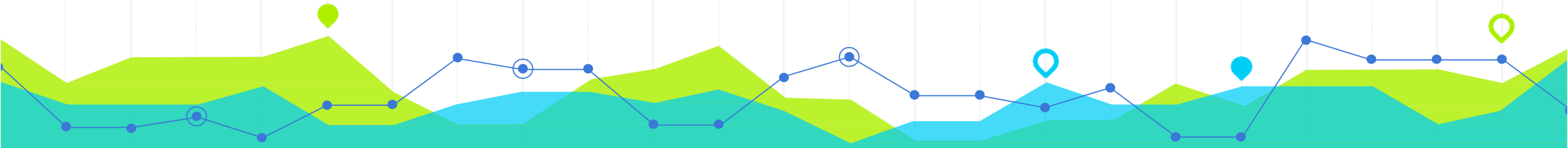
Introduction to Big Data and Machine Learning

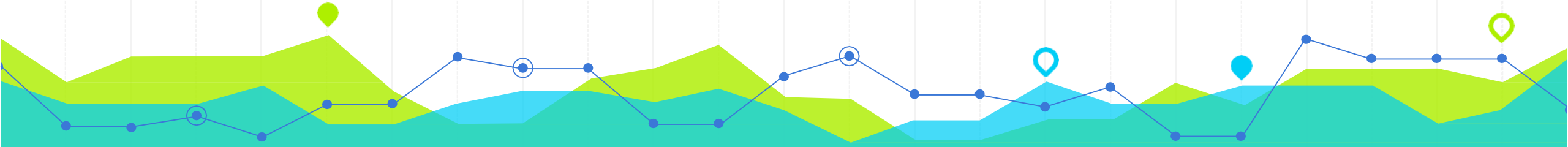


Spatial Data Analysis

Wenzheng Li,

OUTLINE

- Review – first week
- GIS Basics
- Spatial Analysis
- Mapping



Review 0

Day 2: Python Basics

1. Variables and data type

1.1 Creating a variable
1.2 Data types

2. Operators

2.1 Arithmetic operators
2.2 Comparison operators
2.3 Logical operators

3. List

3.1 Defining a list
3.2 List concatenation
3.3 Subscript indices and slices (IMPORTANT)

4. String

5. Dictionary

6. if statement

7. for-loop

7.1 Control statements of a loop: continue and break (Optional)
7.2 The nested loop (loop inside another loop)

8. Defining functions (Optional)

Day 3: Data Management

- Pandas provides two new data types—Series and Dataframe.
  - ***Dataframe****:* Three components—columns, rows, and index
  - ***Series:*** represents a single column of data (one dimension). Think of it as a Pandas-type list.

	HouseID	CommunityID	TotalPrice	TransYear	Bedroom
0	BJFT84326414	1544	1400010.56	2012	2
1	BJCP84958845	2606	1800066.00	2012	3
2	BJDX84905788	2264	1350038.34	2012	2
3	BJFT00386624	3621	1800006.91	2012	2
4	BJCY84713854	1127	1970019.58	2012	1

5 rows × 30 columns

Day 3: Data Management

- A ***label***: one name in the column list or an index in the row index (the column at far left).
- A ***position***: the corresponding position of column name or index in a sequence, starting from zero.

Label and Position

df

		0	1	2	3	4	5	6	7	8	9	
		HouseID	CommunityID	TotalPrice	TransYear	Bedroom	Livingroom	Bathroom	Size	FloorLevel	WinSouth	...
0	0	BJCP85139027	1735	2080032.42	2012	3	2	2	124.62	3	1	...
1	1	BJCY84814525	2023	1440023.27	2012	1	1	1	48.37	4	1	...
2	2	BJHD61617745									1	...
3	3	BJCY00382544									1	...
4	4	BJCY84554915									1	...

Column **Positions**

Column names (**Labels**)

Row **Positions**

Row index (**Labels**)

Task: select the first 2 rows and the first 2 columns

df.loc[,]

row **Labels** col **Labels**

df.iloc[,]

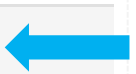
row **Positions** col **Positions**

df.loc[0:1 , ['HouseID' , 'CommunityID']]

df.iloc[0:2 , 0:2]

Day 3: Data Management

```
df_2012["Dist2Subway"] <= 1500
```



Filtering DataFrames

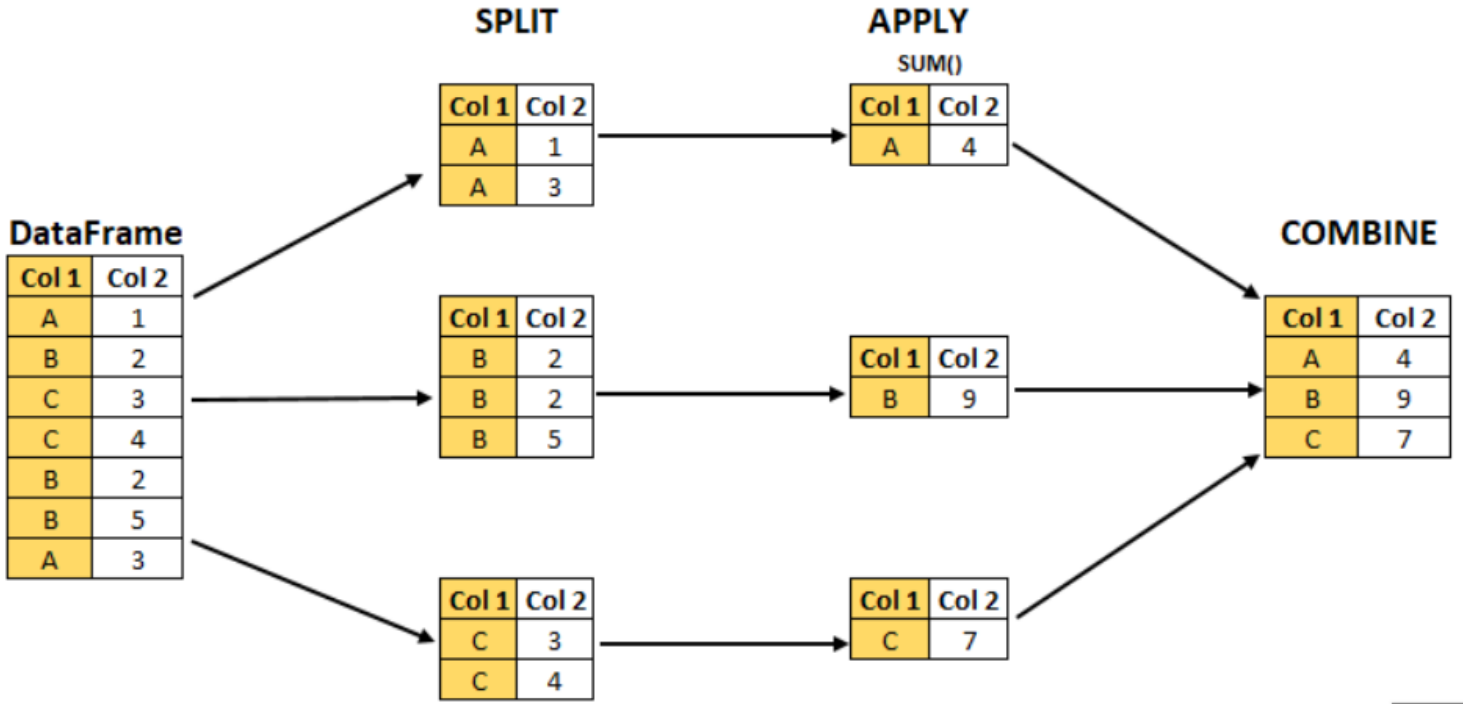
```
0        True
1       False
2        True
3        True
4        True
       ...
```

- `df.loc[df["Dist2Subway"] <= 1500, : ]`:
  - step1, `df["Dist2Subway"] <= 1500` return a series with values of ***False*** or ***True*** (boolean type);
  - step2, it is enclosed by `df.loc[]` and can return a subset of the candidate rows
  - step3, assign the returned DataFrame to a new dataframe called `df_subway`

Day 3: Data Management

By “group by” we are referring to a process involving one or more of the following steps:

- **Splitting** the data into groups based on some criteria.
- **Applying** a function to each group independently.
- **Combining** the results into a data structure.



Day 4: Data Management(II)

- Each column/row in a Pandas (and GeoPandas) DataFrame has a data type, called *dtype* attribute.

Pandas dtype	Python type	Usage
object	str or mixed	Text or mixed numeric and non-numeric values
int64	int	Integer numbers
float64	float	Floating point numbers
bool	bool	True/False values
datetime64	NA	Date and time values
timedelta[ns]	NA	Differences between two datetimes
category	NA	Finite list of text values

Day 4: Data Management(II)

NaN for Missing Value

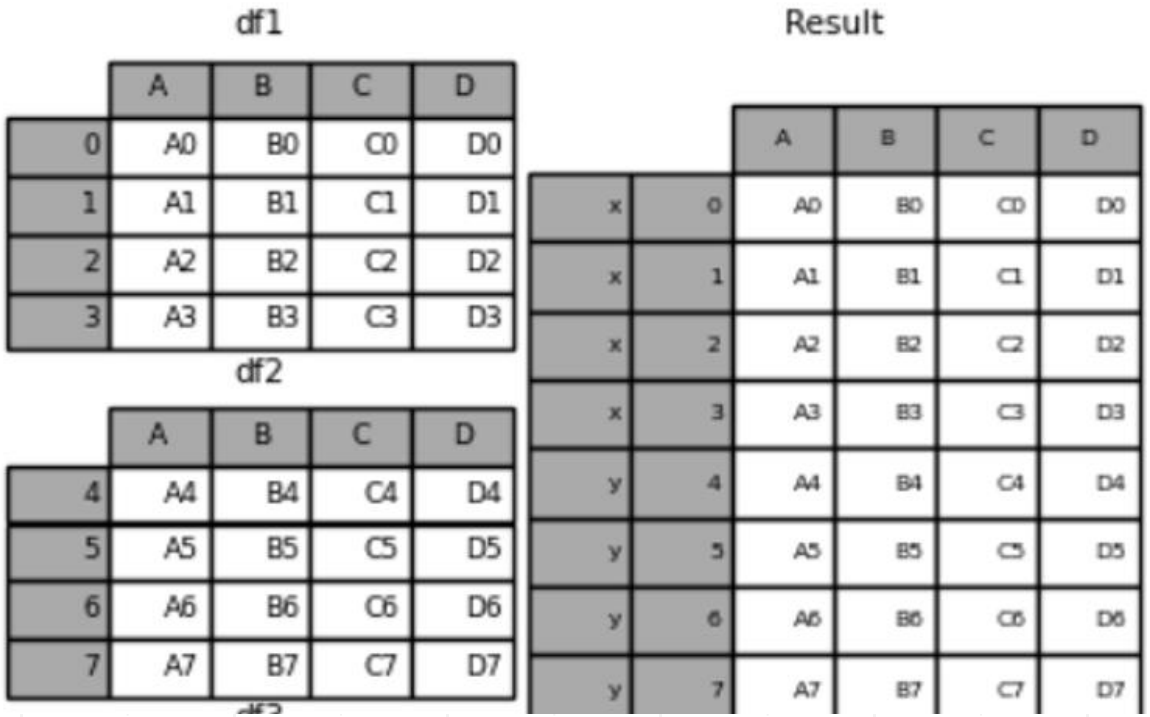
- *NaN* is used for representing missing data in numeric columns.
- The data type of *NaN* **is float**.
- To detect *NaN*, Pandas provides the *.isna()* and *.notna()* functions.
- Some Pandas operations will generate *NaN*. For example, when we concatenate or merge two DataFrames with a different number of columns or keys, the missing columns or rows will be filled with *NaN*.

	2020-01-21	2020-01-22	2020-01-23	2020-01-24	2020-01-25	2020-01-26	2020-01-27	2020-01-28	2020-01-29	2020-01-30
Washington	1	1	1	1	1	1	1	1	1	1
Illinois	NaN	NaN	NaN	1	1	1	1	1	1	2
California	NaN	NaN	NaN	NaN	1	2	2	2	2	2
Arizona	NaN	NaN	NaN	NaN	NaN	1	1	1	1	1
Massachusetts	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Day 4: Data Management (II)

Concatenating multiple Dataframes along the row axis (axis = 0)

- concatenating along the rows:
  - joining df_2 to df_1 vertically using column names as **concatenating/joining identifiers**.



Day 4: Data Management (II)

Concatenating multiple Dataframes along the row axis

	orange	apple	grapes
0	3	0	7
1	2	3	14
2	0	7	6
3	1	2	15

	grapes	mango	banana	pear	pineapple
0	13	10	20	21	30
1	12	13	23	24	33
3	2	2	4	51	30
4	55	9	0	22	36
5	98	76	9	25	31

Merging Dataframes along the column axis

Concat with axis = 0

Concat axis = 0

Concat axis = 1

Concat with axis = 1

	orange	apple	grapes	mango	banana	pear	pineapple
0	3.0	0.0	7	NaN	NaN	NaN	NaN
1	2.0	3.0	14	NaN	NaN	NaN	NaN
2	0.0	7.0	6	NaN	NaN	NaN	NaN
3	1.0	2.0	15	NaN	NaN	NaN	NaN
0	NaN	NaN	13	10.0	20.0	21.0	30.0
1	NaN	NaN	12	13.0	23.0	24.0	33.0
3	NaN	NaN	2	2.0	4.0	51.0	30.0
4	NaN	NaN	55	9.0	0.0	22.0	36.0
5	NaN	NaN	98	76.0	9.0	25.0	31.0

	orange	apple	grapes	grapes	mango	banana	pear	pineapple
0	3.0	0.0	7.0	13.0	10.0	20.0	21.0	30.0
1	2.0	3.0	14.0	12.0	13.0	23.0	24.0	33.0
2	0.0	7.0	6.0	NaN	NaN	NaN	NaN	NaN
3	1.0	2.0	15.0	2.0	2.0	4.0	51.0	30.0
4	NaN	NaN	NaN	55.0	9.0	0.0	22.0	36.0
5	NaN	NaN	NaN	98.0	76.0	9.0	25.0	31.0

Source: Medium

Day 4: Data Management (II)

Merging Dataframes along the column axis

- **Merging along the columns** means merging DataFrame B to DataFrame A horizontally.
- Function: pd.merge( )
- the **pd.concat()** can also be used to merge along columns by changing the argument axis = 1;
- the function **pd.merge()** can **ONLY** be used to merge along the columns.

A

left

	key	A	B
0	K0	A0	B0
1	K1	A1	B1
2	K2	A2	B2
3	K3	A3	B3

B

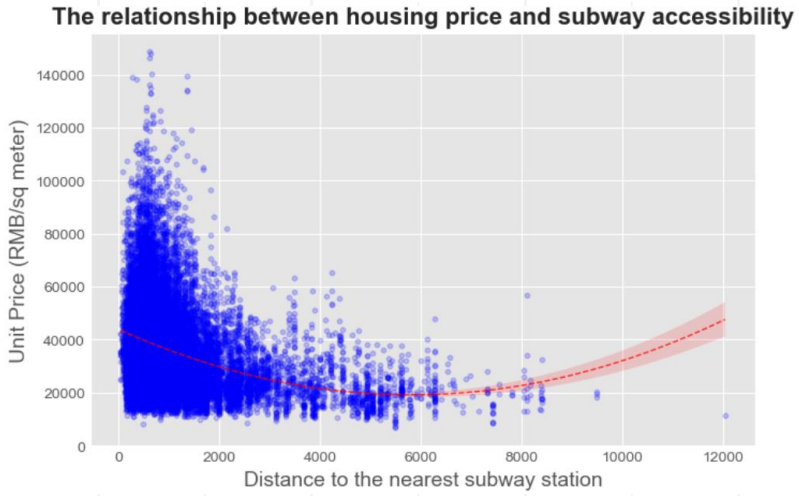
right

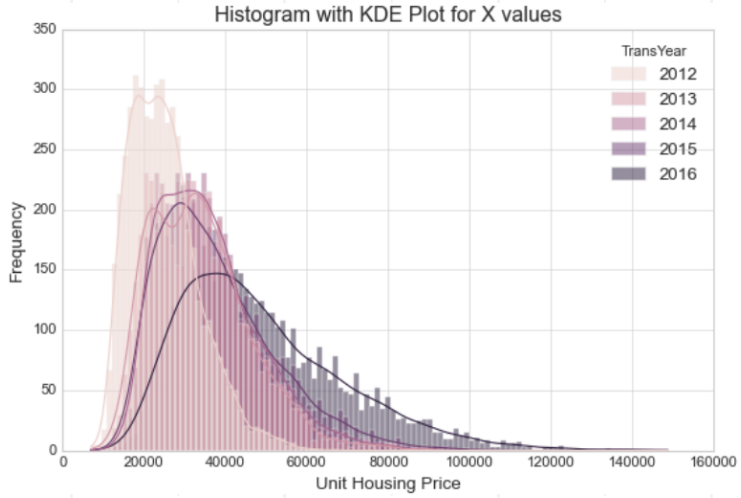
	key	C	D
0	K0	C0	D0
1	K1	C1	D1
2	K2	C2	D2
3	K3	C3	D3

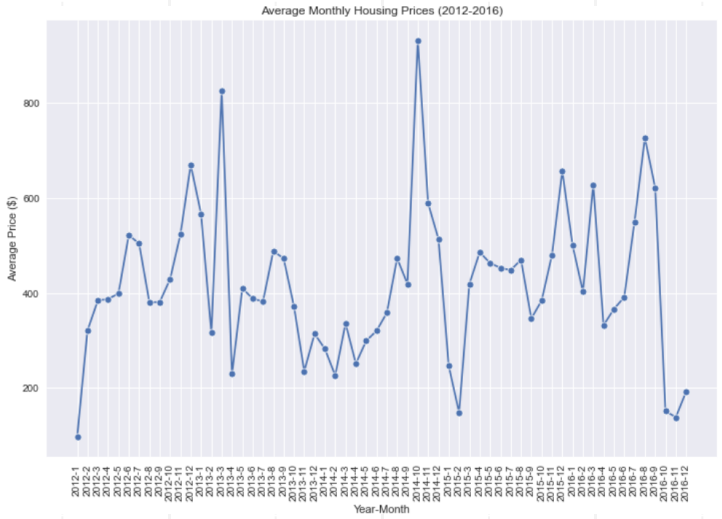
Result

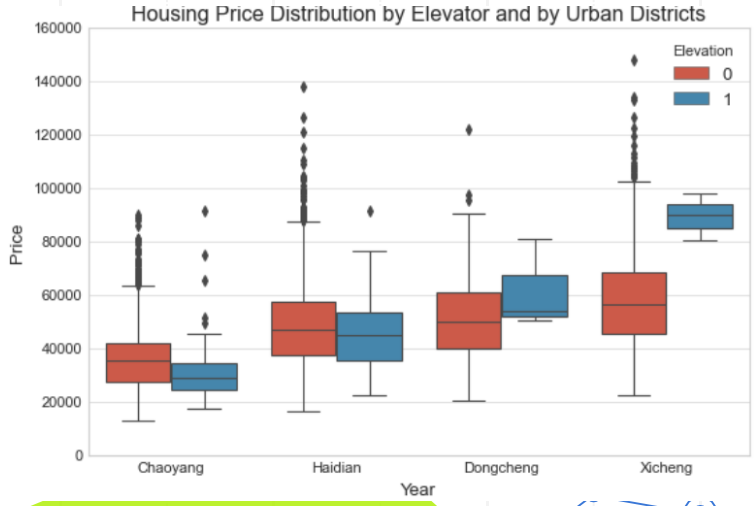
	key	A	B	C	D
0	K0	A0	B0	C0	D0
1	K1	A1	B1	C1	D1
2	K2	A2	B2	C2	D2
3	K3	A3	B3	C3	D3

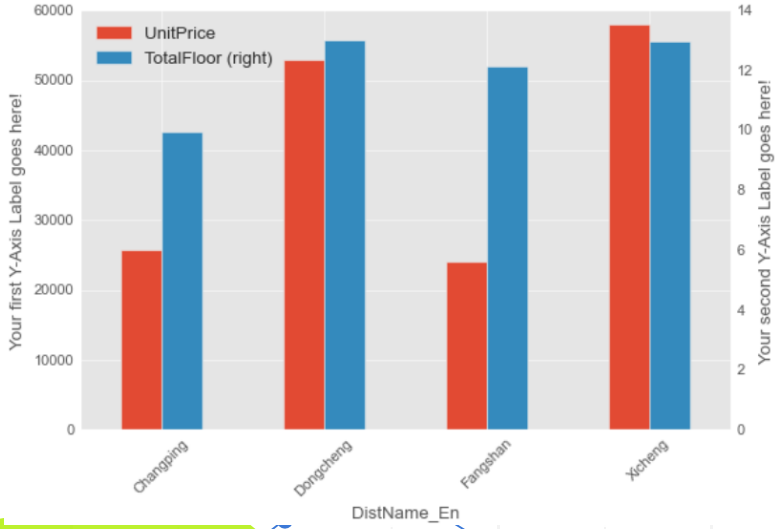
Day 5: Visualization

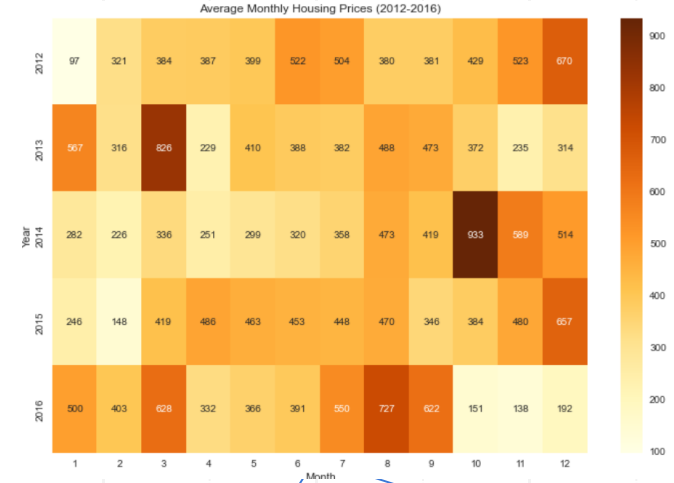


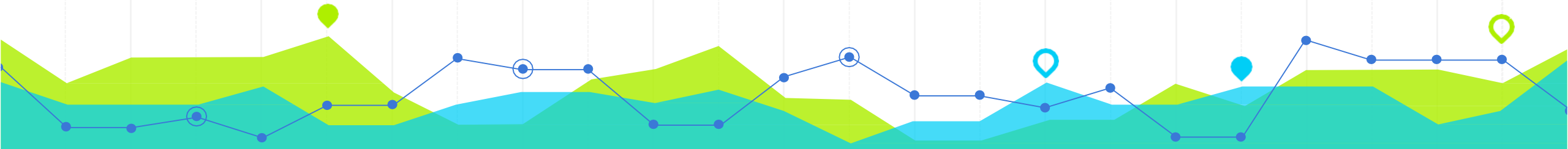






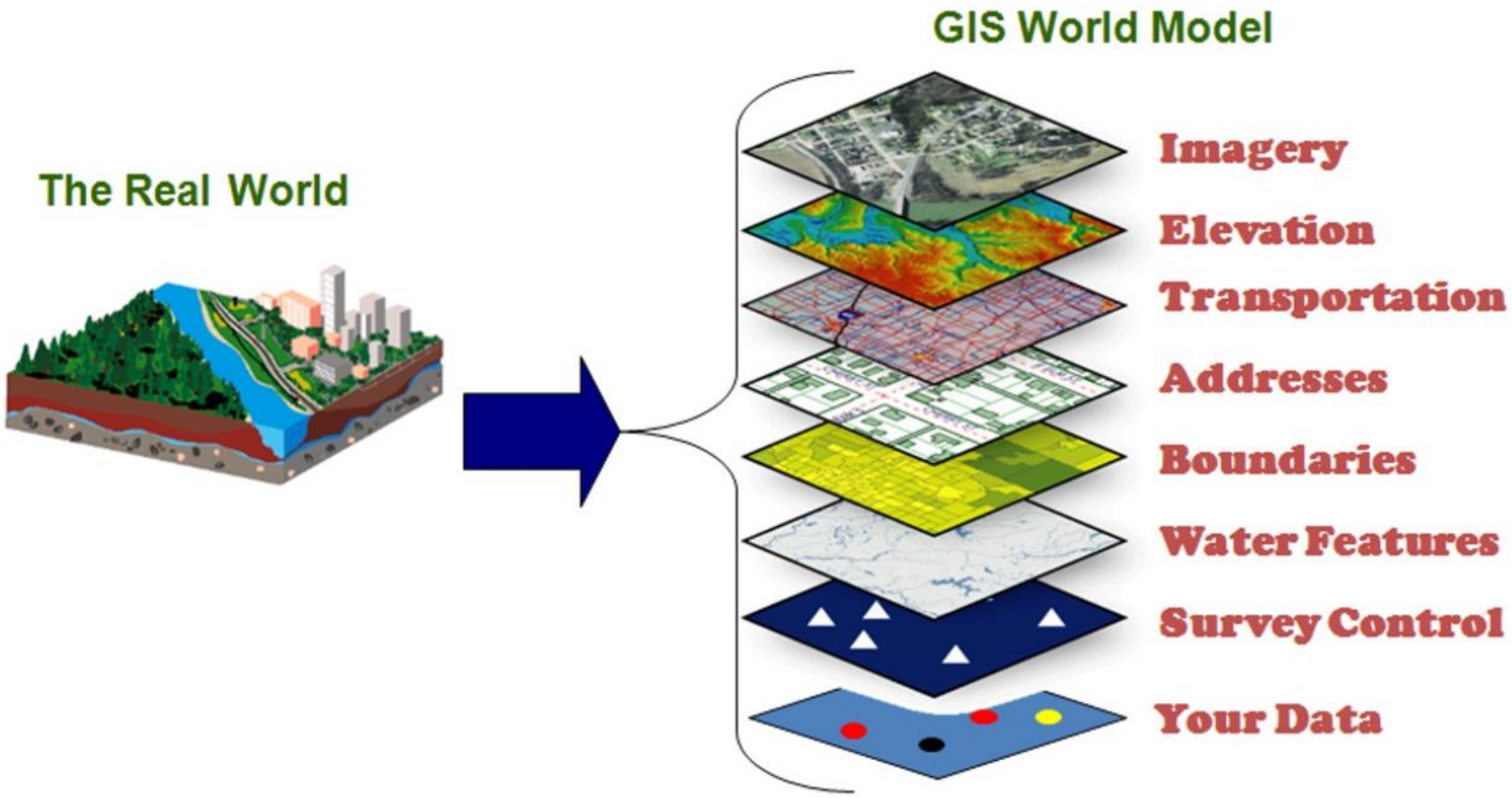






GIS Basics 1

What is Geographic Information System?



Source: https://henrico.us/it/gis/

What is Geographic Information System?

Definition

- **Definition by ESRI(Environmental Systems Research Institute):** A geographic information system (GIS) is a **framework** for gathering, managing, and analyzing **data**. Rooted in the science of geography, GIS integrates many types of data. It analyzes **spatial location** and organizes layers of information into visualizations using maps and 3D scenes. With this unique capability, GIS reveals deeper insights into data, such as patterns, relationships, and situations—helping users make smarter **decisions**.

- **Definition by National Geographic Society:** “A geographic information system(GIS) is a **computer system** for capturing, storing, checking, and displaying **data** related to positions on Earth’s surface.”

What Questions can GIS Answers?

GIS answer/solve geographical questions/problems.

- What is at certain location?
- What has changed in certain location over time?
- What is the spatial pattern? dispersed? agglomerated?
- How to make decision based on the spatial pattern?

What Questions can GIS Answers?

Advancement of human understanding (build theory)

- What is the spatial distribution of asthma cases among children in New Jersey? What factors (demographic, environmental, etc.) might help explain the spatial distribution of asthma cases?

Normative uses:

- e.g., use GIS to conduct a site suitability analysis for a retail establishment (where the best site should be located)
- e.g. Which areas of New York City are prone to flooding? How severe is the impact?

Positive uses:

- e.g., confirm a theory by finding positive evidence in support of it, reject theories when negative evidence is found — a causal inference

Forming your theory/research hypothesis. What questions intrigue you?

What Questions can GIS Answers?

- Emergency management
- Social services
- Economic development
- Transportation
- Urban planning and development
- Water resources
- Business Management
- Environmental sciences and security
- Real estate development and appraisal

.....GIS can be applied in any spatial-related research!

GIS in Planning

Hazard Mitigation

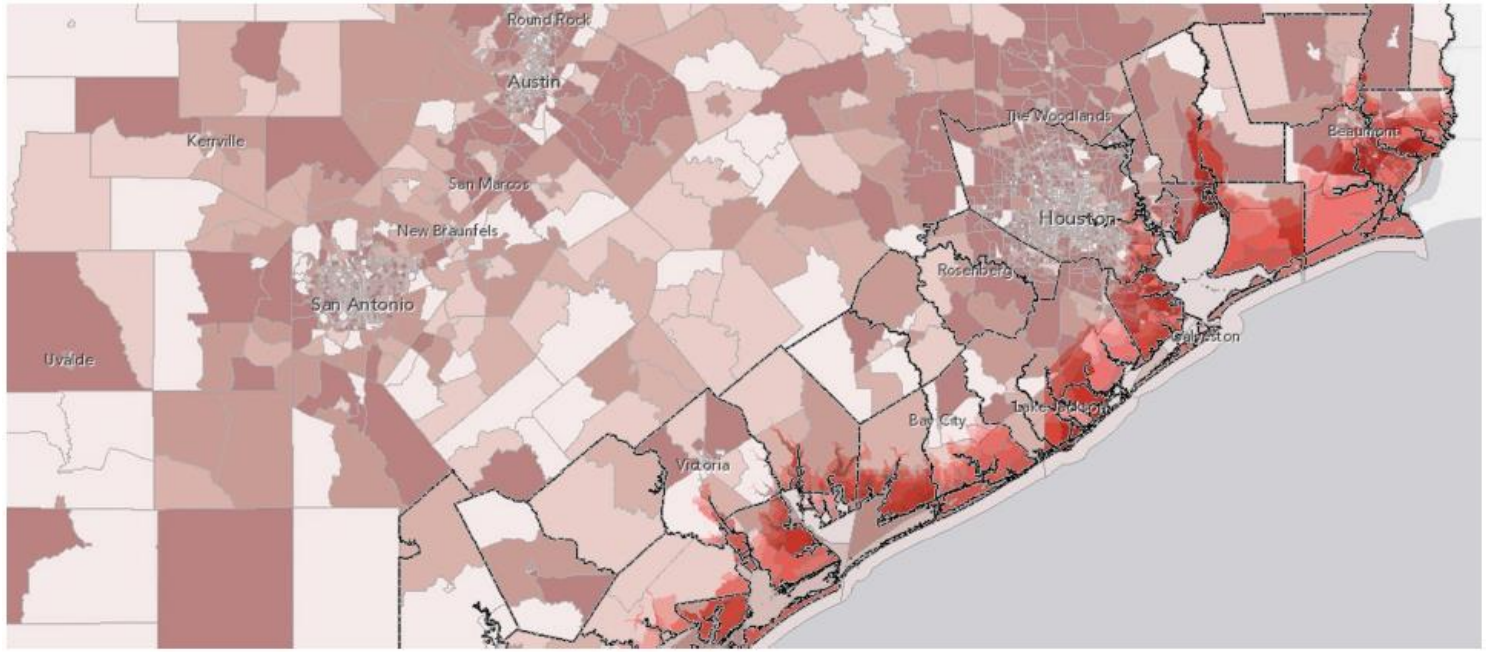


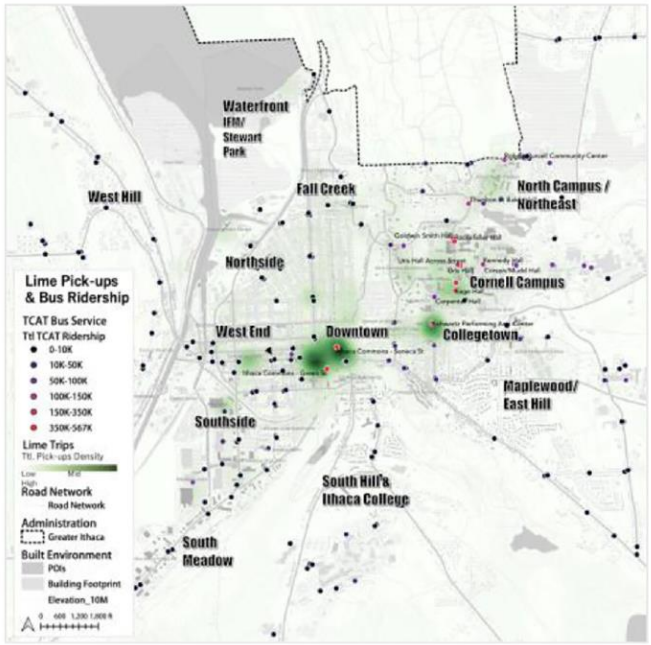
Image source

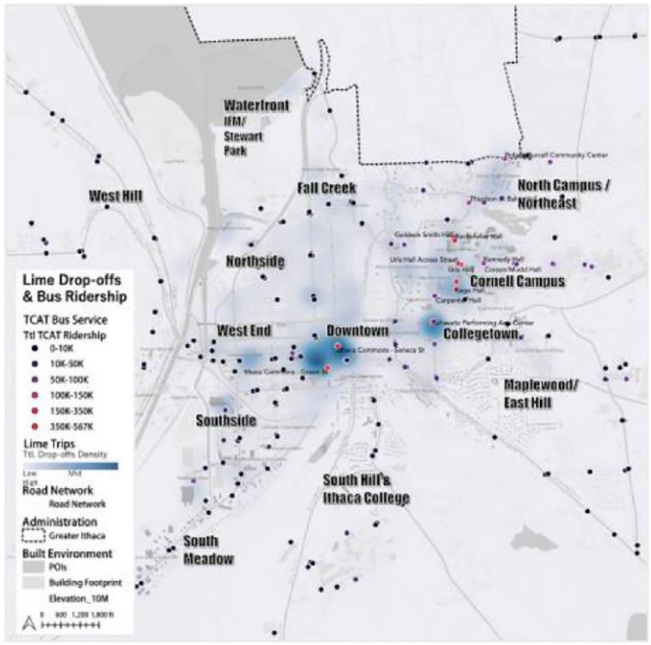
Example from Texas Sustainable and Resilient Planning Atlases: **Mapping the environmental hazards**

- Where are mitigation mostly needed
- Which neighborhoods face the most hazard risks
- Where should future growth occur

GIS in Planning

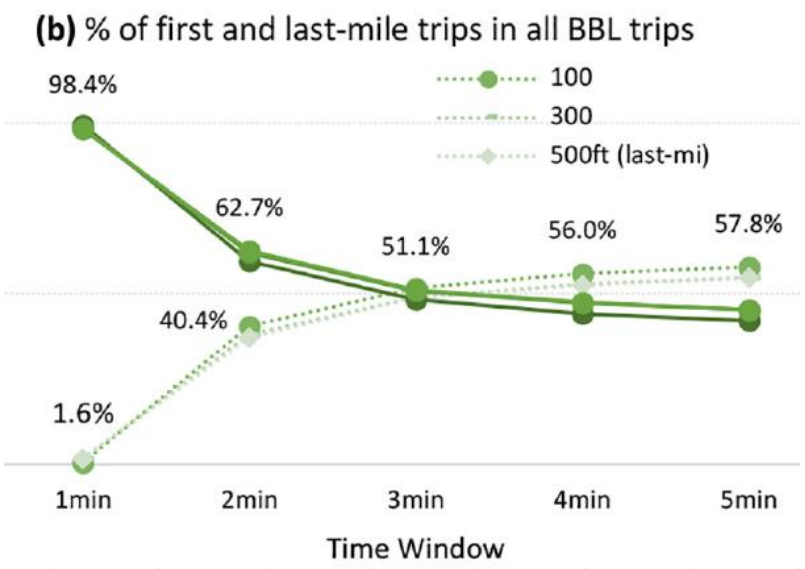
Transportation Planning





Analysis of the connection between mobility options:

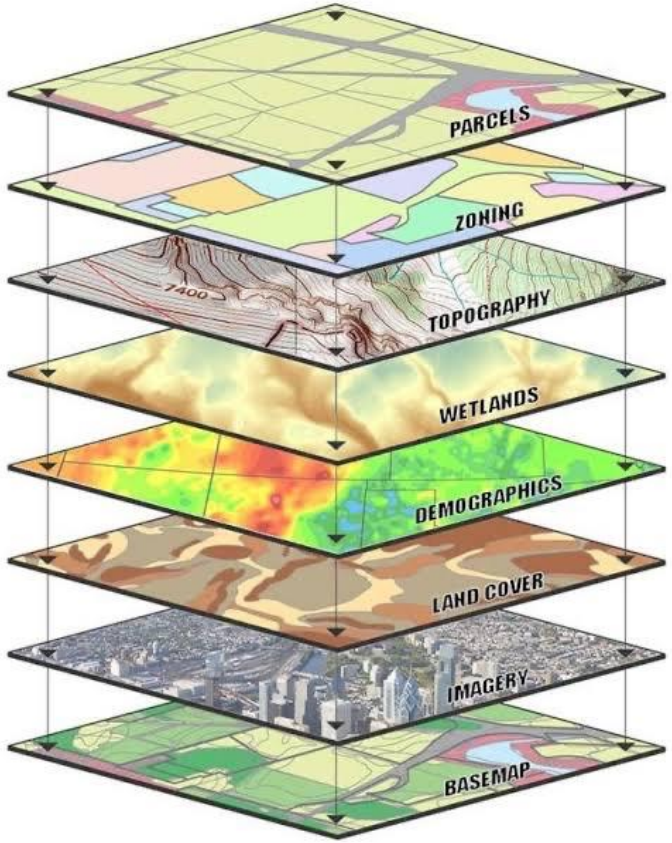
- How Ithaca bikeshare connect with TCAT bus service?



Qiu, W., & Chang, H. (2021). The interplay between dockless bikeshare and bus for small-size cities in the US: A case study of Ithaca. *Journal of Transport Geography*, *96*, 103175.

GIS in Planning

Infrastructure Planning



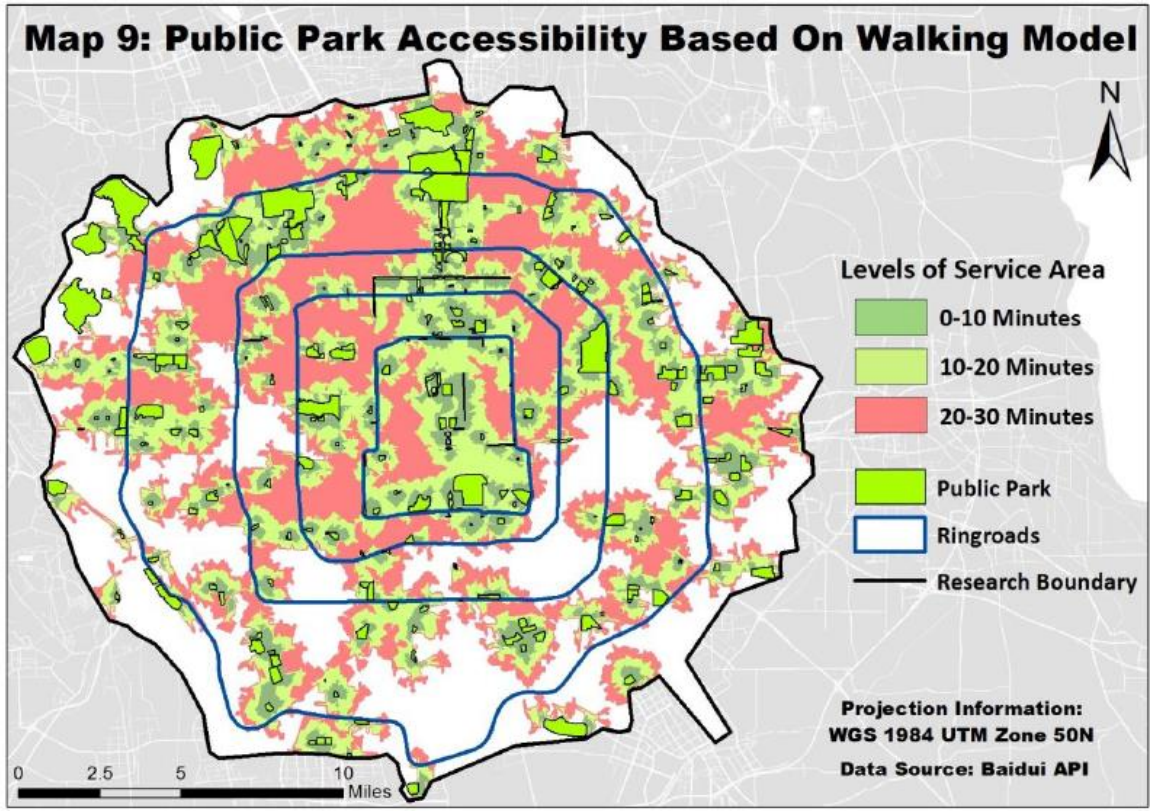
Land use suitability analysis:

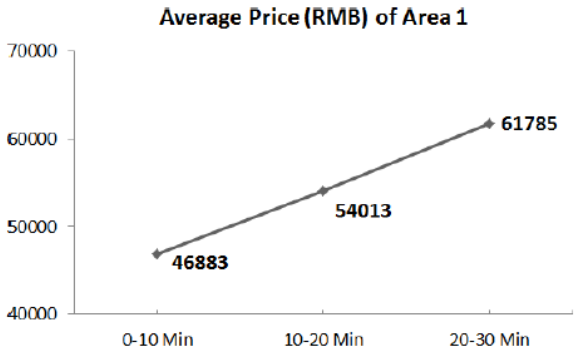
- Location-allocation modeling(systematic analysis)
- Where should we locate an educational facility?

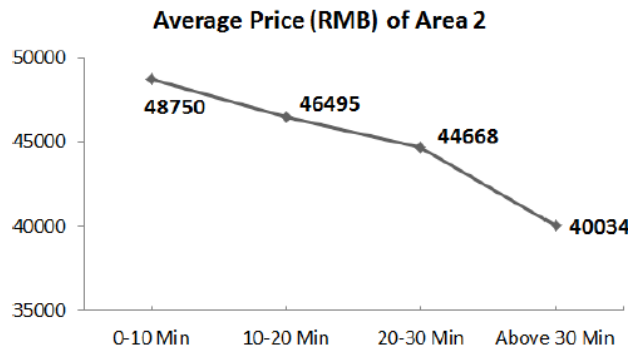
Source: CRP 5080

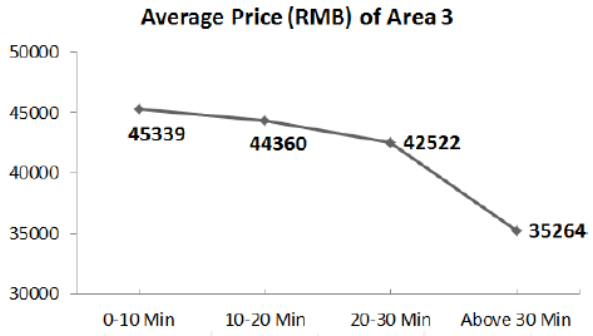
GIS in Planning

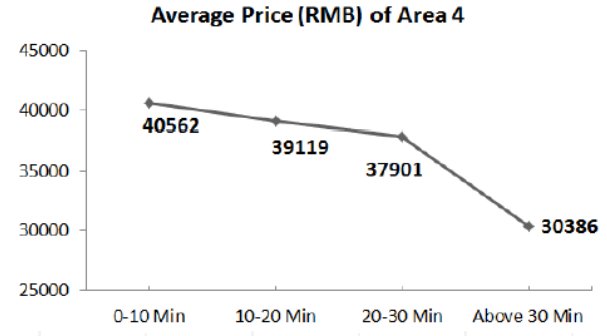
Environmental Planning



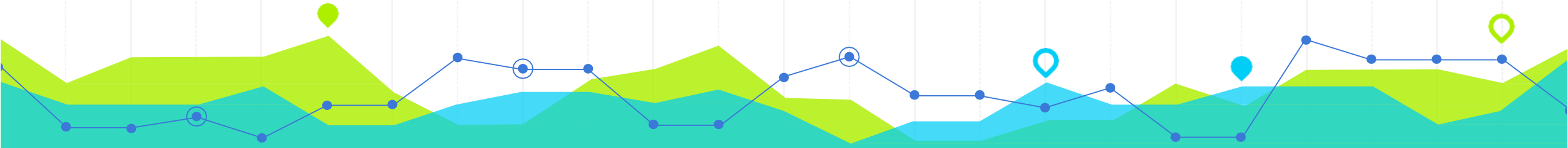








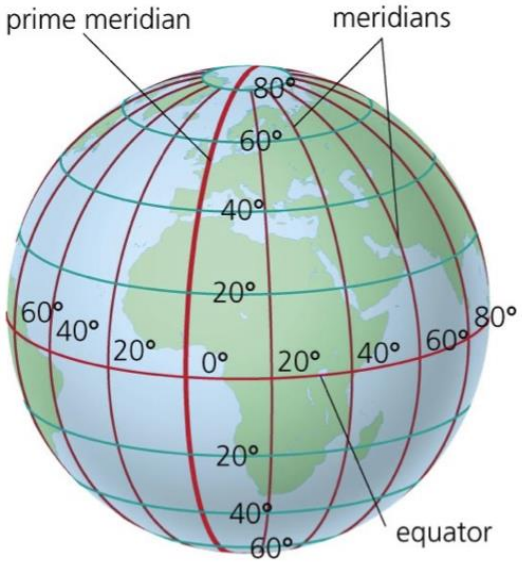
Source: CRP 5080



Spatial Data 1.1

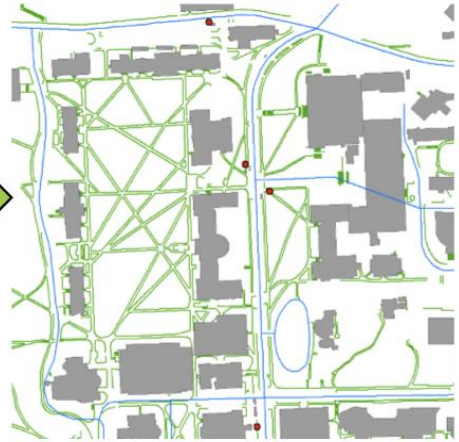
Spatial Data Representation

Spatial data allows abstract representations of geographic features in real-world locations to be digitally represented and stored in a database



Real World locations

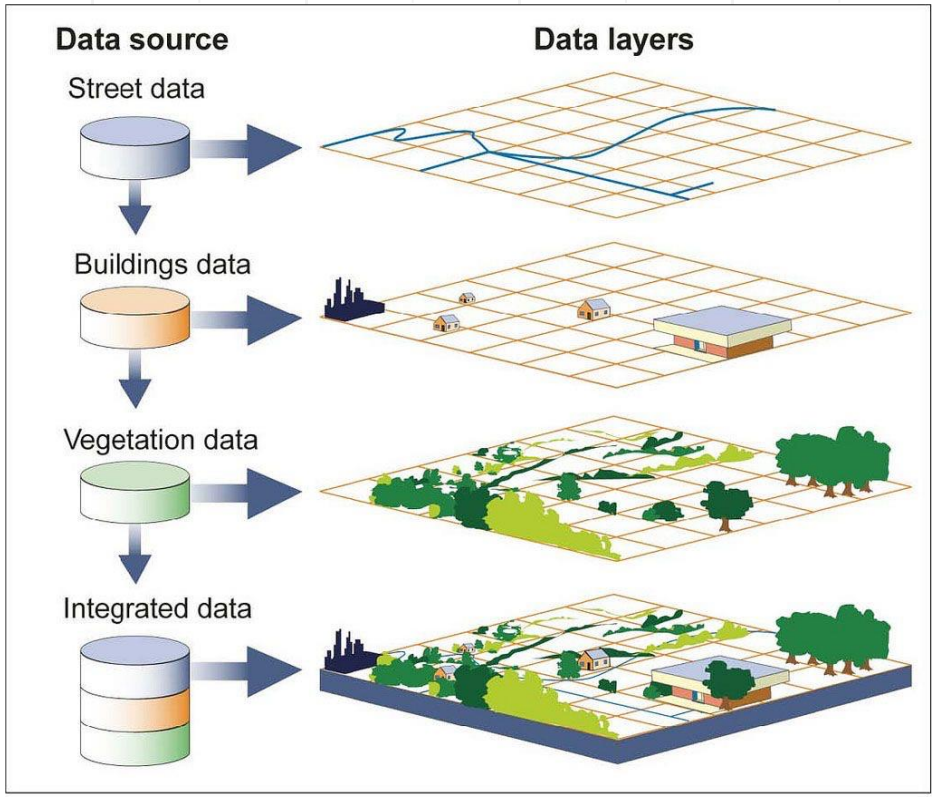
Geographic features



Abstract representations

Spatial Data Representation

Spatial Data representation is a translation process: turning real-world data into map and other visualization.



Source: GAO.

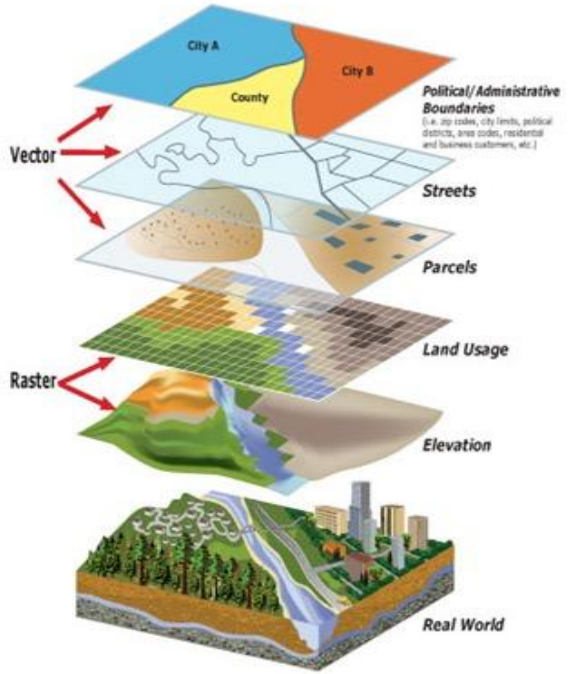


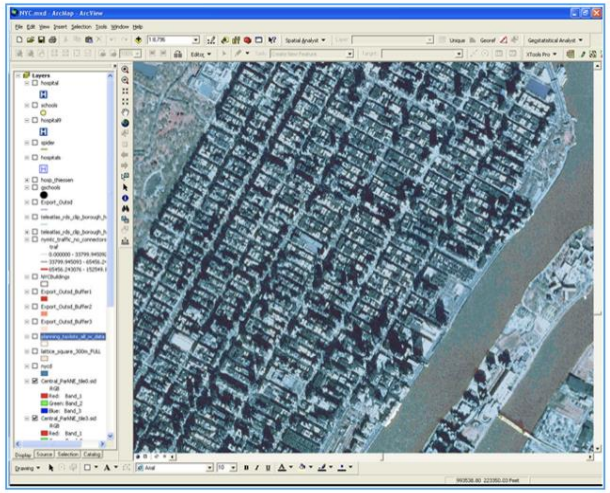
Figure 2: An example of map layers used together in GIS
San Berardino County GIS Dept, 2012. Used for educational purposes only. http://gis.sbcounty.gov/

Spatial Data Representation

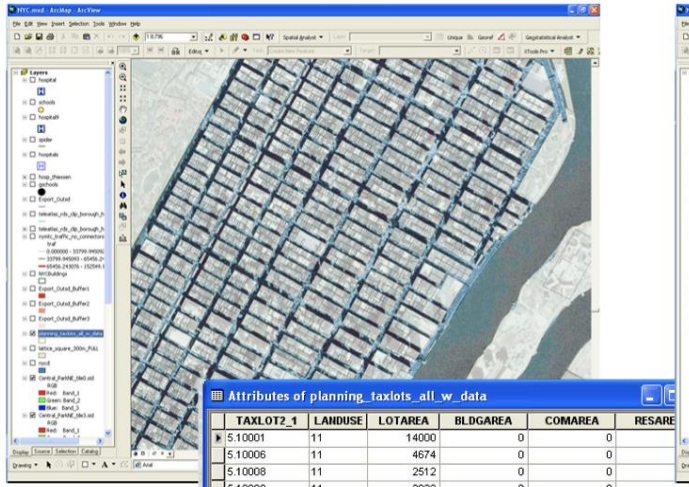
GIS integrates all these layers' of information and provide tools for analysis.
There are three types of information stored within spatial data:

- Topological relationships: Connect, adjacent, contained within, overlap, near, etc.
- Spatial features: point, line, polygons, etc.
- Non-spatial attributes: population, rates of poverty, GDP, etc.

Aerial Photo of Manhattan



Geographic Data: Tax lots

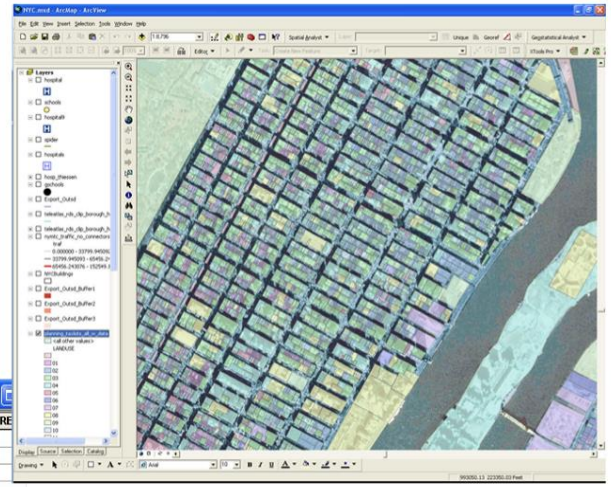


Attributes of planning_taxlots_all_w_data

TAXLOT2_1	LANDUSE	LOTAREA	BLDGAREA	COMAREA	RESARE
5.10001	11	14000	0	0	
5.10006	11	4674	0	0	
5.10008	11	2512	0	0	
5.10009	11	3932	0	0	
5.10010	05	5716	18965	18965	
5.10017	11	910	0	0	
5.10018	11	500	0	0	
5.10019	11	585	0	0	
5.10021	11	6700	0	0	

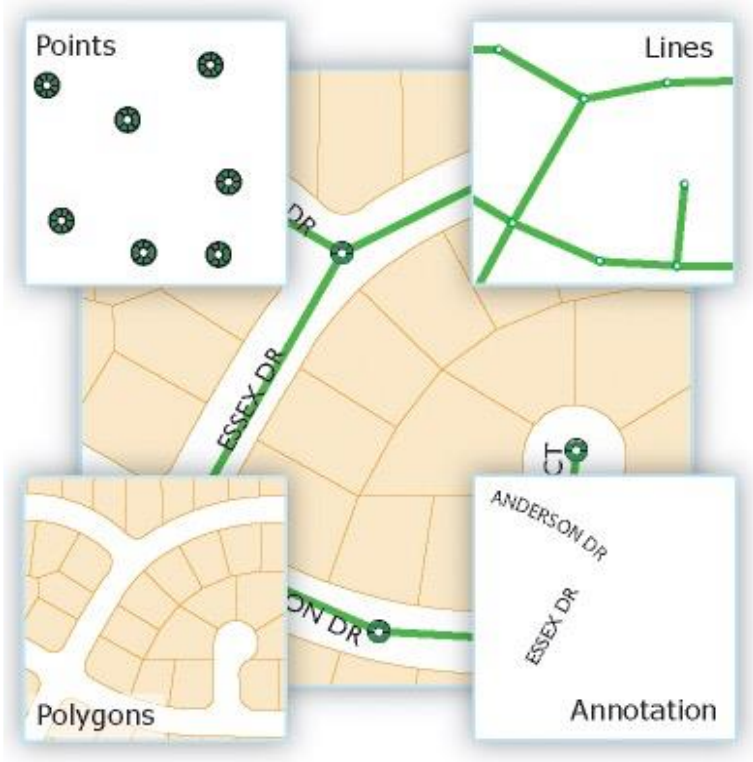
Data Attributes: Land use, building area, etc.

Map Based on Descriptive Data

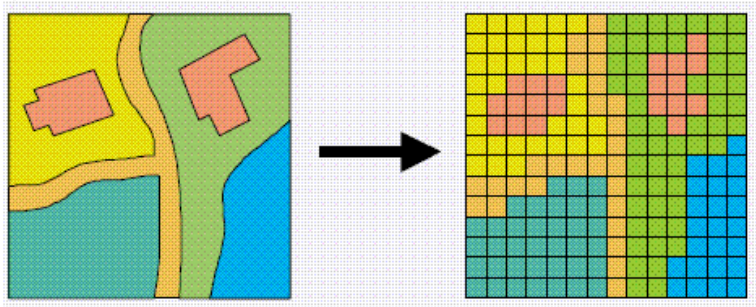


Types of Spatial Data

- Vector/Feature Data:
  Points, Lines, Polygons

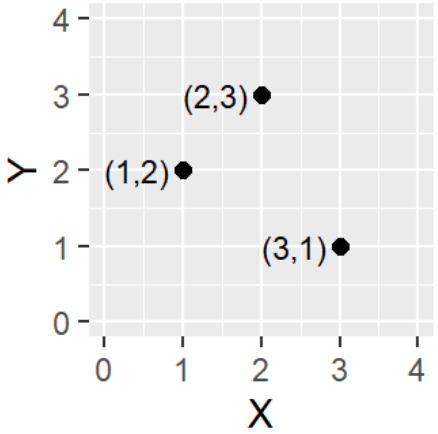


- Raster Data:
  Grid of Fixed-Size Pixels



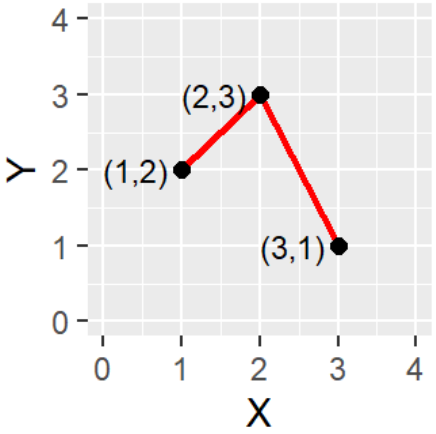
Types of Spatial Data – vector data

- Points



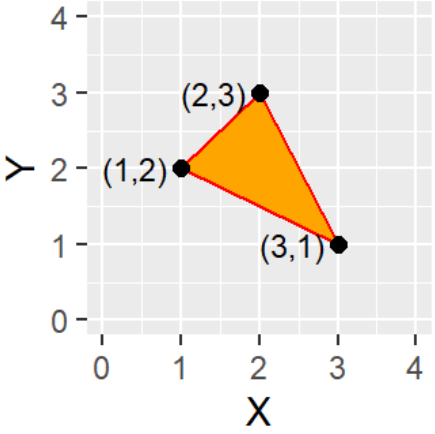
Points – tree planting

- Lines/polylines



Lines – Bike lanes

- Polygons



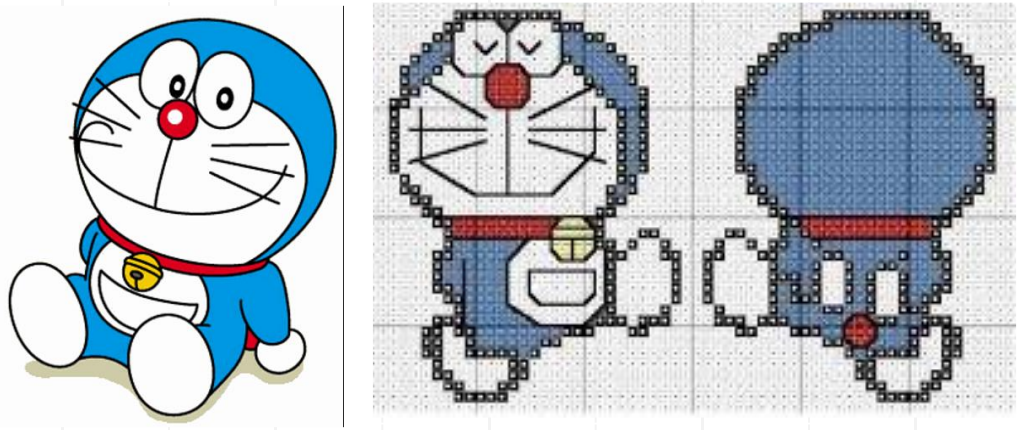
Polygon – Special tax district

Types of Spatial Data in GIS

Vector data

Advantages of vector data:

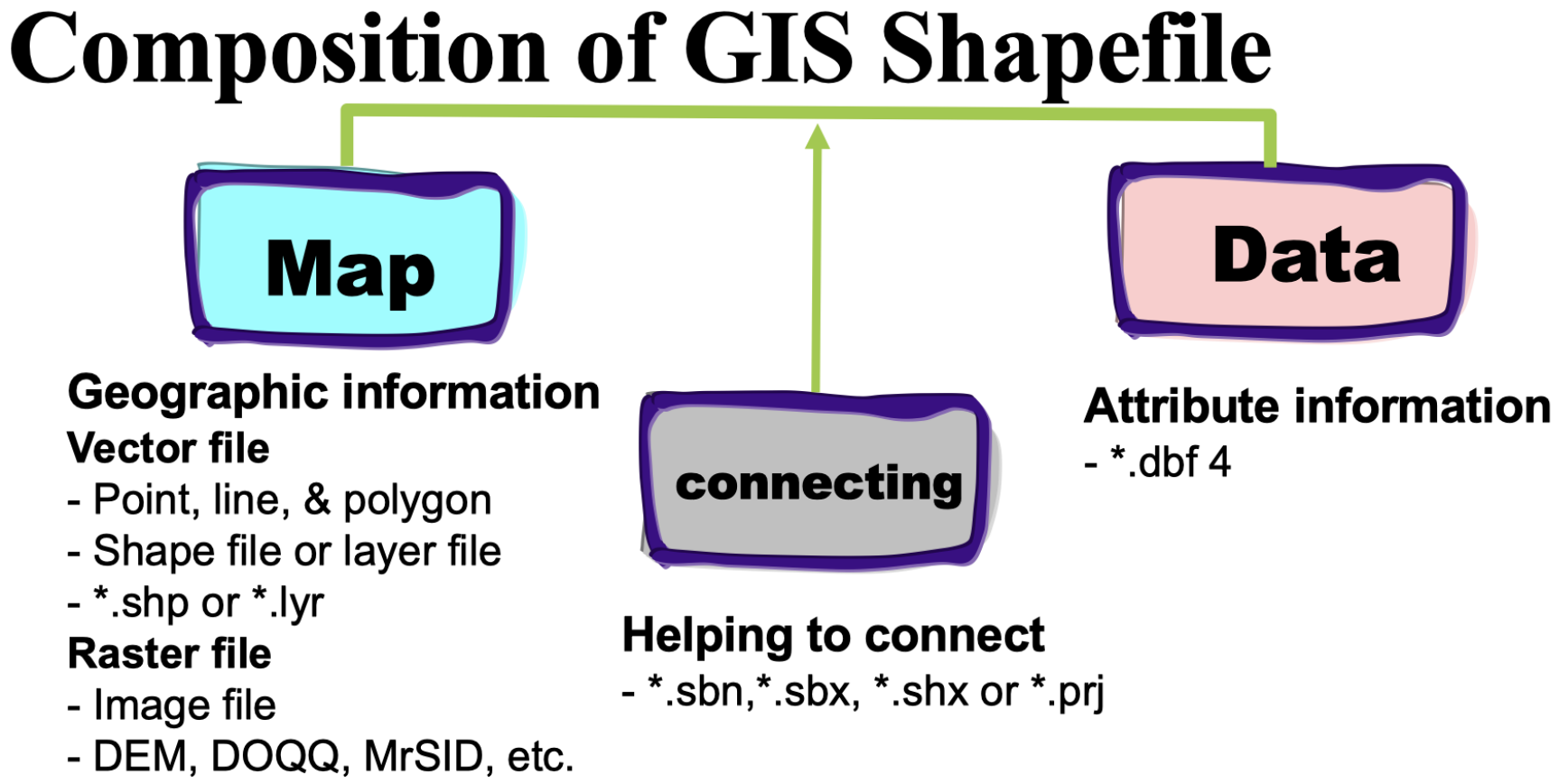
- Accuracy & Aesthetically pleasing: Precision of points, lines, and polygons (pattern/shape recognizable)



- Increased ability to alter the scale of observation and analysis: Each coordinate pair associated with a point, line, and polygon represents an infinitesimally exact location.
- Topology(spatial relationship) is inherent in the vector model: Enable simplified spatial analysis

Types of Spatial Data in GIS

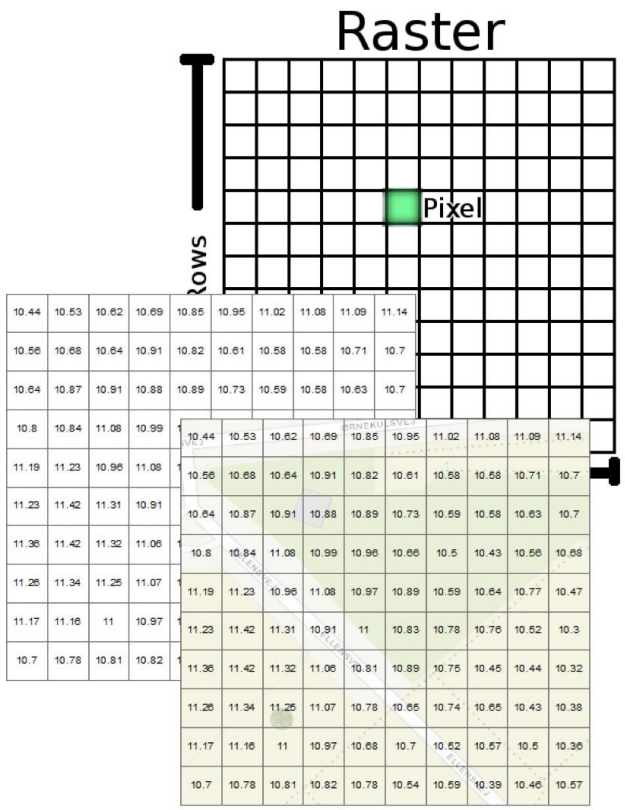
Filetype : We will primarily use shapefiles. It is comprised of several support files.



Types of Spatial Data in GIS

Raster Data

Grid of Fixed-size Pixels

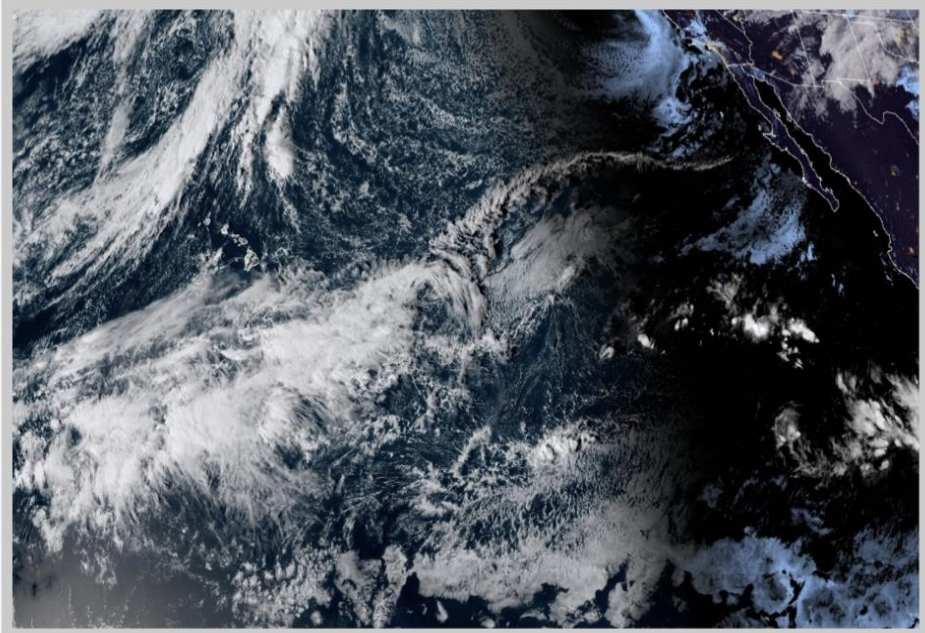


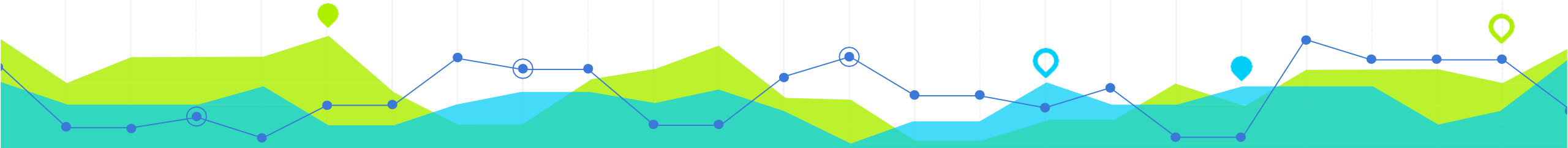
- A matrix of cells, or pixels, organized into rows and columns (or a grid).
- All cells must be the same size, the size of the cell determines its resolution.
- Each cell contains a value, representing information such as temperature or elevation.

Types of Spatial Data in GIS

Raster Data

The process of capturing raster data from an aero-plane or satellite is calledremote sensing. Here are a few examples of raster data:

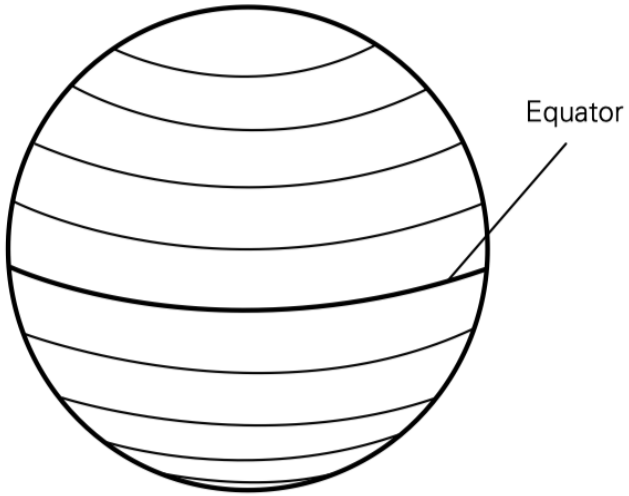
1.Digital aerial photographs	2.Imagery from satellites
	
Downtown Ithaca, 1971	East Central Pacific, 2020.10.25



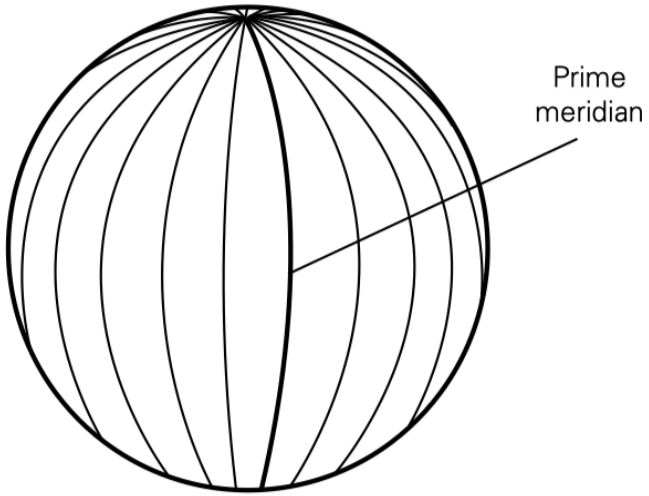
Map Projection 1.2

What is GCS (geographic coordinate system)?

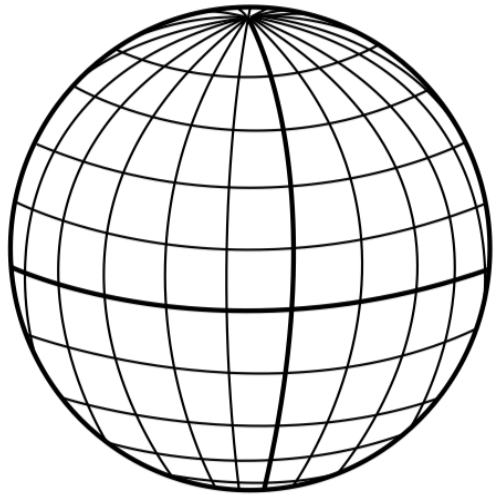
A geographic coordinate system (GCS) is a reference framework that defines the locations of features on a **model** of the earth. It's shaped like a globe—spherical. Its units are angular, usually **degrees.**



Parallels of latitude



Meridians of longitude

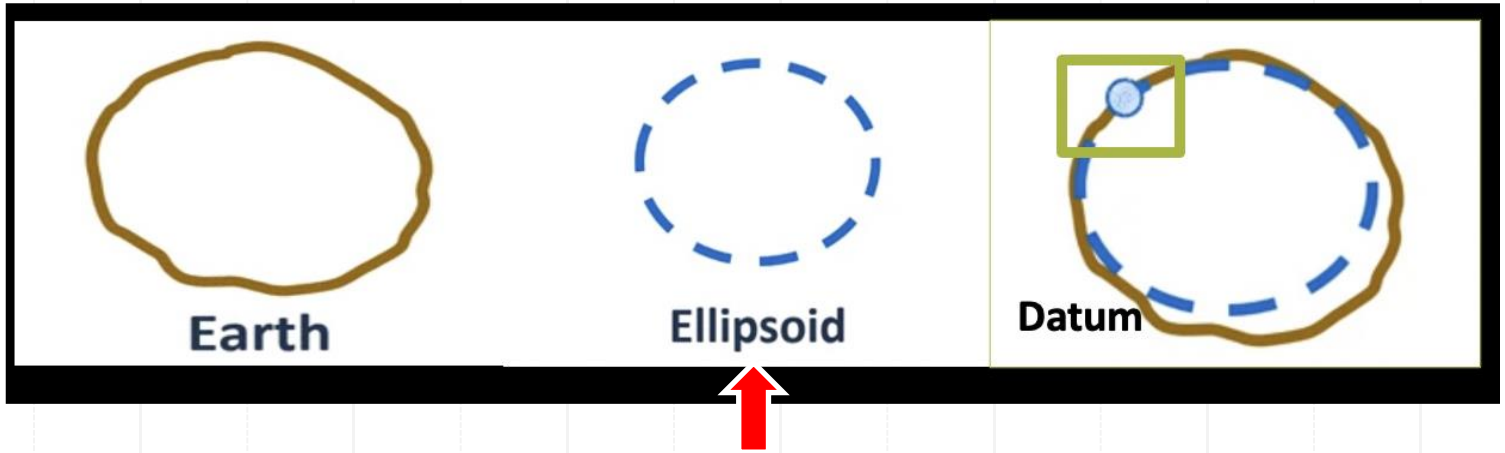


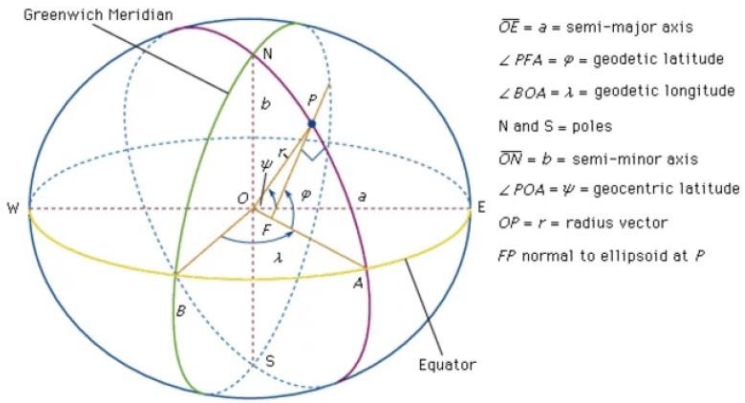
Graticular network

What is GCS (geographic coordinate system)?

Datum: An important parameter of the GCS

- Each Ellipsoid (the model) is designed to approximate the Earth's shape for one part of the planet.
- Datum defines point on ellipsoid linked to point on earth (the origin, from which all other points are calculated)
- Datum determines the ellipsoid model thus GCS

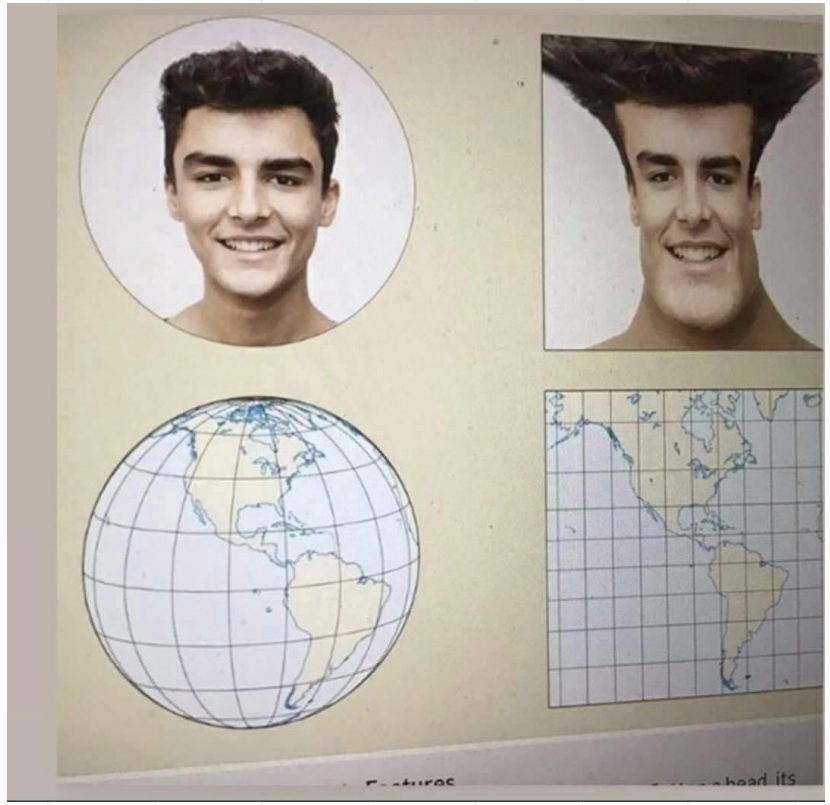




[A Simple Explanation of Datum]

What is PCS (projected coordinate system)?

A **projected coordinate system (PCS)** is flat. It contains a GCS, but it converts that GCS into a flat surface, using math (the projection algorithm) and other parameters. Its units are linear, most commonly in **meters.**



[Why all world maps are wrong]

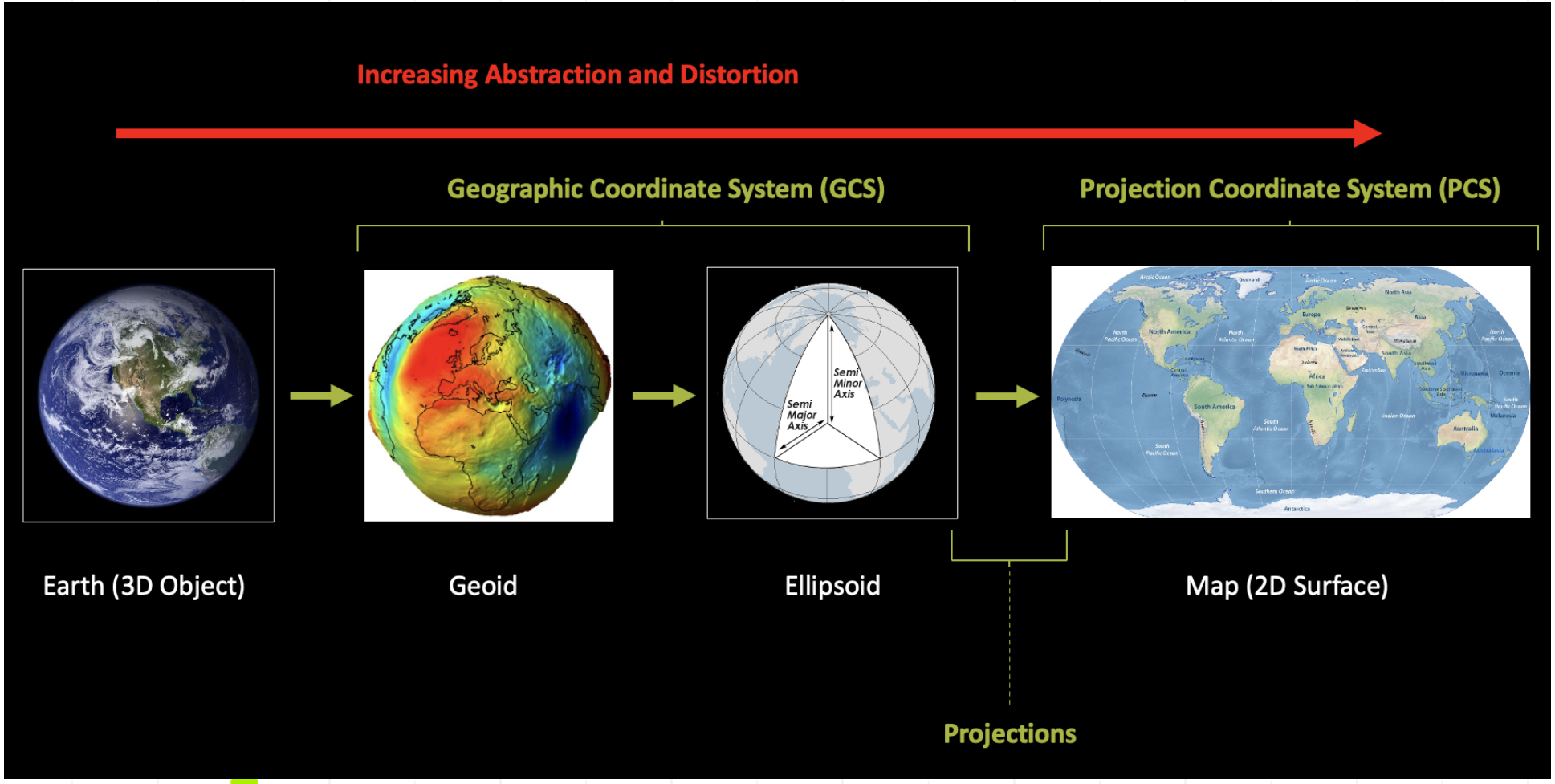
Don’t be confused

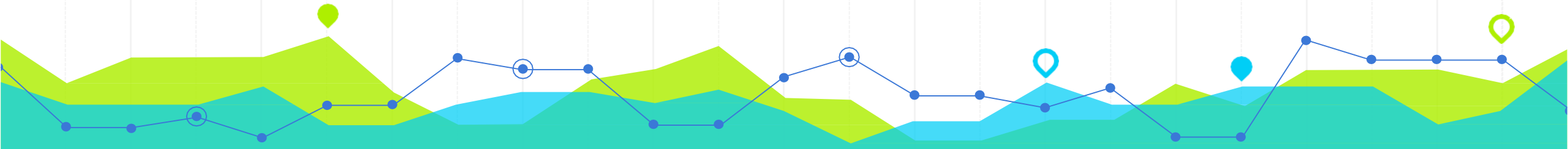
GCS and PCS

What is the relationship and difference between GCS and PCS?

- A GCS defines where the data is located on the earth’s surface. In angular units.
- A PCS tells the data how to draw on a flat surface, like on a paper map or a computer screen. In linear units.
- A projected coordinate system (PCS) is a **GCS** that has been flattened using a map projection. A projection is a mathematical algorithm that is used to create a PCS.
- A projection defines a set of parameter values which vary depending on the projection (false easting, central meridian, standard parallel, and so on).
- Your data must have a GCS before it knows where it is on earth. There has to be GCS first, and then a PCS.

From Earth to GCS to PCS





GeoPandas 2

Geopandas

Pandas and Geopandas

- The majority of functionalities of pandas also work in geopandas.

Data Types — from Pandas to Geopandas

- From Dataframe to Geodataframe
- From Series to Geoseries

Create a GeoDataFrame — three components

```
gdf_combine = gpd.GeoDataFrame(data = df_combine, crs = 'epsg:4326', geometry = geometry_list)
```

- a **DataFrame**
- a **crs** (coordinate reference system)
  - presented by EPSG code, e.g., "epsg: 4326")
- a **geometry list**
  - defines geometric object types of each observation , e.g., points, lines, or polygons

Geopandas

Dist2CBD	DistName_En	SubwayDum	wgs_lng	wgs_lat
9345.20091	Xicheng	Level 2	116.389876	39.839173
9345.20091	Xicheng	Level 2	116.389876	39.839173
9345.20091	Xicheng	Level 2	116.389876	39.839173

A Dataframe with geographical coordinates

Dist2CBD	DistName_En	SubwayDum	wgs_lng	wgs_lat	geometry
9345.20091	Xicheng	Level 2	116.389876	39.839173	POINT (116.38988 39.83917)
9345.20091	Xicheng	Level 2	116.389876	39.839173	POINT (116.38988 39.83917)
9345.20091	Xicheng	Level 2	116.389876	39.839173	POINT (116.38988 39.83917)

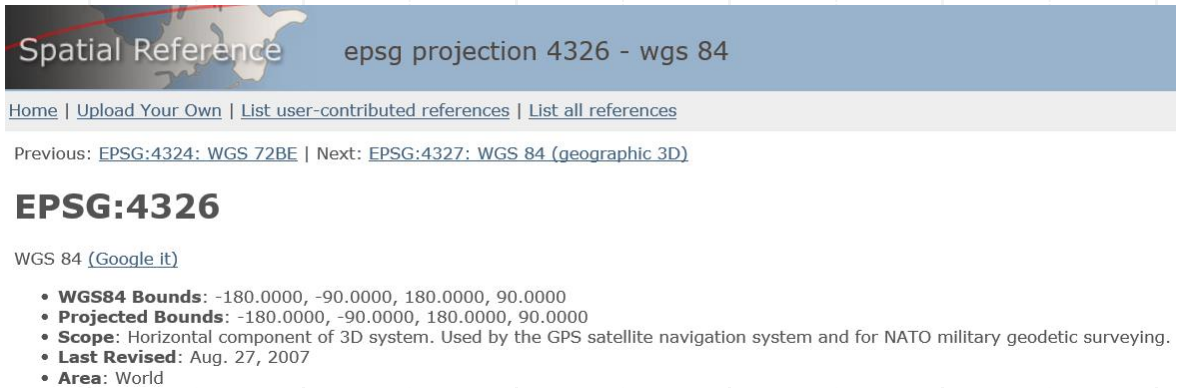
A GeoDataframe with a geometry list

Projection and EPSG

The common way to specify coordinate reference system (crs) is to use the *epsg* code:

- Example: WGS84---**EPSG:4326** (longitude and latitude)

```
gdf_combine = gpd.GeoDataFrame(data = df_combine, crs = 'epsg:4326', geometry = geometry_list)
```

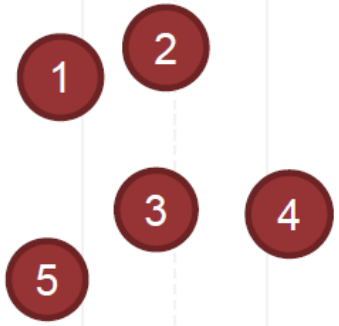


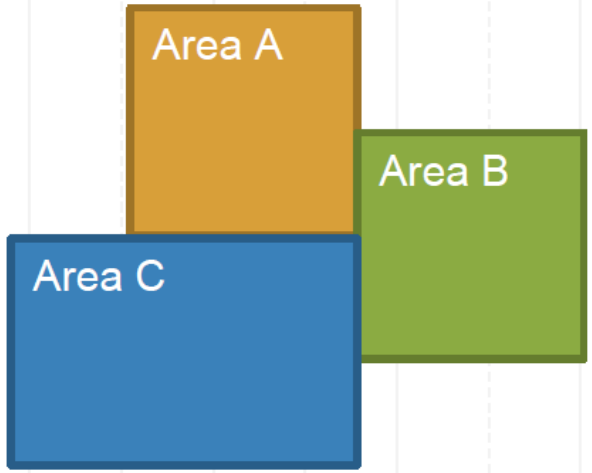
- EPSG is a collection of definitions of coordinate reference systems (map projections and datums) and coordinate transformations which may be global, regional, national or local in application.

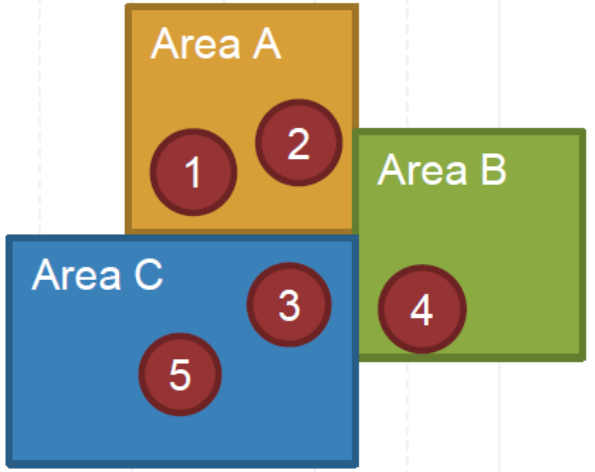
- Each coordinate system corresponds to a EPSG code.

Spatial Join

- Two geometry objects are merged based on their **spatial relationship** to one another.





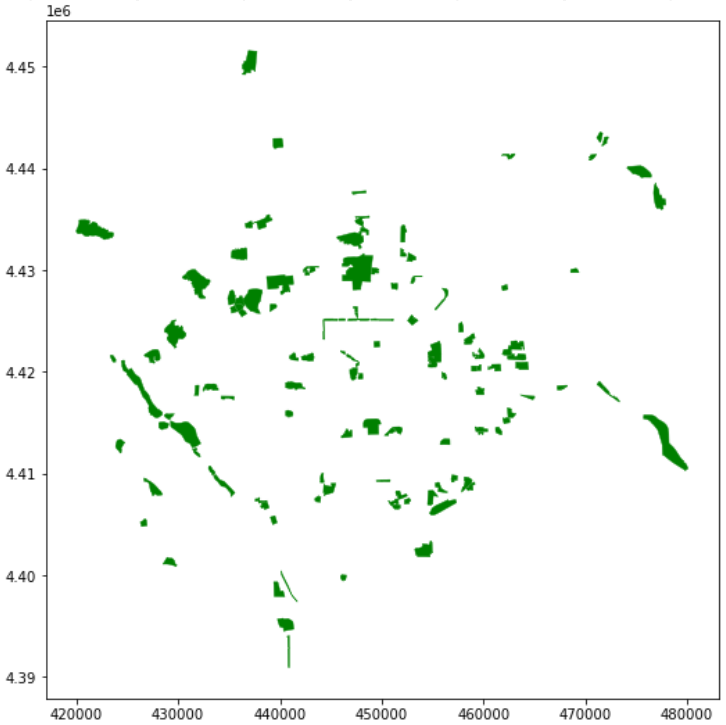


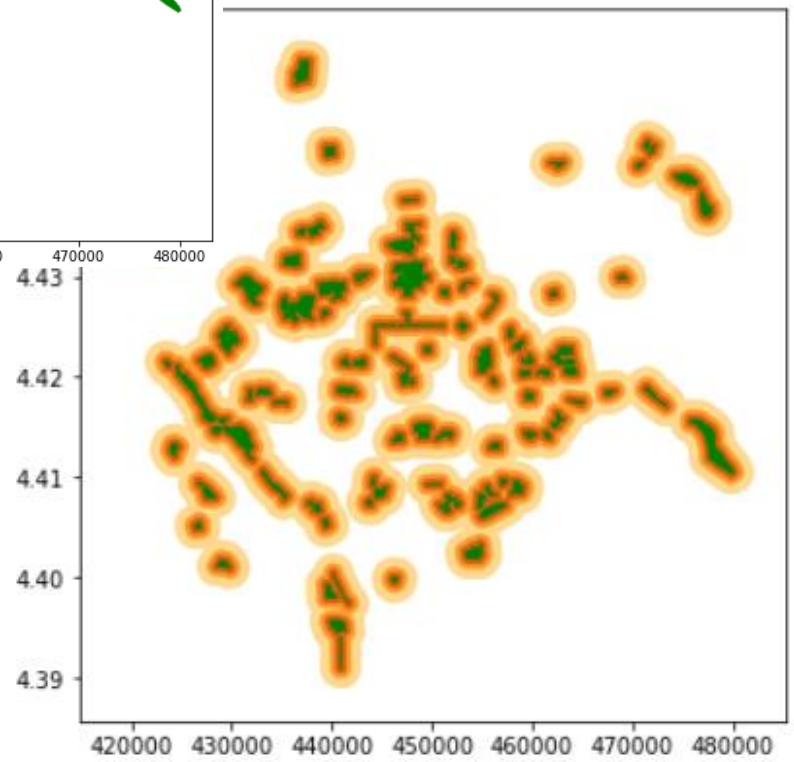
	Point
0	1
1	2
2	3
3	4
4	5

Spatial Join

	Polygon
0	Area A
1	Area B
2	Area C

	Point	Polygon
0	1	Area A
1	2	Area A
2	3	Area C
3	4	Area B
4	5	Area C





Buffer Analysis

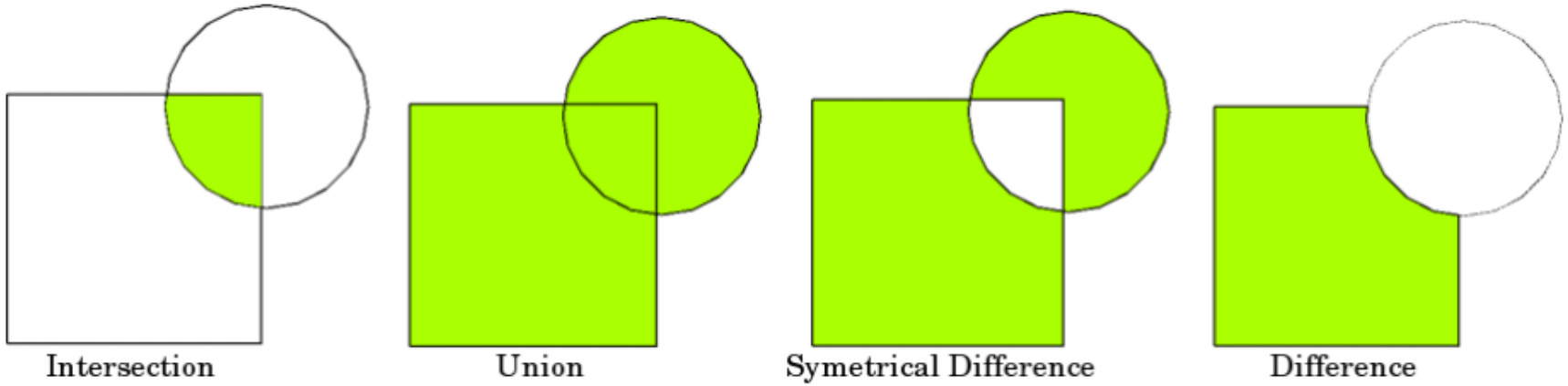
Creating a buffer
gdf.buffer( ): It returns a GeoSeries of geometries representing all points within a given distance of each geometric object.

you need to know which crs (projection) you are using to get the correct output you want. That means **if your data is not projected into projection where meters are used, then the output will not be in meters**.

Obtain and plot the 500m, 1000m, and 2000m buffers for each urban park in the city of Beijing.

Overlay

Overlay analyses are GIS operations in which two or more vector layers are combined to produce new geometries. Typical overlay operations include union, intersection, and difference - named after the results of the combination of two layers.



Source: QGIS Documentation

Overlay

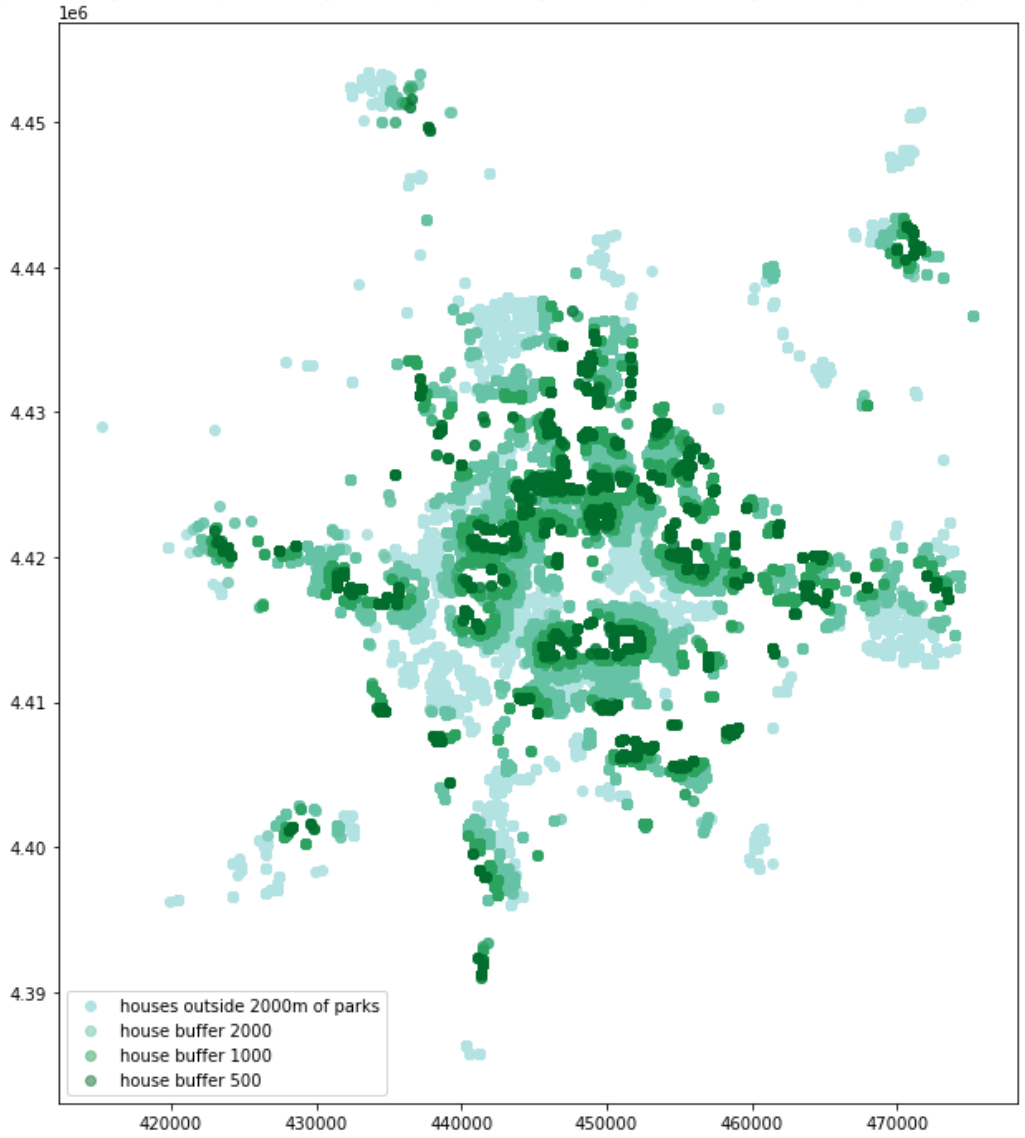
Let's say we want to identify those public housing within the 800m threshold and those that are not. We can use the intersection function in `gpd.overlay()`.

- create 800m buffer area of subway stations using `.buffer()`
- convert public houisng from polygon to point using `.centroid()`
- obtian the public housing within the 800m threshold using `intersection` option.
- obtain the public housing outside the 800m threshold using `difference` option.

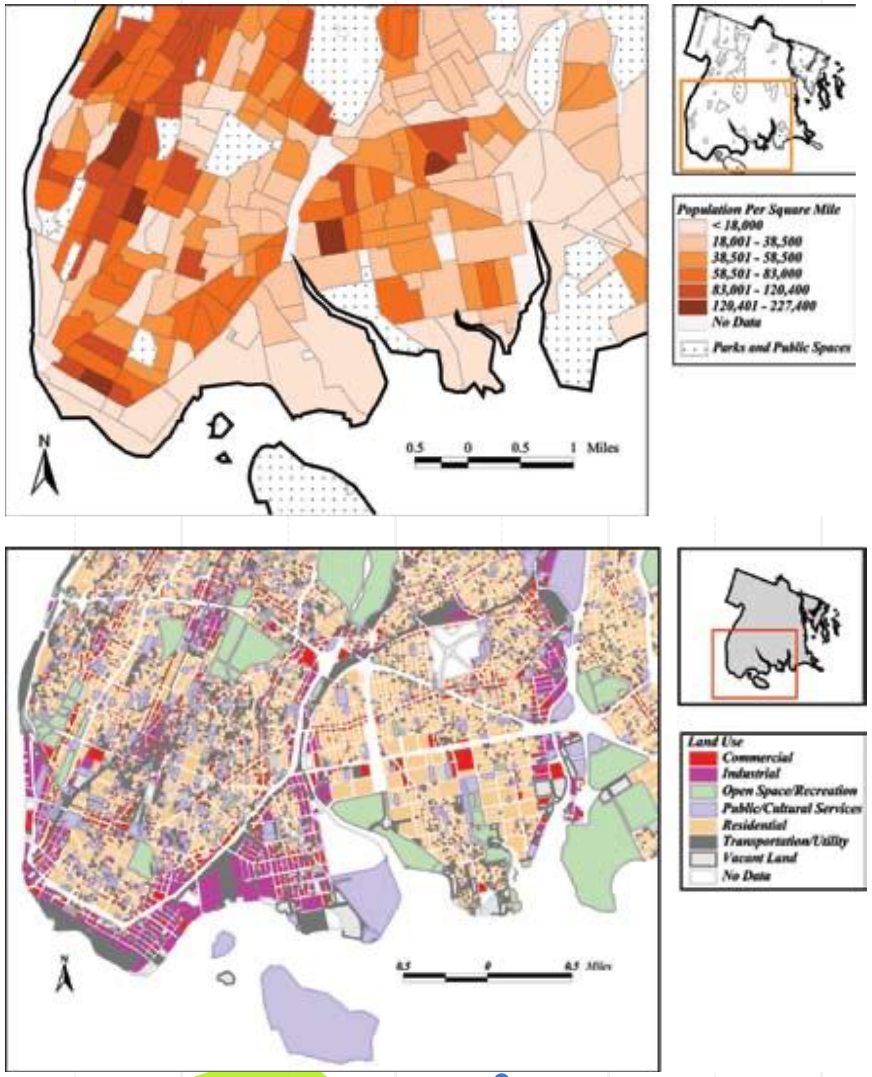
Overlay

Imagine one wishes to identify the houses located within 500m, 1000m, and 2000m, respectively. We can use gpd.overlay() to obtain the housing subsets we want.

```
gpd.overlay(gdf_combine_prj, park_500, how='intersection')
gpd.overlay(gdf_combine_prj, park_1000, how='intersection')
gpd.overlay(gdf_combine_prj, park_2000, how='intersection')
```

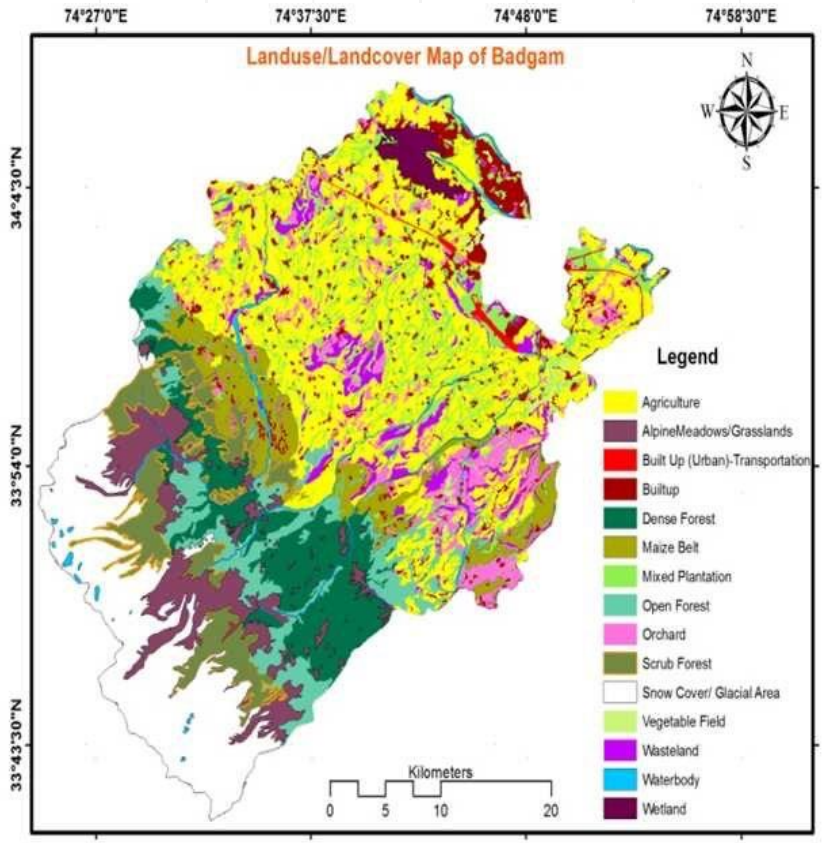


Maps – Quantitative vs. Qualitative Map



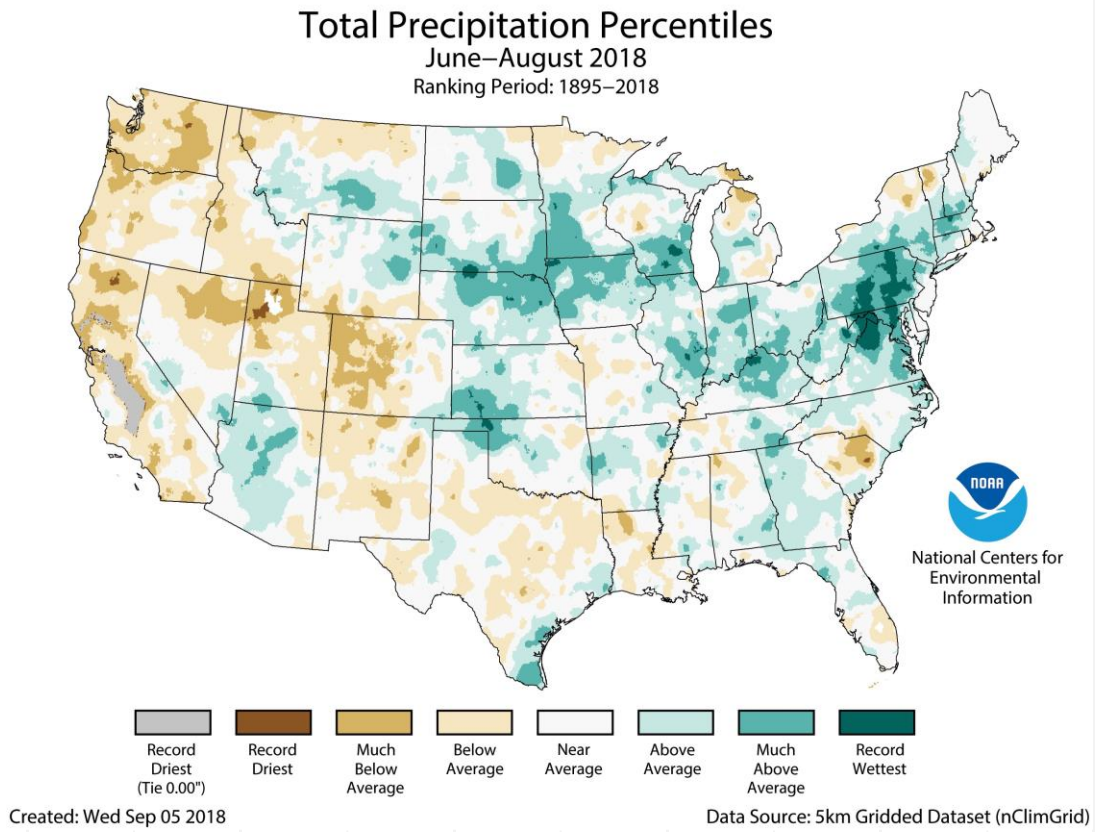
- Quantitative maps have as their basis the numerical relationships of the variables being mapped.
- Qualitative maps, by contrast, are based on descriptive information, and show location and boundaries of differences of kind or type.
- Qualitative maps make use of **nominal** or **ordinal** measurement scales, while quantitative maps make use of **interval** or **ratio** measurement scales.

Qualitative Thematic Maps — Measurement scales: Nominal



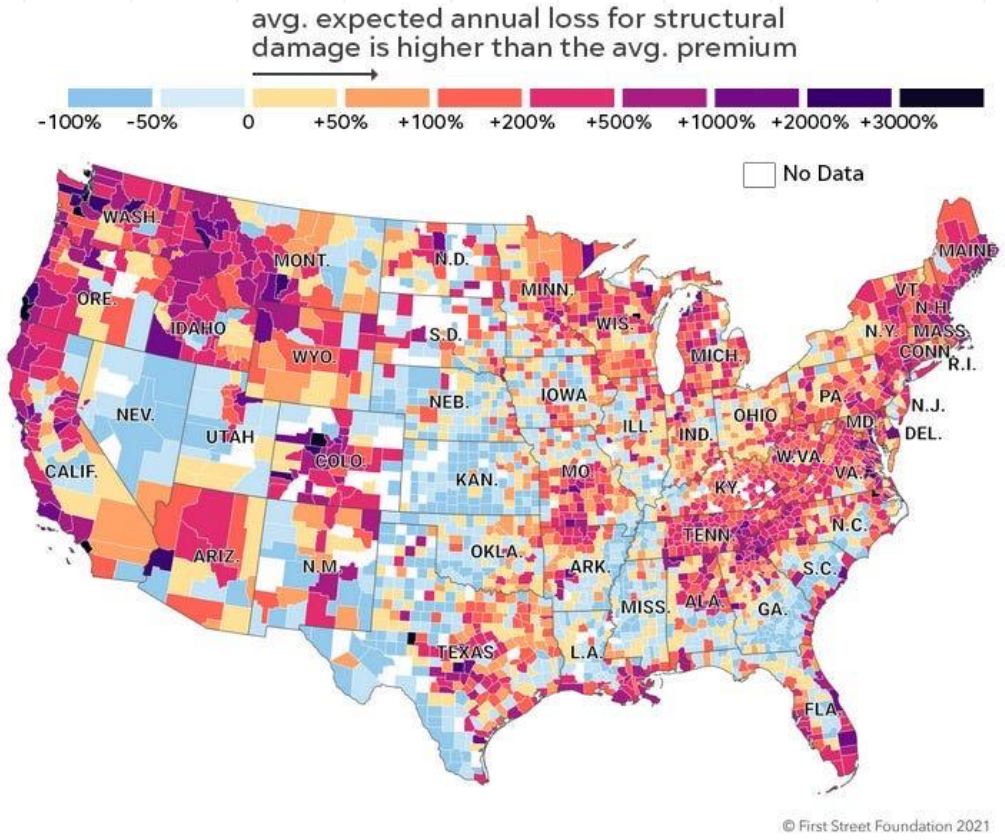
Objects are classified to groups. The groups have names, not numeric values.

Qualitative Thematic Maps — Measurement scales: Ordinal



- Implies a hierarchy of rank—a ranking of classes.

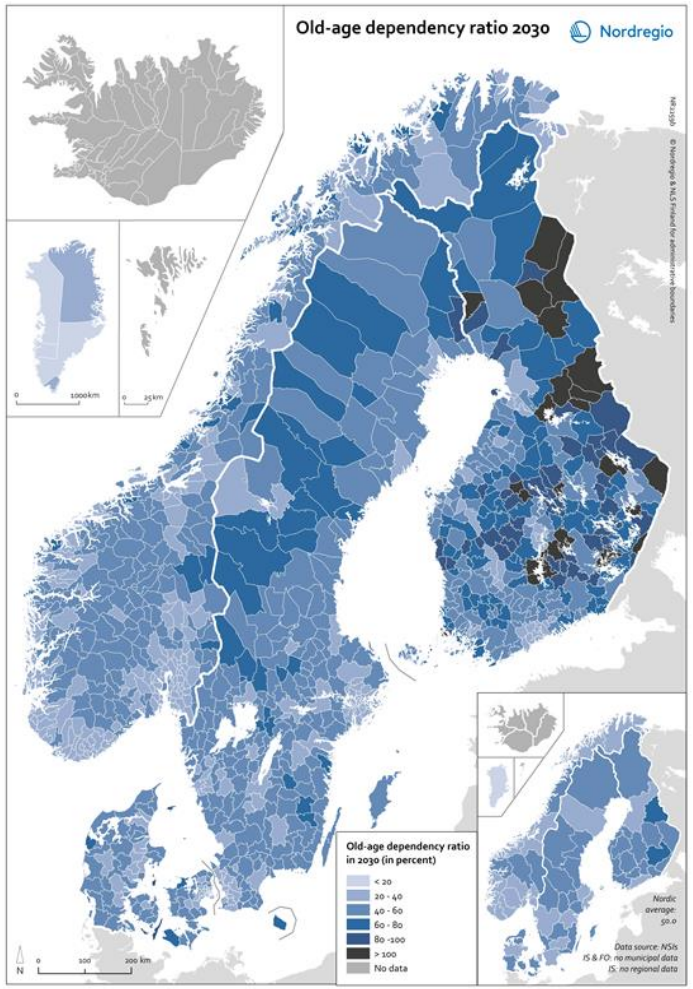
Quantitative Thematic Maps — Measurement scales: Interval



- Arrange the classes in ranks, and the intervals between ranks are known.

Quantitative Thematic Maps — Measurement scales: Ratio

- Ratio–scale magnitudes are absolute, and have a known starting point of zero.



Final Project/Poster (due??)

3

Spatial Analysis

Think about your final project!

- determine relationships
- detect and quantify patterns, assess trends
- make predictions and decisions.

- Step 1: Frame the **question** you want to answer (proposal)

  • What is the spatial distribution of asthma cases among children in New Jersey? What factors (demographic, environmental, etc.) might help explain the spatial distribution of asthma cases?

  • Where is the best site a retail establishment/school/public park should be located?
    - What is the distribution of the current private/public facilitates?
    - Where are the places that are underserved, and why?

  • Which areas/communities are the most severely affected by flooding? (considering social-economic status, infrastructural development, flooding frequency/severity)

Spatial Analysis

Think about your final project!

- Step 2: Find and prepare the data to make it ready for analysis (We will cover this part later this week).

- Step 3: Explore the data both with and without mapping
  - Non-mapping: descriptive tables, scatterplots, histograms, heatmap, etc.
  - Maps (generated by spatial methods: spatial join, spatial overlay, buffer, etc.)

- Step 4: Perform the spatial analysis, using the right tool or set of tools to answer the question
  - Spatial autocorrelation (Global/Local), point pattern analysis
  - Regression analysis, spatial regressions
  - Clustering/classification analysis
  - Predictions?

- Step 5: Share your results to communicate findings or allow others to repeat the process

Q&A

What is Geographic Information System?

Three pillars of GIS

- Technology
- Spatial Data
- Visual Communication

How to become a well-rounded GISicentist?

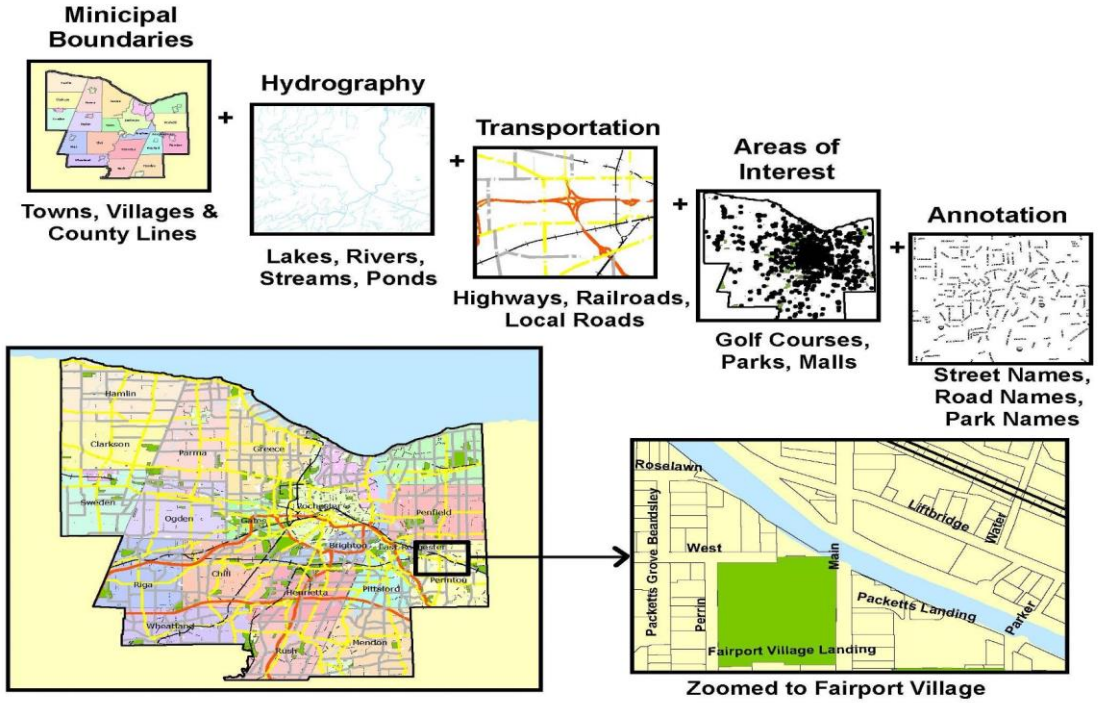
- Know Technology
- Know Analysis

GISystems

GIScience

GIStudies

Geoprocessing



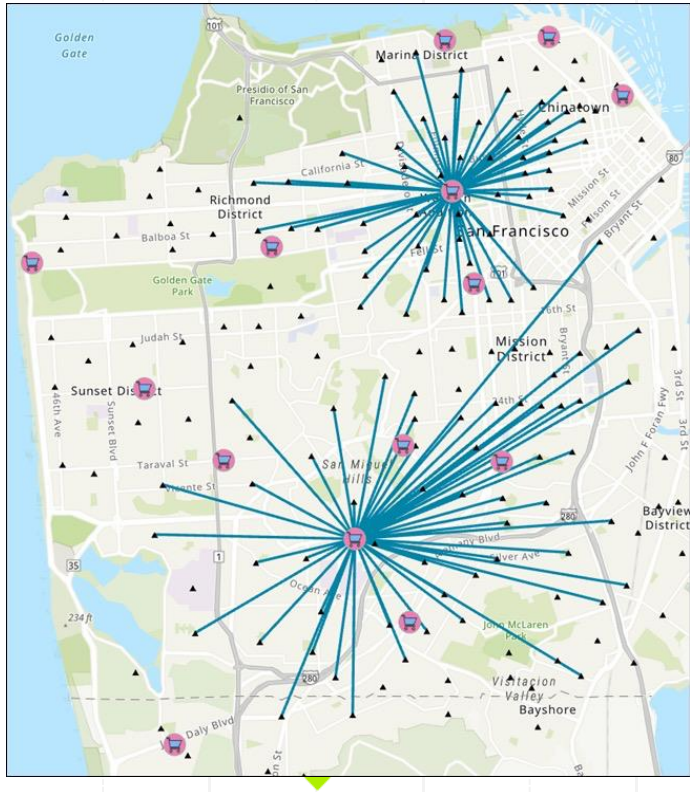
First step: **Overlay** spatial features/data to tell a story, make an argument, solve or understand a problem, etc, about **locations**.

Second step: Share your spatial analysis in reports, maps, tables, and charts through visual communication.

GIS in Planning

Business Planning

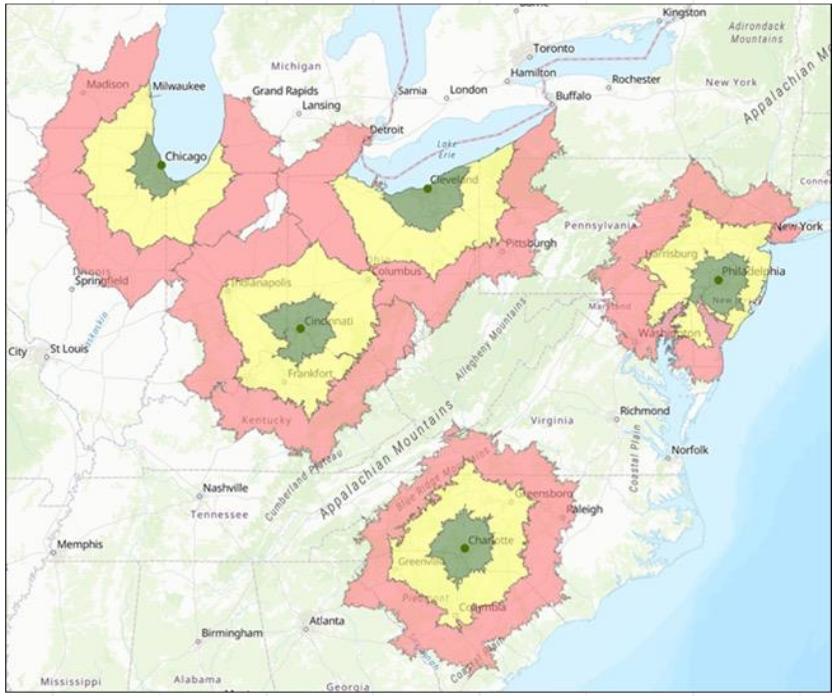
Where should we open a new branch of business?



Applying network analysis to business planning: What regions does the business cover? evaluate accessibility (green points represent warehouses and color code represent commuting time)

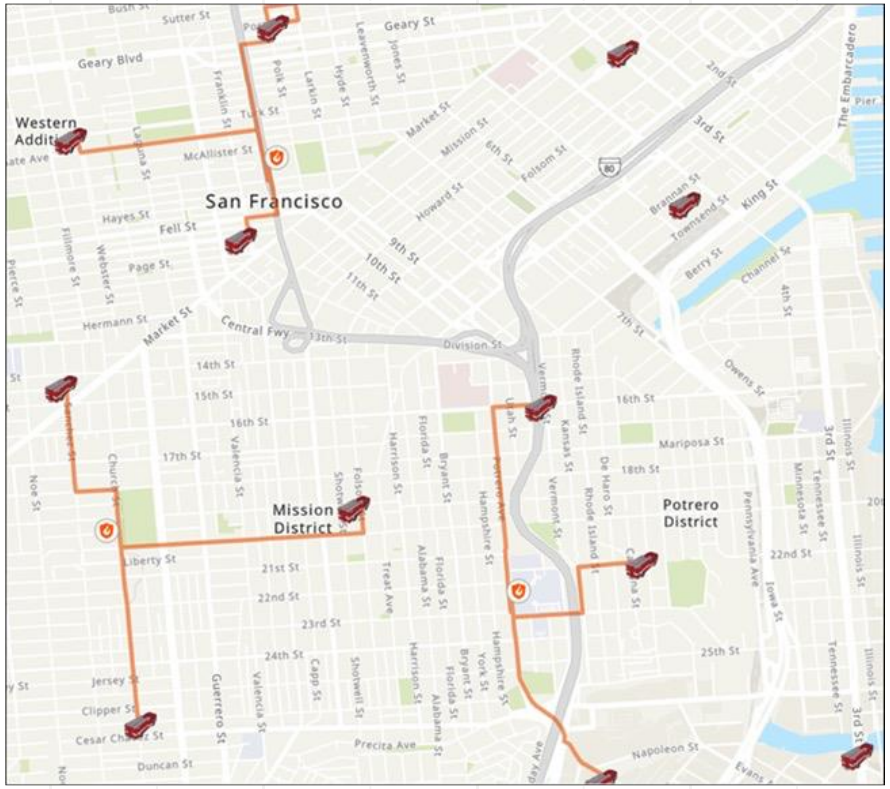
Further analysis the demographic composition

•



GIS in Planning

Emergency Service Planning



Applying network analysis to emergency service planning:

- Optimal planning of fire stations to cover all the fire incidents can be reached within 5 mins.

Types of Spatial Data in GIS

Vector data

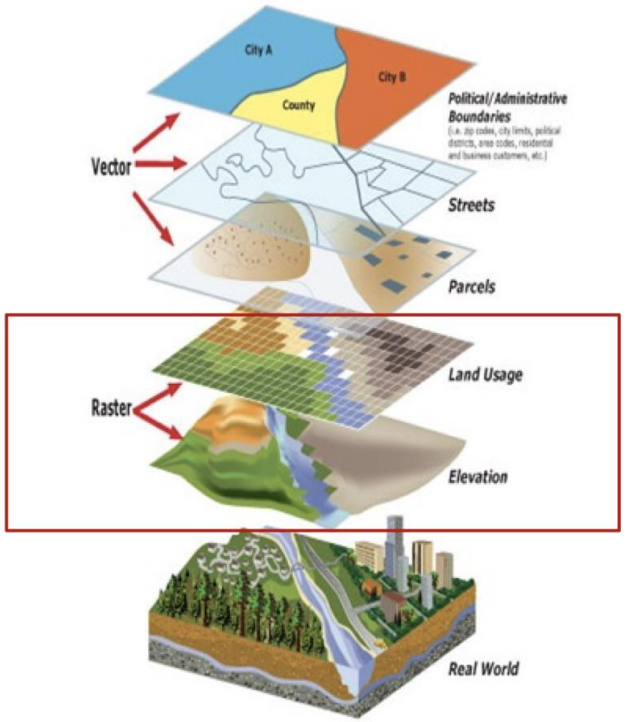
Disadvantages of vector data:

- Storage and data structure much more complex: The location of each vertex needs to be stored explicitly.
- Processing Intensive: For effective analysis, vector data must be converted into a topological structure. This is often processing intensive and usually requires extensive data cleaning. Topology is static, and any updating or editing of the vector data requires re-building of the topology.
- Speed: Algorithms for manipulative and analysis functions are complex and may be processing intensive. Often, this inherently limits the functionality for large data sets
- Spatial limitations: Continuous data, such as elevation data, is not effectively represented in vector form. Usually substantial data generalization or interpolation is required for these data layers.

Types of Spatial Data in GIS

Raster Data

Advantages of raster data:



An example of map layers used in GIS
Source: San Berardino County GIS Dept, 2012

- Storage: Each grid location represented in the raster image correlates to a single value (or series of values if attributes tables are included).
- Speed: Due to the nature of the data storage technique data analysis is usually easy to program and quick to perform.
- Processing: The inherent nature of raster maps is ideally suited for mathematical modeling and quantitative analysis.
- Monitoring continuous data: useful when tracking physical changes to the landscape overtime, since raster data, like satellite imagery, is collected much more frequently than vector data.

Maps

Maps where the color of each shape is based on the value of an associated variable.
Check a wonderful introduction of Choropleth Maps in the following link made by *Story Map Journal application in ArcGIS Online*.
https://www.arcgis.com/apps/MapJournal/index.html?appid=75eff041036d40cf8e70df99641004ca