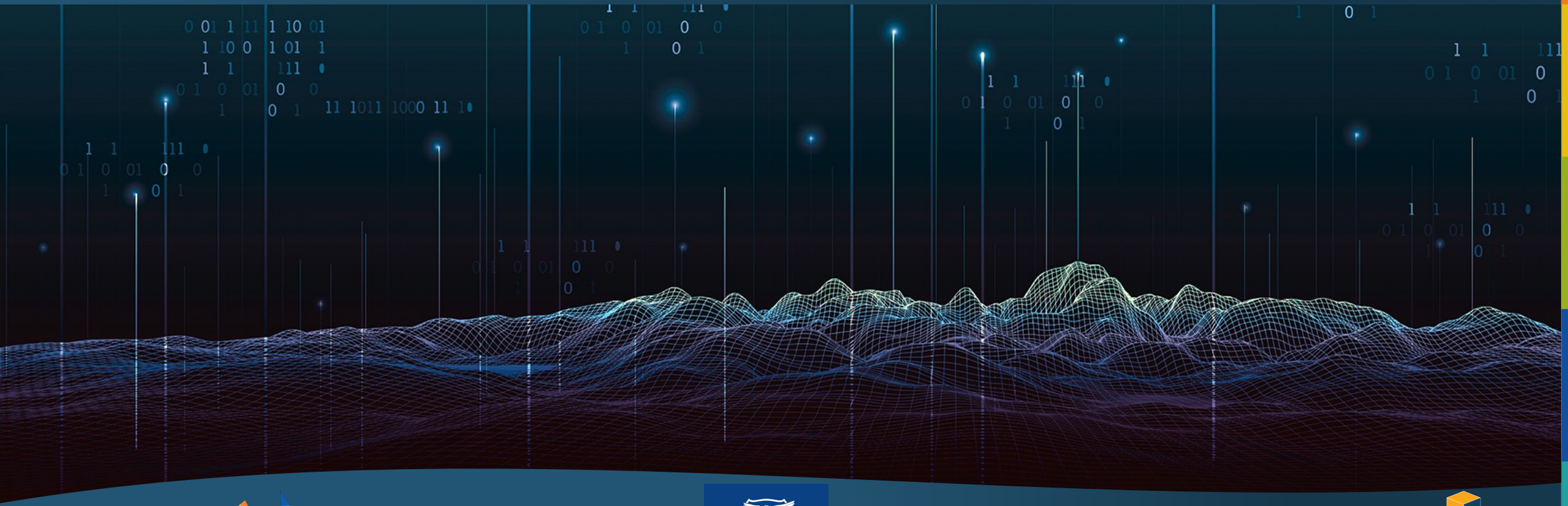
Spatial data and data integration

Australian Urban Research Infrastructure Network (AURIN)



AURIN
aurin.org.au

THE UNIVERSITY OF MELBOURNE

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

Acknowledgement of Country

We acknowledge the Traditional Owners of the land on which this event is taking place and pay respect to their Elders (past and present) and families.

- The Geosocial work package
- Finding and using spatial data
- Deciding on integration
- Integration methods
- Producing a new data product
- Exercise: Producing a map

The Geosocial work package

Finding and using spatial data

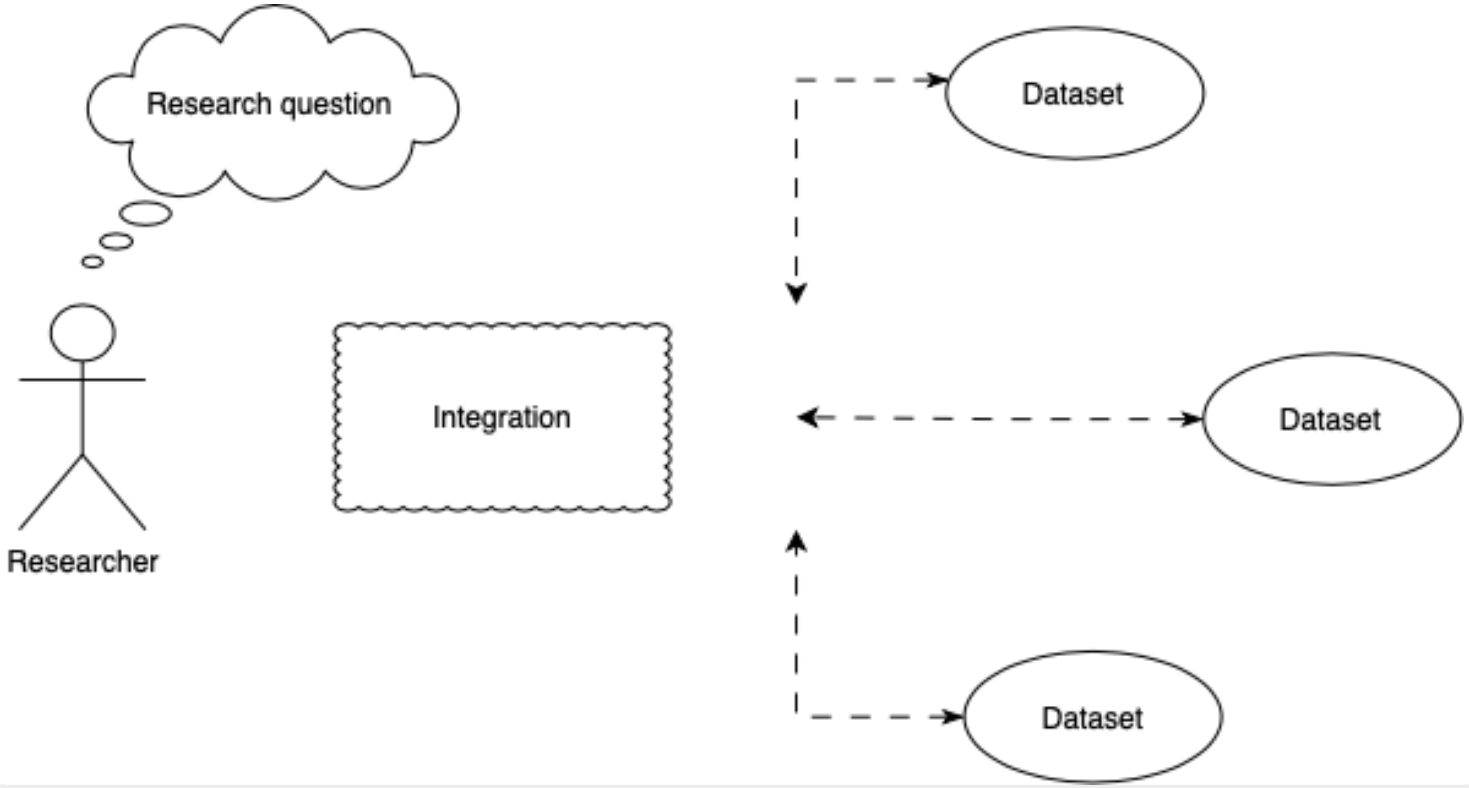
Deciding on integration

Integration methods

Producing a new data product

Exercise: Producing a map

Problem: The researchers want to bring data on people and places together, but don't know how to do it and what the issues might be.



The Geosocial work package: Motivation

Solution: Data integration service which will allow researchers to enhance **people-centred survey** data with **spatially structured data** capturing information on places where these people live.

The Geosocial work package

Finding and using spatial data

Deciding on integration

Integration methods

Producing a new data product

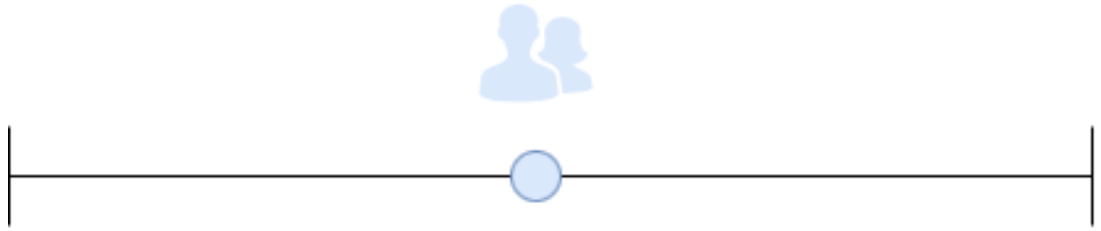
Exercise: Producing a map

Survey data

Cross-sectional data: Cross-sectional data consists of data on one or more variables collected at the same point in time. (Gujarati, 2011)

Examples:

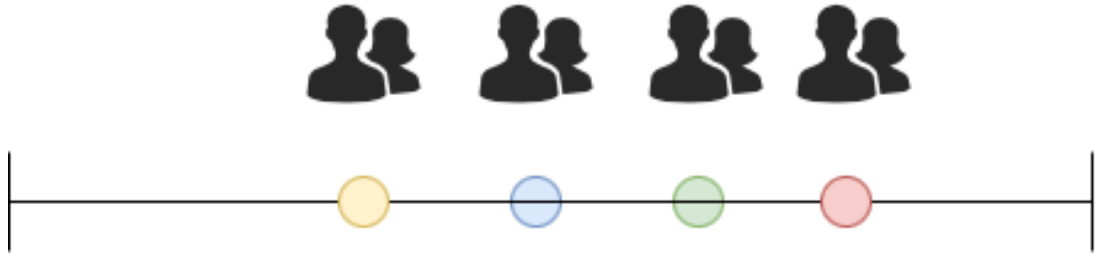
- Population census
- Consumer Expenditure Surveys
- Opinion polls



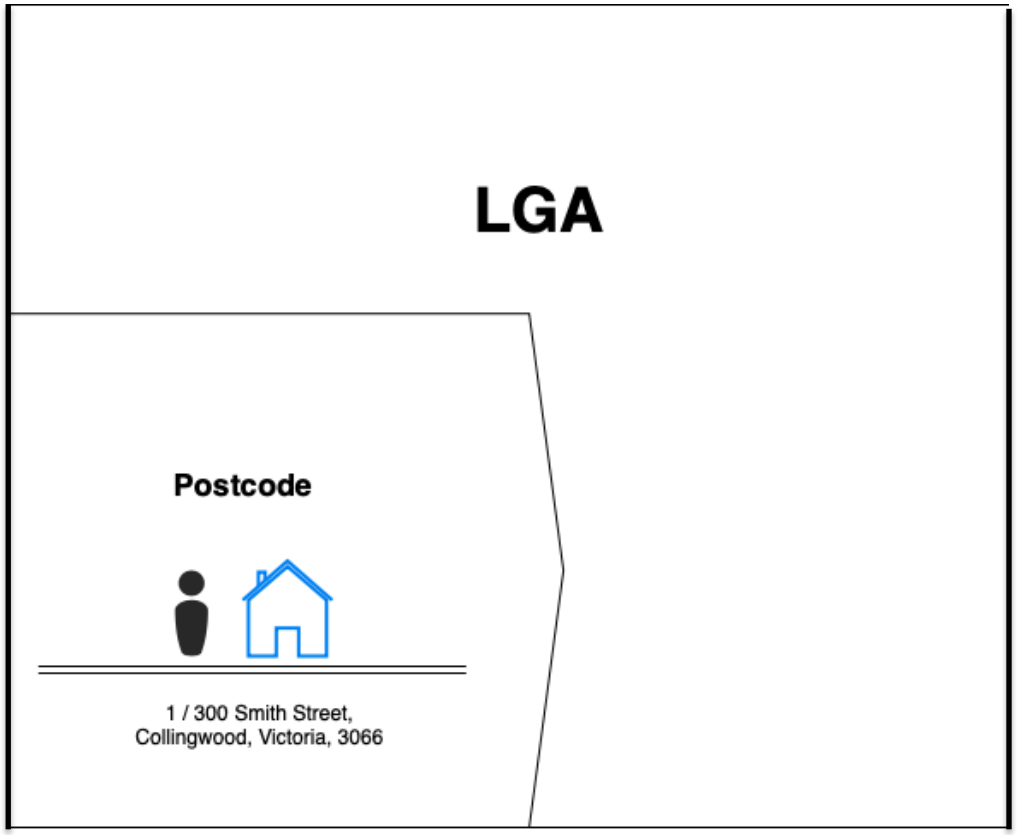
Longitudinal data (panel data): repeatedly collect data from the same sample over an extended period of time

Examples:

- Household, Income and Labour Dynamics in Australia (HILDA)
- Longitudinal Surveys of Australian Youth (LSAY)

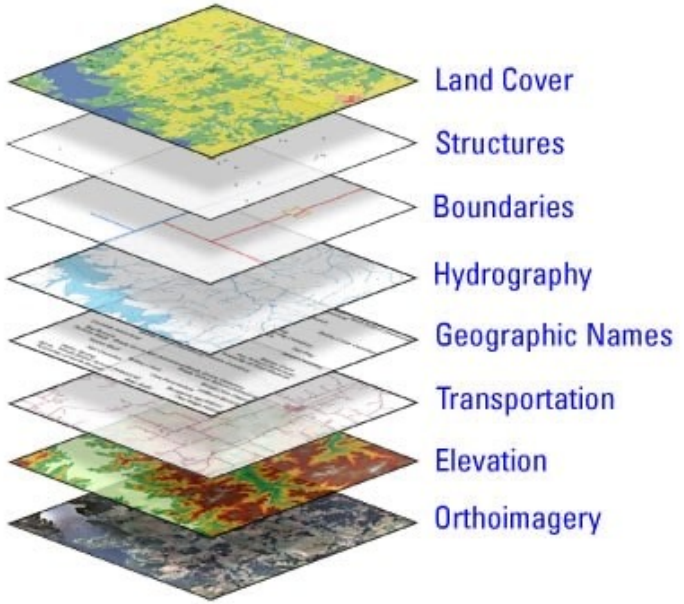


What is spatial data?

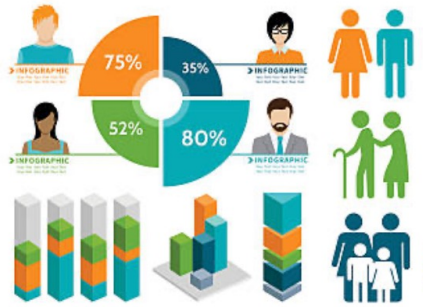


Data that have an implicit or explicit association with a location relative to Earth

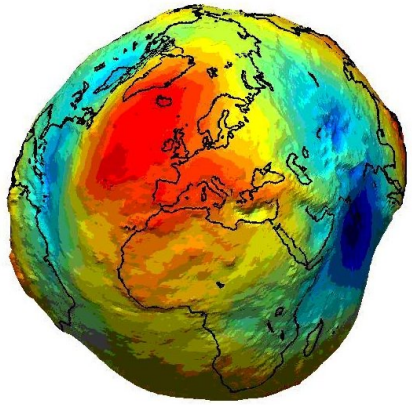
Physical characteristics



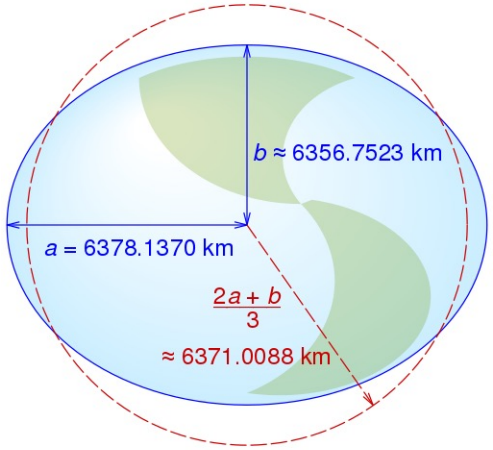
Socio-demographic characteristics



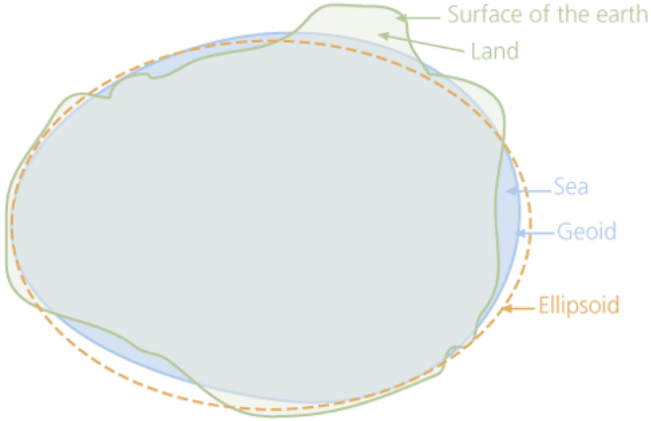
World Geodetic System 1984 (WGS84)

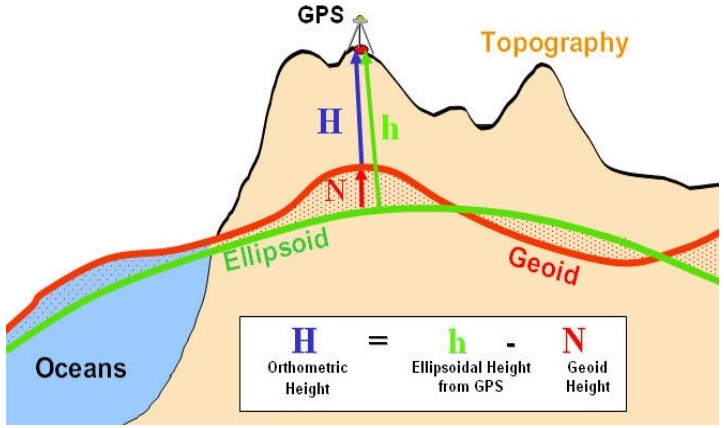


Geoid



Ellipsoid





Geoid vs Ellipsoid

Finding and using spatial data

Spatial analysis: represents a collection of **techniques** and **models** that explicitly use the spatial referencing of each data case.

Spatial analysis needs to make assumptions about or draw on data describing spatial relationships or spatial interactions between cases. (Chorley, 1972; Haining 1994).

The Geosocial work package

Finding and using spatial data

Deciding on integration

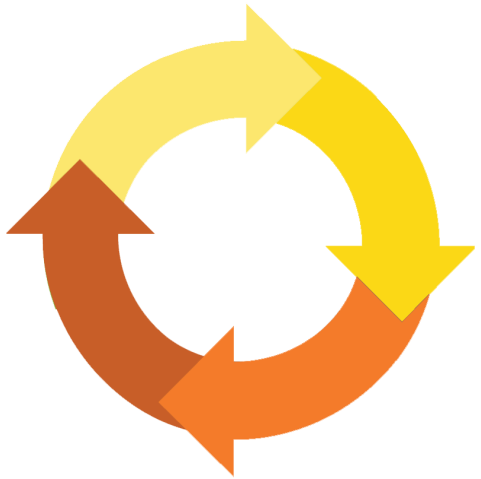
Integration methods

Producing a new data product

Exercise: Producing a map

Deciding on integration:

Step 1: Decide a research question.
Step 2: Find datasets that are suitable to the research question.
Step 3: Define a methodology to answer your research question
Step 4: Robustness: Evaluate the methodology



Deciding on integration:

Step 1: **Research question**: How is **leaving a high school** associated with Australia's **economic development**?

Step 2:

Database: **PHIDU - Education (LGA) 2015-2016:**

Variable: School Leaver Participation In Higher Education 2016 Enrolled in higher education.

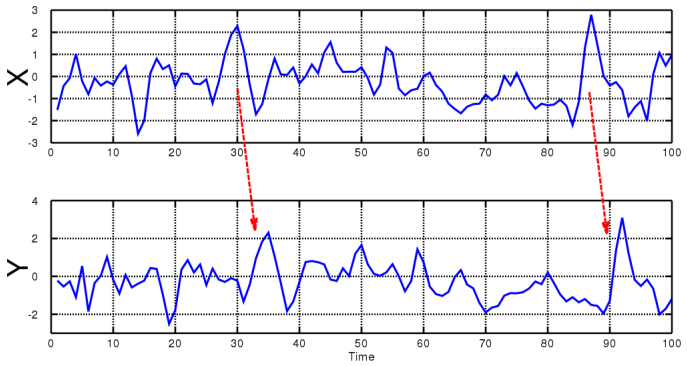
Database: Socio-Economic Indexes for Areas (SEIFA) – SA2 - 2016

Variable:

Step 3: Use a spatial correlation (Moran's I) to evaluate the correlation between leaving the school on economic development (SEIFA).

Step 4: Try other ways to estimate spatial correlation: Geary's C and Getis-Ord's index

Considerations



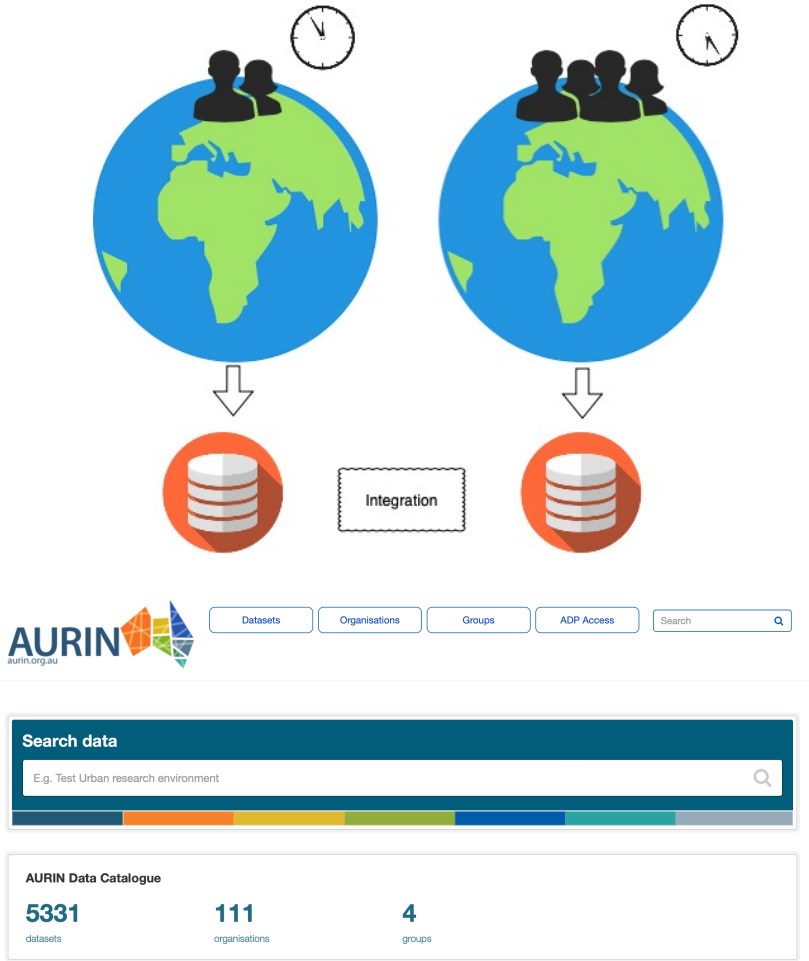
Causality/ Temporary lags

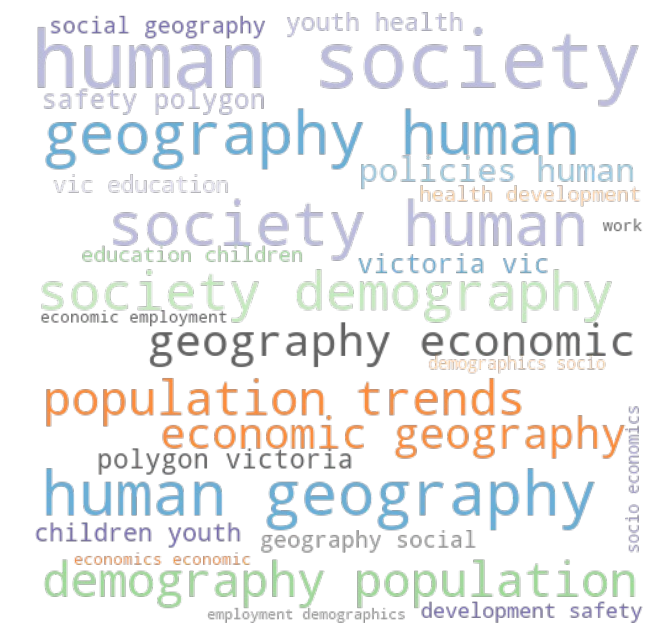
Spatial

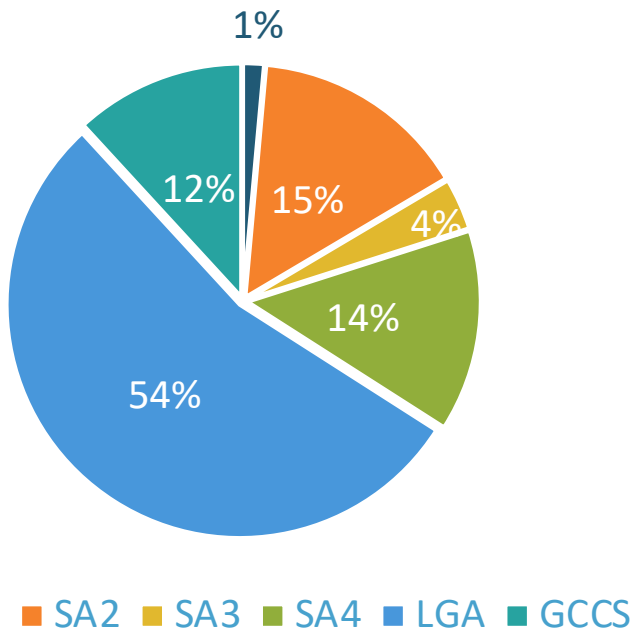
Semantic

Images from: BiObserver, Earthdatascience & 52north

Deciding on integration:







More information: https://data.aurin.org.au

The Geosocial work package

Finding and using spatial data

Deciding on integration

Integration methods

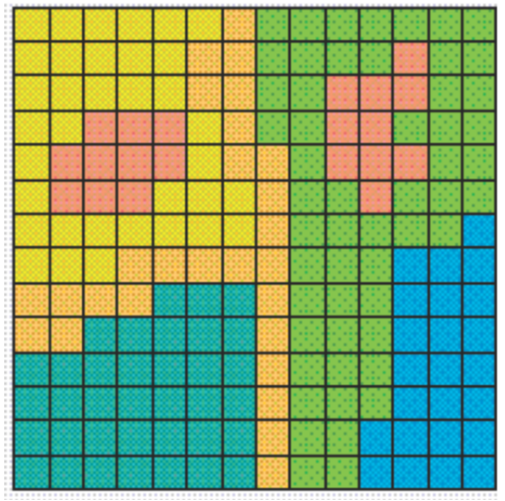
Producing a new data product

Exercise: Producing a map

Finding and using spatial data

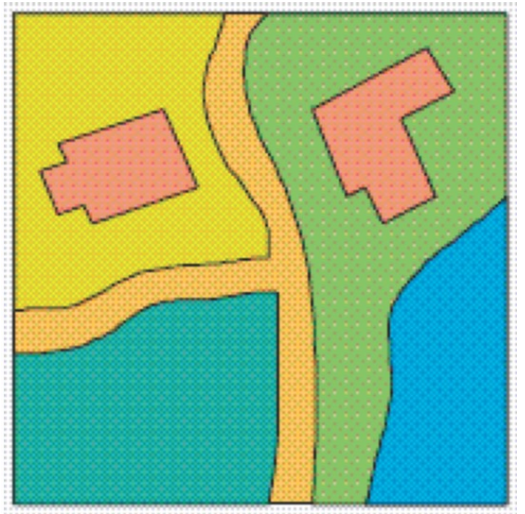
Raster map:

- Raster data is stored as a grid of values that are rendered on a map as pixels.
- Each pixel value represents an area on the Earth's surface.



Vector map:

- Consists of objects described by coordinates in a given coordinate system.
- The vector model uses points and line segments to identify locations on the earth.



Images from: ESRI

Vectorial map:

Type of elements:

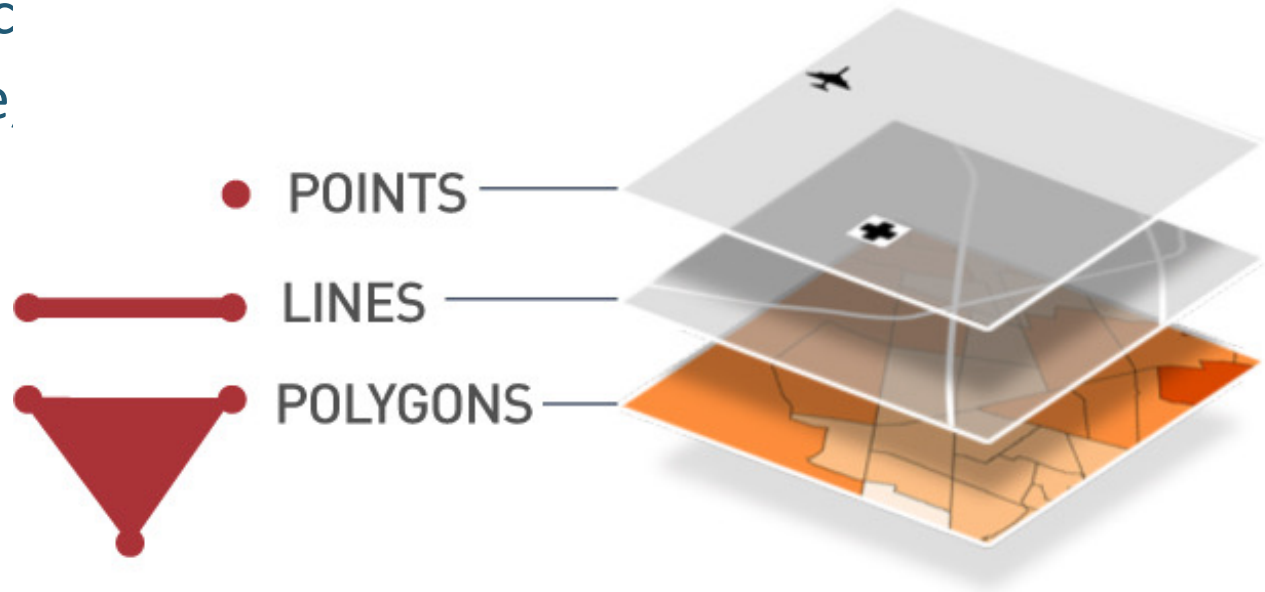
Point: Addresses, locations, points of interest, etc

Lines: streets, freeways, borders, etc

Polygons: Countries, cities, Cadastre

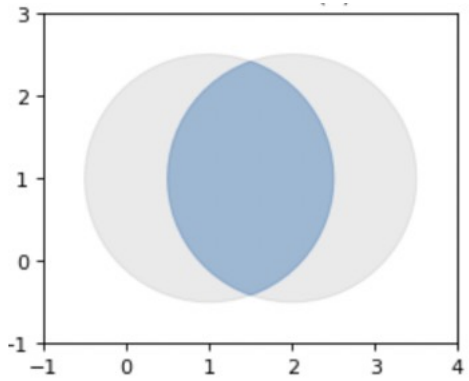
Advantages:

- Spatial operations
- Spatial aggregation
- Spatial Join

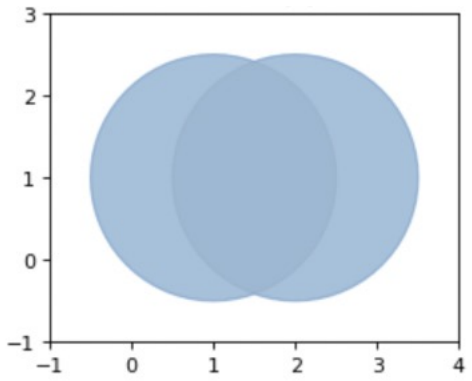


Images from: Berkeley

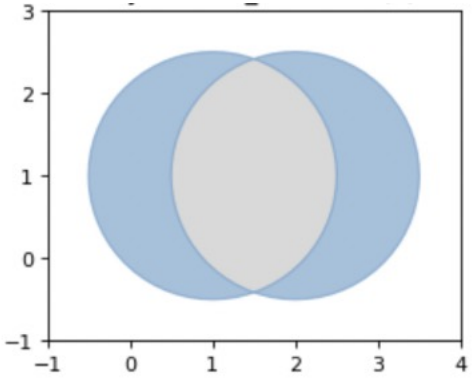
Spatial operations



Intersection

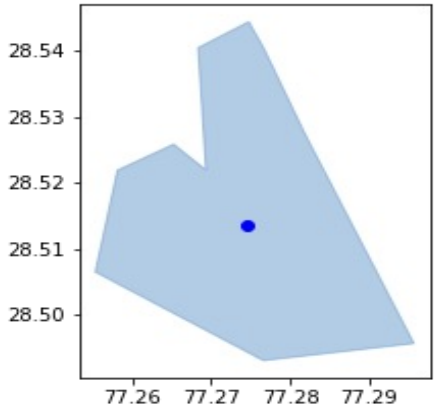


Union

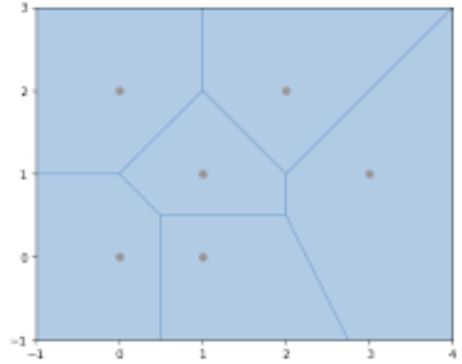


Difference

Contour



Centroid

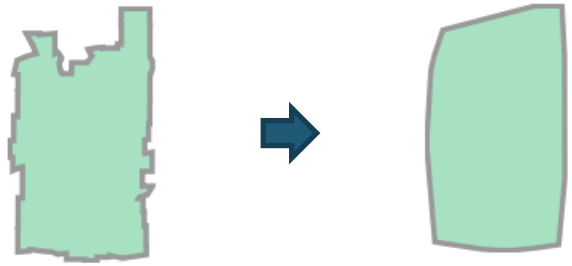


Within

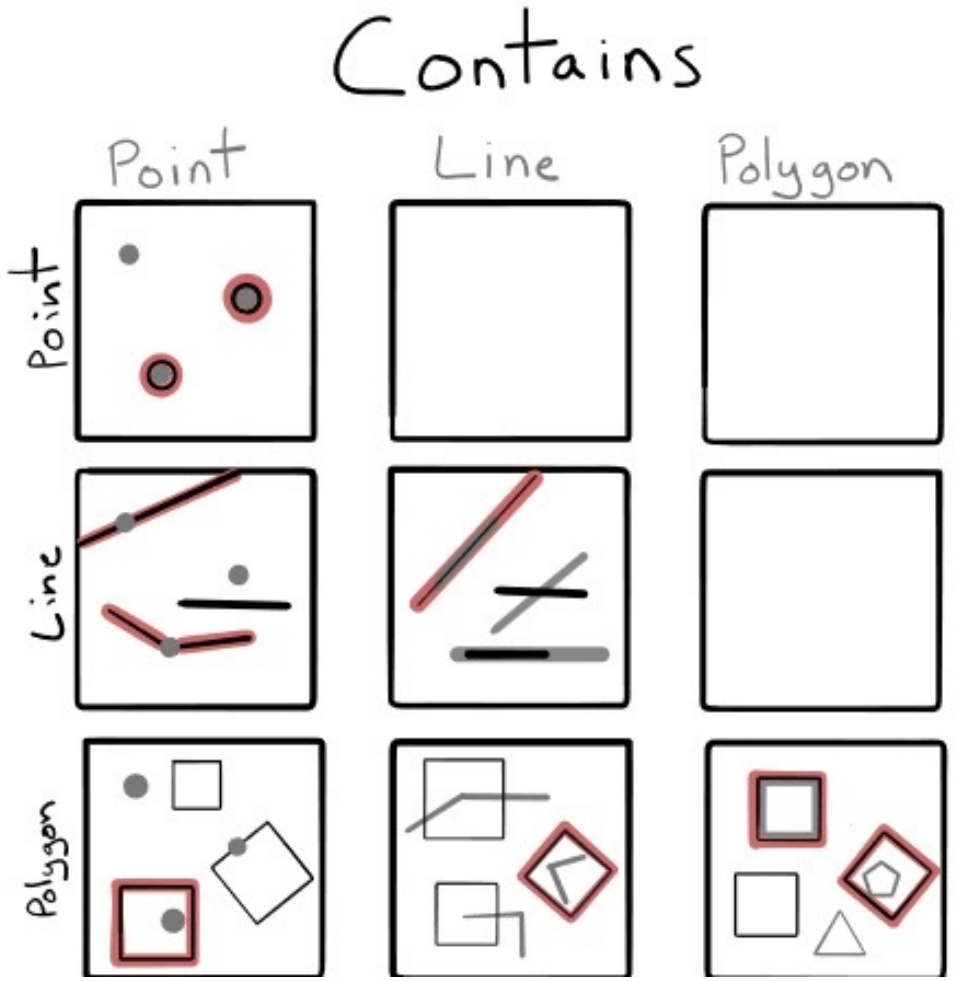
Images from: Pysal

Spatial operations

Buffer: Create a new layer that covers this in a zone of influence whose radius is the one indicated in the analysis tool.

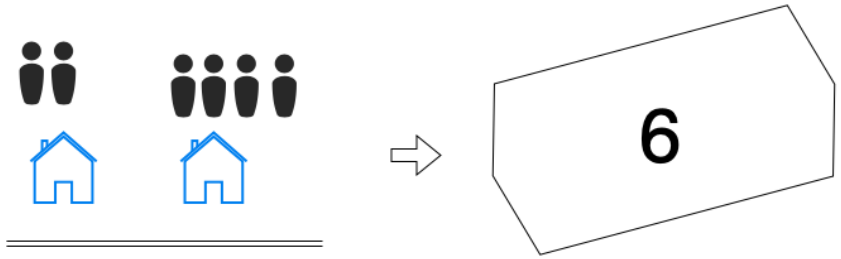


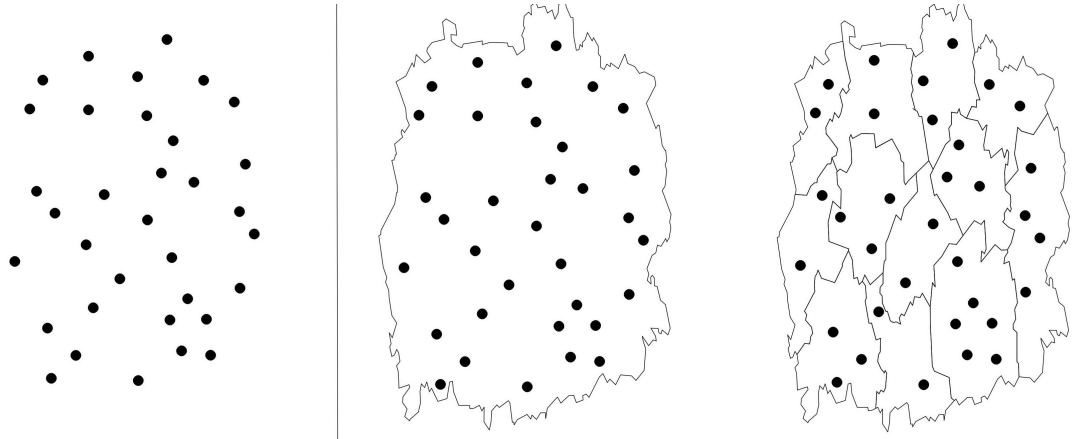
Convex capsule: a set of points that contains the intersection of all convex sets within a polygon.



Images from: Pysal & PyGis

Spatial aggregation





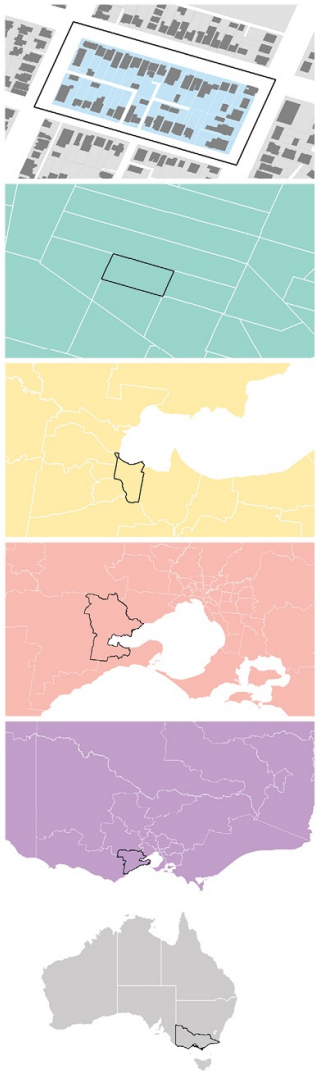
Potential areal data

Possible aggregations

Images from: Dusek, Tamás. (2004)

Spatial aggregation

Mesh Blocks	- The "building blocks" of the census - 2016 Census contains 358,122 Mesh Blocks - Generally follow man-made or natural boundaries (ie. roads, rivers, etc)
SA1s	- Aggregates of Mesh Blocks - 2016 Census contains 57,523 SA1s - Generally contain population of 200 - 800 people
SA2s	- Aggregates of SA1s & Mesh Blocks - 2016 Census contains 2,310 SA2s - Generally contain population of 3,000 to 25,000 people
SA3s	- Aggregates of SA2s, SA1s & Mesh Blocks - 2016 Census contains 358 SA3s - Generally contain population of 30,000 - 130,000 people
SA4s	- Aggregated of SA3s, SA2s, SA1s and Mesh Blocks - 2016 Census contains 18 SA4s - Generally contain population of 100,000 - 500,000 people
States and Territories	- All States and Territories of Australia



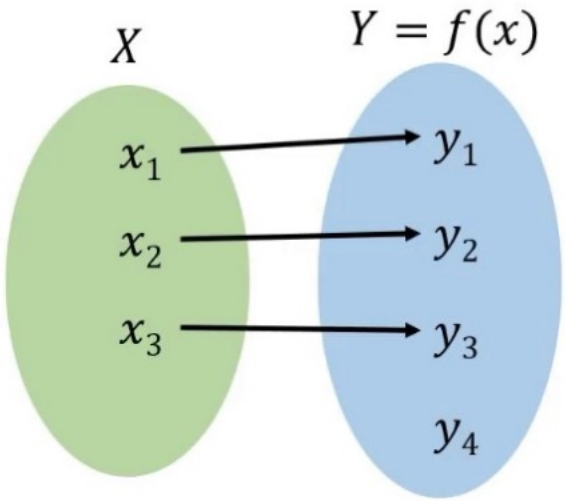
Images from: ABS

Concordance/Correspondence:

Objective: mathematically convert statistical data to and from geographic regions

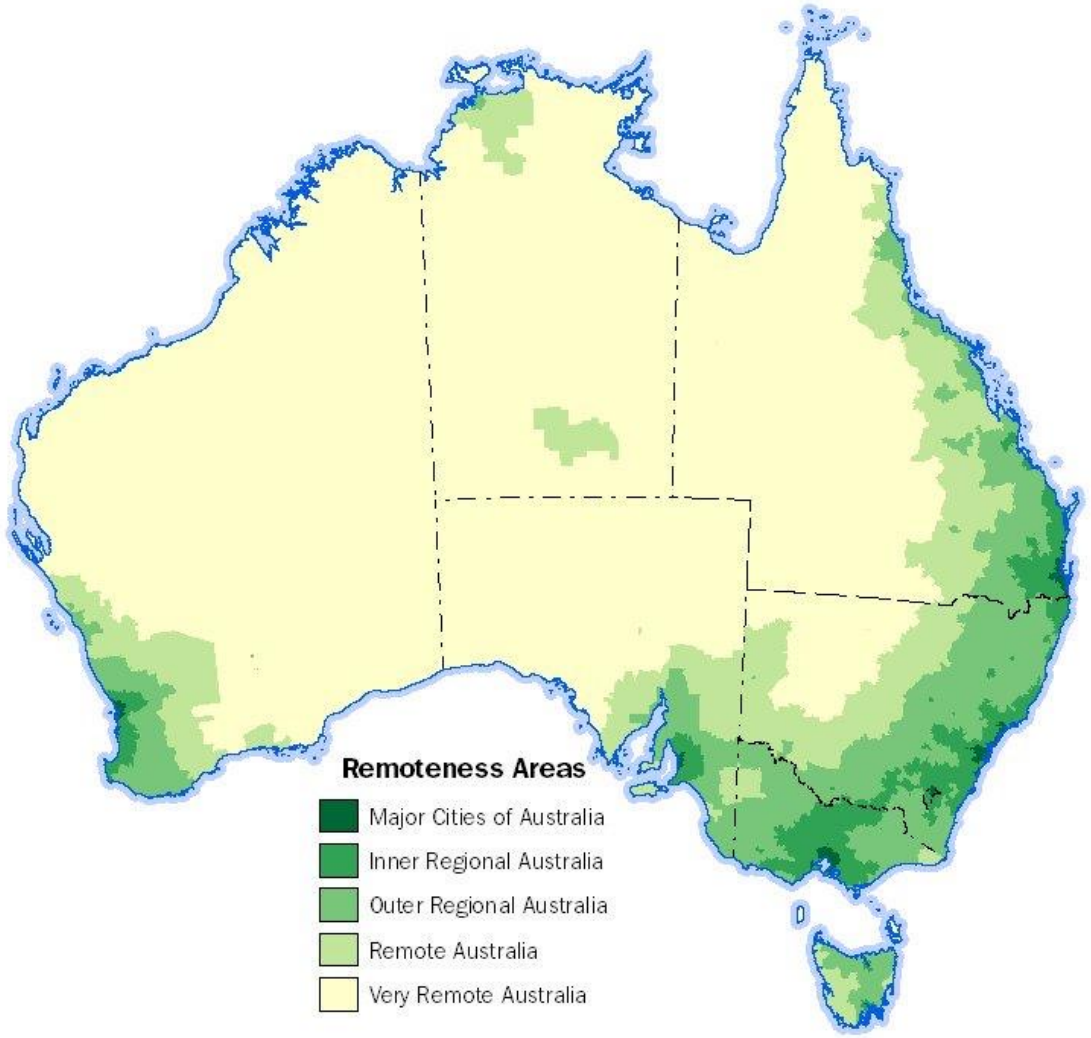
Main Structure and Greater Capital City Statistical Areas

2016 Mesh Blocks to 2021 Mesh Blocks
2016 Statistical Areas Level 1 to 2021 Statistical Areas Level 1
2016 Statistical Areas Level 2 to 2021 Statistical Areas Level 2
2016 Statistical Areas Level 3 to 2021 Statistical Areas Level 3
2016 Statistical Areas Level 4 to 2021 Statistical Areas Level 4
2016 Greater Capital City Statistical Areas to 2021 Greater Capital City Statistical Areas



Images from: ABS

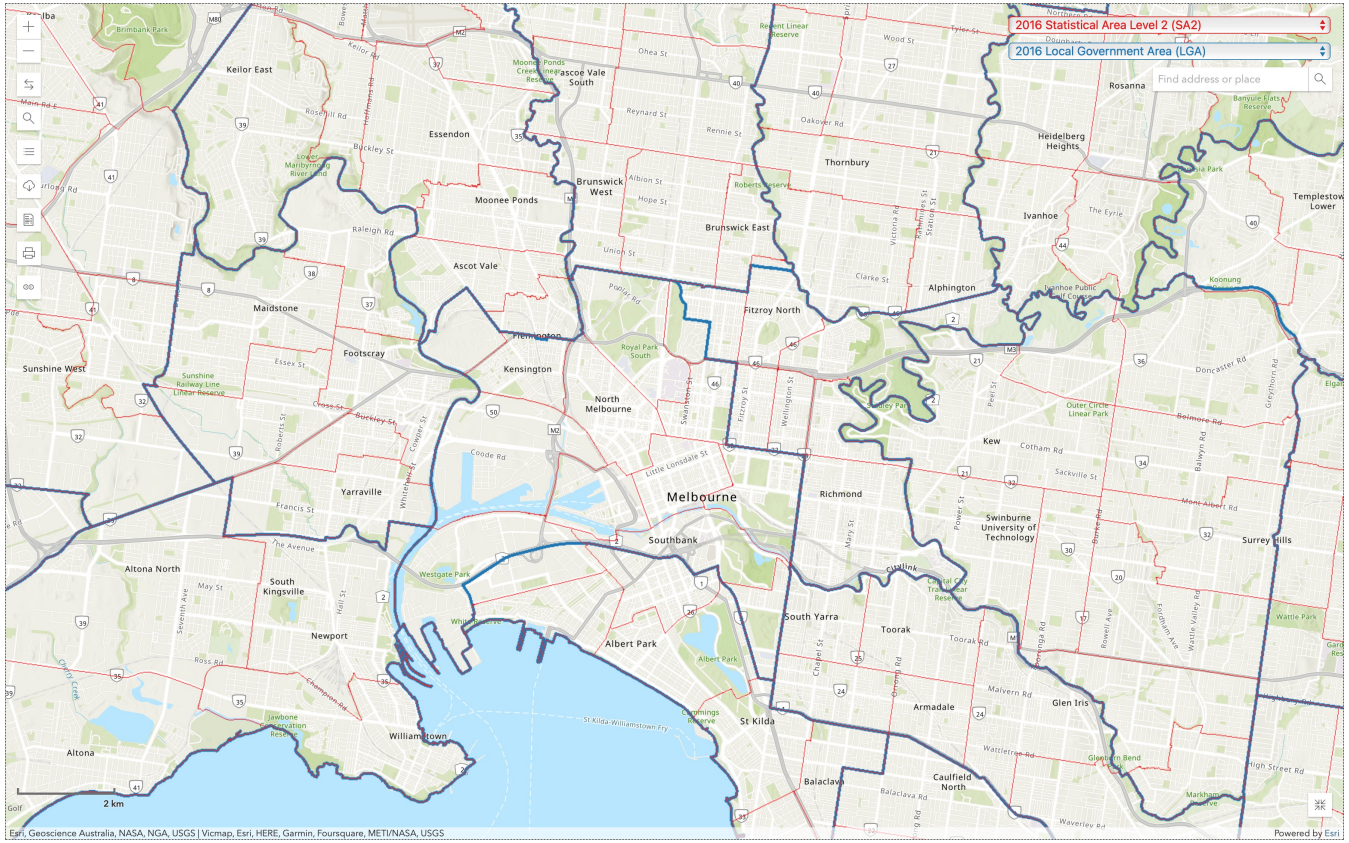
Spatial aggregation: Choropleth map



Images from: ABS

Example: Spatial aggregation

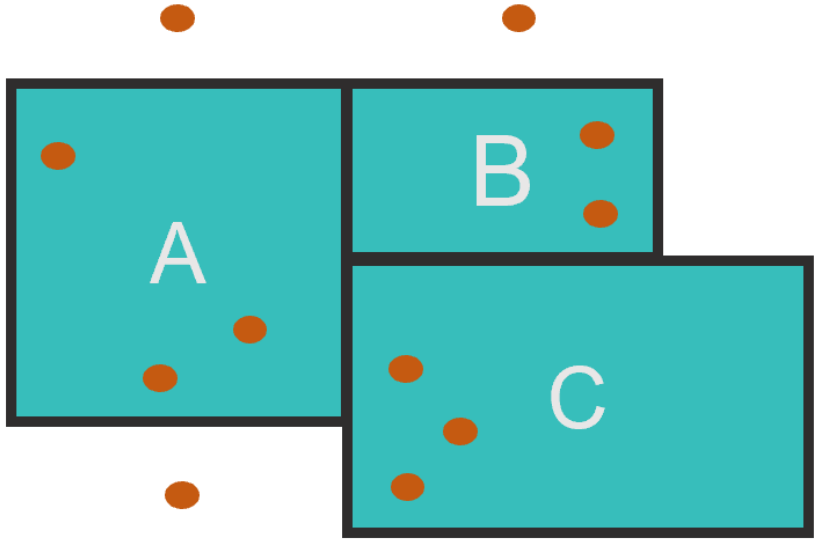
Socio-Economic Indexes for Areas (SEIFA) – SA2 - 2016 → **PHIDU - Education (LGA) 2015-2016:**



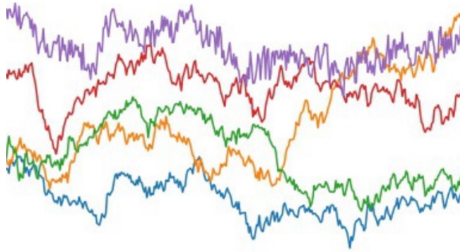
Images from: ABS

Spatial Join

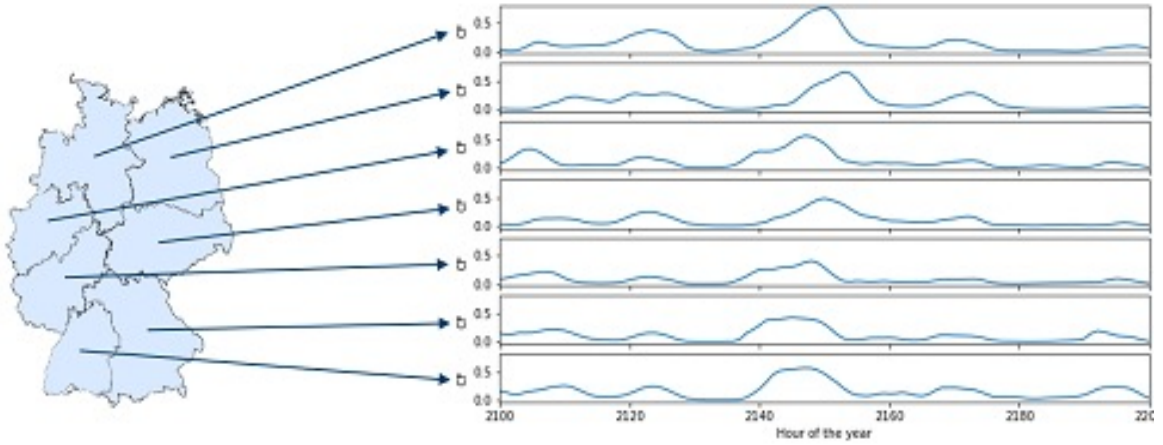
Geospatial



Temporal

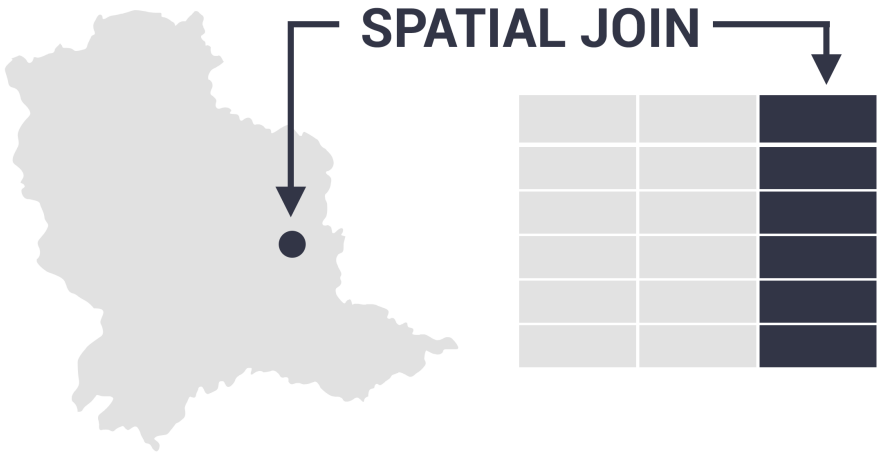


Spatiotemporal

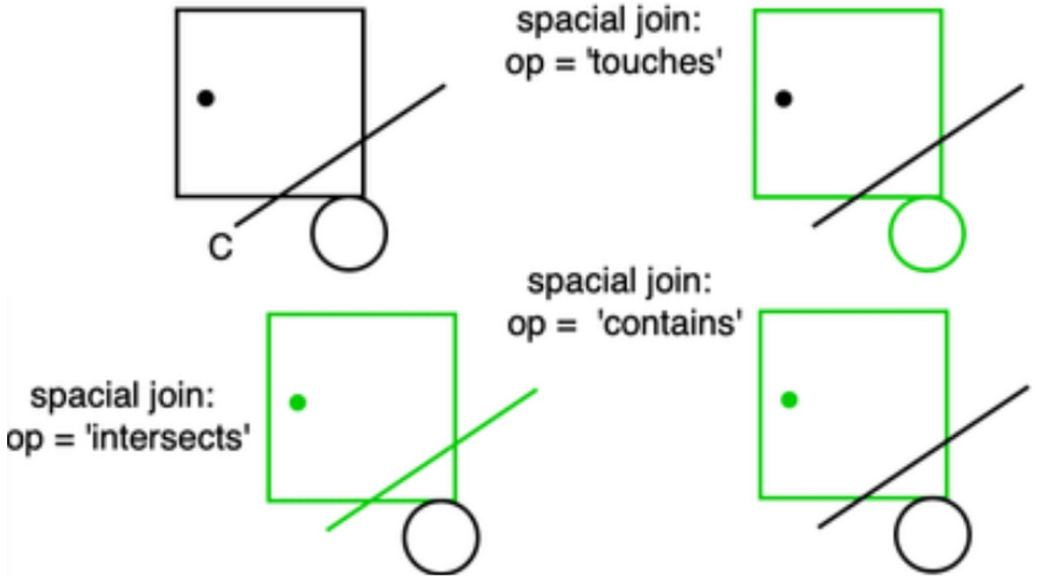


Images from: Metis & Interwork

Spatial Join



Types of spatial join

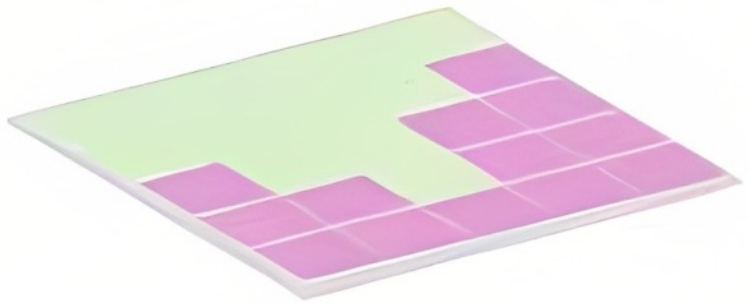


Images from: Pygis & Harder, Dorian

Spatial autocorrelation - Example

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

Waldo Tobler



High Spatial Autocorrelation
(Clustering)

Low Spatial Autocorrelation
(Checkerboard)

Images from: GisGeography

Limitations

- **Spatial aggregation:** In some cases, it is impossible to delve into the smallest detail of the problem.
- **Measurement error:** Data that does not correspond to the true values.
- **Assumptions**: Assumptions that are not entirely realistic.
- **Computing capacity:** High computational costs.

The Geosocial work package

Finding and using spatial data

Deciding on integration

Integration methods

Producing a new data product

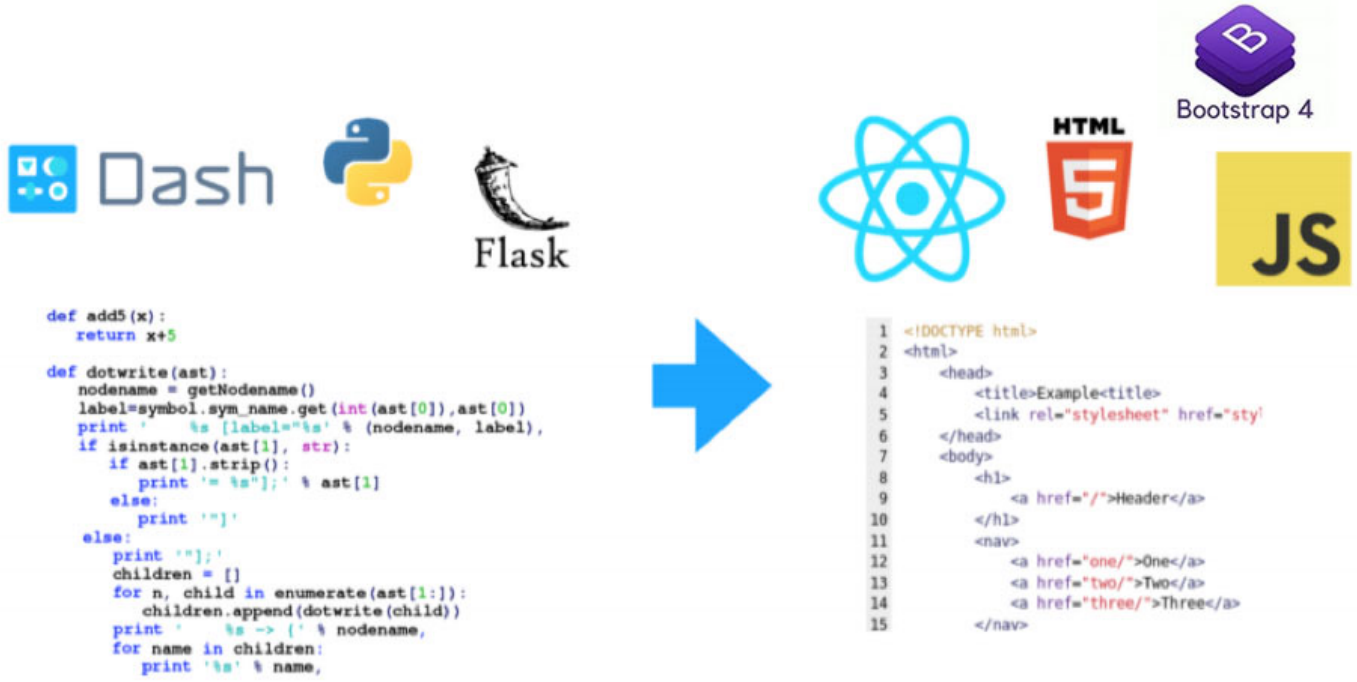
Exercise: Producing a map

Producing a new data product

A report, dashboard or application with its own user interface (UI), an API, or command-line SQL access

Image from: “Product Data and Your Digital Catalog” by Handshake

Producing a new data product



Examples:

- **R - Shiny:** https://shiny.rstudio.com/gallery/
- **Python Dash:** https://dash.gallery/Portal/
- **Java** - https://kepler.gl

Image from: “Product Data and Your Digital Catalog” by Handshake

The Geosocial work package

Finding and using spatial data

Deciding on integration

Integration methods

Producing a new data product

Exercise: Producing a map

Exercise – R

- **Part 1:** Spatial operations
  - Union
  - Intersects
  - Difference
  - Within
  - Buffer
  - Convex hull
  - Centroid
  - Boundaries
  - Area (Mts^2)

- **Part 2**: Integration methods:
  - ID join: Education & LGA by LGA_ID
  - Spatial join: SEIFA – LGA database

- **Part 3:** Geospatial aggregation:
  - SEIFA → SA2 → LGA

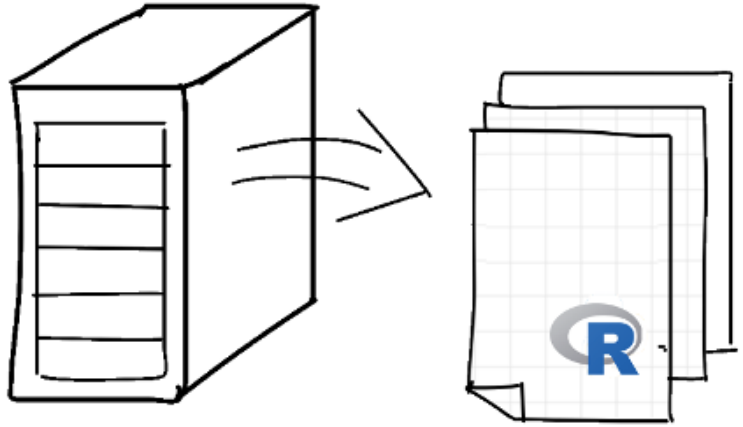
- **Part 4:** Producing a map

Exercise – R

Choices for the R environment

a) Cloud environment

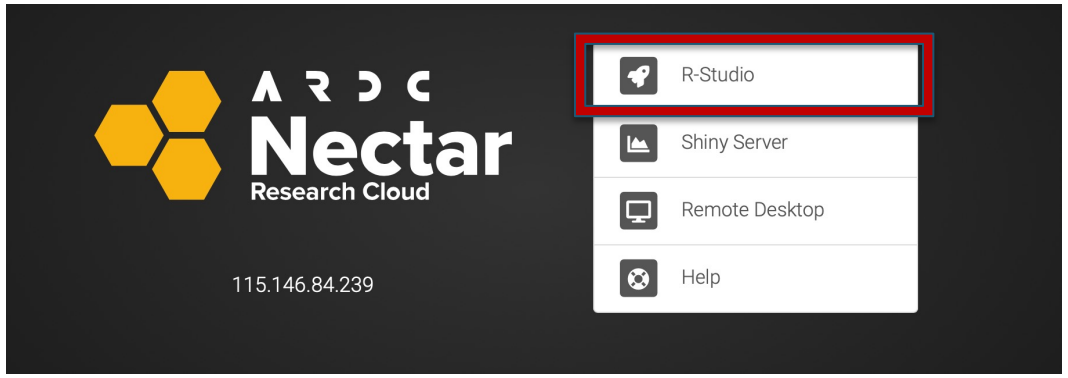
- Compatibility
- Less time spent installing R and dependencies.
- Better Package compatibility
- New users



b) Local environment

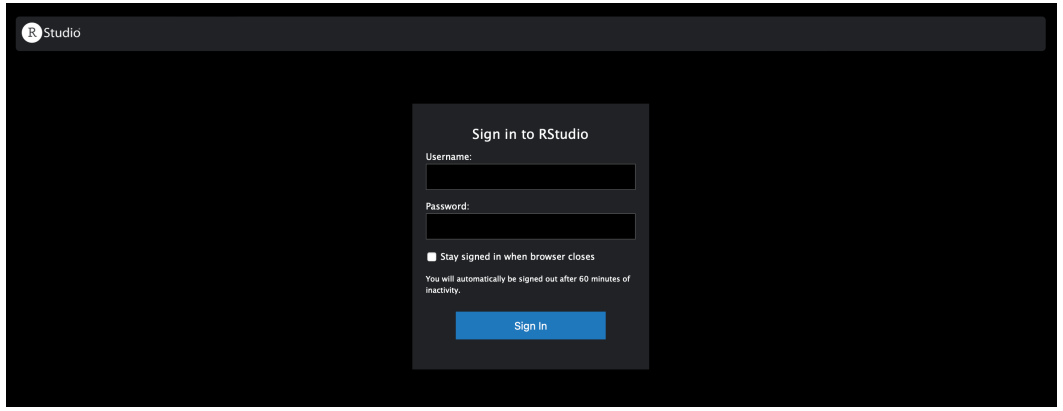
- Customisable
- Privacy
- Large data sets
- Experienced users

Cloud environment



Group 1:

- **Host: http://203.101.230.180**
- **User: user1….user15**
- **Password: hass2023**



Group 2:

- **Host: http://203.101.231.227**
- **User: user1….user15**
- **Password: hass2023**

Local environment

Step 1: Clone the repository on GitHub

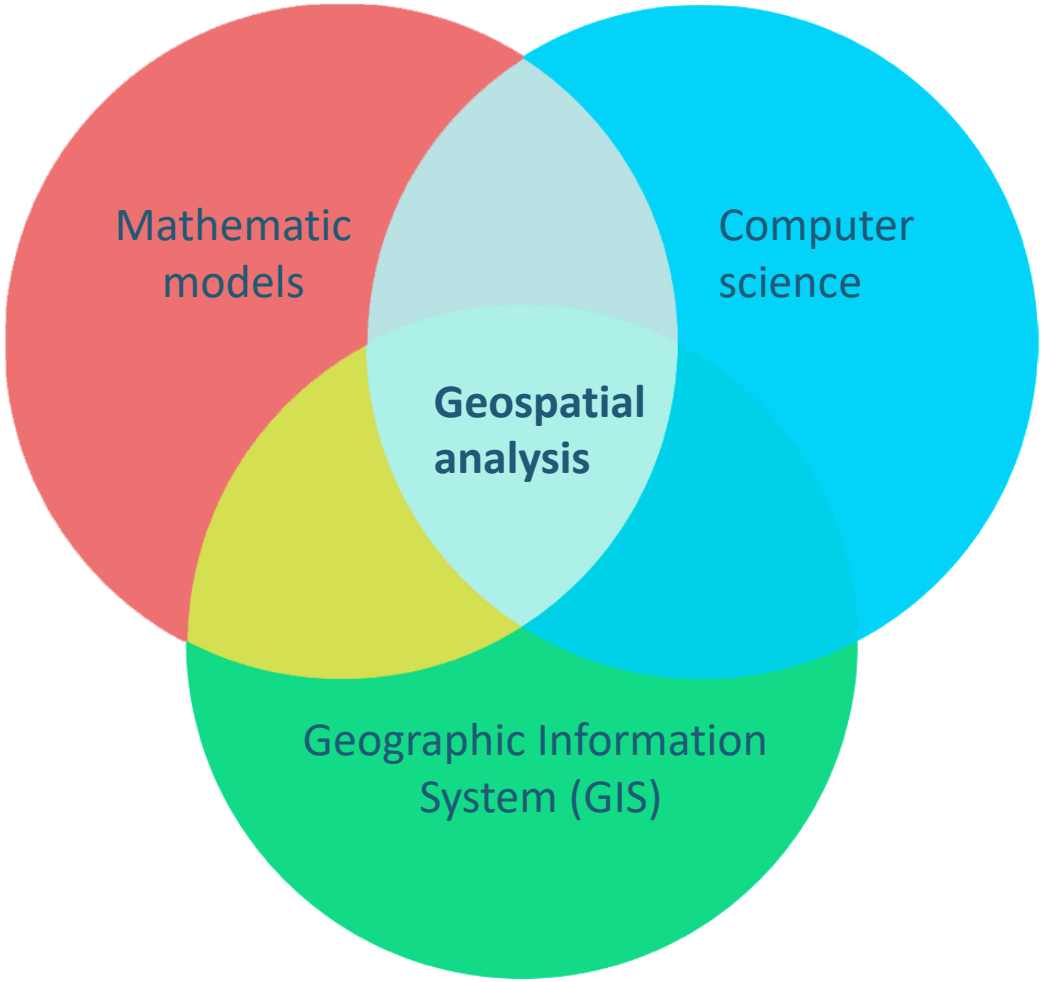
https://github.com/AURIN-OFFICE/HASS_Summer_School

Step 2: Install the necessary libraries

```
############### ---------------- Workshop -------------- ###############
###### ------ Clean variables ----- #########
rm(list=ls())
###### ------ Install libraries ----- #########
# install.packages(c('sf','tidyverse','ggplot2','leaflet.extras'))
```

Thank you

Finding and using spatial data



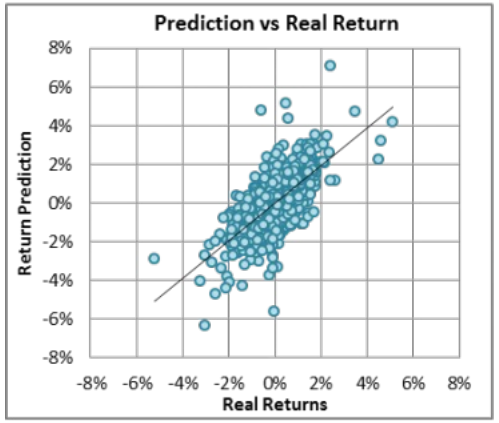
Supervised vs unsupervised learning

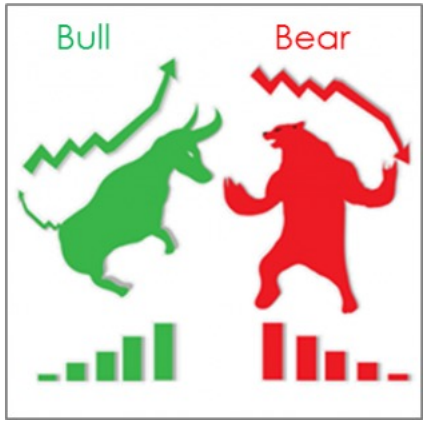
AURIN
aurin.org.au

Supervised models:

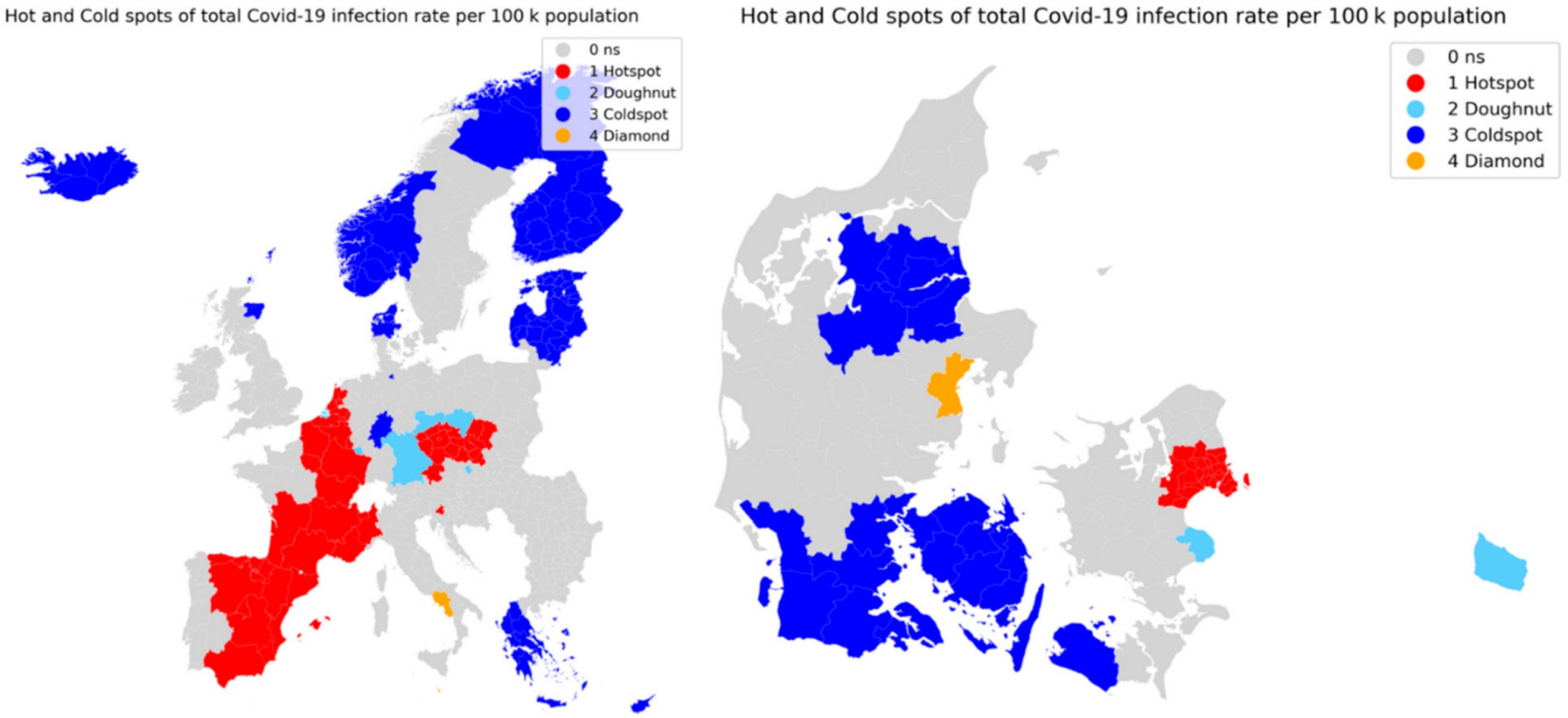
Supervised learning uses as training a data set that contains a mark or label on the data. In this type of learning, the distinction between the independent and dependent variables is clear.

Regression vs Classification





Spatial autocorrelation - Example



Source: Hass, Frederik Seeup, and Jamal Jokar Arsanjani. "The geography of the COVID-19 pandemic: A data-driven approach to exploring geographical driving forces." International Journal of Environmental Research and Public Health 18.6 (2021): 2803.

Unsupervised models

Unsupervised models

Unsupervised learning does not require prior labelling of the data, it uses all the information to make associations between the data or group them.

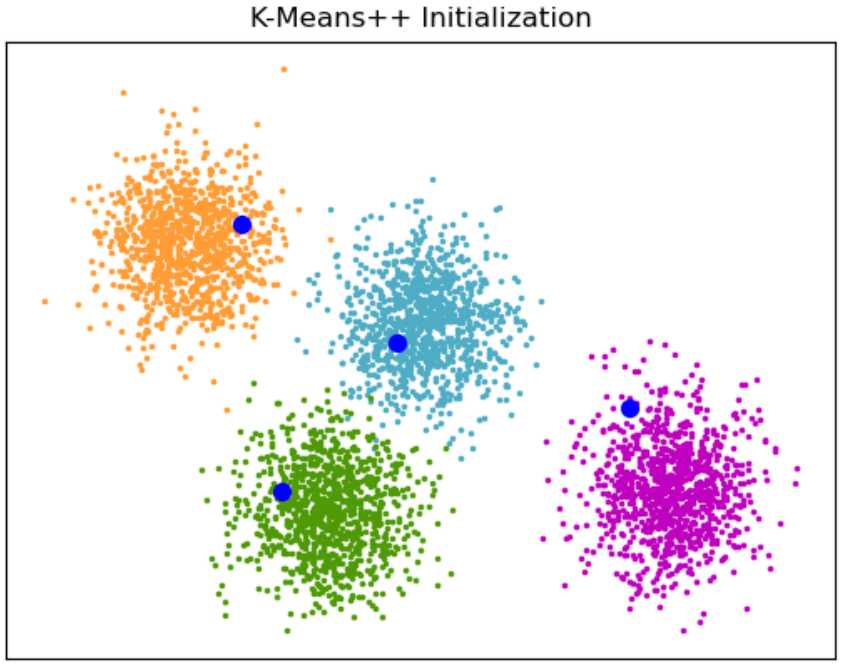


Image from: Scikit-learn