANG
9 cases



Healing ozone layer

Map of hotspots and haze cloud



CBD Singapore

Spatial and Spatio-temporal data (location stamped data) are used everywhere!

media

personal devices

paper maps

our own senses/memories

What is Spatial Data?

All spatial data consist of *positional information*, answering the question **“where is it”** (on Earth, body, Sun, moon, etc.)?

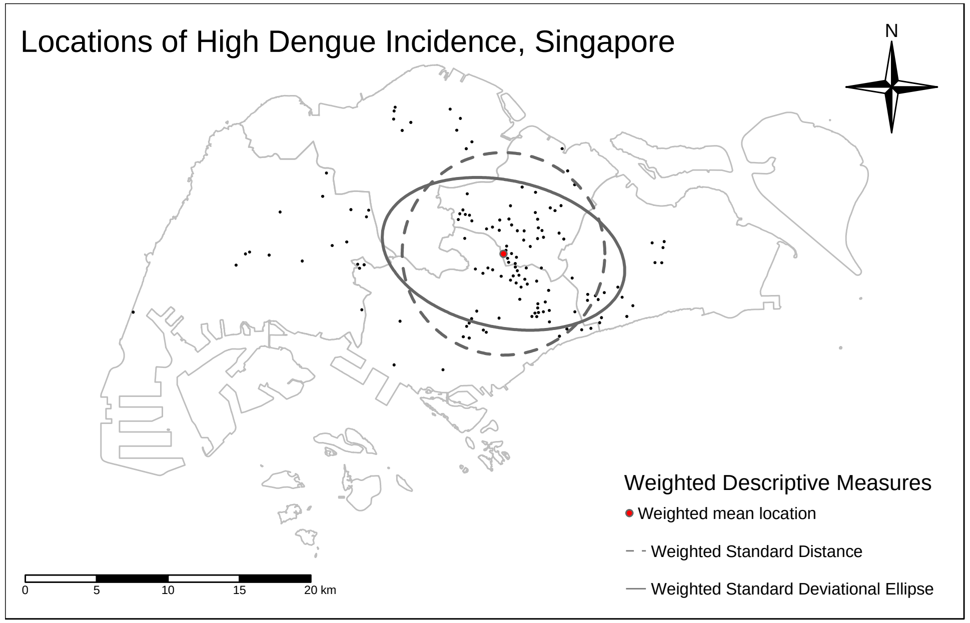
Many empirical data contain not only information about the attribute of interest (i.e. the response/variable being studied), but also other variables that denote the **geographic location** where the response was observed.

Spatial data have a spatial reference: they have **coordinates** and a system of **reference** for those coordinates (a.k.a coordinate reference system, CRS).

Eg: Locations of volcano peaks on Earth. We could list the coordinates for all known volcanoes as pairs of **longitude/latitude** decimal degree with respect to the **prime meridian** at Greenwich and zero latitude at the **equator** (known as The **World Geodetic System** - WGS84).

Key feature of spatial data

One of the key features of spatial data is the **auto-correlation** of observations in space. Observations in close spatial proximity tend to be more similar than for observations that are more spatially separated.



Spatial data analysis was simultaneously developed by many disciplines…

Outside mainstream statistics in geology, geography, meteorology and other subject areas.

Traditional statistical areas such as linear model response surface theory: spatial statistics

Spatial Econometric approaches

Thus, the main challenge for us is to combine these varied tools to suit our purposes!

Types of Spatial Data:

THEORETICAL CLASSIFICATION

Types of Spatial Data: Theoretical Classification

Denote a spatial process in d dimensions as:

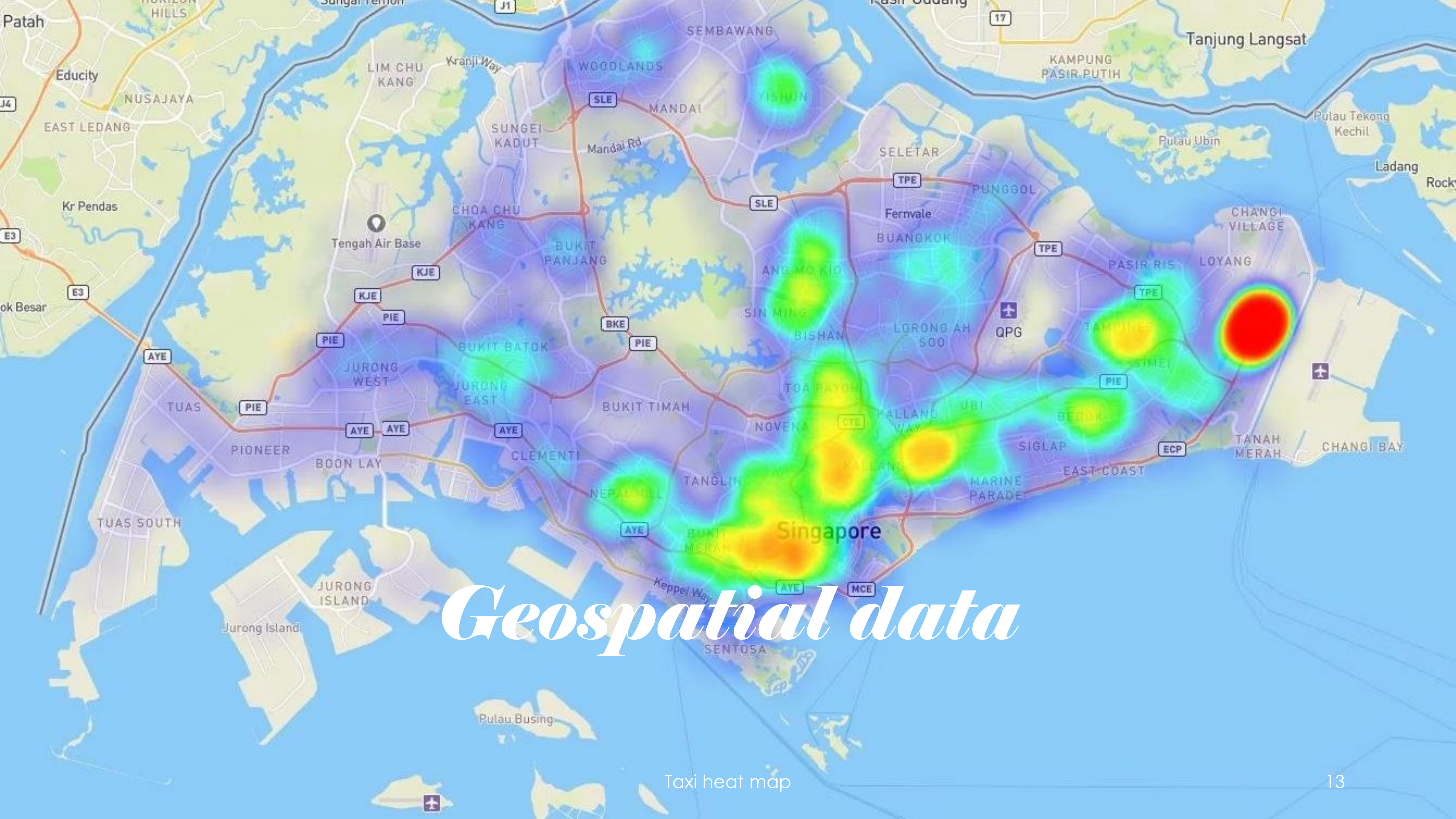
$$\{Z(\mathbf{s}) : \mathbf{s} \in D \subset R^d\}$$

where Z is the observed attribute at location **s**, a (d × 1) vector of coordinates. The spatial data types are distinguished through characteristics of the domain *D*.

1. Geospatial/Geostatistical/Earth data

2. Point patterns

3. Areal/lattice/regional data



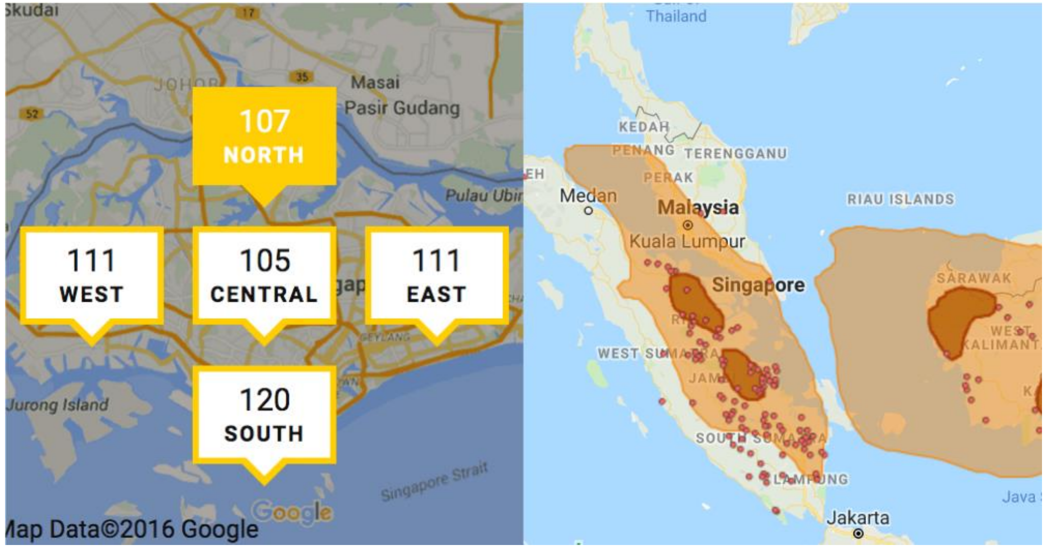
Geospatial data

Taxi heat map

Geostatistical data

Domain ***D* is continuous** s.t. Z(**s**) can be observed anywhere in *D* . i.e., between any two sample locations, you can theoretically place an infinite number of other samples.

- For example, consider measuring air temperature or PSI value, which at least in theory, can be recorded at any location, however, we only observe a finite number of observations.
- Other examples: Ozone layer concentration of a certain mineral, ground commination levels, etc.



Geostatistical data

- Due to the continuity of *D*, Geo-statistical data are also known as **"spatial data with continuous variation".** However, keep in mind that continuity is associated with the domain and not the attribute itself (which may be discrete or continuous or even nominal or ordinal).
  - Example: temperature can be measured using a Celsius scale (continuous) or an ordinal scale (discrete).
- Since the spatial domain is continuous, it cannot be sampled exhaustively and an important task in the analysis is the **reconstruction of the surface** of the attribute Z over the entire domain, i.e., mapping of Z(**s**).

Point patterns

Point patterns

The important feature of a point pattern is the **random domain** and the attribute itself will be **degenerate/binary** in nature (it is there or not there!).

Example: locations of lightening strikes, locations at which weed emerge in a garden, locations of lunar craters, etc.

If along with the location of an event, if we observe a stochastic attribute, then is it called a “**marked**” pattern. For example if we observe the size of the lunar crater along with the location.

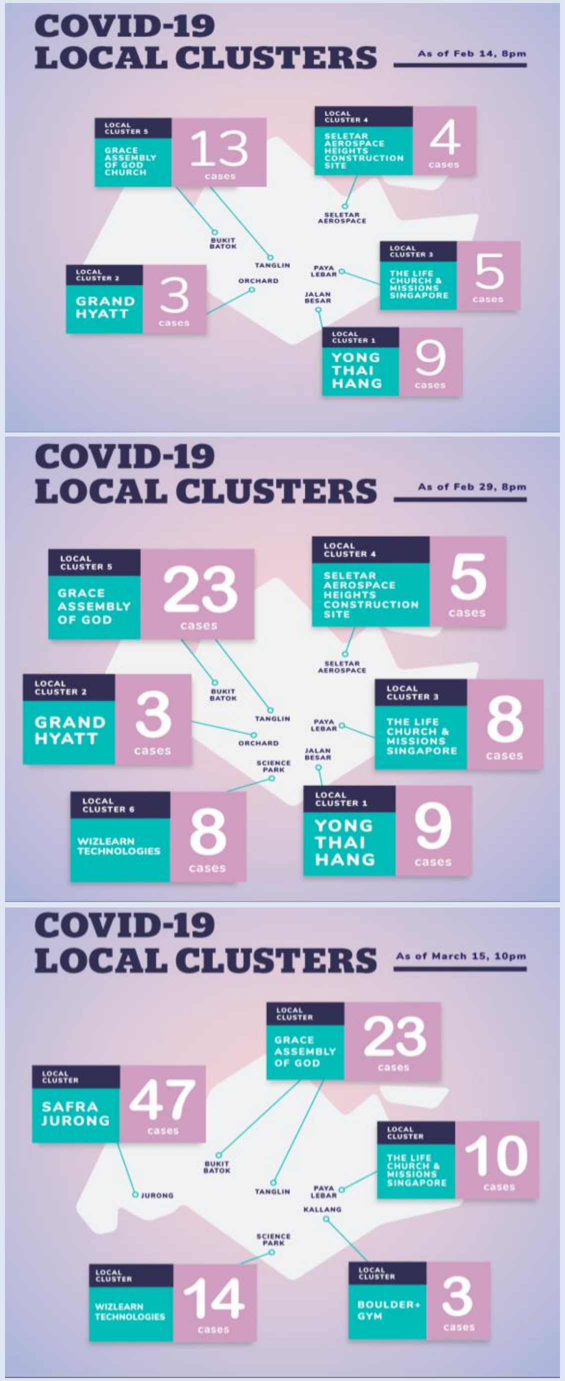
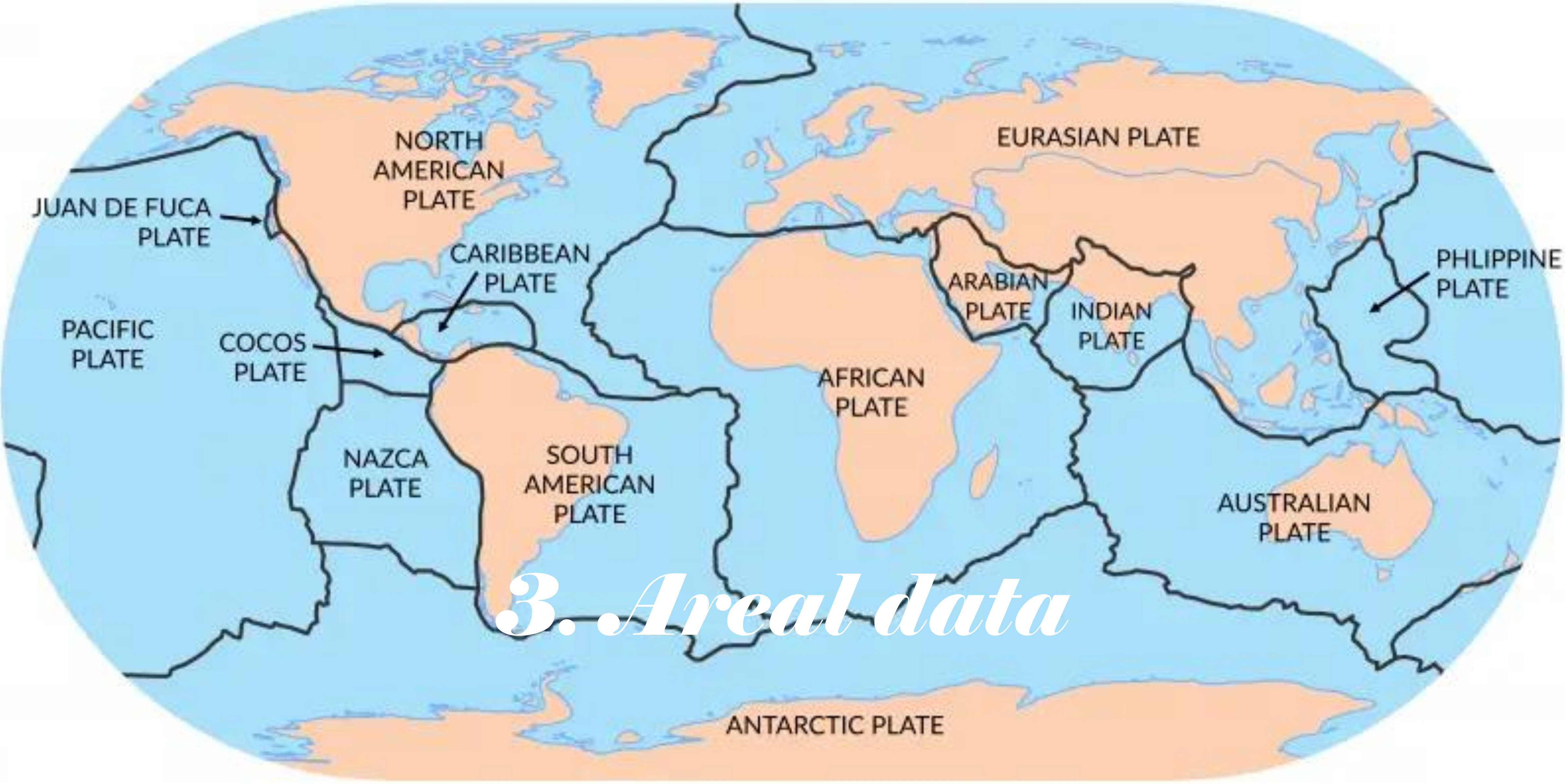


Image source: https://www.todayonline.com

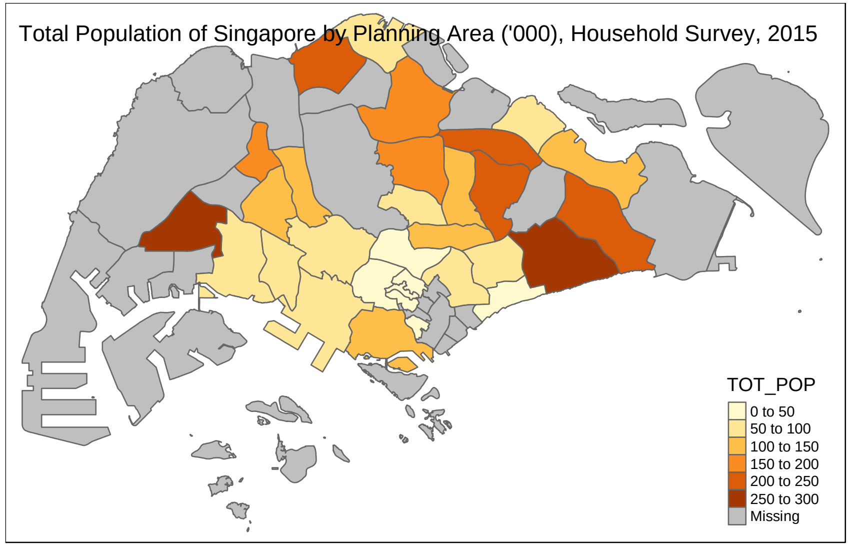
3. Areal data



Tectonic plates

Areal/lattice/regional data

- These are spatial data where the **domain D is "fixed and discrete"** (non-random and countable).
  - Eg: postal codes, GRCs, planning areas, remotely sensed data reported by pixels (such as data coming from satellites).
- Spatial locations with areal data are often referred to as "**sites**" or "**areal units**".
- One of the main differences between point data and areal data is that, in practice areal data are **spatially aggregated** over areal regions. (Mathematically this refers to an integration of a continuous spatial attribute).
  - yield measures on an agricultural plot
  - event counts (such as deaths, crimes, voter turnout, etc.) for various sites (such as postal codes, regions, states, etc.)



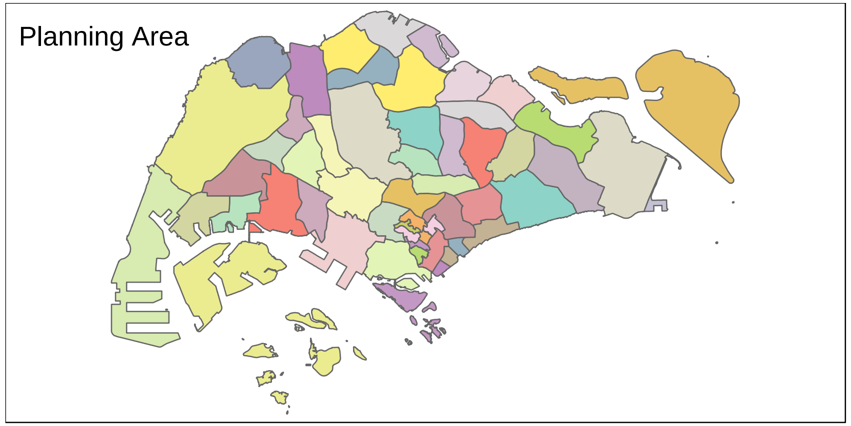
Areal/lattice/regional data

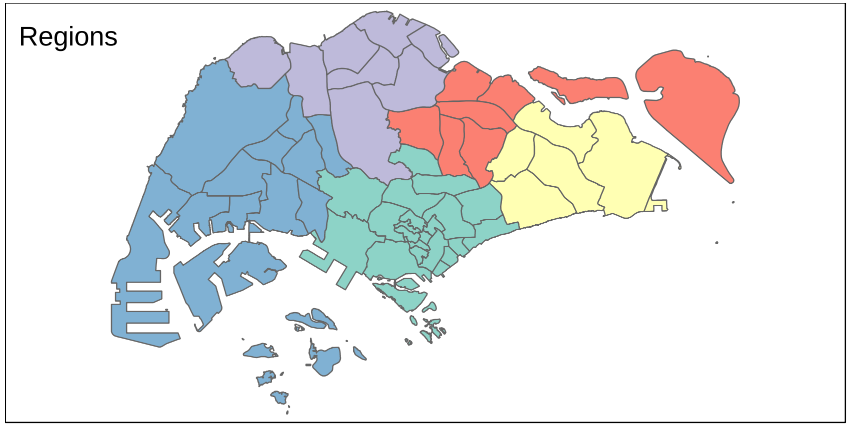
- If areal units are **irregular**, a more precise term would be "regional data".
- If areal units are **regular**, a more precise term would be "lattice data".
- Given the discrete nature of the collection of sites, areal data can be **exhaustive** (another differentiating feature compared to point data or geospatial data).
  - For example voter turnout data provide the number for every electoral unit and the issue of predicting the number for any other are does not arise.

Areal/lattice/regional data: MAUP

Modifiable Areal Unit Problem:

Coined by geographers during the 1970s, the modifiable areal unit problem (MAUP) is one of the most **stubborn problems** in spatial analysis when spatially aggregated data are used. Data tabulated for different spatial scale levels or according to different zonal systems for the same region will not provide consistent analysis results.





The statistical methodology to be applied will inherently depend on the type of spatial data that we have…

THE DIFFERENT STATISTICAL TECHNIQUES WILL BE COVERED FROM CHAPTER 5 ONWARDS

Types of Spatial Data:

PRACTICAL CLASSIFICATION

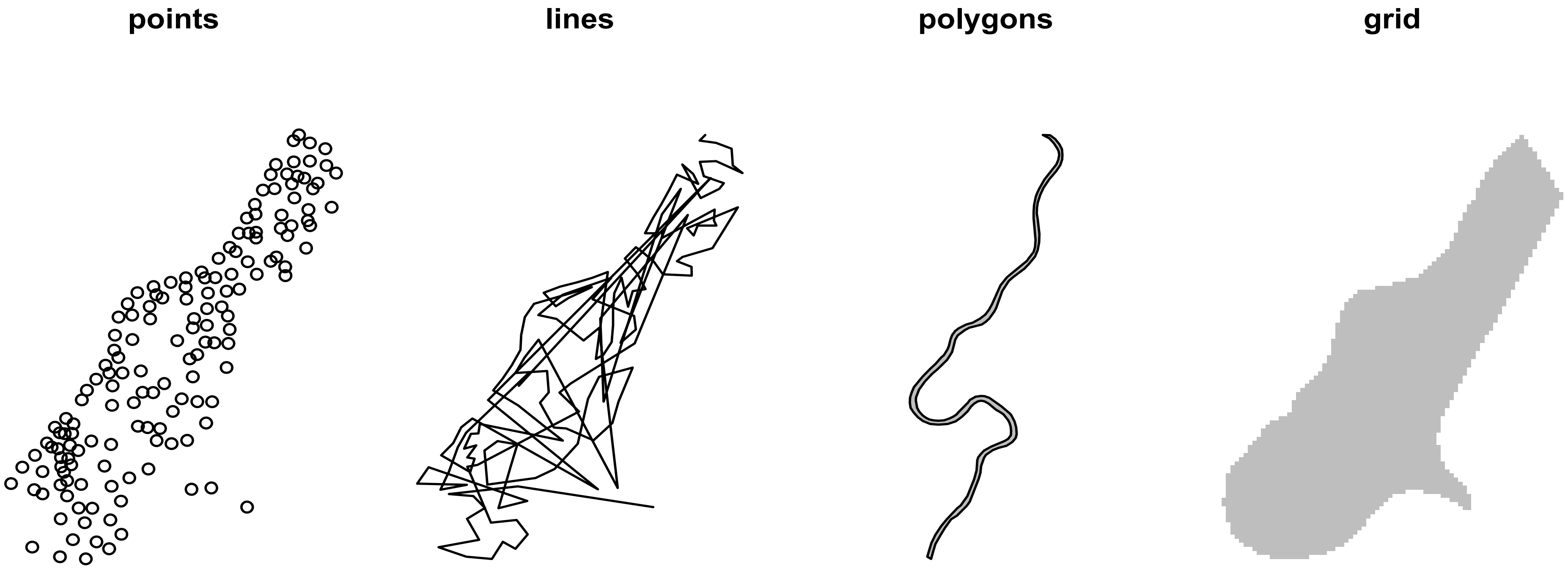
Types of Spatial Data: Practical Classification

This is how spatial data are "classified" and contained in software programs(such as R, ArcGIS, etc) that handle spatial datasets.

1. **Points**: a single point location, such as a GPS reading or a Geo-coded address.
2. **Lines**: a set of ordered points, connected by straight line segments, for example, the contour lines that shows altitude of a certain mountainous region, road network, river network, etc.
3. **Polygons**: an area marked by one or more enclosed lines such as administrative regions, for example, collection of islands, planning areas, regions, GRCs, etc.
4. **Grids/raster**: a collection of rectangular cells organised in a regular lattice such as remote sensing instruments that register data on a regular grid.

The first three (points lines and polygons), are collectively known as **vector** data.

Types of Spatial Data: Practical Classification

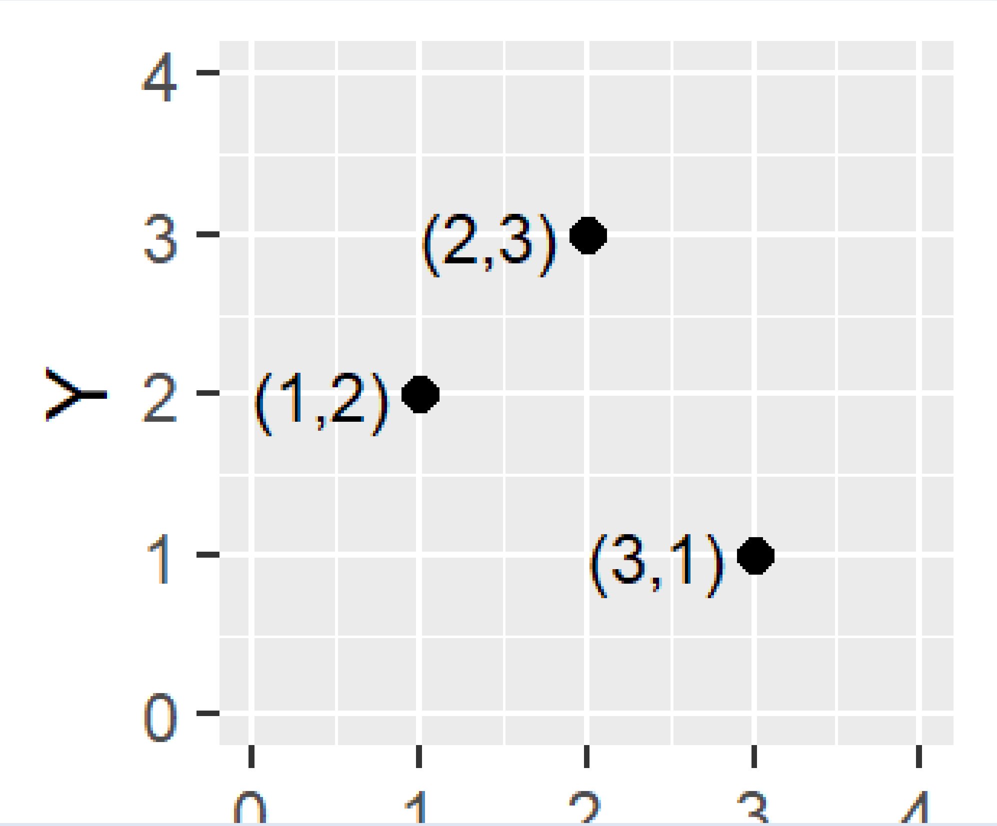


Vector Data (points, lines, polygons)

- The geographic vector data model is based on observations located within a coordinate reference system (CRS).
- Observation can represent self-standing features (e.g., the location of a bus stop) or they can be linked together to form more complex geometries such as lines and polygons.
- Most vector geometries contain only 2-dimensions (3-dimensional CRSs contain an additional z value, e.g.: height above sea level, depth, etc).

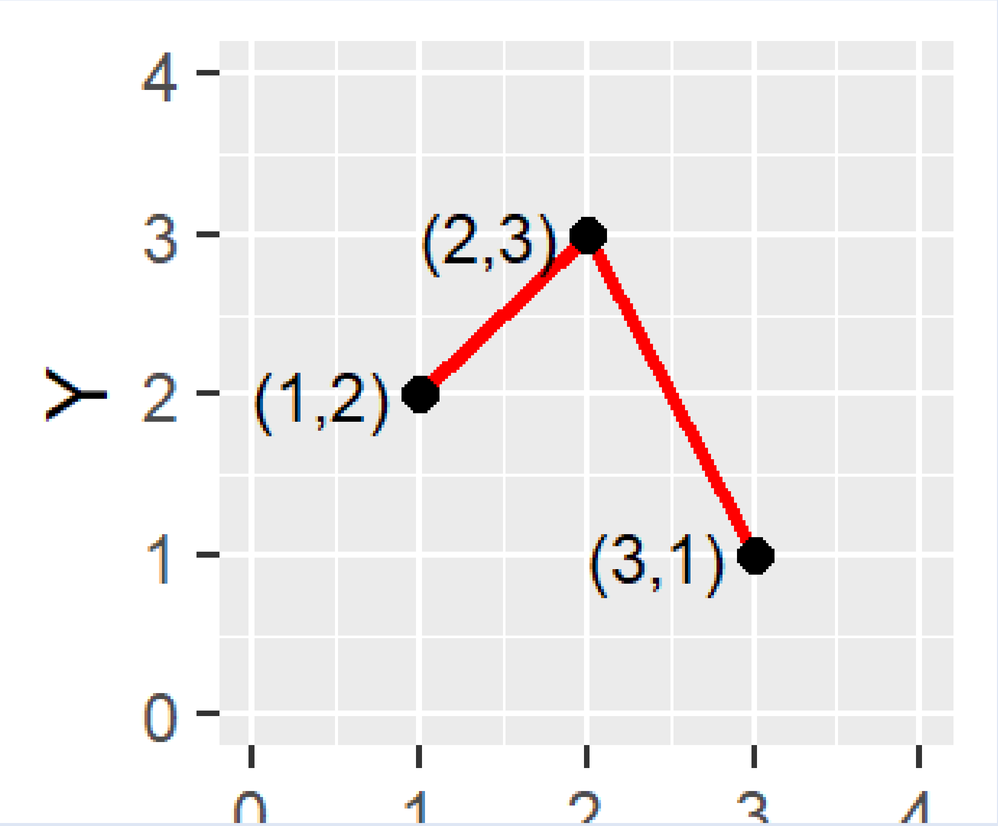
Point

- A point is composed of **one coordinate pair** representing a specific location in a coordinate system.
- Points are the most **basic** geometric primitives having **no length or area.**



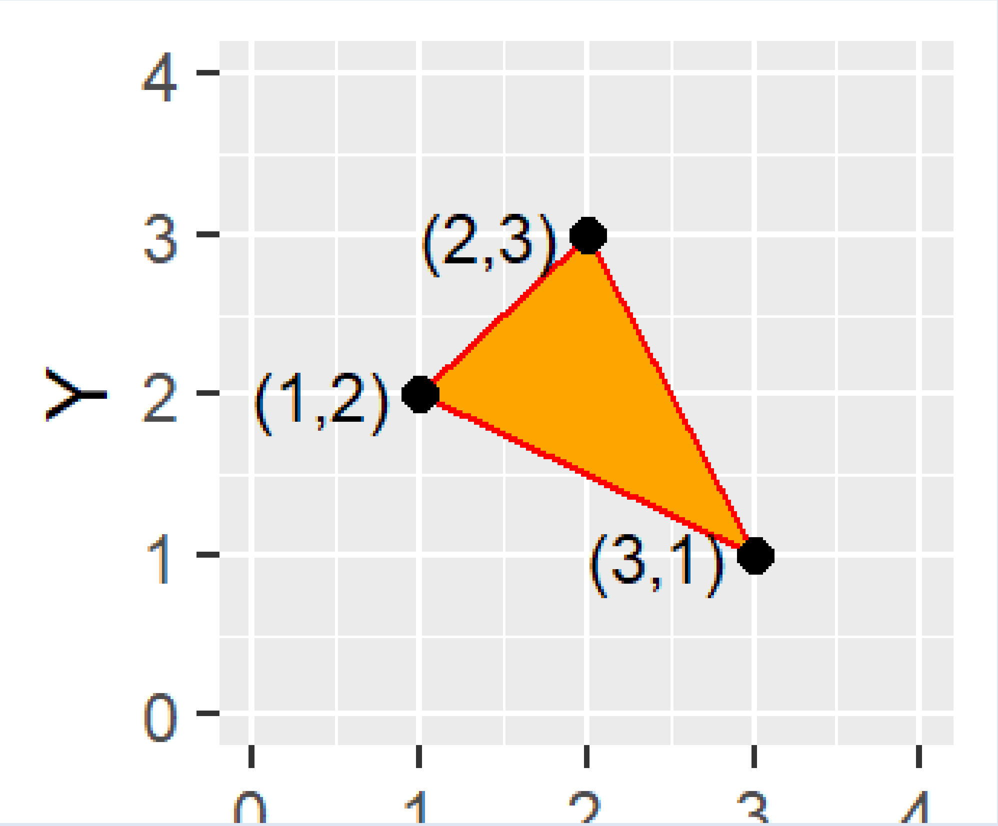
Polyline

- A polyline is composed of a sequence of two or more coordinate pairs called **vertices**.
- A vertex is defined by coordinate pairs, just like a point, but what differentiates a vertex from a point is its explicitly defined relationship with neighbouring vertices. A vertex is connected to at least one other vertex.
- Roads and rivers are commonly stored as polylines



Polygon

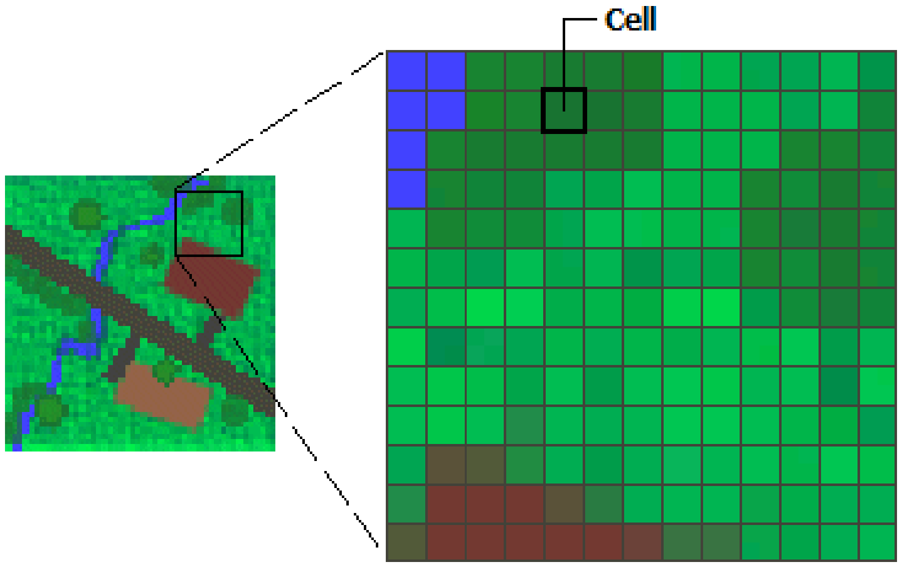
- A polygon is composed of three or more line segments whose starting and ending coordinate pairs are the same.
- Sometimes you will see the words *lattice* or *areal unit* used in lieu of 'polygon'.
- The **area that a polygon encloses is explicitly defined**. If it isn't, then you are working with a polyline feature. If this does not seem intuitive, think of three connected lines defining a triangle: they can represent three connected road segments (thus polyline features), or they can represent the grassy strip enclosed by the connected roads (in which case an 'inside' is implied thus defining a polygon).



```
> head(world)
Simple feature collection with 6 features and 10 fields
geometry type:  MULTIPOLYGON
dimension:      XY
bbox:           xmin: -180 ymin: -18.28799 xmax: 180 ymax: 83.23324
CRS:            EPSG:4326
  iso_a2      name_long     continent region_un        subregion              type    area_km2       pop  lifeExp gdpPercap                           geom
1     FJ           Fiji       Oceania   Oceania        Melanesia Sovereign country    19289.97    885806 69.96000  8222.254 MULTIPOLYGON (((180 -16.067...
2     TZ       Tanzania        Africa    Africa   Eastern Africa Sovereign country   932745.79  52234869 64.16300  2402.099 MULTIPOLYGON (((33.90371 -0...
3     EH Western Sahara        Africa    Africa  Northern Africa     Indeterminate    96270.60        NA       NA        NA MULTIPOLYGON (((-8.66559 27...
4     CA         Canada North America  Americas Northern America Sovereign country 10036042.98  35535348 81.95305 43079.143 MULTIPOLYGON (((-122.84 49,...
5     US  United States North America  Americas Northern America           Country  9510743.74 318622525 78.84146 51921.985 MULTIPOLYGON (((-122.84 49,...
6     KZ     Kazakhstan          Asia      Asia     Central Asia Sovereign country  2729810.51  17288285 71.62000 23587.338 MULTIPOLYGON (((87.35997 49...
```

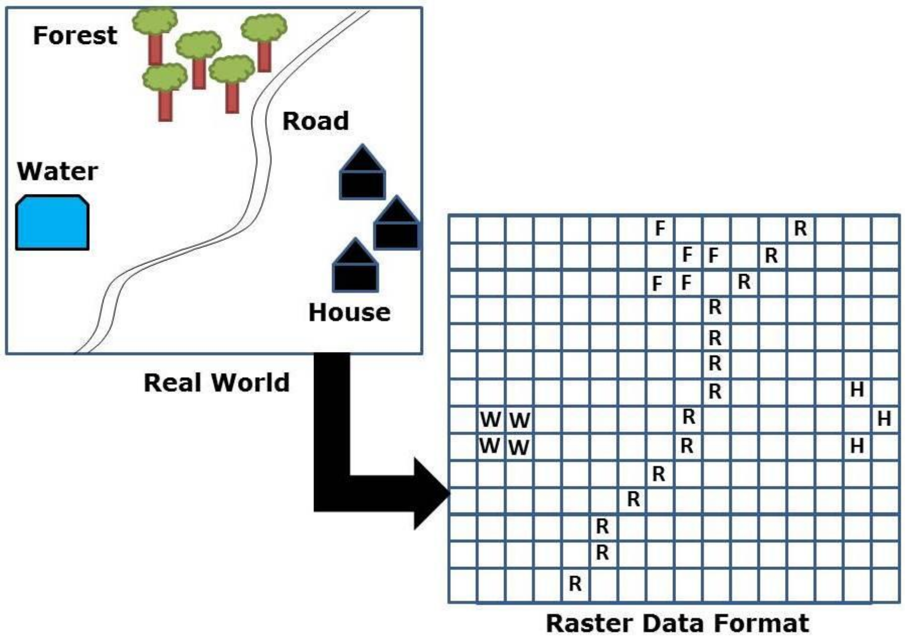
Vector dataset example

Raster Data



- The geographic raster data model usually consists of a raster header and a matrix (with rows and columns) representing **equally spaced cells** (often also called pixels or tiles).
- The **raster header defines the coordinate reference system**, the extent and the origin.
- The header defines the extent via the number of columns, the number of rows and the cell size resolution.
- Starting from the origin, we can access and modify each single cell by either using the **cell ID** or by specifying the rows and columns.

Raster Data

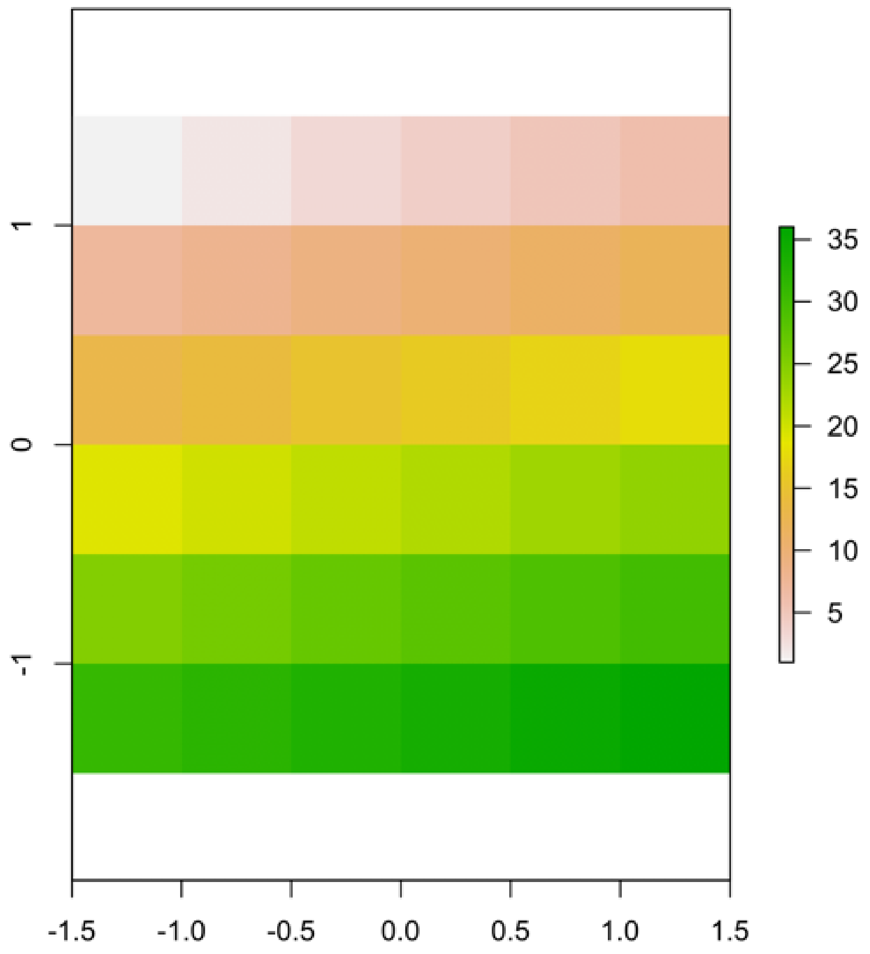


Real World Feature Representation in Raster Data Format

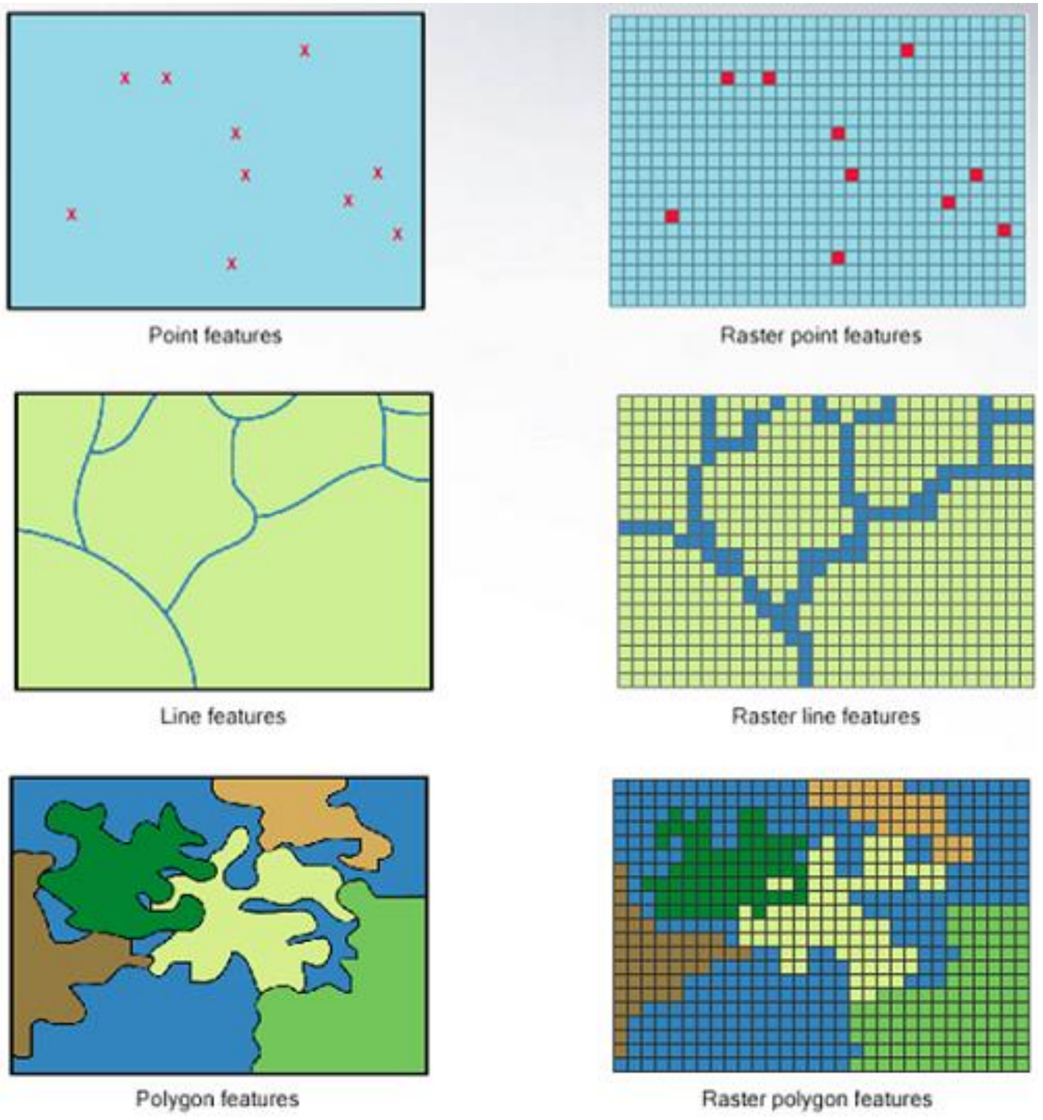
- This matrix representation avoids storing explicitly the coordinates for the four corner points (in fact it only stores one coordinate, namely the origin) of each cell corner as would be the case for rectangular vector polygons. This and map algebra makes raster processing much more efficient and faster than vector data processing (think of satellite imaging).
- In contrast to vector data, the cell of one raster layer can only hold a single value. The value might be numeric or categorical.
- Raster datasets are commonly used for representing and managing imagery, surface temperatures, digital elevation models, and numerous other entities.

Raster dataset example

```
> new_raster
class       : RasterLayer
dimensions  : 6, 6, 36  (nrow, ncol, ncell)
resolution  : 0.5, 0.5  (x, y)
extent      : -1.5, 1.5, -1.5, 1.5  (xmin, xmax, ymin, ymax)
crs         : +proj=longlat +datum=WGS84 +ellps=WGS84 +towgs84=0,0,0
source      : memory
names       : layer
values      : 1, 36  (min, max)
```

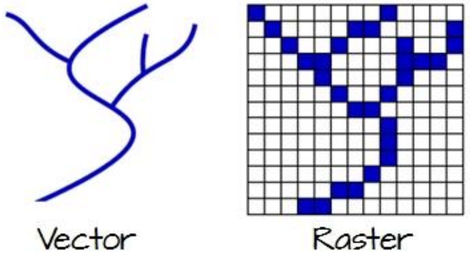


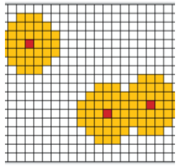
Vector vs. Raster



Vector Data vs. Raster Data

Vector data have discrete, well-defined borders, meaning that vector datasets usually have a **high level of precision**.





The raster data divides the surface up into cells of constant size (grids or tiles). Raster datasets are the basis of background images used in **web-mapping** and have been a vital source of geographic data since the origins of **aerial photography and satellite-based remote sensing devices**. Rasters aggregate spatially specific features to a given resolution

Which to use? The answer likely depends on your domain of application:

Vector data tends to dominate the social sciences because human settlements tend to have discrete borders.

Raster dominates many environmental sciences because of the reliance on remote sensing data.

There is **much overlap** in some fields and raster and vector datasets can be used together: ecologists and demographers, for example, commonly use both vector and raster data. Furthermore, it is possible to **convert between the two forms**.

Plan for the term: practical topics

First, we will look at how to manipulate vector data in R using the *sf* package.

- A whole lesson is dedicated on understanding the geometry operations and handling the coordinate reference system in datasets.
- The art of cartography in R using the *tmap* package.

Spatial descriptive summary measures: spatial mean, spatial sd, etc.

Theoretical analysis of spatial data

- Point pattern analysis
- Areal data analysis
  - Modelling spatial relationships: spatial econometrics
- Geostatistical analysis

Take home points…

- Why and what of spatial data?
- Key features of spatial data and analysis
- Types of spatial data:
  - Theoretical classification
    - Geospatial
    - Point
    - Areal
  - Practical classification
    - Vector: points, lines and polygons
    - Raster
- MAUP
- Vector vs. raster

References

- *Statistical Methods for Spatial Data Analysis* by Oliver Schabenberger and Caro A. Gotway, 1st edition, (2005) Chapter 1.
- *Geocomputation with R*, by Robin Lovelace, Jakub Nowosad, Jannes Muenchow. https://geocompr.robinlovelace.net