



Introduction to GIS

Raphaëlle ROFFO

Sciences Po - Urban School



Session 2

Sourcing and loading data in GIS

Today's plan

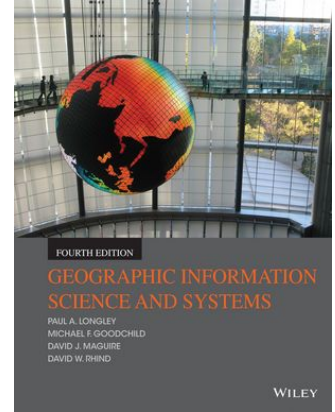
1. Last week recap
2. Common GIS data formats
3. Data sourcing
4. Looking at data in kepler.gl
5. Loading data in QGIS



Session 1 and tutorial:
questions?



Session 1 Recap: GIS as “The science of where”



“Geographic Information Systems are computer-based tools that analyze, store, manipulate and visualize geographic information, usually in a map.”

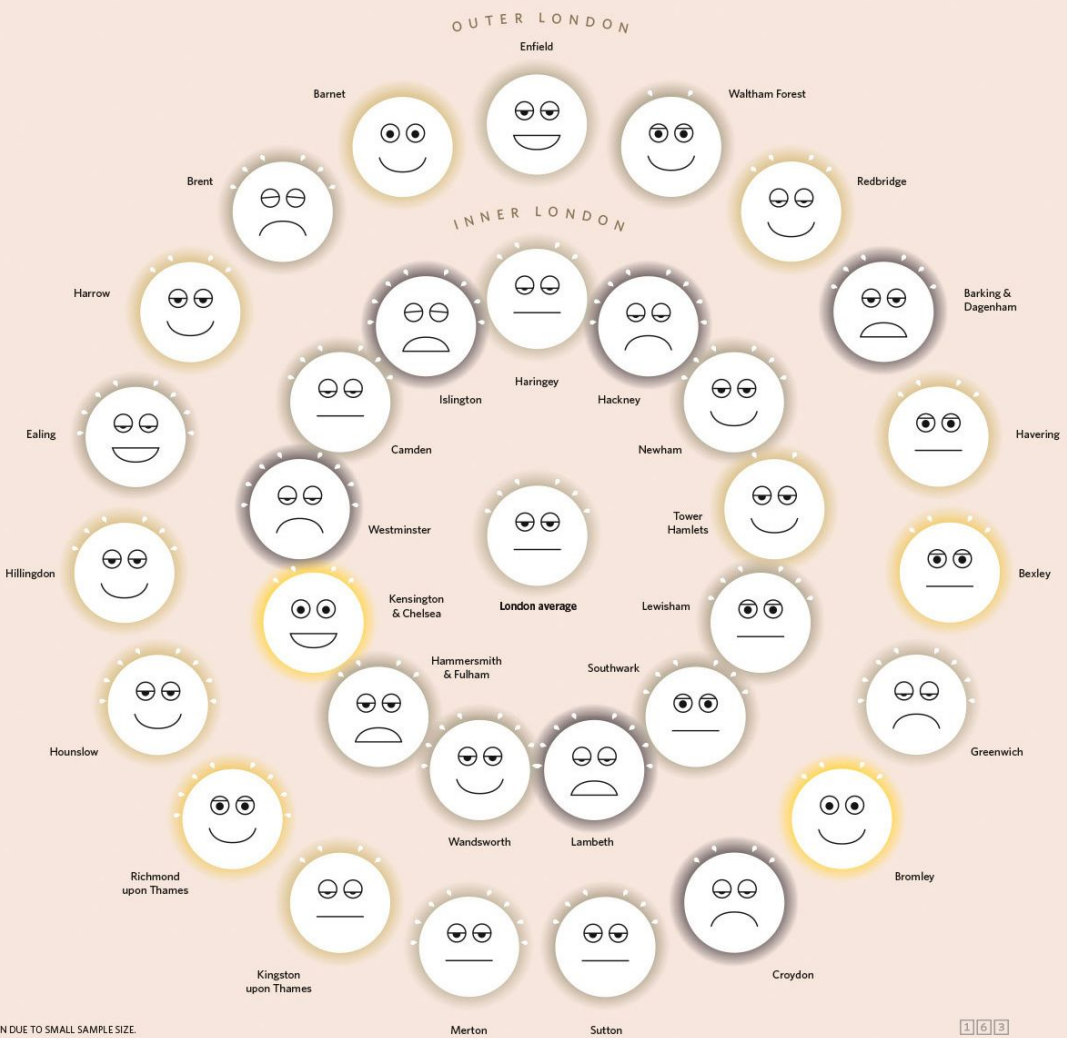
Michael Goodchild

Session 1 Recap: GIS use cases

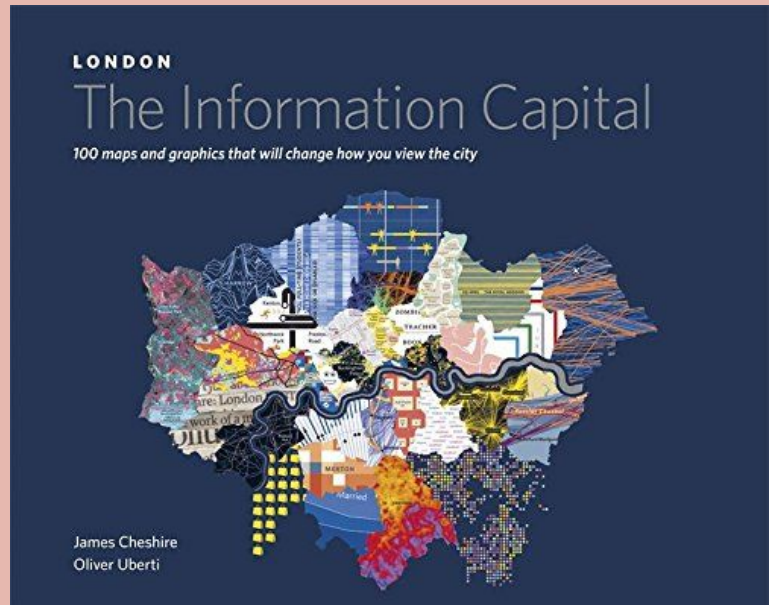
- How various social and environmental phenomena happen across geographies
- Detecting patterns
- Exploring how 2 variables may influence each other
- Especially relevant to inform urban policy making

Session 1 Recap: GIS use cases

- Finding the best location for a new public library, such that it benefits most the young and the elderly from low income households.
- Carrying out an environmental impact analysis ahead of a new urban project such as a road
- Assessing a community's vulnerability to flooding
- Designing a new bus route that will reduce the time that the most vulnerable parts of the population will take to reach a hospital.



WN DUE TO SMALL SAMPLE SIZE.



The basis of it all: DATA

Urban and environmental policies are incredibly common GIS use cases and open datasets of great quality are available to you.

→ But how do you **find data** and how to you make sure you're using **good quality** datasets?

1. Identify your research question (Requirements gathering)



2. Finding the DATA





2. Common GIS data formats





Spatial data representation

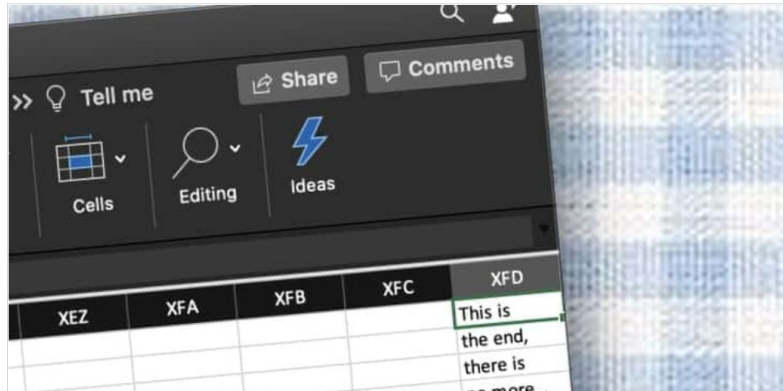
Data formats **do** matter !

UK loses 16,000 COVID-19 cases due to Excel spreadsheet snafu

Today's lesson is: use the right tool for the job.



Graham Cluley • [@gcluley](#)
11:47 am, October 5, 2020



Some 16,000 Coronavirus cases went missing after the Excel spreadsheet they were being recorded in reached its maximum limit, and did not allow the automated process to add any more names.

As a result, it's possible that some people who might have been infected by COVID-19 may not have been properly traced in a timely fashion.

Geospatial data can efficiently be stored in spatial databases

Data size	No. bytes	Example dataset	Example storage system
1 megabyte	1,000,000	Single data set in a small project database	Text files, .csv files, single shapefiles, .kml files
1 gigabyte	1,000,000,000	Entire street network of a large city or small country	ESRI geodatabase QGIS geopackages
1 terabyte	1,000,000,000,000	Elevation of entire Earth surface recorded at 30 m intervals	Enterprise spatial database: e.g. Oracle Spatial, PostgreSQL/PostGIS
1 petabyte	1,000,000,000,000,000	Satellite image of entire Earth surface at 1 m resolution	Distributed/cloud storage, e.g. Hadoop, Hive

Representing Spatial Data

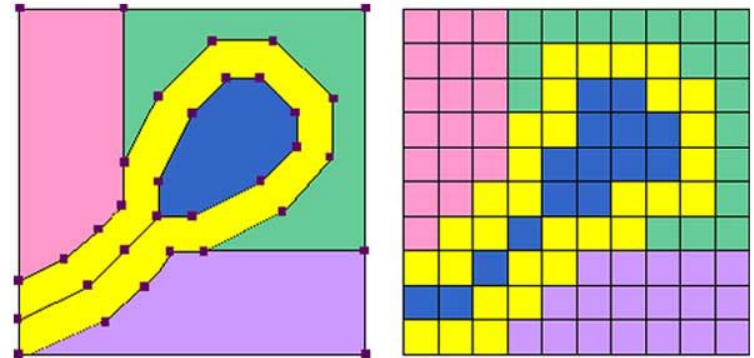
Object view versus field view

- **Object view:** The world is filled with discrete objects around which we can draw boundaries
 - Forests, cities, lakes, rivers, buildings etc.
 - **Field view:** The world is a continuous surface containing a finite number of variables, each of which can be measured at each location
 - Landuse, elevation, population density etc.
- The most appropriate representation depends on data type, context, scale

Vector vs Raster

Spatial data can be represented in two ways:

- **Vector** = Object view = geometries: Point, Line, Polygon
- **Raster** = Field view = pixels, fields of continuous vlike a photo (each pixel is assigned a value)



Vector

Raster

Source: David DiBiase et al, [The Nature of Geographic Information](#)

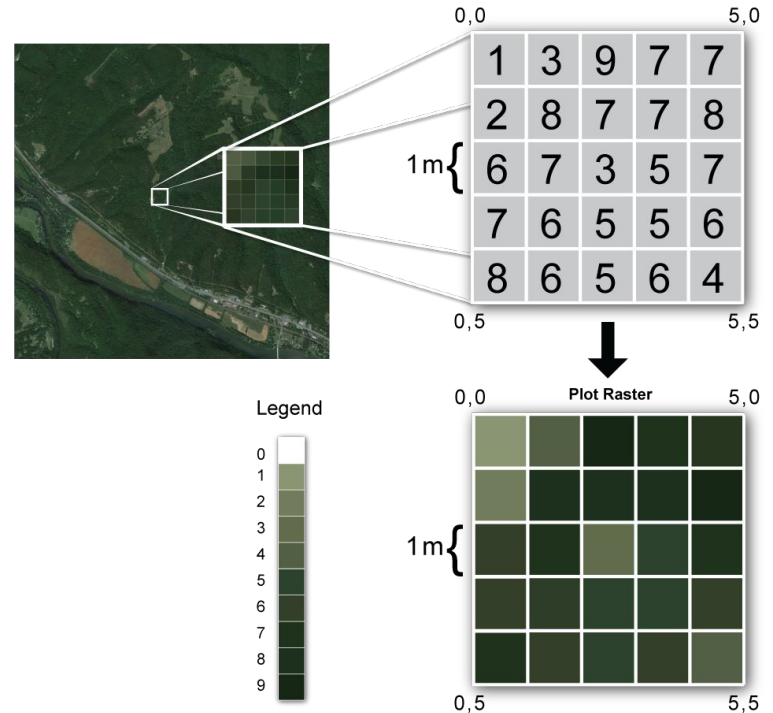


Raster data

Raster data structure

Raster = a grid made of cells (like pixels), which each hold a value or class.

- Satellite imagery is usually made of 3 bands (red, green, blue) and for each cell there is a value for red, for green and for blue.
- Pixels can hold elevation values, or rainfall, or land cover, etc.

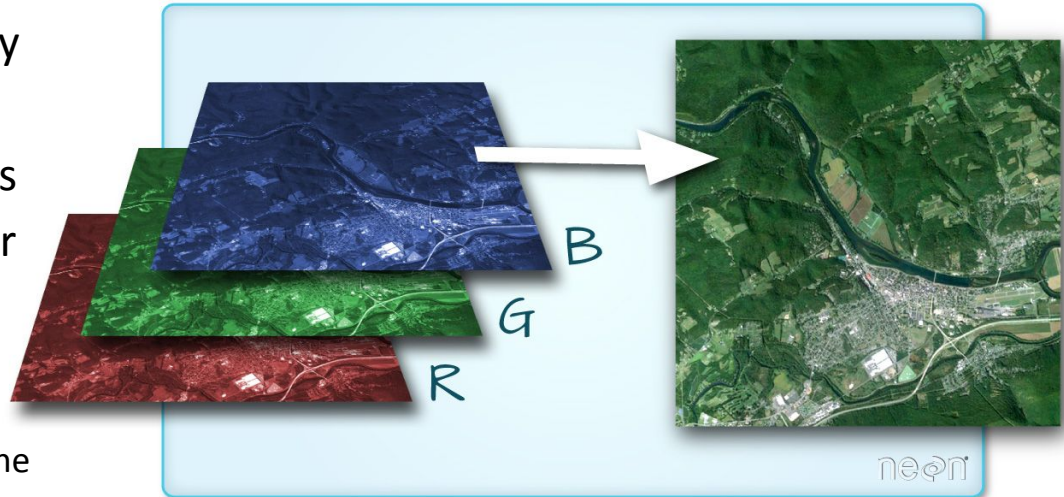


Source: US National Ecological Observatory Network (NEON)

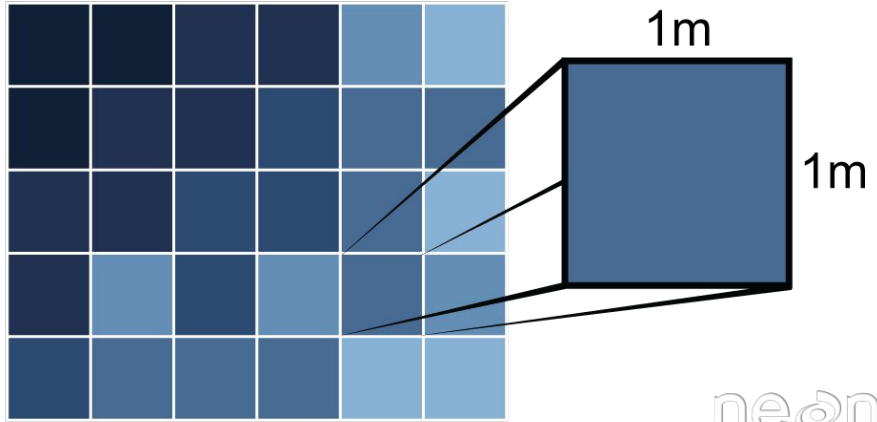
Raster data structure

Multispectral satellite imagery contains 3 bands (red, green, blue) and for each cell there is an intensity* value for red, for green and for blue.

*the solar radiance at each given wavelength that gets reflected from the ground

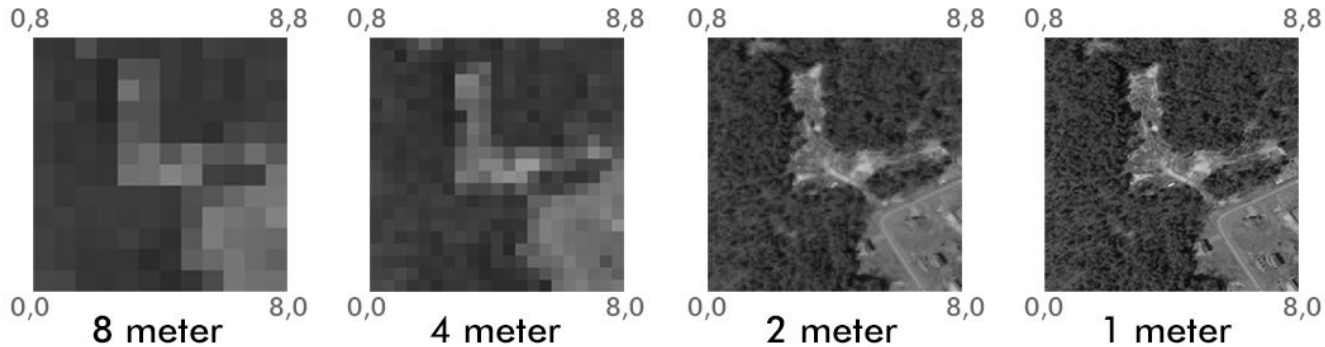


Raster resolution



neon

Raster over the same extent, at 4 different resolutions



Raster formats

All sorts of formats exist! Here is a non-exhaustive list:

- ECW (ER Mapper Enhanced Compression Wavelet), a compressed format to store aerial or satellite imagery ***.ecw**
- GeoTIFF (Geo Tagged Image File Formats) ***.tiff** or ***.tif**, which may come bundled with other files: ***.tfw** to store the raster geolocation, a ***.xml** may be provided to store metadata, an ***.aux** auxiliary file to store projections and a ***.ovr** pyramid file to improve display performance.
- DEM (Digital Elevation Model): ***.dem**, ***.ddf**
- ASCII grid (matrix of float) ***.asc**
- But also ***.hgt**, ***.h4**, ***.hdf**, ***.netCDF**, ***.bil**, ***.hdr** and many many more

Main advantages of raster format

- It's a very simple data structure, and easy to interpret
- You can perform map algebra and this is very fast and powerful.
- You can overlay complex datasets and combine them efficiently.
- It's well suited to represent continuous surfaces (e.g. elevation, temperature, or levels of lead contamination in the soil) and perform surface analysis.

Bonus: you can convert vector into raster. That way you can uniformly store points, lines, polygons, and surfaces in a single matrix (e.g. for land use)

Main limitations of raster format

- With coarse resolution, you can have a very pixelated look to your layer and that makes it difficult to display features such as roads or narrow rivers
- It can get very large and heavy with higher resolution, which uses up processing and storage.



Vector

Vector representation

Vector = Object view = geometries

A coordinate-based data model that represents geographic features as points, lines, and polygons.

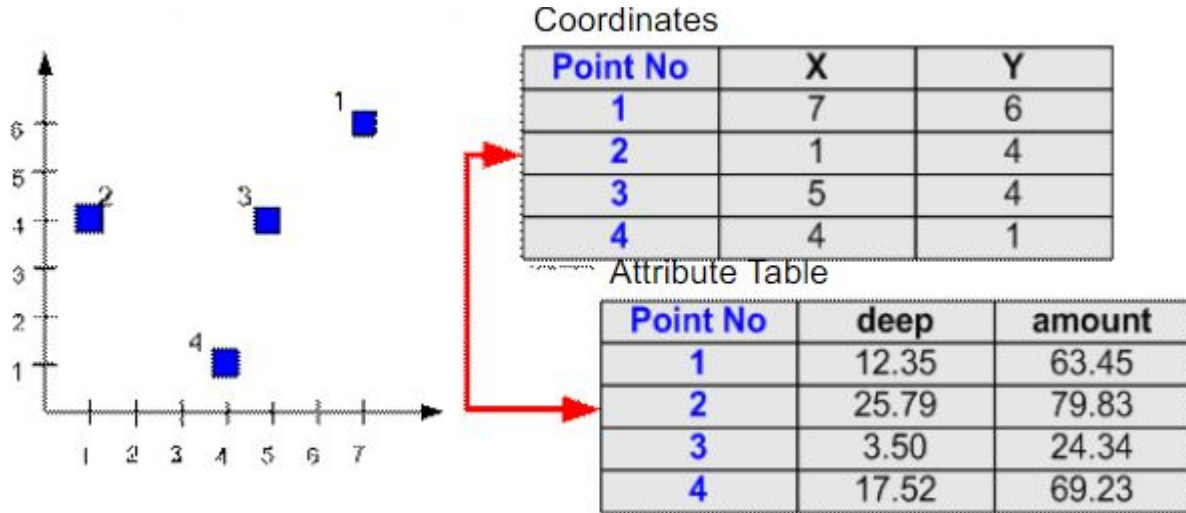
Vector formats

Most common formats:

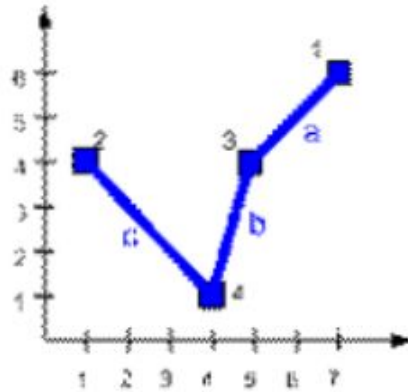
- Shapefiles (careful, it comes in 4 parts!) ***.shp**
- Geopackages (QGIS, open-source) ***.gpkg** / Geodatabases (ESRI ArcGIS) ***.gdb**
- PostgreSQL/PostGIS databases (directly from a database connection)
- Tabular data with some kind of lat/long or Eastings/Northings coordinates (most commonly: ***.csv** files)
- GeoJSON (format commonly used in web development) ***.geojson**

Data you load into QGIS can be saved into a different format (typically, you can create a point layers from a CSV file and save it as a *.shp or a layer in a *.gpkg)

Vector representations: Points



Vector representations: Lines

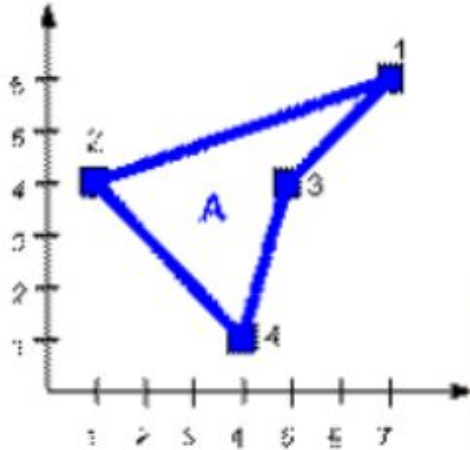


Node No	X	Y
1	7	6
2	1	4
3	5	4
4	4	1

Line	First Node	Last Node
a	1	3
b	3	4
c	4	2

Line	Flow	Capacity
a	960	2200
b	1250	2000
c	1100	2000

Vector representations: Polygons



Node No	X	Y
1	7	6
2	1	4
3	5	4
4	4	1

Polygon	Node sequence
A	1,3,4,2,1

Polygon	Area	Population
A	15.23	12.35

What things represented as Point, Line or Polygon?

It depends on scale (more on next slide) and function.

- **Line:** road, pipeline, water line, rivers, bus route
- **Points:** buildings, post offices, bus stops, hospitals, police stations, wells at a 1:25,000 or 1:10,000 scale.
- **Polygons:** at a scale of 1:1,000 : can be a lake, park, campus

Large scale vs small scale

RF scale (Representative Fraction) such as **1:10,000**
(1cm on the map represents 10,000 cm=100m in real life)

- Large scale maps: neighborhoods, a localized area, small towns, etc.
- Small scale maps: larger geographic area with few details on them. The RF scale of a **small scale map** would have a **much larger number** to the right of the colon such as 1 : 1,000,000.

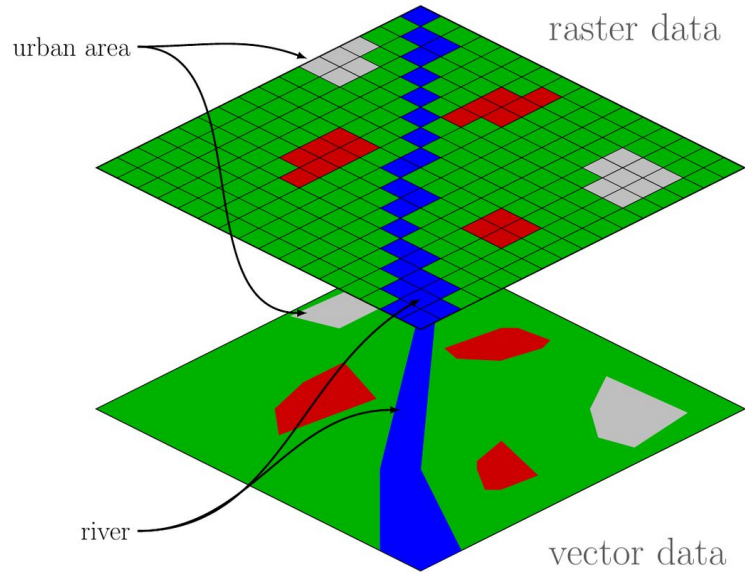
Main advantages of vector format

- Visually, the graphical output of vector data is more suited to display objects (buildings, lakes, administrative regions...).
- The geographic accuracy is also much higher than on a raster grid because you're not tied to the raster resolution.
- You open up the tools that derive from geometry (clipping, buffering etc) and you can also use network analysis tools.

Main limitations of vector format

- It's difficult to store and display continuous data as vectors. It's possible to vectorize continuous data but in many cases it would require substantial generalization, after which your data may not may render your data useless.
- Vector operations such as feature edits or geoprocessing can be very greedy/intensive and take a long time to run

So, Vector or Raster?





3. Data sourcing



Discussion

1. Where would you look for data to use in a research project?
- 

Where do you find data?

- National data stores (France: <https://data.gouv.fr/en/> and <https://geo.data.gouv.fr/en/>, Spain: <https://datos.gob.es/en> , UK: <https://data.gov.uk/> , etc)
- EU Data portal <https://data.europa.eu/euodp/en/data/>
- National Surveying agencies (IGN in France, the Ordnance Survey in the UK etc)
- City-level data stores: Paris Data Store <https://opendata.paris.fr/> or <https://data.iledefrance.fr/explore/>, London DataStore <https://data.london.gov.uk/> , etc.
- See a list of open data portals here: <https://dataportals.org/search>
- Environmental Agencies, Statistics Offices etc
- Administrative boundaries by country: GADM <https://gadm.org/>

Also check out Robin Wilson's list (many great resources listed, but some items are outdated):

<https://freegisdata.rtwilson.com/>

Share, share, share !


These portals can be difficult to navigate.

In Slack, check out **#resources** for data portals already shared by last year's students, and do contribute with the ones you come across!



Discussion

2. How would you make sure this dataset is well suited to answer your research question?



What to look for?

Metadata is crucial to assessing the suitability of a given dataset for your analysis.

The metadata of a dataset documents the who, what, when, where, how, and why of a data resource. It will determine how you use the data, what analysis you can carry out, what level of uncertainty it carries. Metadata used to take the form of marginalia on old paper maps, then it used to be sent by proprietary software companies to the user, and things changed with the growing popularity of open software.

Over the past 15 years, the EU Inspire (<https://inspire-geoportal.ec.europa.eu/>) directive has pushed standardization of data related to the environment across the EU, in order to allow comparison across countries or cities. The problem of the language barrier is still not fully solved, but at least existing data is increasingly standardized and documented.

What to look for?

Important data quality criteria:

- Provenance of the data (lineage)
- Bias
- Completeness
- Currency (how up to date? Like census every 10 years)
- Timeliness (does the data arrive at the right time for you)
- Consistency (shed vs hut.... Is it the same type of object? Are there rules as to what is what?)
- Relevance and fitness for use
- Accuracy (like Ordnance survey has a 200 year history and also provides hundreds of pages of metadata)
- Reliability (trustworthiness)

3. Exploratory Spatial Data Analysis (ESDA)

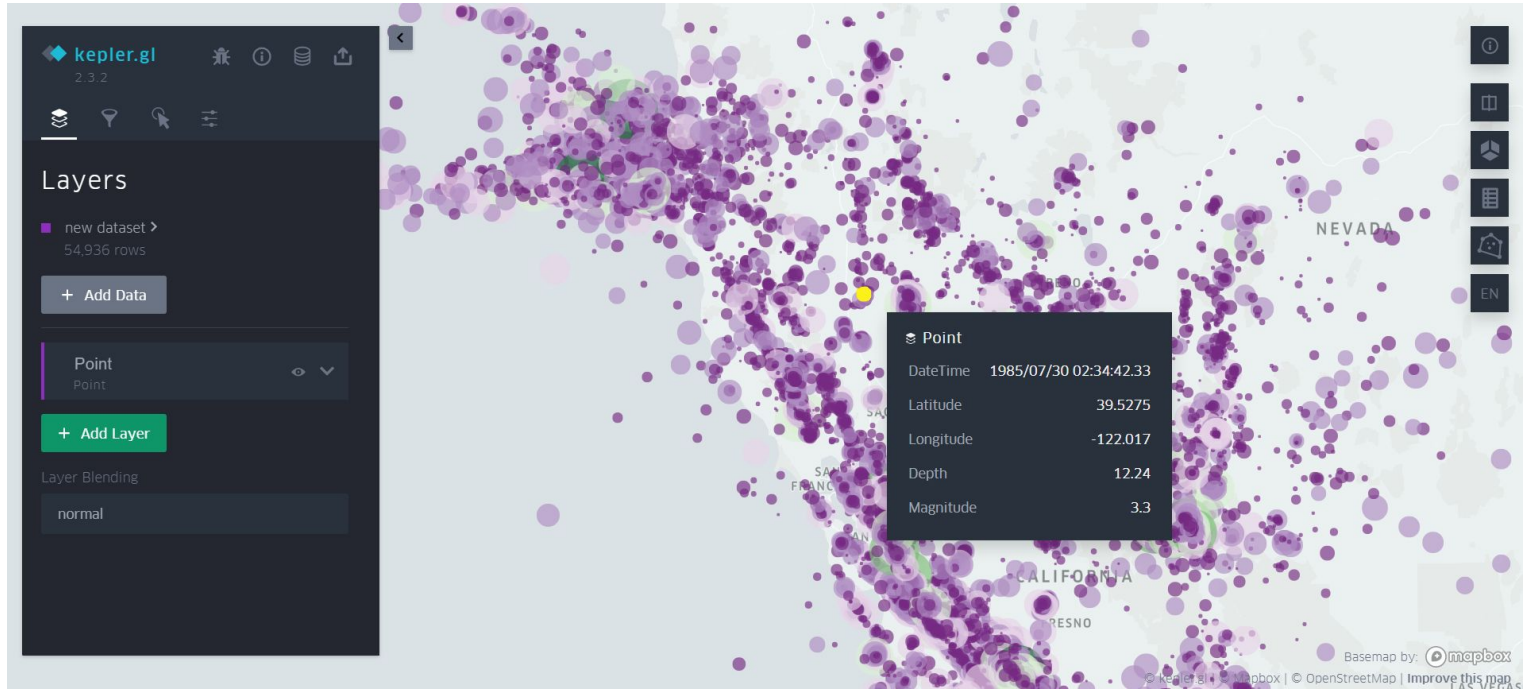




4. Looking at your data (example with kepler.gl)

kepler.gl demo

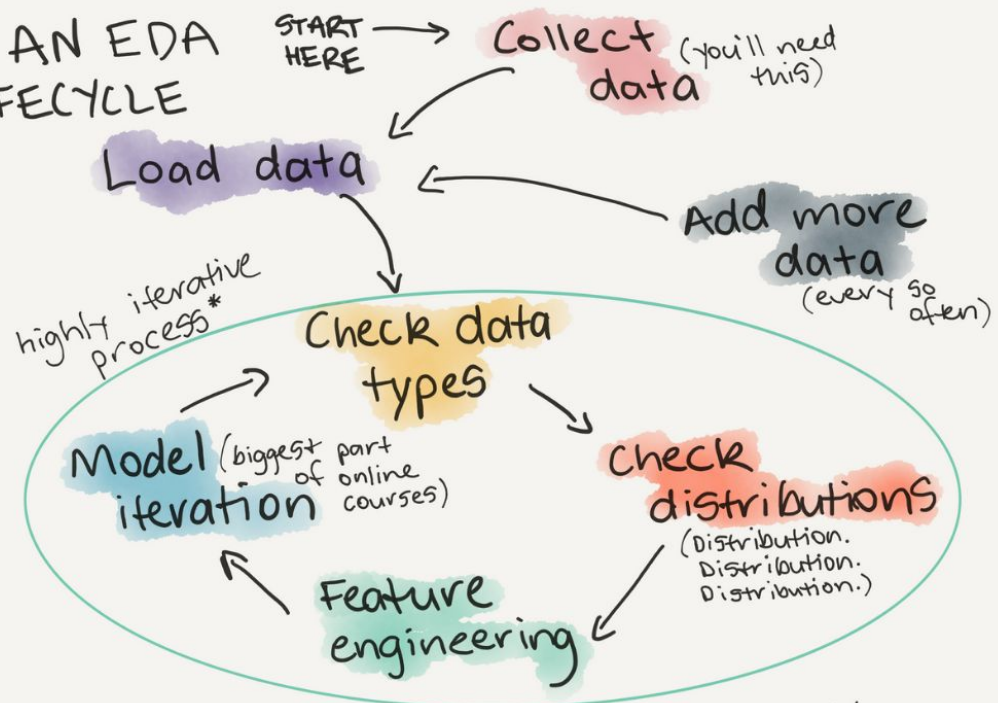
<https://kepler.gl/demo/earthquakes>



4. Refining the research objective and goals



AN EDA LIFECYCLE



* Large potential for Stockholm syndrome to set in



5. Tutorial

Tutorial

In [this week's tutorial](#) you'll learn:

- The best practices for setting up a QGIS project using geopackages.
- How to load data into QGIS
- How to save your project
- How to export data in a different format

Homework

1. Do the [QGIS tutorial](#) on loading data in QGIS and saving it in different formats
2. Play around with other demo datasets in kepler.gl (play with symbology!). Think about the insights you get from this simple visualisation, the research hypotheses that it allows you to formulate, and what research questions it could inform.