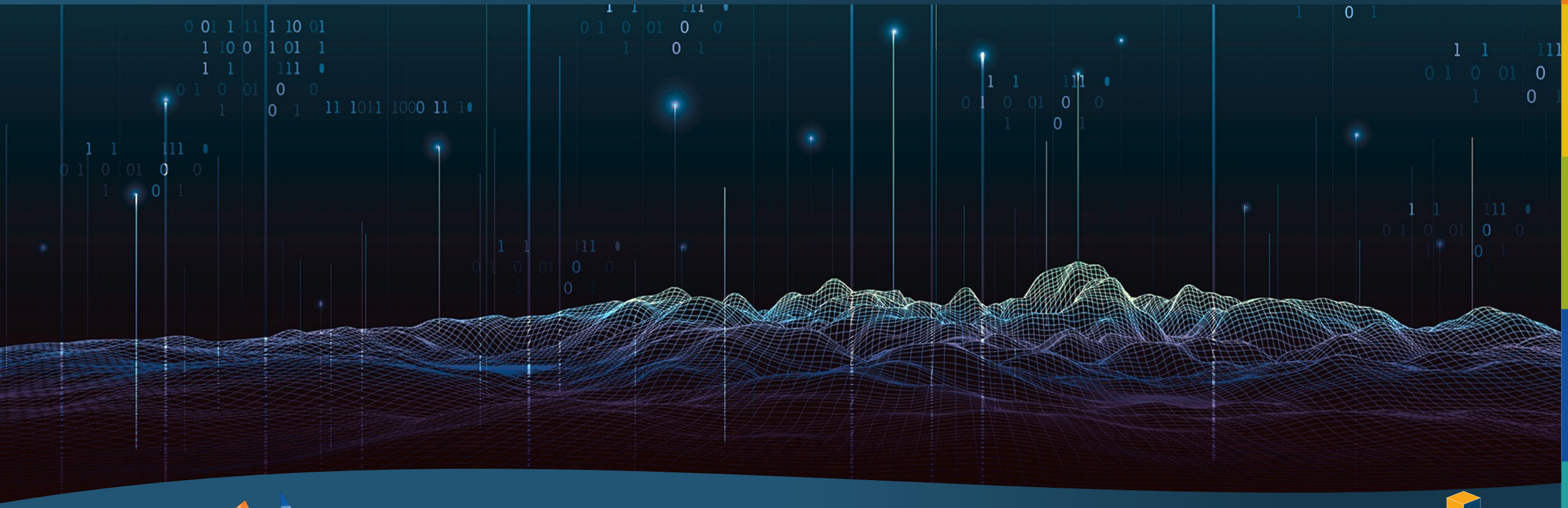
Spatial data and data integration

Australian Urban Research Infrastructure Network (AURIN)



AURIN
aurin.org.au

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.

Introduction to geospatial data

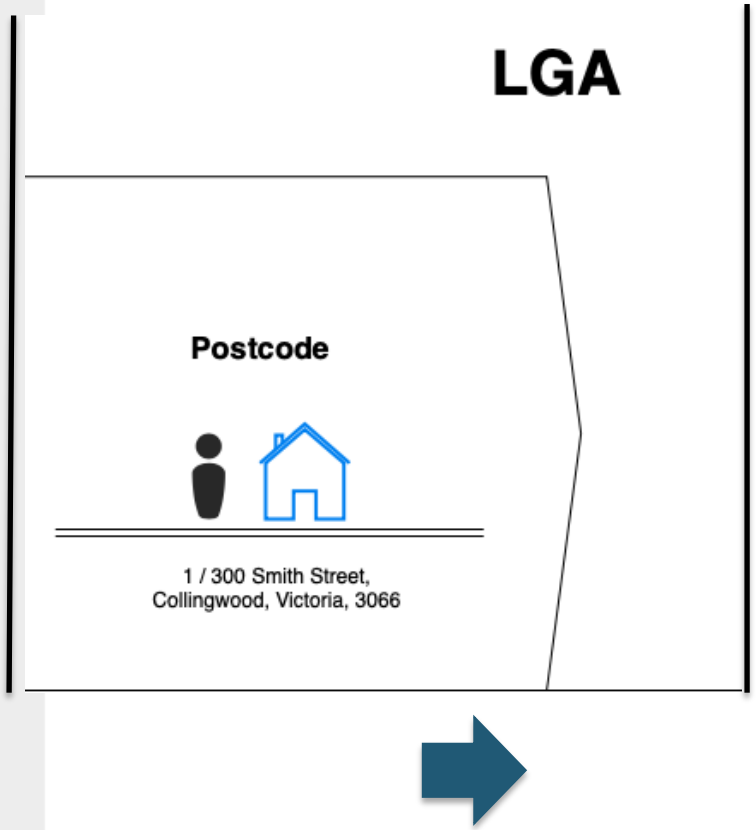
Finding and using spatial data

Deciding on integration

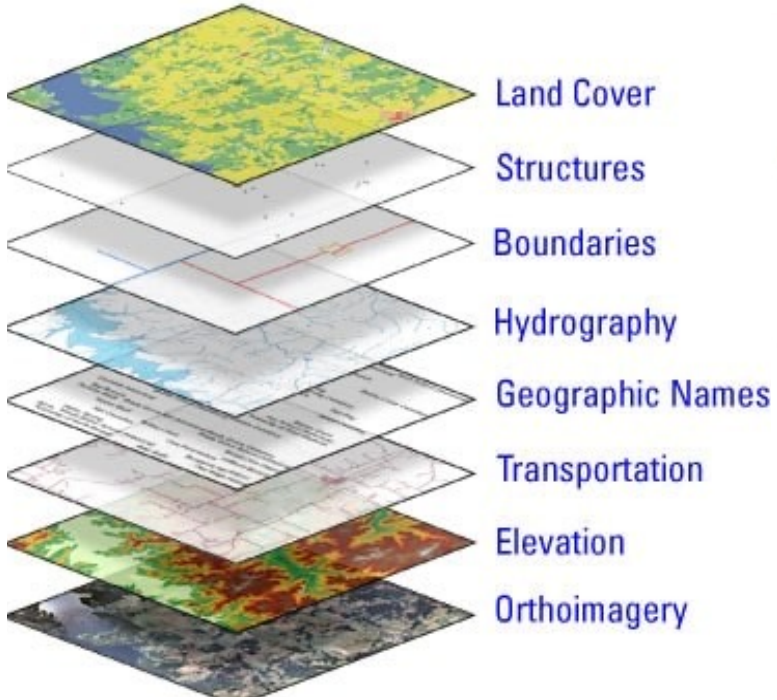
Producing a new data product

The Geosocial work package

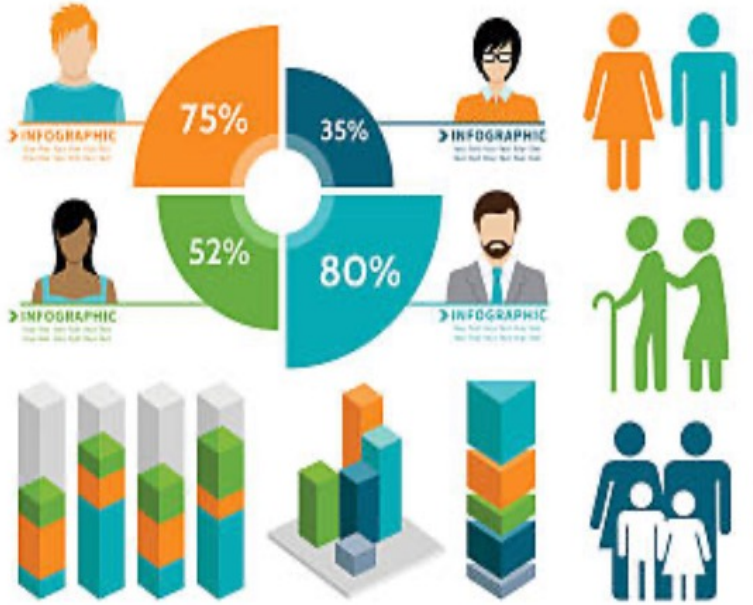
What is spatial data?



Physical



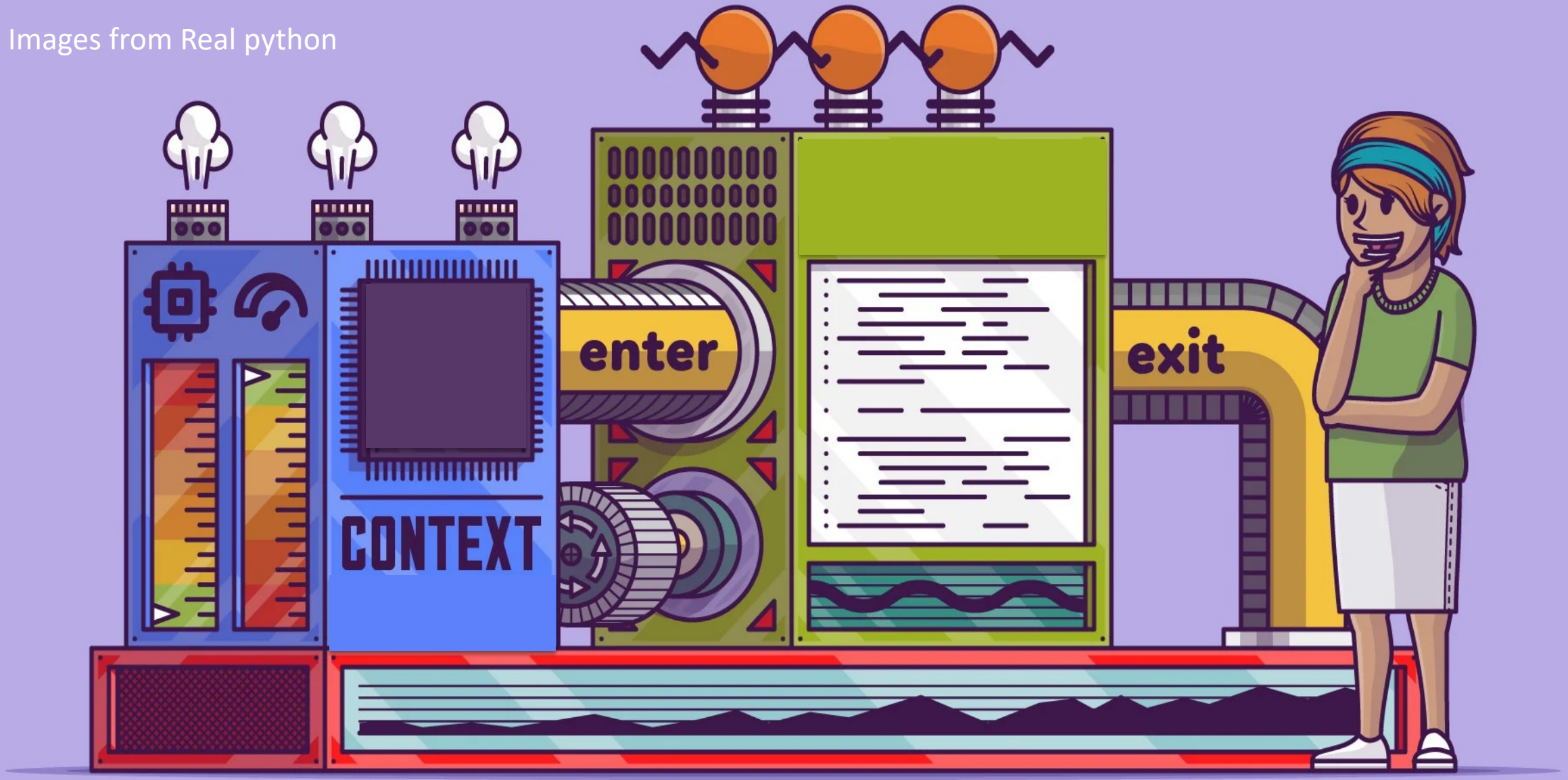
Socio-demographic



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

Data enrichment

Images from Real python

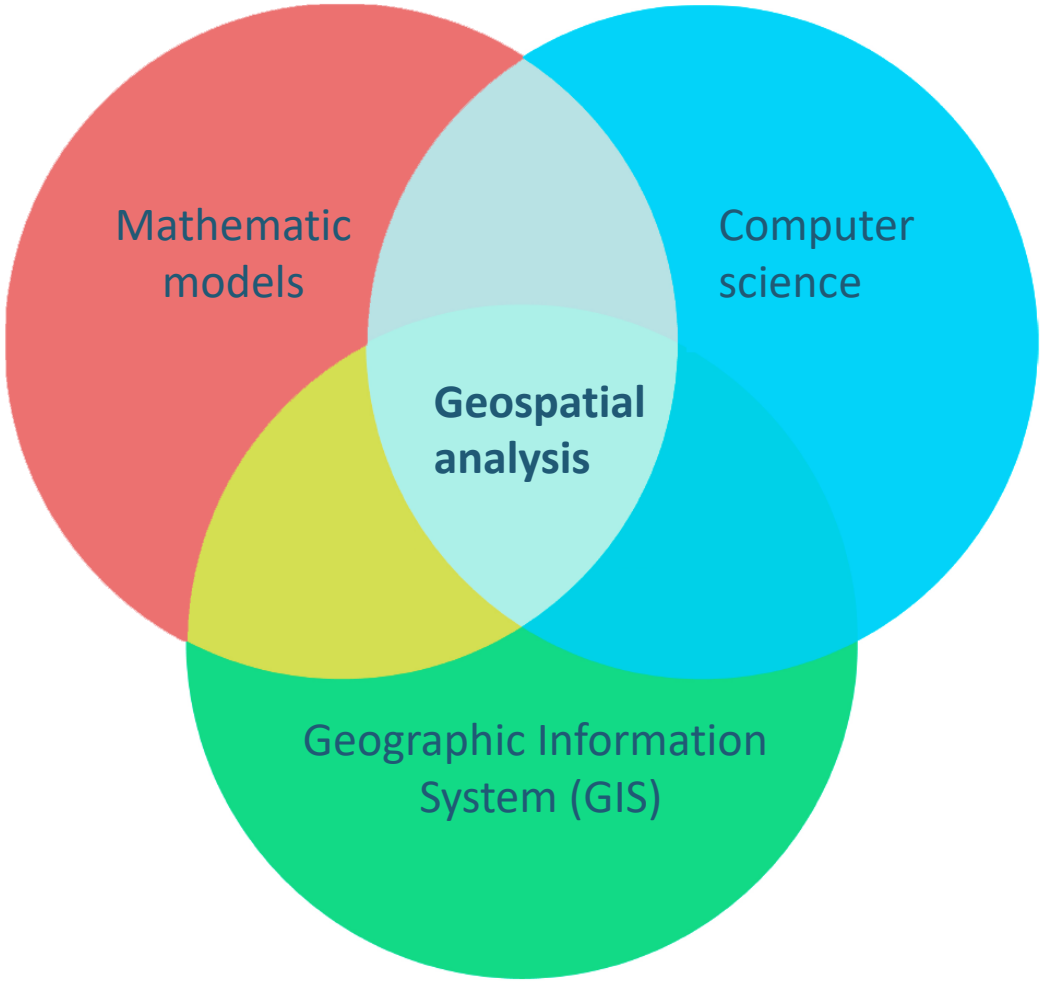


Geospatial analysis

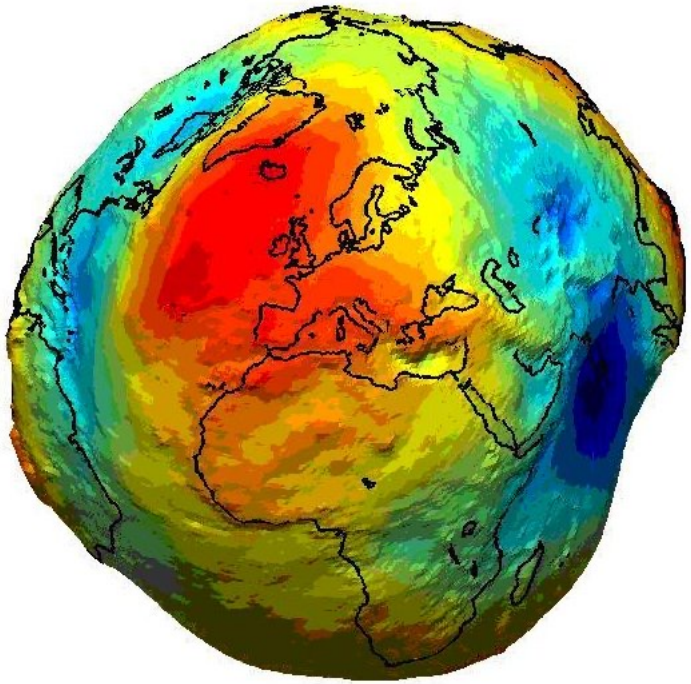
Geospatial analysis: represents a collection of **techniques** and **models** that explicitly use the spatial referencing of each data case.

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

Geospatial Analysis



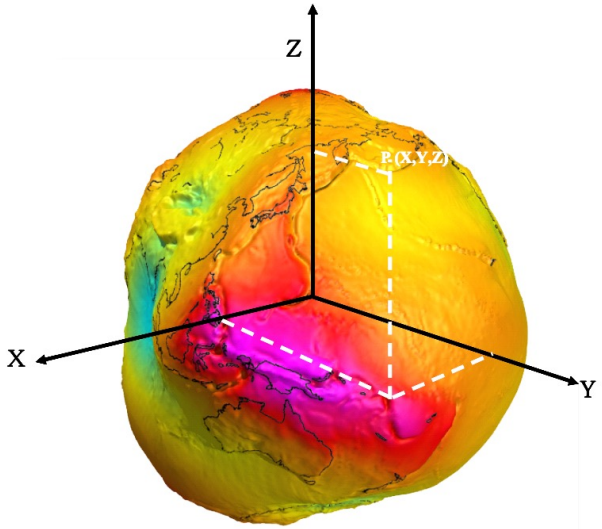
Geoid



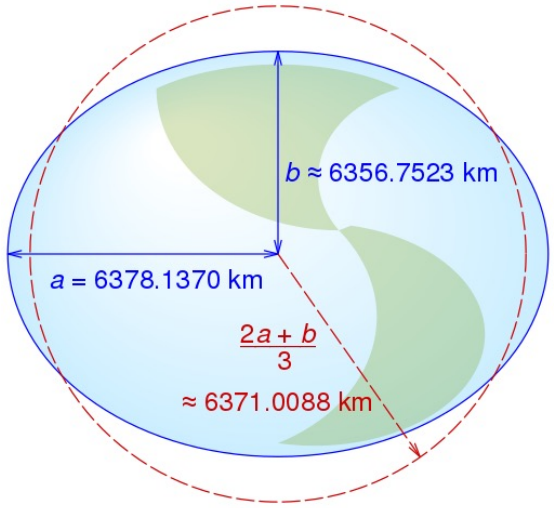
A spheroid shape with a slight flattening at the poles represents the Earth's vertical composition profiles:

- It is a theoretical representation
- Gravity has the same value on the entire surface
- There is a correspondence with the mean sea level which is taken as zero level
- Allows you to calculate altitudes

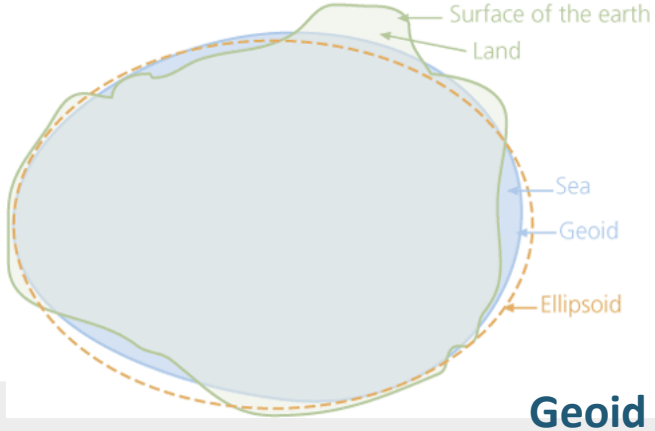
World Geodetic System 1984 (WGS84)

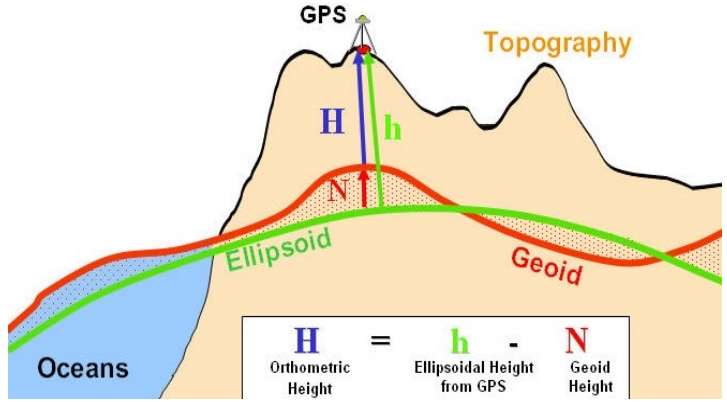


Geoid



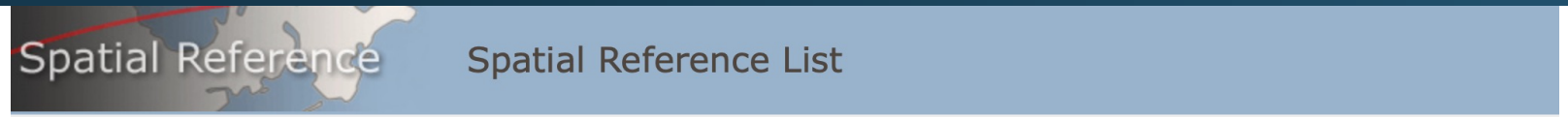
Ellipsoid



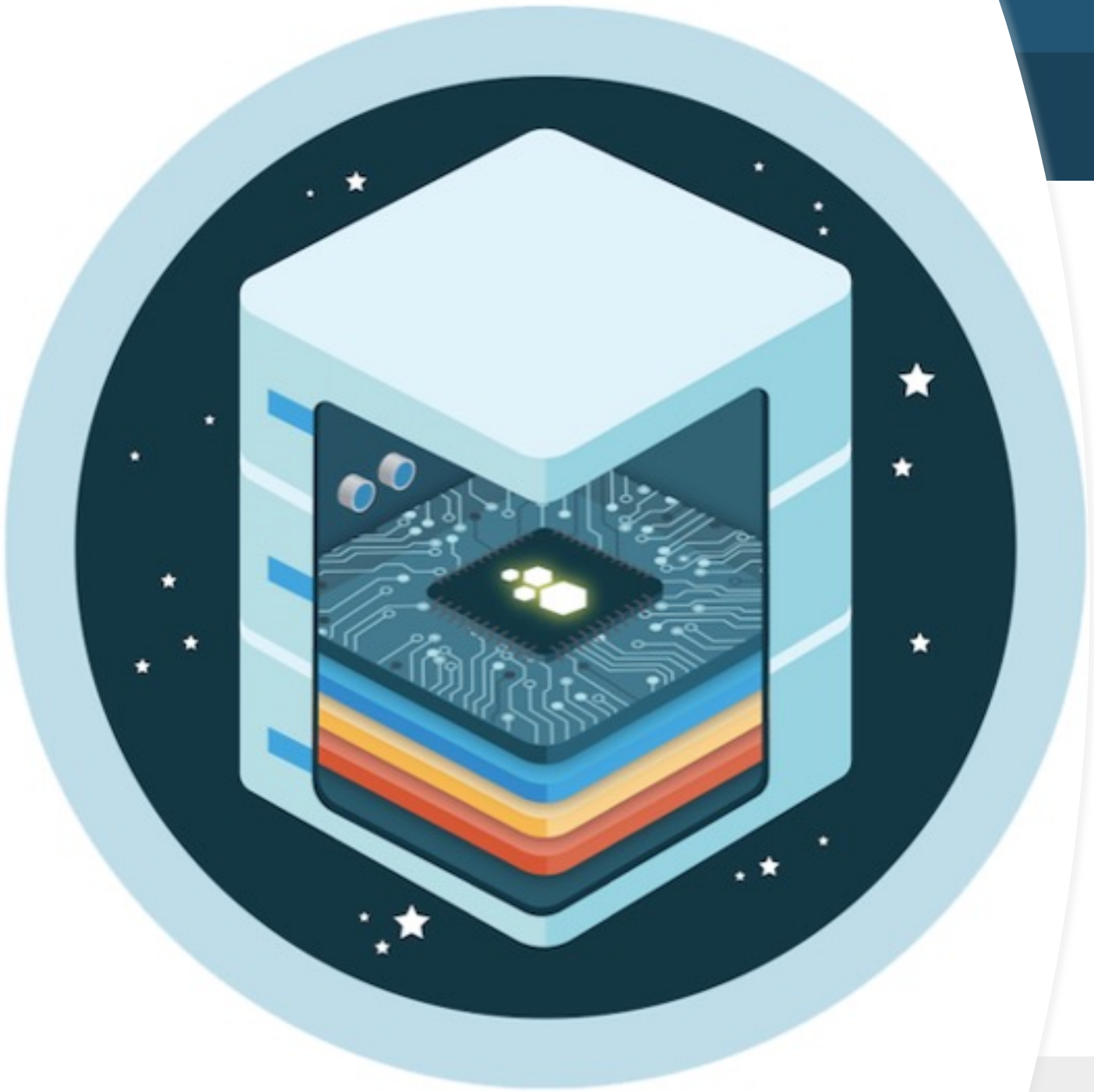


Geoid vs Ellipsoid

EPSG



Source: Spatial reference

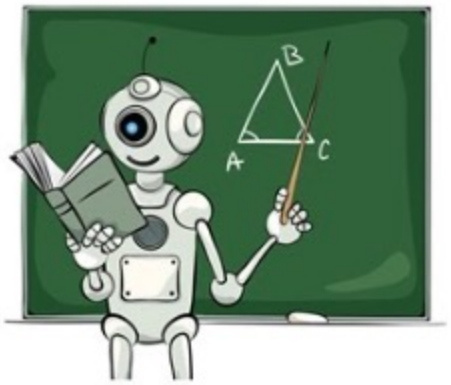


Machine Learning

Machine Learning (ML) is the use of algorithms and statistical models to enable computer systems to learn from data and improve on specific tasks without being explicitly programmed.

Machine learning

Supervised models: They require a training sample that is previously marked with a label.



Unsupervised models: They do not require a previously marked label. They are algorithms that look for patterns within data without prior knowledge.

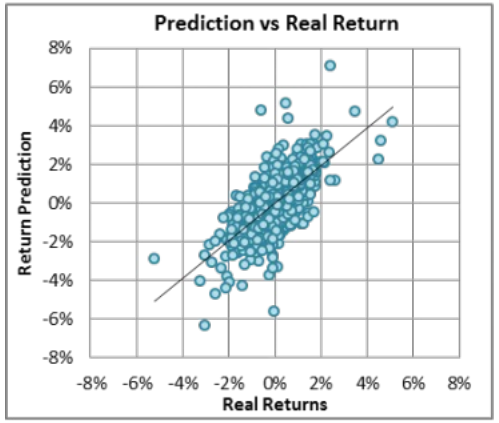
Supervised vs unsupervised learning

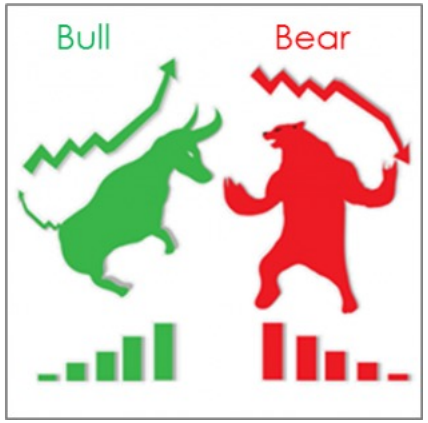
AURIN
aurin.org.au

Supervised models:

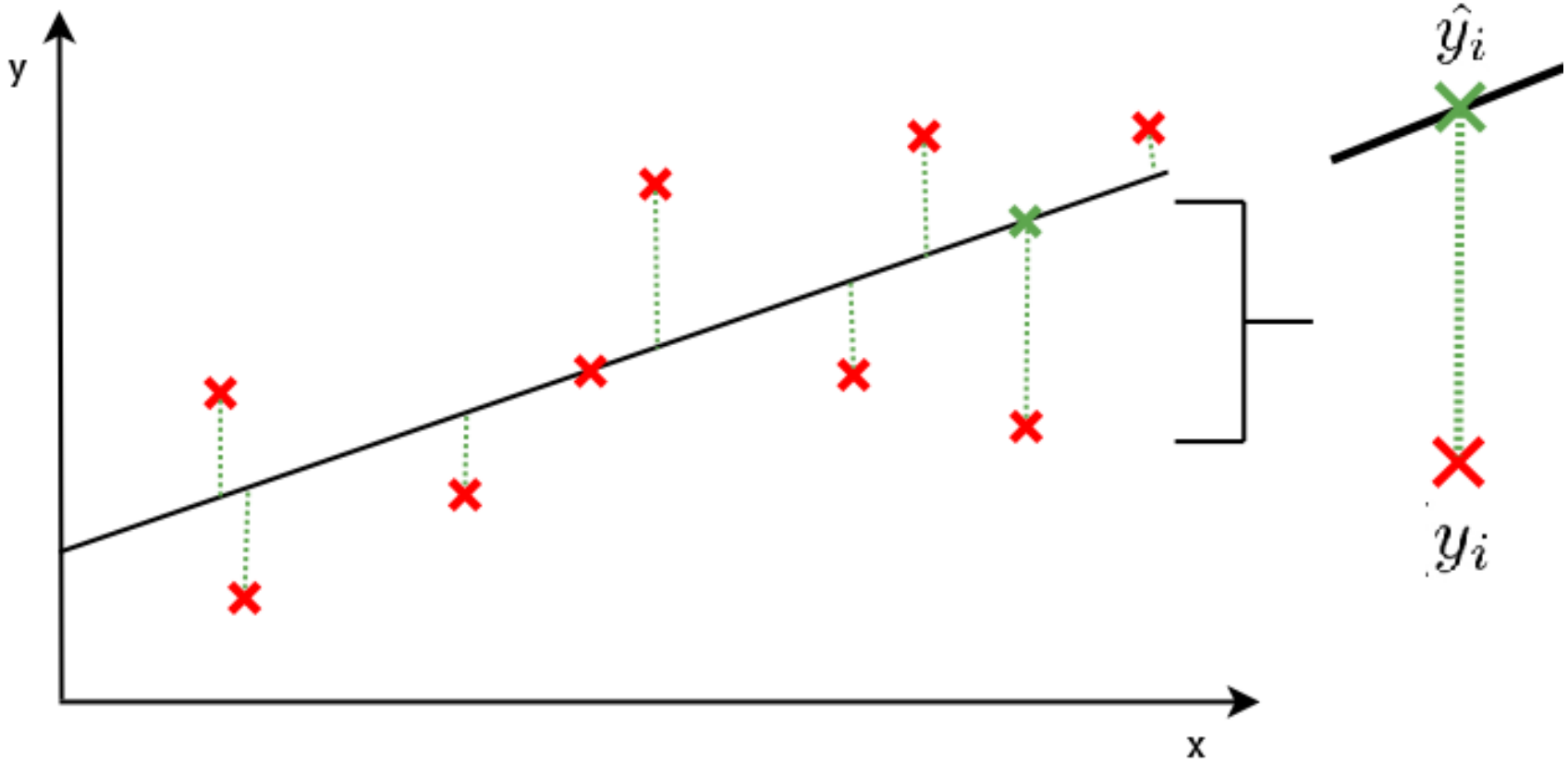
Supervised learning uses as training a data set that contains a mark or label on the data. In this type of learning, the distinction between the independent and dependent variables is clear.

Regression vs Classification





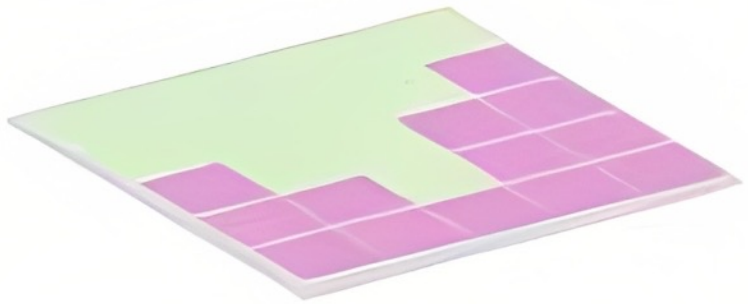
Supervised vs unsupervised learning



Spatial autocorrelation - Example

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

Waldo Tobler



High Spatial Autocorrelation
(Clustering)

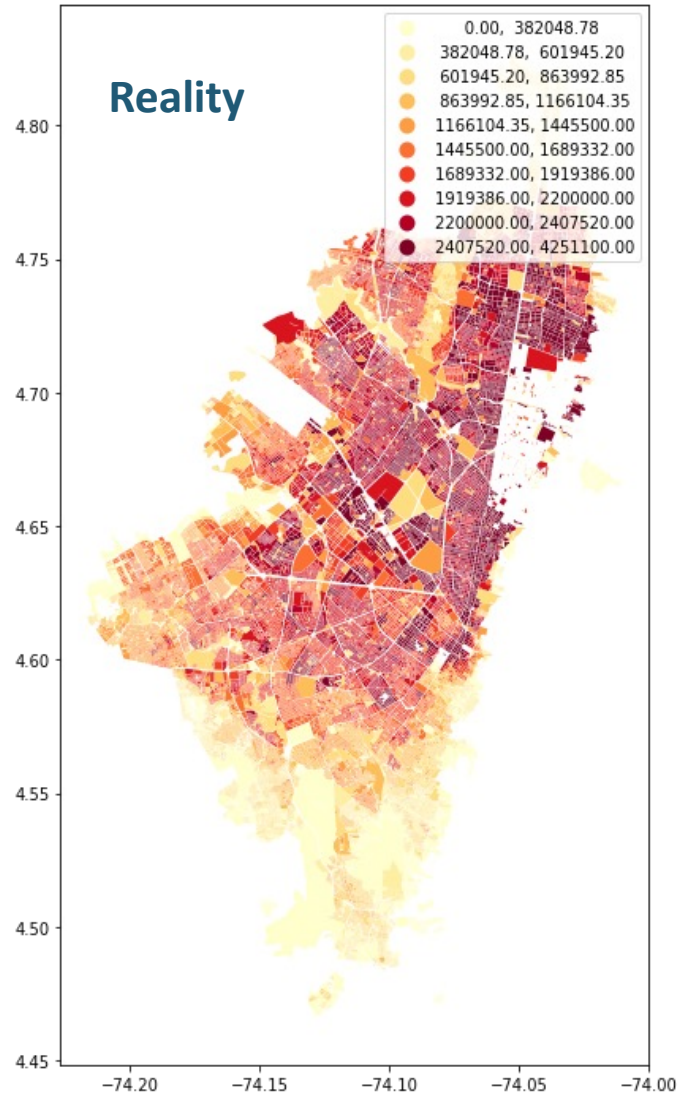
Low Spatial Autocorrelation
(Checkerboard)

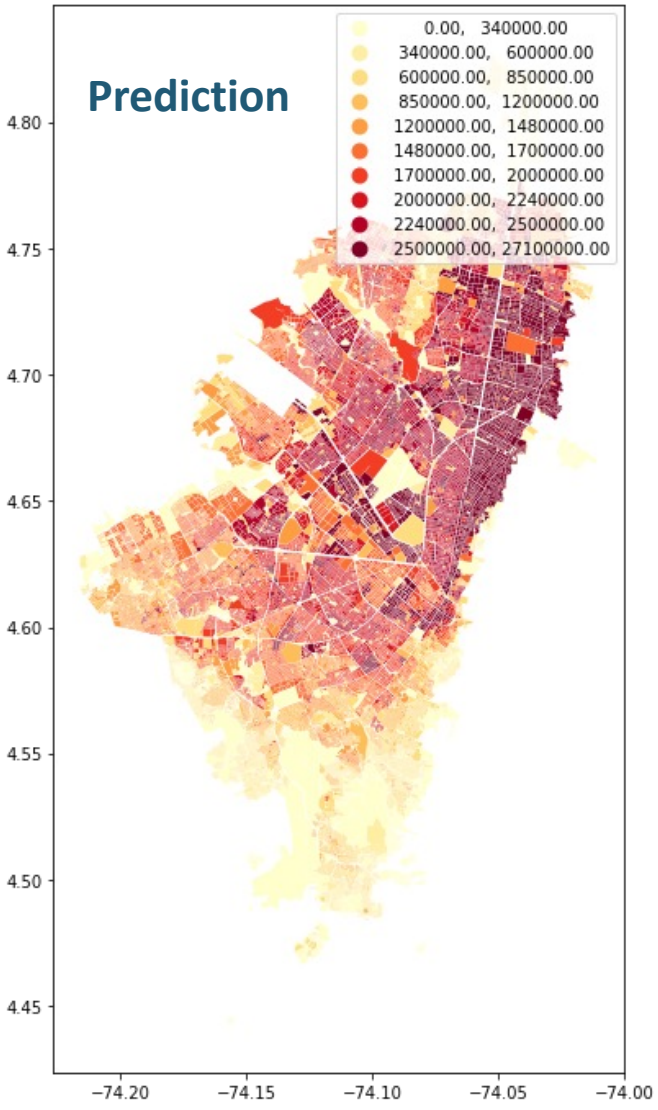
Images from: GisGeography

High spatial autocorrelation

Image from: Domain

Example: Housing price - Geographically weighted regression





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.

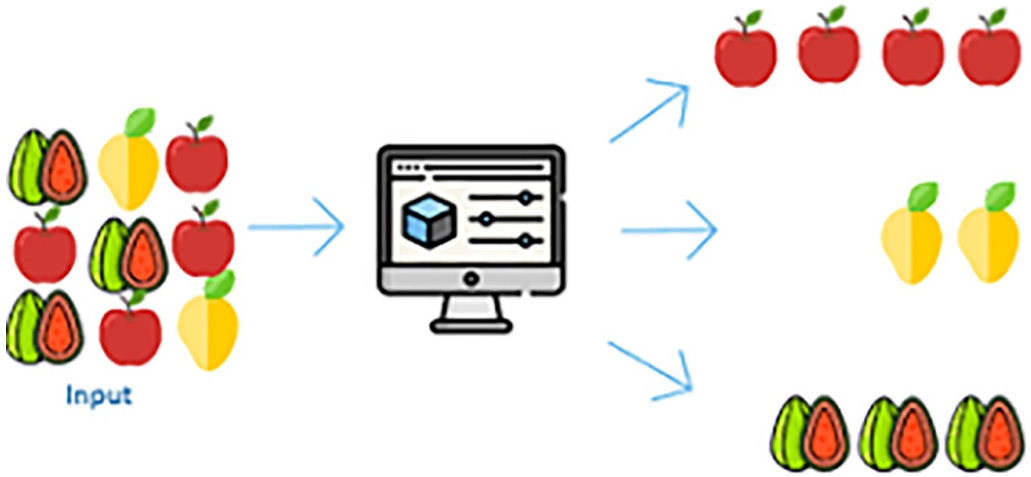
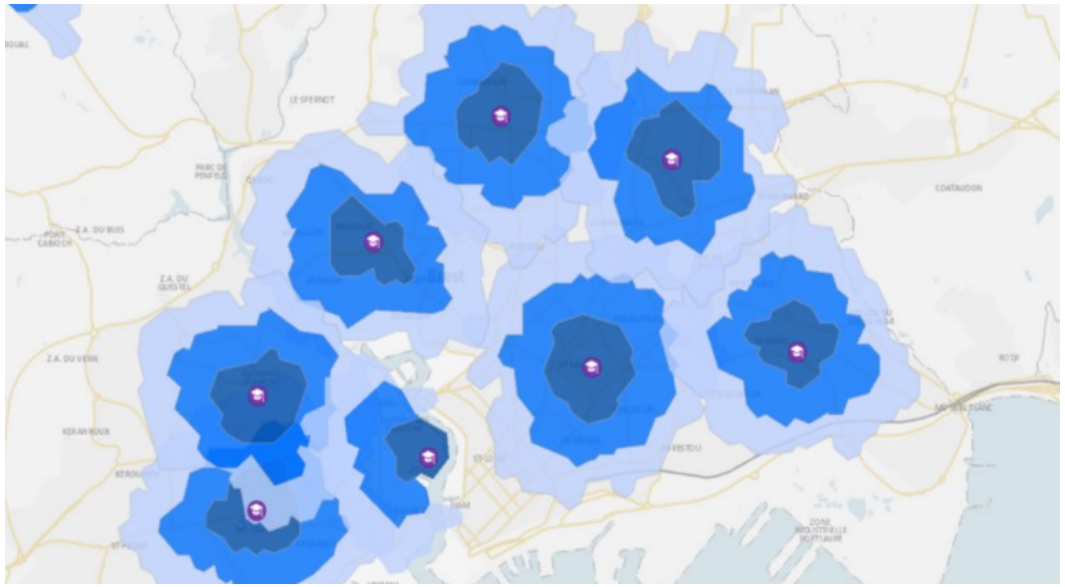


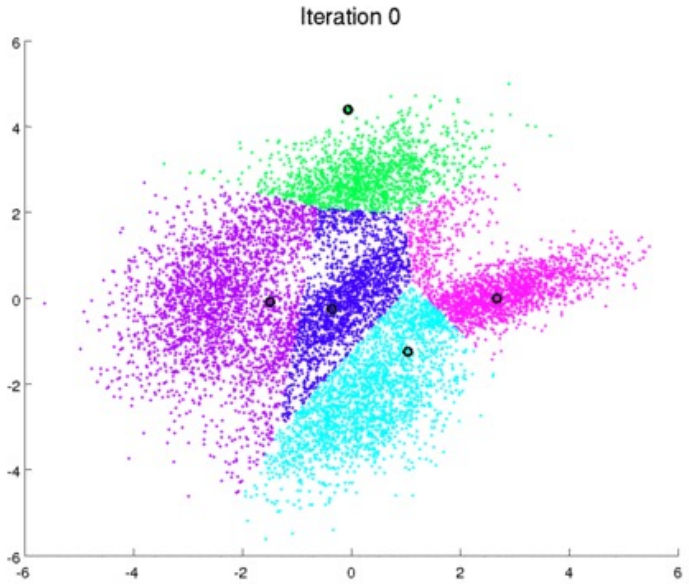
Image from: Scikit-learn

Unsupervised models

Geospatial clustering



K-means



K-prototypes

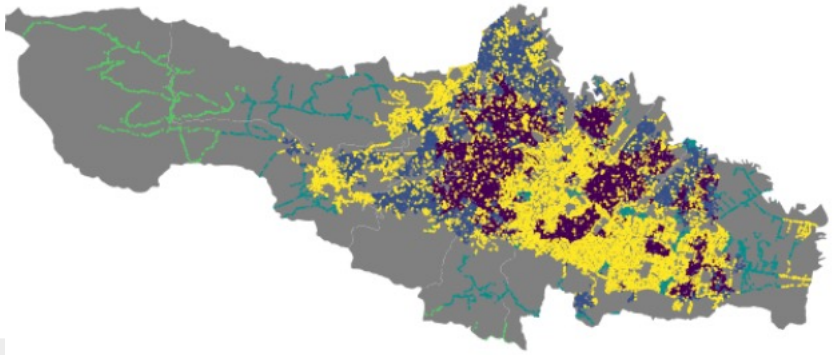
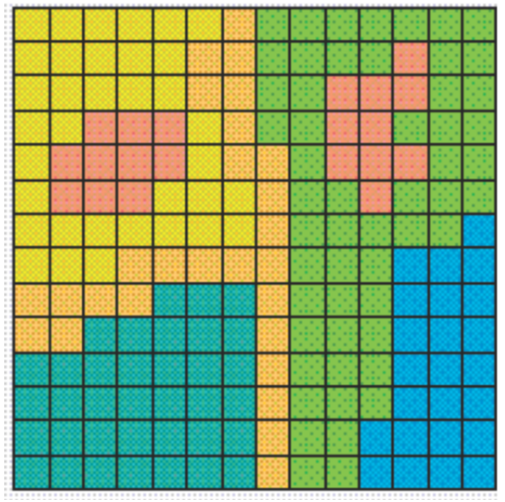


Image from: Scikit-learn

Finding and using spatial data

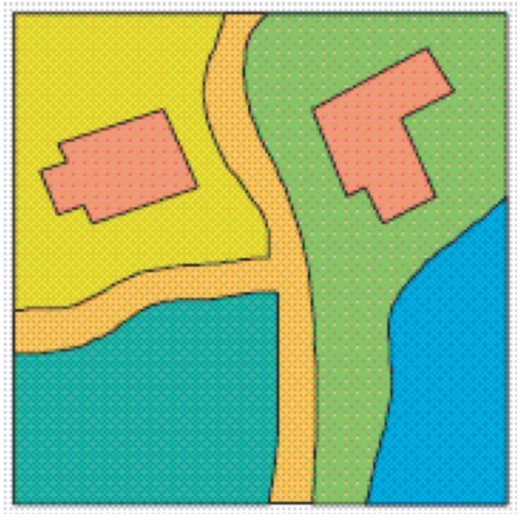
Raster map:

- 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.

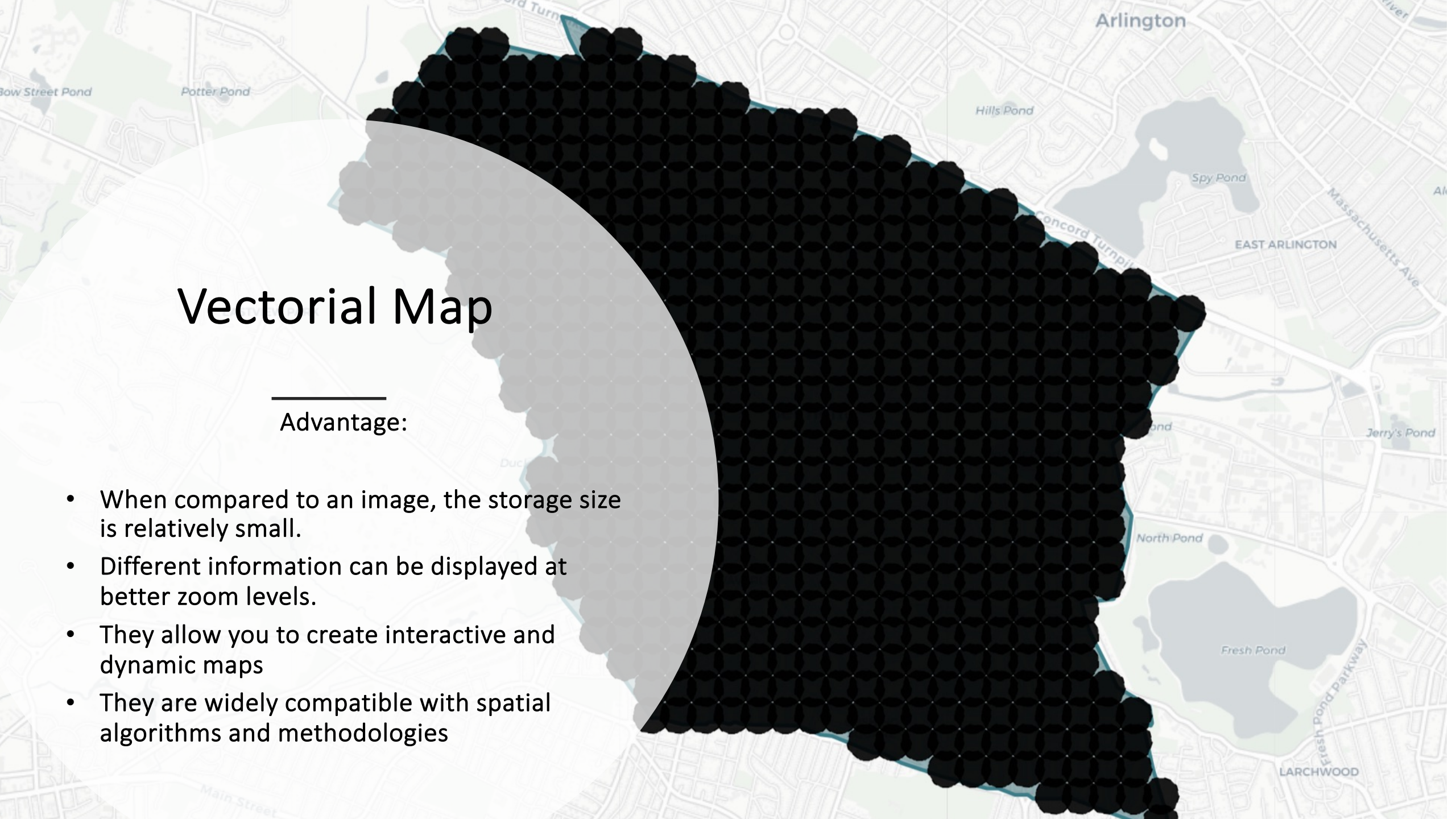


Images from: ESRI

Vectorial Map

Advantage:

- When compared to an image, the storage size is relatively small.
- Different information can be displayed at better zoom levels.
- They allow you to create interactive and dynamic maps
- They are widely compatible with spatial algorithms and methodologies



Vectorial map:

Type of elements:

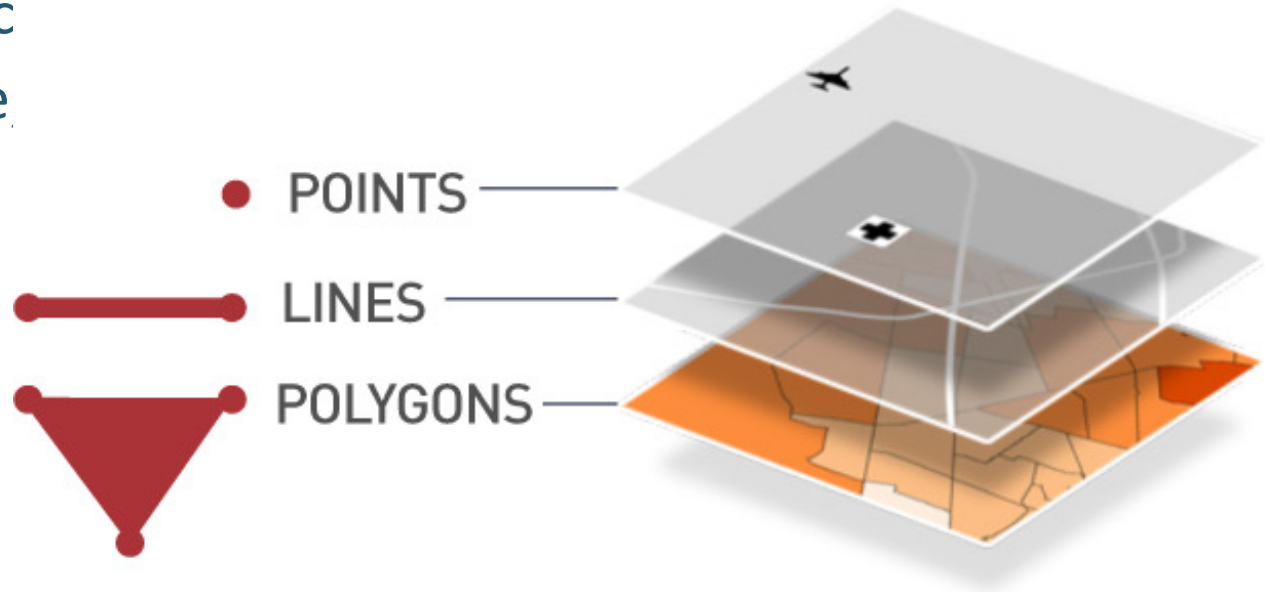
Point: Addresses, locations, points of interest, etc

Lines: streets, freeways, borders, etc

Polygons: Countries, cities, Cadastre

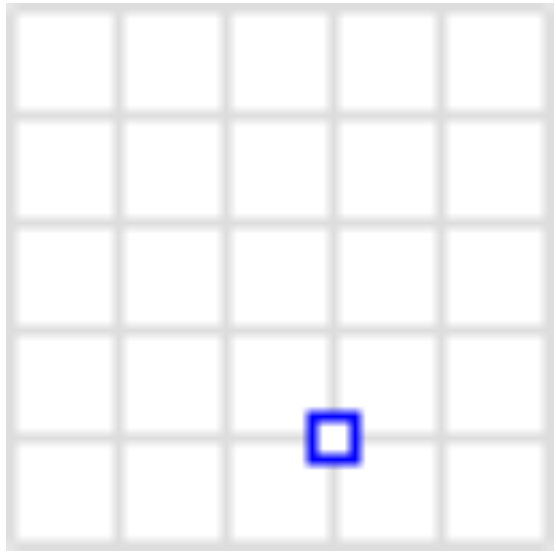
Advantages:

- Spatial operations
- Spatial aggregation
- Spatial Join



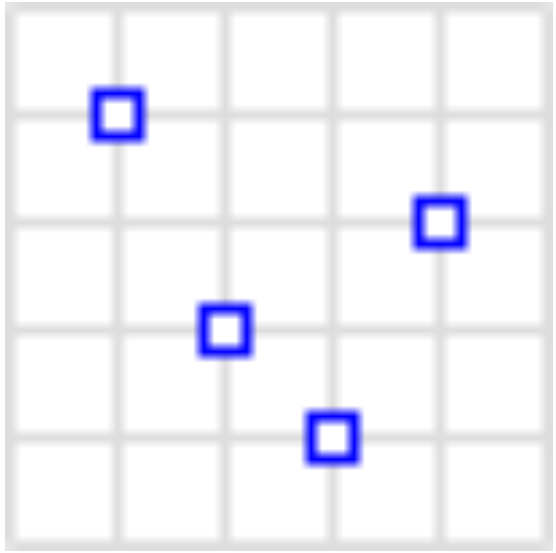
Images from: Berkeley

Vectorial map:



Point

It represents a point on the Earth's surface (latitude, longitude)

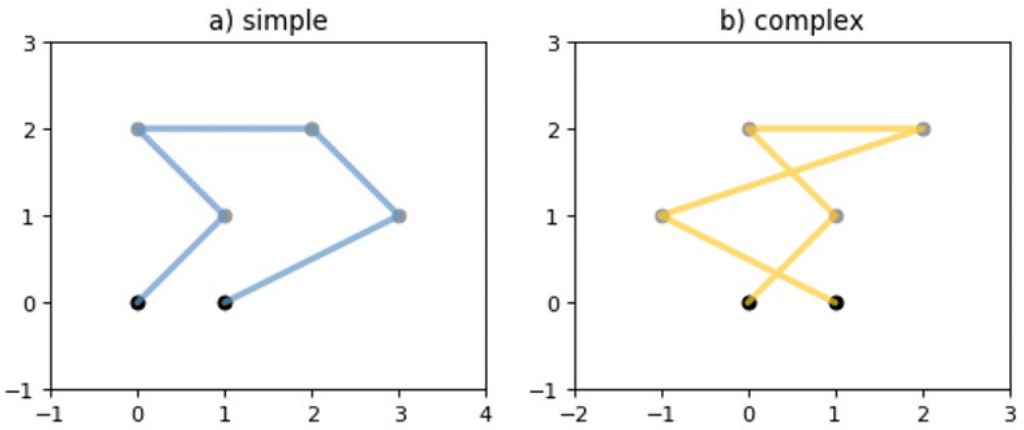


MultiPoint

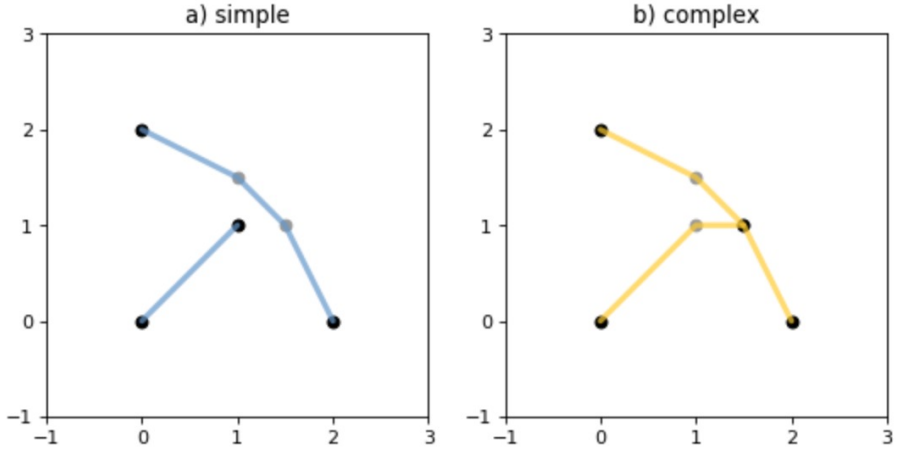
Constructed by multiple points that cannot be connected.

Images from: Berkeley

Vectorial map:



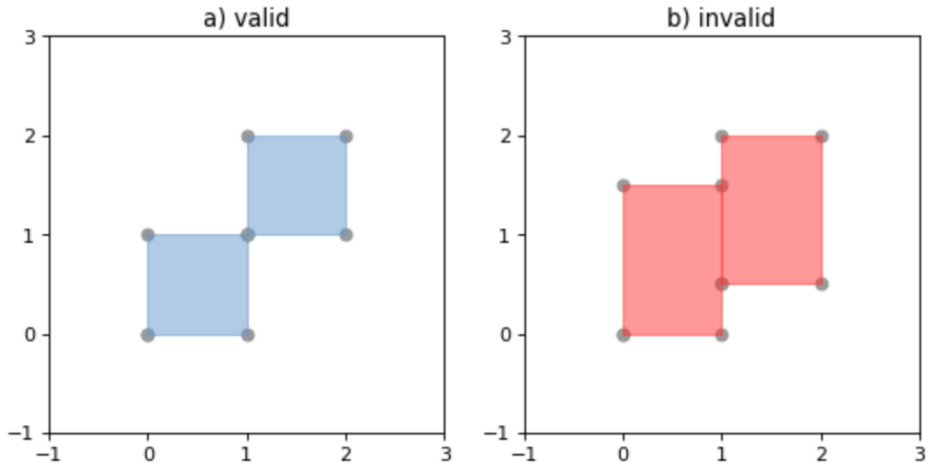
Linestring



Multilinestring

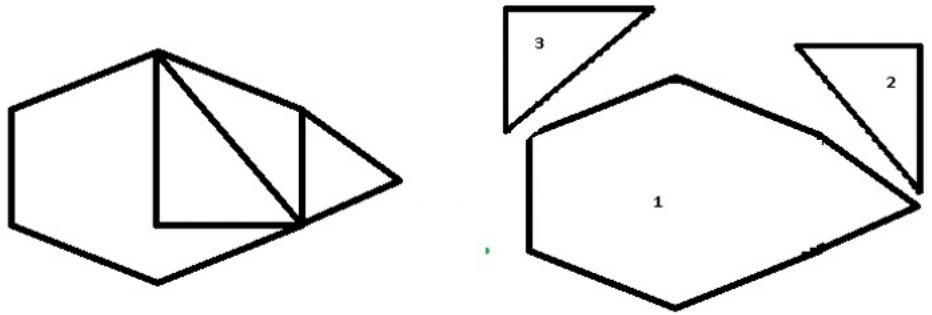
Images from: Berkeley

Vectorial map:



Polygon

It is a closed figure made up of coordinates that enclose a region in the plane.

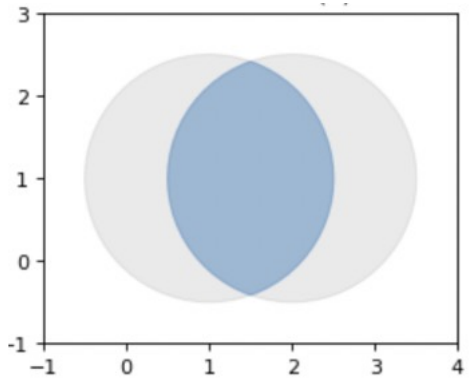


Multipolygon

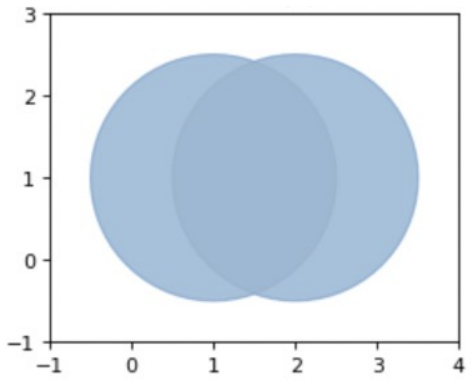
A collection of polygons. It is useful for gathering a group of Polygons into one geometry

Images from: Berkeley

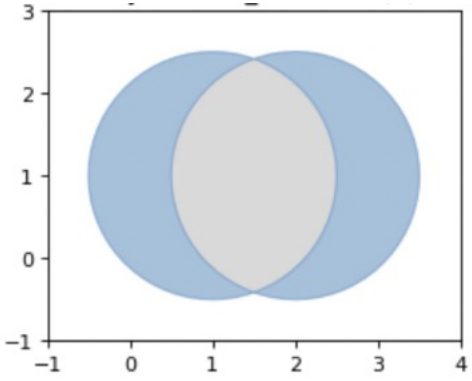
Spatial operations



Intersection

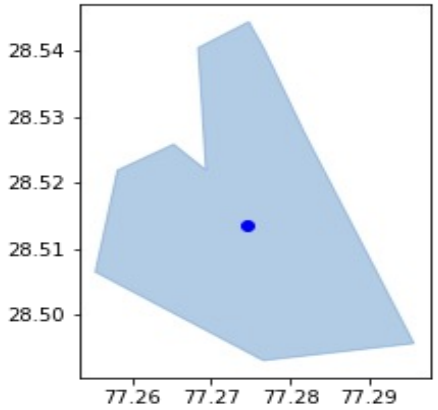


Union

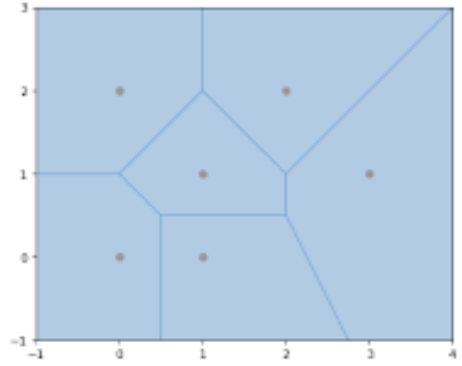


Difference

Contour



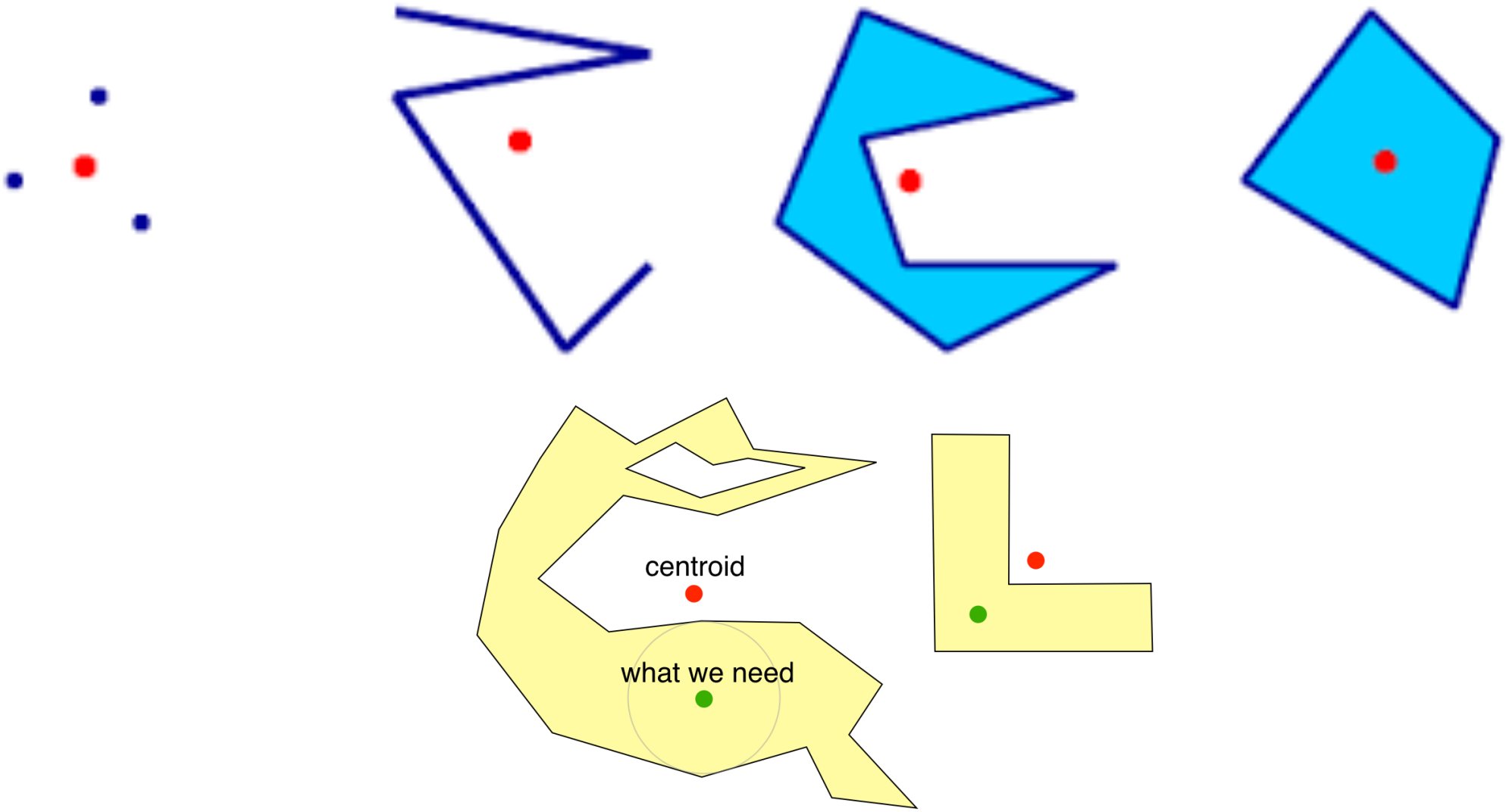
Centroid



Within

Images from: Pysal

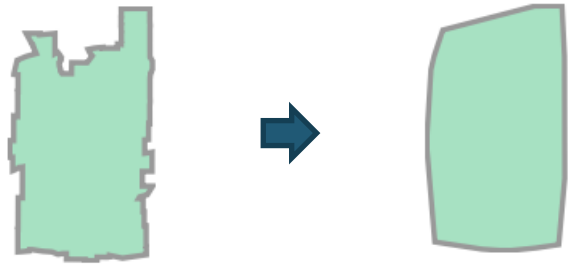
Spatial operations



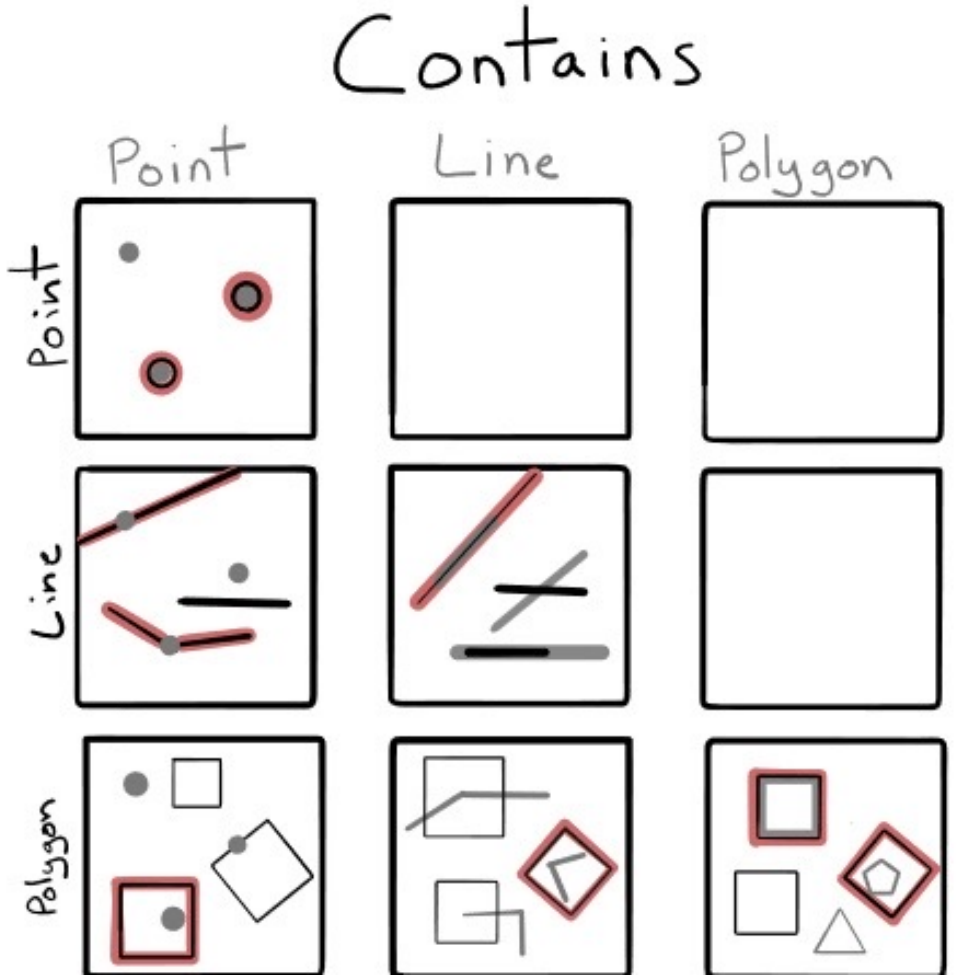
Images from: OpenGeo and MapBox

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.



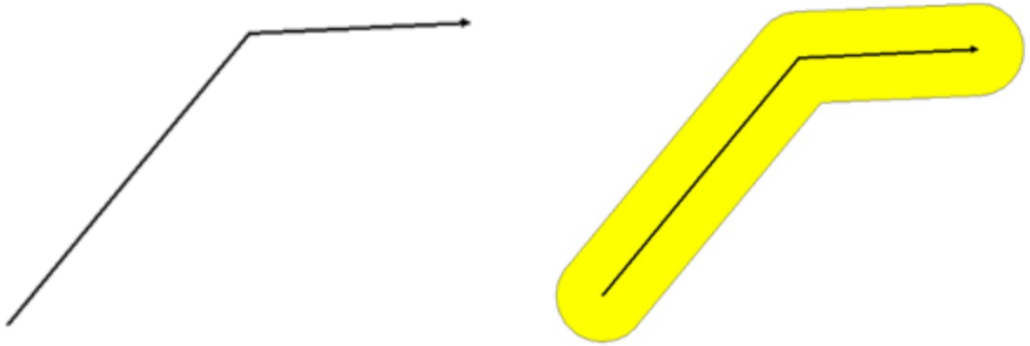
Convex capsule: a set of points that contains the intersection of all convex sets within a polygon.

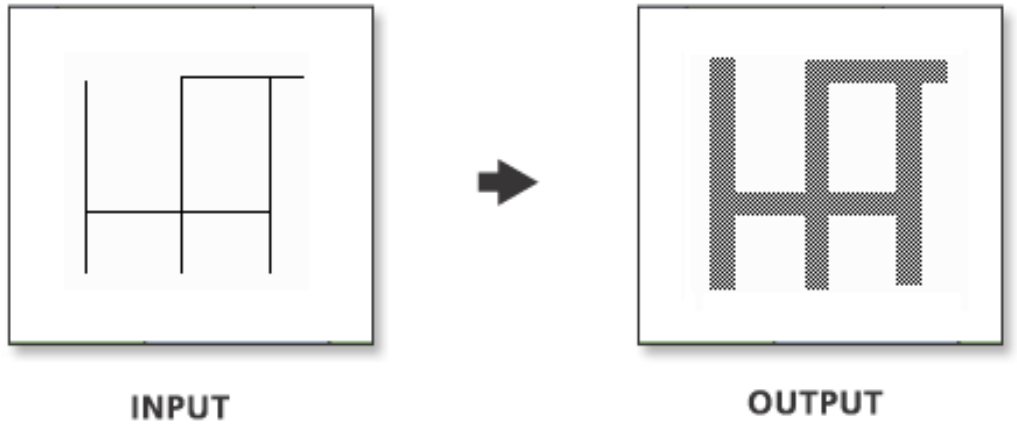


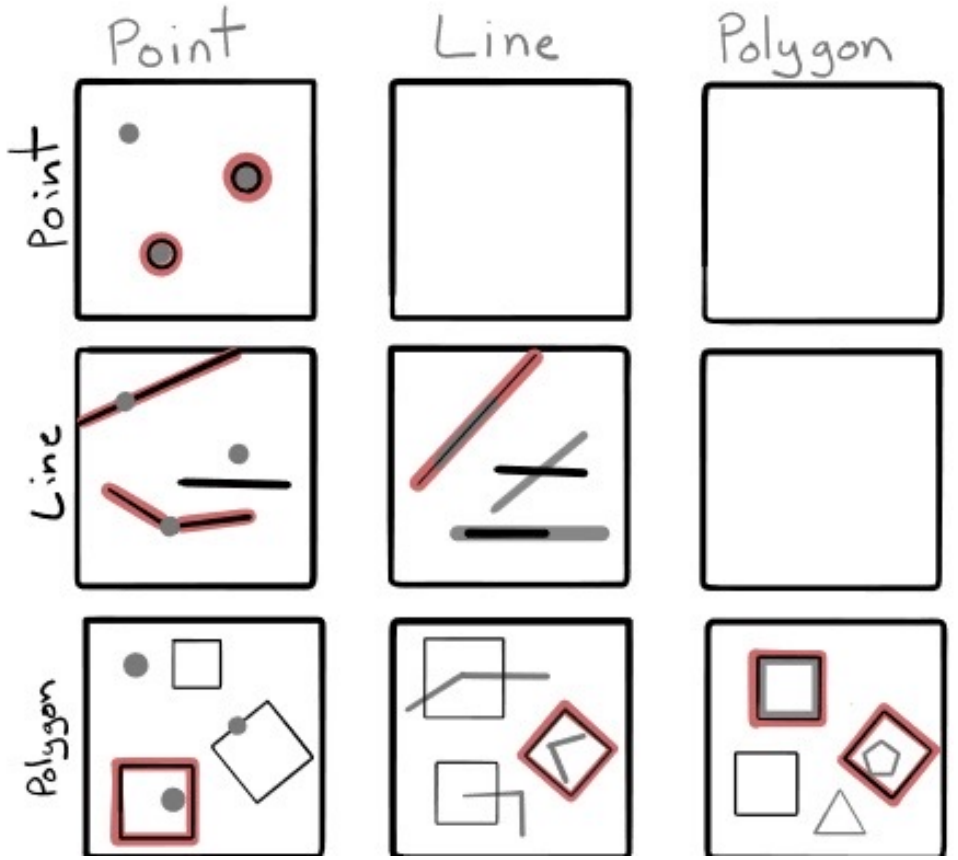
Images from: Pysal & PyGis

Spatial operations

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



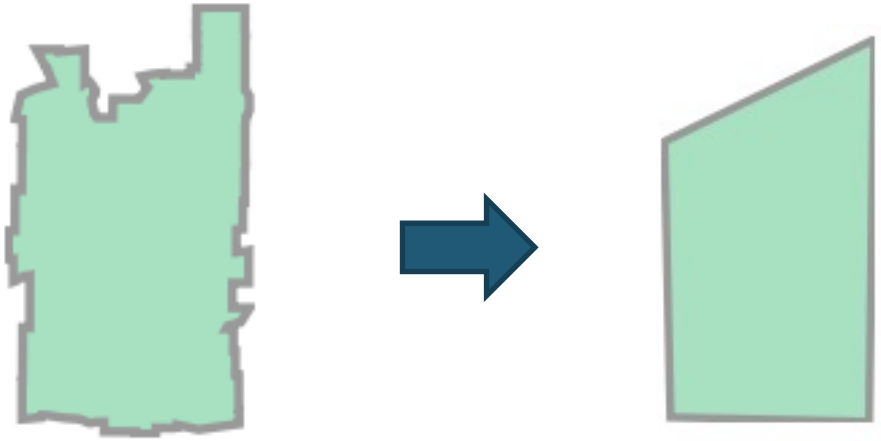


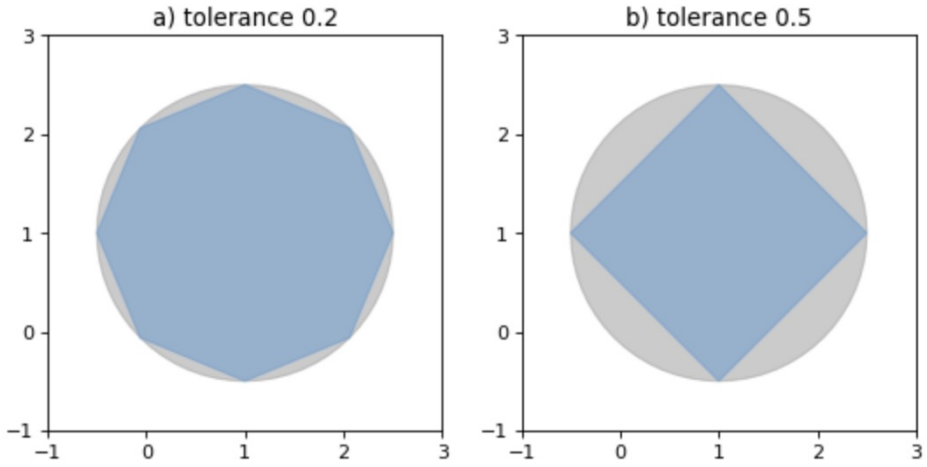


Images from: Pysal & PyGis

Spatial operations

Simplification



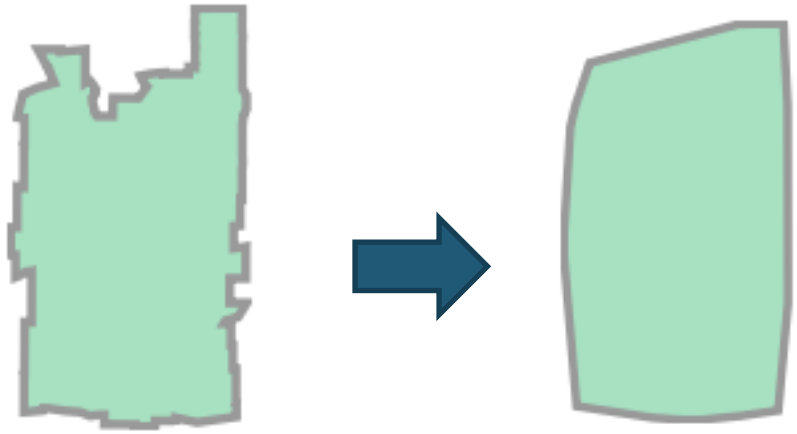


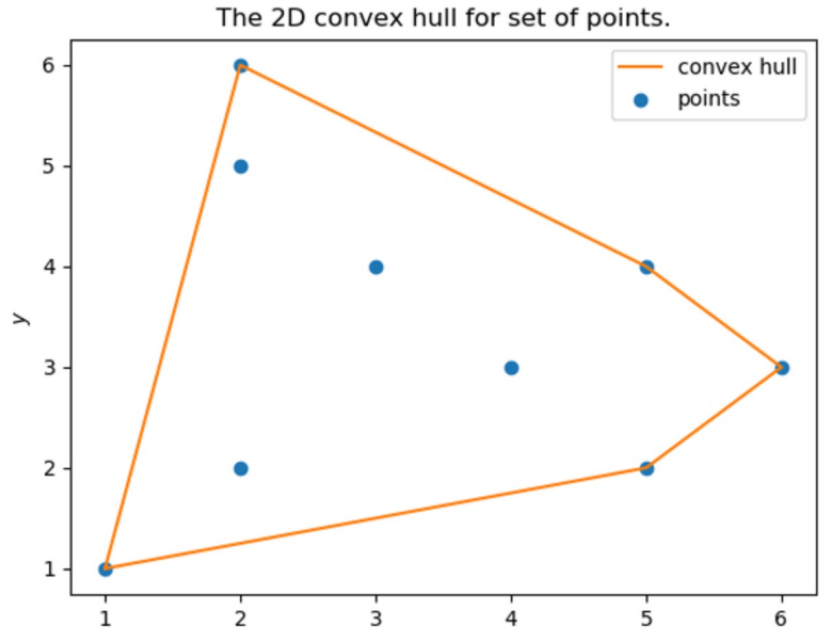
This function simplifies the geometry using a specified tolerance.

All points of the simplified object will be within the tolerance distance from the original geometry.

Images from: Pysal & PyGis

Spatial operations

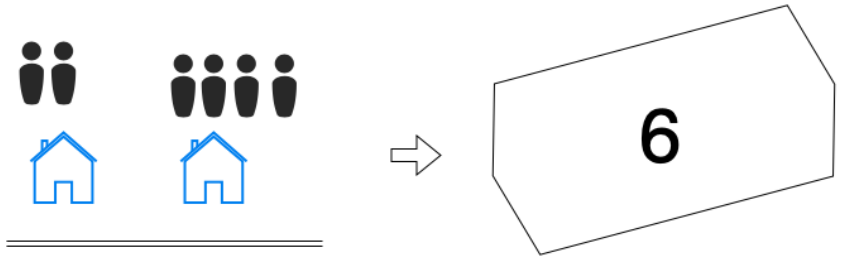


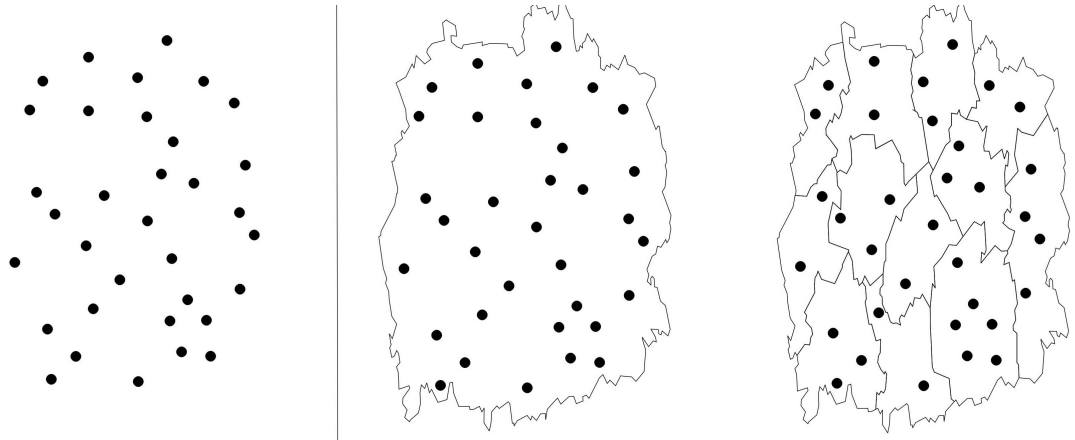


Convex capsule: a set of points that contains the intersection of all convex sets within a polygon.

Images from: Pysal & PyGis

Spatial aggregation





Potential areal data

Possible aggregations

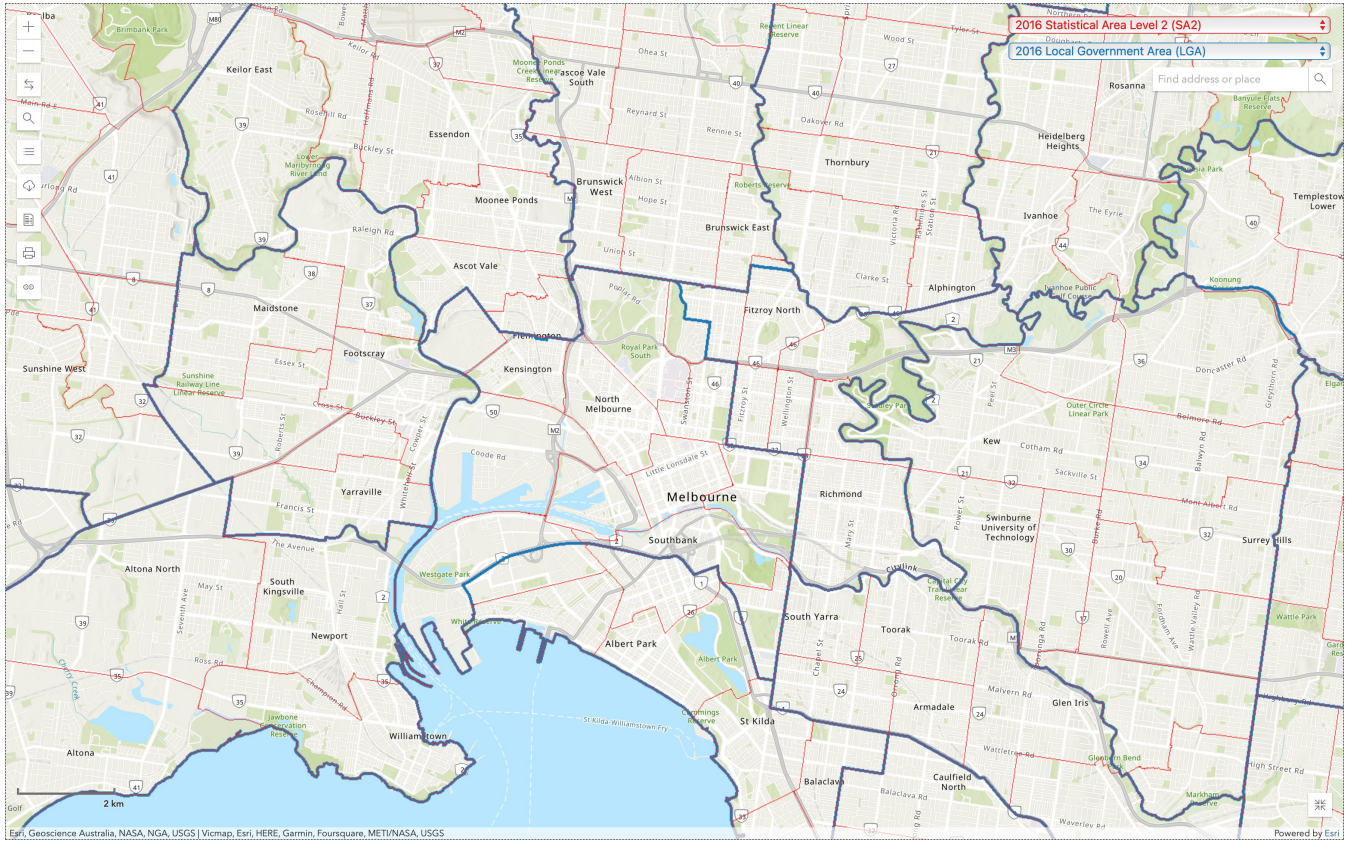
Images from: Dusek, Tamás. (2004)

Example: Spatial aggregation

AURIN
aurin.org.au

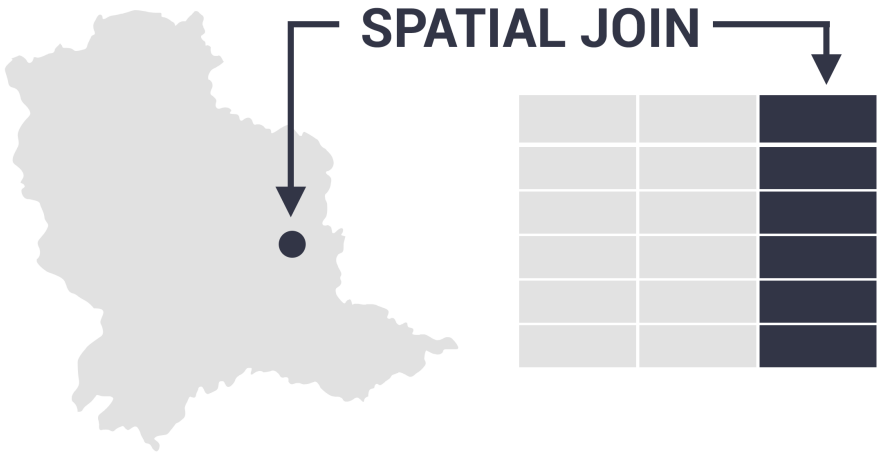
Socio-Economic Indexes for Areas (SEIFA) – SA2 - 2016

PHIDU - Education (LGA) 2015-2016:

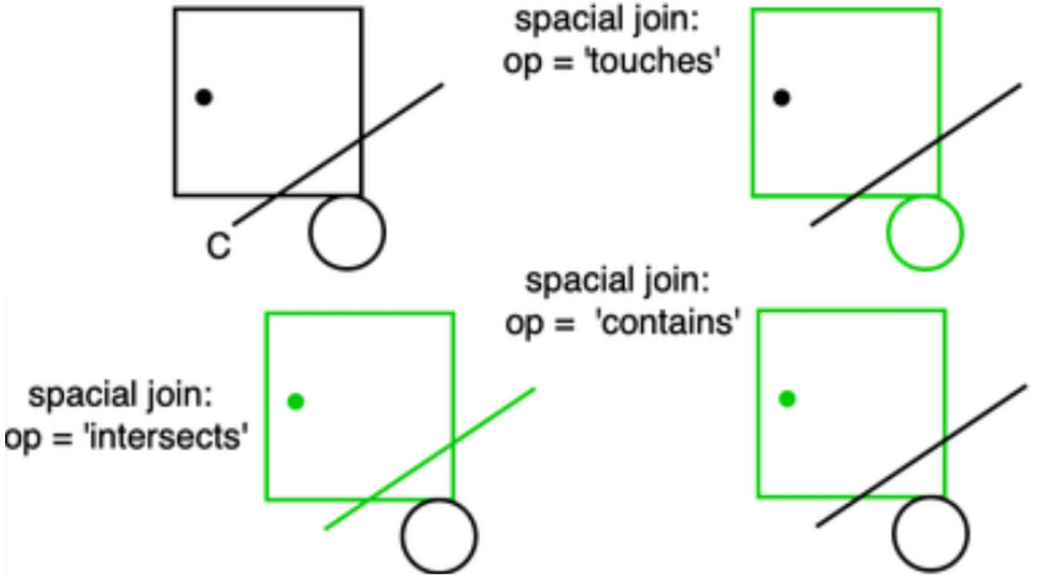


Images from: ABS

Spatial Join



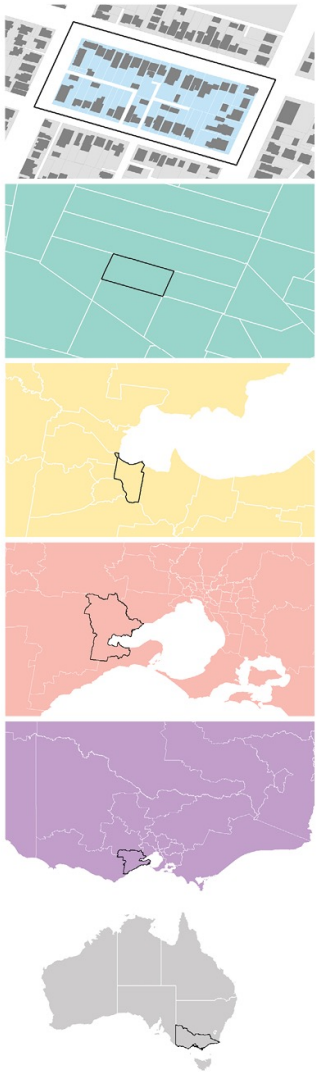
Types of spatial join



Images from: Pygis & Harder, Dorian

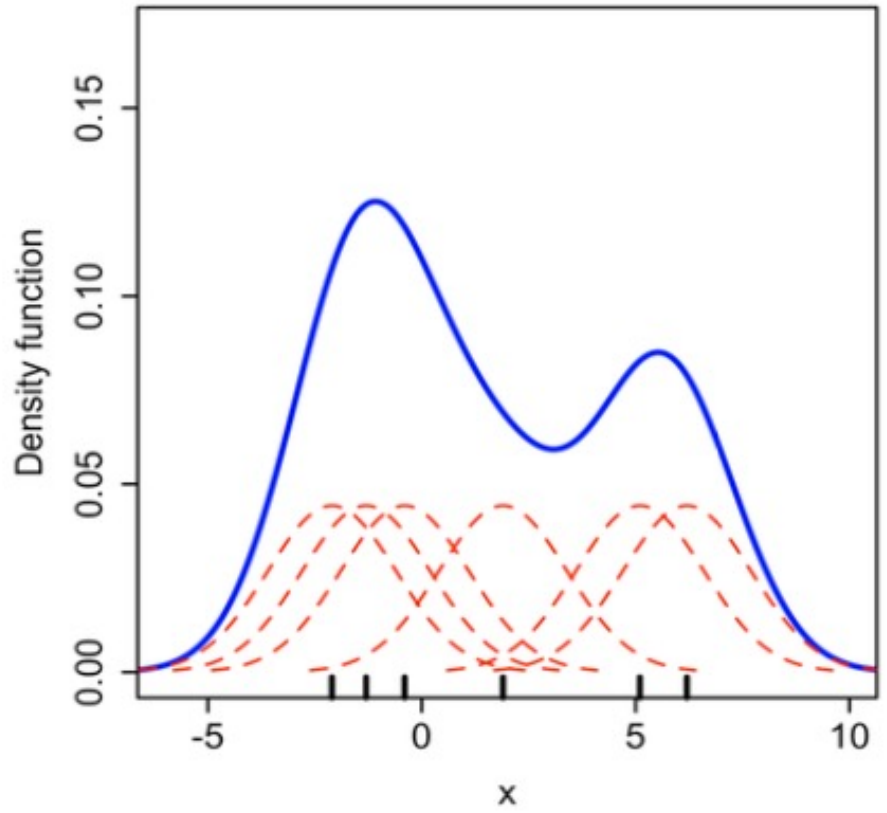
Spatial aggregation: Spatial boundaries

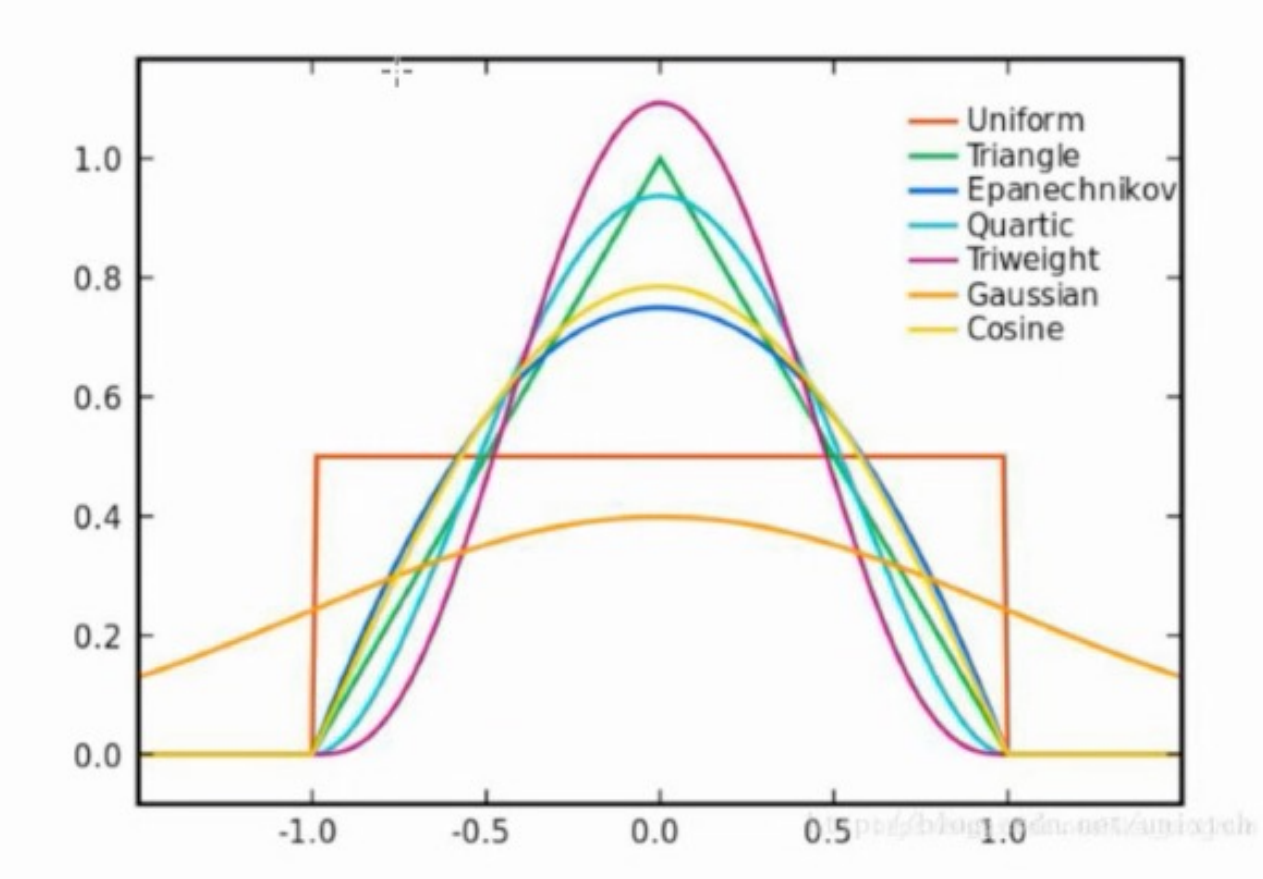
Mesh Blocks	- The "building blocks" of the census - 2016 Census contains 358,122 Mesh Blocks - Generally follow man-made or natural boundaries (ie. roads, rivers, etc)
SA1s	- Aggregates of Mesh Blocks - 2016 Census contains 57,523 SA1s - Generally contain population of 200 - 800 people
SA2s	- Aggregates 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

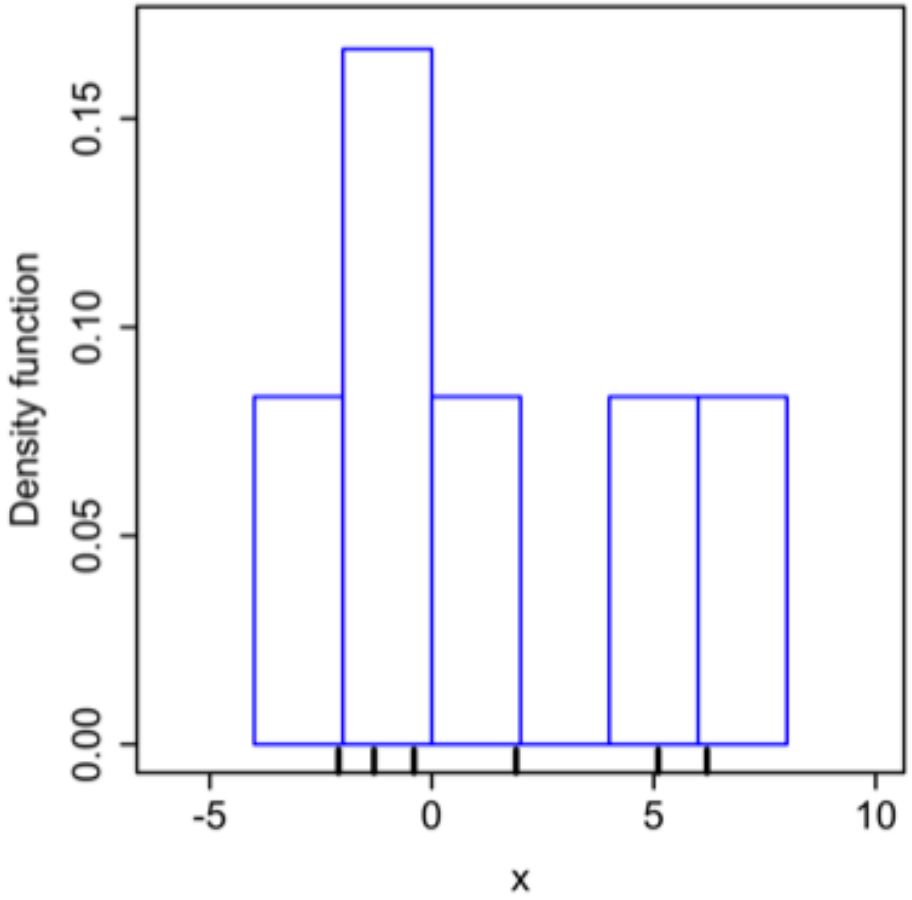
Kernel Density

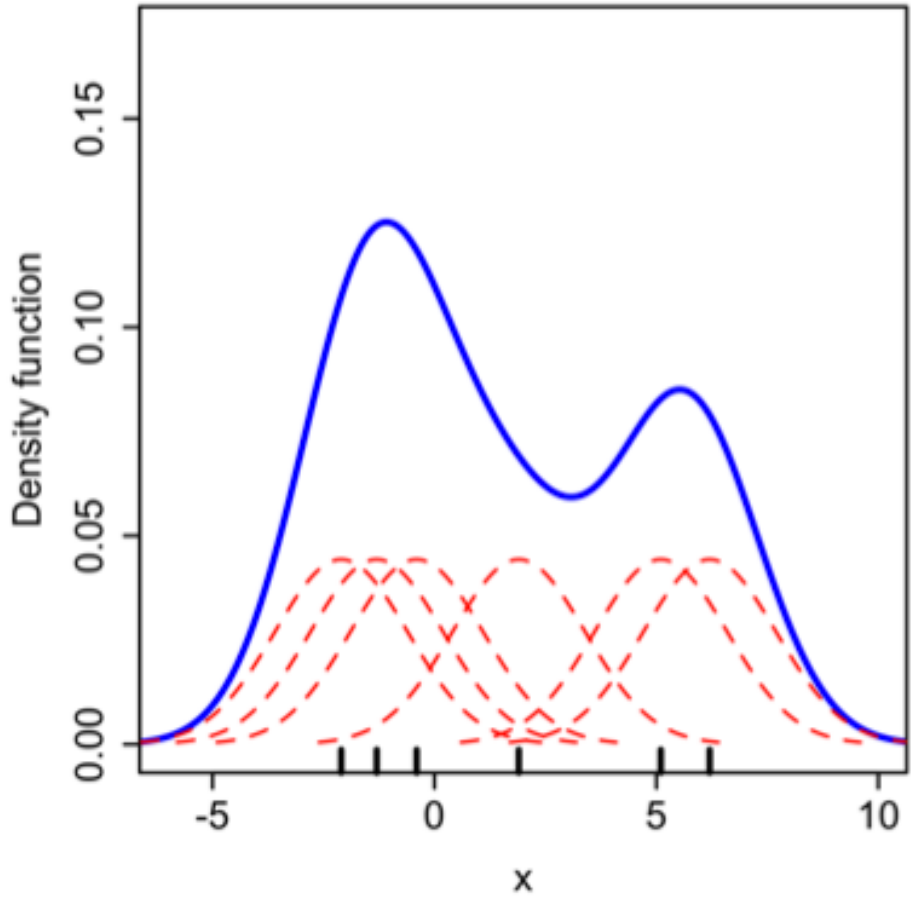




Images from: ABS

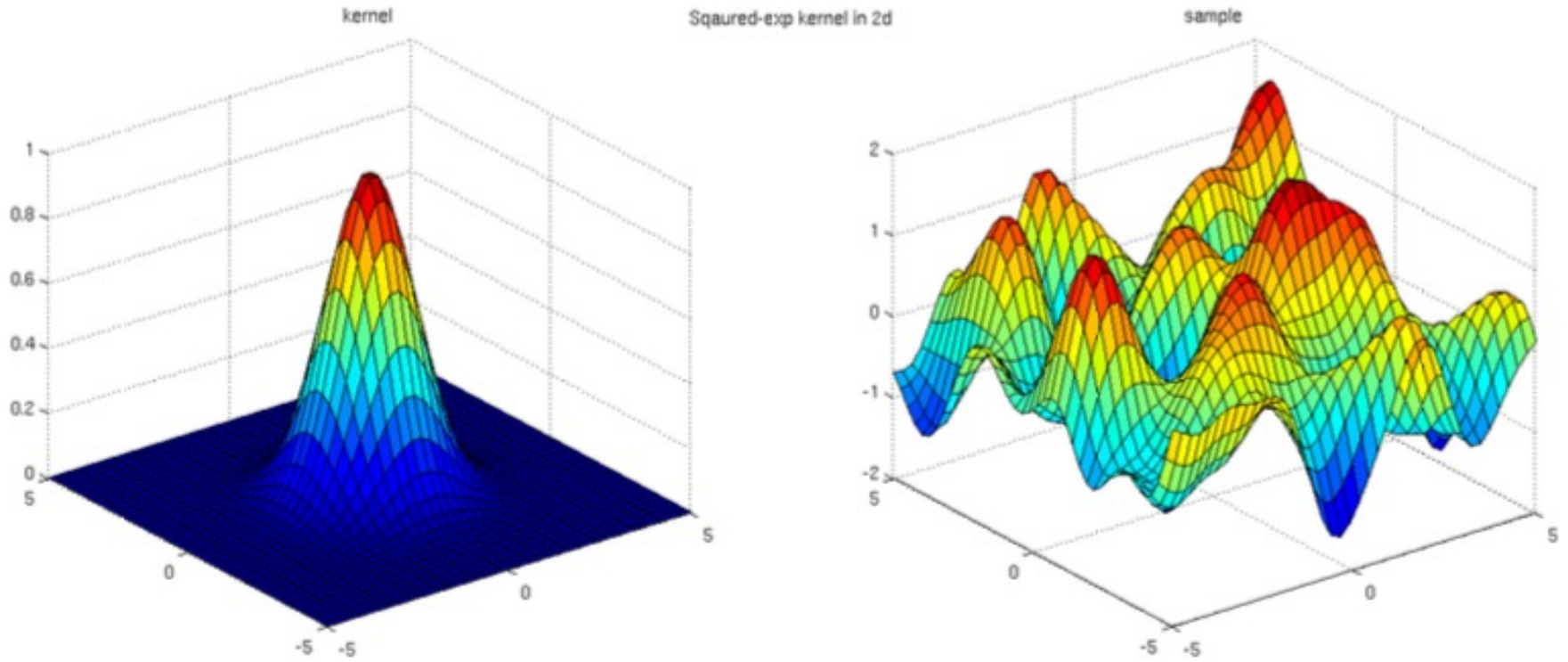
Kernel Density





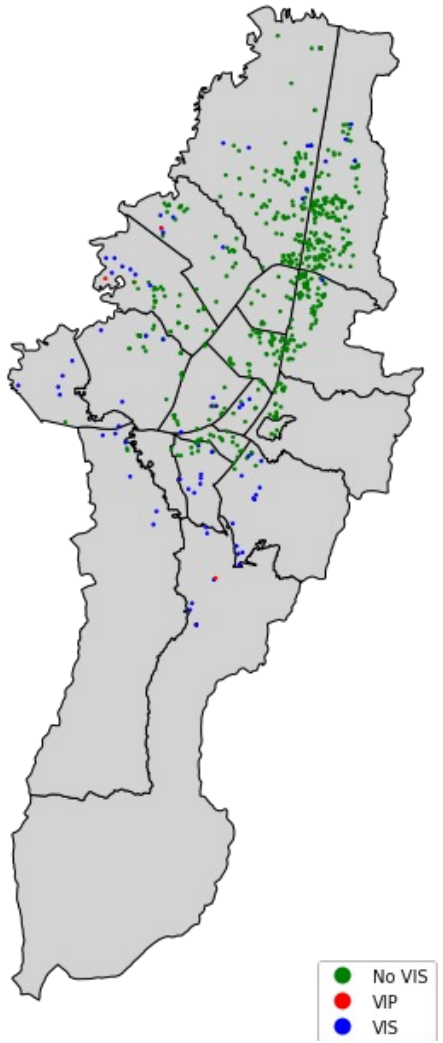
Images from: ABS

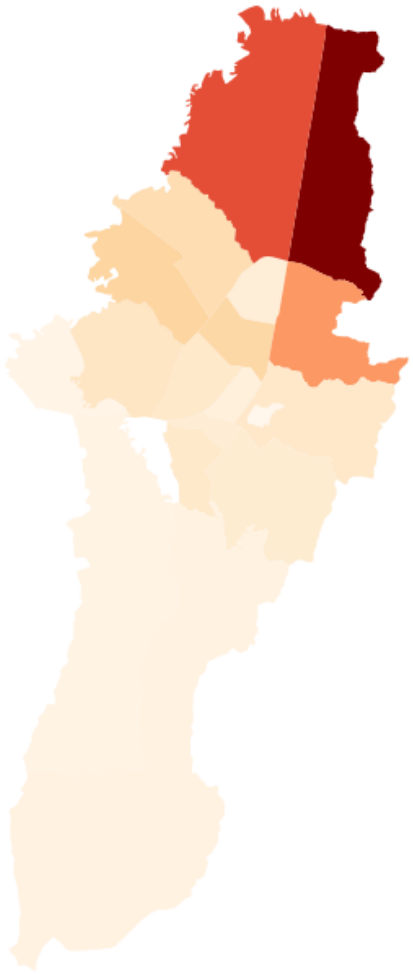
Kernel Density

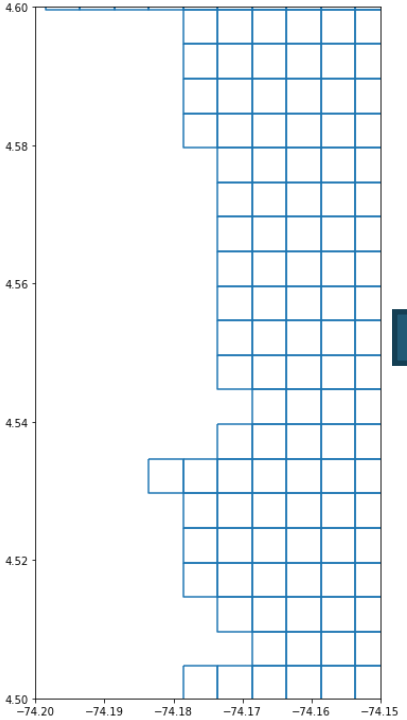


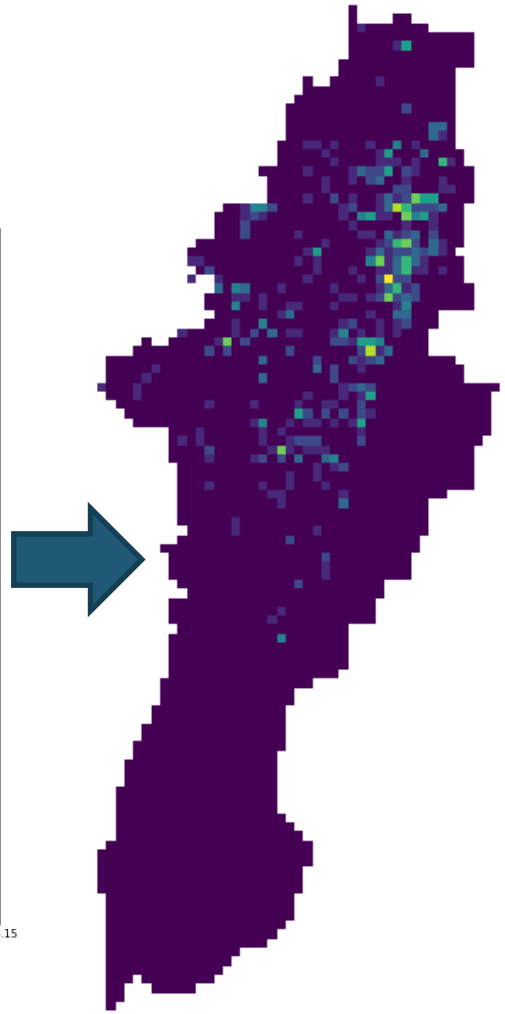
Images from: ABS

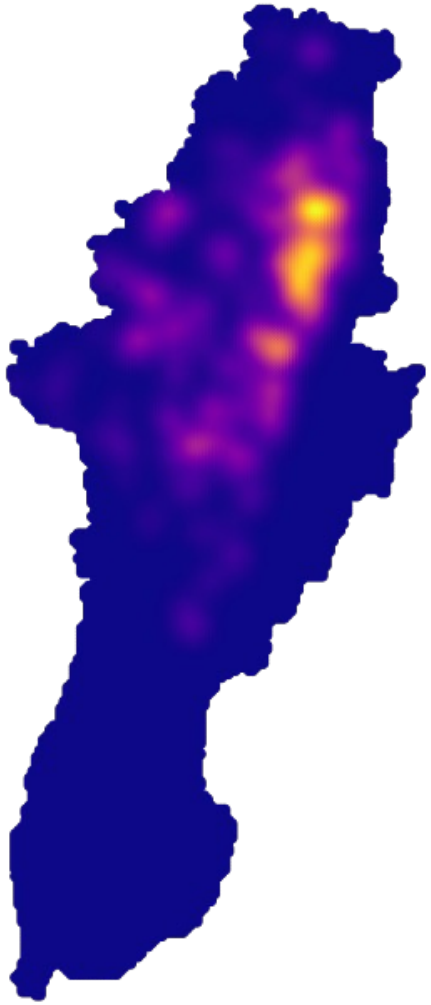
Kernel Density











By boundary

Grid

KDA

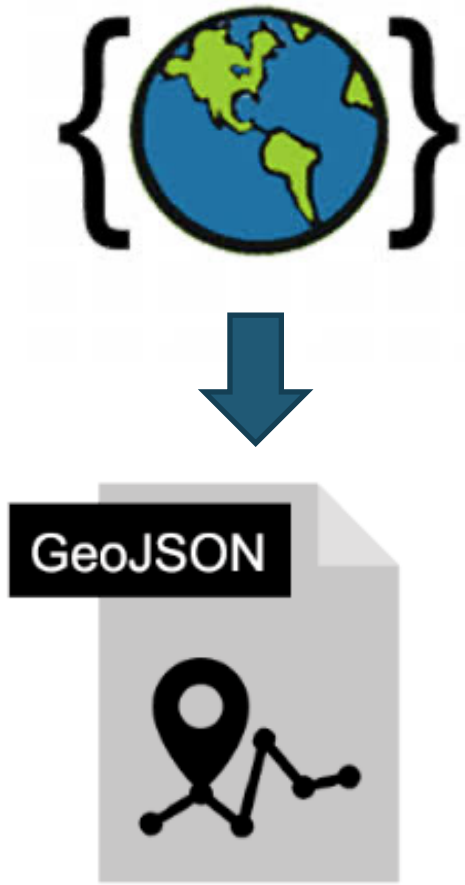
Images from: ABS

GeoJson

A format that is used to represent geographical features and their respective non-spatial attributes.

Characteristics:

- Format widely used in mapping applications.
- Fast, light and simple data exchange.
- GeoJSON uses a geographic coordinate reference system, WGS84, and units in decimal degrees.
- Maintained by a community of developers on the Internet.



GeoJson

```
{
   "type": "Feature",
   "geometry": {
      "type": "Point",
      "coordinates": [125.6, 10.1]
   },
   "properties": {
      "name": "Dinagat Islands"
   }
}
```

ESRI Shapefile (SHP)

The ESRI Shapefile (SHP) is a digital storage vector format that is used to save the location of geographic elements and the attributes associated with them. This format was developed by ESRI, a company that creates and markets software for Geographic Information Systems and was specifically designed for use with their ArcView GI product.

Images from: ESRI

This is a multi-file format, meaning it is created by multiple computer files.

Shapefile

- .shp - stores the geometric entities of the objects.
- .shx - stores the index of the geometric entities.
- .dbf - The database, stored in dBASE format, contains information about object attributes.

Other formats

File: Transform a data frame containing shapes to a complex object in a byte string

- Pickle
- RDS
- RDATA

Database: Datastore based on the use of a database management system

- Postgres
- MongoDB

Exercise – R

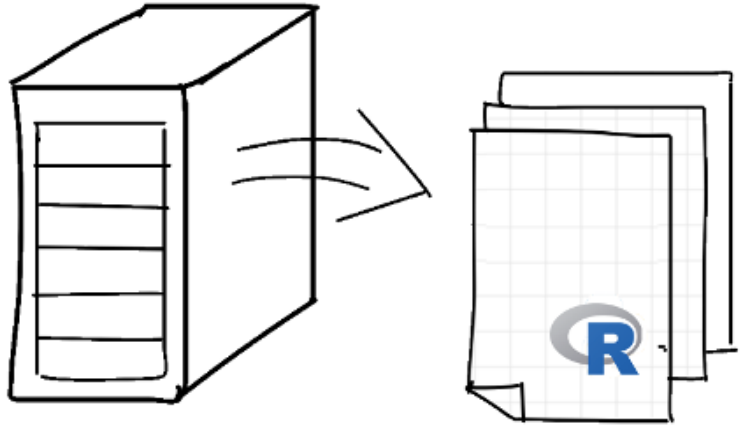
- **Part 1:** Spatial operations
  - Union
  - Intersects
  - Difference
  - Within
  - Buffer
  - Convex hull
  - Centroid
  - Boundaries
  - Area (Mts^2)

Exercise – R

Choices for the R environment

a) Cloud environment

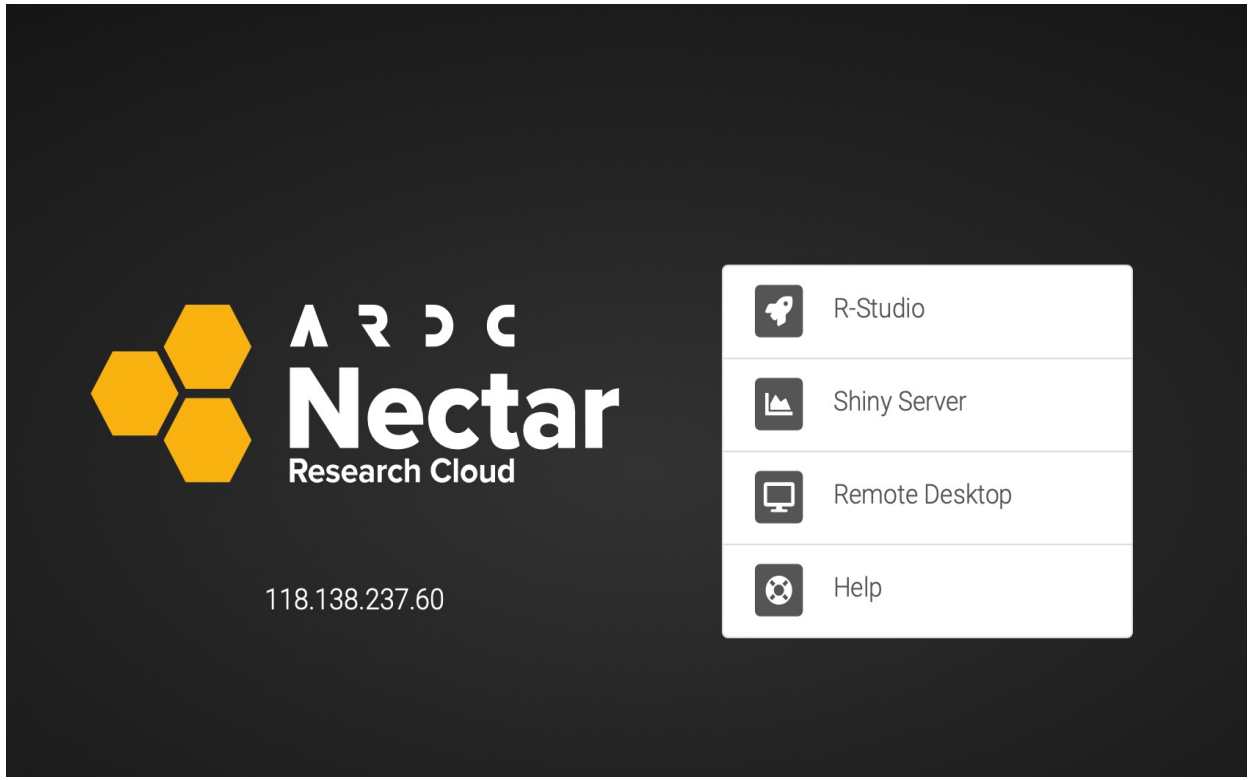
- Compatibility
- Less time spent installing R and dependencies.
- Better Package compatibility
- New users

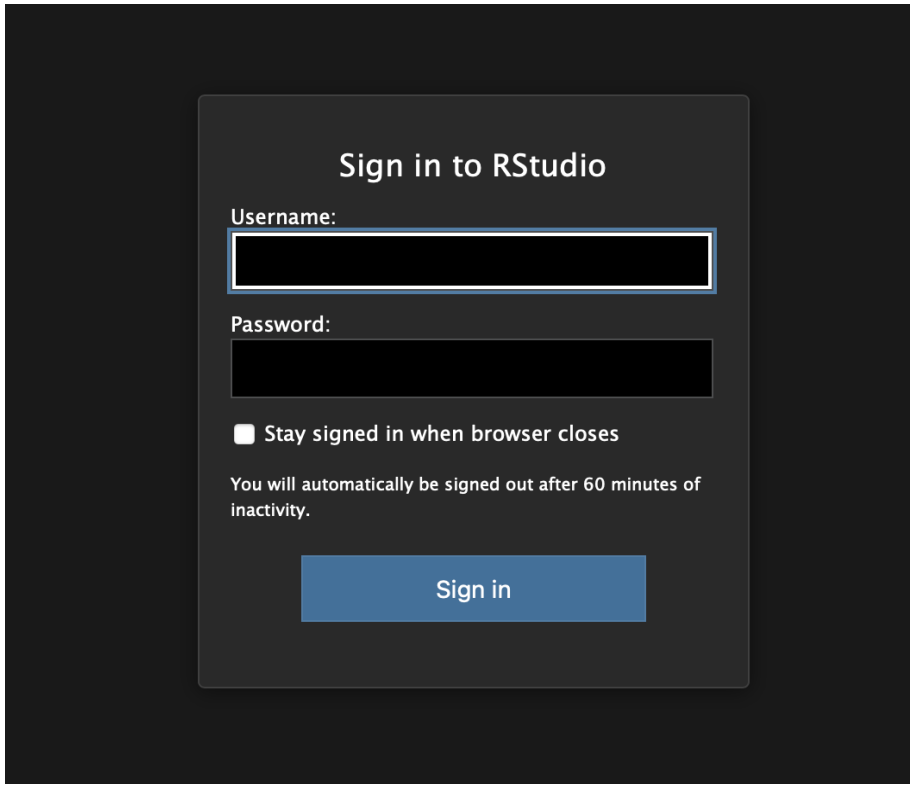


b) Local environment

- Customisable
- Privacy
- Large data sets
- Experienced users

Exercise – R


ARDC
Nectar
Research Cloud
118.138.237.60
R-Studio
Shiny Server
Remote Desktop
Help


Sign in to RStudio
Username:
Password:
Stay signed in when browser closes
You will automatically be signed out after 60 minutes of inactivity.
Sign in

Cloud environment

Group 1:
Host: http://118.138.239.12
User: user1….user15
Password: hass2024

Group 2:
Host: http://118.138.237.228
User: user1….user15
Password: hass2024

Group 3:
Host: http://118.138.234.153/
User: user1….user15
Password: hass2024

Local environment

Step 1: Clone the repository on GitHub

https://github.com/AURIN-OFFICE/HASS2024

Step 2: Install the necessary libraries

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

Introduction to geospatial data

Finding and using spatial data

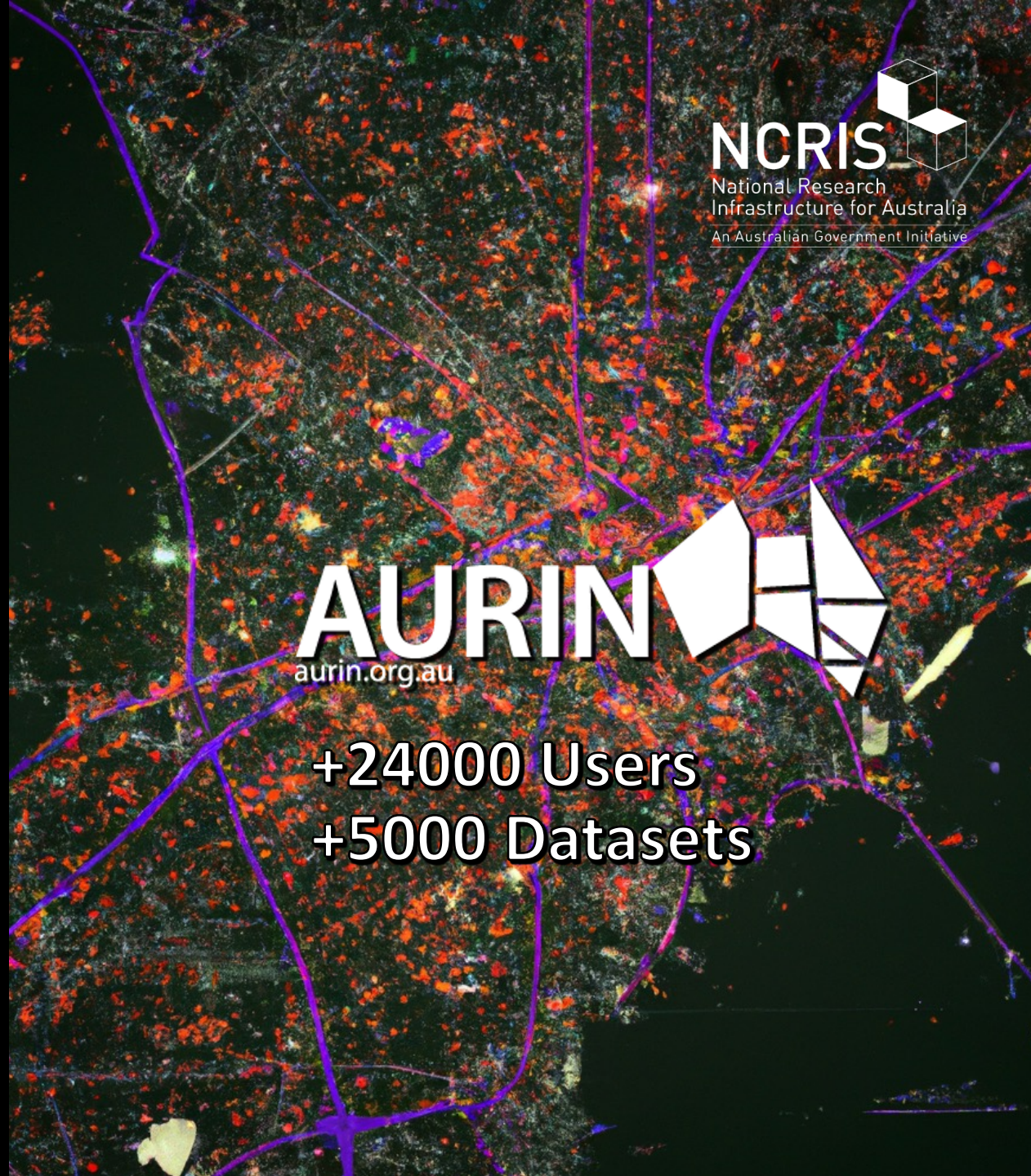
Deciding on integration

Producing a new data product

The Geosocial work package

AURIN in a Glance

- Established in 2010
- NCRIS funded
- National digital research infrastructure
- Critical data and analytics
- AURIN capabilities
- Collaboration and partnership
- AURIN transformation

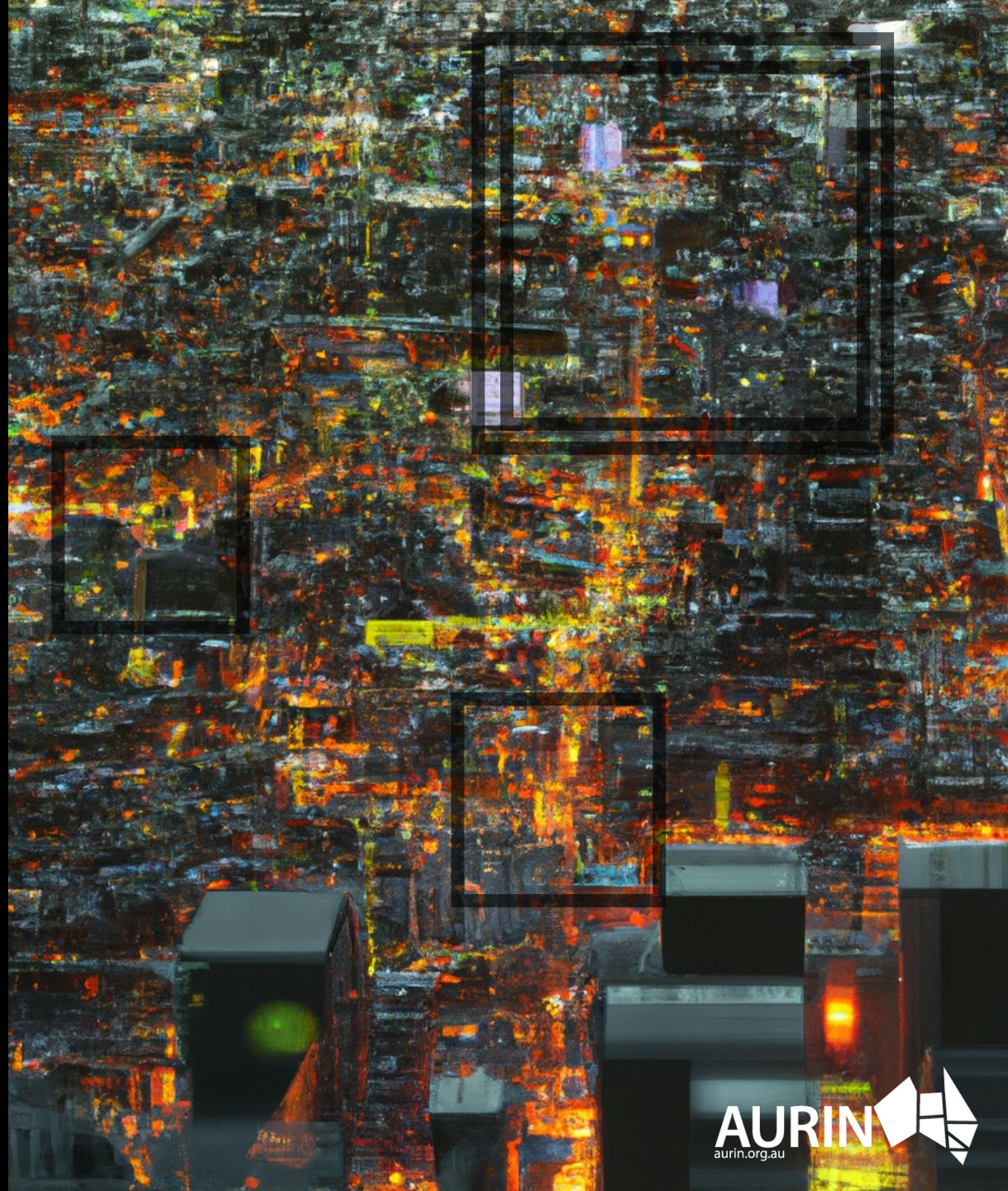


Addressing key urban challenges in Australia.

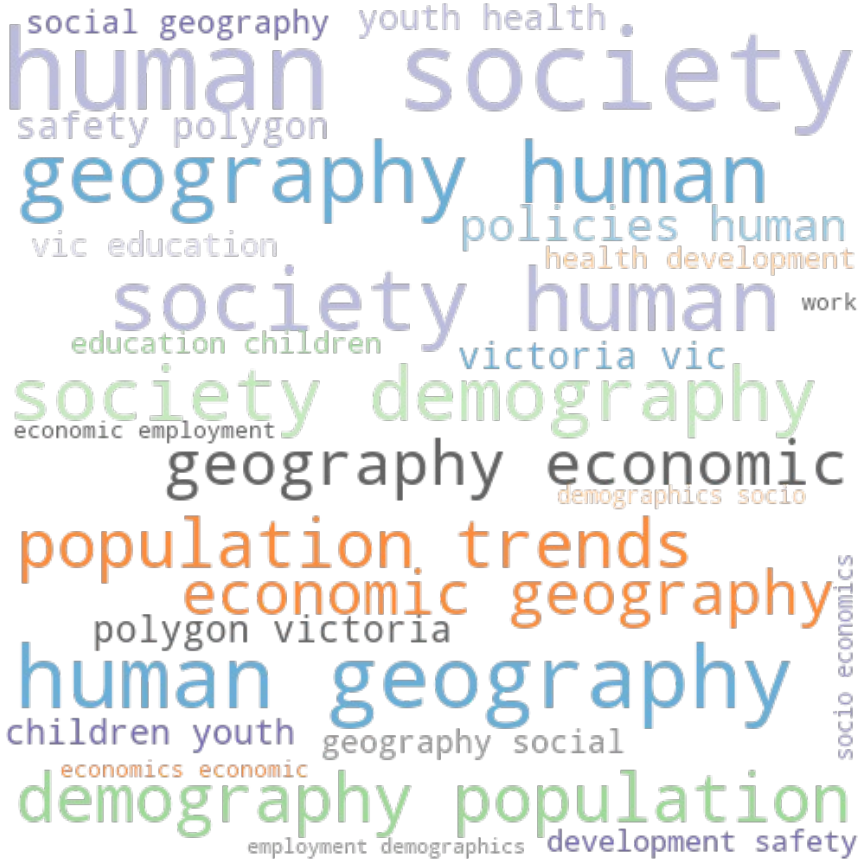
- Australian key urban challenges
  - Climate change
  - Energy transition
  - Demographic transformation

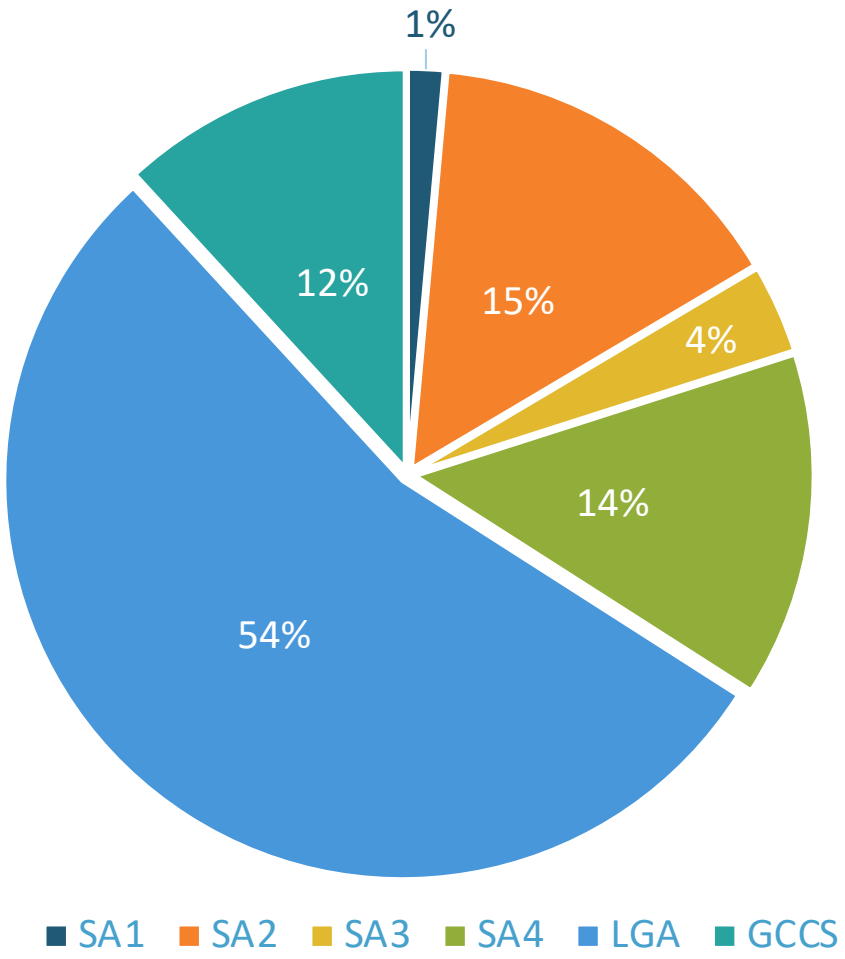
- Urban digital twins as enablers
  - Digital representations
  - Analytics and modelling capabilities
  - More informed decision-making tools

- Does this make a city smart?



AURIN DATA

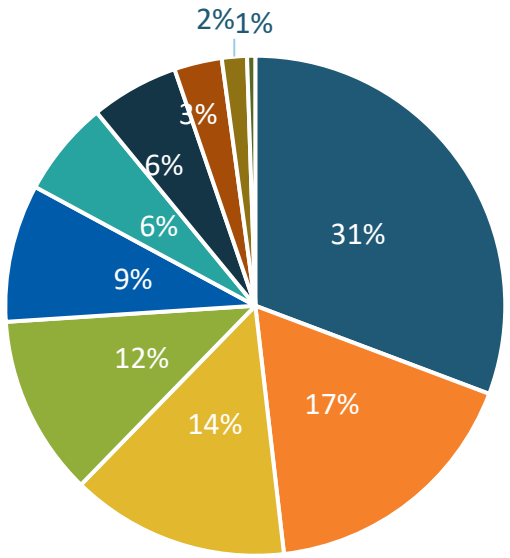




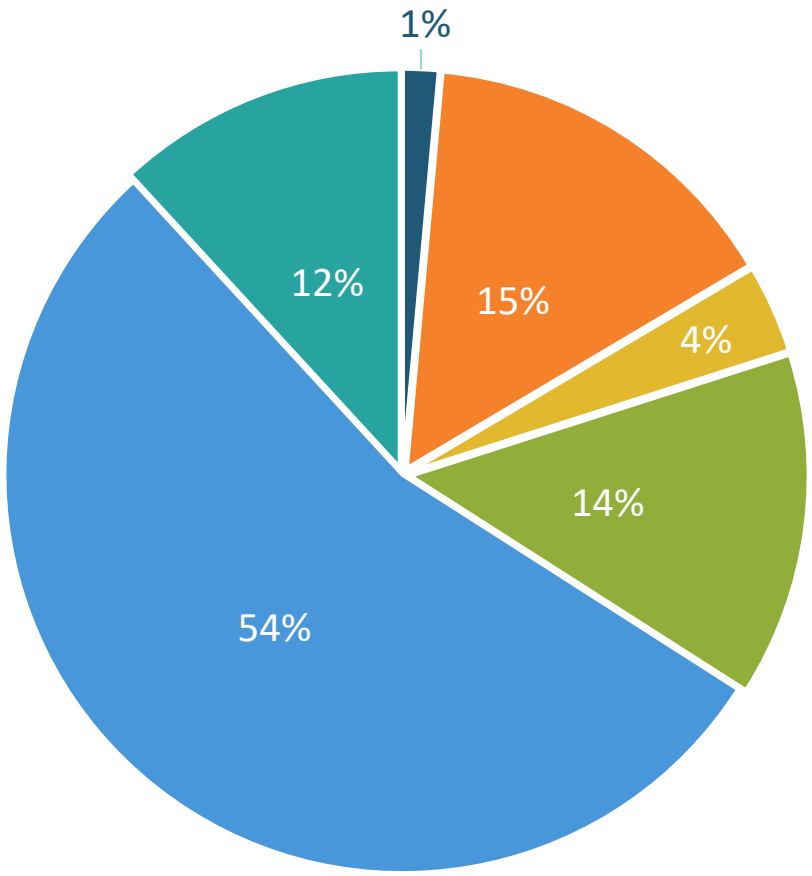
Datasets: 5,290

AURIN DATA

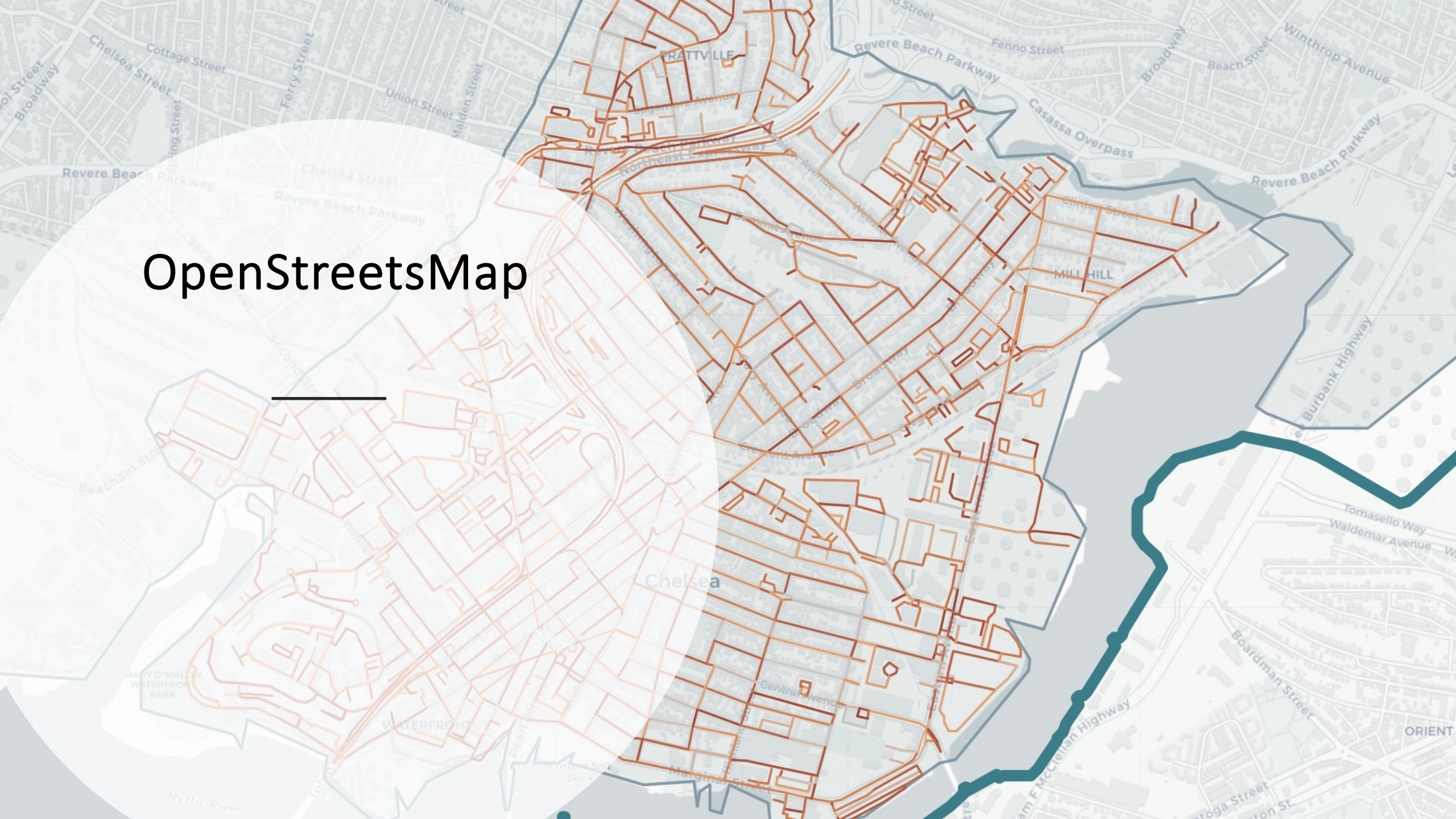
Classification by theme (NPL)



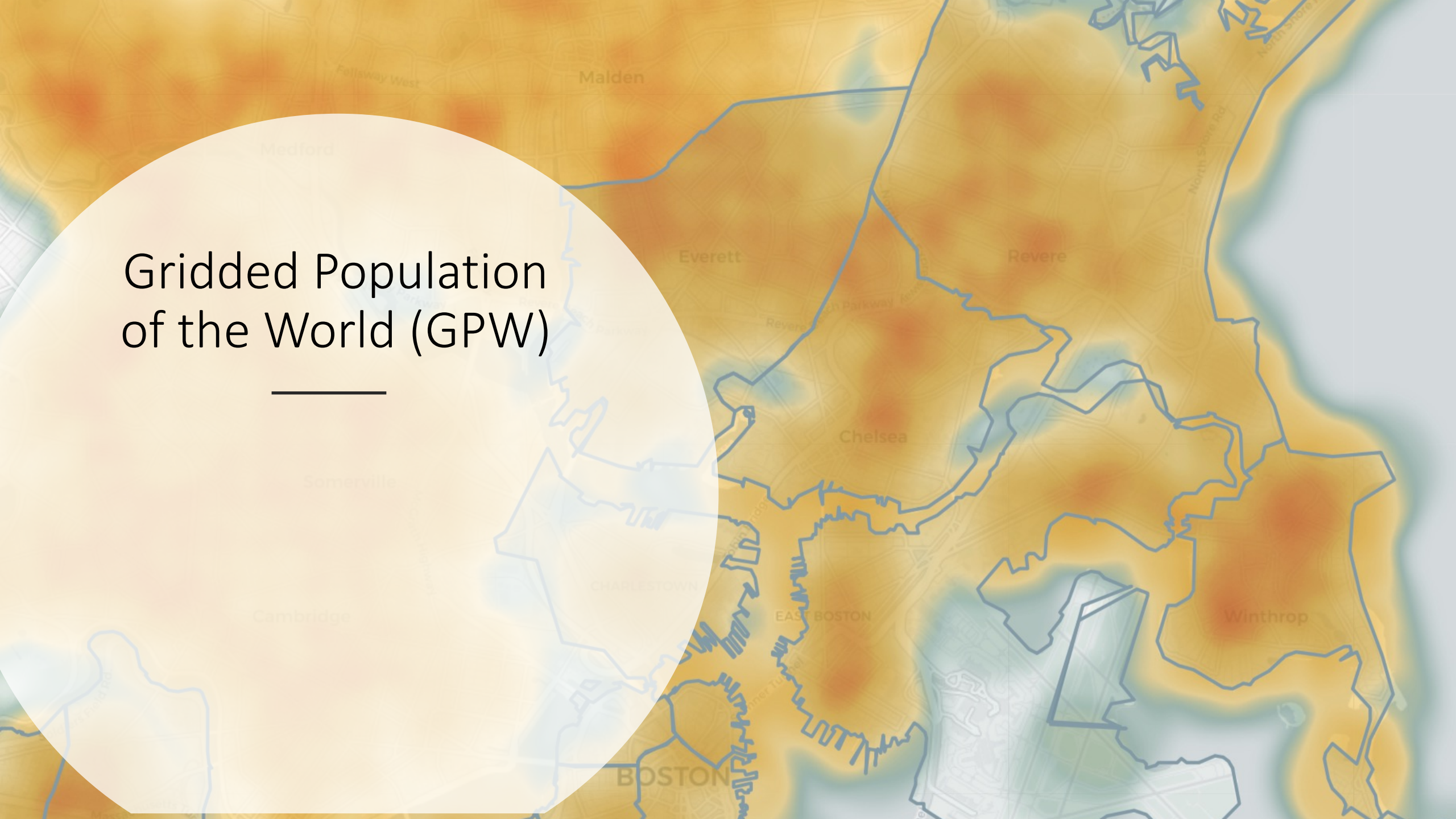
Transcribe
Transport
Health
Housing
Demographics
Social policy
Economy
Climate
Planning
Land use
Infrastructure



SA1
SA2
SA3
SA4
LGA
GCCS

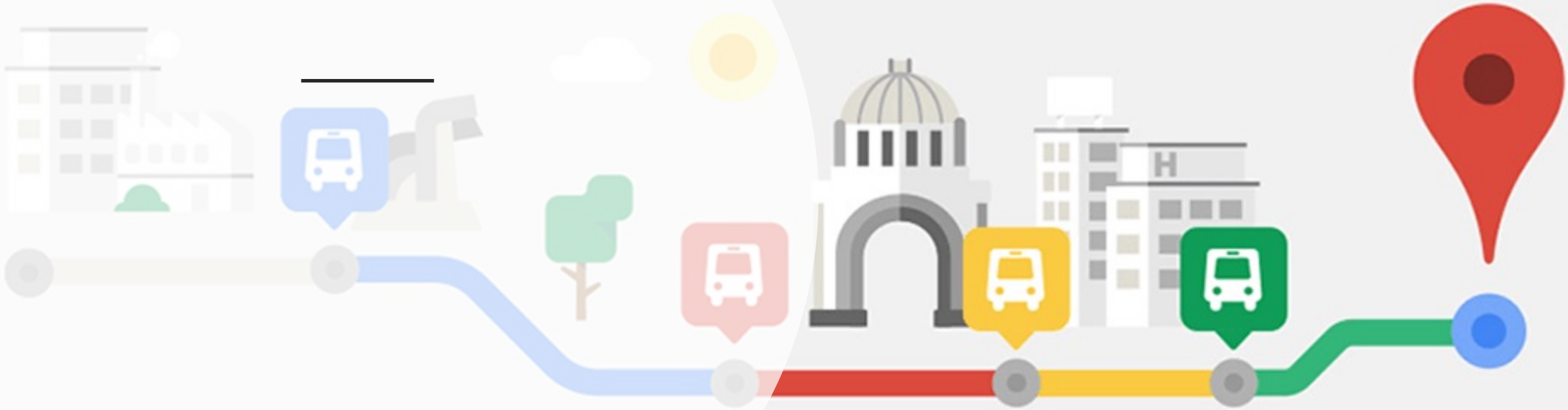

OpenStreetsMap
Revere Beach Parkway
Chelsea Street
Cottage Street
Ferry Street
Union Street
Malden Street
Broadway
Fenno Street
Winthrop Avenue
Beach Street
Casassa Overpass
Northeast Expressway
Broadway
Central Avenue
Burbank Highway
Tomasello Way
Waldemar Avenue
Boardman Street
Gen Edward F McClellan Highway
Beacham Street
Second Street
Mystic River
WATERFRONT
MARY O'MALLEY WATERFRONT PARK
ORIENT

Gridded Population of the World (GPW)



Google™ | Transit

GTFS



Introduction to geospatial data

Finding and using spatial data

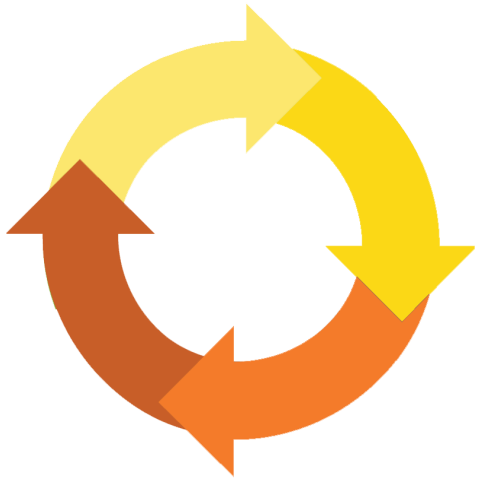
Deciding on integration

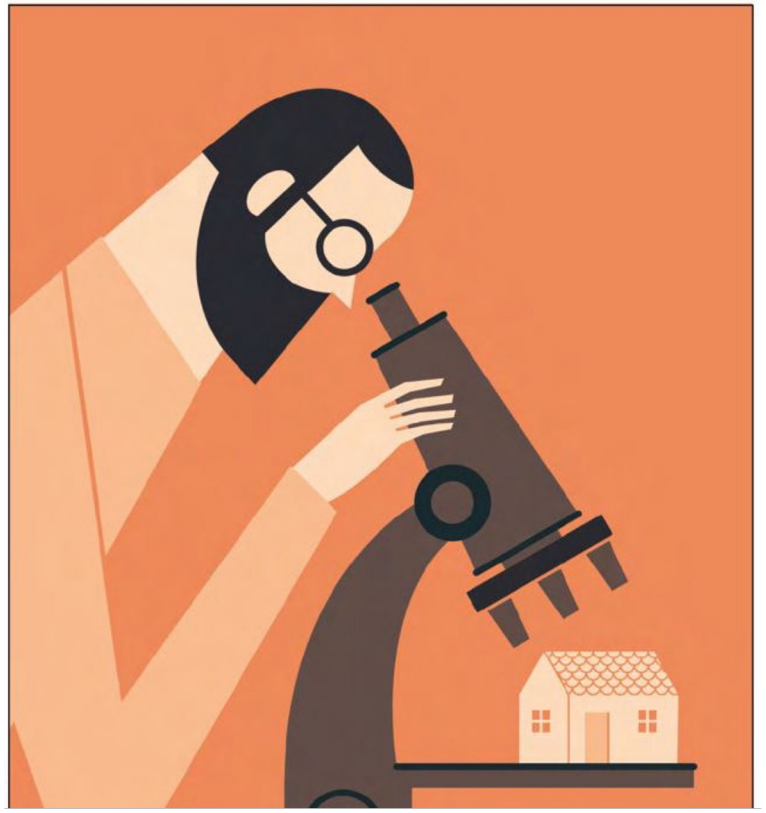
Producing a new data product

The Geosocial work package

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.

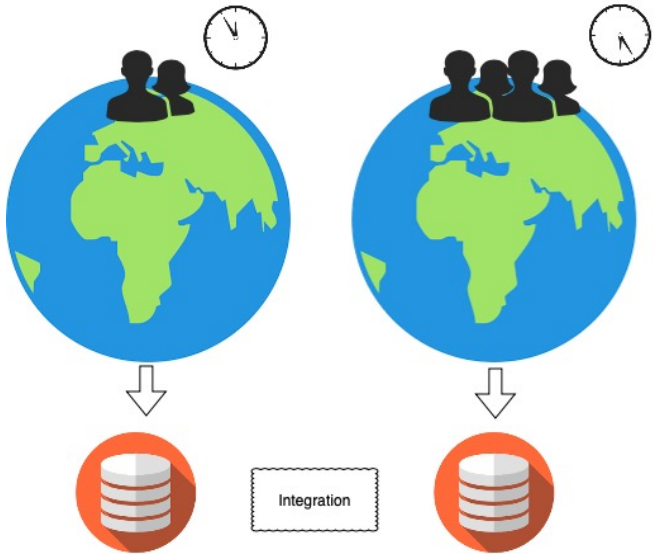
Database: Socio-Economic Indexes for Areas (SEIFA) – SA2 - 2016

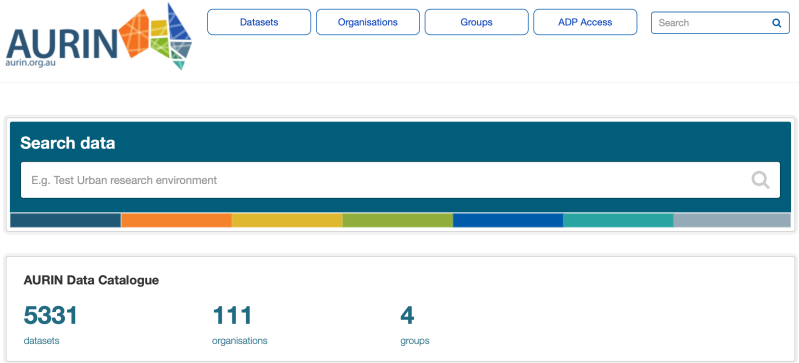
Variable:

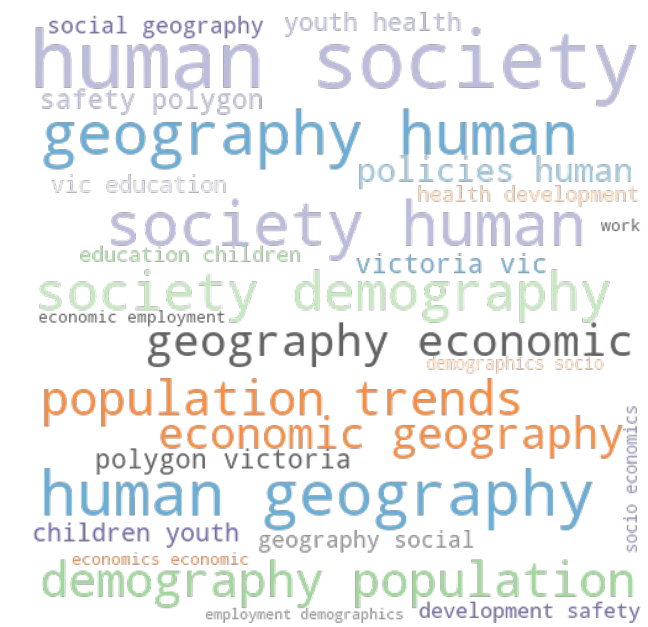
Step 3: Use a spatial correlation (Moran's I) to evaluate the correlation between leaving the school on economic development (SEIFA).

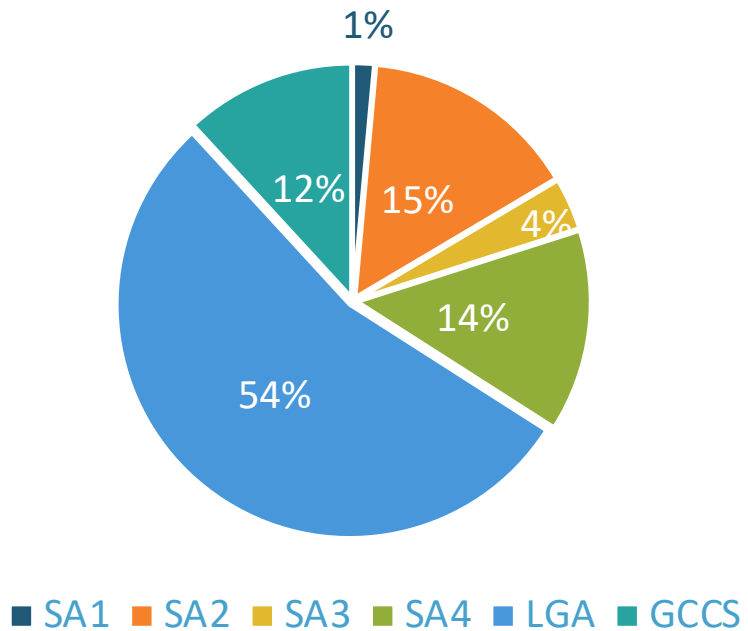
Step 4: Try other ways to estimate spatial correlation: Geary's C and Getis-Ord's index

Deciding on integration:









Exercise – R

- **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

Introduction to geospatial data

Finding and using spatial data

Deciding on integration

Producing a new data product

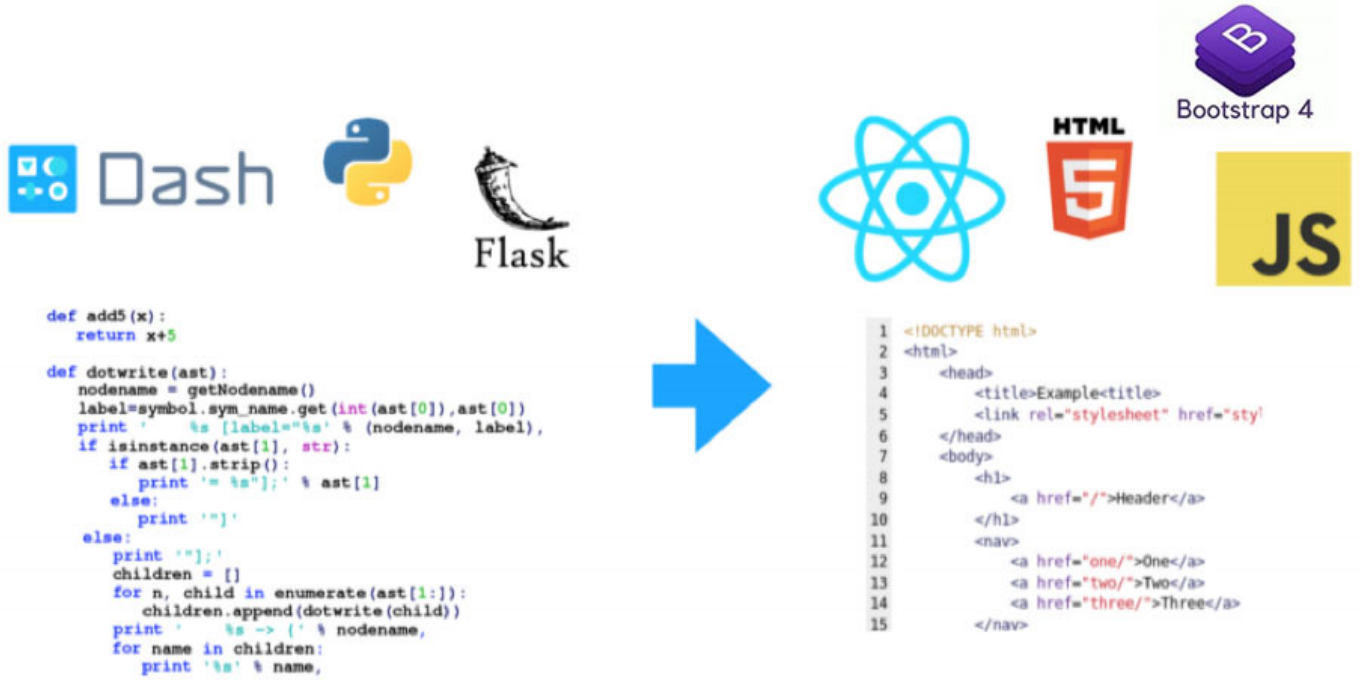
The Geosocial work package

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

Producing a new data product

CSS

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

Producing a new data product

AURIN
aurin.org.au

Bootstrap 4

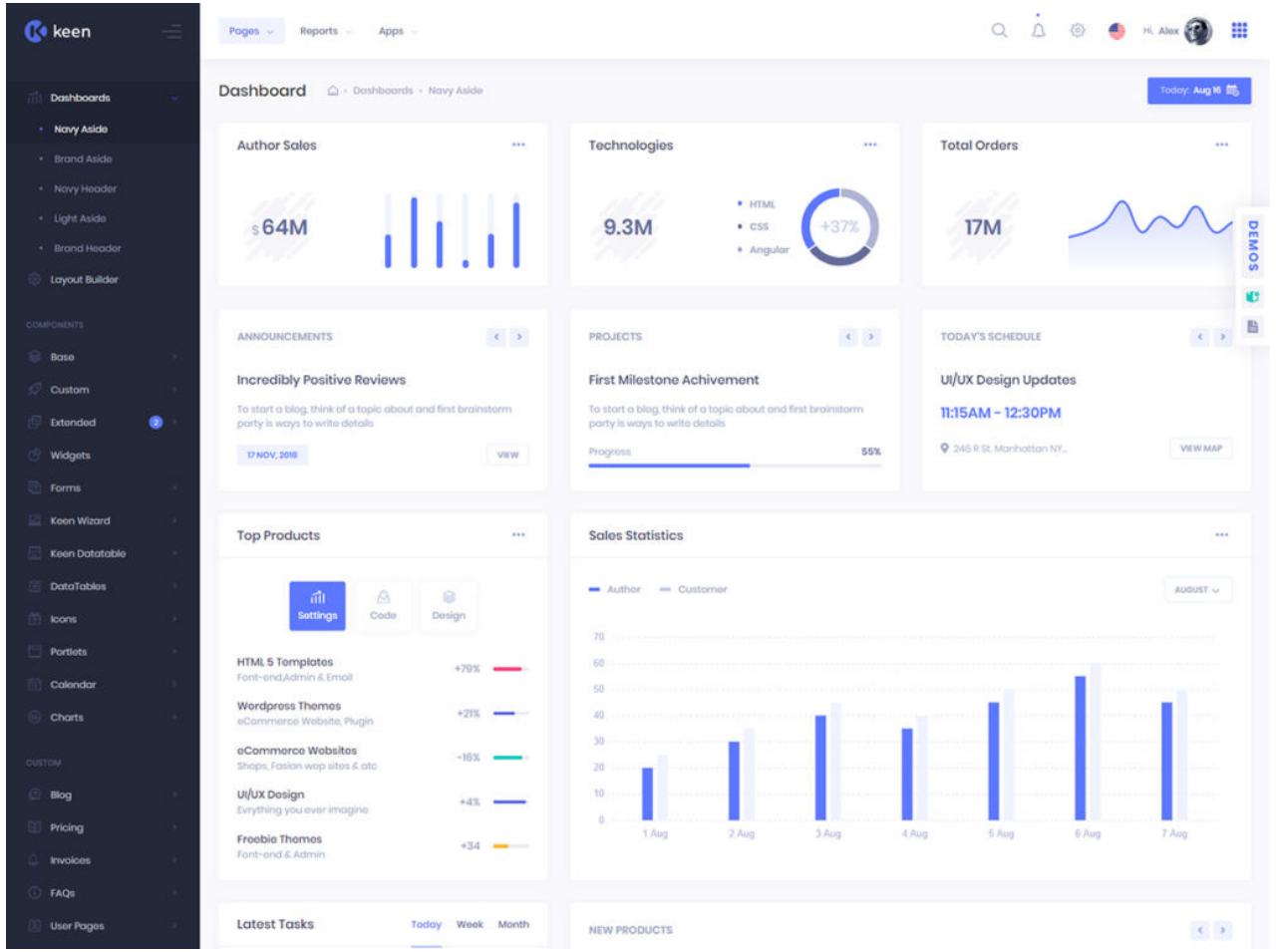


Image from: “Product Data and Your Digital Catalog” by Handshake

Producing a new data product

Javascript

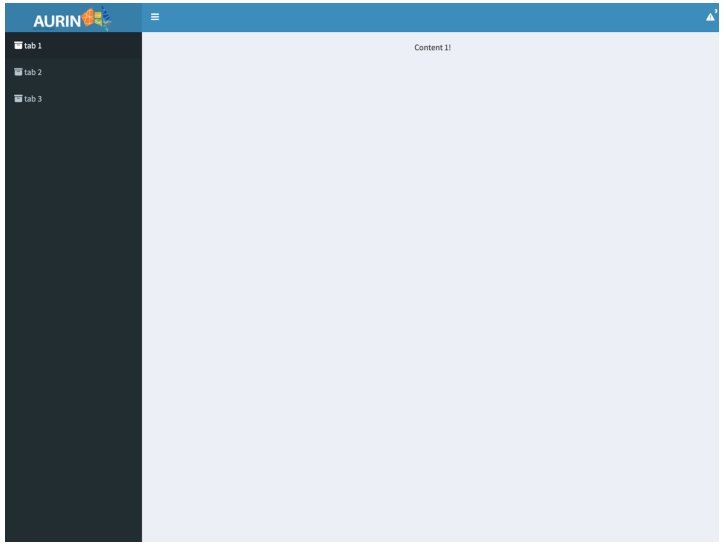
```
class Car {
    constructor(doors, engine, color) {
        this.doors = doors;
        this.engine = engine;
        this.color = color;
    }

    carStats() {
        return `This car has ${this.doors} do
```

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

Producing a new data product

Frontend: Dash



Outputs:

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

Introduction to geospatial data

Finding and using spatial data

Deciding on integration

Producing a new data product

The Geosocial work package

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.

IRISS project

The Integrated Research Infrastructure for the Social Sciences (IRISS) project addresses the fragmentation of the Australian social science research infrastructure, establishing a new foundation for integrating data, analysis and platforms for social science research in Australia.

GeoSocial

Longitudinal survey

Geospatial data

The GeoSocial solution allows researchers to link Australia's largest longitudinal surveys with geospatial statistical data derived from the Australian Census of Population and Housing. GeoSocial will empower Australia's large cross-disciplinary social research community to identify patterns, make predictions, and inform social policy using rich integrated GeoSocial data.

How data linkage works?

GeoSocial utilizes the geographical identifier from the longitudinal survey and converts it to a Statistical Areas Level 3 (SA3s) for linking with geospatial statistical data obtained from the Australian Census of Population and Housing. The Geosocial output retains the original format of the longitudinal survey, with the addition of geospatial variables as a new column. It is the responsibility of the user to:

- Request access to the Longitudinal Surveys of Australian Youth datasets.
- Set up a safe environment according to the data custodians' policies.
- Install R and required dependencies

The GeoSocial solution is composed of the following elements:

- **Toolbox:** R library that has all the R functions you need for data linkage.
- **Parameters:** File with all the relevant information for data linkage, including data locations, API credentials, wave and cohort information.
- **Script:** Used to execute the workflow which will use the toolbox to read and merge the data based on user preferences.

GeoSocial does not collect or retain any personally identifying information.

GeoSocial

START

Thank you