

# **Lesson 2: The Devil is in the Data: Collection, representation, and modelling**

**Dr. Kam Tin Seong  
Assoc. Professor of Information Systems**

**School of Computing and Information Systems,  
Singapore Management University**

**2019-01-01 (updated: 2021-08-22)**

# What will you learn from this course?

- Understand how GIS represent real world geography
- Understand the characteristics, advantages, and disadvantages of vector and raster data models
- Understand the basic concepts of georeferencing and geocoding

# Getting to Know Geospatial Data Models

- Basic concept of geographical data
- Basic geospatial data models
- Vector data models
  - Spaghetti data models
  - Topological data models
- Raster data models

# Discrete Objects and Continuous Fields

- Two ways of conceptualizing geographic variation
  - The most fundamental distinction in geographic representation
- Discrete objects
  - The world as a table-top
  - Objects with well-defined boundaries

# Discrete Objects

- Countable, persistent through time, perhaps mobile
- Human-made objects
  - Office buildings, houses, bus stops, etc
- Biological organisms
  - Animals, trees



# Continuous Fields

- Properties that vary continuously over space
  - Value is a function of location
  - Property can be of any attribute type, including direction
- Elevation as the archetype
  - A single value at every point on the Earth's surface
  - The source of metaphor and language
  - Any field can have slope, gradient, peaks, pits



# Basic Spatial Data Models

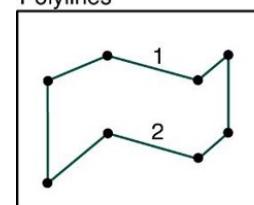
- Vector, implementation of discrete object conceptual model
  - Point, line and polygon representations
  - Widely used in cartography, and network analysis
- Raster, implementation of field conceptual model
  - Array of cells used to represent objects
  - Useful as background maps and for spatial analysis

# Vector Data Models

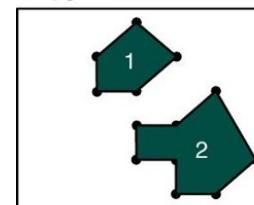
- Used to represent points, lines, and areas
- All are represented using coordinates
  - One per point
- Lines as polylines
  - Straight lines between points
- Areas as polygons
  - Straight lines between points, connecting back to the start
  - Point locations recorded as coordinates

Points	Point number	(x,y) coordinates
+1	1	(2,4)
+2	2	(3,2)
+3	3	(5,3)
+4	4	(6,2)

Polylines	Polyline number	(x,y) coordinates
	1	(1,5) (3,6) (6,5) (7,6)
	2	(1,1) (3,3) (6,2) (7,3)

Polygons	Polygon number	(x,y) coordinates
	1	(2,4) (2,5) (3,6) (4,5) (3,4) (2,4)
	2	(3,2) (3,3) (4,3) (5,4) (6,2) (5,1) (4,1) (4,2) (3,2)

# Vector Database

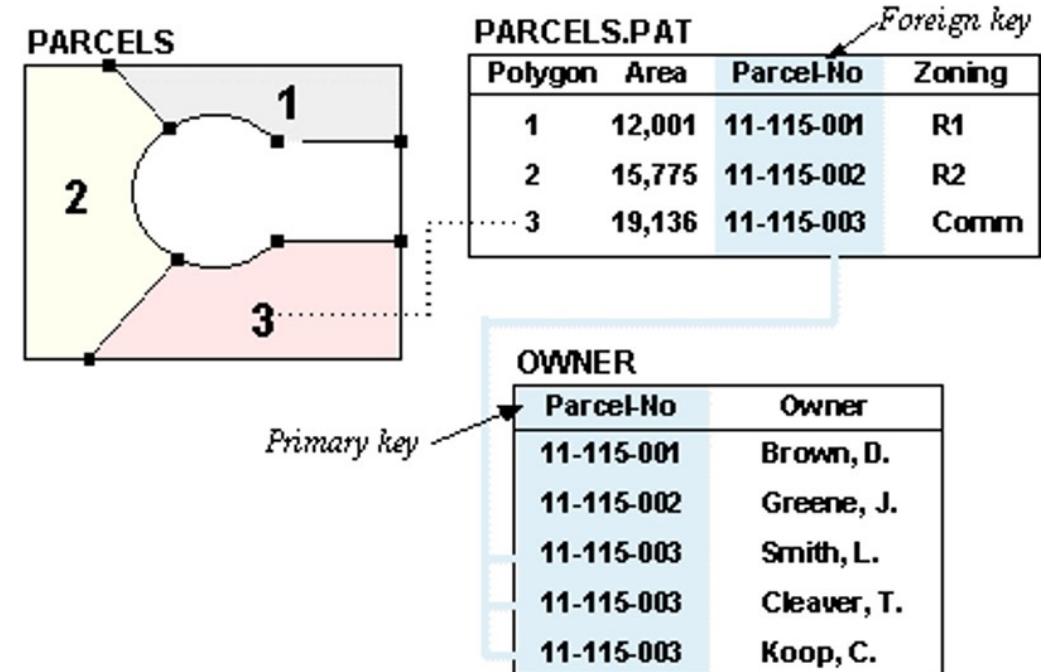
Building footprints are represented by polygon features, road reserves represented by polyline features, and convenient stores are represented by point features



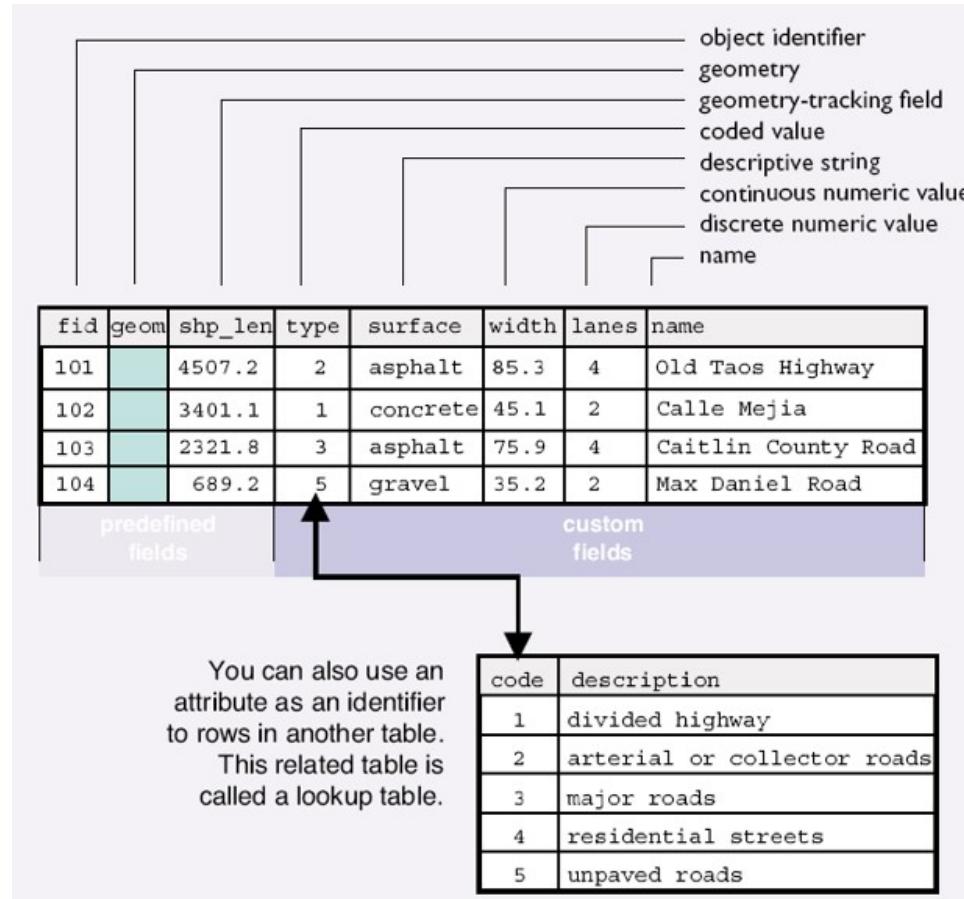
# Georelational Vector Data Model

Stores spatial and attribute data separately in a split system: spatial data ("geo") in graphic fields and attribute data ("relational") in a relational database

## Georelational Vector Data Model



# Attribute Table

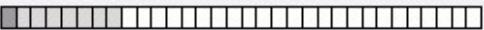


# Types of Attributes: Numerical

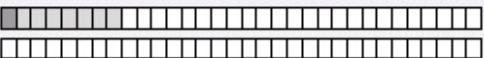
10.0  
2.3  
float  
double  
-4.7

**short integer**

## long integer



A float value contains one sign bit, seven exponent bits, and 24 mantissa bits.



A double value contains one sign bit, seven exponent bits, and 56 mantissa bits.



A short integer value contains one sign bit and 15 binary bits with a range of approximately -32 thousand to 32 thousand.



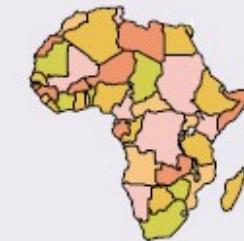
A long integer value contains one sign bit and 31 binary bits with a range of approximately -2 billion to 2 billion.

Graduated  
symbols



Any type of numeric value can be drawn with graduated symbols, which vary in proportion to a value.

## Classified values



A classification is a statistical subdividing of the numeric values of a set of objects. Classified values are drawn with color ramps.

# Other Types of Attributes

<i>Sea</i> <i>Blvd</i> <i>Arkansas</i> <i>Red</i> <i>45th</i>	<table border="1"><tr><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td><td>F</td><td>G</td><td>H</td><td>I</td><td>J</td><td></td><td></td><td></td><td></td></tr></table> <p>Text values contain any number of characters. Each character is stored in a byte (8 bits). All text values in a field have the same number of characters with trailing blanks.</p>	A	B	C	D	E	F	G	H	I	J					<p>Description</p>  <p>Egypt</p> <p>Text shows names and other qualities of features.</p>
A	B	C	D	E	F	G	H	I	J							
<i>12/1/61</i> <i>1/30/00</i> <i>7/16/97</i>	<table border="1"><tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table> <p>Date values are based on a standard time format.</p>														The date value is translated into the current day and time in the local time zone.	
<i>635432</i> <i>object ID</i> <i>689764</i>	<table border="1"><tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table> <p>An object ID value is a long unique identifier generated in geodatabases.</p>														Object IDs are used for database joins and establishing relationships between objects.	
	<table border="1"><tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table> <p>BLOB values contain complex objects like images and video.</p>														BLOB values let you add any kind of multimedia content to your geodatabase tables.	

# Vector Data File Format: shapefile

- A shapefile is a simple, nontopological format for storing the geometric location and attribute information of geographic features.
- Geographic features in a shapefile can be represented by points, lines, or polygons (areas).
- To find out more about shapefile, visit this [link](#).



 polbnda.dbf	DBF File
 polbnda.prj	PRJ File
 polbnda.shp	SHP File
 polbnda.shp	XML File
 polbnda.shx	SHX File

## Limitation of shapefile

- It is proprietary (closed and controlled by ESRI).
- It's a multofile format (.shp,.dbf.,.shx,.prj, encoding, other indexes, etc.) (everybody knows the problem with GIS newbies sending you solely the .shp... 😊).
- Attribute names are limited to 10 characters and only 255 attributes are allowed in .dbf.
- Limited data types. Data types are limited to float, integer, date and text with a maximum 254 characters.
- Unknown character set. There is no way to specify the character set used in the database.
- It's limited to 2GB of file size. Although some tools are able to surpass this limit, they can never exceed 4GB of data.

## Limitation of shapefile

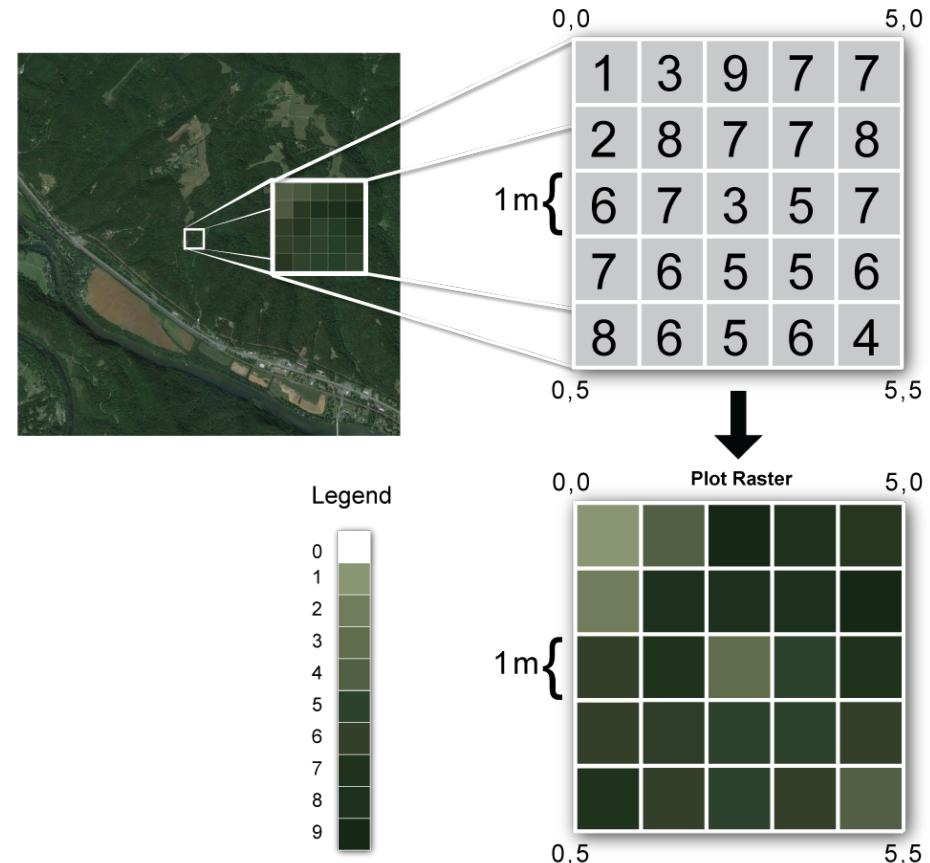
- Uses ESRI's [WKT](#). Can result in inconsistencies.
- Single geometry type per file. There is no way to save mixed geometry features.
- More complicated data structures are impossible to save. It's a "flat table" format.
- There is no way to store 3D data with textures or appearances such as material definitions. There is also no way to store solids or parametric objects.
- Projections definition. They are incompatible or missing.
- Line and polygon geometry type, single or multipart, cannot be reliably determined at the layer level, it must be determined at the individual feature level.

## Other Vector GIS File Formats

- MapInfo TAB format - MapInfo's vector data format using TAB, DAT, ID and MAP files.
- Personal Geodatabase - Esri's closed, integrated vector data storage strategy using Microsoft's Access MDB format
- Keyhole Markup Language (KML) - XML based open standard (by OpenGIS) for GIS data exchange.
- Geography Markup Language (GML) - XML based open standard (by OpenGIS) for GIS data exchange.
- GeoJSON - a lightweight format based on JSON, used by many open source GIS packages.
- TopoJSON, an extension of GeoJSON that encodes topology.

# Raster Data Model

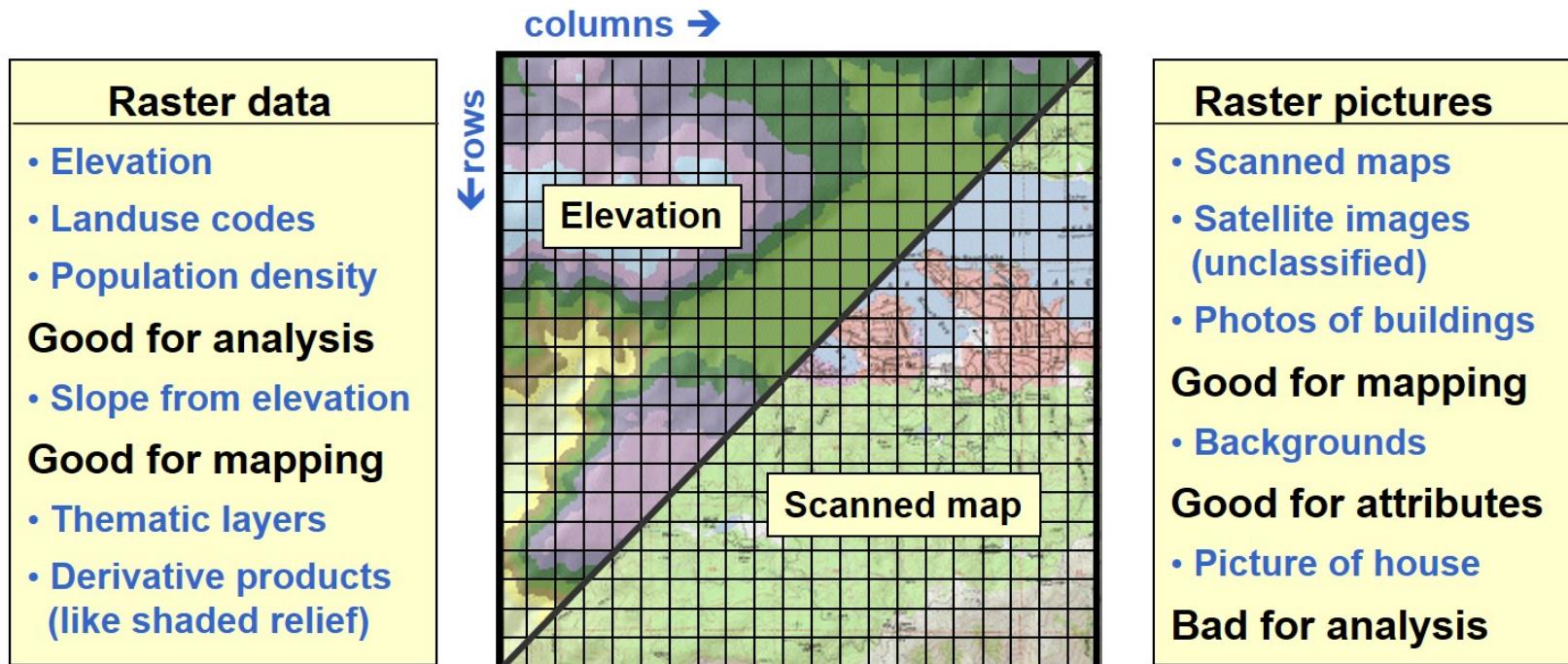
- Divide the world into square cells
- Register the corners to the Earth
- Represent discrete objects as collections of one or more cells
- Represent fields by assigning attribute values to cells
- More commonly used to represent fields than discrete objects
- [What is raster data?](#), ESRI. A good read to learn the basic of raster data:



neon

# Raster Database

- All raster formats are basically the same
- Cells organized in a matrix of rows and columns
- Content is more important than format: data or picture?

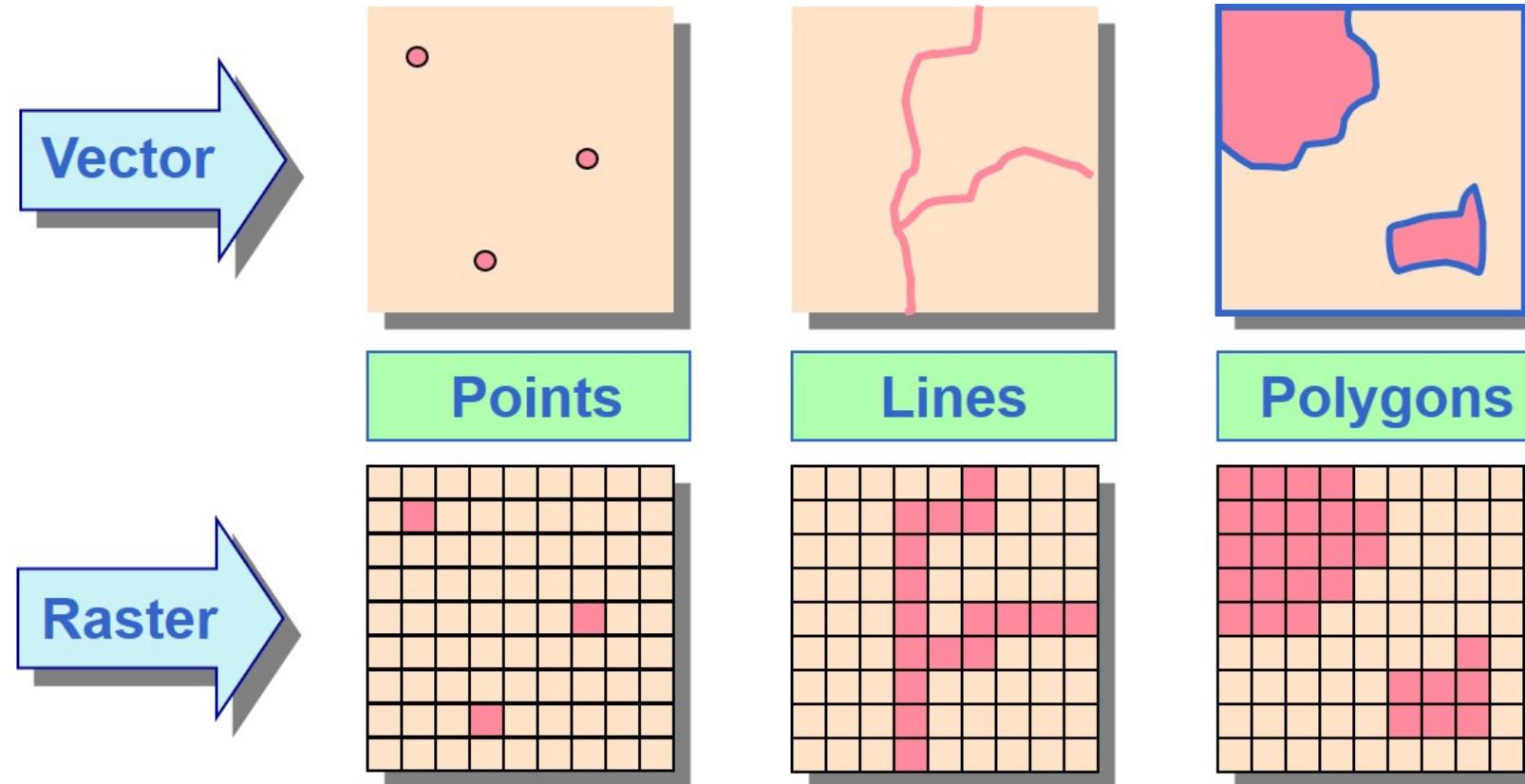


## Characteristics of raster data model

- Pixel size
  - The size of the cell or picture element, defining the level of spatial detail.
  - All variation within pixels is lost.
- Assignment scheme
  - The value of a cell may be an average over the cell, or a total within the cell, or the commonest value in the cell.
  - It may also be the value found at the cell's central point.

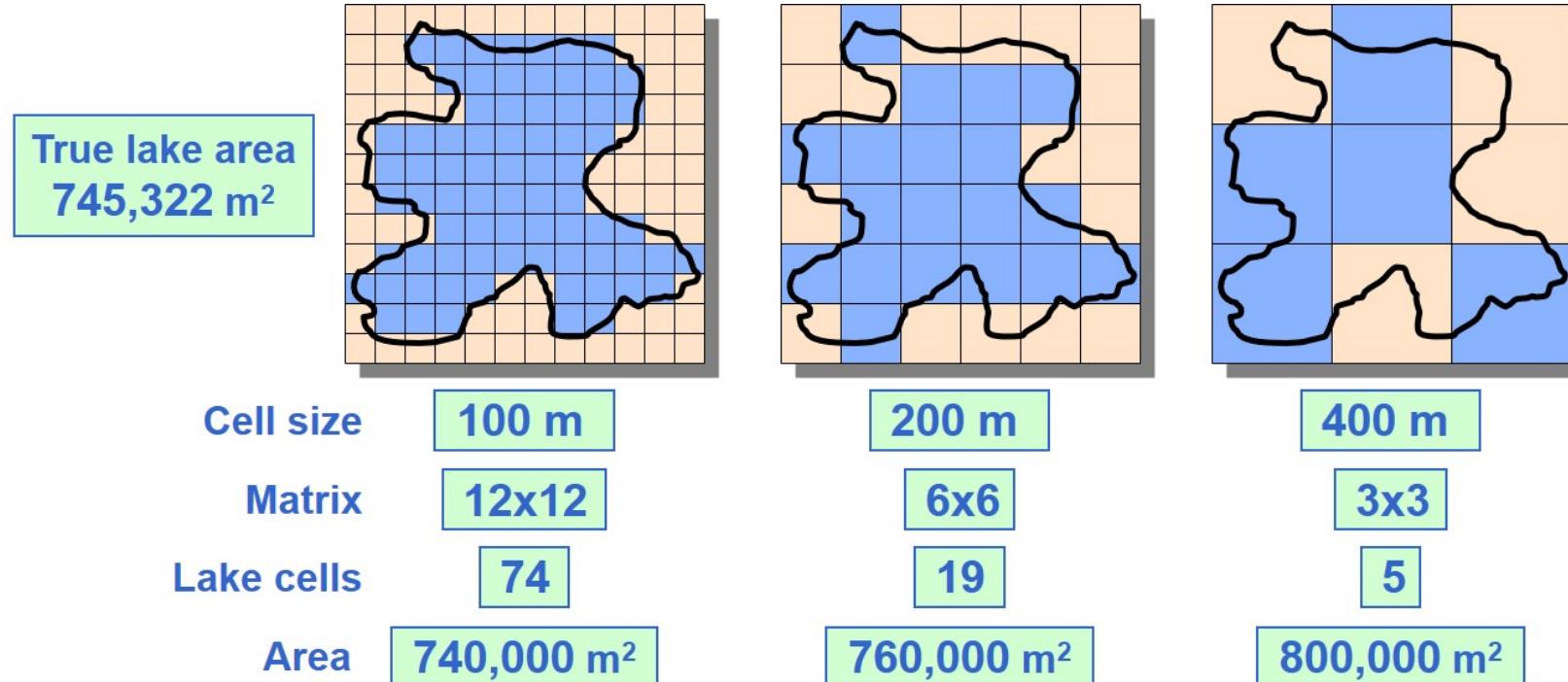
## Features as raster

- Features lose uniqueness with raster representation (a line becomes a collection of cells, not one feature)

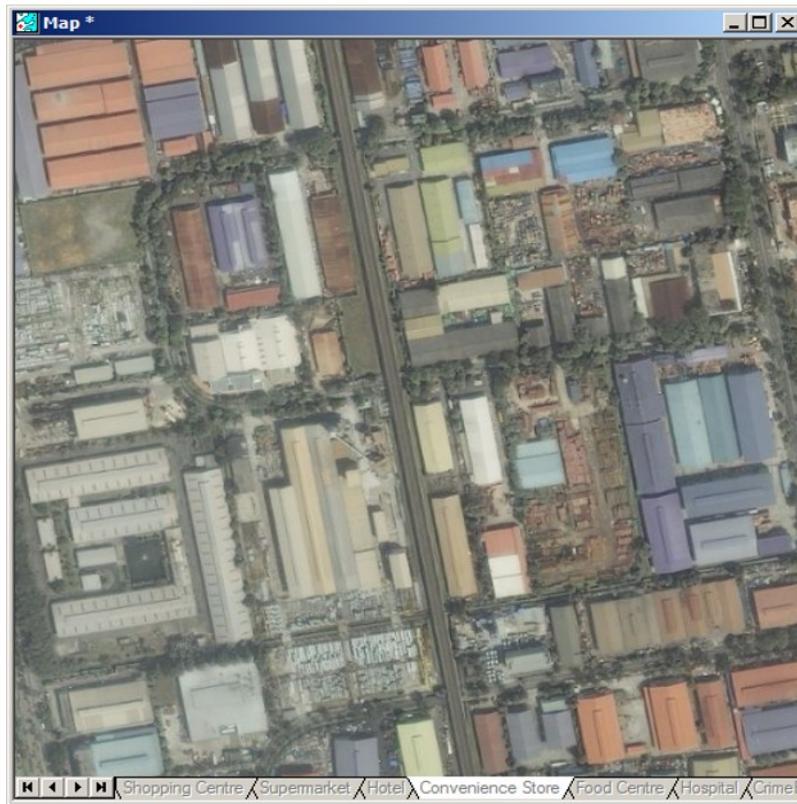


# Raster Resolution

- Rasters always generalize spatial data
  - A function of cell size (smaller cells = higher resolution).
  - Impacts accuracy, processing speed, storage space.



# Raster Resolution and Spatial Details



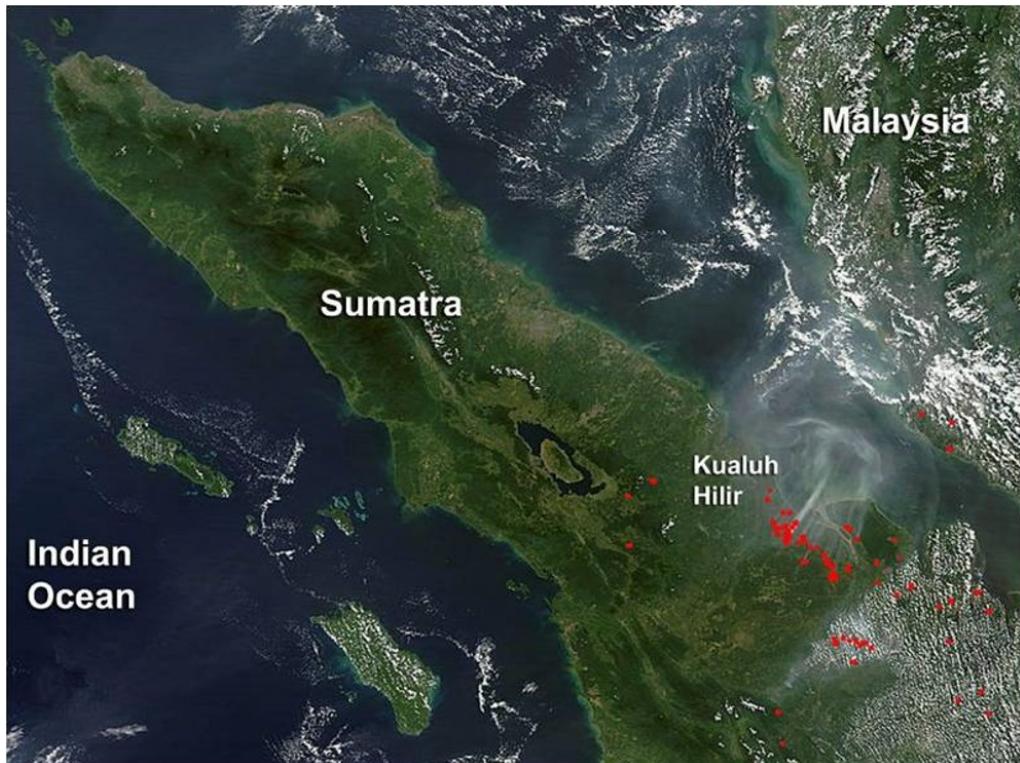
1:5000



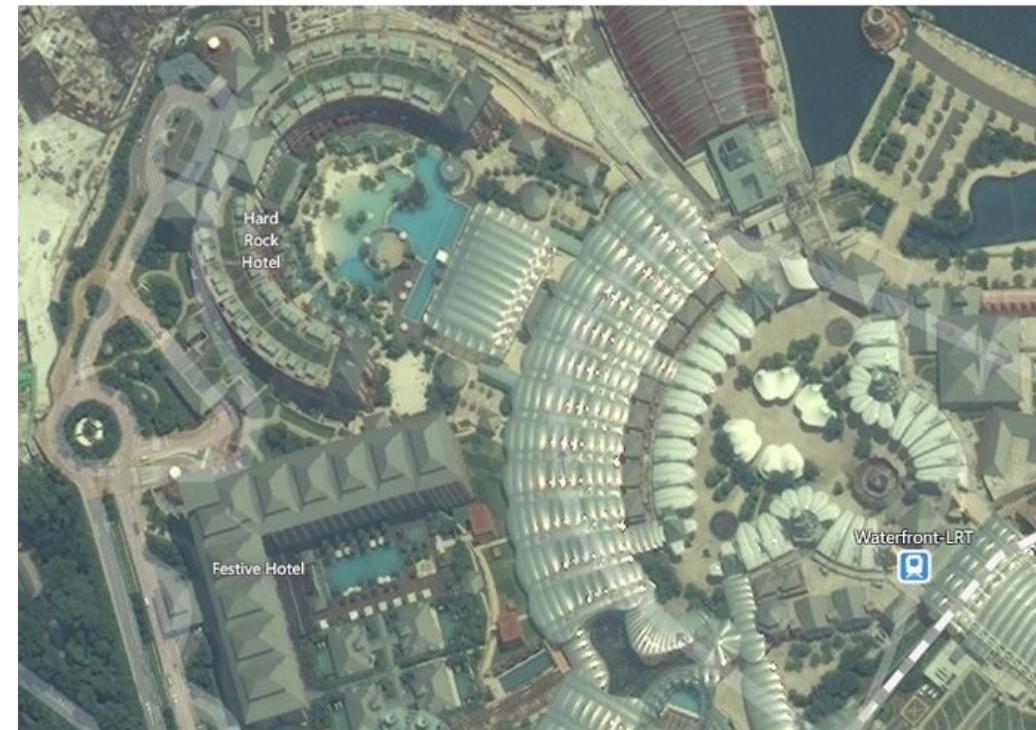
1:500

# Raster Resolution and Spatial Extent

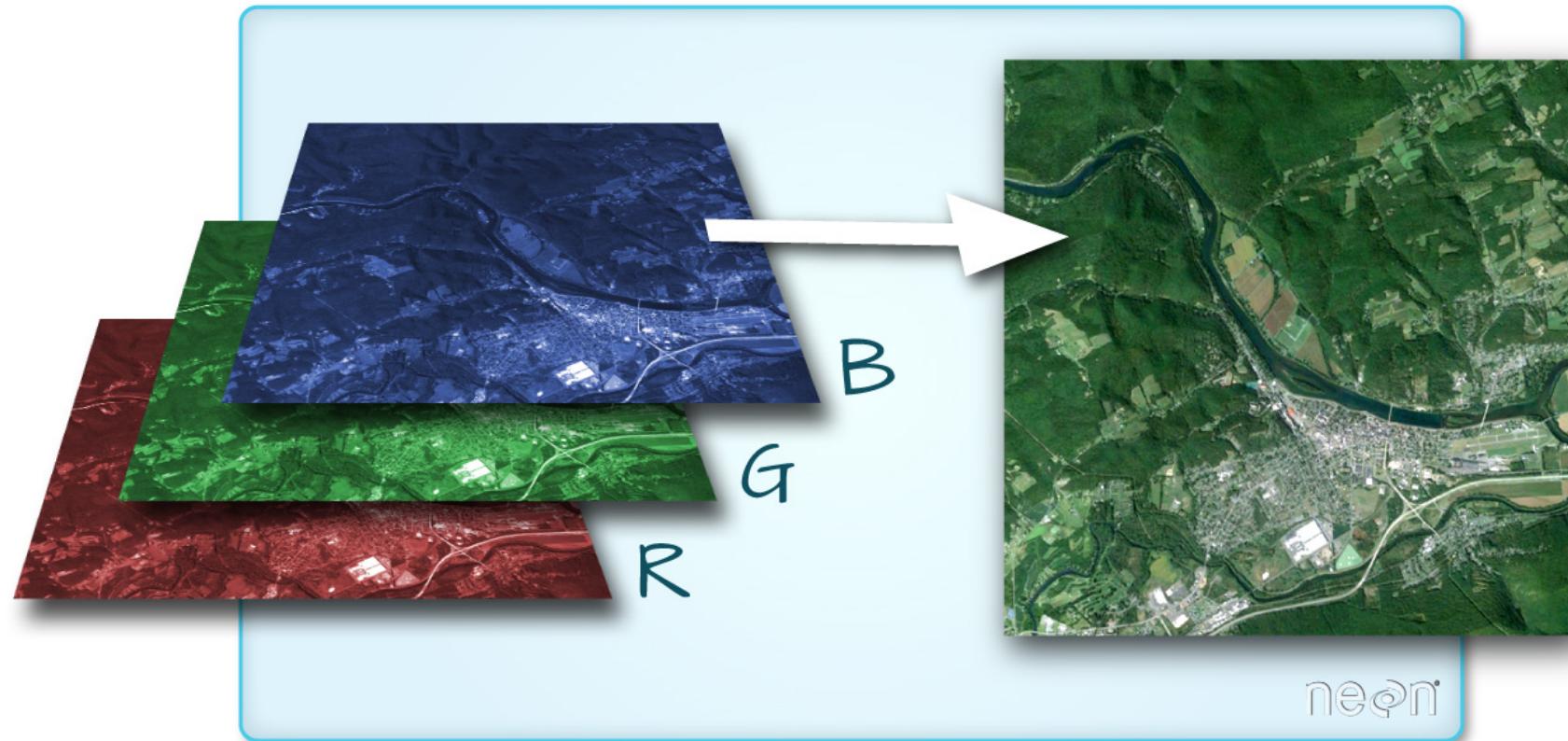
Small-scale satellite image



Large-scale satellite image



# Multi-band Raster Data



## Raster Data Format

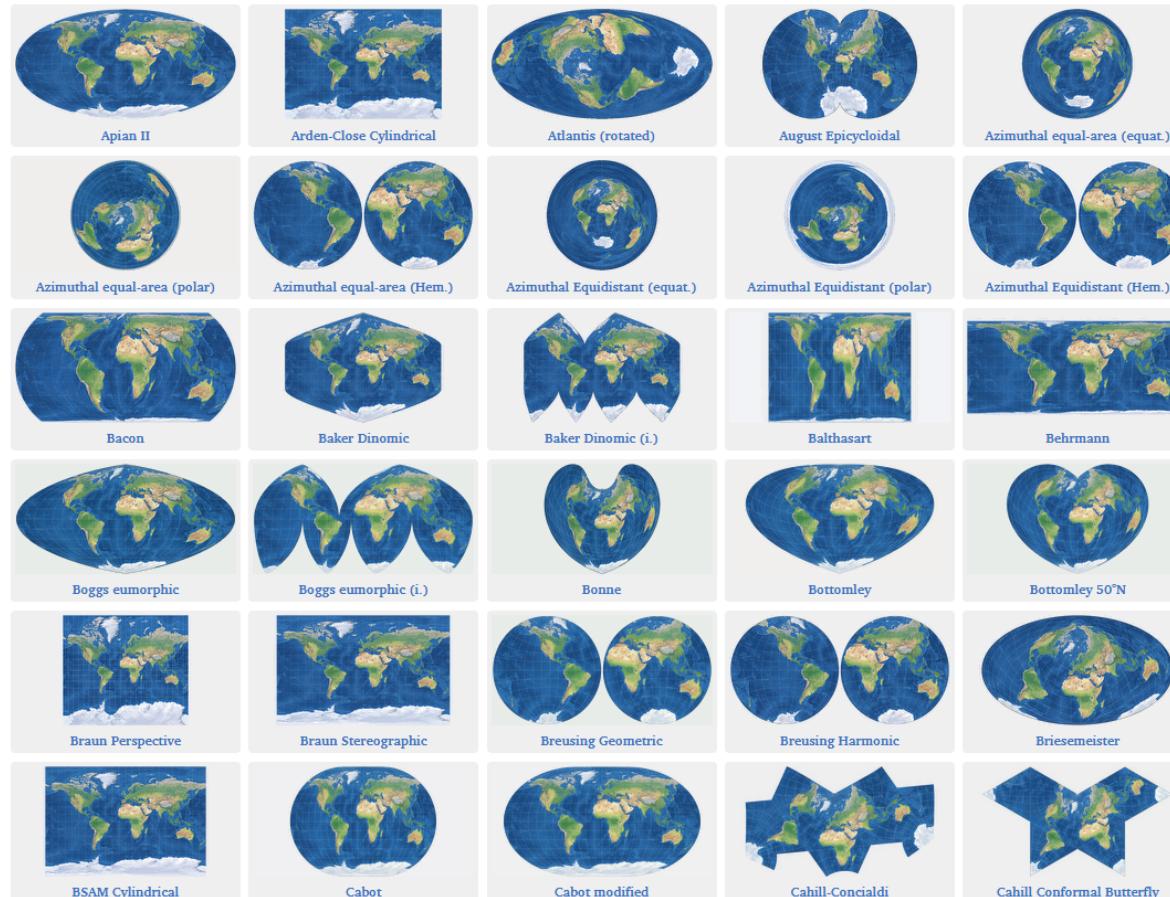
- [GeoTIFF](#): TIFF variant enriched with GIS relevant metadata.
- [JPEG2000](#): Open-source raster format. A compressed format, allows both lossy and lossless compression.
- [BIL, BIP, and BSQ raster files](#): image format linked with satellite derived imagery, namely: BIL (Band Interleaved by Line), BIP (Band Interleaved by Pixel) and BSQ (Band Sequential). To learn more about these three data types, read [Fundamental of raster data](#), ESRI.
- [ADRG](#): National Geospatial-Intelligence Agency (NGA)'s ARC Digitized Raster Graphics.
- [Digital raster graphic \(DRG\)](#): digital scan of a paper USGS topographic map.
- [ESRI grid](#): proprietary binary and metadataless ASCII raster formats used by Esri.
- [IMG](#): ERDAS IMAGINE image file format.
- [ECW](#)): Enhanced Compressed Wavelet (from ERDAS). A compressed wavelet format, often lossy.
- [MrSID](#): Multi-Resolution Seamless Image Database (by Lizardtech). A compressed wavelet format, allows both lossy and lossless compression.

# Unique Properties of GIS Data

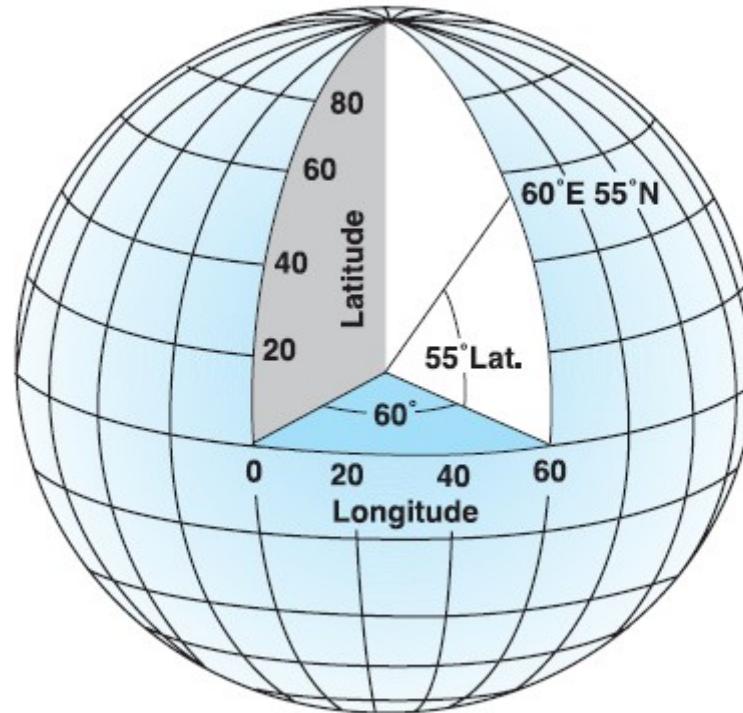
- Geographical reference
- Cartographic generalisation
- GIS data accuracy and uncertainty

# Coordinate Systems and Map Projections

## What is a coordinate system?



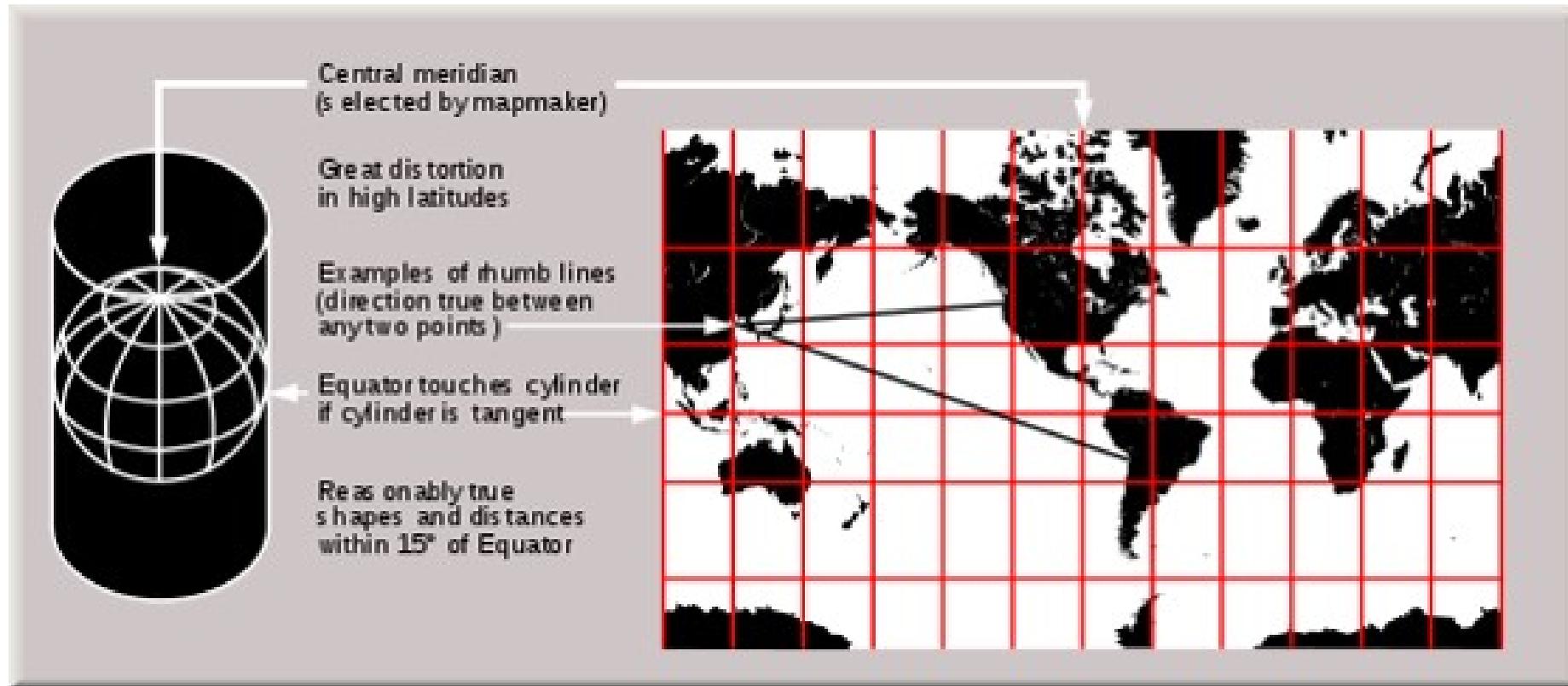
# Geographical Coordinate Systems



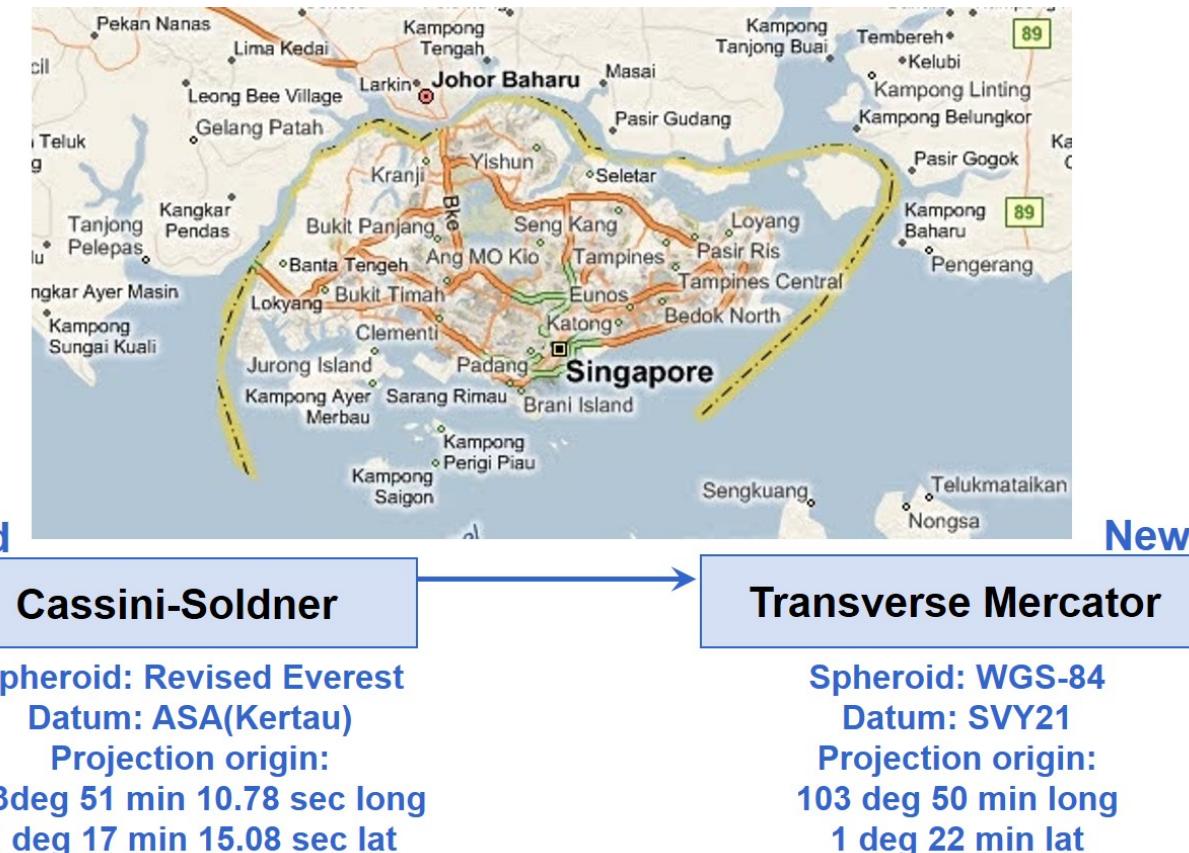
Reference: [http://en.wikipedia.org/wiki/Map\\_projection](http://en.wikipedia.org/wiki/Map_projection)

# Projected Coordinate Systems

- Based on a map projection such as transverse Mercator, Albers equal area, or Robinson.



# Singapore Projected Coordinate System



- [epsg.io](https://epsg.io) provides a comprehensive list of country coordinate systems such as svy21.

# EPSG Reference System

Link to [epsg.io](https://epsg.io/?q=singapore)

The screenshot shows a web browser displaying the epsg.io website at <https://epsg.io/?q=singapore>. The search bar contains the query "singapore". The results page lists three coordinate reference systems:

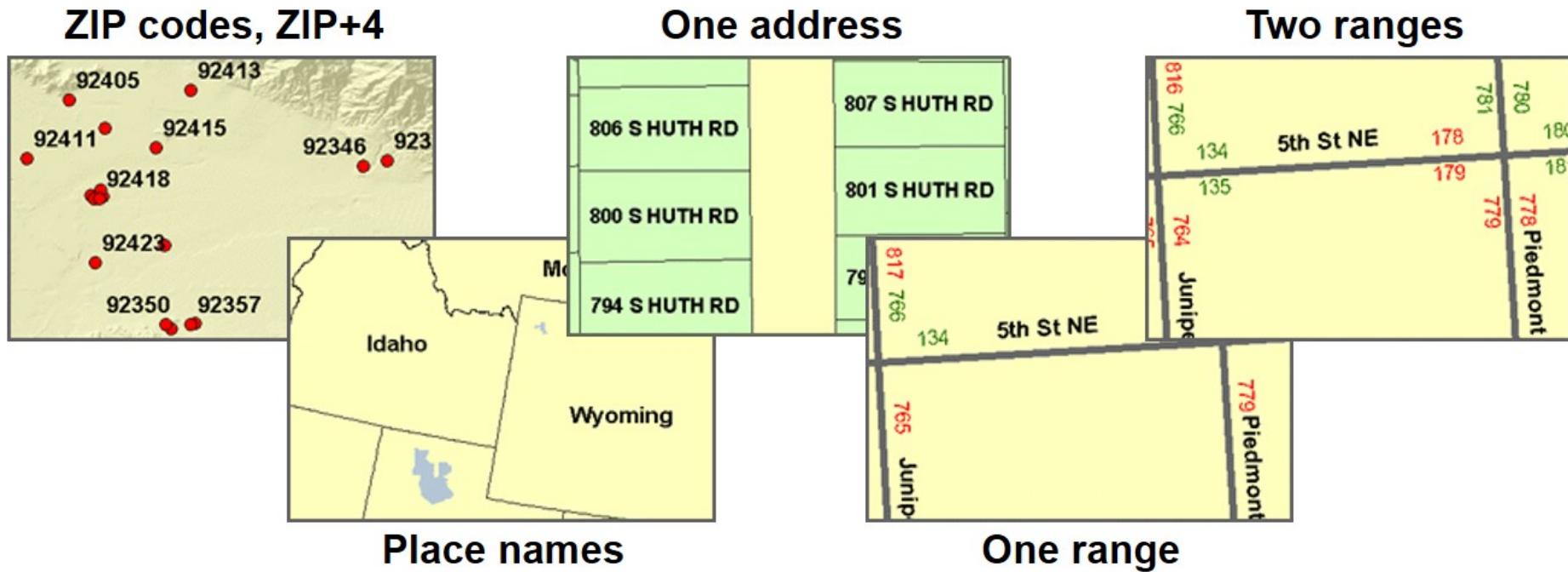
- SVY21 / Singapore TM**  
EPSG:3414 with transformation: 8886  
Area of use: Singapore - onshore and offshore. (accuracy: 1.0)  
[Transform coordinates](#) | [Get position on a map](#)
- Kertau 1968 / Singapore Grid**  
EPSG:24500 with transformation: 1158  
Area of use: Singapore - onshore and offshore. (accuracy: 15.0)  
[Transform coordinates](#) | [Get position on a map](#)
- SVY21 / Singapore TM + SHD height**  
EPSG:6927  
Area of use: Singapore - onshore and offshore.  
[Transform coordinates](#) | [Get position on a map](#)

On the right side, there is a sidebar titled "Type of results" listing various categories and their counts:

- Coordinate reference systems (18)
  - Projected (5)
  - Geodetic (5)
  - Geodetic 3D (2)
  - Geocentric (3)
  - Vertical (1)
  - Compound (2)
- Operation (9)
  - Transformation (4)
  - Compound (1)
  - Conversion (4)
- Datum (4)
  - Vertical (1)
  - Geodetic (3)
- Area (6)

# What is geocoding

- Reference data: features with address attributes Points, lines, polygons.

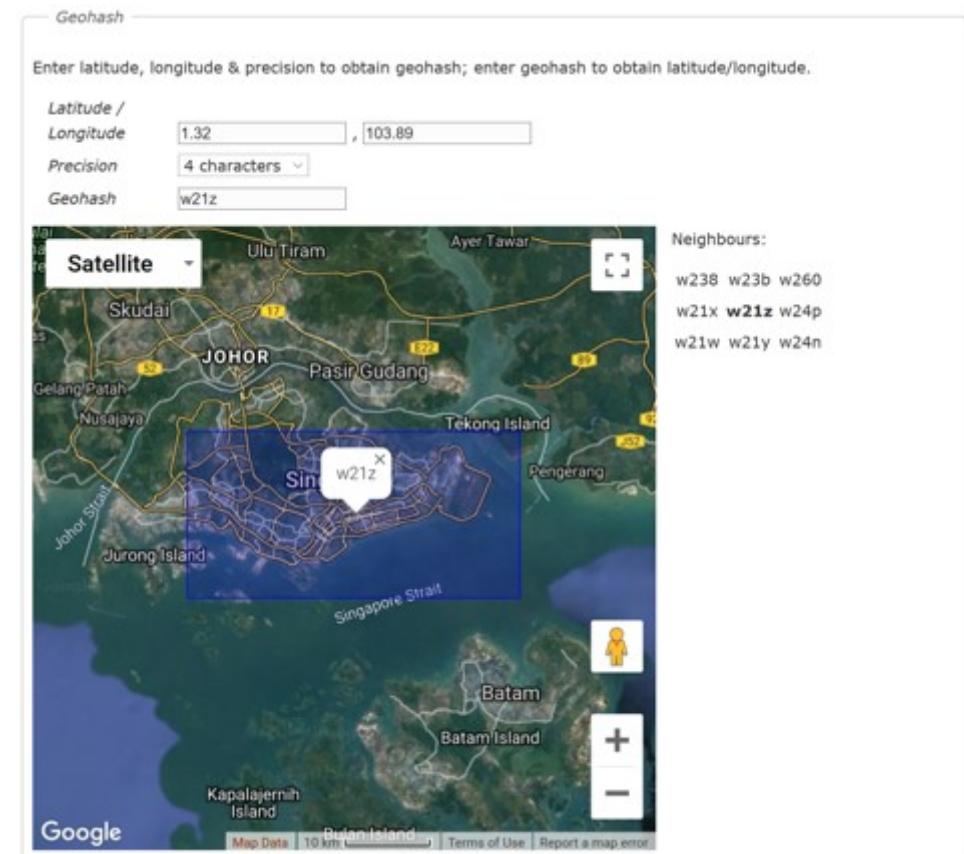


Source: <http://en.wikipedia.org/wiki/Geocoding>

# GeoHash

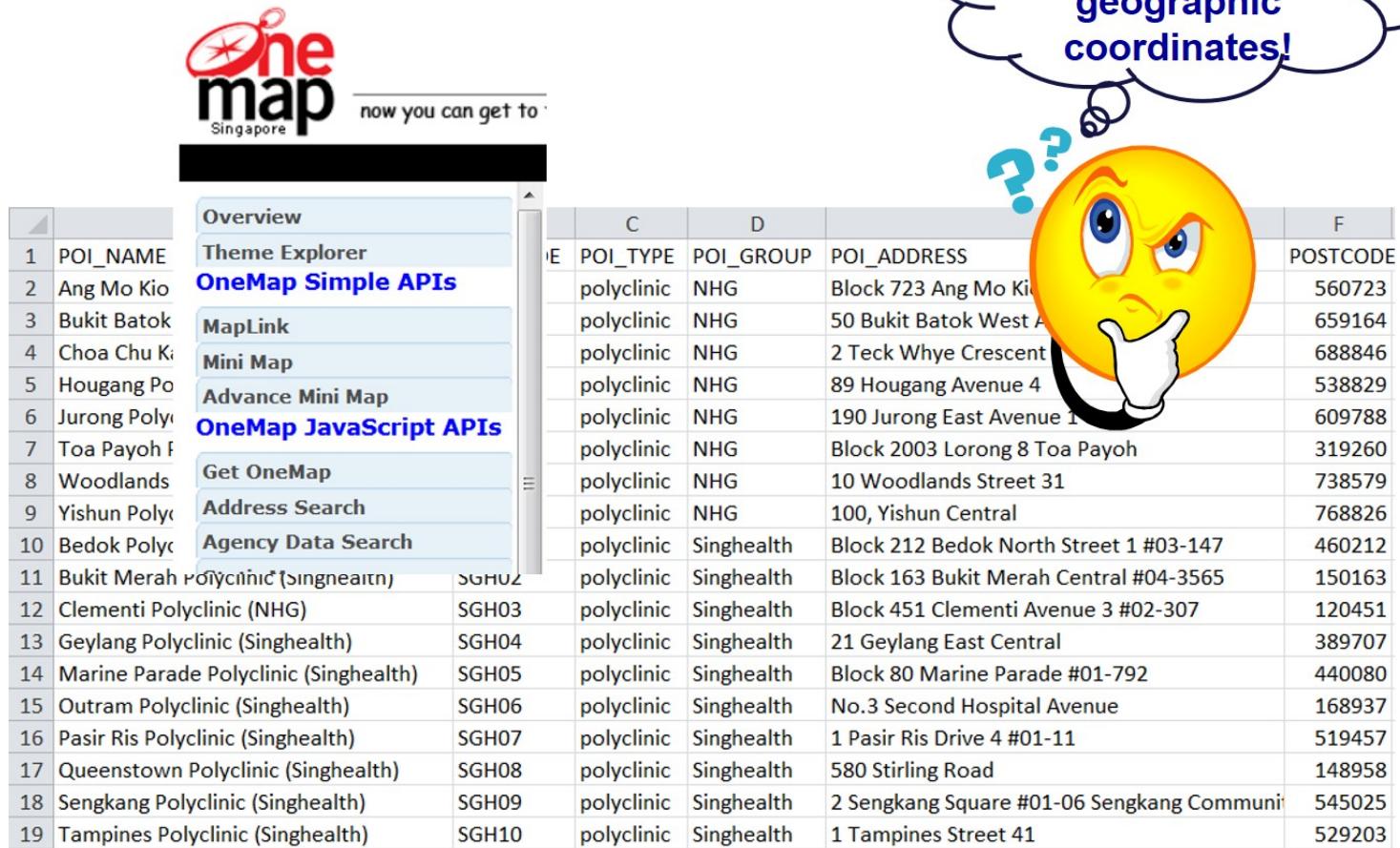
- GeoHash is a public domain geocode system invented in 2008 by Gustavo Niemeyer.
- For more information, visit this [link](#) and

## GeoHash of Singapore



Source: <https://www.movable-type.co.uk/scripts/geohash.html>

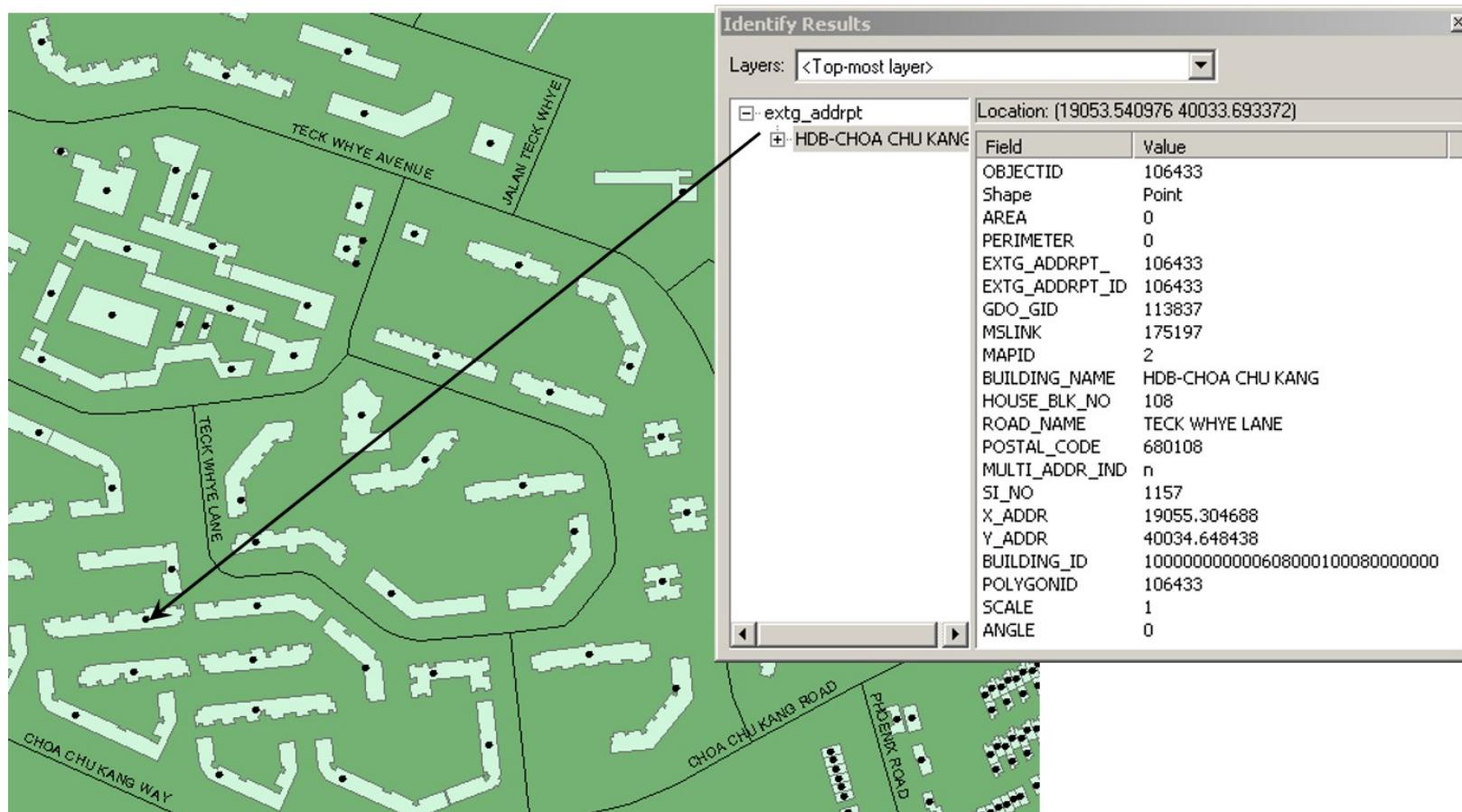
# Geocoding in Singapore: SLA's Address-Point Data



The screenshot shows the OneMap Singapore website interface. On the left, there's a sidebar with links like Overview, Theme Explorer, OneMap Simple APIs, MapLink, Mini Map, Advance Mini Map, and OneMap JavaScript APIs. The main area displays a table of address-point data. A thought bubble above a confused emoji contains the text: "Oh no, these data do not have geographic coordinates!".

	POI_NAME	POI_TYPE	POI_GROUP	POI_ADDRESS	POSTCODE
1	Ang Mo Kio	polyclinic	NHG	Block 723 Ang Mo Kio	560723
2	Bukit Batok	polyclinic	NHG	50 Bukit Batok West Avenue 1	659164
3	Choa Chu Kang	polyclinic	NHG	2 Teck Whye Crescent	688846
4	Hougang Po	polyclinic	NHG	89 Hougang Avenue 4	538829
5	Jurong Polycl	polyclinic	NHG	190 Jurong East Avenue 1	609788
6	Toa Payoh F	polyclinic	NHG	Block 2003 Lorong 8 Toa Payoh	319260
7	Woodlands	polyclinic	NHG	10 Woodlands Street 31	738579
8	Yishun Polyc	polyclinic	NHG	100, Yishun Central	768826
9	Bedok Polyc	polyclinic	Singhealth	Block 212 Bedok North Street 1 #03-147	460212
10	Bukit Merah Poly	polyclinic	Singhealth	Block 163 Bukit Merah Central #04-3565	150163
11	Clementi Polyclinic (NHG)	polyclinic	Singhealth	Block 451 Clementi Avenue 3 #02-307	120451
12	Geylang Polyclinic (Singhealth)	polyclinic	Singhealth	21 Geylang East Central	389707
13	Marine Parade Polyclinic (Singhealth)	polyclinic	Singhealth	Block 80 Marine Parade #01-792	440080
14	Outram Polyclinic (Singhealth)	polyclinic	Singhealth	No.3 Second Hospital Avenue	168937
15	Pasir Ris Polyclinic (Singhealth)	polyclinic	Singhealth	1 Pasir Ris Drive 4 #01-11	519457
16	Queenstown Polyclinic (Singhealth)	polyclinic	Singhealth	580 Stirling Road	148958
17	Sengkang Polyclinic (Singhealth)	polyclinic	Singhealth	2 Sengkang Square #01-06 Sengkang Communi	545025
18	Tampines Polyclinic (Singhealth)	polyclinic	Singhealth	1 Tampines Street 41	529203
19					

# Structure of SLA's Address-Point data



# SLA Onemap Geocoding API

The screenshot shows the OneMap Singapore website. The main navigation bar includes links for Search, Introduction, Basemap, Authentication Service (POI), OneMap REST APIs, Coordinates Converters, Themes, Planning Area, Population Query, Routing Service, and Static Map. The 'OneMap REST APIs' link is highlighted with an orange box. Below the navigation, there's a sidebar with a search icon and the text 'Search'. The main content area is titled 'OneMap REST APIs' and has a sub-section titled 'Search'. It describes the API as providing address data search results with latitude, longitude, and x, y coordinates. A note states that each page of JSON response is restricted to a maximum of 10 results. A table titled 'Parameters' lists five variables: 'searchVal' (Required, Keywords entered by user), 'returnGeom' (Y/N, Checks if user wants to return geometry), 'getAddrDetails' (Y/N, Checks if user wants to return address details for a point), and 'pageNum' (Optional, Specifies the page to retrieve search results from).

## Usage:

```
/commonapi/search?searchVal={SearchText}&returnGeom={Y/N}&getAddrDetails={Y/N}&pageNum={PageNumber}
```

## Examples:

### With optional variables:

```
https://developers.onemap.sg/commonapi/search?searchVal=revenue&returnGeom=Y&getAddrDetails=Y&pageNum=1
```

## Sample Response(Only 2 Results):

```
{
  "found":5,
  "totalNumPages":1,
  "pageNum":1,
  "results":[
    {
      "SEARCHVAL":"INLAND REVENUE AUTHORITY OF SINGAPORE (IRAS)",
      "BLK_NO":"55",
      "ROAD_NAME":"NEWTON ROAD",
      "BUILDING":"INLAND REVENUE AUTHORITY OF SINGAPORE (IRAS)",
      "ADDRESS":"55 NEWTON ROAD, SINGAPORE 307987",
      "POSTAL":"307987",
      "X":28983.7537272647,
      "Y":33554.4361084122,
      "LATITUDE":"1.31972890510723",
      "LONGITUDE":"103.842158118267",
      "LONGITUDE":"103.842158118267"
    },
    ...
  ]
},
```

# Geocoding options for QGIS users

- Geocode Tools of [MMQGIS](#) plugin.



# Sources of GIS data

- Field surveying
  - Conventional land surveying
  - GPS surveying
- Digitising
- Remote sensing
  - Airborn
  - Satellite
- Digital data
  - Internet map services
  - Open Data.gov

# Field surveying

## Land surveying



## GPS surveying



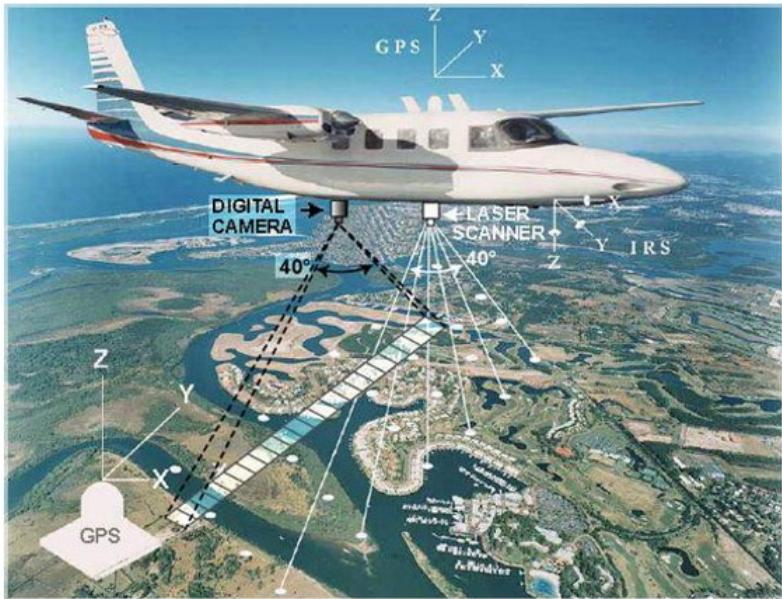
## Mobile mapping

# Digitising

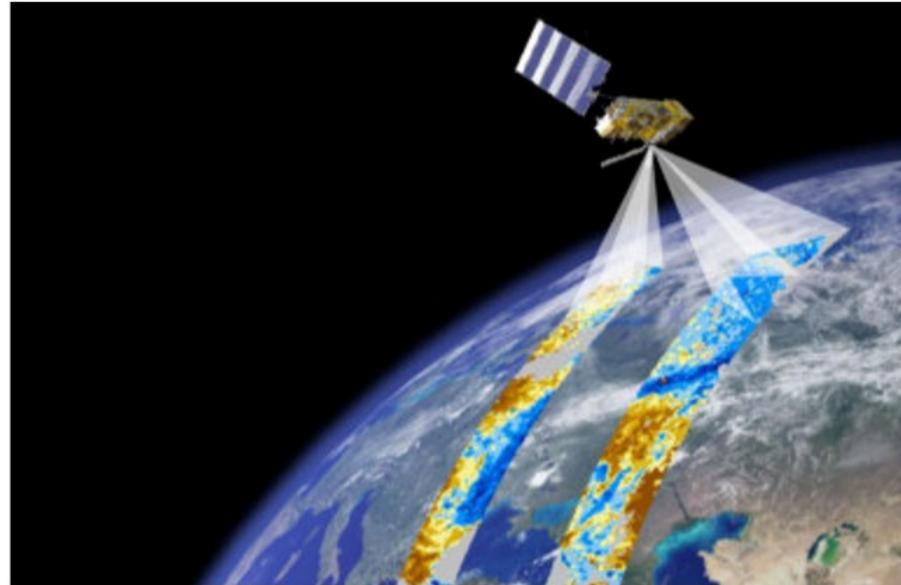


# Remotely Sensed Data

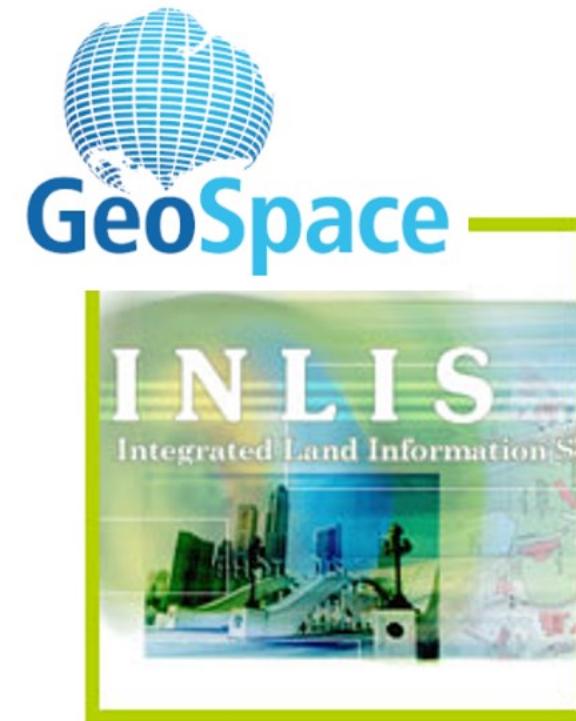
## Airborne



## Satellite



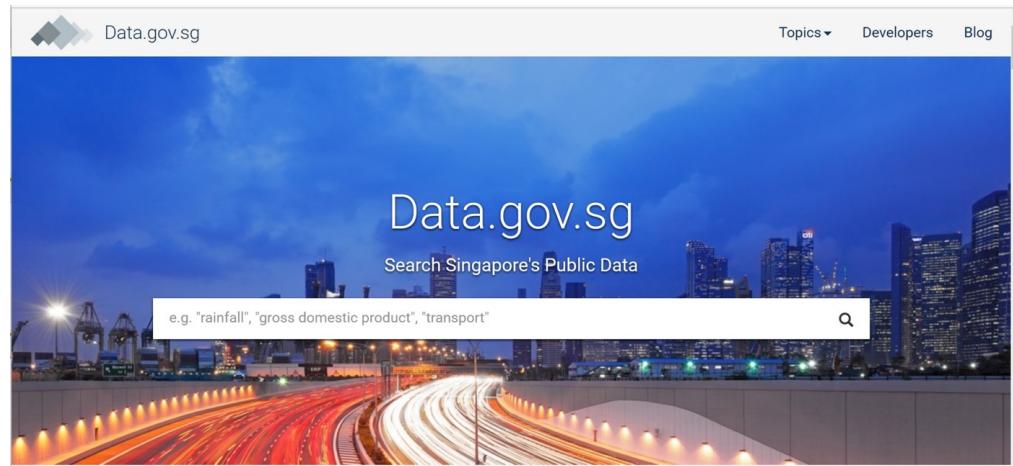
# Geospatial Digital Data from SLA



Reference: <http://www.asiageospatialforum.org/2011/proceeding/pps/leekim.pdf>

# Geospatial Digital Data from Open Gov

data.gov.sg



LTA DataMall



# OpenStreetMap (OSM)

The screenshot shows the OpenStreetMap website interface. At the top, there are navigation links for 'View' and 'Edit', and a 'History' section. On the right, there are 'log in' and 'sign up' buttons. The main area displays a detailed map of Singapore and parts of Malaysia, including Johor Bahru and surrounding areas. The map is overlaid with various data layers, including road networks, place names, and administrative boundaries. A large red banner with white text 'The poor's man alternative!' is prominently displayed across the bottom right of the map area. On the left side of the page, there is a sidebar with a magnifying glass icon and the text 'OpenStreetMap The Free Wiki World Map'. Below this, there is a search bar with placeholder text 'examples: "Alkmaar", "Regent Street, Cambridge", "CB2 5AQ", or "post offices near Lünen" more examples... Where am I?'. Further down, there is a message about the map being free and created by users, along with links to download and use it under an open license, and to create a user account. At the bottom of the sidebar, there are links for 'Help', 'Help Centre', 'Documentation', and 'Community'.

- To learn more about OSM Singapore, visit this [link](#).
- To obtain the latest OSM extract for Singapore, visit this [link](#)

# Open Global Digital Data

Natural Earth



GDAM: Global Administration Boundary Maps

