Introducing the IRISS Geosocial Data Integration Service: Integrating information on people, places, time, and space.

Australian National University

ADA AUSTRALIAN DATA ARCHIVE

ACSPRI Australian Consortium for Social & Political Research Inc.

AURIN aurin.org.au

THE UNIVERSITY OF MELBOURNE

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

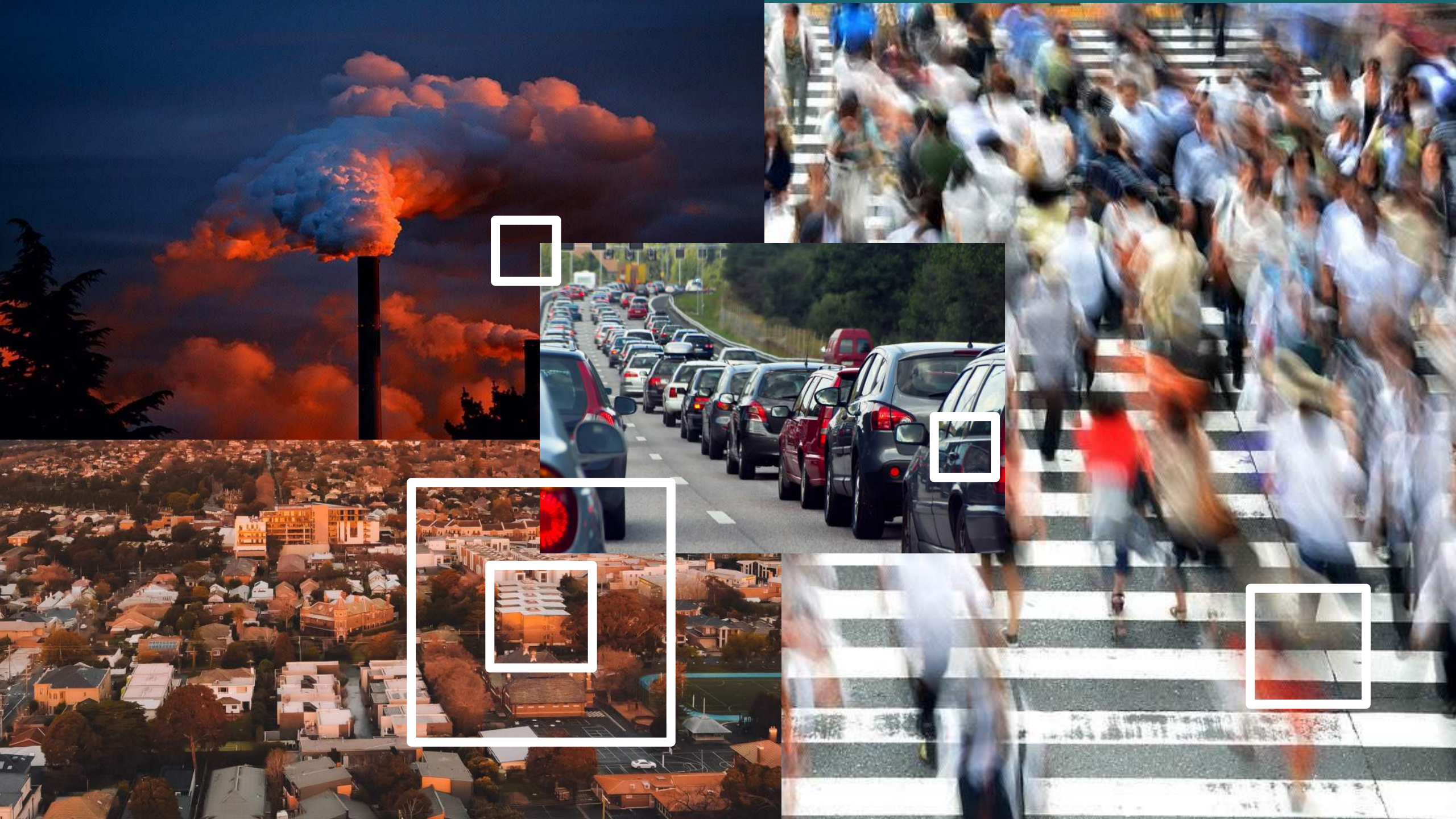
ARDC Australian Research Data Commons

Integrated Research Infrastructure for Social Science (IRISS)

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
- Motivation
- Spatial data and data integration
- Service design
- Demonstrator



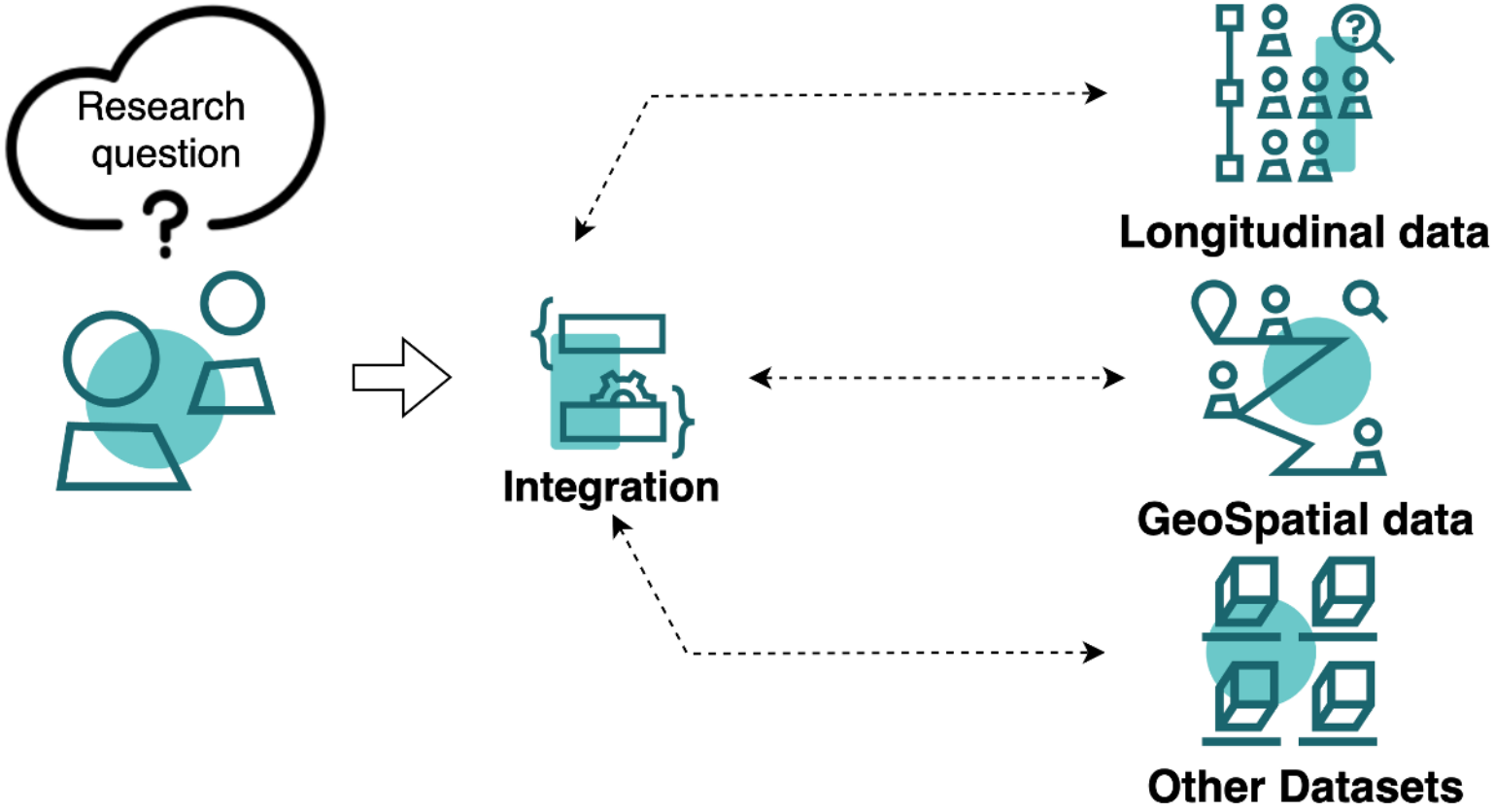
The Integrated Research Infrastructure for the Social Sciences (IRISS)

Objective: Address 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.

- **WP1:** Project Management
- **WP2:** VASSSAL (Vocabulary Access Service for Social Science in Australia)
- **WP3:** GeoSocial
- **WP4:** Demonstrator Projects
- **WP5:** SPIRE (Survey Project Integrated Research Environment)
- **WP6:** CARDSS (Curation of Australian Research Data in the Social Sciences)

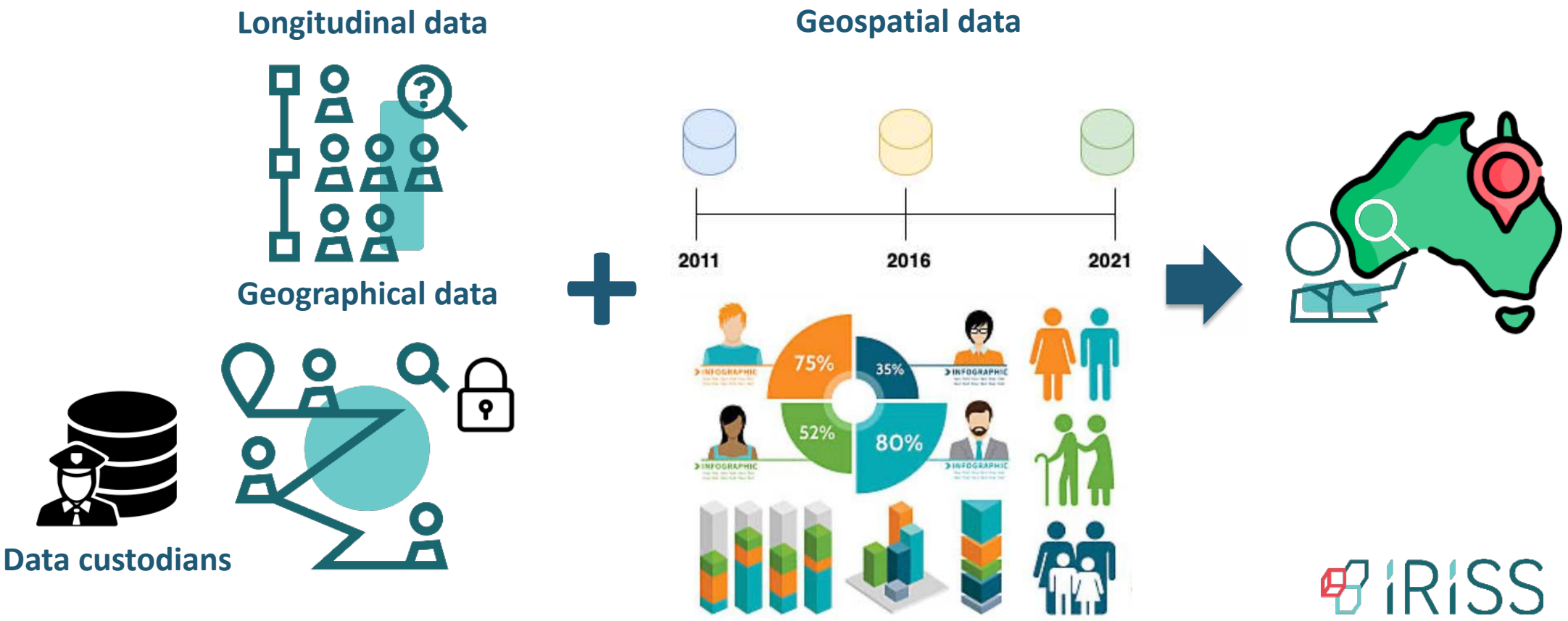
GeoSocial: Problem

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.



Motivation

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.



Opportunities and barriers

Current landscape:

- Fragmented data
- Lacking ‘good’ documentation (particularly concerning technical data/issues)
- (Spatial) data integration demands deep technical and methodological knowledge

Consequences:

- Duplication
- Lack of consistency given the need for individualised approaches; lack of reproducibility/scalability
- Time-consuming

Personas/users

Low skills level

- Easy access to the data.
- Certainty regarding data meanings.
- Less room for analytic errors.
- Increased data usability and utility to untrained users.
- Reduction of the risk of data breaches.

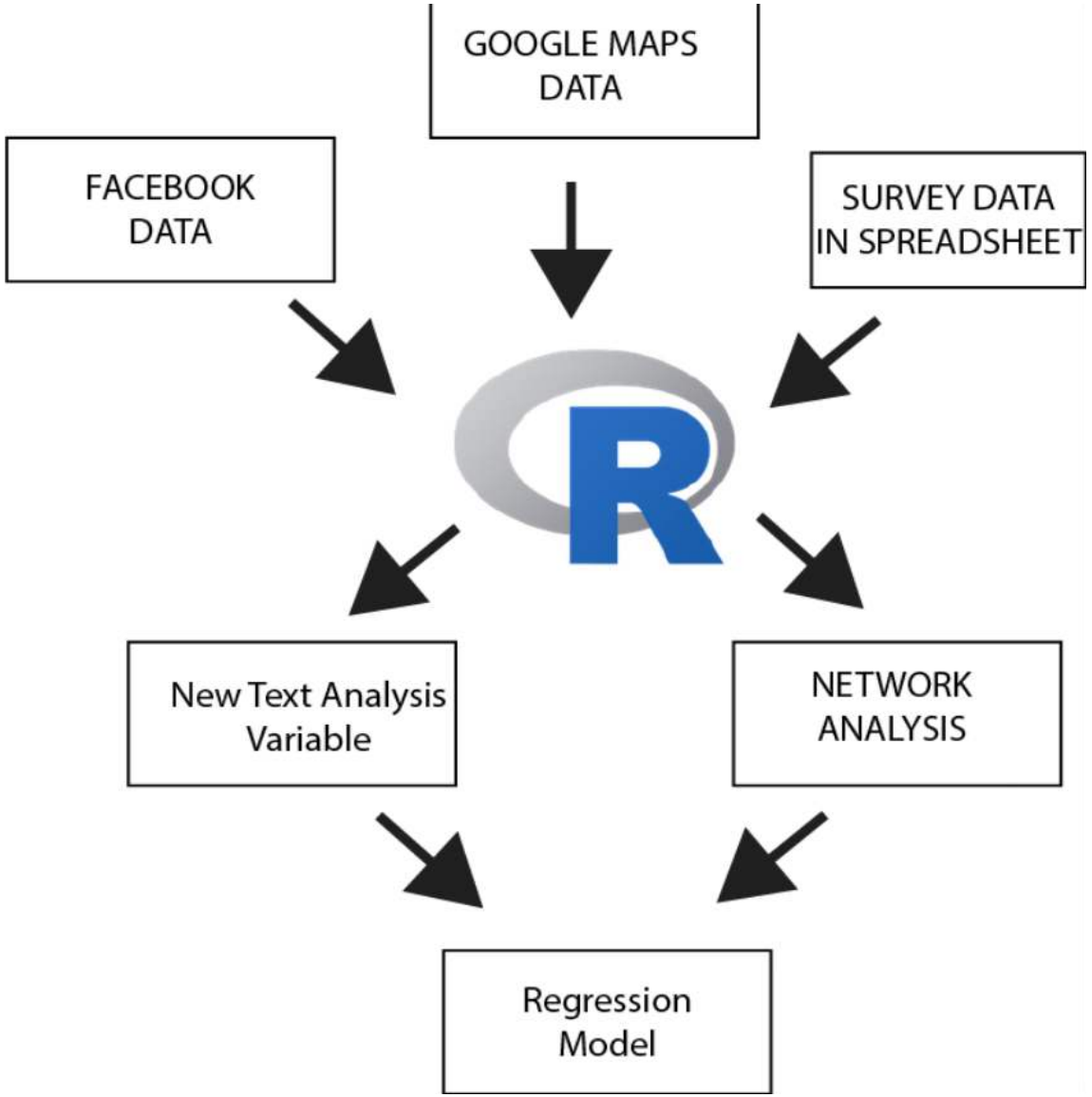
Mid-level user

- Confident with understanding and tweaking R scripts
- Experienced in the use of Stata software
- Limited understanding of geospatial data
- Needs to integrate longitudinal and geospatial data for analysis
- May consult with researchers to achieve goals

Advanced user

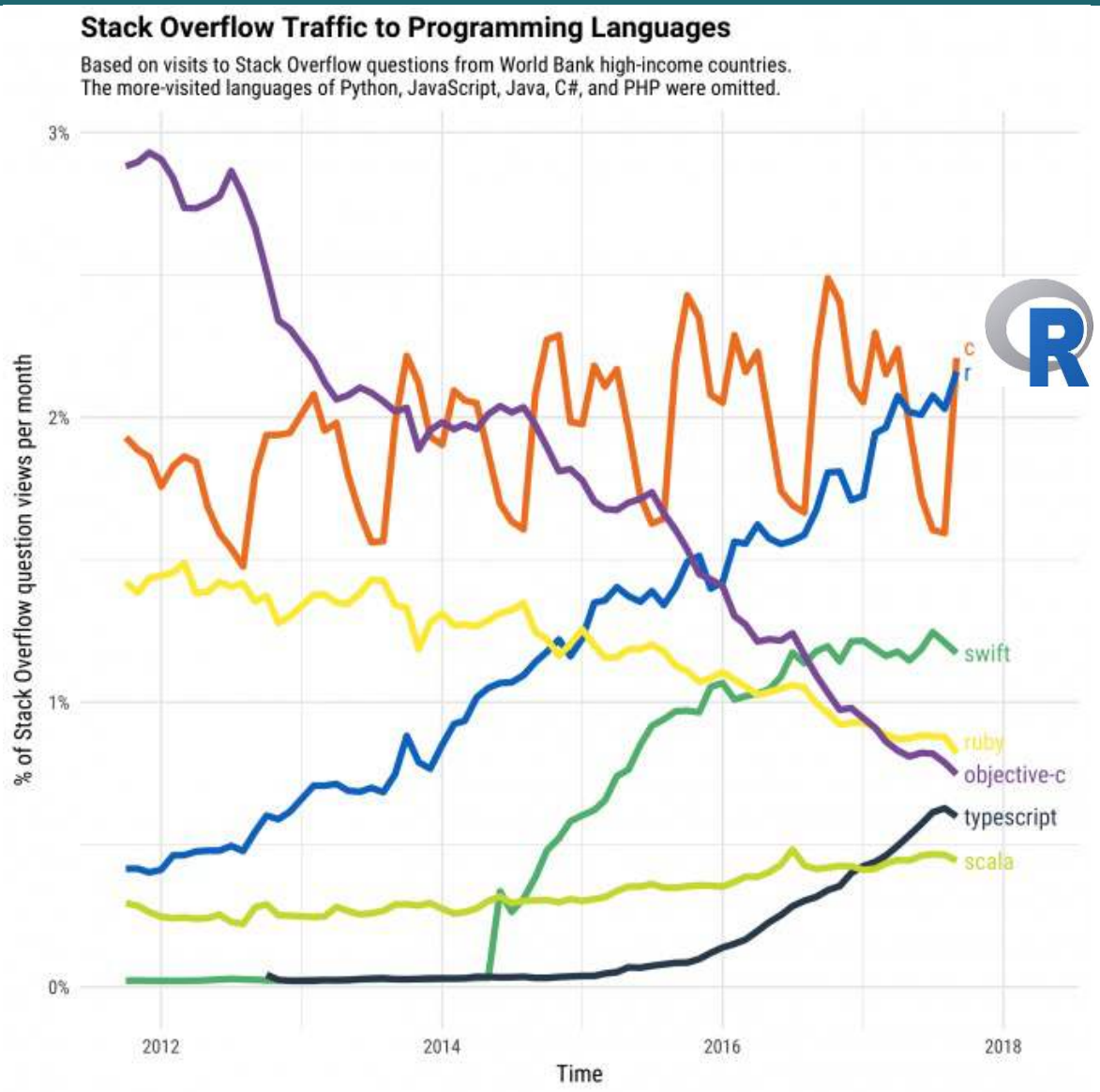
- Confident with using Python and/or R for data wrangling, integration, and analysis
- Good understanding of geospatial data
- Needs to integrate longitudinal and geospatial data for analysis
- Supports other social science researchers

Preferred language



Why R?

- **Installation:** Easy and fast
- **Customizable:** Can be easily tailored to specific needs
- **Community:** Active community provides libraries and modules that are frequently updated and monitored
- **Documentation:** well-documented and strict publishing rules.
- **Cost:** 0$!!!

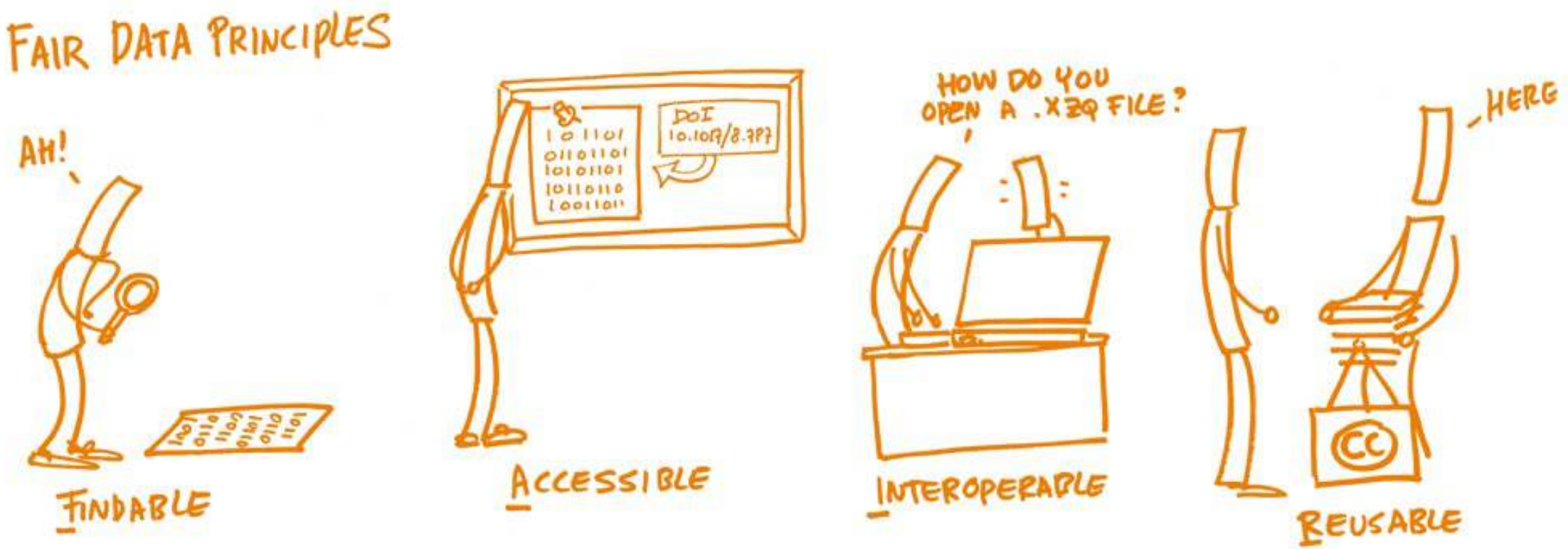


Requirements:

- Accessible and usable for Mid-level users and advanced users.
- Delivered through a code language and executed using a script.
- Clear documentation and examples.
- Follow the standards and procedures established by the data custodians of the longitudinal survey.
- Login-free access allows users to discover the service benefits quickly.

Requirements:

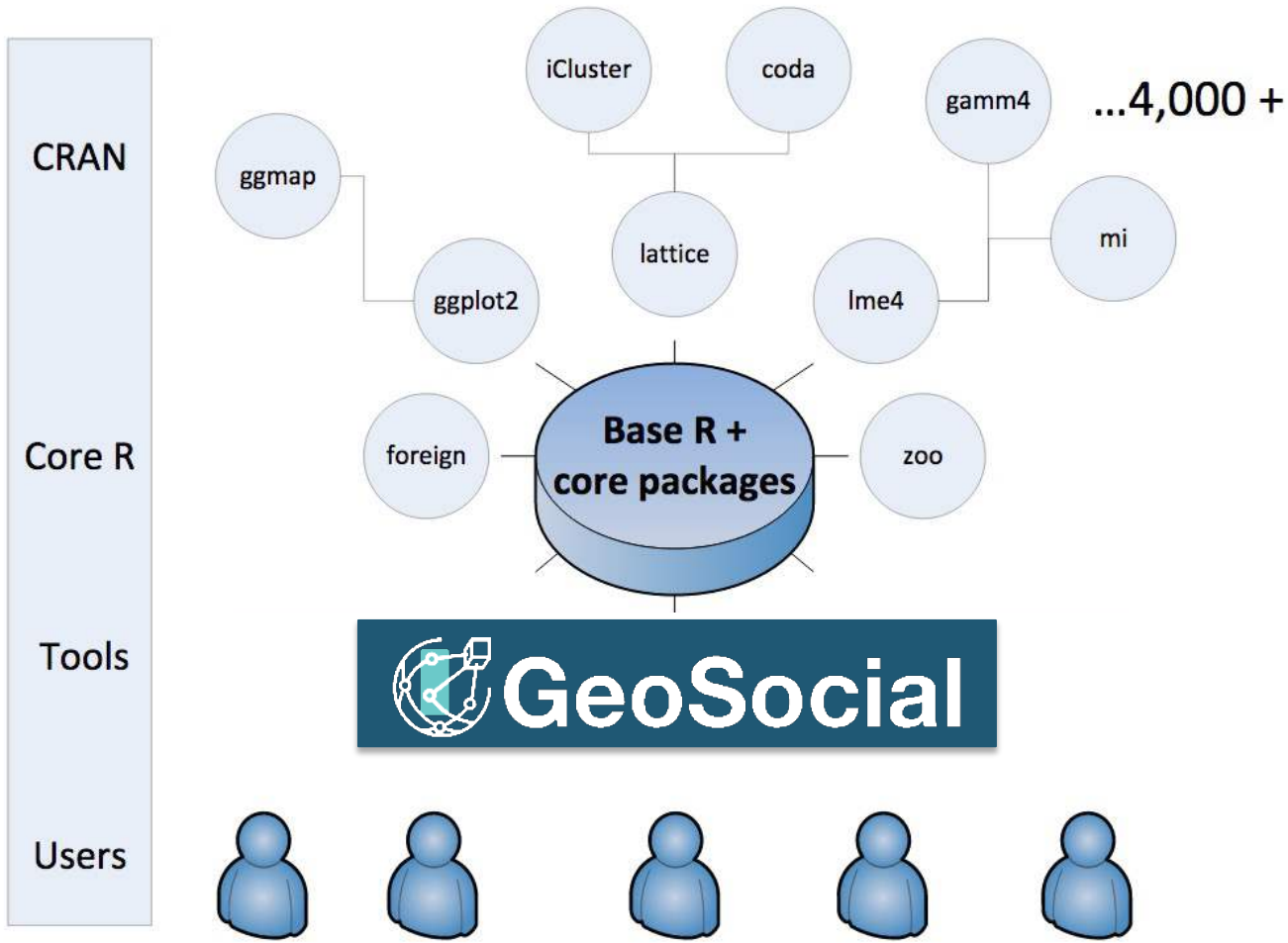
- Design and develop the following FAIR principles for research software



Images from Foster Open Science

R Library

An R library contains code, data, and documentation in a standardised collection format that can be installed by users of R.



Introduction

Motivation

Spatial data and data integration

Service design

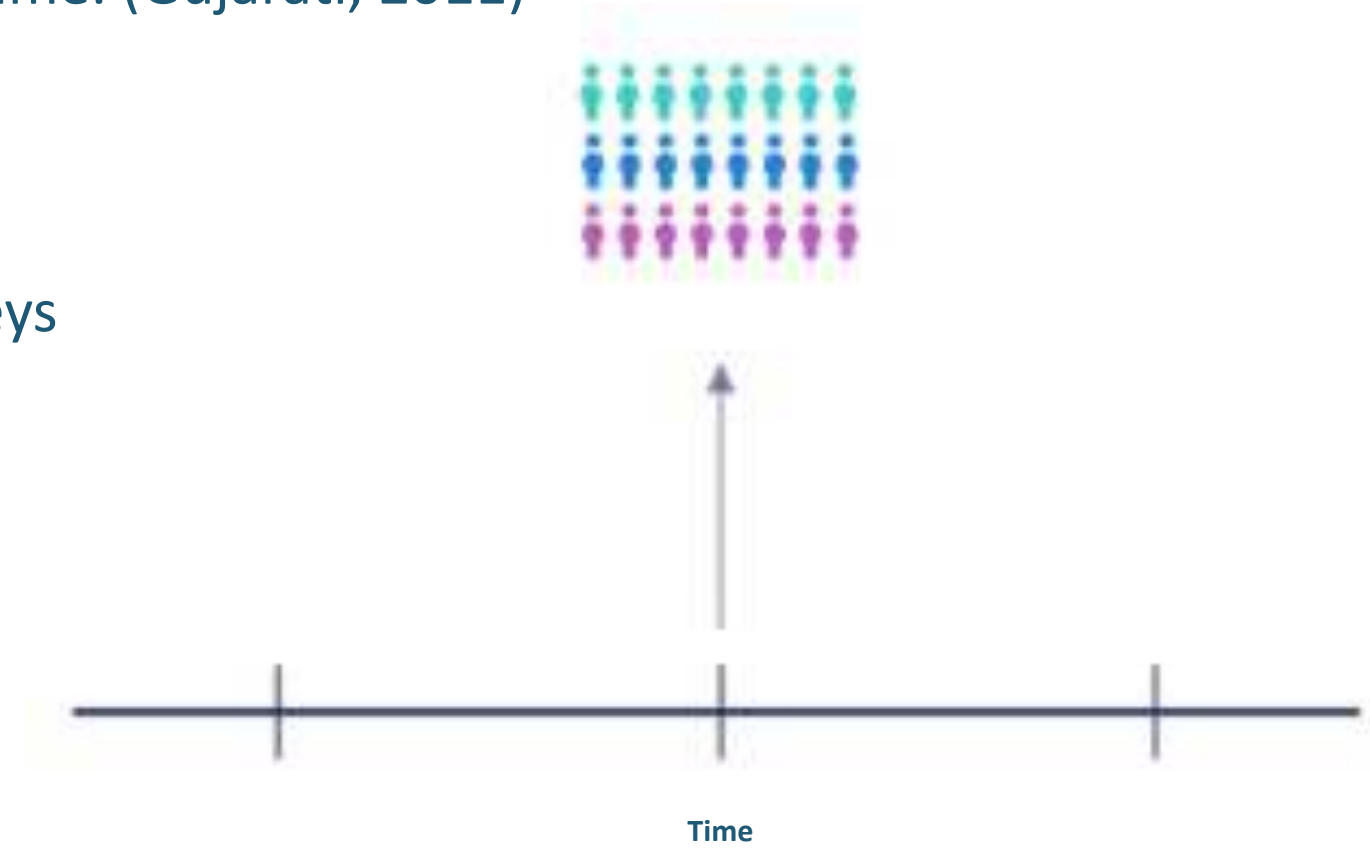
Demonstrator

Cross-sectional data

Cross-sectional: 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

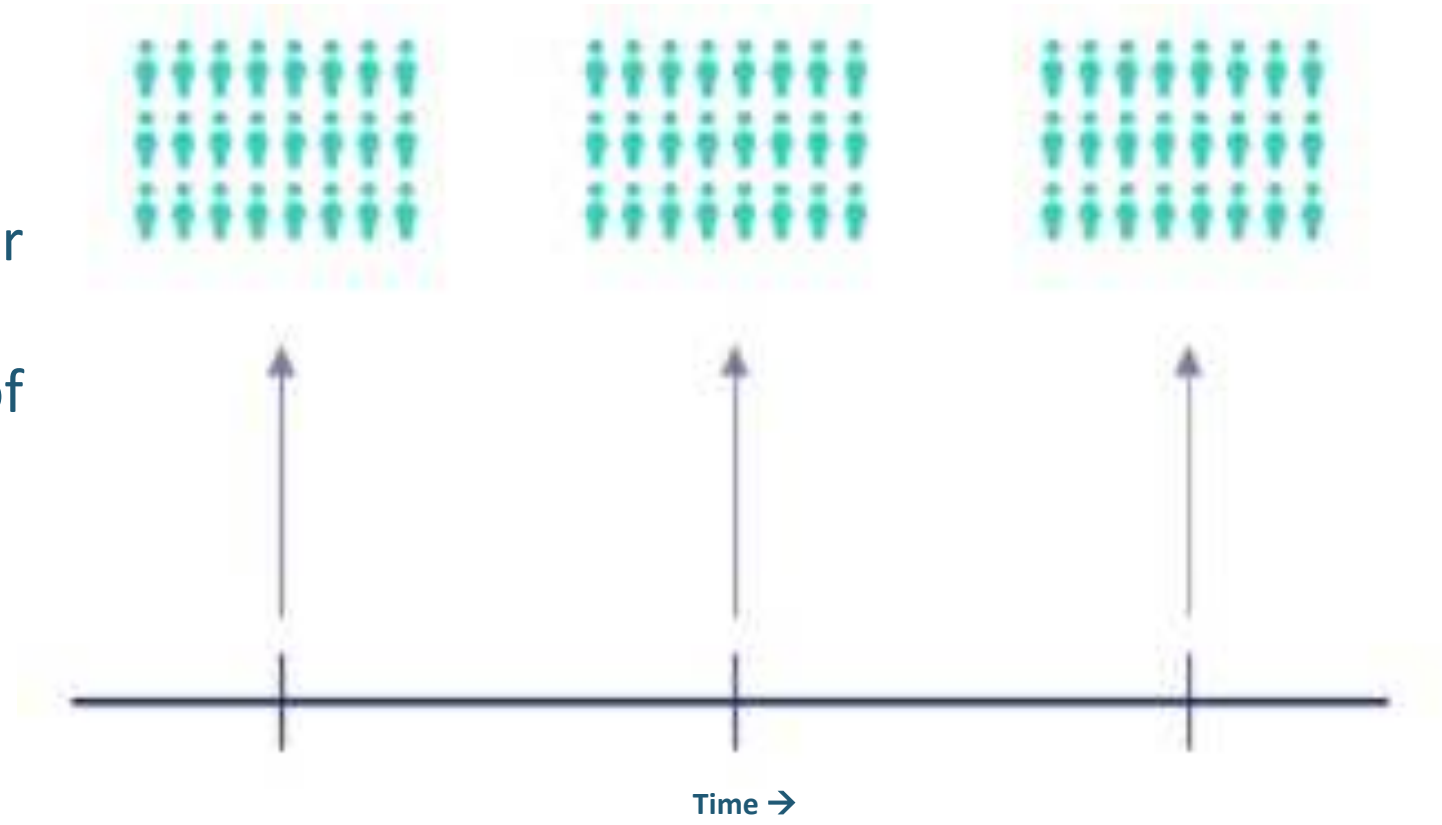


Longitudinal Survey data

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

Examples:

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



Survey data

Cross-sectional

- Conducted at a given point in time.
- Samples are generated randomly.
- It is difficult for studies to establish a cause-and-effect relationship.
- Cross-sectional study is comparatively cheaper.

Longitudinal

- Conducted at various points in time
- High attrition rates
- It can be used to study cause-and-effect relationships.
- Minimize the random effects and their associated noise.
- It can take several years and often involves high expenses.

Images from QuestionPro

Longitudinal studies: Australia

AURIN
aurin.org.au

The University Of Queensland
AUSTRALIA

LIVING IN AUSTRALIA
HILDA
Household, Income and Labour Dynamics in Australia

AES AUSTRALIAN ELECTION STUDY

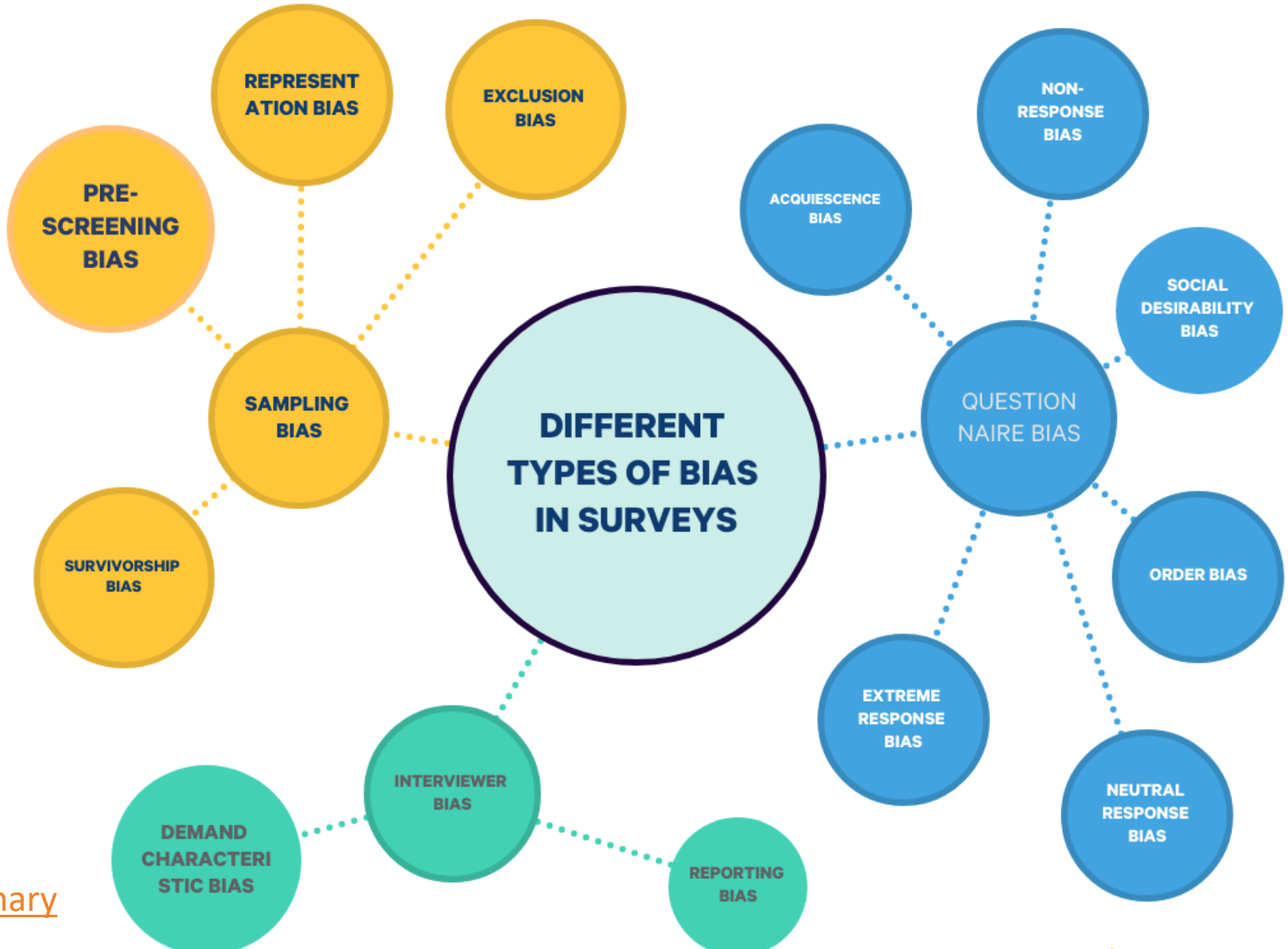
Longitudinal Surveys of Australian Youth

Growing Up in Australia
The Longitudinal Study of Australian Children

AuSSA
The Australian Survey of Social Attitudes

iRiSS

BIAS



Images from Trustmary

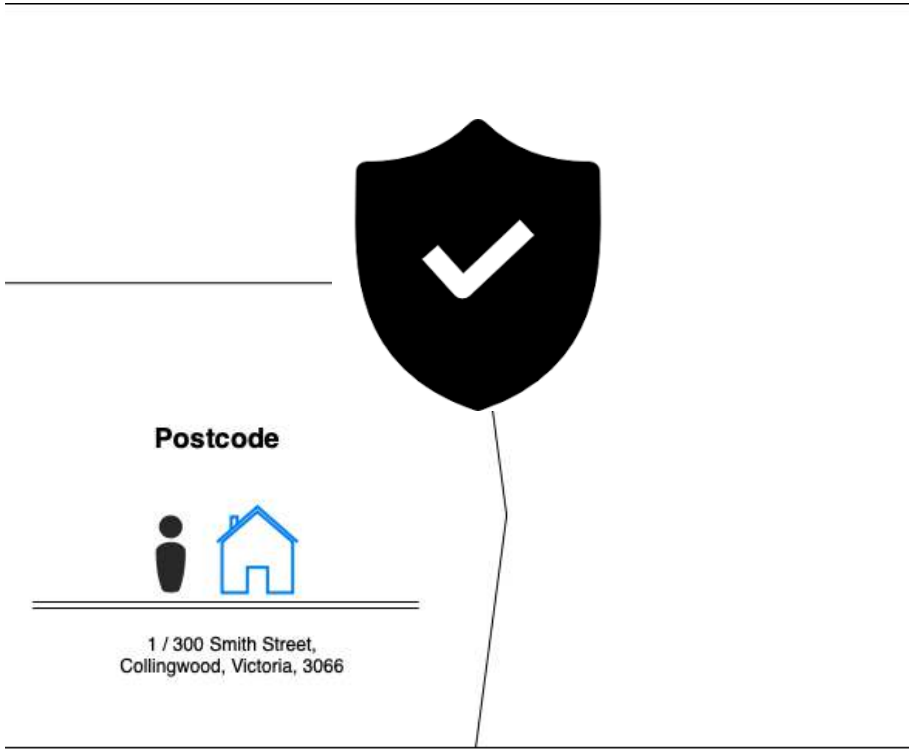
Data custodians

Data custodians established certain conditions and procedures for accessing their resources.

- **Assess project risk:** Mitigate risks in a data integration project, assess the level of risk and specify strategies.
- **Comply with policy and legislation:** Ensure all legislative and ethical obligations are met before releasing data.
- **Enter project agreements:** Enter formal agreements with the nominated integrating authority.

Source: ABS

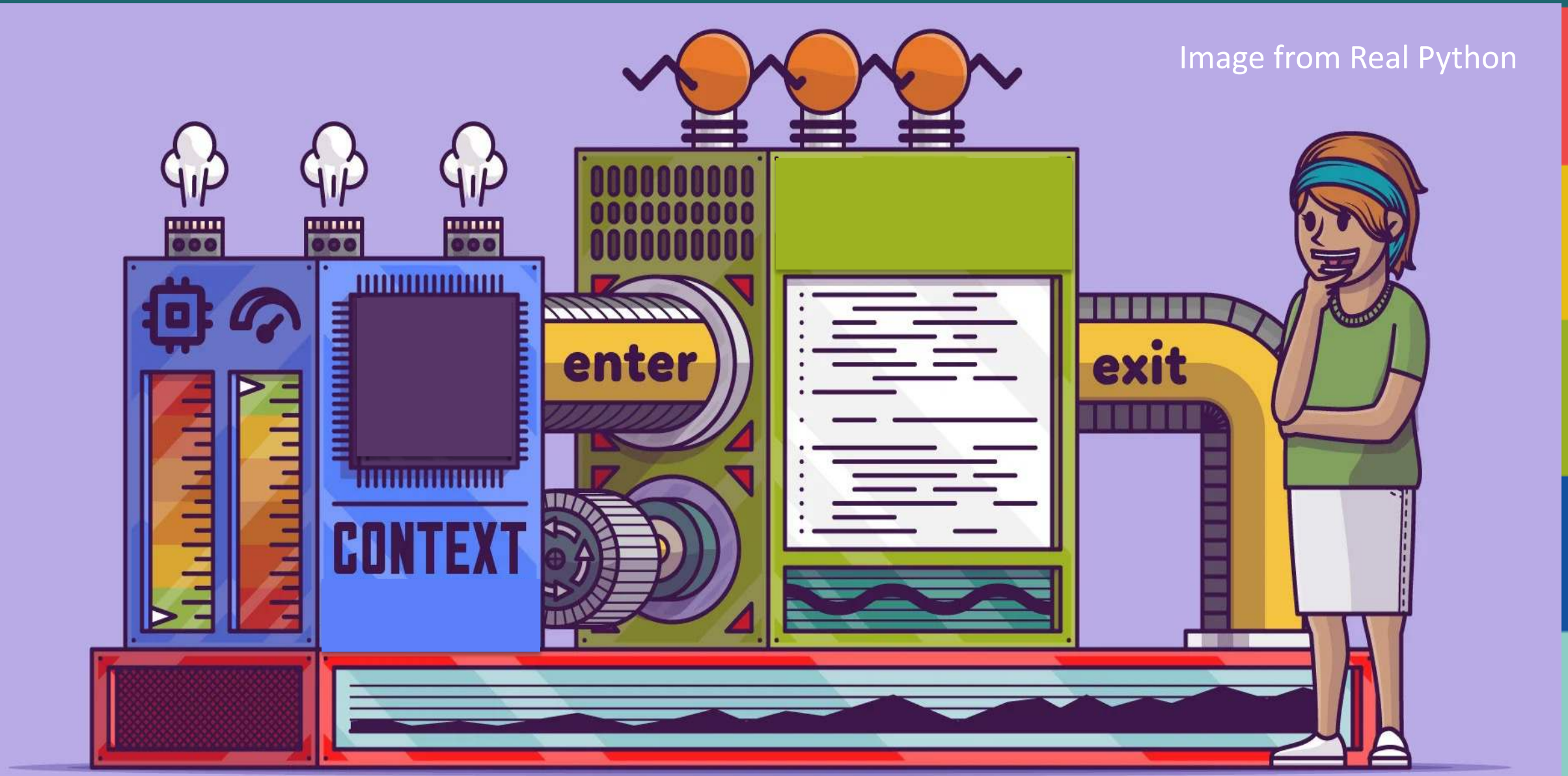
Data custodians



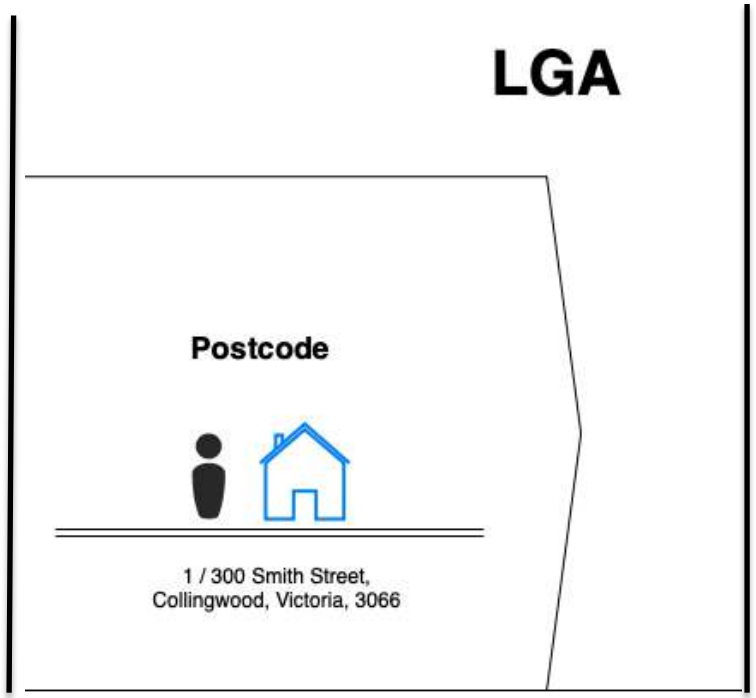
- **Safe storage:** Ensure that the integrating authority can provide safe storage of unit record data under the data custodian's requirements and data storage policies.
- **Safely transmit unit record data:** Ensure the safe transmission of data to integrating authorities, consistent with Australian Privacy Principles and the Australian Government Protective Security Policy Framework.

Source: ABS

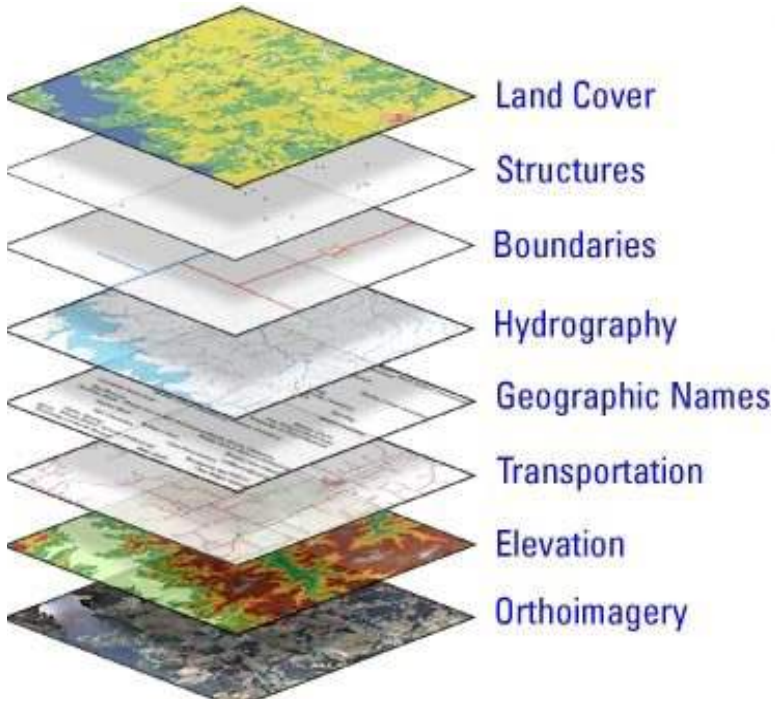
Data Enrichment



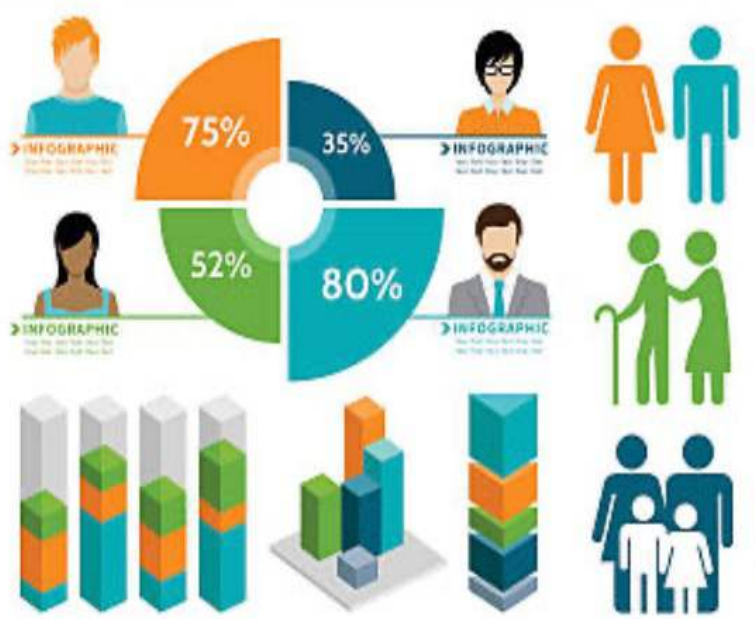
What is spatial data?



Physical



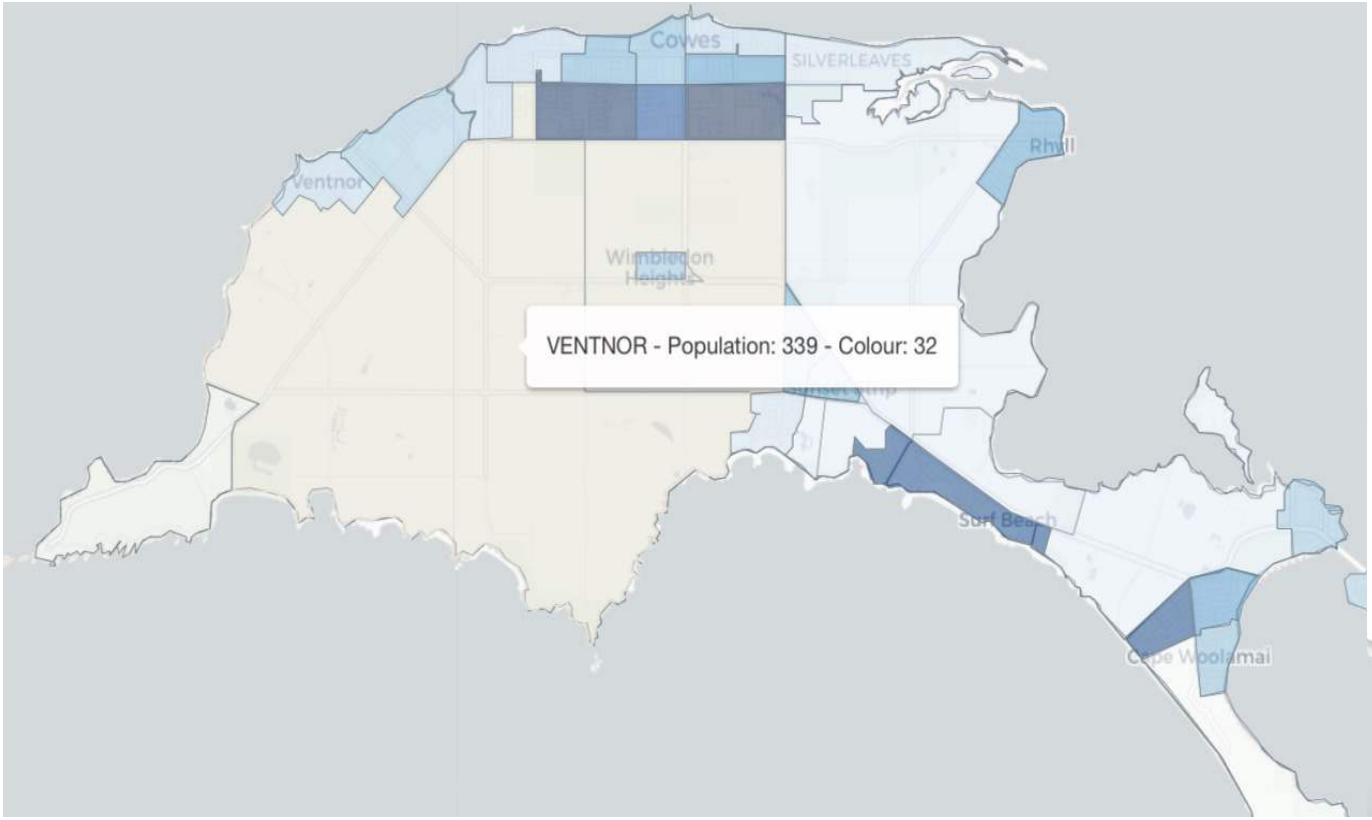
Socio-demographic



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

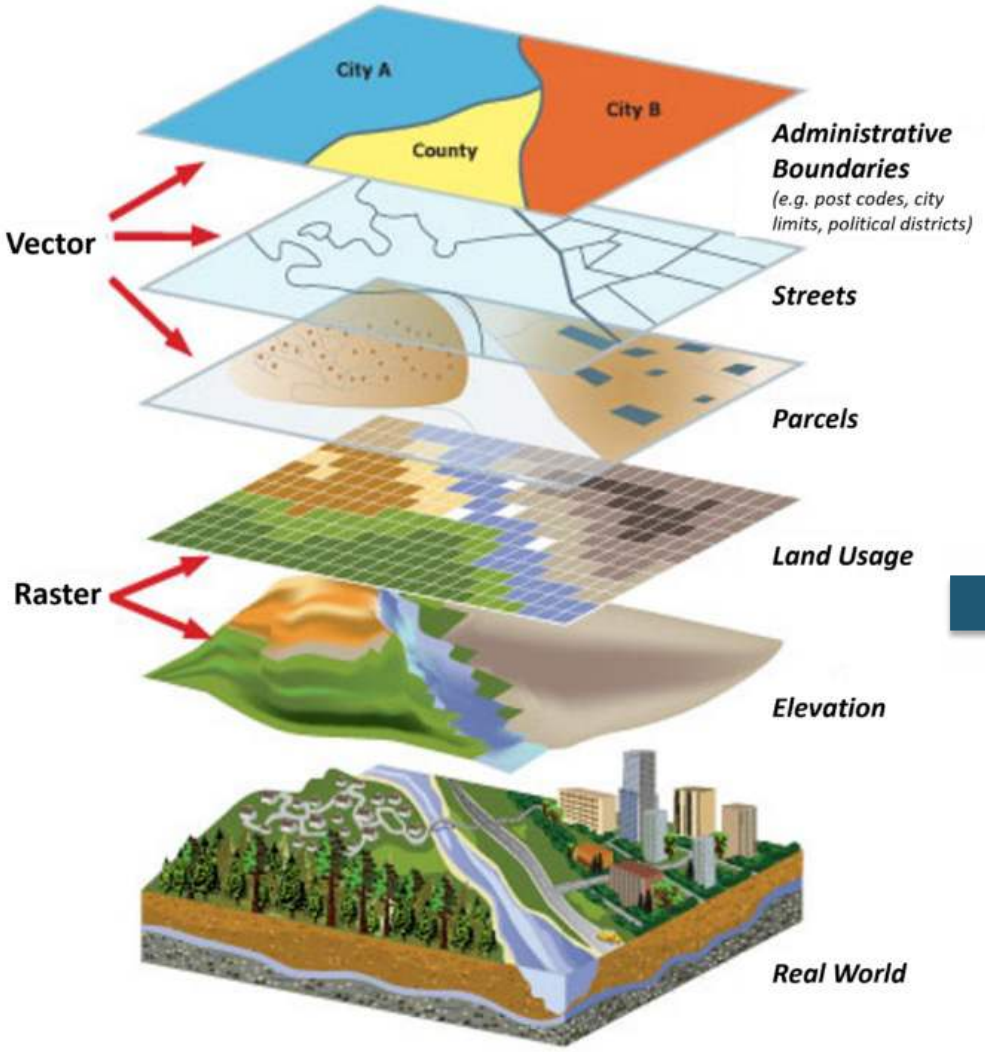
Why use Geospatial data?

Understand the population



Images from: Author

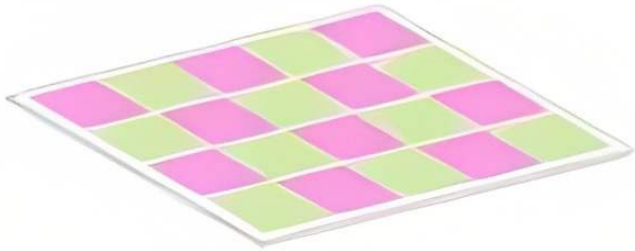
Why use Geospatial data?



"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

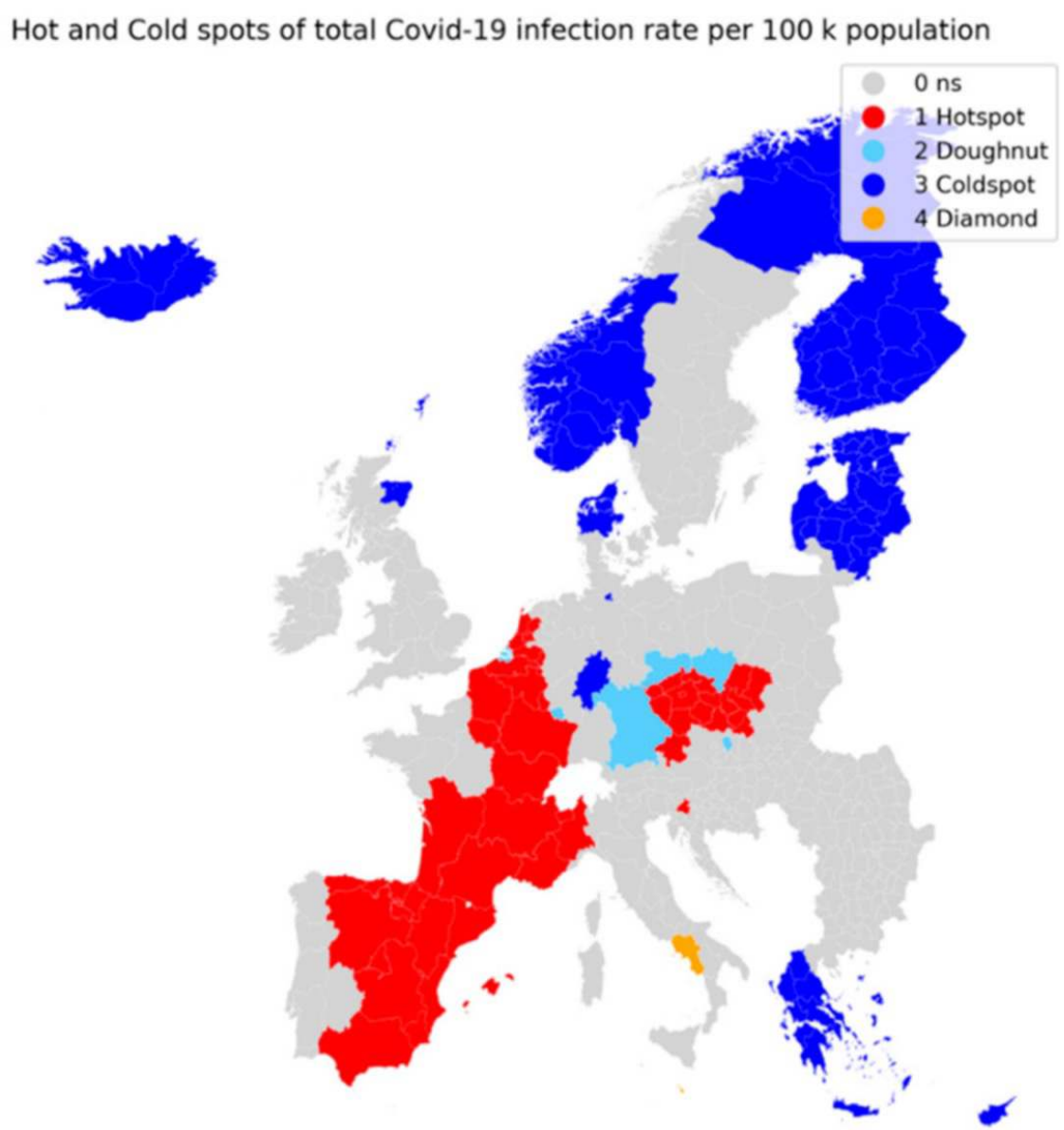
High spatial autocorrelation

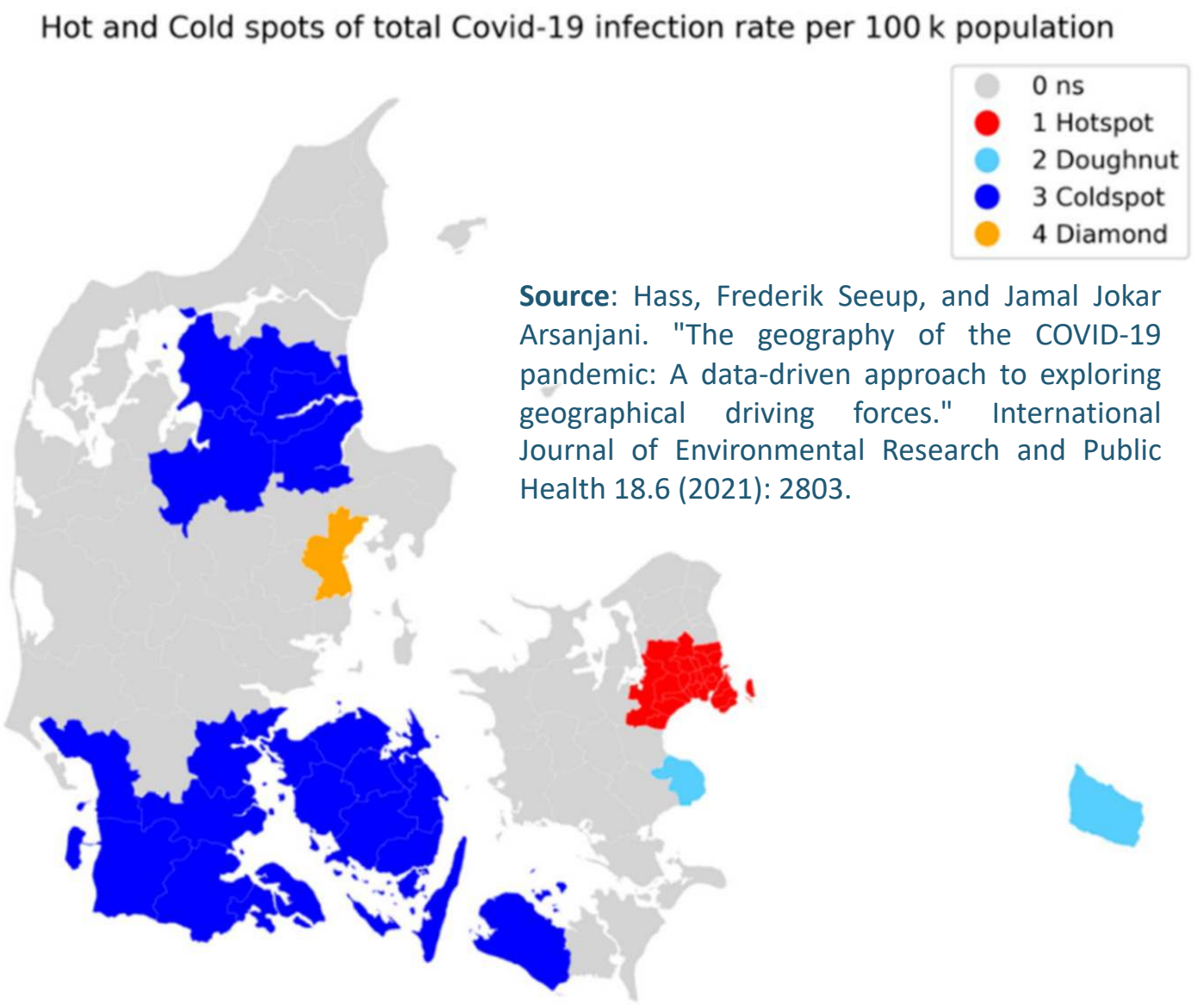
Image from: Domain

Why use Geospatial data?

AURIN
aurin.org.au

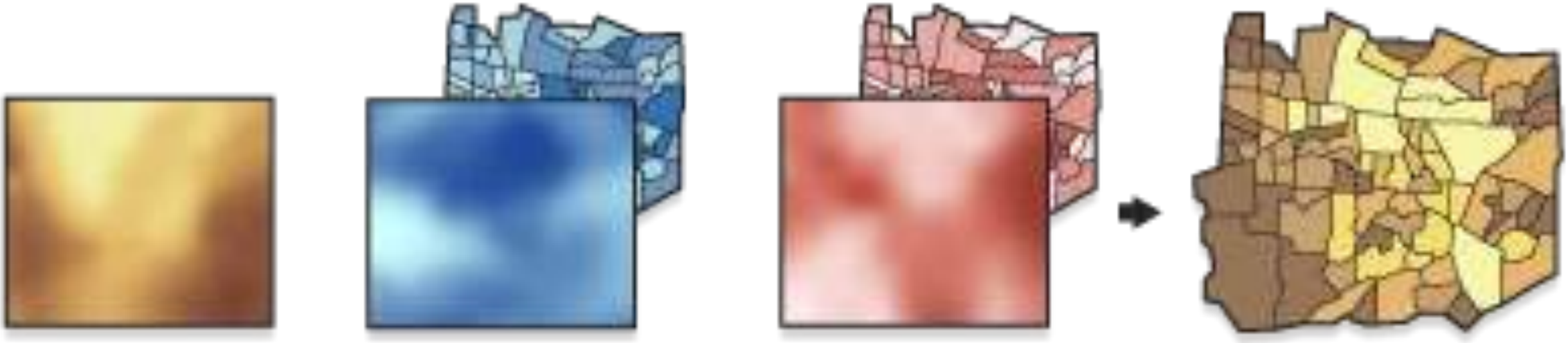
The University Of Queensland
AUSTRALIA





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.

Why use Geospatial data?



Geographically Weighted Regression

Images from: ESRI

Why use Geospatial data?

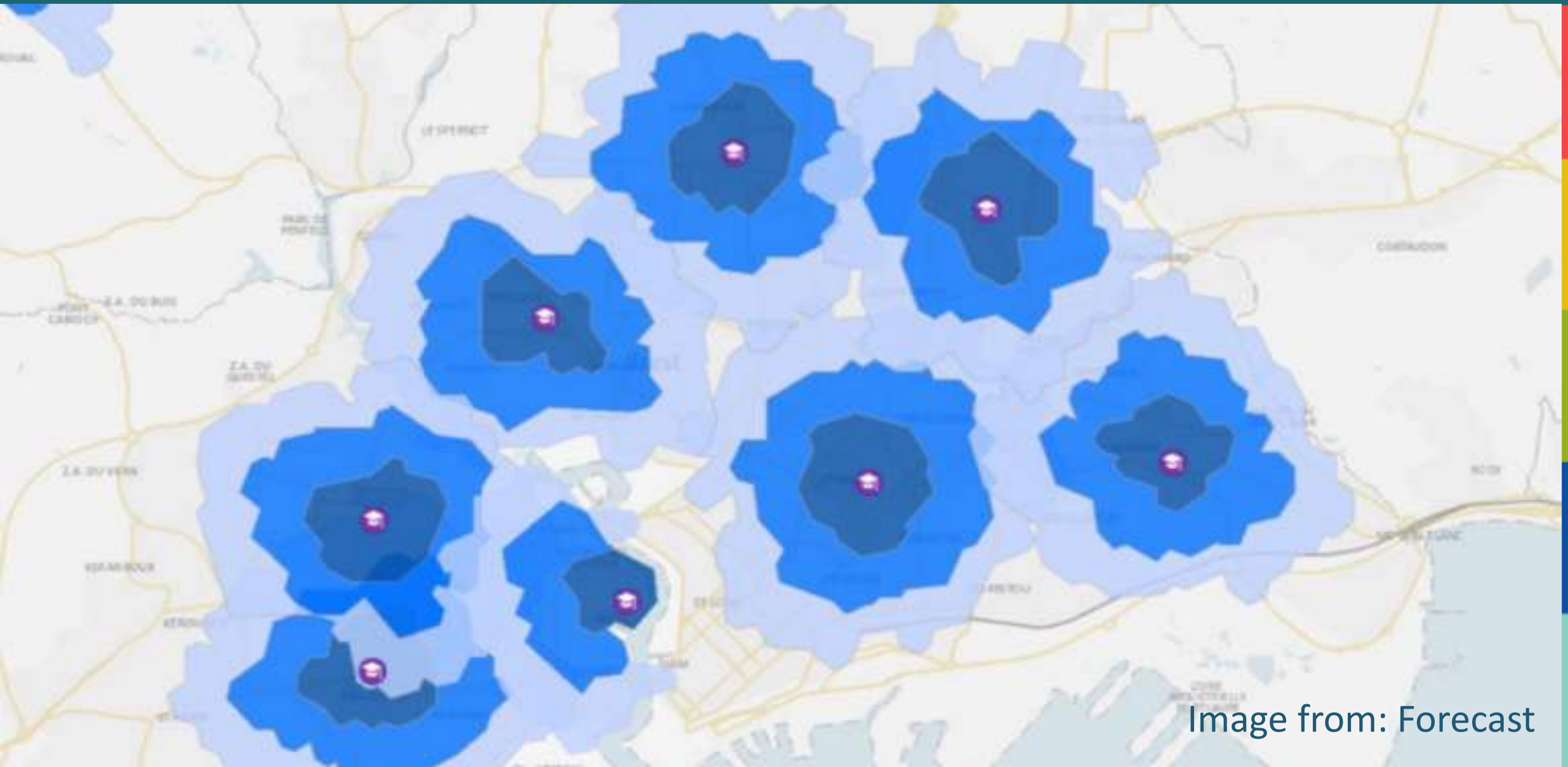


Image from: Forecast

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:

- Empower cross-disciplinary social research using rich integrated geosocial data
- Inform policy and facilitate social planning across wide-ranging issues relating to population health, social well-being and community cohesion

Introduction

Motivation

Spatial data and data integration

Service design

Demonstrator

Spatial analysis

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

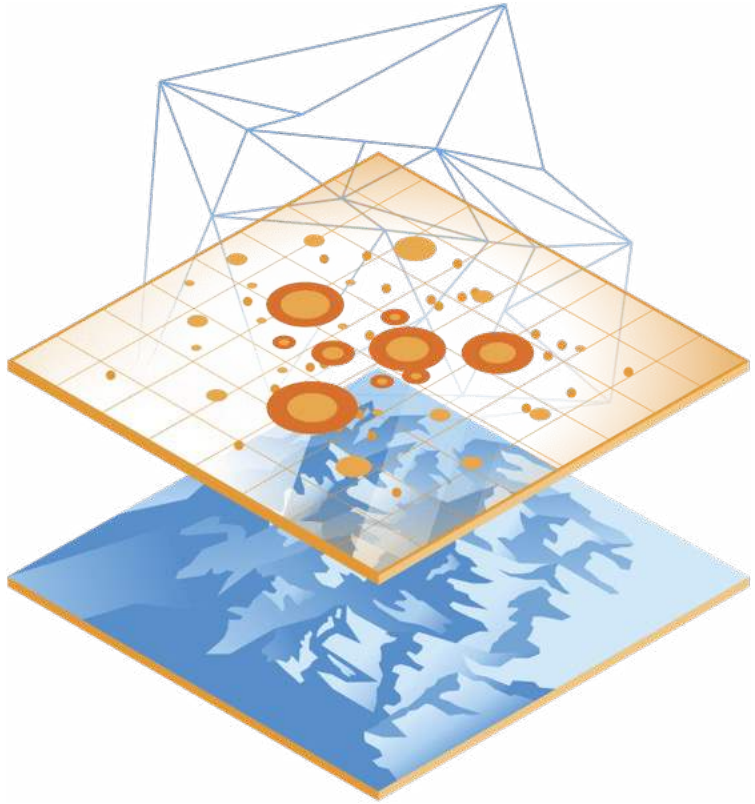
Spatial analysis needs to make assumptions about or draw on data describing spatial relationships or spatial interactions between cases. (Chorley, 1972; Haining 1994).

Geodetic System

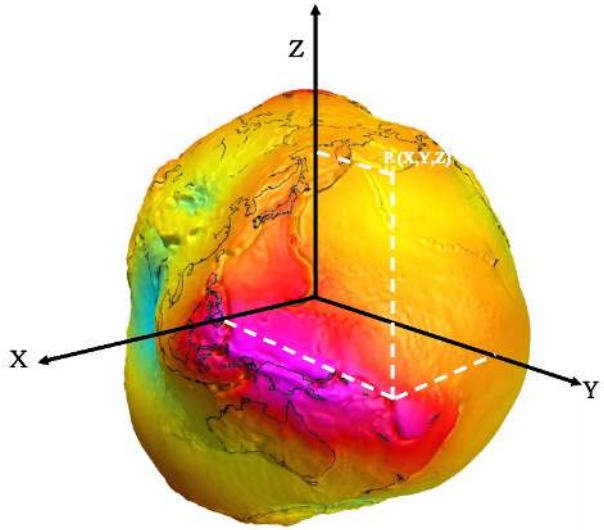
A geodetic reference system is necessary to assign coordinates to points on the Earth's surface.

Each geographic data will have a unique geographic reference associated with it that can help locate precisely where it occurs on a map.

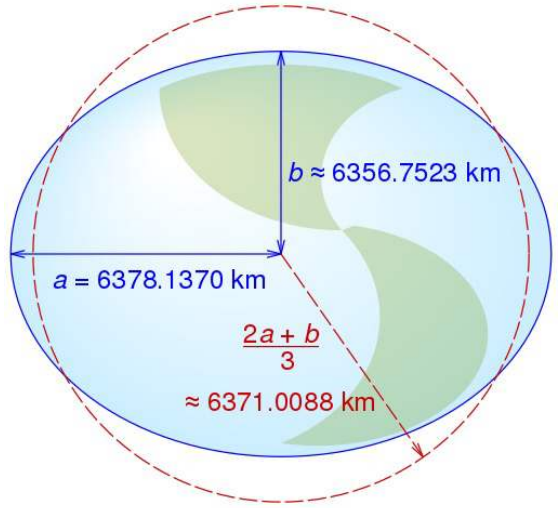
The Geodetic System is an essential tool for accurately mapping and understanding our planet.



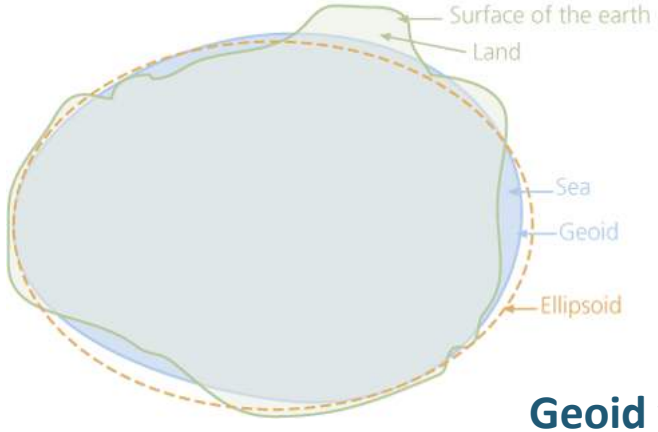
World Geodetic System 1984 (WGS84)

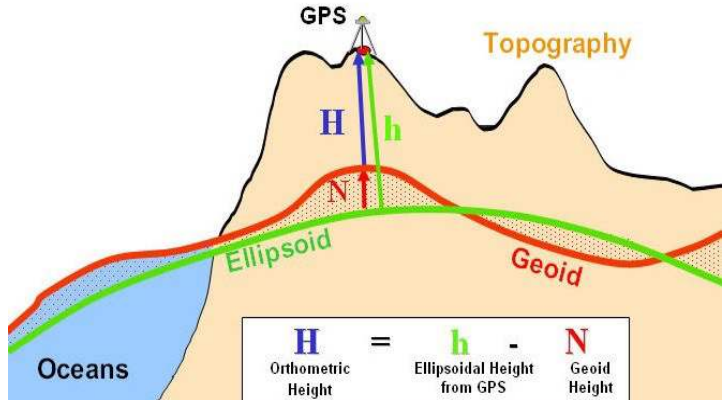


Geoid



Ellipsoid



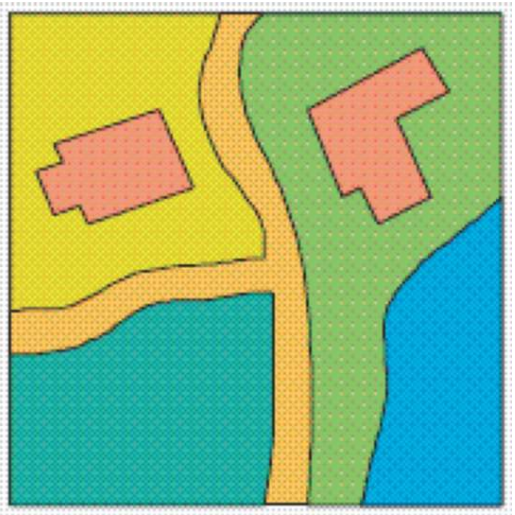


Geoid vs Ellipsoid

Finding and using spatial data

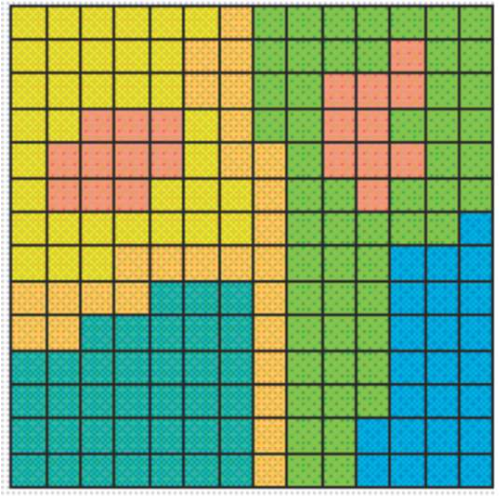
Vector map:

- 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.



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.



Images from: ESRI

Vectorial map:

Type of elements:

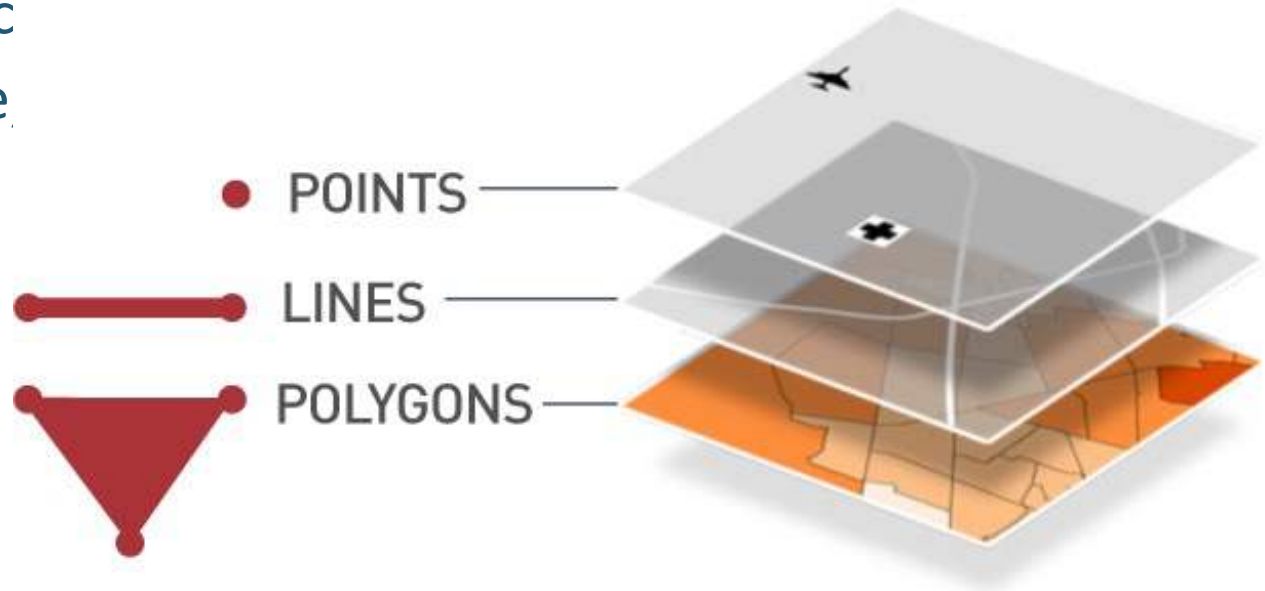
Point: Addresses, locations, points of interest, etc

Lines: streets, freeways, borders, etc

Polygons: Countries, cities, Cadastre

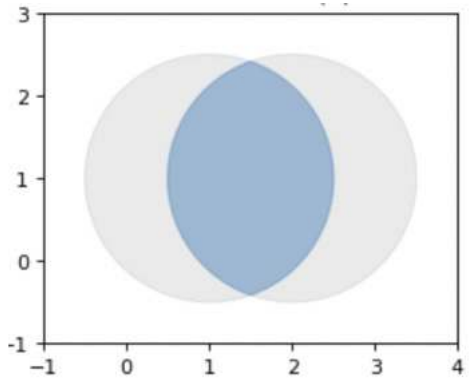
Advantages:

- Spatial operations
- Spatial aggregation
- Spatial Join

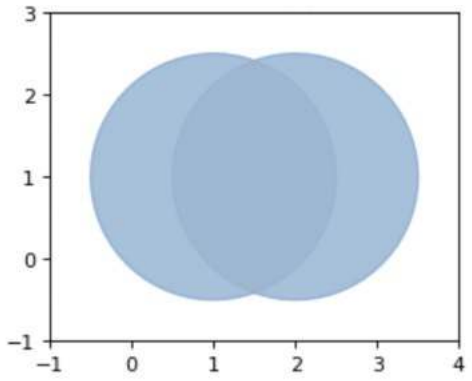


Images from: Berkeley

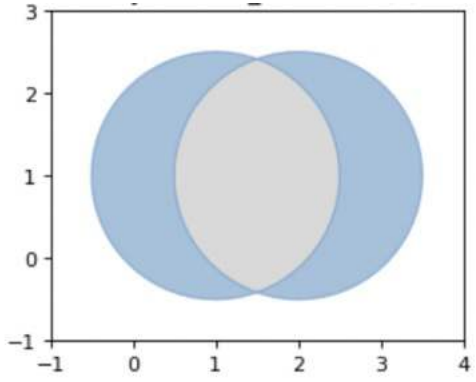
Spatial operations



Intersection

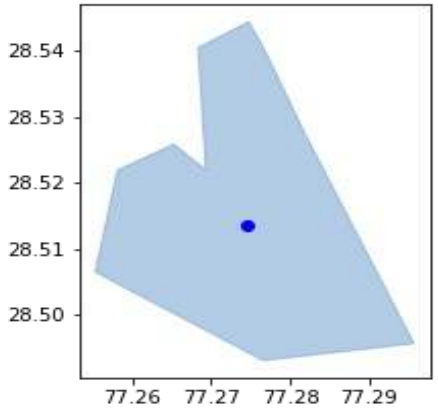


Union

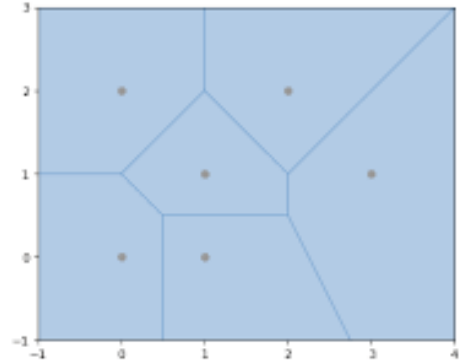


Difference

Contour



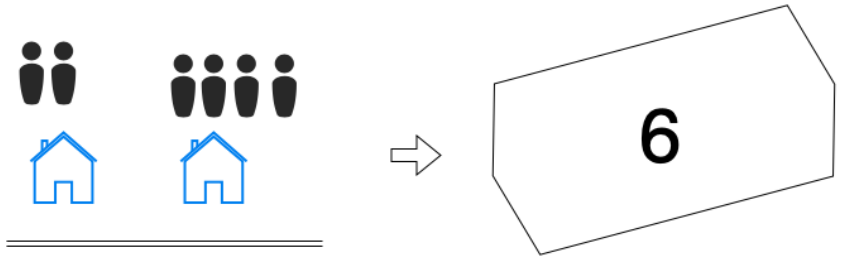
Centroid

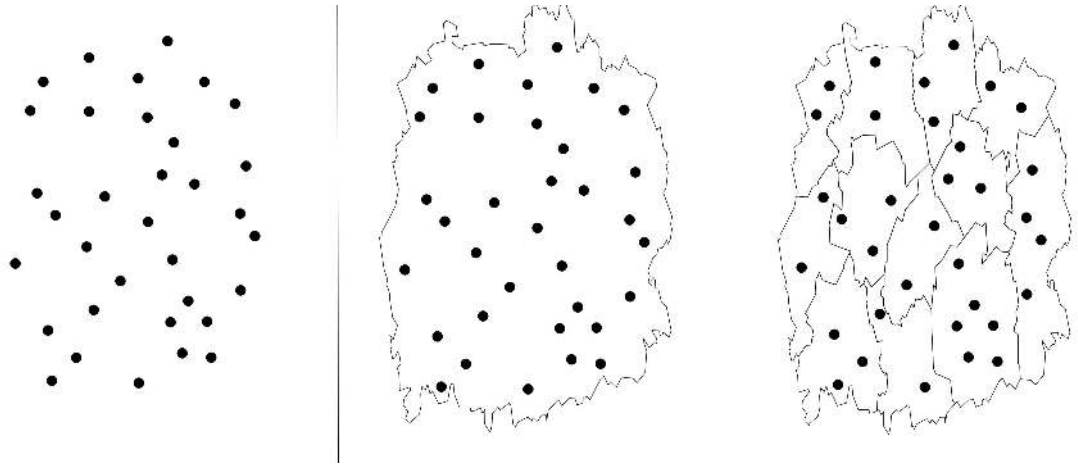


Within

Images from: Pysal

Spatial aggregation



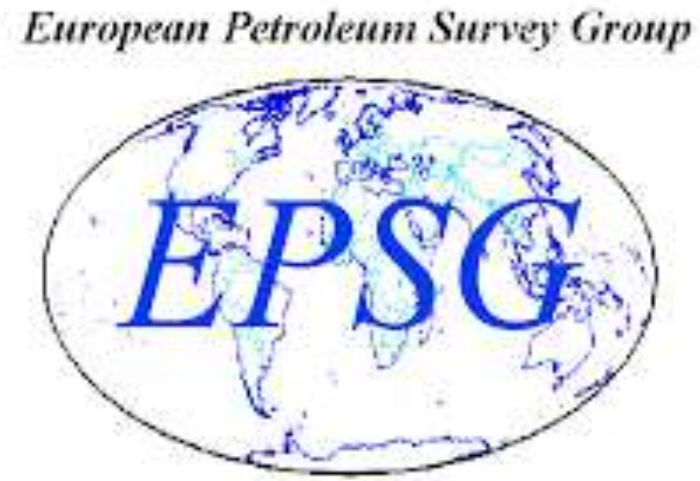


Potential areal data

Possible aggregations

EPSG

The European Petroleum Survey Group (EPSG) created a database of geodetic parameters that includes details on ancient and modern reference systems, map projections, and ellipsoids worldwide.



Source: Spatial reference

EPSG

Spatial Reference

Spatial Reference List

Home | List all references

| Next Page
Search References: Search

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(CSRS98) / New Brunswick Stereo (deprecated)
- EPSG:2037: NAD83(CSRS98) / UTM zone 19N (deprecated)
- EPSG:2038: NAD83(CSRS98) / 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

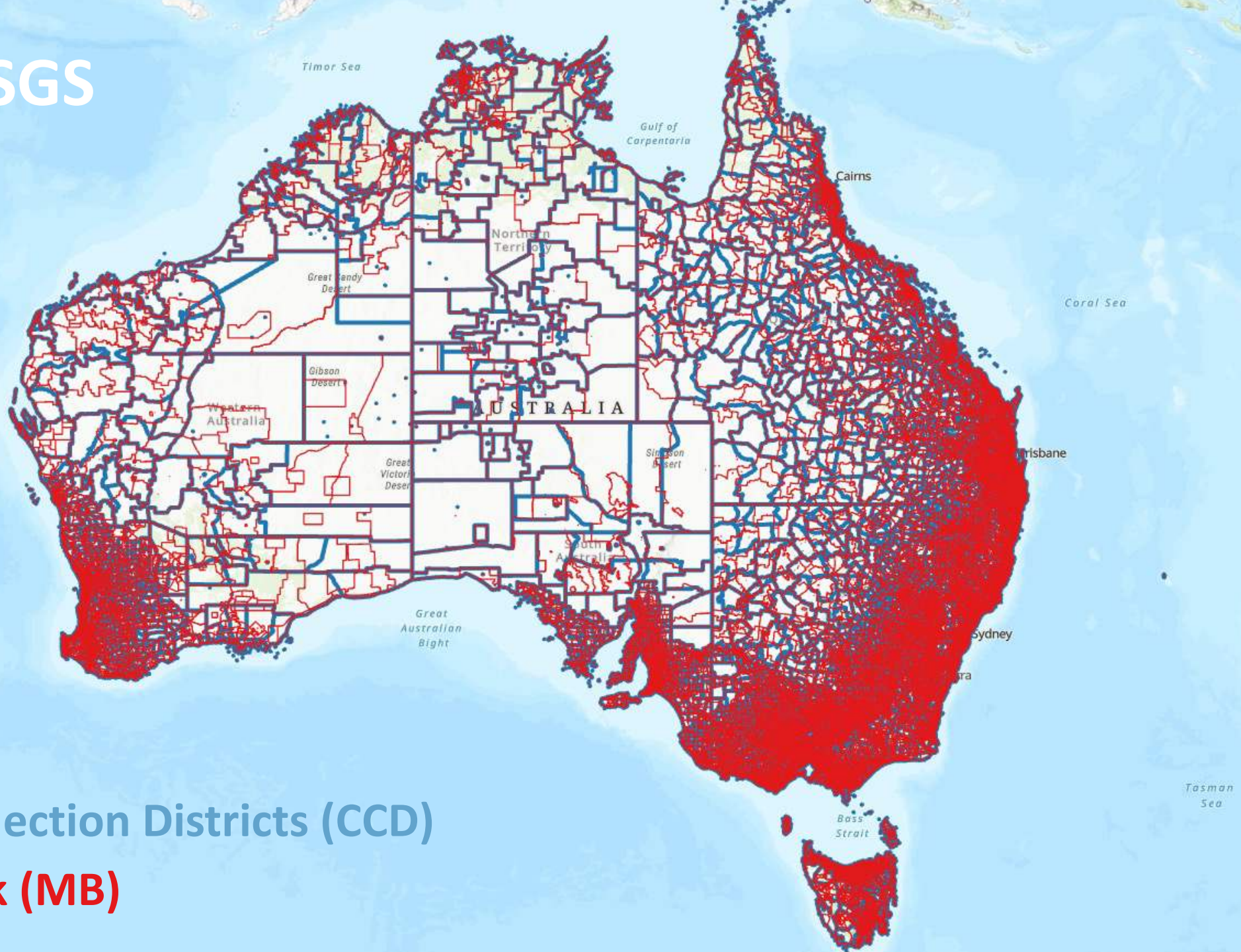
Australia: ABS

- Australian Standard Geographical Classification (ASGC): 1984-2006
- Australian Statistical Geography Standard (ASGS): 2011 to 2021

2006 - ASGC	2011 - ASGS
Statical Area Level 4 (106)	Statical division (69)
Statical Area Level 3 (351)	Statical subdivision (217)
Statical Area Level 2 (2,214)	Statical Local area (1426)
Statical Area Level 1 (54,805)	Collection district
Mesh Blocks (347,627)	

Source: ABS

ASGC vs ASGS

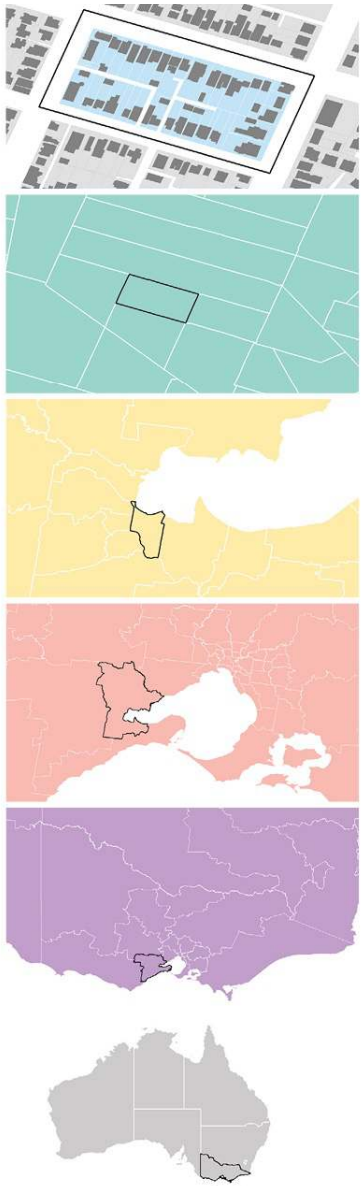


2006 Census Collection Districts (CCD)

2011 Mesh Block (MB)

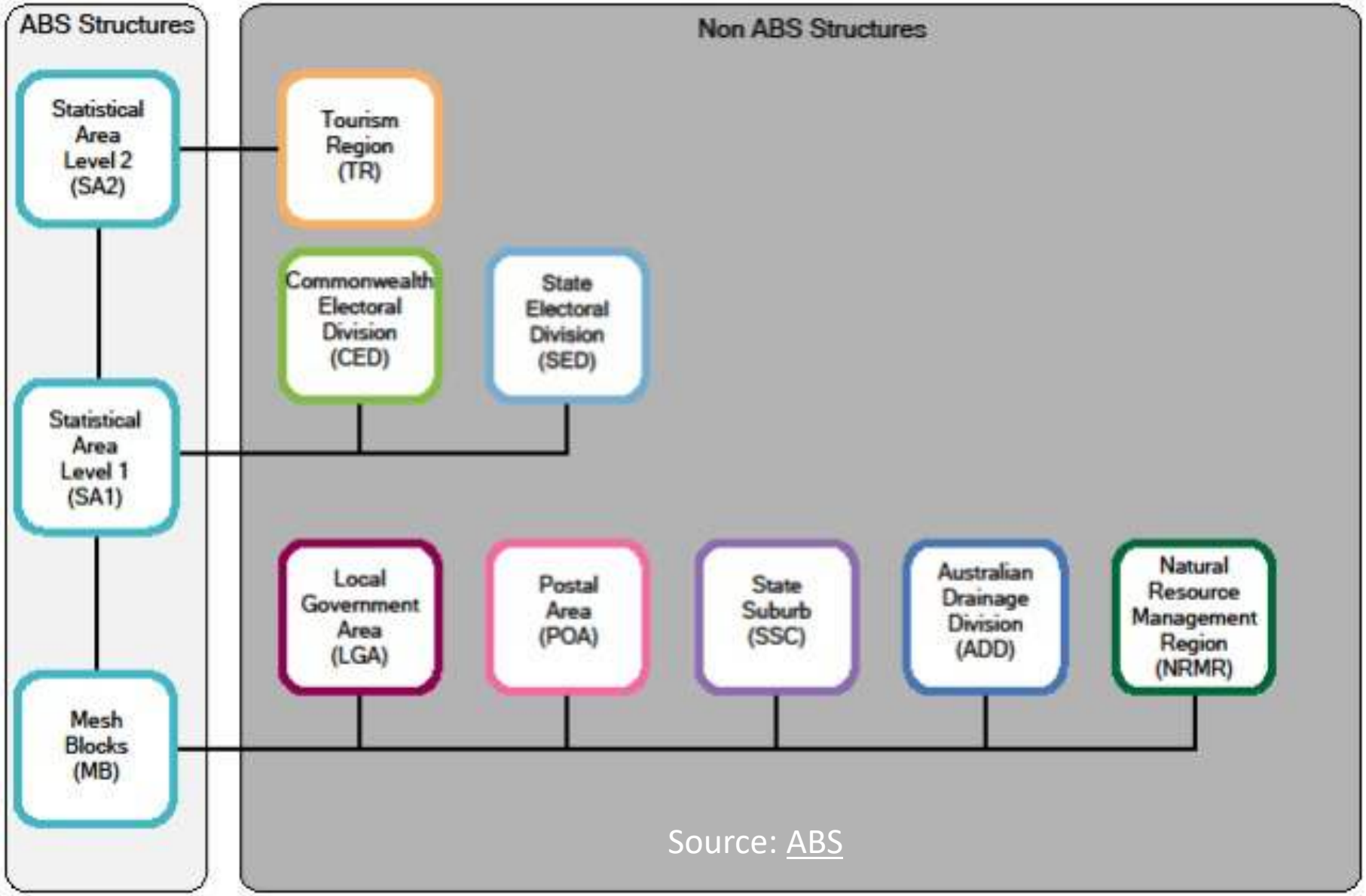
ABS Structure 2011-2021

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



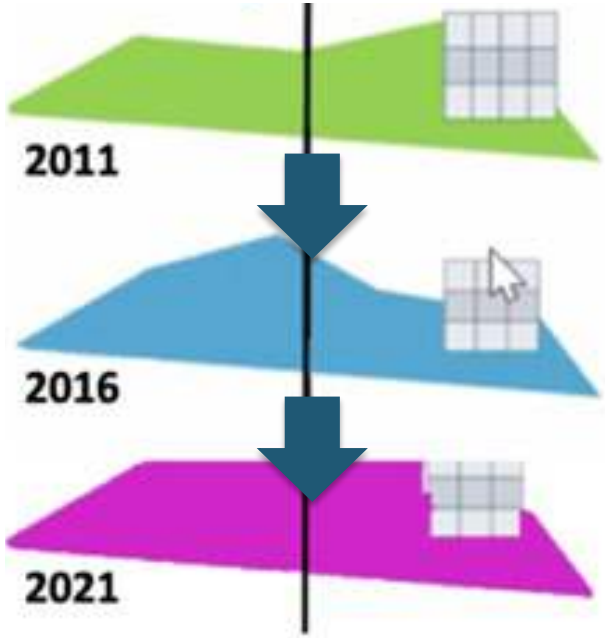
Images from: ABS

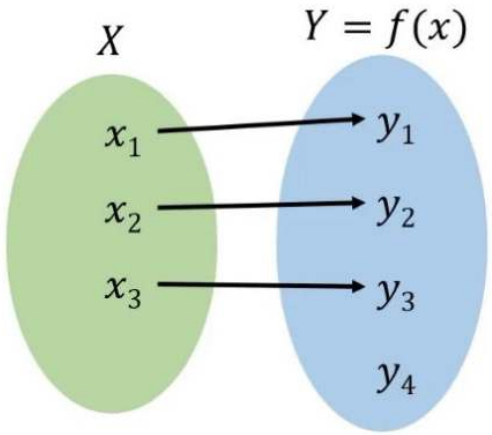
Non-ABS Structure



Source: ABS

Concordance/Correspondence





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

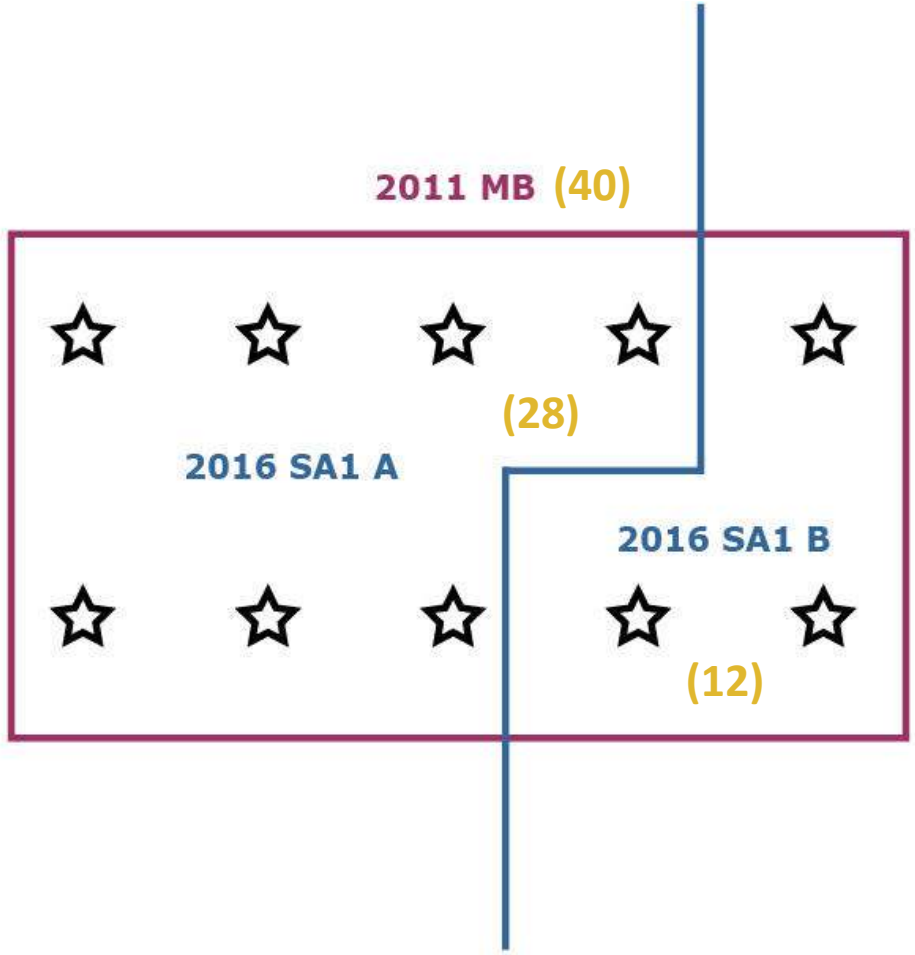
2016 Statistical Areas Level 3 to 2021 Statistical Areas Level 3

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

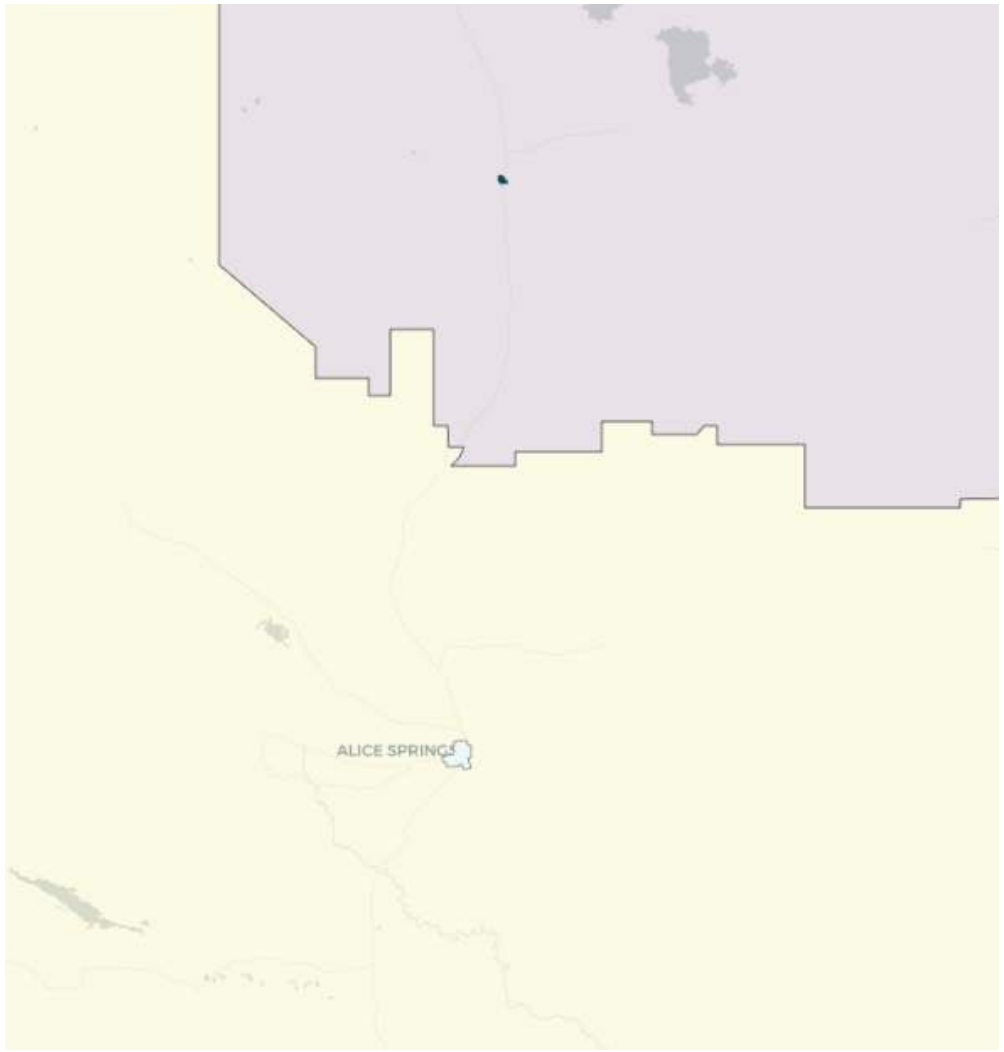
Population Weighted Grid Correspondences



2016 → SA1 A: 28 / 40 which gives a ratio of 0.7 or 70 per cent.
2016 → SA1 B: 12 / 40 which gives a ratio of 0.3 or 30 per cent.

Images from: ABS

Example



Correspondence between Postcode 0870 (2011) to SA3 70201 (2011)

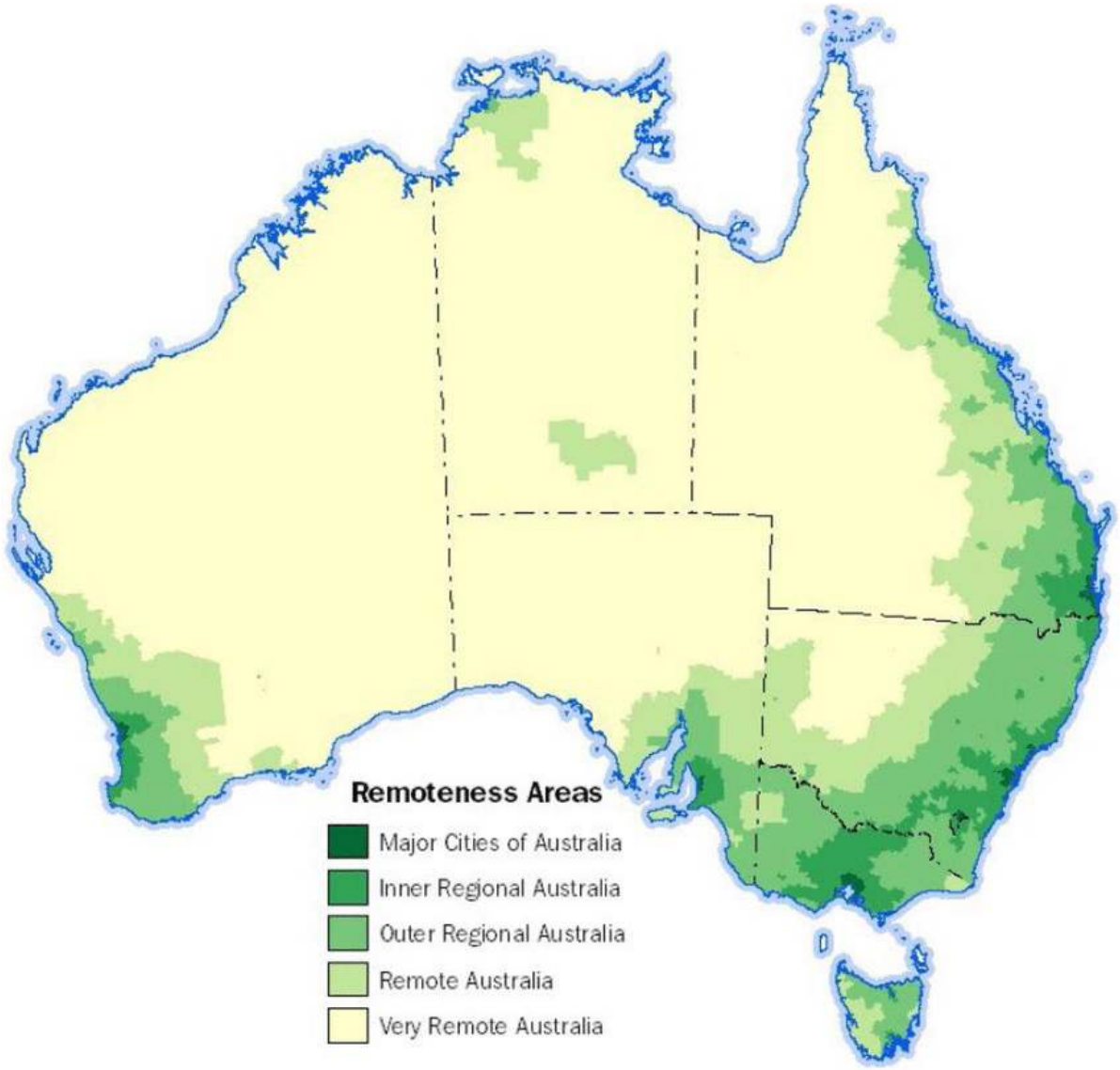
Postcode to SA3 example

Postcode 2011	SA3 code 2011	SA3 name 2011	Percentage
4053	30201	Bald Hills - Everton Park	37.1939731
4053	30202	Chermside	31.1308593
4053	30404	The Gap - Enoggera	18.4242249
4053	30503	Brisbane Inner - North	0.0537311
4053	31401	Hills District	13.1972116

Remoteness Areas for Australia 2021

AURIN
aurin.org.au

The University Of Queensland
AUSTRALIA



iRiSS

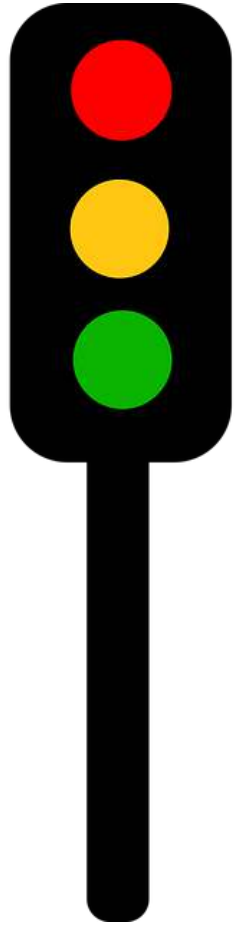
Quality indicator

The quality indicator categorises the ratio of each concordance into one of three values:

- **Ratio > 0.9** = accurate conversion of geographic data.

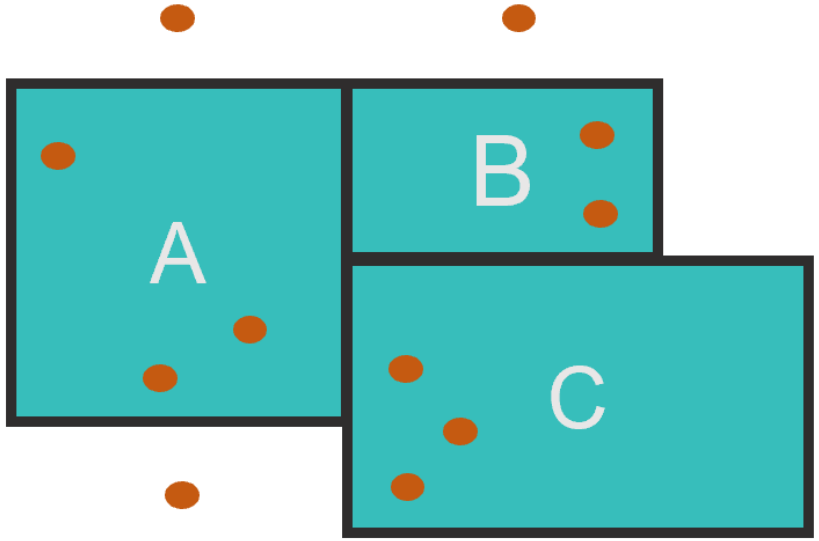
- **Acceptable (0.75 - 0.9):** Data conversion may vary in quality and accuracy, caution is advised.

- **Poor (< 0.75):** Converted data may not reflect the actual characteristics of many geographic regions involved due to inaccurate conversion likelihood. Use it with caution.



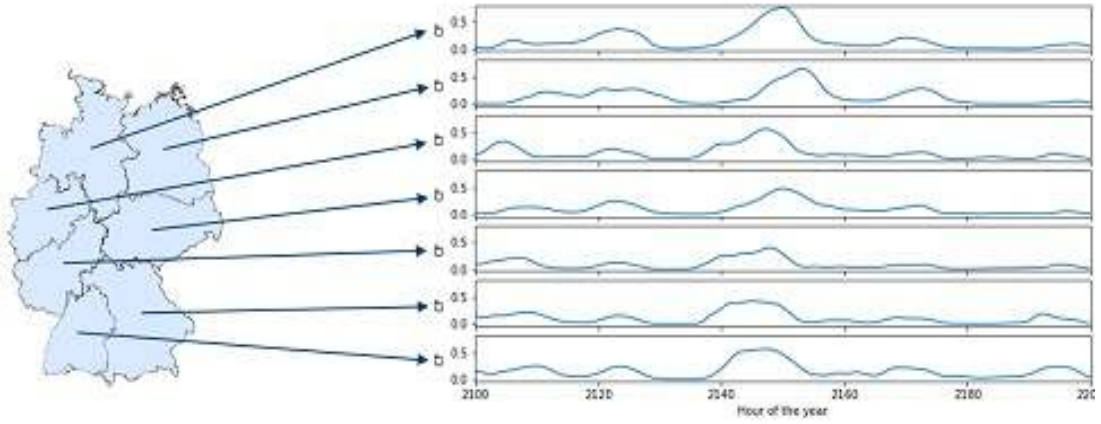
Data integration

Geospatial



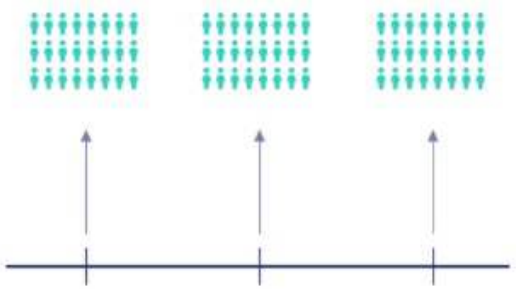
Temporal

Spatiotemporal

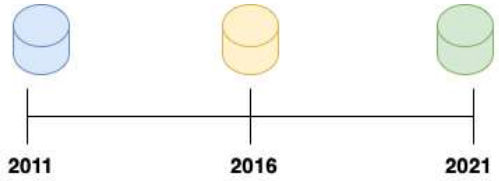


Images from: Metis & Interwork

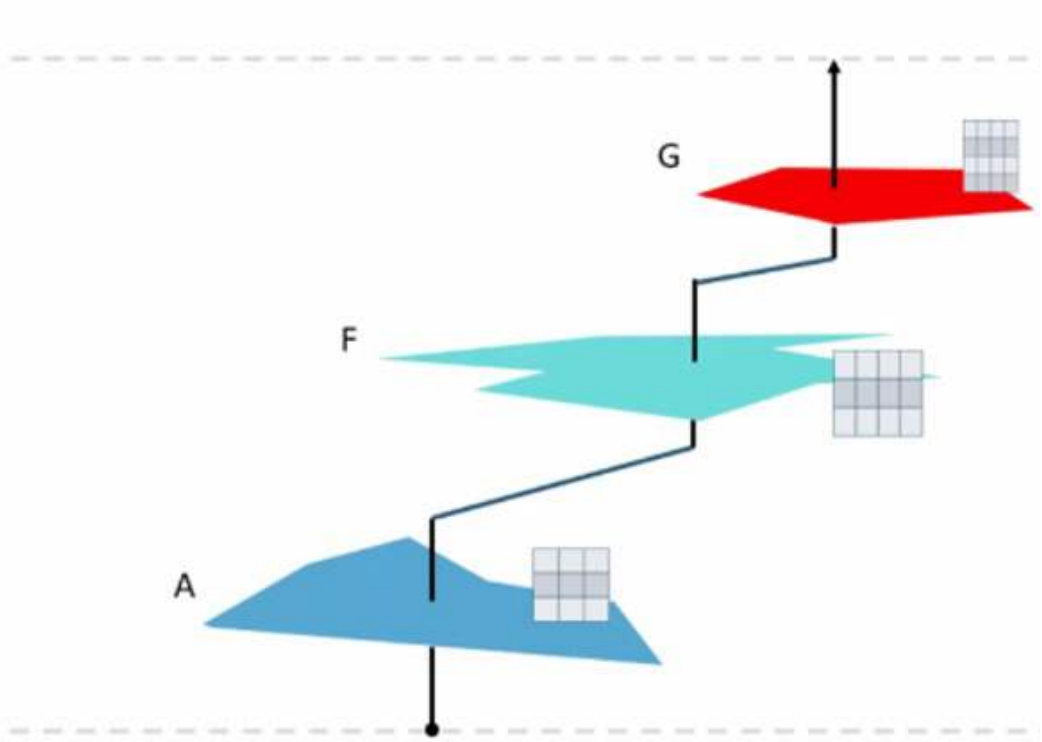
Type of data linkage



Longitudinal data

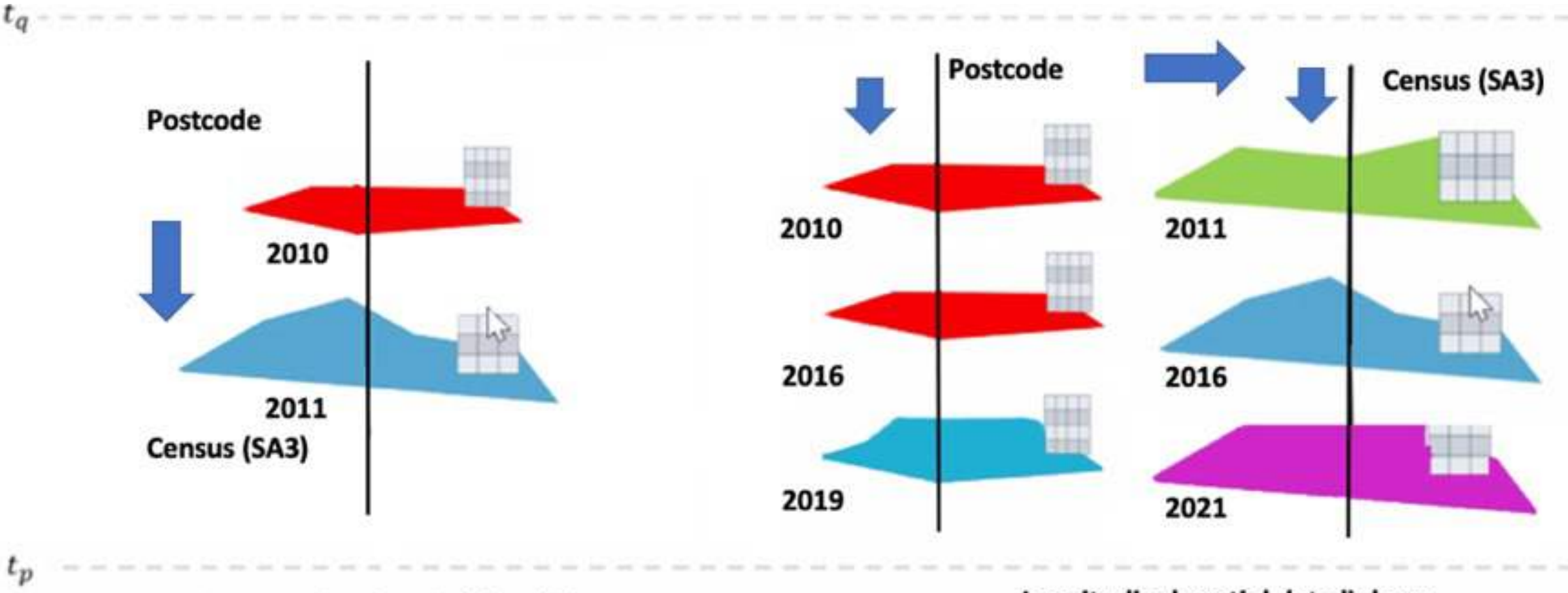


Geographical data



Datasets contain geographic features that may differ in their characteristics, units, and scales.

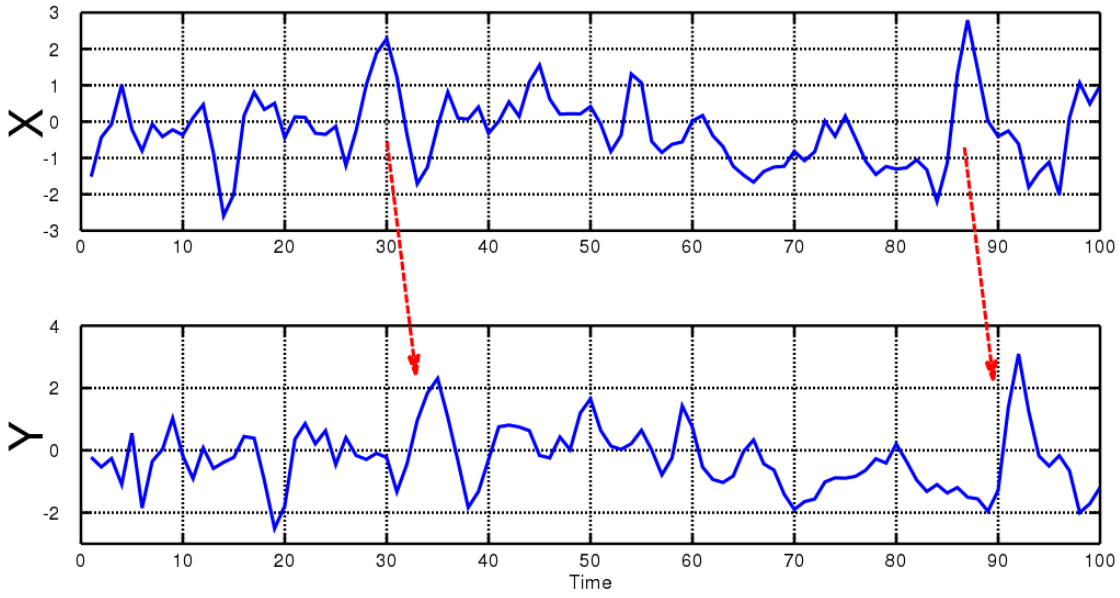
Type of data linkage



Cross-sectional spatial data linkage

Longitudinal spatial data linkage

Methodological Considerations

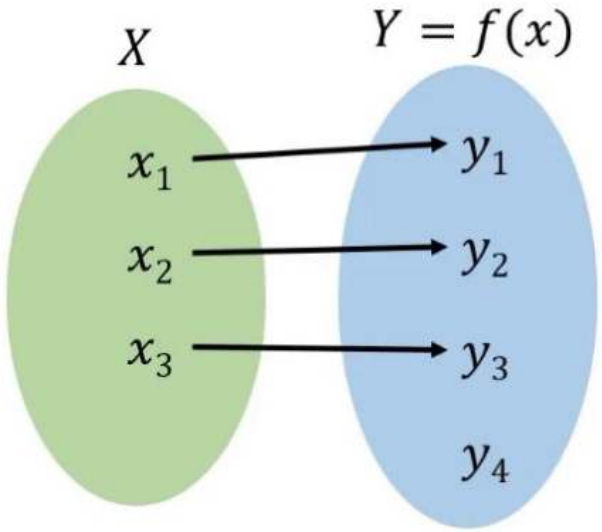


Causality/Temporary lags

Spatial

Images from: BiObserver, Earthdatascience & 52north

Methodological Considerations



Concordance/Correspondence

Comparison between two spatial variables collected in two different moments of the time.

Semantics

Classifications and vocabularies change with time and are not in a machine-readable format.

Methodological Considerations

AURIN
aurin.org.au

The University Of Queensland
AUSTRALIA

Census 2011	Census 2016
01 Negative income 02 Nil income 03 $1-$199 ($1-$10,399) 04 $200-$299 ($10,400-$15,599) 05 $300-$399 ($15,600-$20,799) 06 $400-$599 ($20,800-$31,199) 07 $600-$799 ($31,200-$41,599) 08 $800-$999 ($41,600-$51,999) 09 $1,000-$1,249 ($52,000-$64,999) 10 $1,250-$1,499 ($65,000-$77,999) 11 $1,500-$1,999 ($78,000-$103,999) 12 $2,000 or more ($104,000 or more) && Not stated @@ Not applicable VV Overseas visitor	01 Negative income 02 Nil income 03 $1-$149 ($1-$7,799) 04 $150-$299 ($7,800-$15,599) 05 $300-$399 ($15,600-$20,799) 06 $400-$499 ($20,800-$25,999) 07 $500-$649 ($26,000-$33,799) 08 $650-$799 ($33,800-$41,599) 09 $800-$999 ($41,600-$51,999) 10 $1,000-$1,249 ($52,000-$64,999) 11 $1,250-$1,499 ($65,000-$77,999) 12 $1,500-$1,749 ($78,000-$90,999) 13 $1,750-$1,999 ($91,000-$103,999) 14 $2,000-$2,999 ($104,000-$155,999) 15 $3,000 or more ($156,000 or more) && Not stated @@ Not applicable VV Overseas visitor

Variable Total Personal Income (weekly) (INCP)

Source: ABS

iRiSS

Methodological Considerations

Census 2011	Census 2016
5 Certificate Level	**5 Certificate III & IV Level**
50 Certificate Level, nfd	510 Certificate III & IV Level, nfd
500 Certificate Level, nfd	511 Certificate IV
51 Certificate III & IV Level	514 Certificate III
510 Certificate III & IV Level, nfd	**6 Secondary Education - Years 10 and above**
511 Certificate IV	611 Year 12
514 Certificate III	613 Year 11
52 Certificate I & II Level	621 Year 10
520 Certificate I & II Level, nfd	**7 Certificate I & II Level**
521 Certificate II	720 Certificate I & II Level, nfd
524 Certificate I	721 Certificate II
6 School Education Level	724 Certificate I
611 Year 12	**8 Secondary Education - Years 9 and below**
613 Year 11	811 Year 9
621 Year 10	812 Year 8 or below
622 Year 9	
067 Year 8 or below	

Variable Level of Highest Educational Attainment (HEAP)

Source: ABS

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.

Introduction

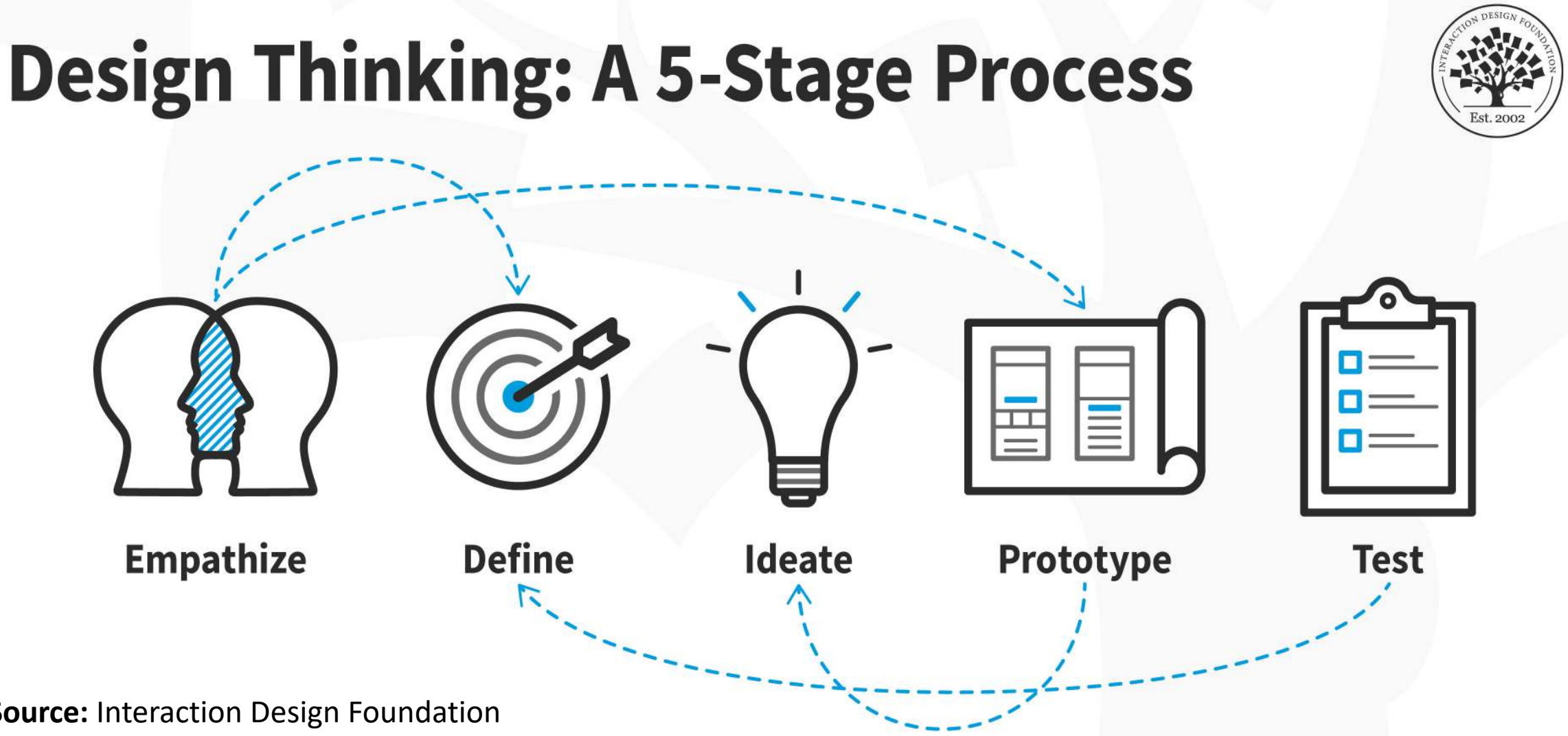
Motivation

Spatial data and data integration

Service design

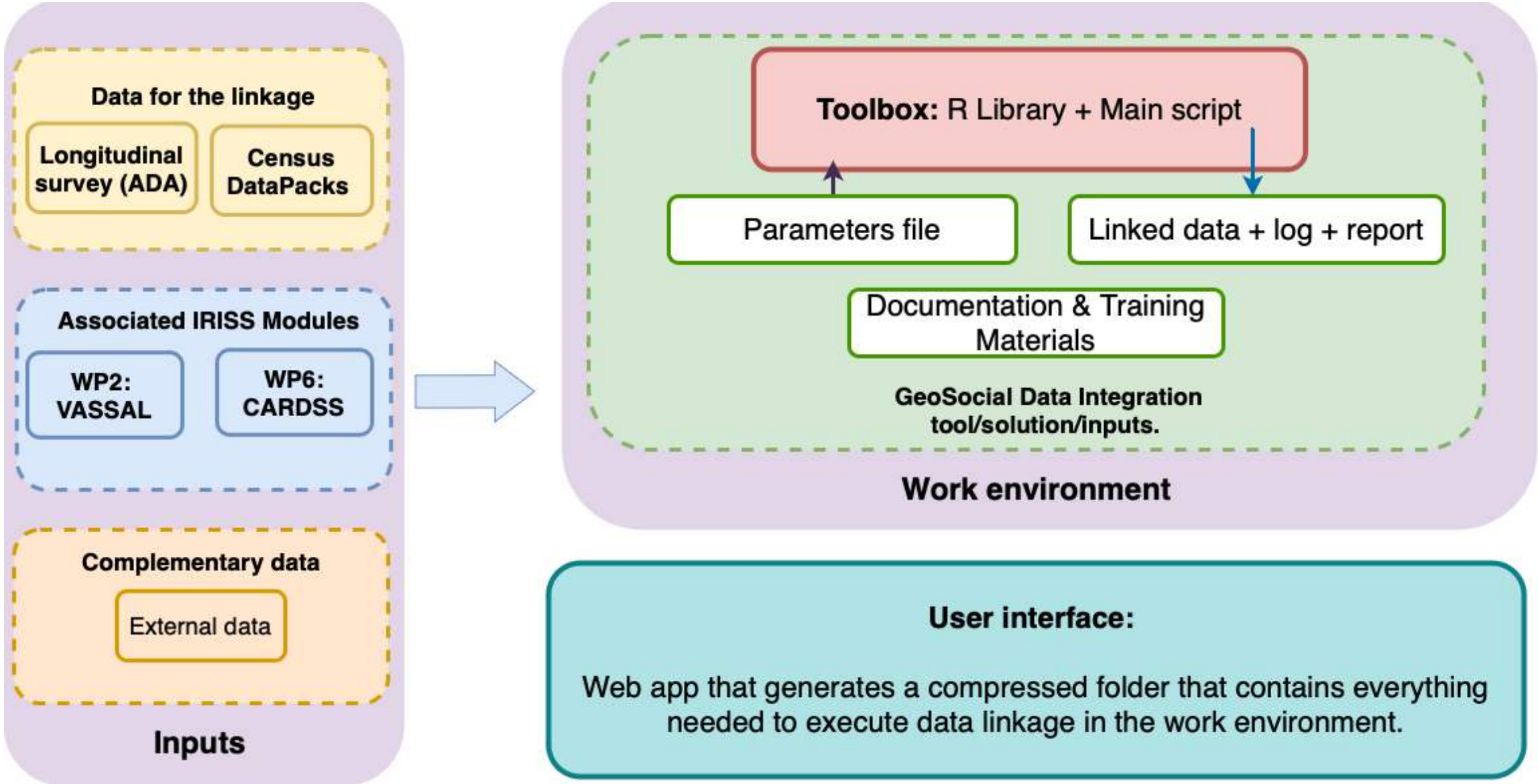
Demonstrator

Design

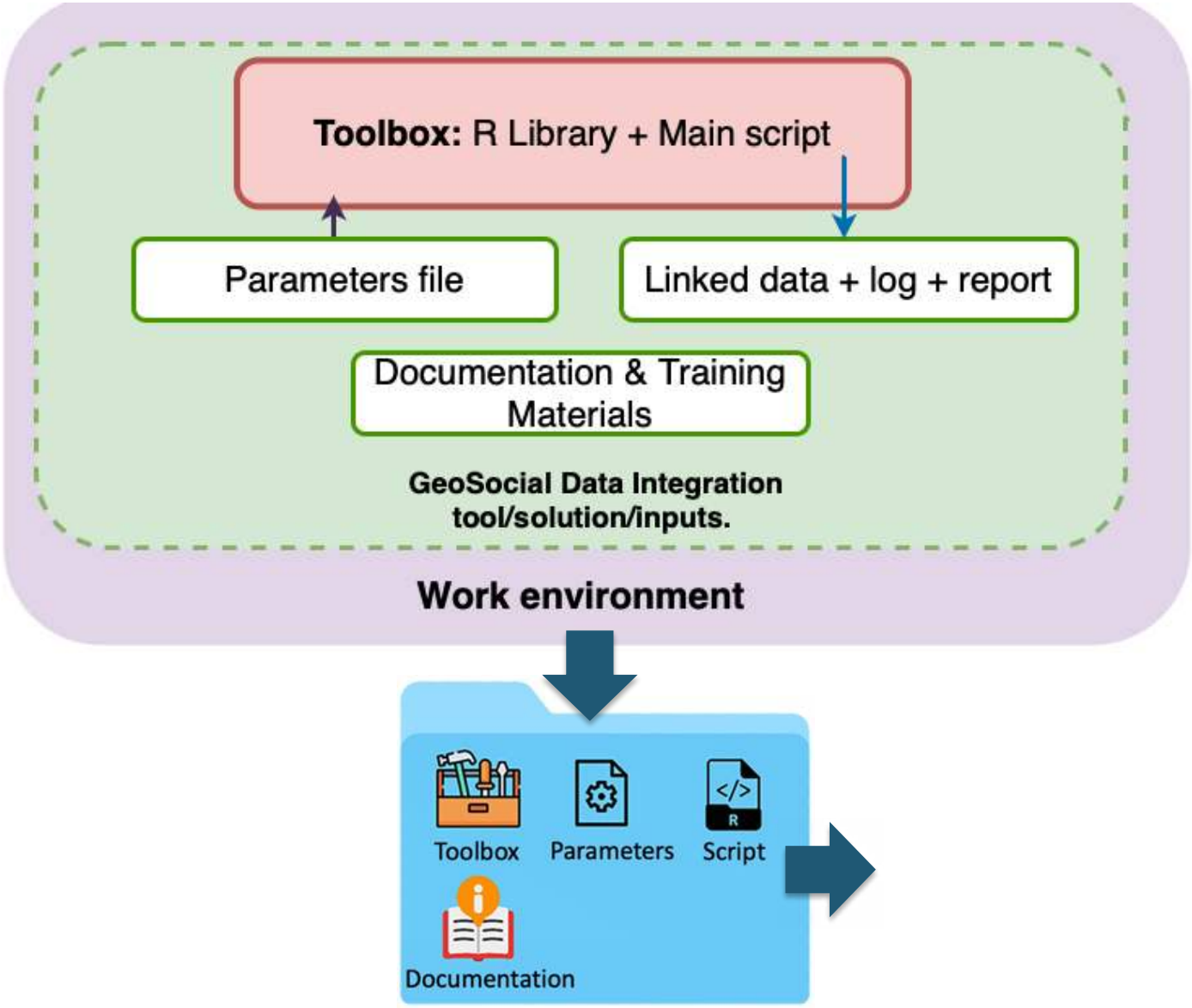


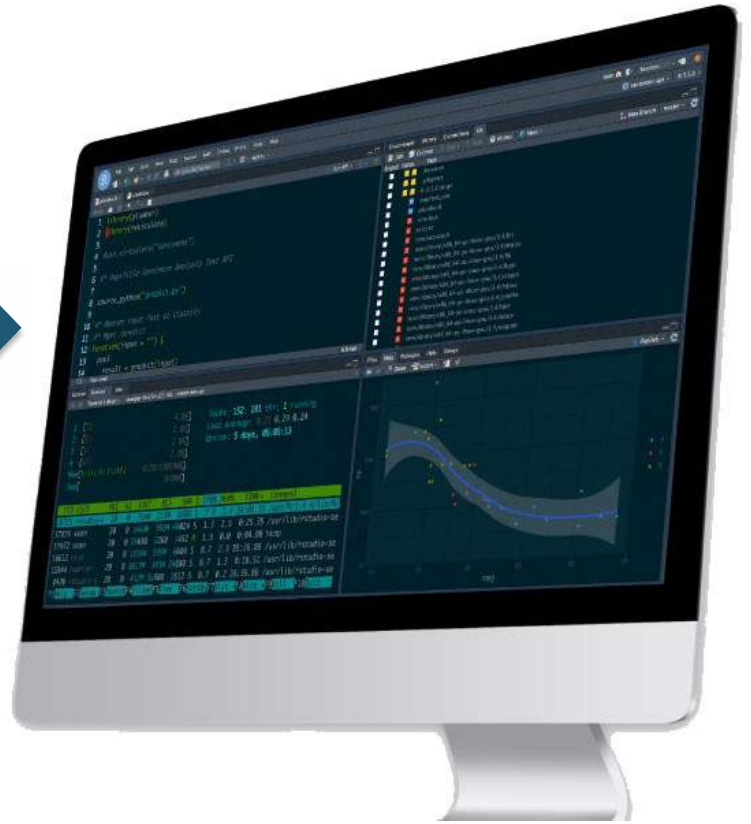
Source: Interaction Design Foundation

Geosocial Service Design



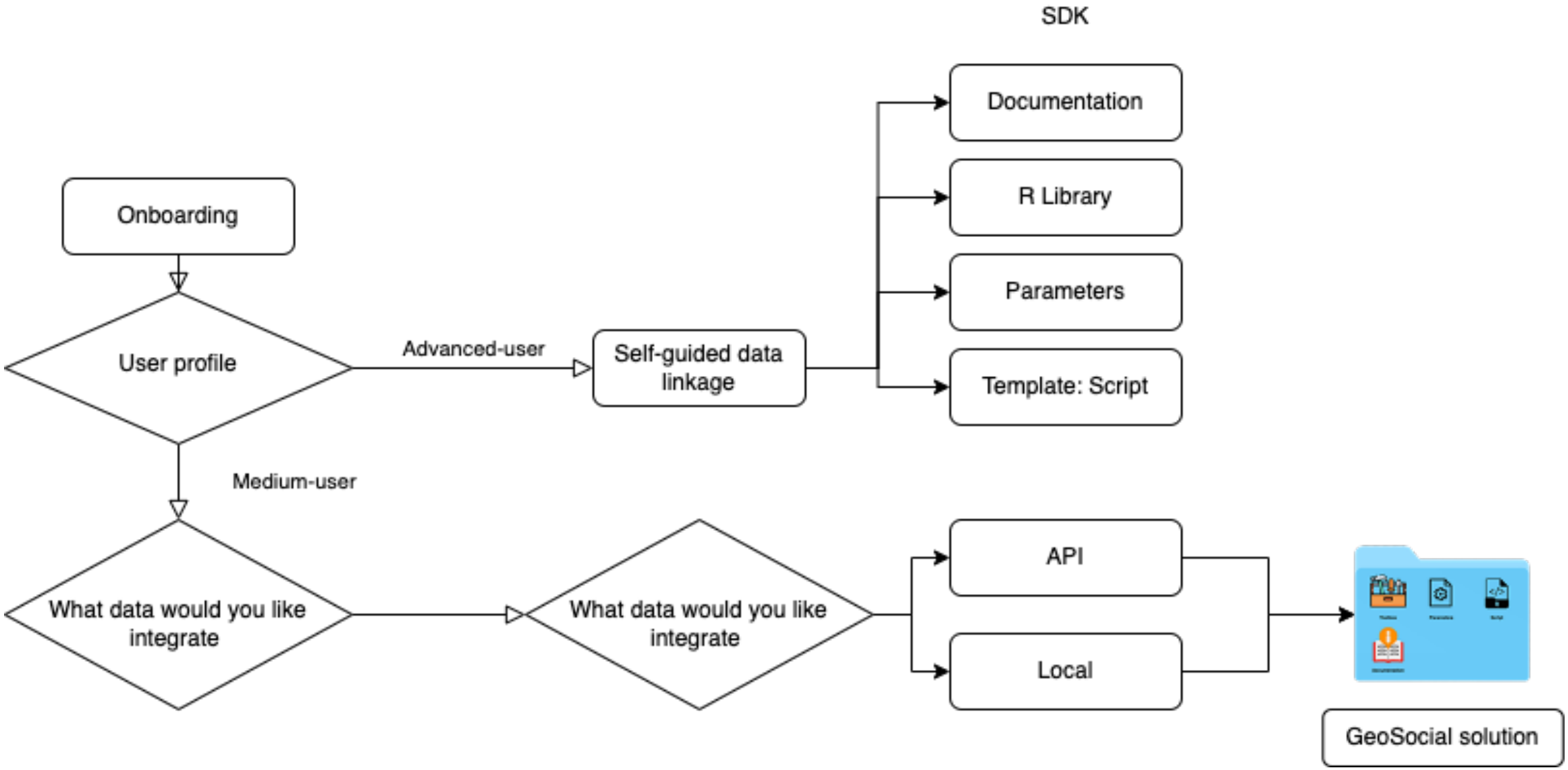
Geosocial Service Design



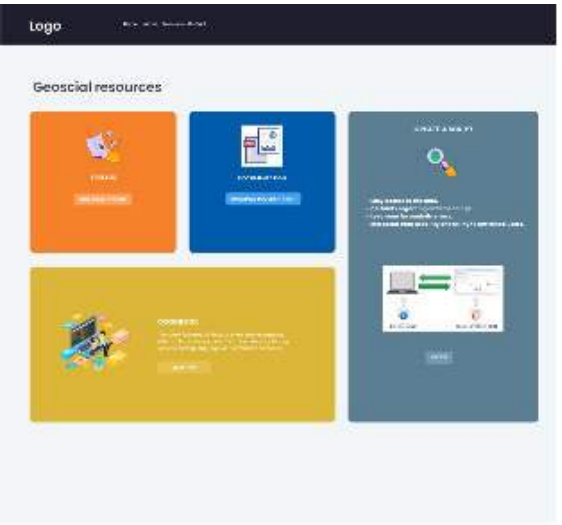


User: work environment

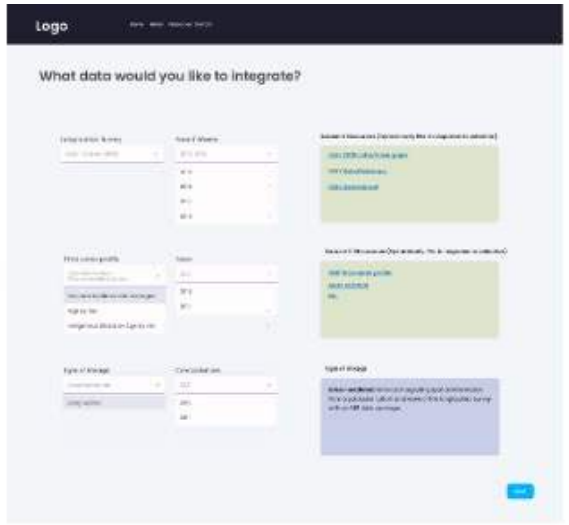
Geosocial Service Design



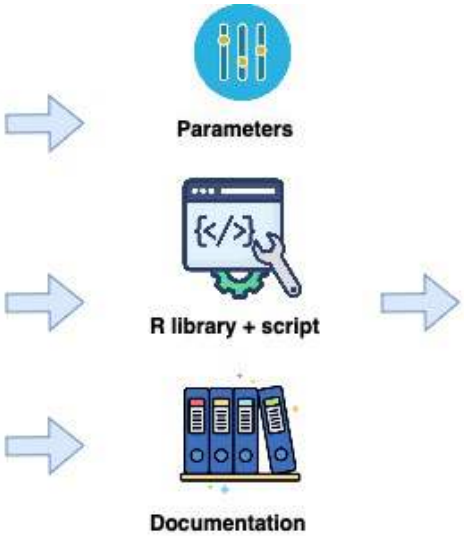
Mid-level user: User interface



Step 1: Visit GeoSocial resources



Step 2: Selection of Parameters: Type of linkage, wave, variables, etc.



Step 3: Download Toolbox

Step 4: Run the code using the work environment

Advanced user

Resources

TOOLBOX

R library with essential functions needed for data linkage.

DOWNLOAD

DOCUMENTATION

Information about each function in the library including a description, arguments, and value.

DOCUMENTATION

PARAMETERS

File with the definitions that are relevant to the data linkage process: survey, cohort etc.

DOWNLOAD

SCRIPT

A template of the workflow which uses the toolbox to read and merge the data based on user preferences.

DOWNLOAD

Benefits:

- Designed to meet the needs of mid to advanced users
- Meet the needs of mid and advanced users
- Interoperable and can connect with other work packages
- Does not interfere with data custodian requirements.
- Transparency
- High level of personalization

Introduction

Motivation

Spatial data and data integration

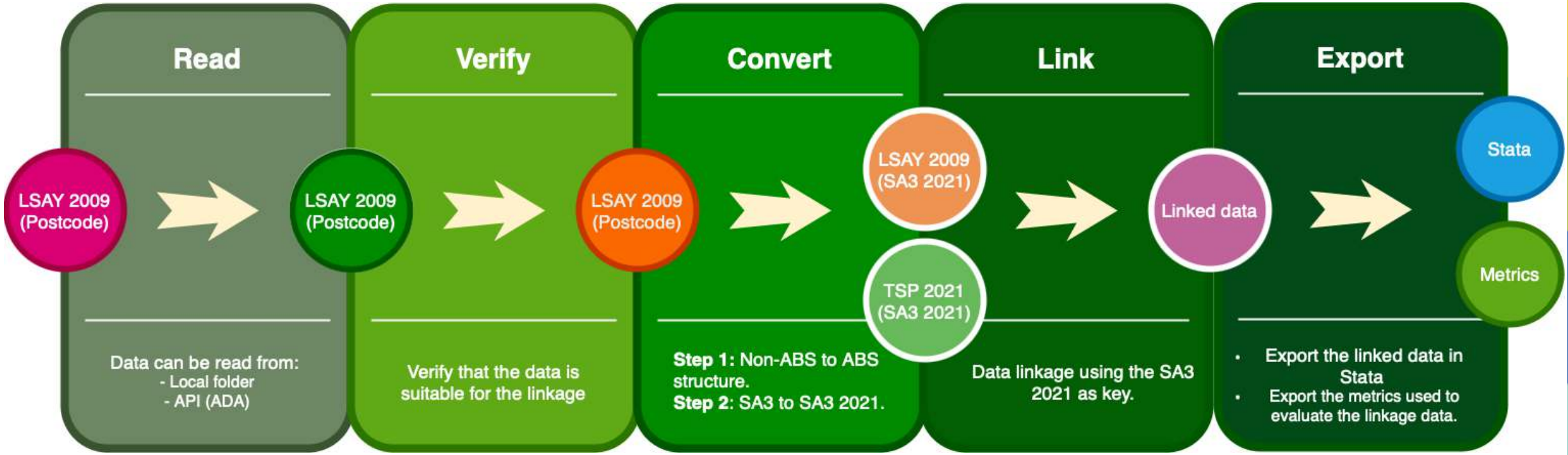
Service design

Demonstrator

Demonstrator:

Longitudinal Survey: LSAY 2009

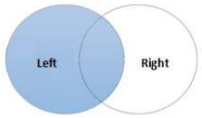
Geospatial dataset: Time series profile 2021



1. Geospatial concordances

a) Non-ABS structure to ABS structure → **b) Concordances** →

LSAY 2009	Non-ABS structure	Potential ABS structure	TSP linkage	TSP 2021 - Census data
Wave 2	Postcode 2011	SA3 2011	SA3 2021	2011



2. Data linkage: Home (SA3) with TSP 2021 sociodemographic characteristics (13 – TSP demonstrator)

Left: Wave 2: 2011

SA3 - 2021	STIDSTD
10102	x
10103	x
10104	x
...	...
...	...
Rows	14251

Right: TSP 2021

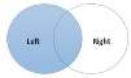
SA3 - 2021	Variable-x_2011	Variable-y_2011	Variable-z_2011	Columns
1	...	...	...	2299
2	...	...	...	2299
3	...	...	...	2299
...	...	...	...	2299
Rows	358	358	358	

→

Wave 2 - TSP 2021 (2011)

SA3 - 2021	STIDSTD	X_ABS_2011	....	Z_ABS_2011	Columns
1	x	...	...	...	2301
2	x	...	...	...	2301
3	x	...	...	...	2301
...	...	...	...	...	2301
...	...	...	...	...	2301
Rows	14251	14251	14251	14251	

3. Data linkage: Home (SA3) + TSP 2021 with LSAY 2009 (filtered by the questions associated with the wave)



Right: Wave 2 - TSP 2021 (2011)

SA3 - 2021	STIDSTD	X_ABS_2011	....	Z_ABS_2011	Columns
1	x	...	...	...	2301
2	x	...	...	...	2301
3	x	...	...	...	2301
...	...	...	...	...	2301
...	...	...	...	...	2301
Rows	14251	14251	14251	14251	

Left: LSAY 2009 – WAVE 2

STIDSTD	X_LSAY_2011	....	Z_LSAY_2011	Columns
x	...	...	...	808
x	...	...	...	808
x	...	...	...	808
...	...	...	...	808
...	...	...	...	808
14251	14251	14251	14251	

→

4: Output

Wave 2 - TSP 2021 (2011)

STIDSTD	SA3 - 2021	Variable-x_2011	....	Variable_LSAY_2011	Columns
x	10102	...	...	...	3108
x	10103	...	...	...	3108
x	10104	...	...	...	3108
...	...	...	...	...	3108
...	...	...	...	...	3108
Rows	14251	14251		14251	

STATA

R Functions

Parameters
ParsingParameter
LoadParameters

Read
LoadTSP2021
LoadLSAY: - SurveyLSAY: - SpatiallyLSAY:

API
TestDataverseConnection
downloadDataverseData

Check
checkLSAY
checkVariableNames
checkNamesDuplicates
checkPostcodeStructure

Convert
PotentialCensus
FilterConcordance
QualityIndicator
TransformPOA
TransformSA3
LSAY_POA_SA3
LSAY_PSA3_SA3
GeoSpatialJoin

Vocabularies
SearchConcept
GetTerm

Utilities
CreateFolders
SummaryReport
SummaryLog
WriteStata

R library

Package 'geosocial'

June 30, 2023

Type Package

Title IRISS WP3 GeoSocial solution - Toolbox

Version 1.0

Author Australian Urban Research Infrastructure Network (AURIN)

Maintainer Pascal Perez <pascal.perez@unimelb.edu.au>,German Gonzalez <german.gonzalez@unimelb.edu.au>, Masoud Rahimi <masoud.rahimi@unimelb.edu.au>

Description

License GLP-3

Encoding UTF-8

LazyData true

Imports haven, dplyr, sjlabelled, openxlsx, rjson, dataverse, tidyverse

Depends haven, dplyr, sjlabelled, openxlsx,rjson,dataverse,tidyverse, R (>= 2.10)

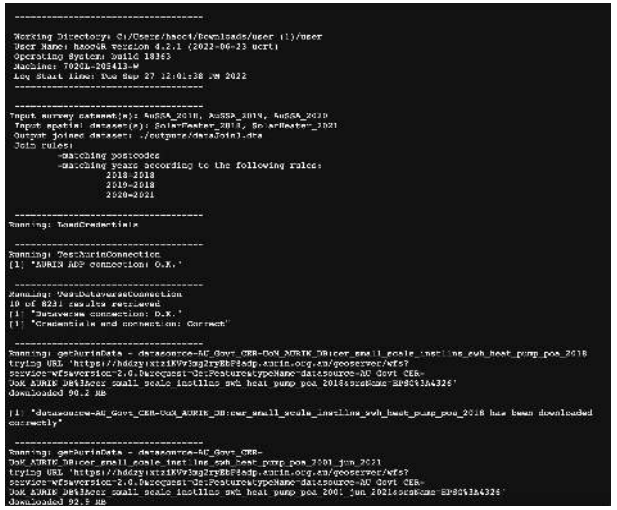
RoxygenNote 7.2.3

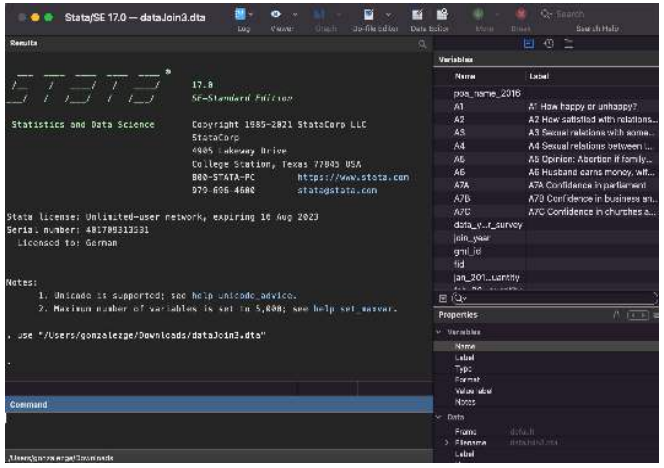
R topics documented:

R Library

Parameters:

R script





User interface

- **Requirements**: Control of all the dependencies and requirements that the application needs, such as specific versions of programming language run times and other software libraries.
- **Isolation**: Containers give developers the ability to create predictable environments that are isolated from other applications.
- **Agility**: Containers accelerated development, improved consistency across environments, empowered autonomous teams improving productivity and quality.

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

Guided data linkage

We have developed a pipeline to guide you through the components involved in the linkage. The guided option provides:

- Easy access to the data.
- Certainty regarding data meanings.
- Less room for analytic errors.
- Increased data usability and utility to untrained users.

SELECT

Self-guided data linkage

We have allowed you to customise your data pipeline and personalize the data linkage. The self-guided option is suitable if you are:

- Confident with using Python and/or R for data wrangling, integration, and analysis.
- Familiar with geospatial data.
- Adding new datasets.
- Supporting other social science researchers.

SELECT

BACK

What data would you like to integrate?

Longitudinal Survey:

LONGITUDINAL SURVEY

Years/Waves:

SELECT MORE THAN ONE

Sub-major topic area:

SELECT MORE THAN ONE

DataPack

SELECT MORE THAN ONE

Census

SELECT MORE THAN ONE

Variables:

SELECT MORE THAN ONE

Survey data documentation

- How to access LSAY data
- LSAY 2009 cohort user guide
- LSAY variable listing and metadata
- LSAY 2009 cohort questionnaires and frequency tables

Geospatial data documentation

- ABS DataPacks
- Understanding Census geography
- ASGS SA3s
- Geographic correspondences

BACK

CONTINUE

Where would you like your integrated data stored?

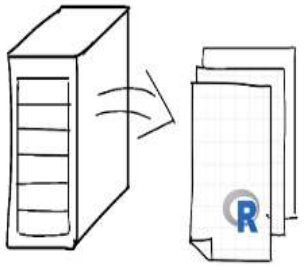
The toolbox allows the user to load the survey data from one of the following sources:

Australian Data Archive

The survey is provided by ADA, a national service for collecting and preserving digital research data.

CLOUD

Local environment



The user provides the survey in the local environment where the toolbox is executed.

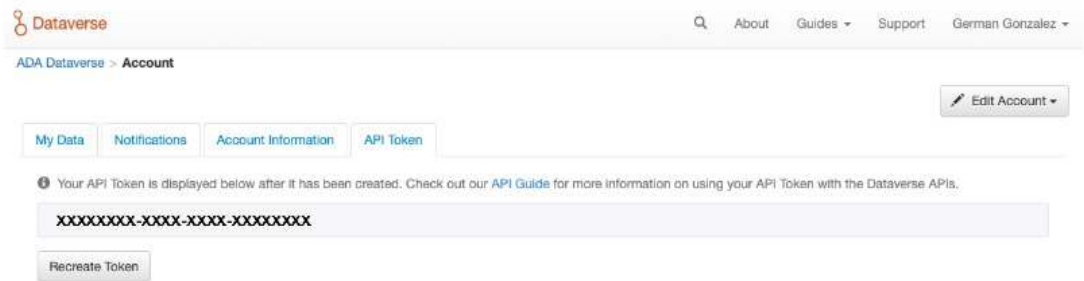
LOCAL

BACK

Where would you like your integrated data stored?

Australian Data Archive (ADA) API

Before generating an API token to use the ADA API, it is necessary to obtain approval to access the LSAY 2009 data through ADA. Click here for information. After getting the approval, you can create a token. Please refer to the image below to locate it.



Please copy and paste the ADA token into the designated field below:

```
1 Please introduce your ADA token
```

We do not collect or upload any information. The token is included in the parameters file that you execute on your computer.

BACK

CONTINUE

Where would you like your integrated data stored?

Local environment

In order to load the LSAY 2009 cohort, you need to indicate where it is located on your computer.

Please indicate the folder where the LSAY 2009 cohort in Stata format is located:

```
1 | Please introduce your absolute path. For example: C:\Users\example\Documents\LSA09\
```

We do not collect or upload any information. The absolute path is included in the parameters file that you execute on your computer..

BACK

CONTINUE

Thank you, we have generated all the necessary components for the data linkage

Step 1: Download GeoSocial

DOWNLOAD

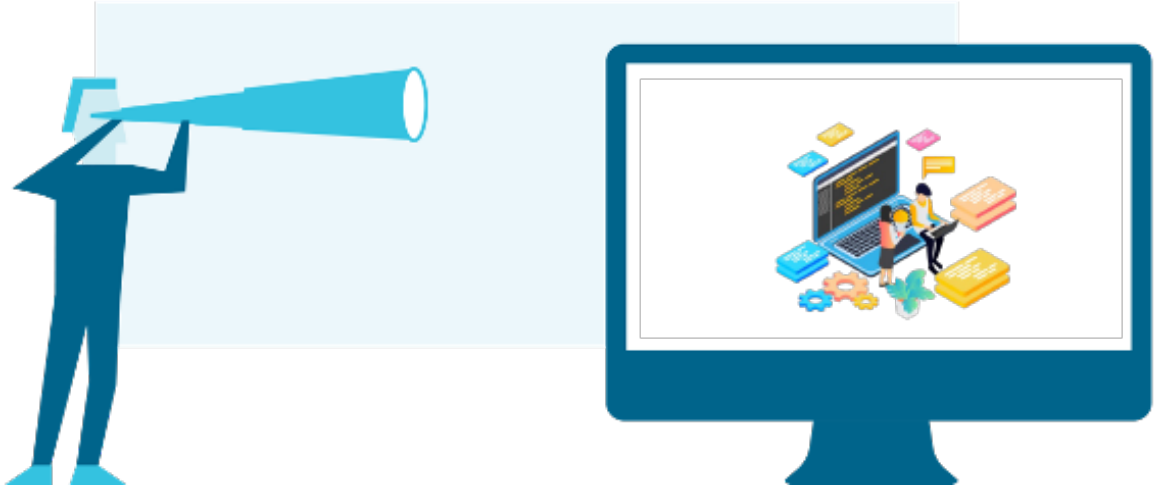
Step 2: Read "readme.pdf"
It will introduce you to the code and explain each chunk of it.

Step 3: Run the code
To start the data linkage, it is necessary to execute the main.R

Step 4: See the outputs
The linked data is stored in a new file containing the GeoSpatial variables.

BACK

Next steps



- Consultations/road testing with a broad range of stakeholders
- Re-engaging with the policy and service agencies that provide and control access to various relevant datasets.
- Collect user training needs, and develop a forward plan for user training and community engagement
- Create a web app that executes the flow to low-skilled users

Thank you

Follow us:@ausiriss