

# Spatial data and data integration

Australian Urban Research Infrastructure Network (AURIN)



# 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**

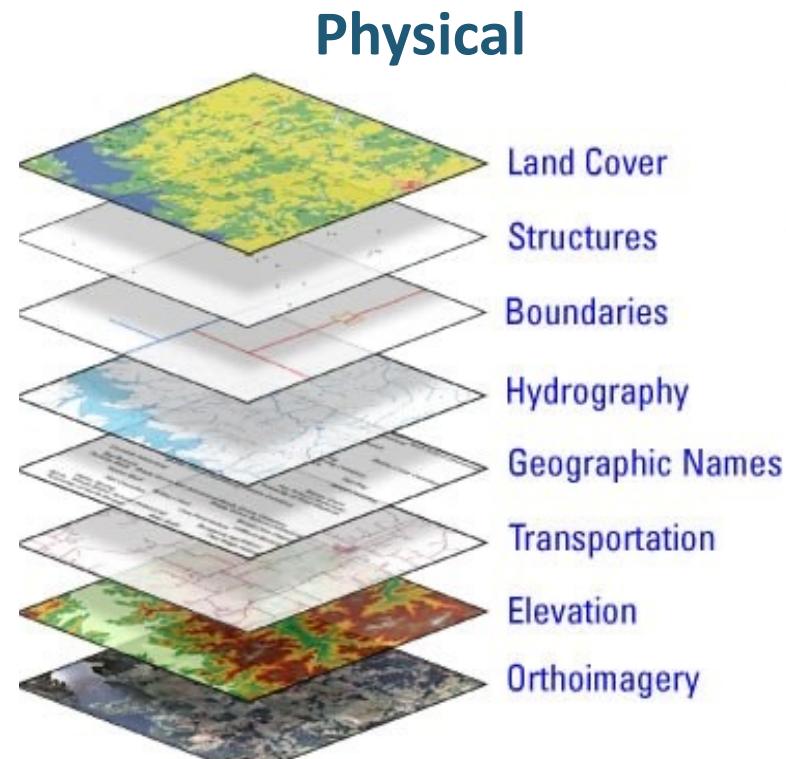
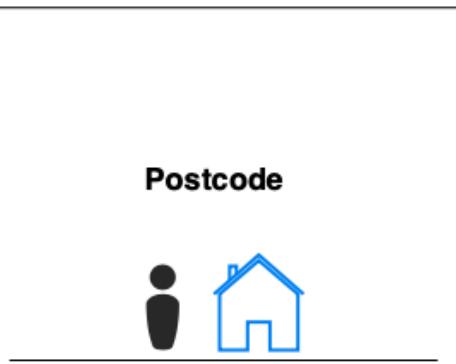
**Deciding on integration**

**Producing a new data product**

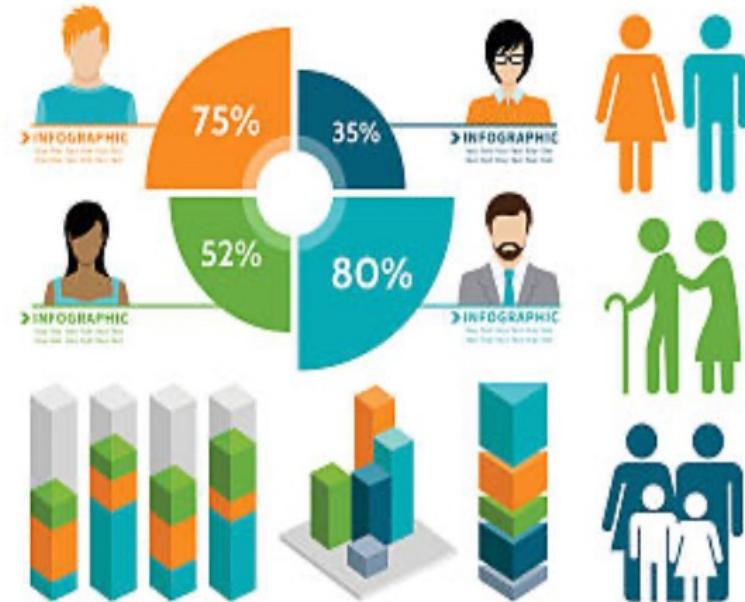
**The Geosocial work package**

# What is spatial data?

LGA



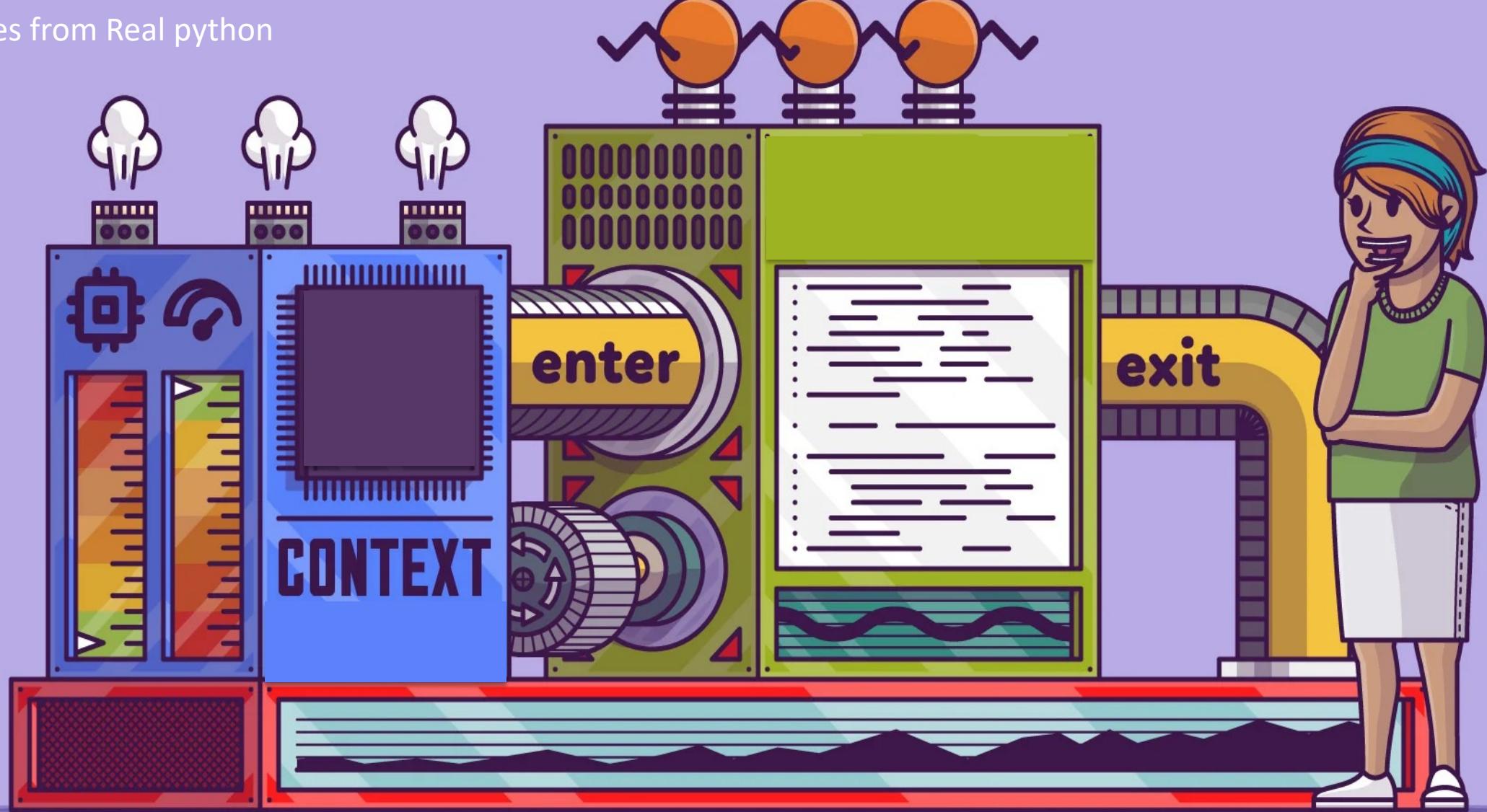
Socio-demographic



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

# Data enrichment

Images from Real python



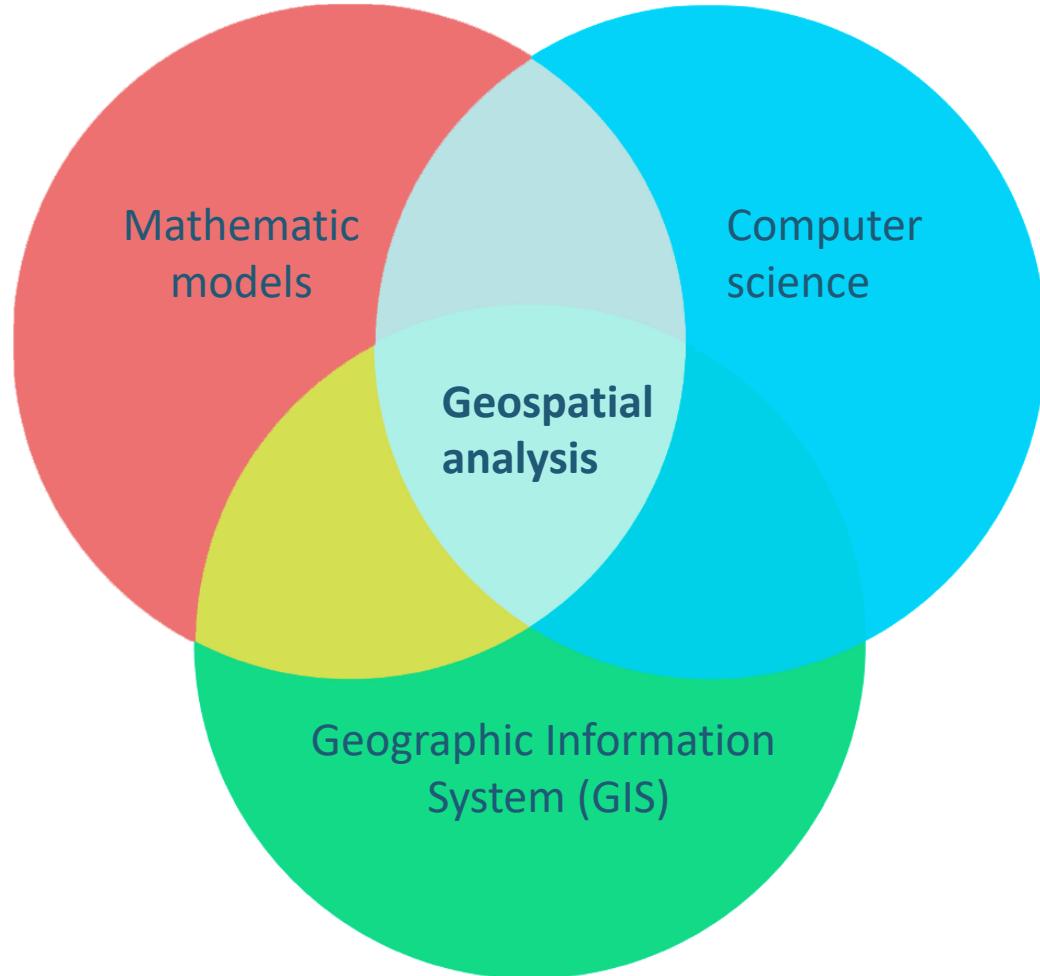
# Geospatial analysis

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

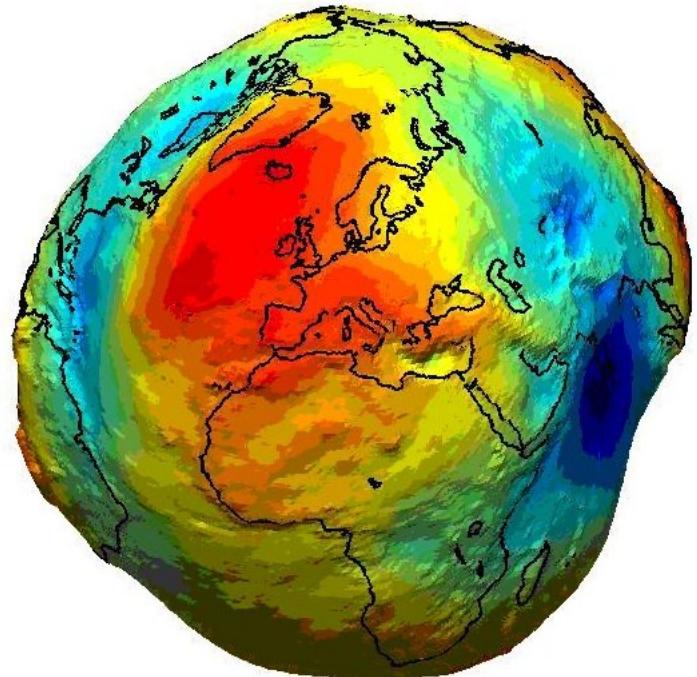
Spatial analysis needs to make assumptions about or draw on data describing spatial relationships or spatial interactions between cases. (Chorley, 1972; Haining 1994).



# Geospatial Analysis



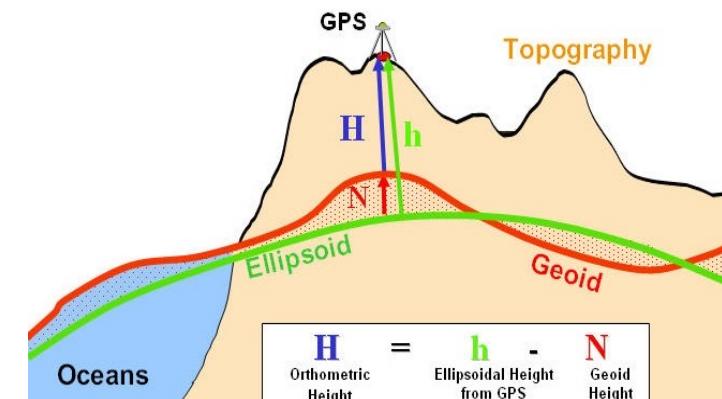
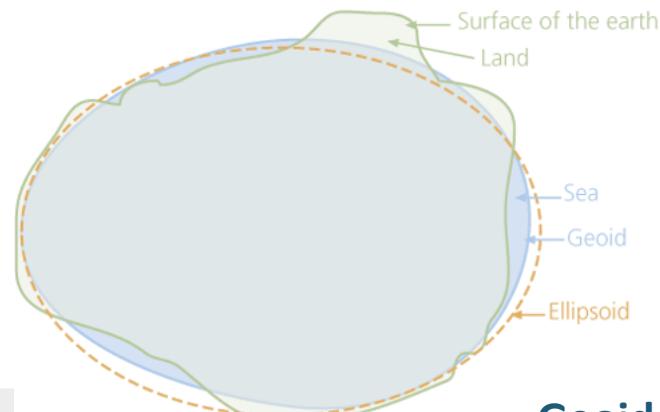
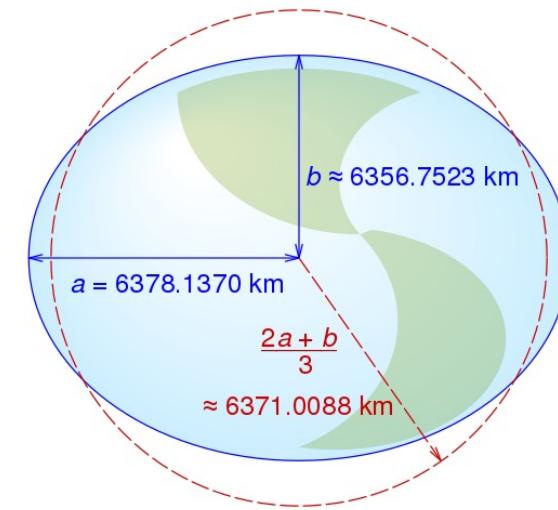
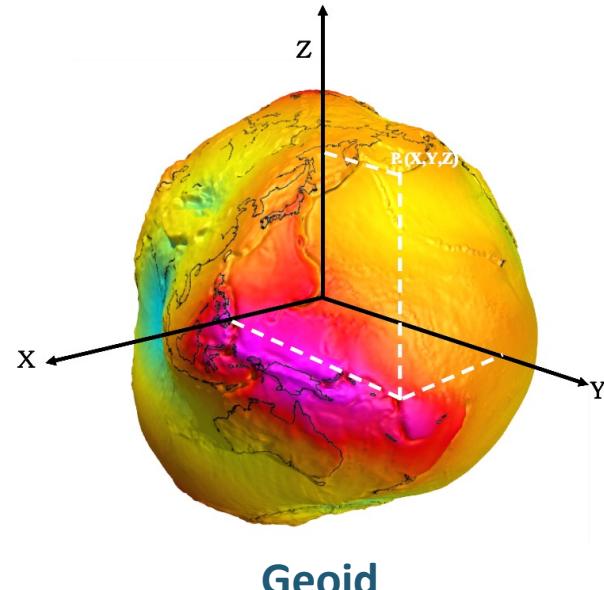
# Geoid



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 vs Ellipsoid



## Spatial Reference List

[Home](#) | [List all references](#)

| [Next Page](#)

Search References:

You are only searching **EPSG** references. [Search All?](#)

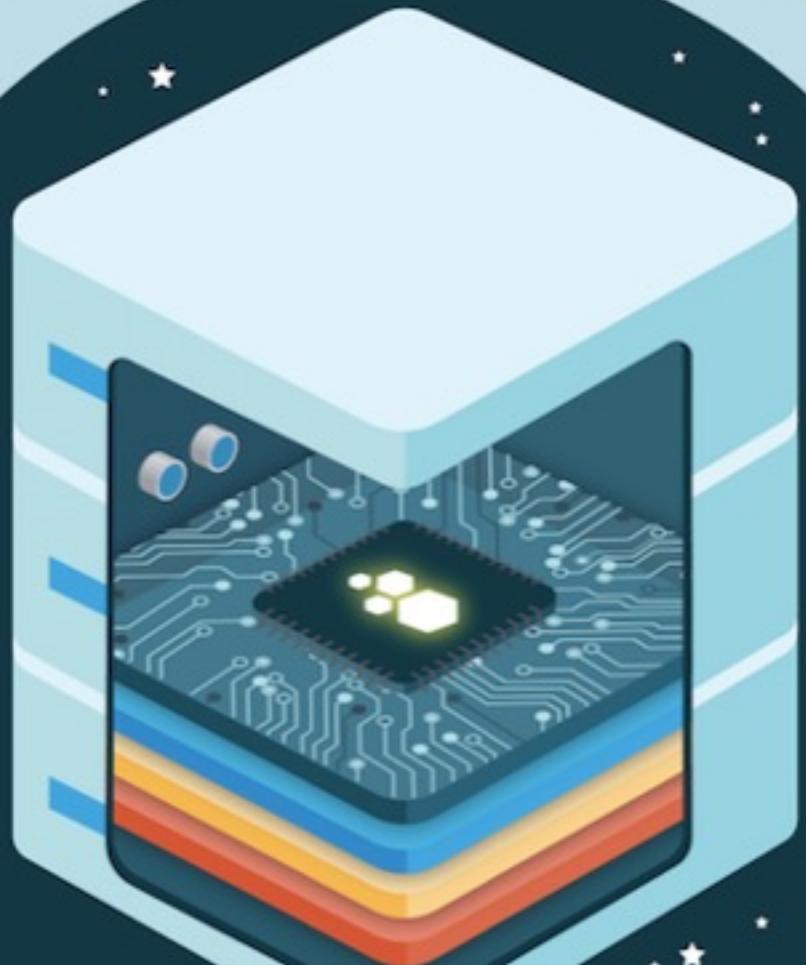
Entries found: 7452

- [EPSG:2000](#): Anguilla 1957 / British West Indies Grid
- [EPSG:2001](#): Antigua 1943 / British West Indies Grid
- [EPSG:2002](#): Dominica 1945 / British West Indies Grid
- [EPSG:2003](#): Grenada 1953 / British West Indies Grid
- [EPSG:2004](#): Montserrat 1958 / British West Indies Grid
- [EPSG:2005](#): St. Kitts 1955 / British West Indies Grid
- [EPSG:2006](#): St. Lucia 1955 / British West Indies Grid
- [EPSG:2007](#): St. Vincent 45 / British West Indies Grid
- [EPSG:2008](#): NAD27(CGQ77) / SCoPQ zone 2 (deprecated)
- [EPSG:2009](#): NAD27(CGQ77) / SCoPQ zone 3
- [EPSG:2010](#): NAD27(CGQ77) / SCoPQ zone 4
- [EPSG:2011](#): NAD27(CGQ77) / SCoPQ zone 5
- [EPSG:2012](#): NAD27(CGQ77) / SCoPQ zone 6
- [EPSG:2013](#): NAD27(CGQ77) / SCoPQ zone 7
- [EPSG:2014](#): NAD27(CGQ77) / SCoPQ zone 8
- [EPSG:2015](#): NAD27(CGQ77) / SCoPQ zone 9
- [EPSG:2016](#): NAD27(CGQ77) / SCoPQ zone 10
- [EPSG:2017](#): NAD27(76) / MTM zone 8
- [EPSG:2018](#): NAD27(76) / MTM zone 9
- [EPSG:2019](#): NAD27(76) / MTM zone 10
- [EPSG:2020](#): NAD27(76) / MTM zone 11
- [EPSG:2021](#): NAD27(76) / MTM zone 12
- [EPSG:2022](#): NAD27(76) / MTM zone 13
- [EPSG:2023](#): NAD27(76) / MTM zone 14
- [EPSG:2024](#): NAD27(76) / MTM zone 15
- [EPSG:2025](#): NAD27(76) / MTM zone 16
- [EPSG:2026](#): NAD27(76) / MTM zone 17
- [EPSG:2027](#): NAD27(76) / UTM zone 15N
- [EPSG:2028](#): NAD27(76) / UTM zone 16N
- [EPSG:2029](#): NAD27(76) / UTM zone 17N
- [EPSG:2030](#): NAD27(76) / UTM zone 18N
- [EPSG:2031](#): NAD27(CGQ77) / UTM zone 17N
- [EPSG:2032](#): NAD27(CGQ77) / UTM zone 18N
- [EPSG:2033](#): NAD27(CGQ77) / UTM zone 19N
- [EPSG:2034](#): NAD27(CGQ77) / UTM zone 20N
- [EPSG:2035](#): NAD27(CGQ77) / UTM zone 21N
- [EPSG:2036](#): NAD83(CRS98) / New Brunswick Stereo (deprecated)
- [EPSG:2037](#): NAD83(CRS98) / UTM zone 19N (deprecated)
- [EPSG:2038](#): NAD83(CRS98) / UTM zone 20N (deprecated)
- [EPSG:2039](#): Israel 1993 / Israeli TM Grid
- [EPSG:2040](#): Locodjo 1965 / UTM zone 30N
- [EPSG:2041](#): Abidjan 1987 / UTM zone 30N
- [EPSG:2042](#): Locodjo 1965 / UTM zone 29N
- [EPSG:2043](#): Abidjan 1987 / UTM zone 29N
- [EPSG:2044](#): Hanoi 1972 / Gauss-Kruger zone 18
- [EPSG:2045](#): Hanoi 1972 / Gauss-Kruger zone 19
- [EPSG:2046](#): Hartebeesthoek94 / Lo15
- [EPSG:2047](#): Hartebeesthoek94 / Lo17
- [EPSG:2048](#): Hartebeesthoek94 / Lo19
- [EPSG:2049](#): Hartebeesthoek94 / Lo21

Source: Spatial reference

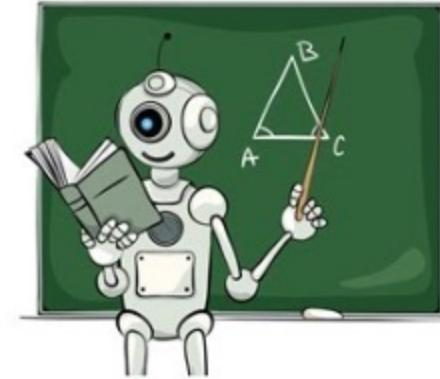
## Machine Learning

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.

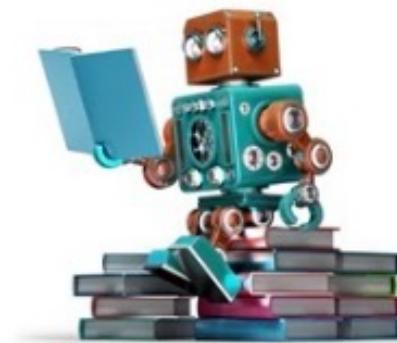


# Machine learning

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



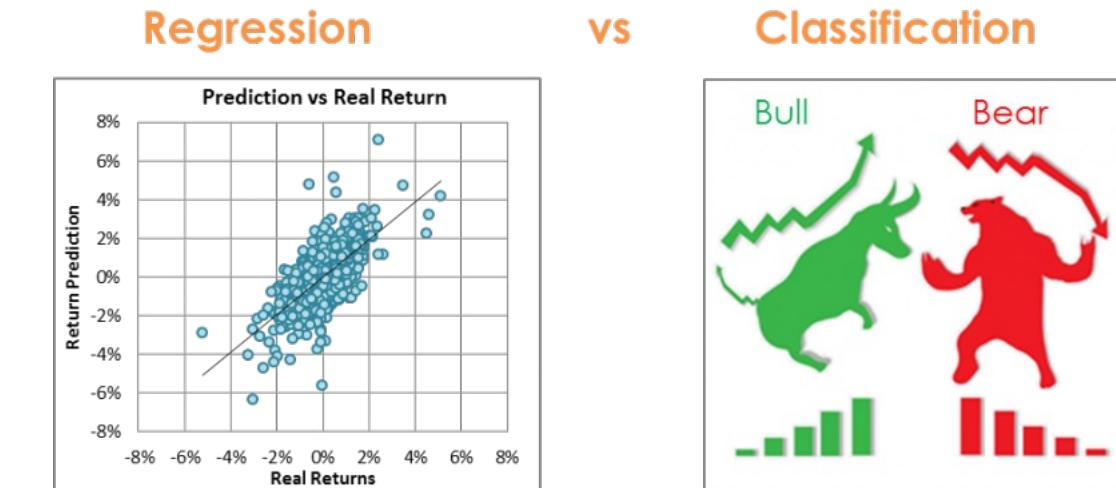
**Unsupervised models:** They do not require a previously marked label. They are algorithms that look for patterns within data without prior knowledge.



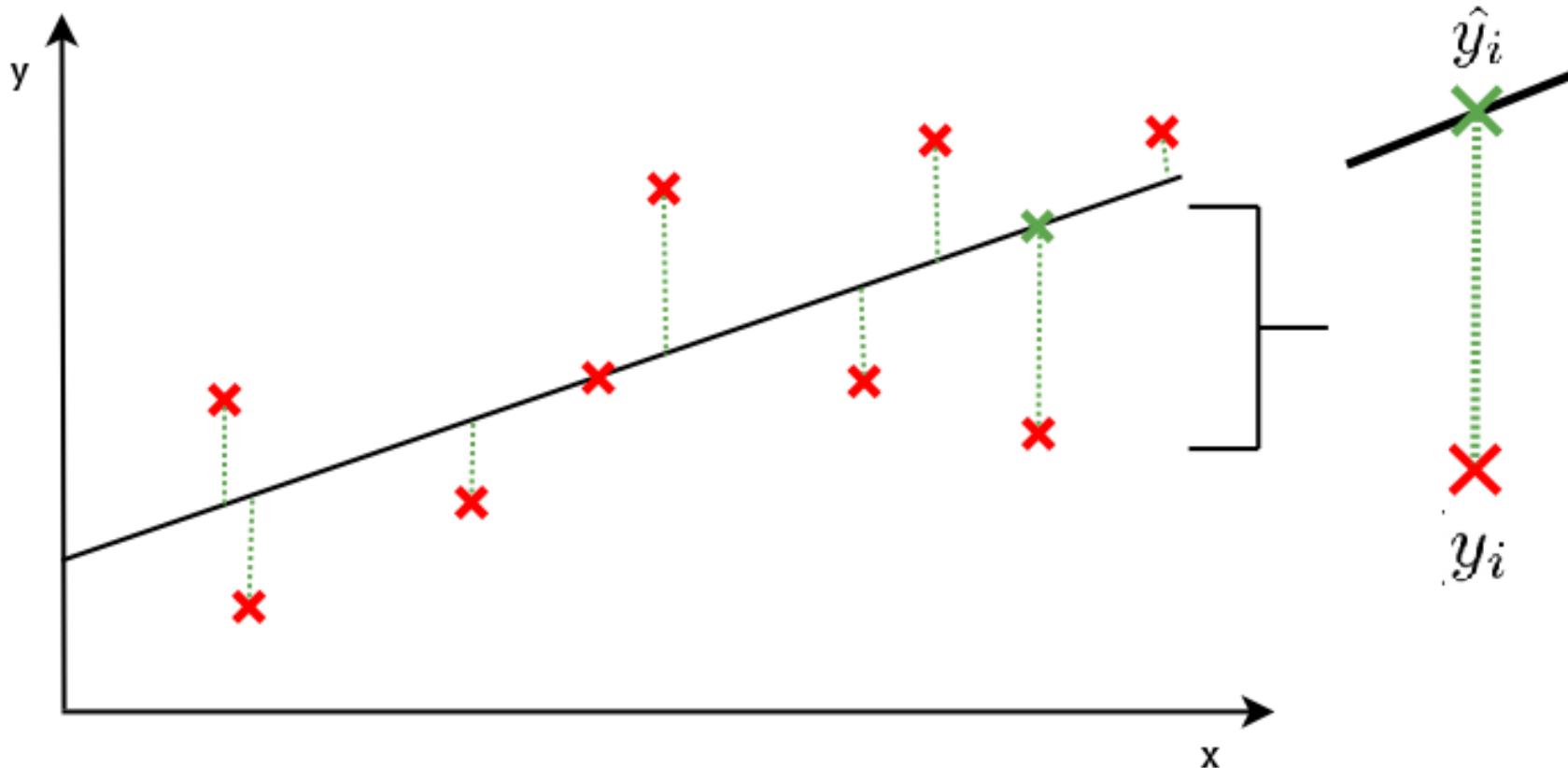
# Supervised vs unsupervised learning

## 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.



# Supervised vs unsupervised learning



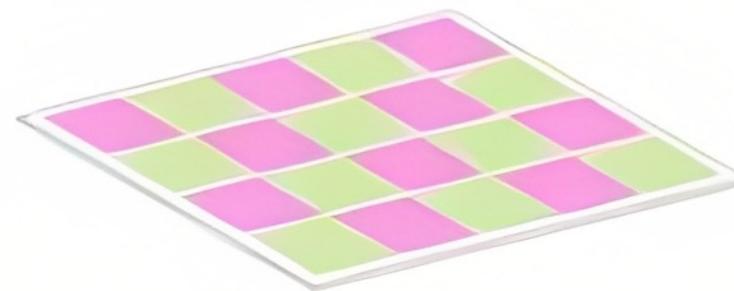
# Spatial autocorrelation - Example

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

Waldo Tobler



High Spatial Autocorrelation  
(Clustering)



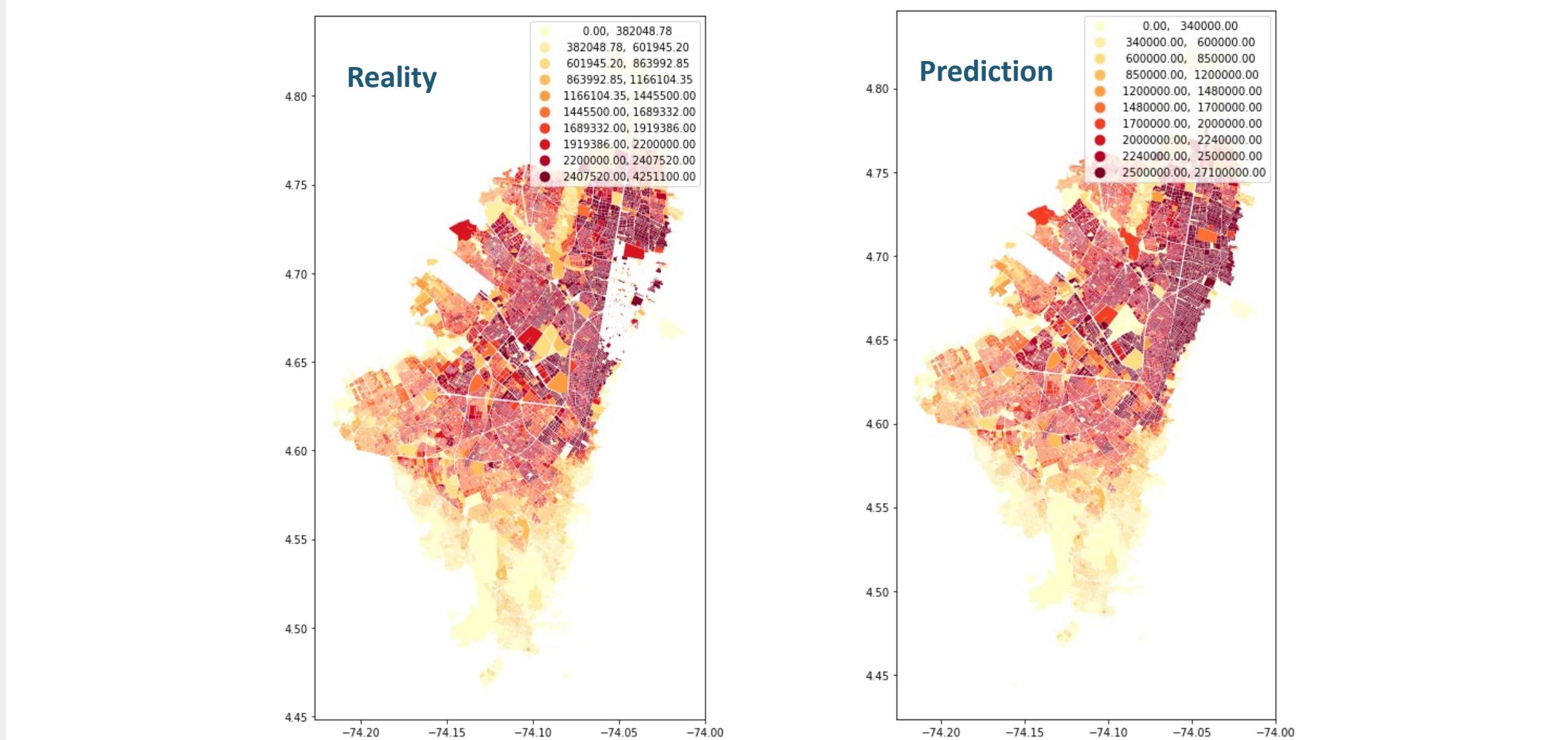
Low Spatial Autocorrelation  
(Checkerboard)

# High spatial autocorrelation



Image from: Domain

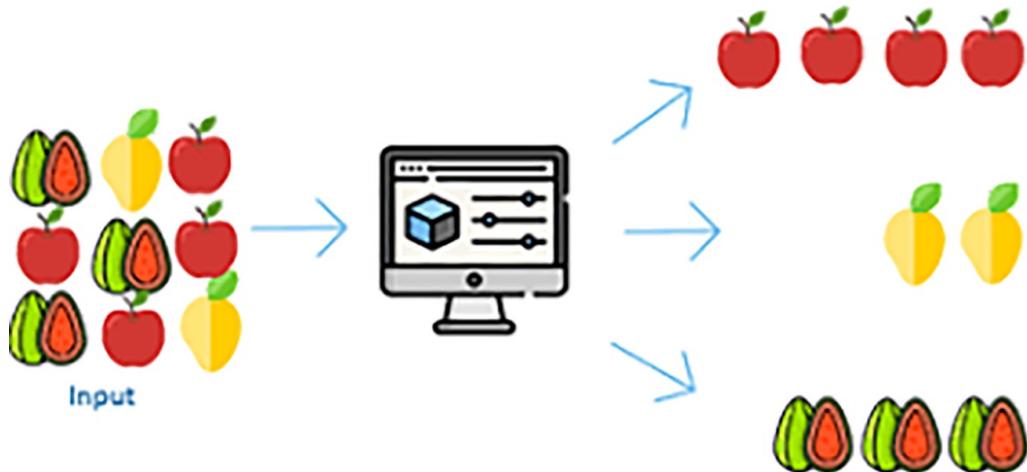
# Example: Housing price - Geographically weighted regression



# Unsupervised models

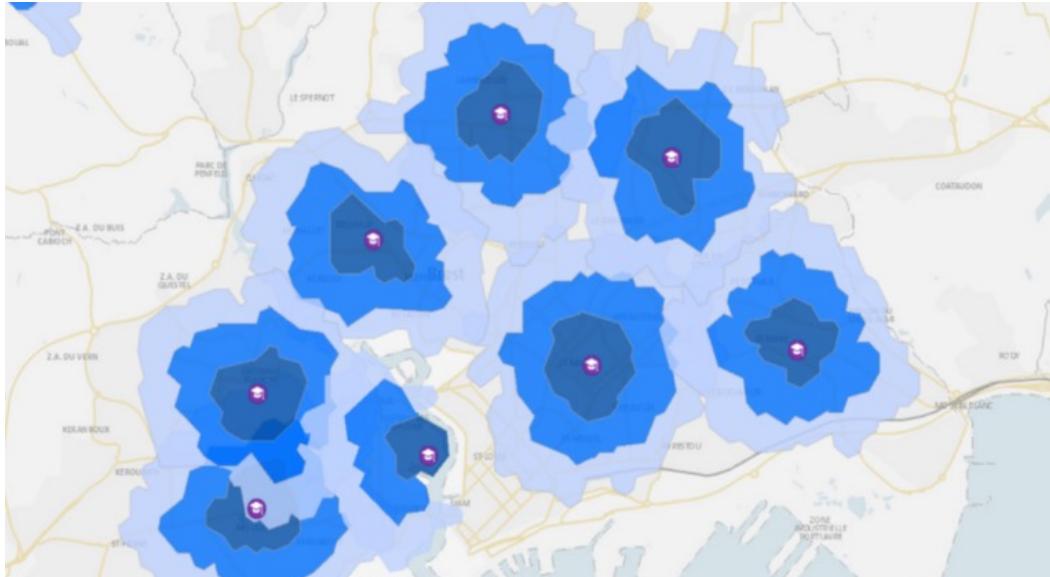
## Unsupervised models

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

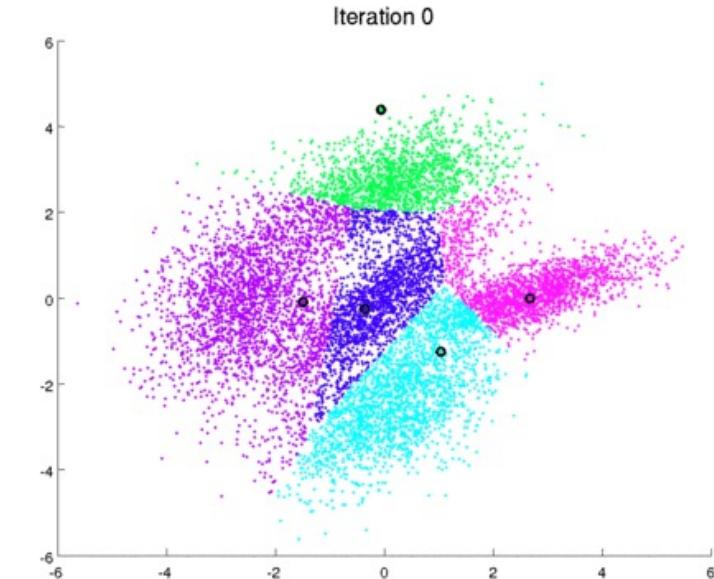


# Unsupervised models

Geospatial clustering



K-means



K-prototypes

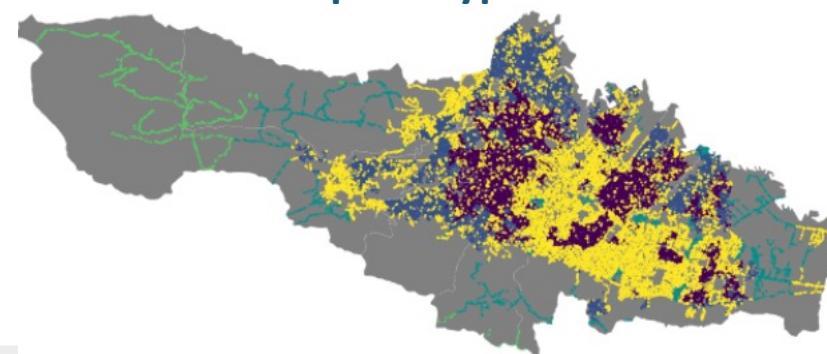
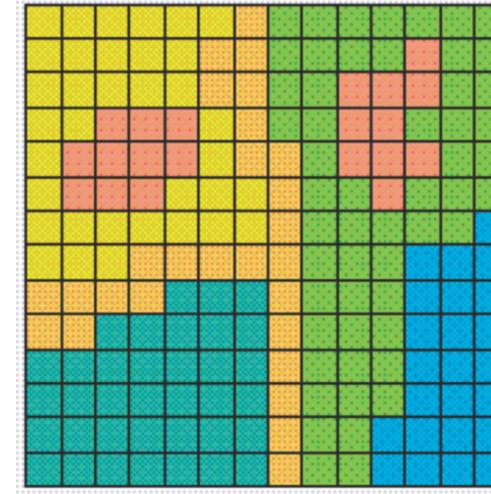


Image from: Scikit-learn

# Finding and using spatial data

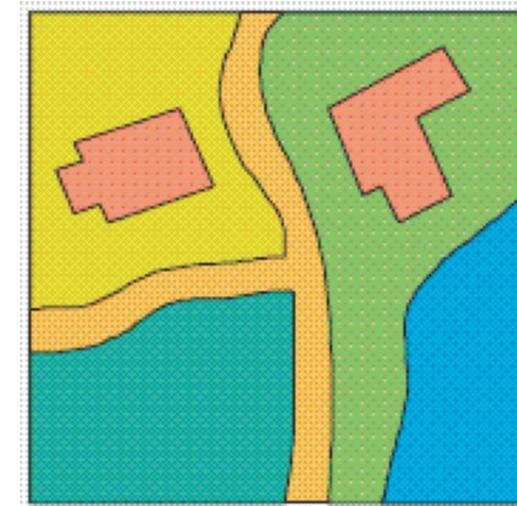
## Raster map:

- 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.



# 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:

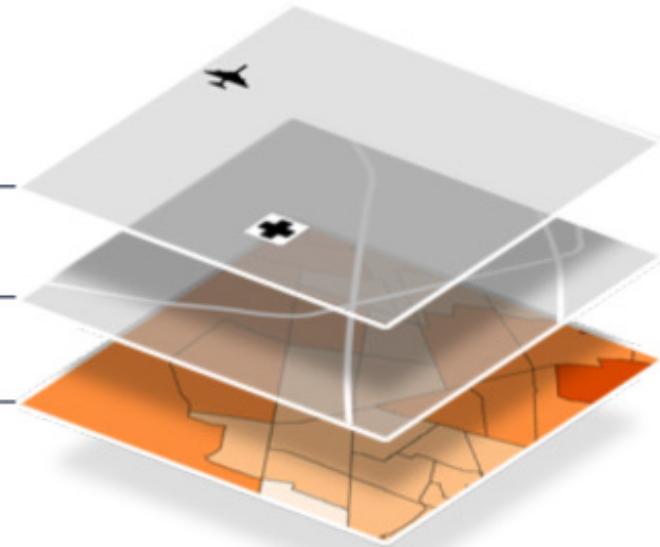
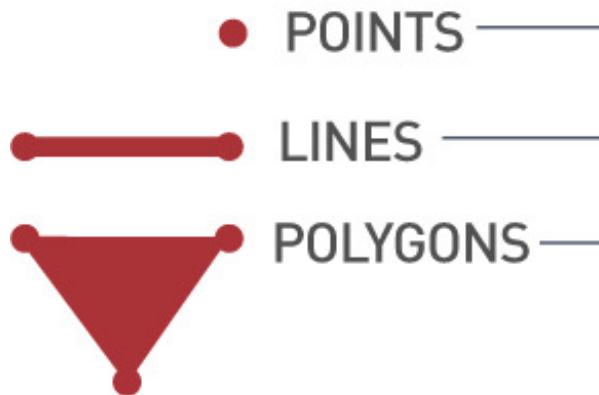
**Point:** Addresses, locations, points of interest, etc

**Lines:** streets, freeways, borders, etc

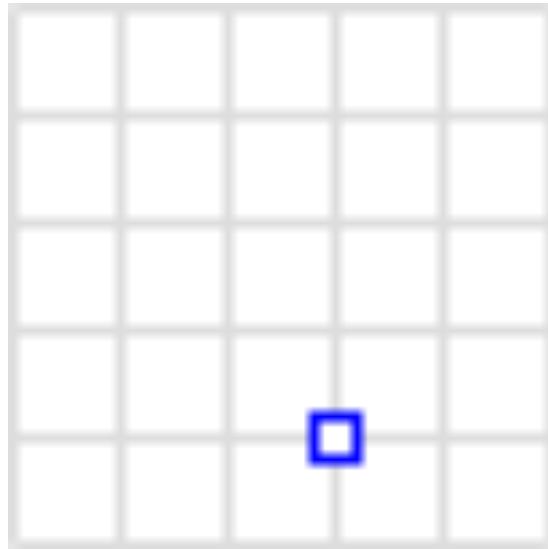
**Polygons:** Countries, cities, Cadastre

## Advantages:

- Spatial operations
- Spatial aggregation
- Spatial Join

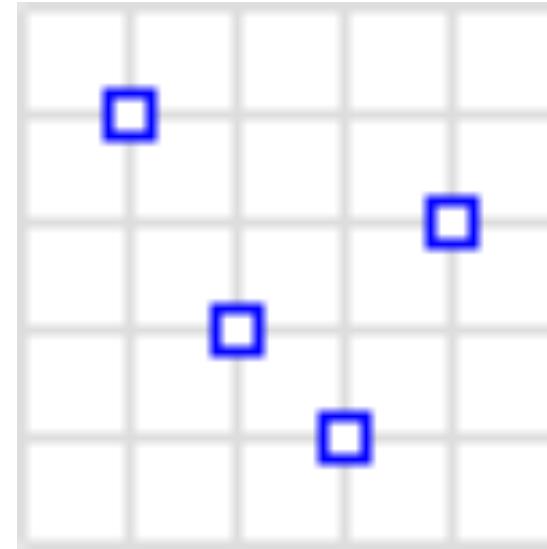


# Vectorial map:



**Point**

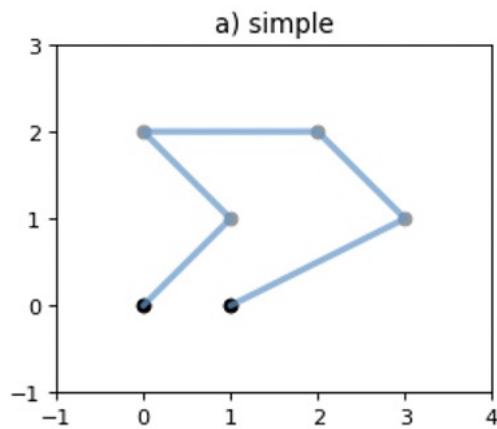
It represents a point on the Earth's surface (latitude, longitude)



**MultiPoint**

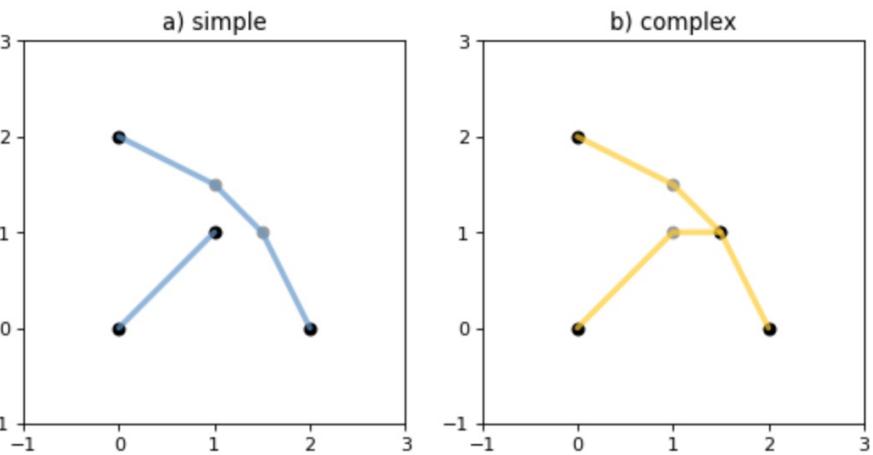
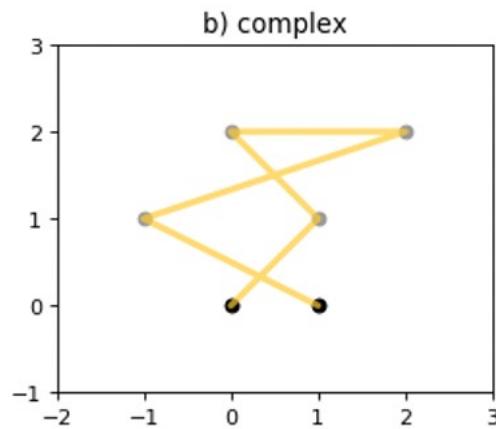
Constructed by multiple points that cannot be connected.

# Vectorial map:



**Linestring (Linea)**

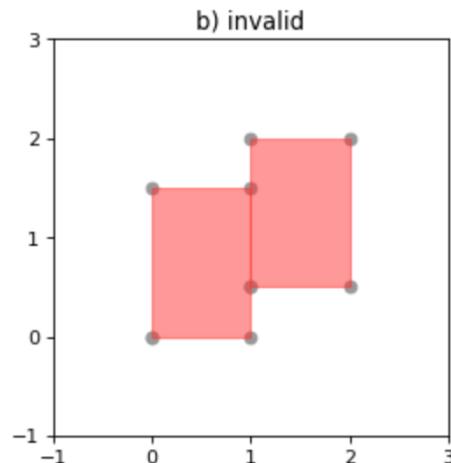
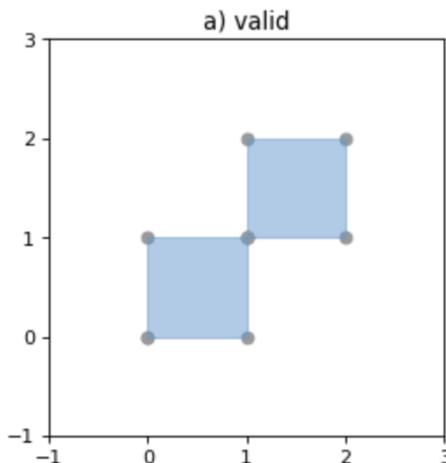
El objeto LineString construido representa una o más splines lineales conectadas entre los puntos. Un LineString puede cruzarse (es decir, ser complejo y no simple).



**Multilinestring (Multilínea)**

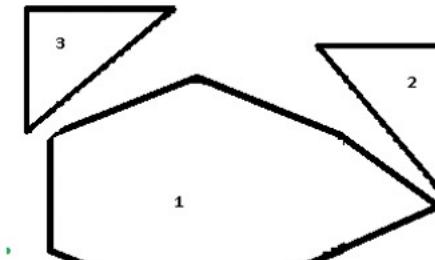
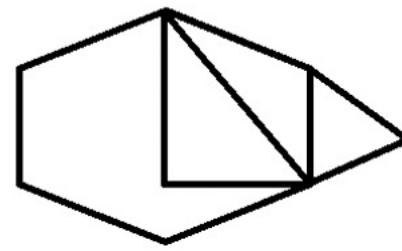
El objeto multilinestring construido representa una o más líneas conectadas entre los puntos.

# Vectorial map:



## Polygon

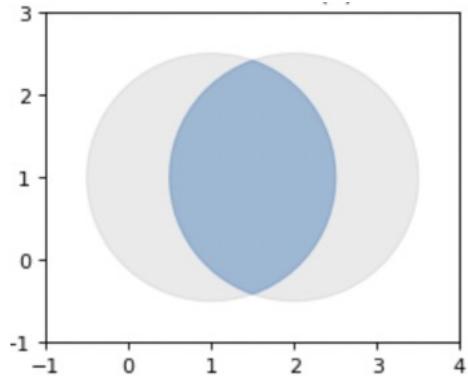
It is a closed figure made up of coordinates that enclose a region in the plane.



## Multipolygon

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

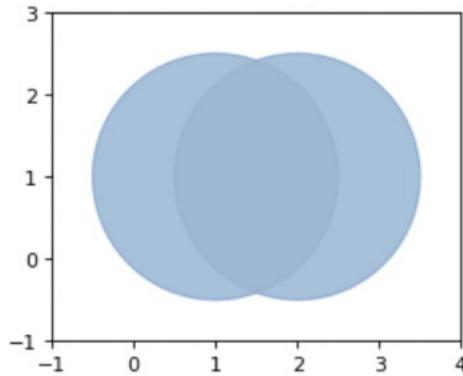
# Spatial operations



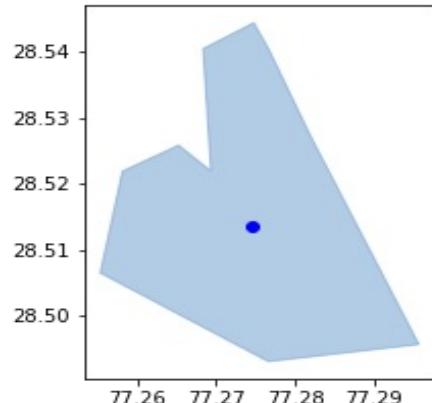
Intersection



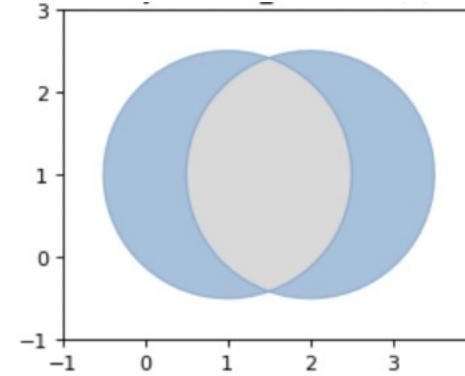
Contour



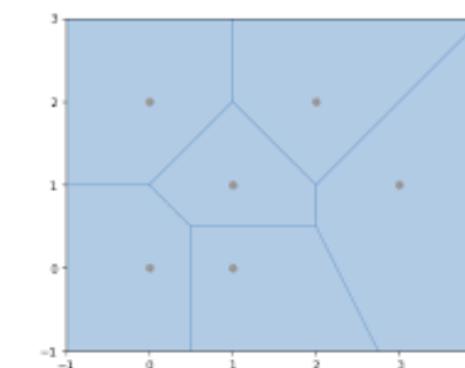
Union



Difference

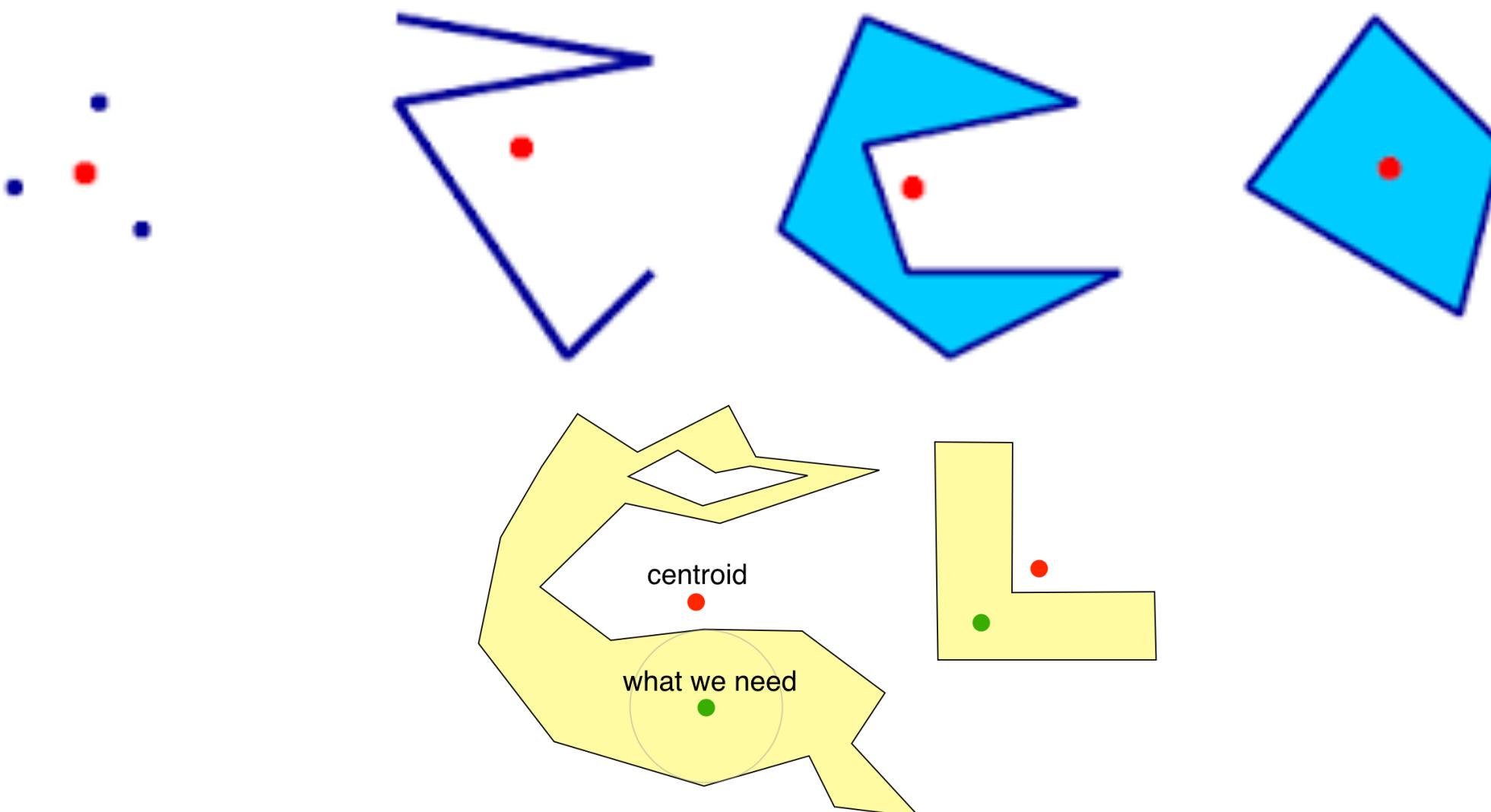


Centroid

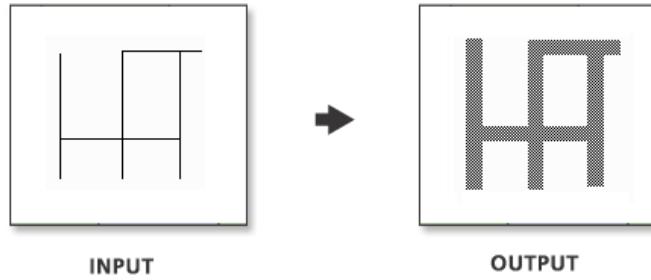


Within

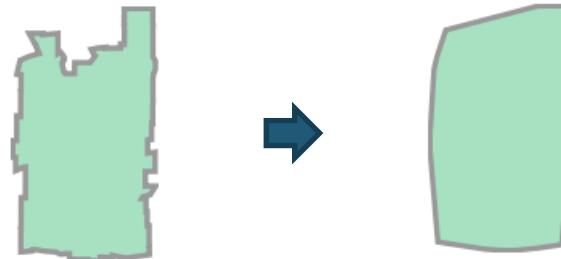
# Spatial operations



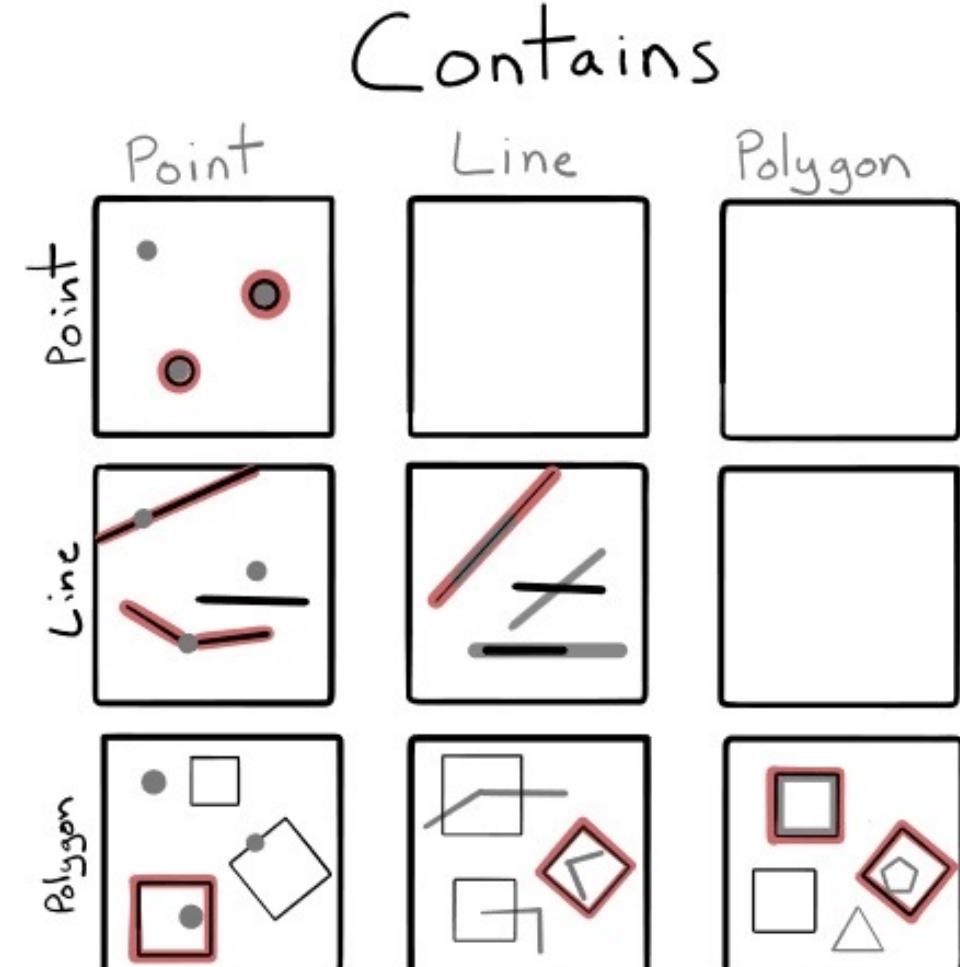
# Spatial operations



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

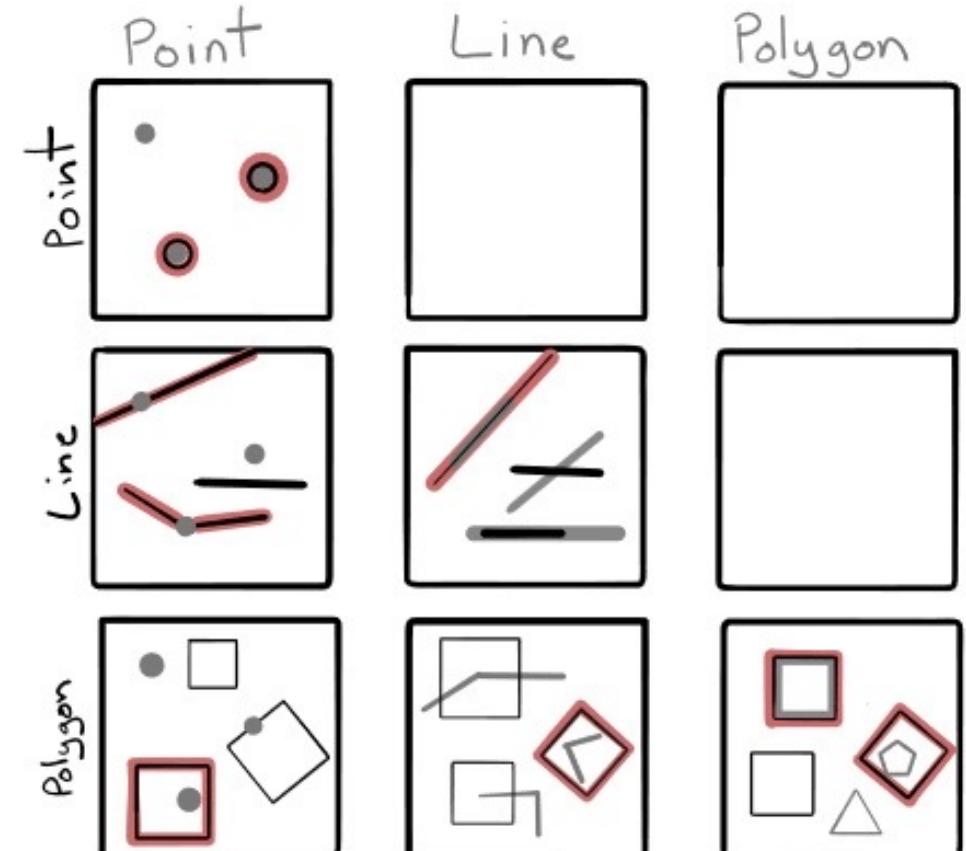
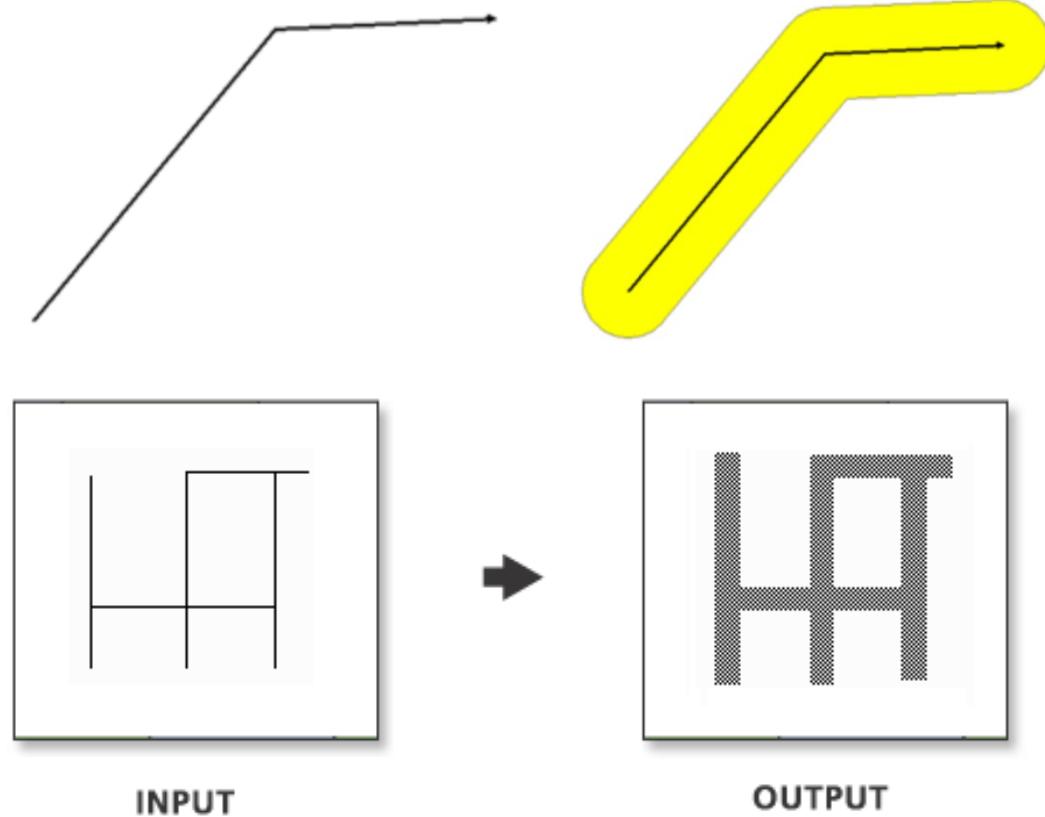


**Convex capsule:** a set of points that contains the intersection of all convex sets within a polygon.



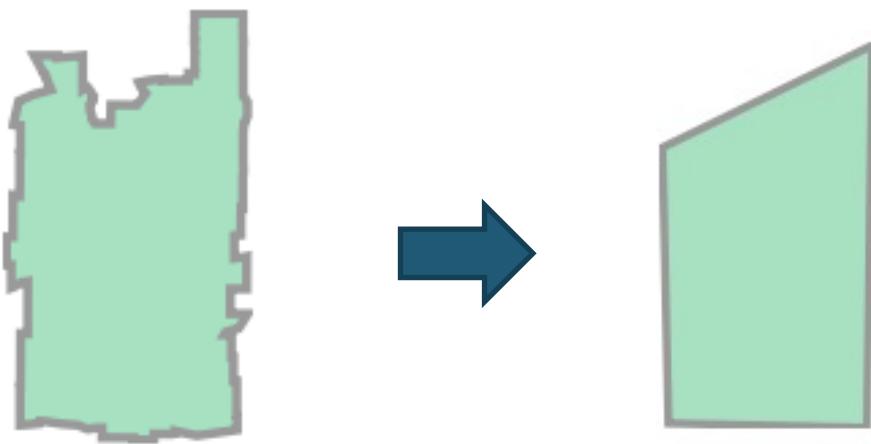
# Spatial operations

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

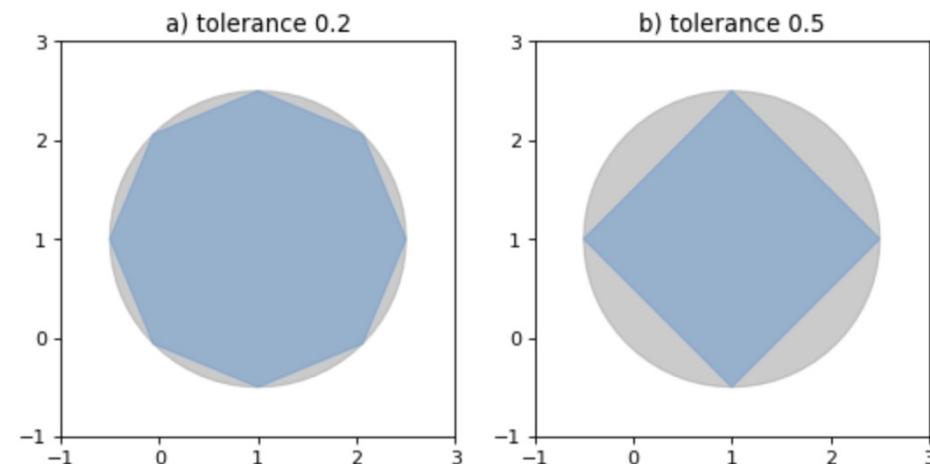


# Spatial operations

## Simplification

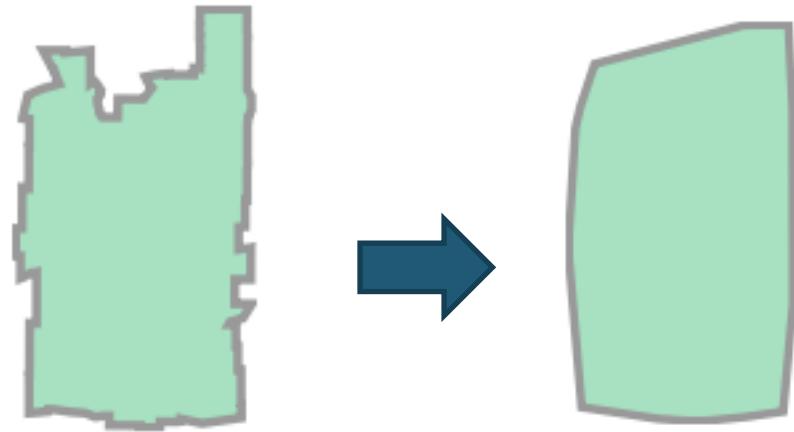


This function simplifies the geometry using a specified tolerance.

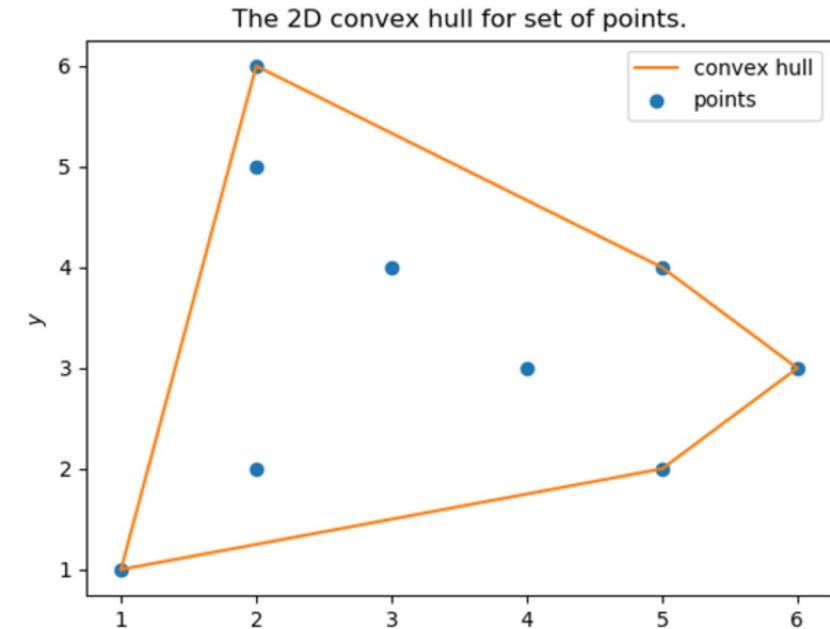


All points of the simplified object will be within the tolerance distance from the original geometry.

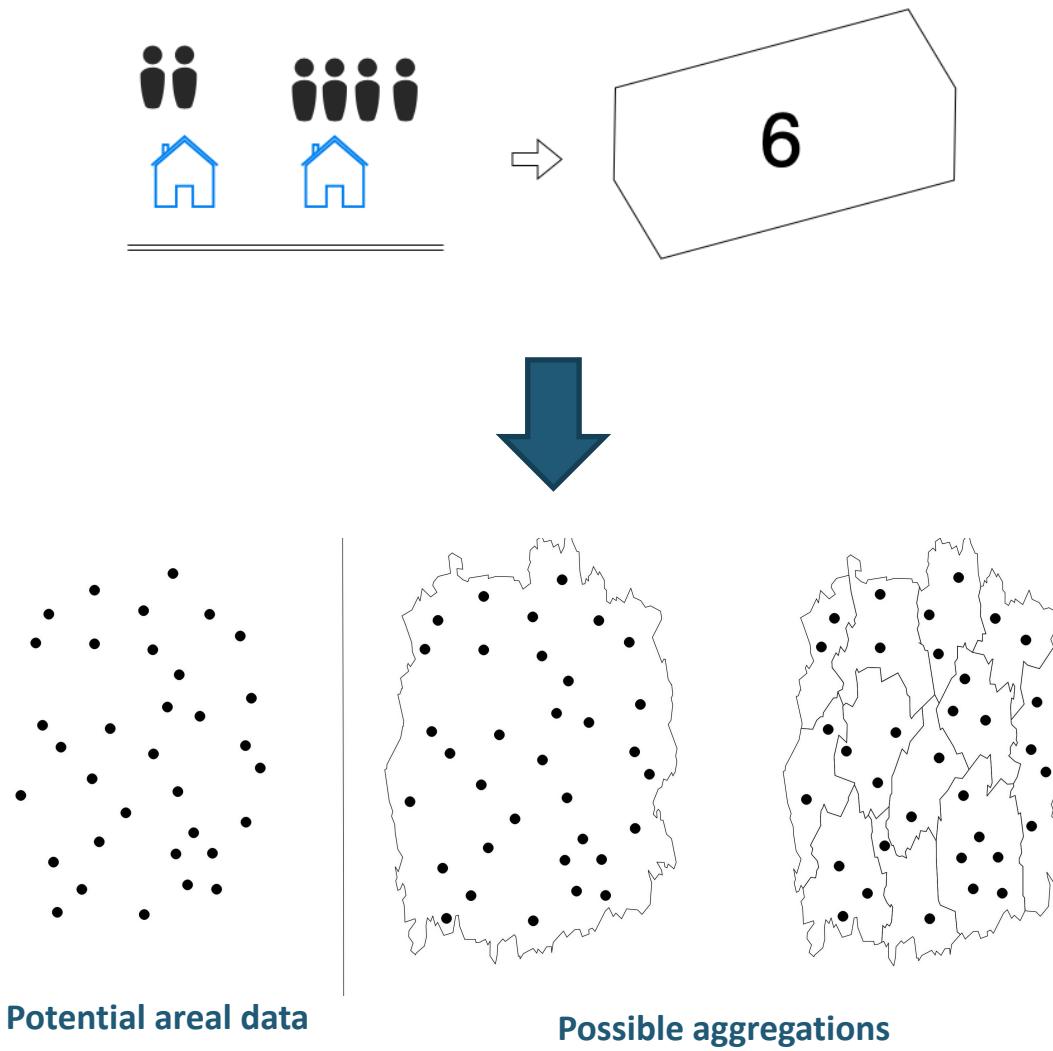
# Spatial operations



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



# Spatial aggregation

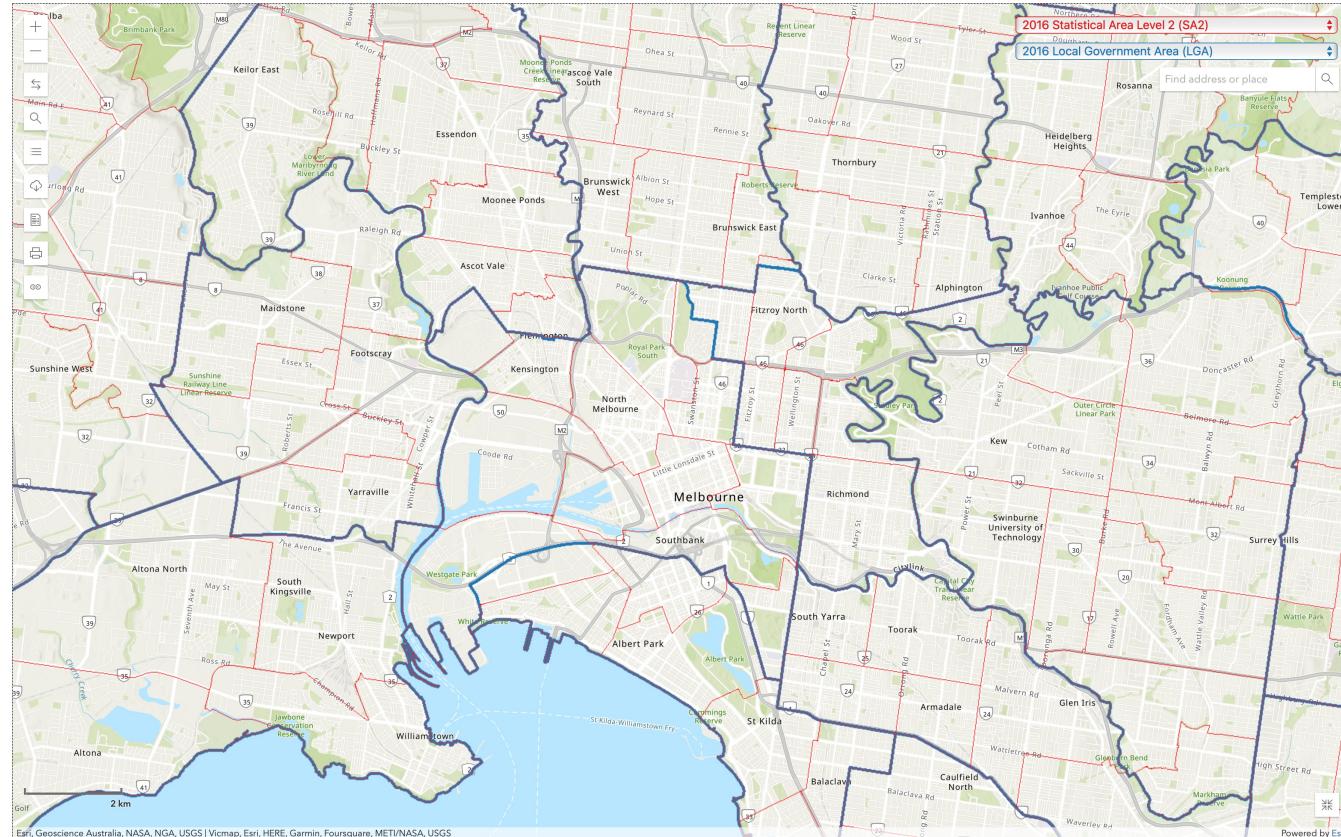


# Example: Spatial aggregation

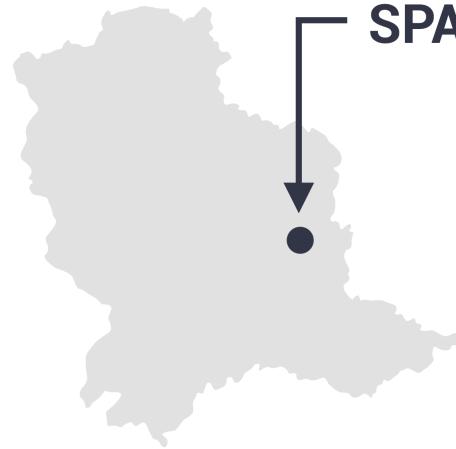
Socio-Economic Indexes for Areas (SEIFA) – SA2 - 2016



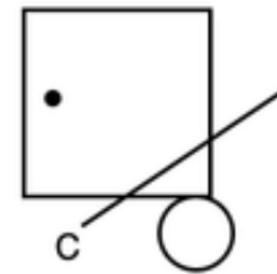
PHIDU - Education (LGA) 2015-2016:



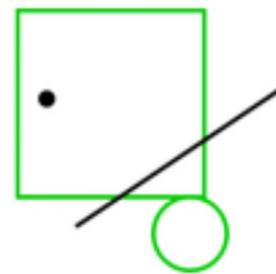
# Spatial Join



## Types of spatial join

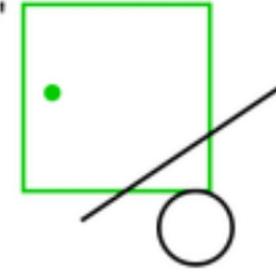
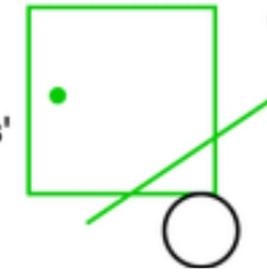


spacial join:  
op = 'touches'

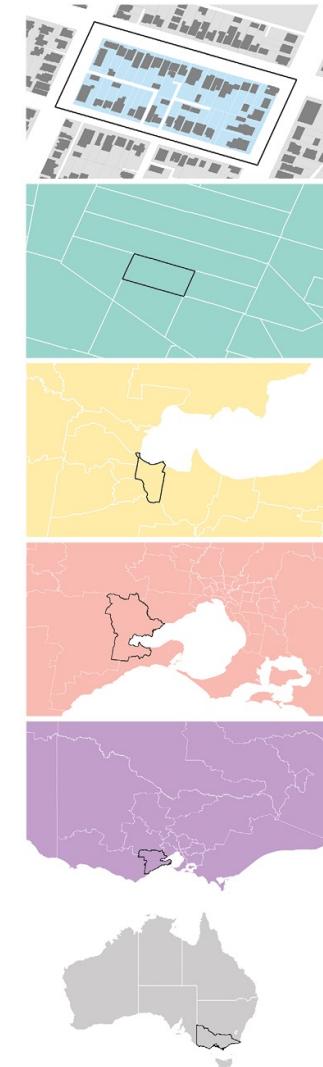
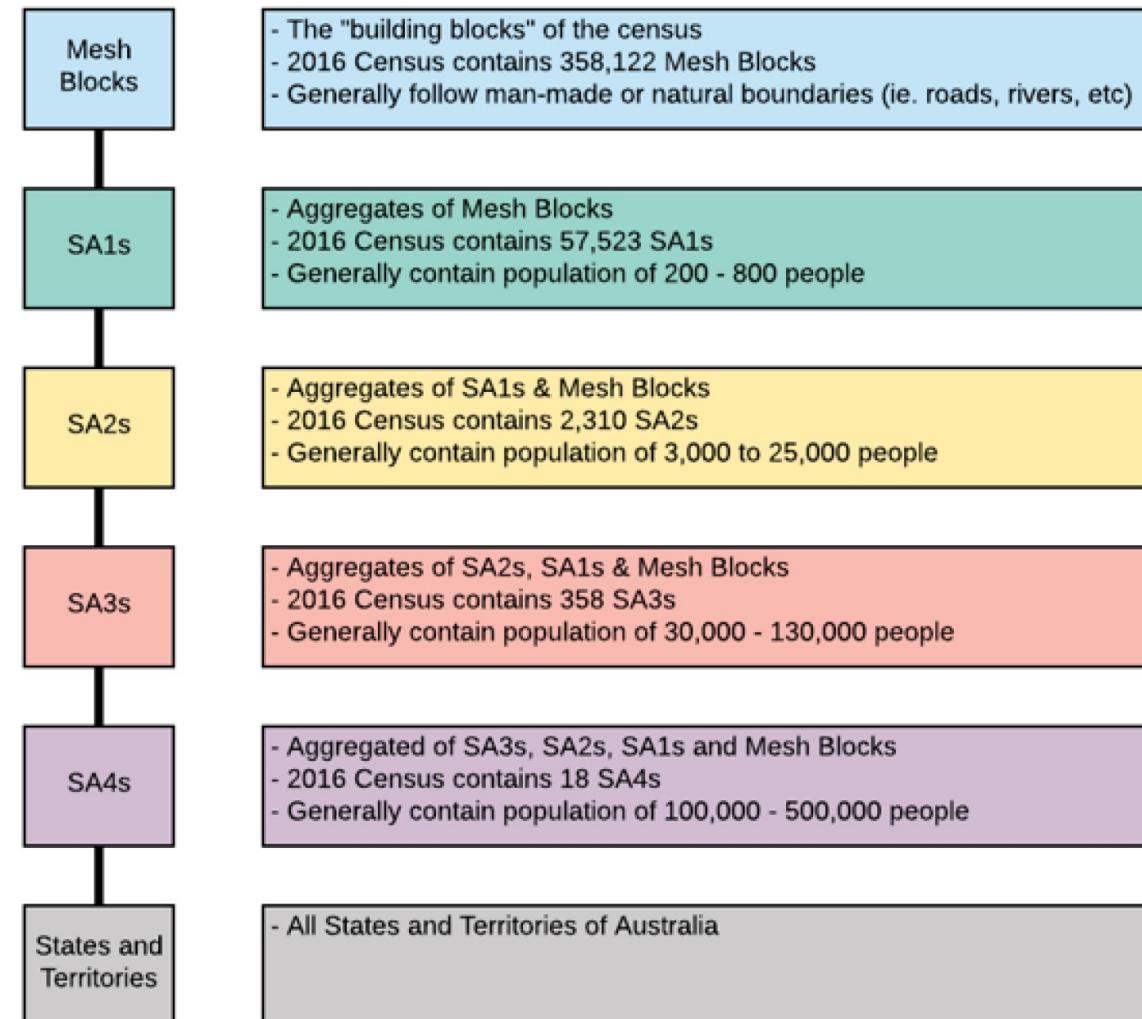


spacial join:  
op = 'contains'

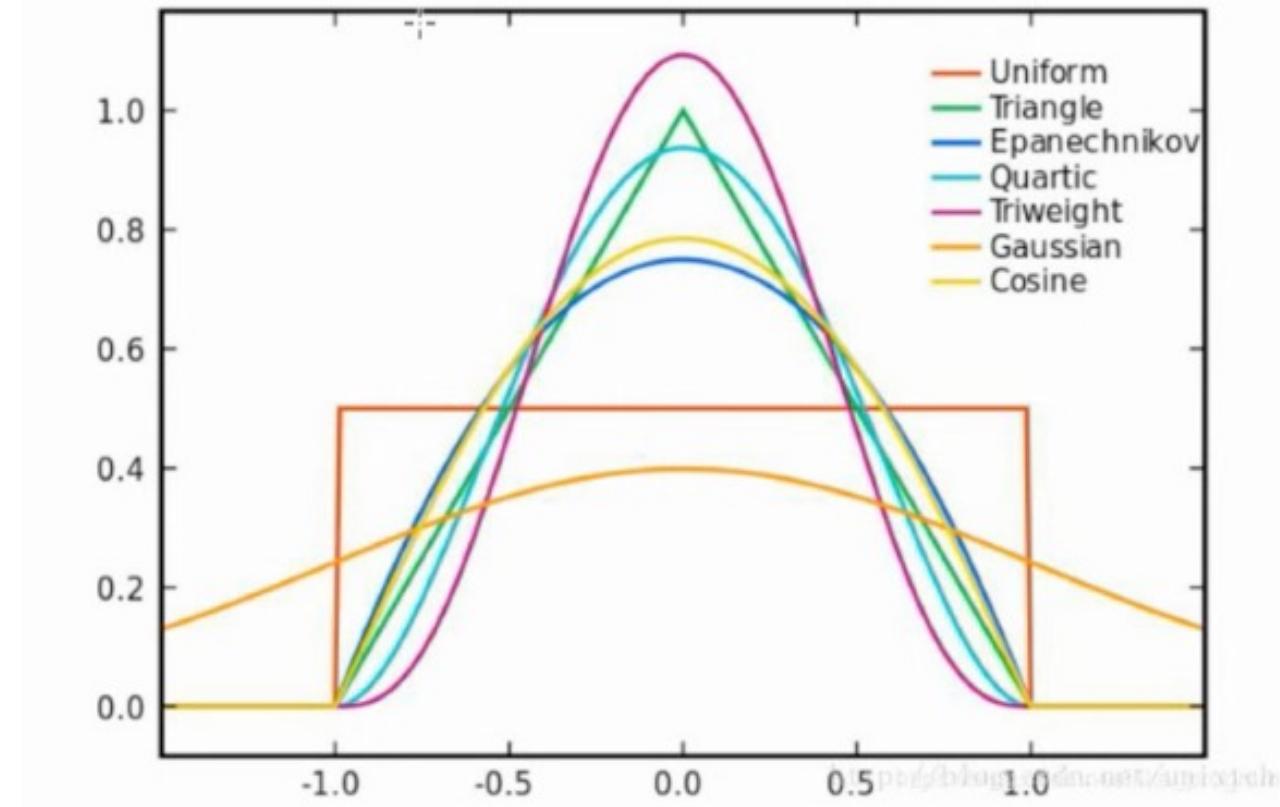
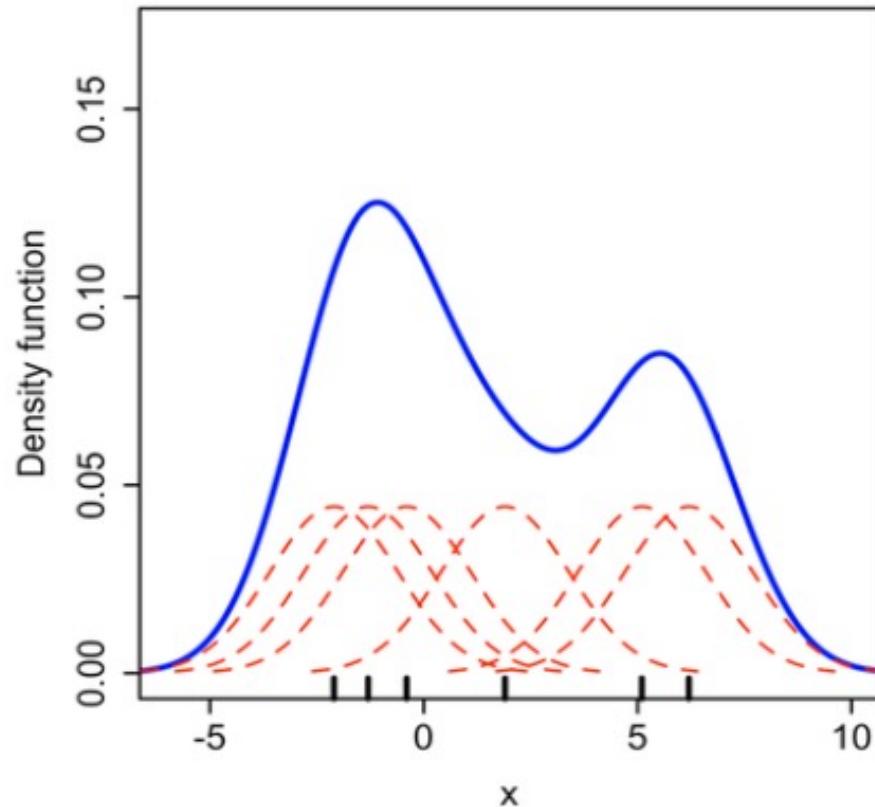
spacial join:  
op = 'intersects'



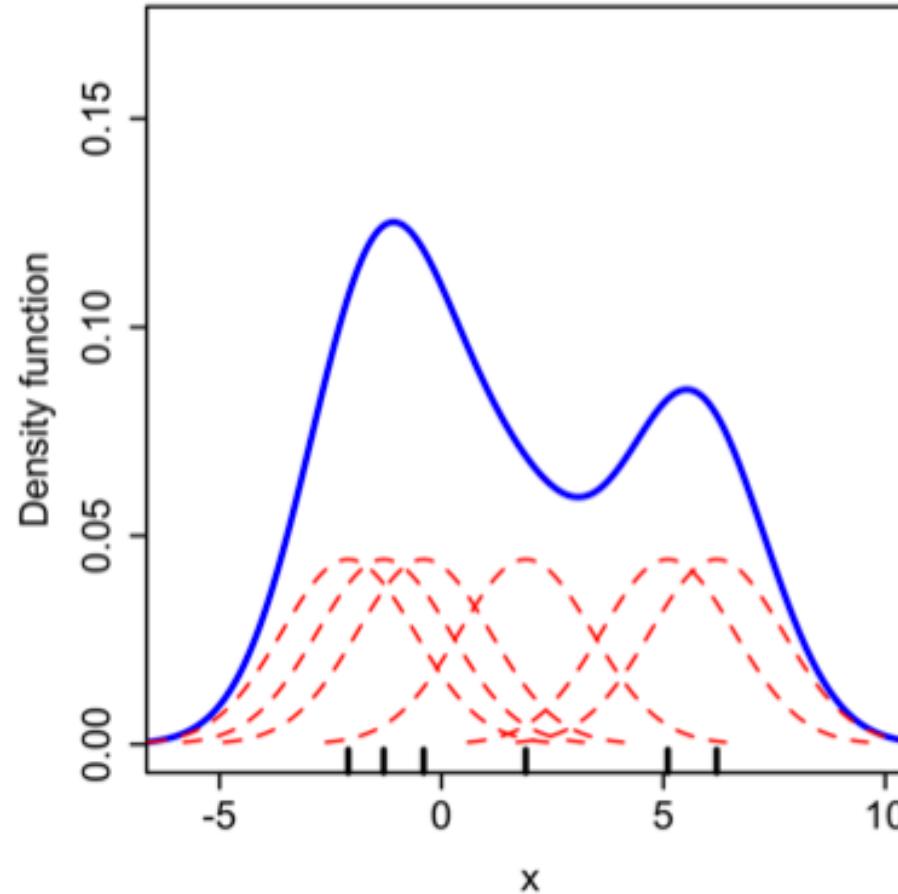
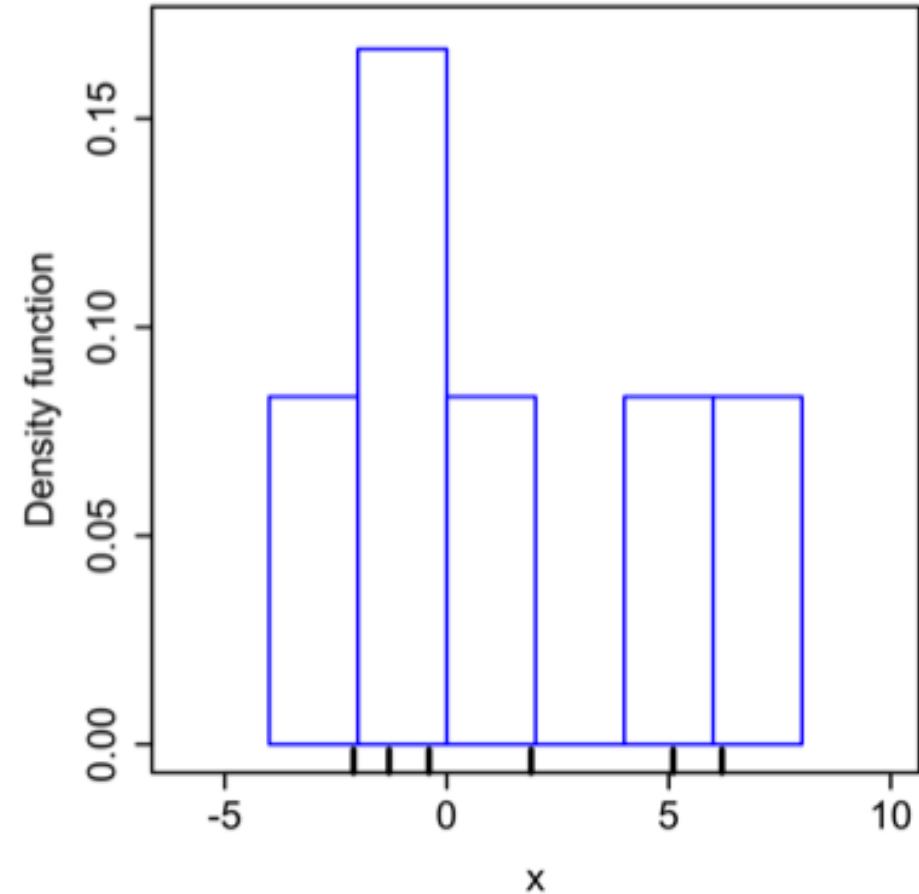
# Spatial aggregation: Spatial boundaries



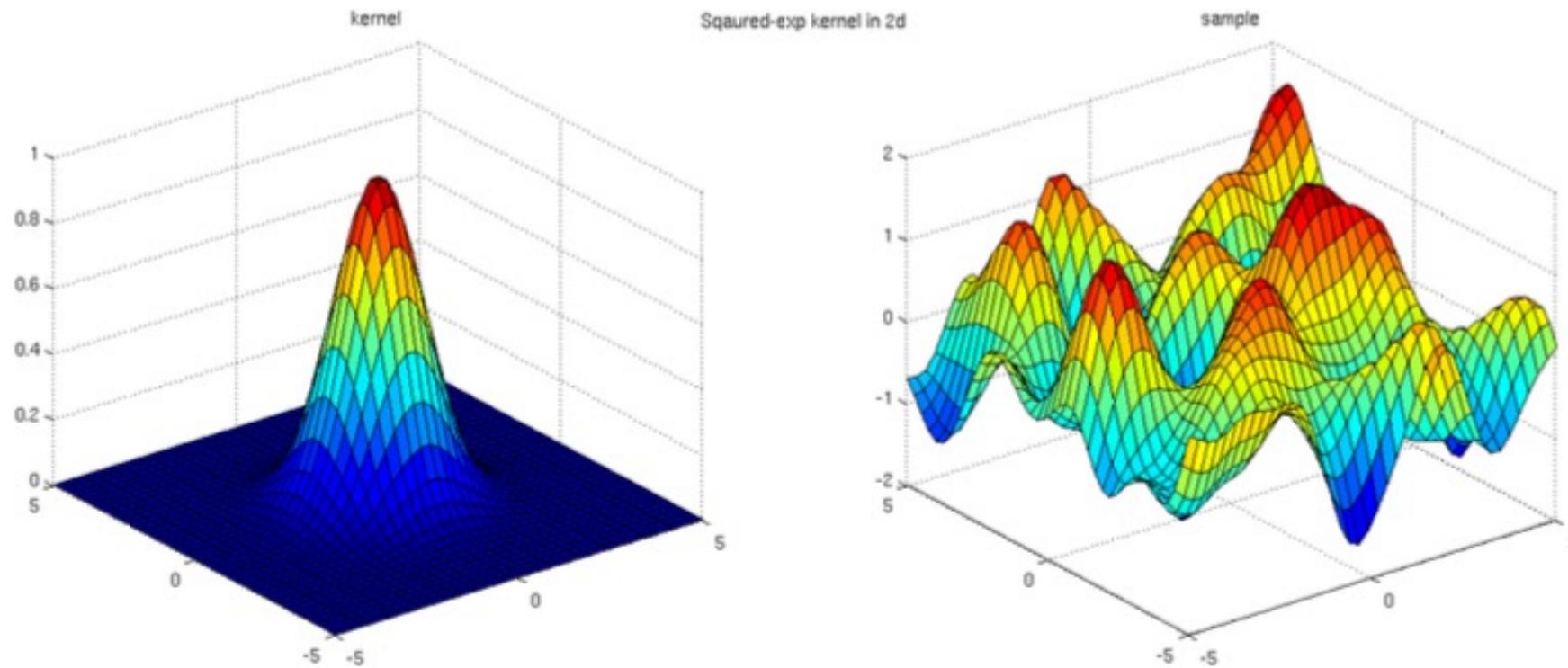
# Kernel Density



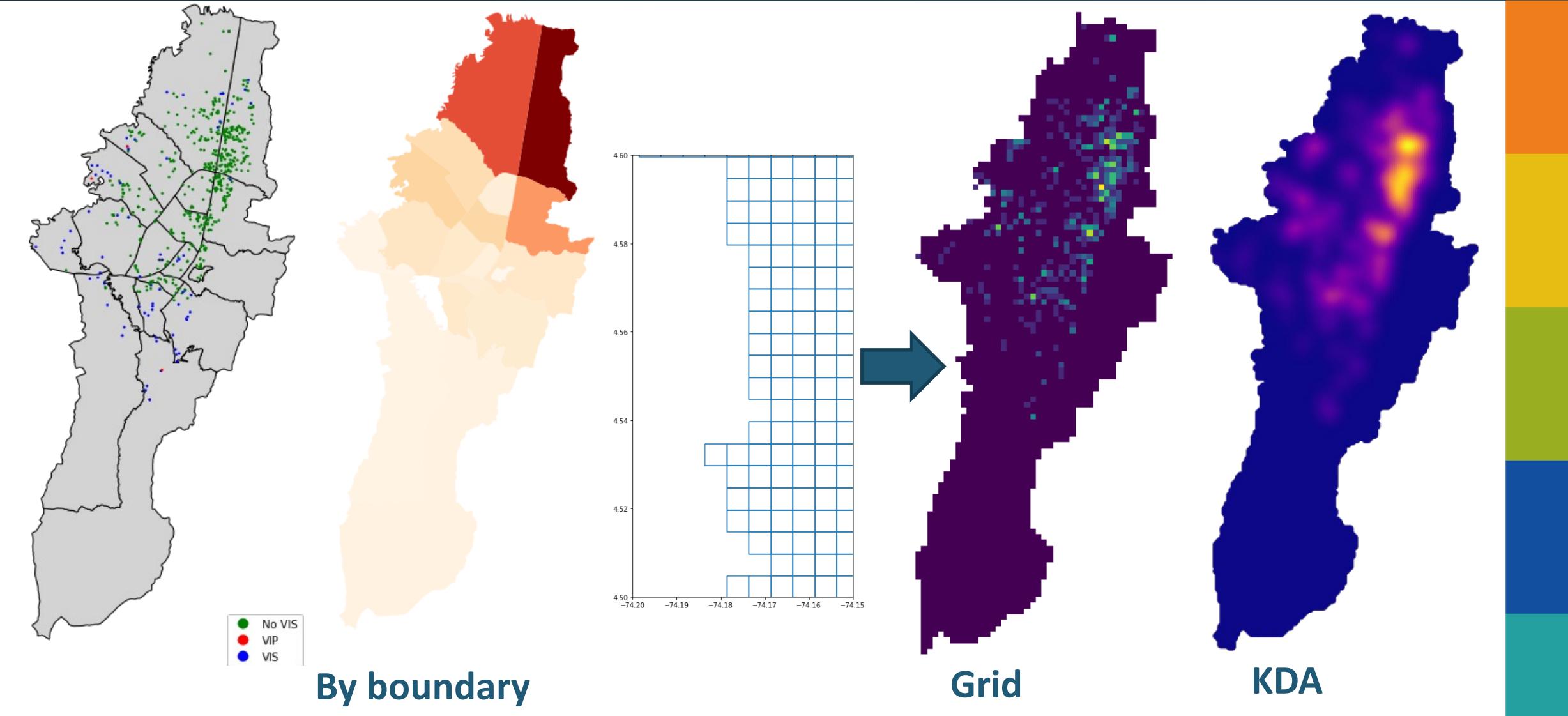
# Kernel Density



# Kernel Density



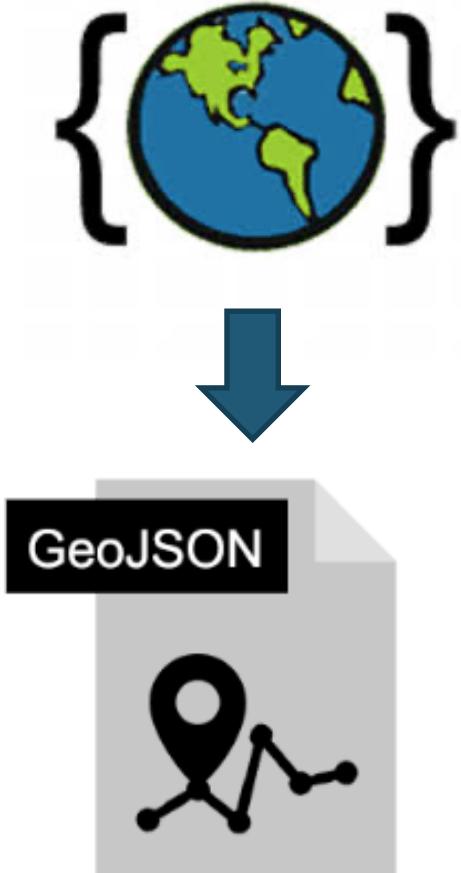
# Kernel Density



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.





## Shapefile

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

- .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



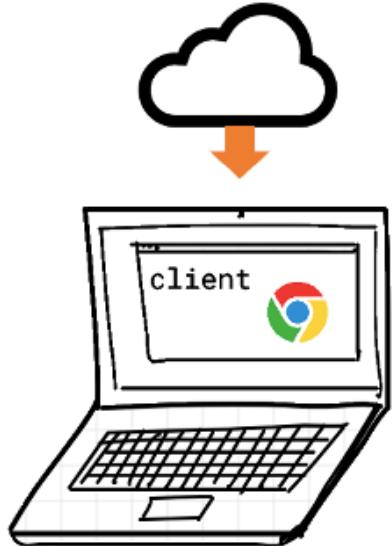
**mongo DB**

# Exercise – R

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

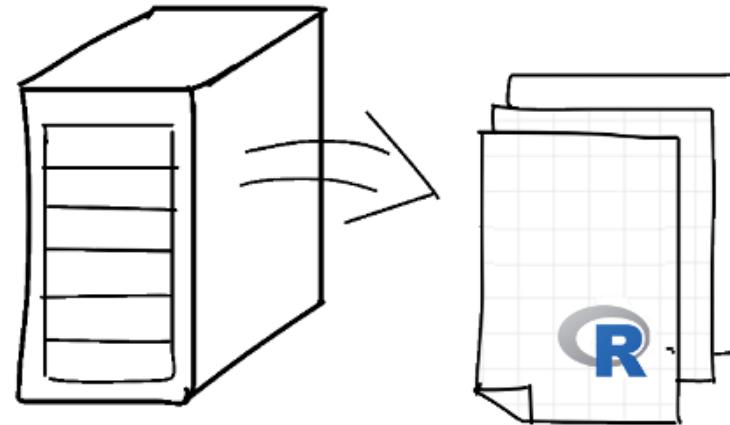


## Choices for the R environment



**a) Cloud environment**

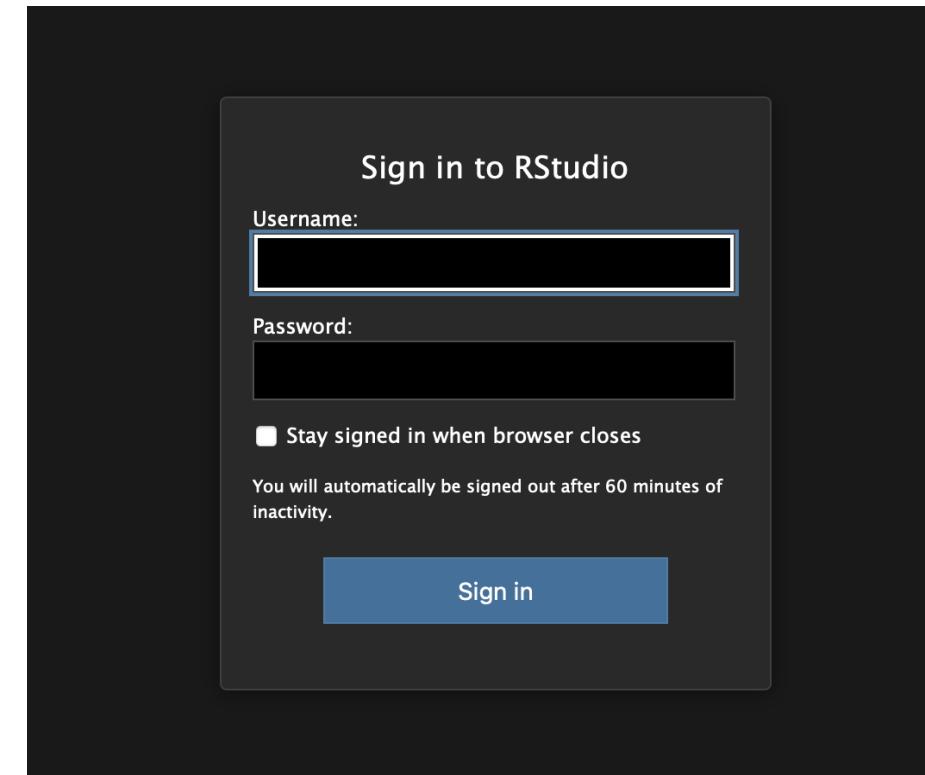
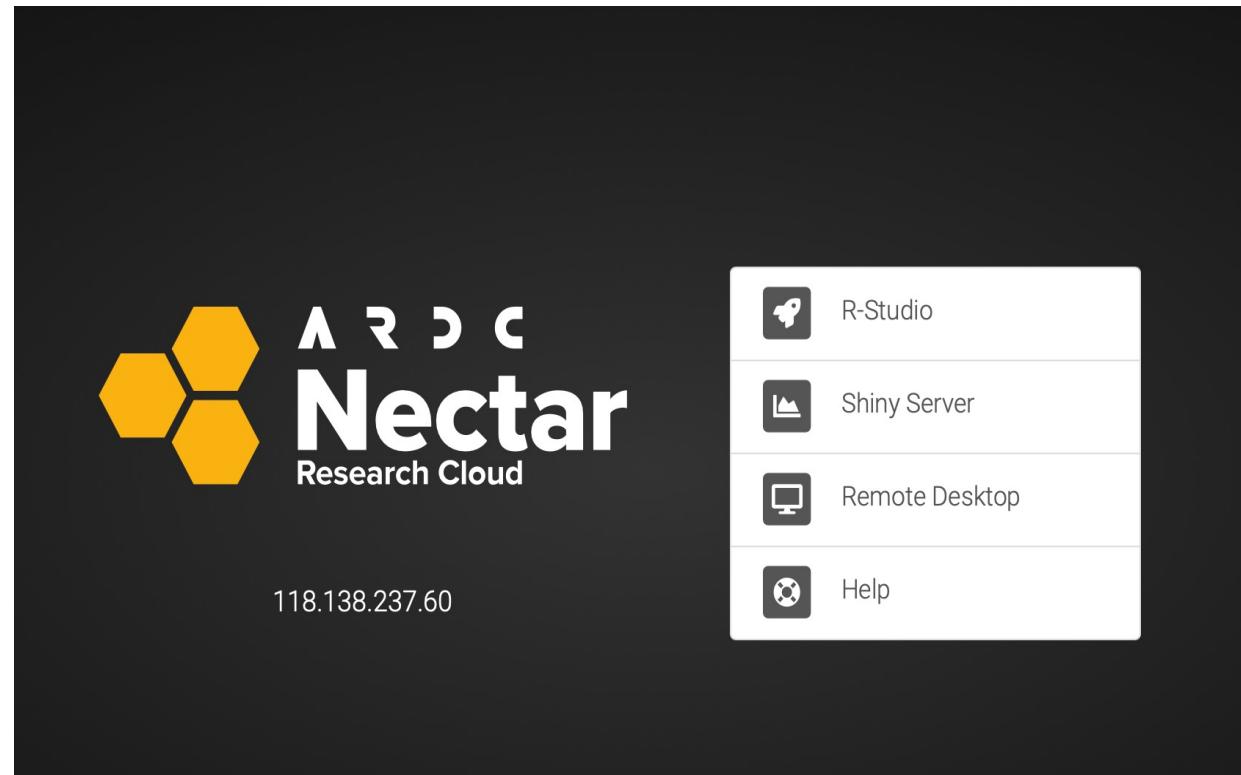
- Compatibility
- Less time spent installing R and dependencies.
- Better Package compatibility
- New users



**b) Local environment**

- Customisable
- Privacy
- Large data sets
- Experienced users

# Exercise – R



# 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**

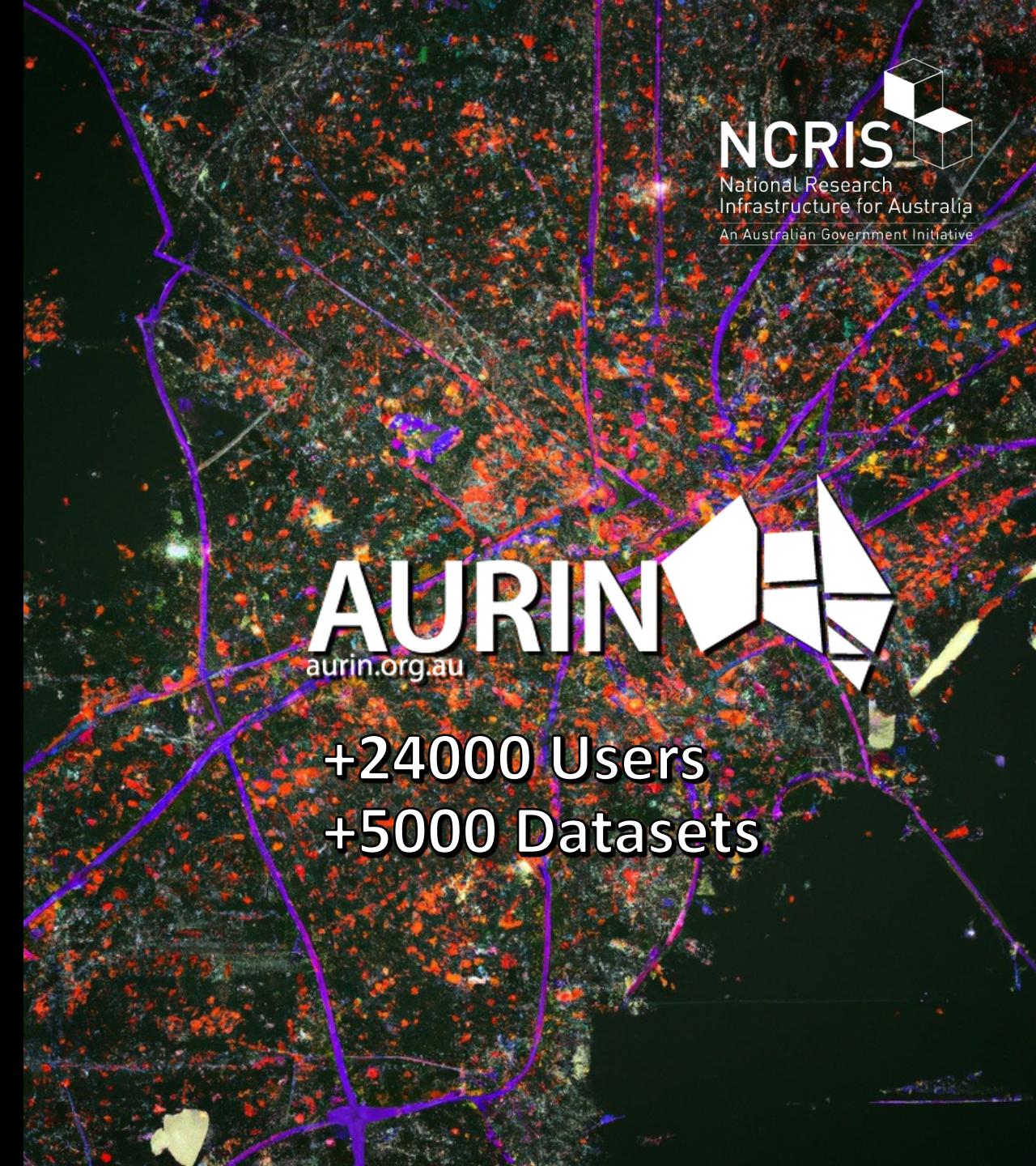
**Deciding on integration**

**Producing a new data product**

**The Geosocial work package**

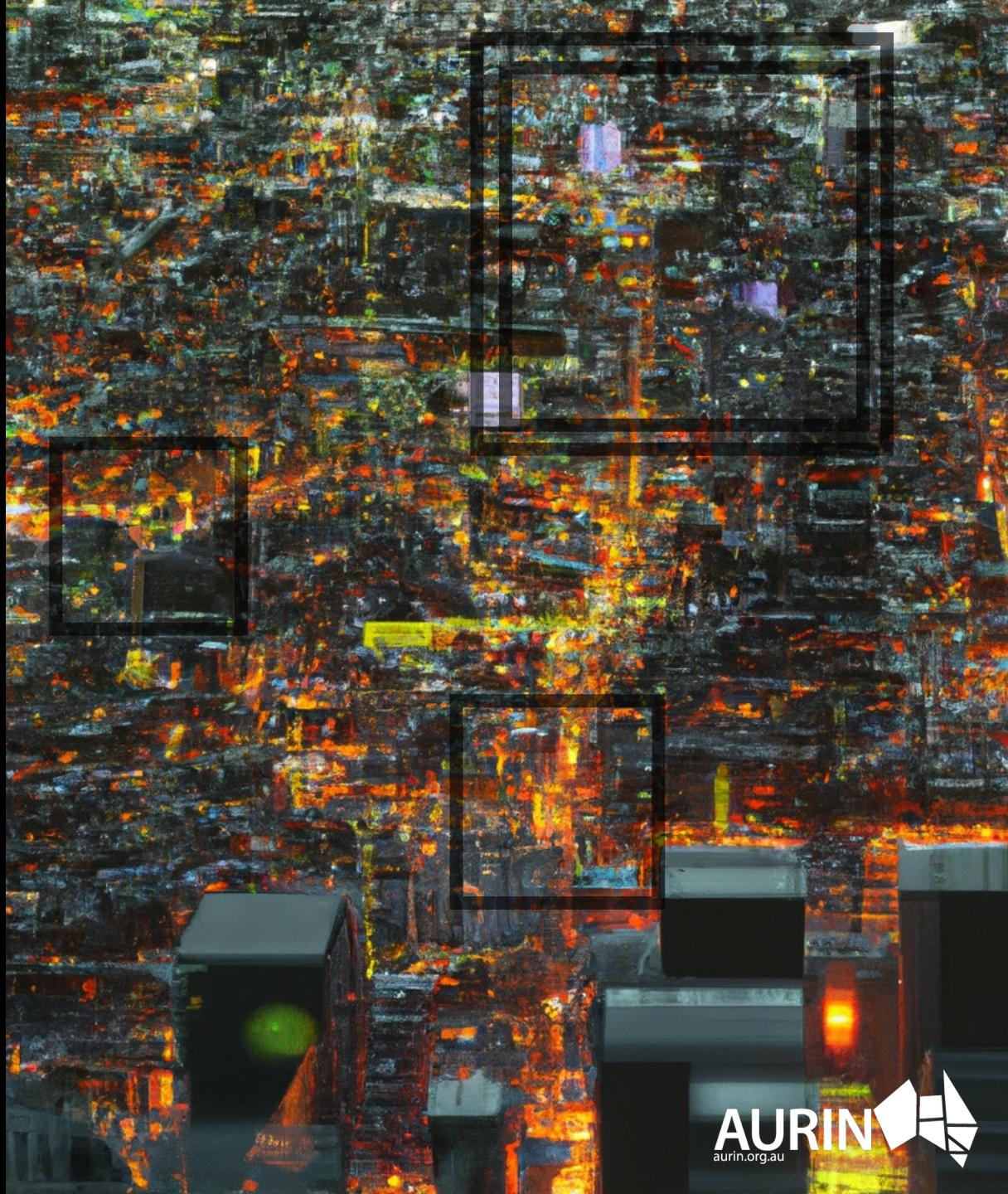
# AURIN in a Glance

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



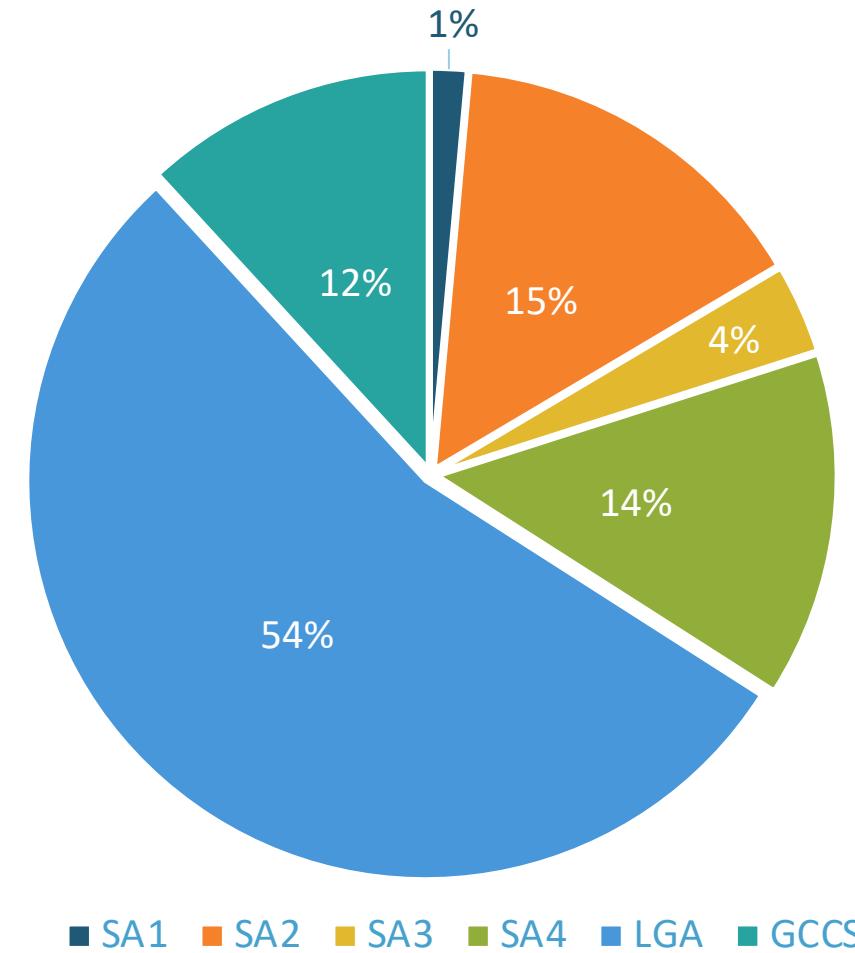
# Addressing key urban challenges in Australia.

- 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?

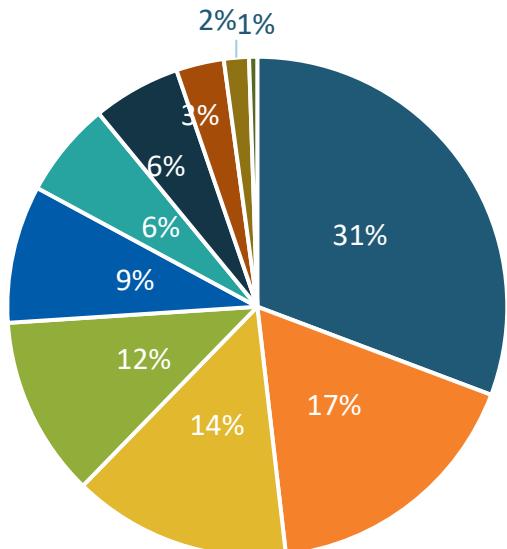




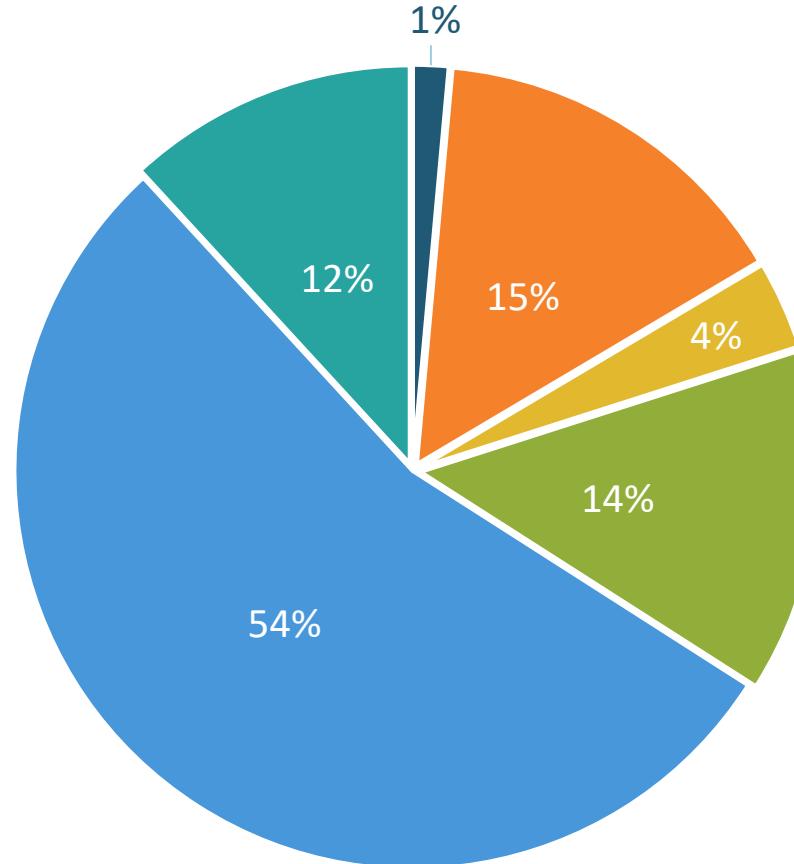
Datasets: 5,290



Classification by theme (NPL)



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



- SA1
- SA2
- SA3
- SA4
- LGA
- GCCS

# OpenStreetMap

OpenStreetMap es un proyecto abierto y colaborativo que permite crear mapas editables y libres. Los mapas se crean utilizando información geográfica capturada con dispositivos GPS móviles, ortofotografías y otras fuentes libres. Esta cartografía, tanto las imágenes creadas como los datos vectoriales almacenados en su base de datos, se distribuyen bajo licencia abierta.

- [Ver enlace](#)

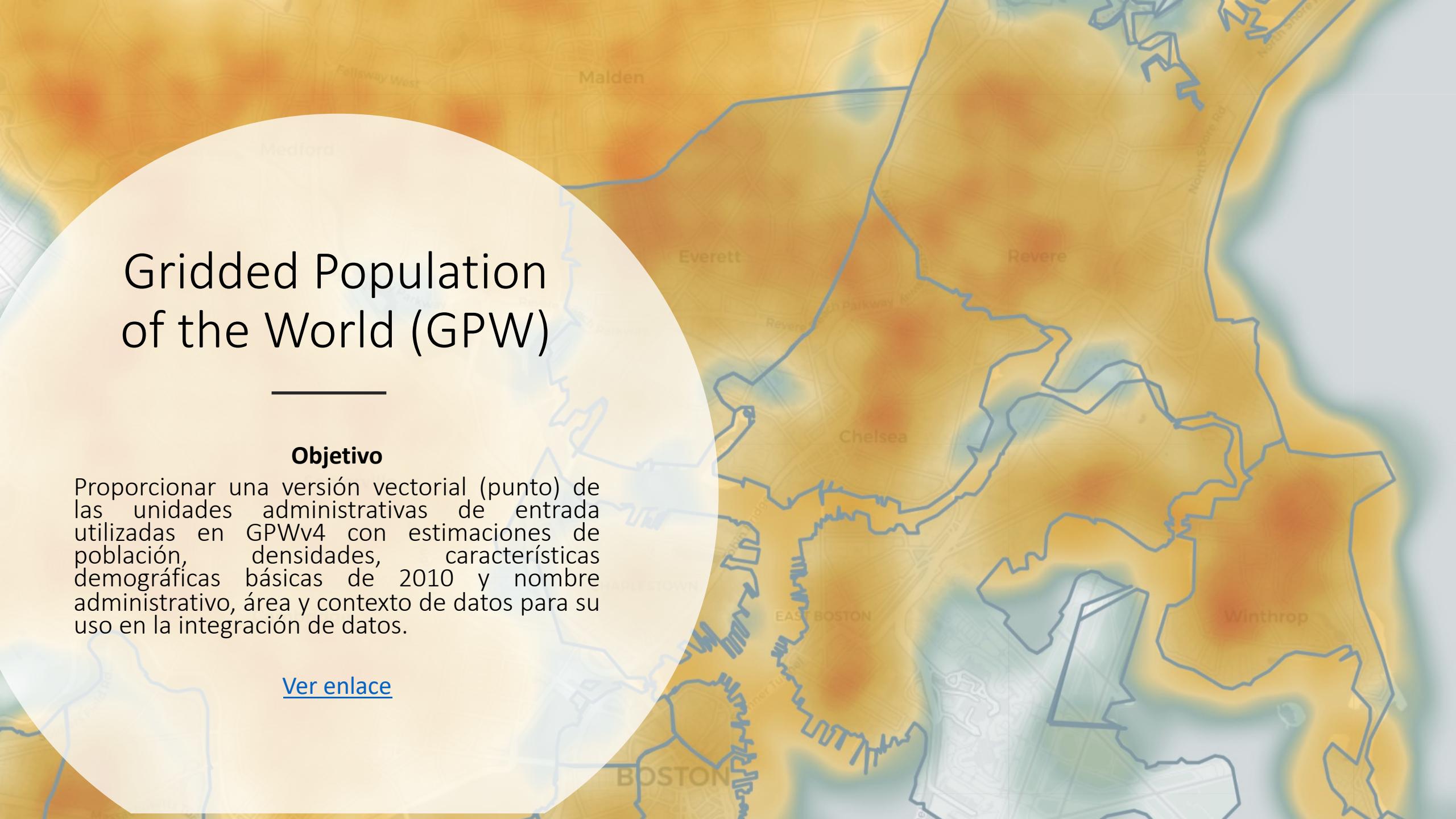
# Gridded Population of the World (GPW)

---

## Objetivo

Proporcionar una versión vectorial (punto) de las unidades administrativas de entrada utilizadas en GPWv4 con estimaciones de población, densidades, características demográficas básicas de 2010 y nombre administrativo, área y contexto de datos para su uso en la integración de datos.

[Ver enlace](#)



# Google™ Transit

## GTFS

### Objetivo

El formato común para los horarios de transporte público y la información geográfica relacionada. Los "Feeds" GTFS permiten que las empresas de transporte público publiquen sus datos y que los desarrolladores programen aplicaciones que consuman esos datos de transporte público de manera interoperable.

[Descripción](#)

[Referencias](#)

[Tiempo real](#)

[GTFS por país](#)

[Colombia](#)



**Introduction to geospatial data**

**Finding and using spatial data**

**Deciding on integration**

**Producing a new data product**

**The Geosocial work package**

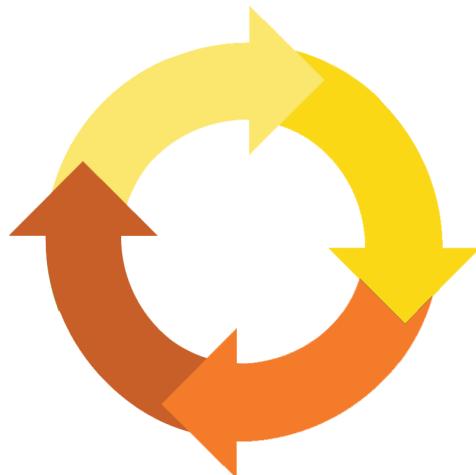
# Deciding on integration:

**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.

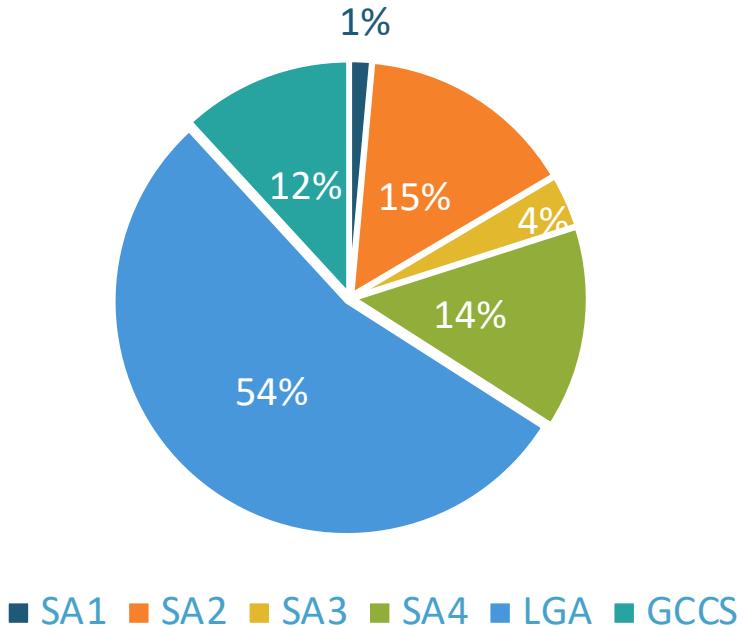
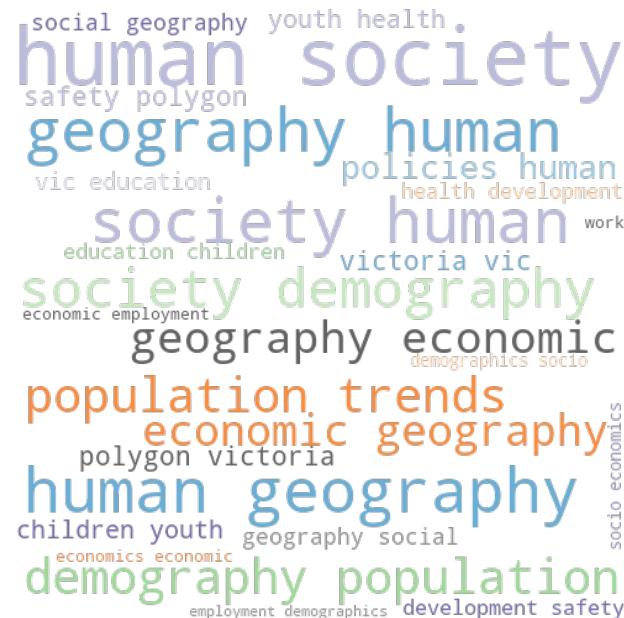
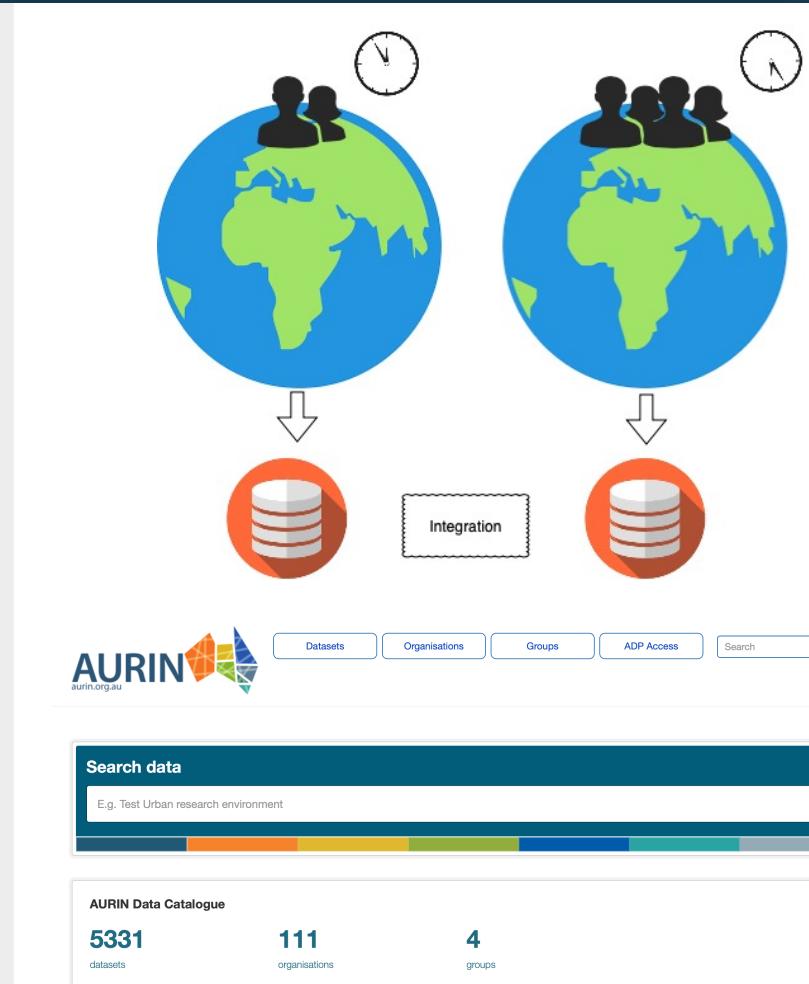
**Database:** Socio-Economic Indexes for Areas (SEIFA) – SA2 - 2016

**Variable:**

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

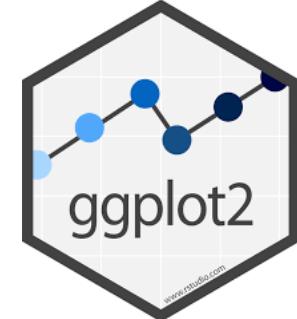
**Step 4:** Try other ways to estimate spatial correlation: Geary's C and Getis-Ord's index

# 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



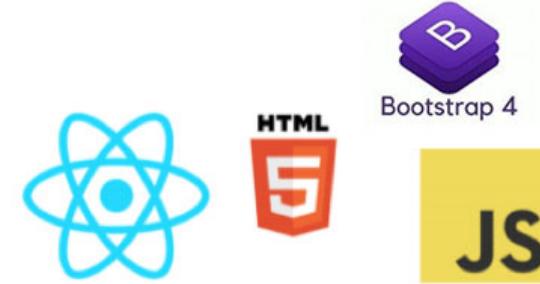
Image from: "Product Data and Your Digital Catalog" by Handshake

# Producing a new data product



```
def add5(x):
    return x+5

def dotwrite(ast):
    nodename = getNodename()
    label=symbol.sym_name.get(int(ast[0]),ast[0])
    print '%s %s %s' % (nodename, label),
    if isinstance(ast[1], str):
        if ast[1].strip():
            print '%s';% ast[1]
        else:
            print ''
    else:
        print ']';
        children = []
        for n, child in enumerate(ast[1:]):
            children.append(dotwrite(child))
        print '%s -> [%s' % nodename,
        for name in children:
            print '%s' % name,
```



```
1 <!DOCTYPE html>
2 <html>
3     <head>
4         <title>Example</title>
5         <link rel="stylesheet" href="styl
6     </head>
7     <body>
8         <h1>
9             <a href="/">Header</a>
10        </h1>
11        <nav>
12            <a href="one/">One</a>
13            <a href="two/">Two</a>
14            <a href="three/">Three</a>
15        </nav>
```

## Examples:

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

# Producing a new data product

## CSS

Lenguaje de hojas de estilo en cascada (Cascading Style Sheets) se usa para estilizar elementos escritos en un lenguaje de marcado como HTML. CSS separa el contenido de la representación visual del sitio.



# Producing a new data product

## Bootstrap 4

Framework CSS desarrollado inicialmente (en el año 2011) por Twitter que permite dar forma a un sitio web mediante librerías CSS que incluyen tipografías, botones, cuadros, menús y otros elementos que pueden ser utilizados en cualquier sitio web. Bootstrap fue liberado bajo licencia MIT en el año 2011 y su desarrollo continua en un repositorio de GitHub.

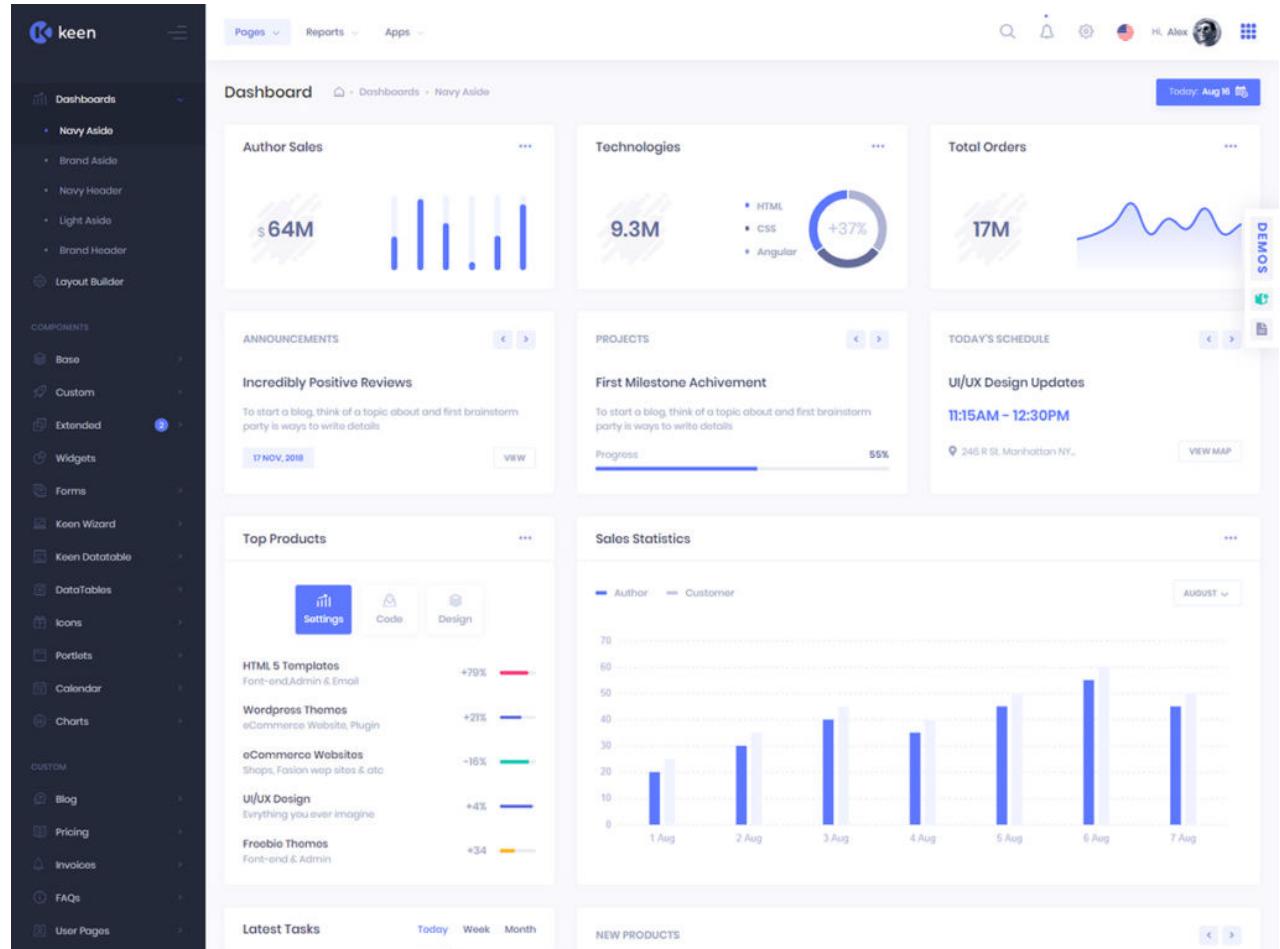


Image from: "Product Data and Your Digital Catalog" by Handshake

## Javascript

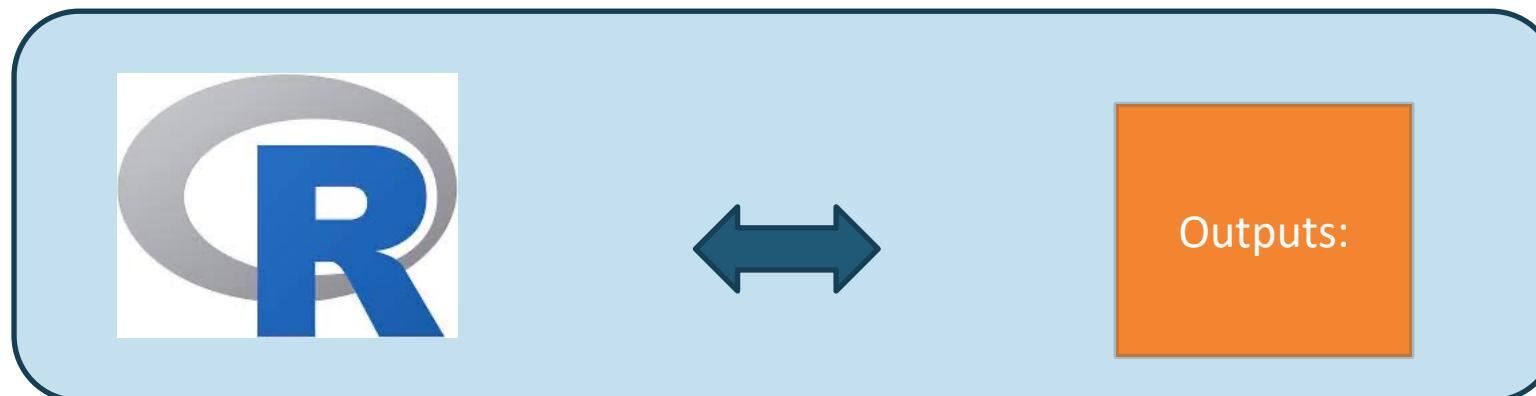
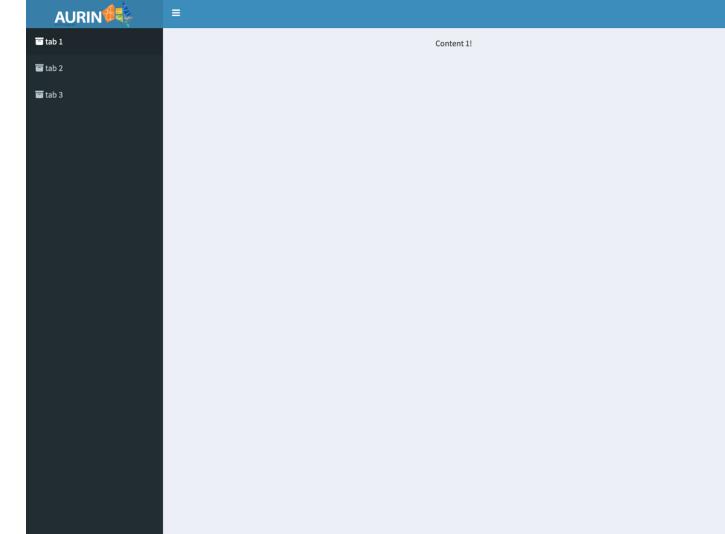
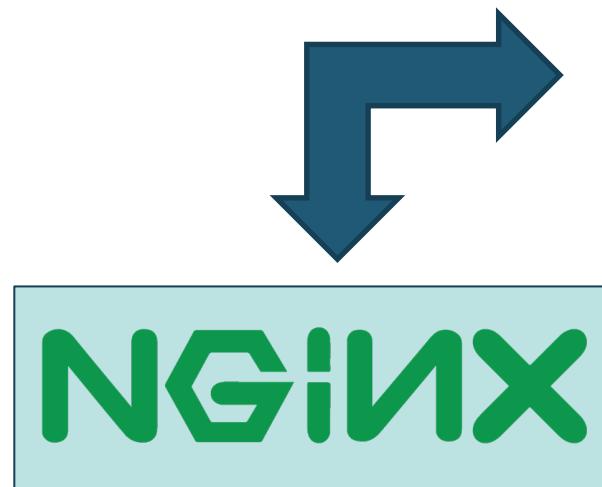
JavaScript es un lenguaje sencillo destinado a añadir algunas características interactivas a las páginas web.

```
class Car {  
  constructor(doors, engine, color) {  
    this.doors = doors;  
    this.engine = engine;  
    this.color = color;  
  }  
  
  carStats() {  
    return `This car has ${this.doors} do
```



# Producing a new data product

## Arquitectura Frontend: Dash



**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