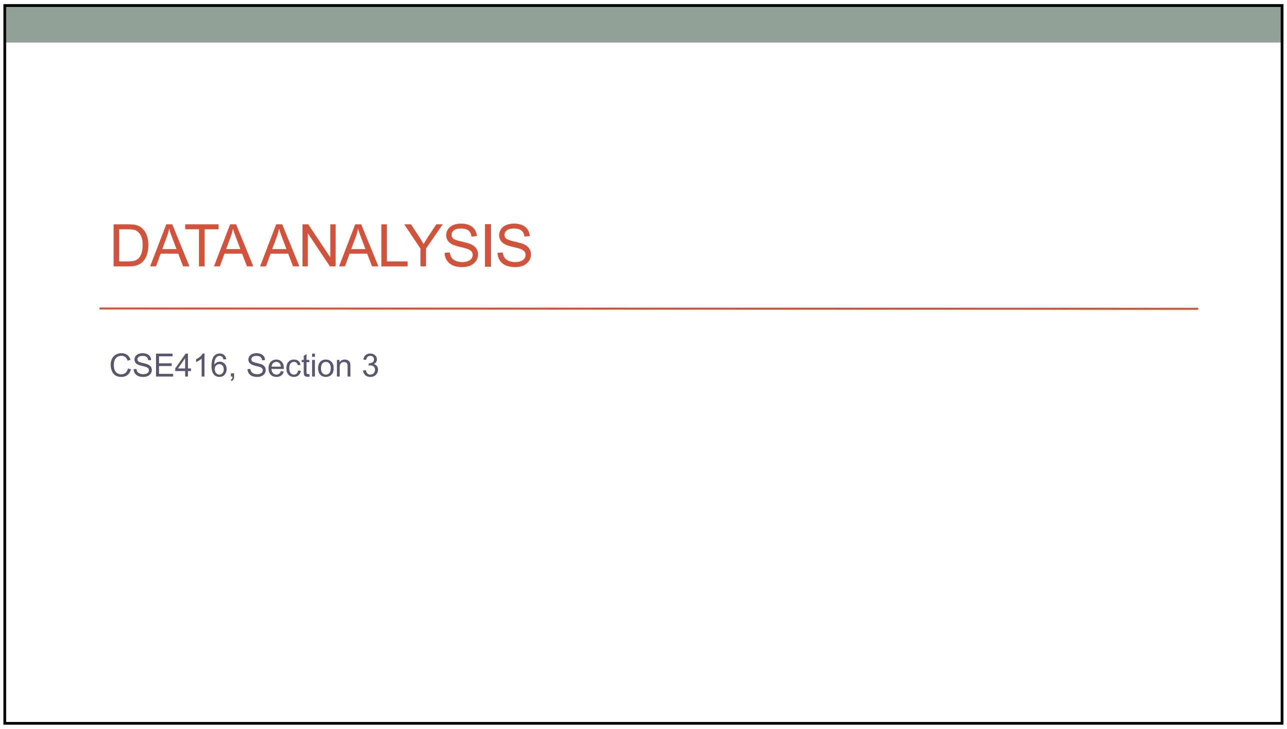
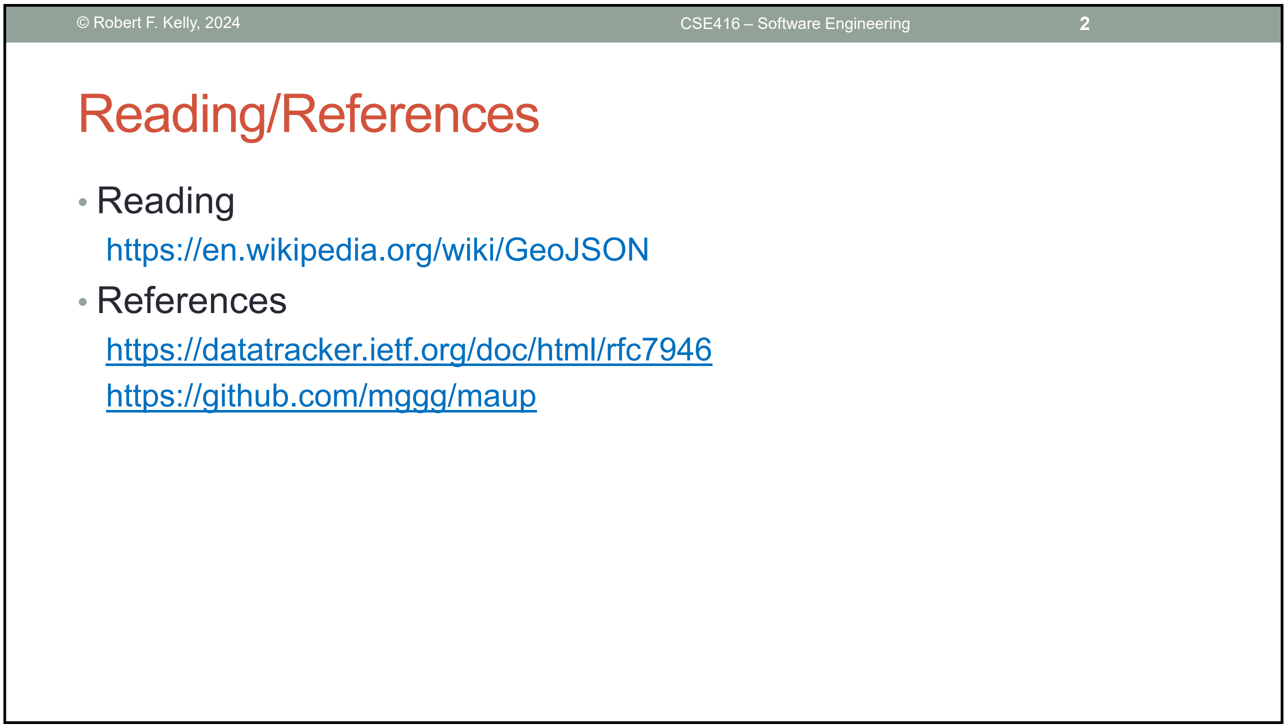


DATA ANALYSIS

CSE416, Section 3

1



Reading/References

- Reading

  https://en.wikipedia.org/wiki/GeoJSON

- References

  https://datatracker.ietf.org/doc/html/rfc7946

  https://github.com/mggg/maup

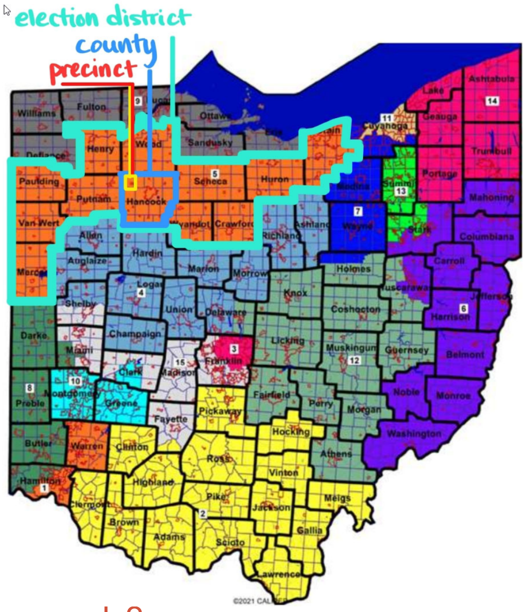
2

Session Objectives

- Begin to plan early project activities
- Understand data requirements
- Understand options for accessing and processing currently available data

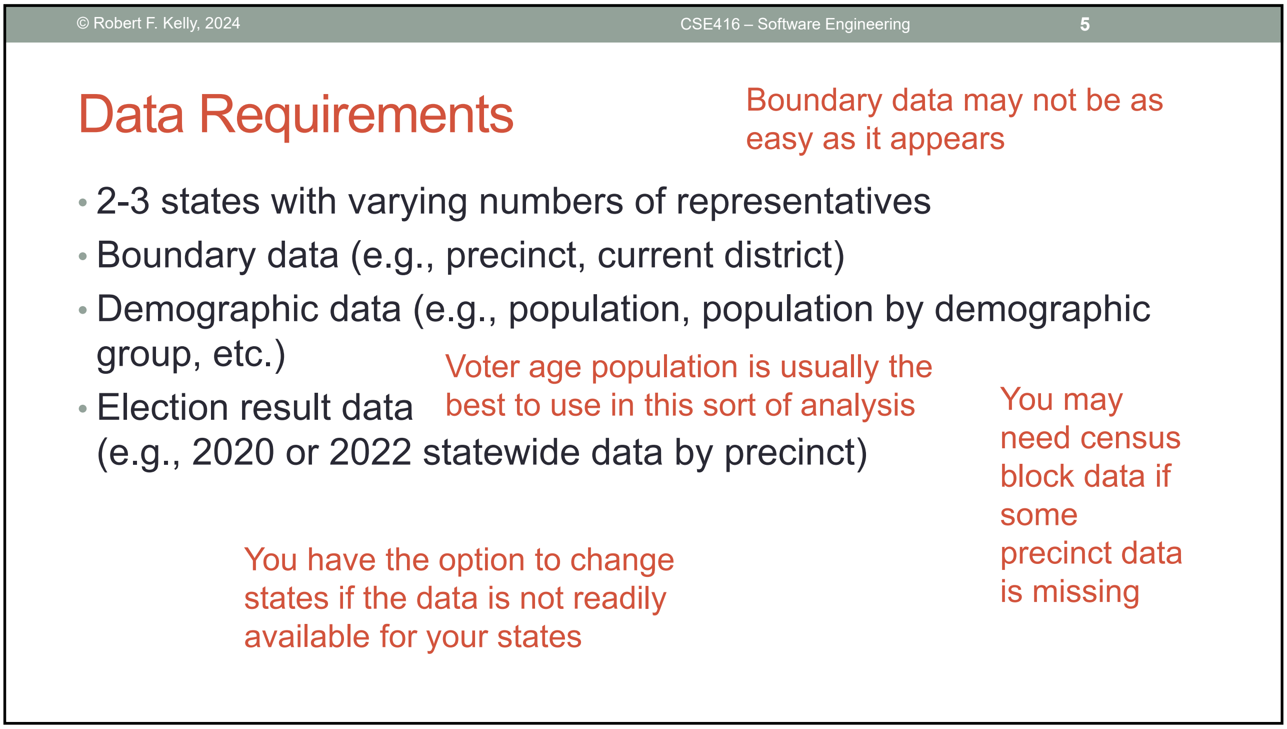
Project Goal

- Evaluate whether the Fair Representation Act for multi-member election districts (MMD) will increase fairness and lessen the effects of Gerrymandering in Congressional elections



What data do you need to achieve these goals?

Image: statenews.org



CSE416 – Software Engineering 5

Data Requirements

Boundary data may not be as easy as it appears

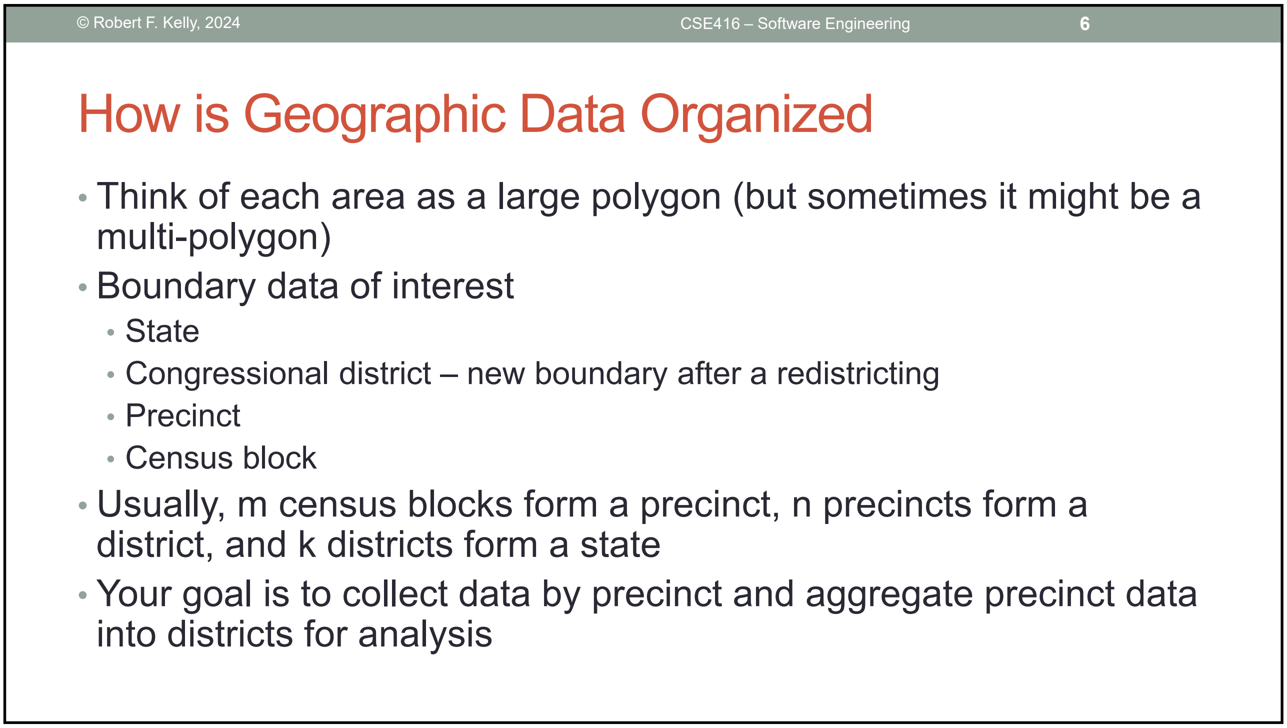
- 2-3 states with varying numbers of representatives
- Boundary data (e.g., precinct, current district)
- Demographic data (e.g., population, population by demographic group, etc.)
- Election result data (e.g., 2020 or 2022 statewide data by precinct)

Voter age population is usually the best to use in this sort of analysis

You may need census block data if some precinct data is missing

You have the option to change states if the data is not readily available for your states

5



CSE416 – Software Engineering 6

How is Geographic Data Organized

- Think of each area as a large polygon (but sometimes it might be a multi-polygon)
- Boundary data of interest
  - State
  - Congressional district – new boundary after a redistricting
  - Precinct
  - Census block
- Usually, m census blocks form a precinct, n precincts form a district, and k districts form a state
- Your goal is to collect data by precinct and aggregate precinct data into districts for analysis

6

Shapefiles

- Geospatial vector data format
- Developed and maintained by ESRI
- Introduced in early 1990s
- Collection of files
  - Usually stored as a zip file
  - Mandatory files (.shp, .shx, and .dbf) and other files
- Represents points, lines, polygons
- Formatted as fixed length header, followed by one or more variable length records

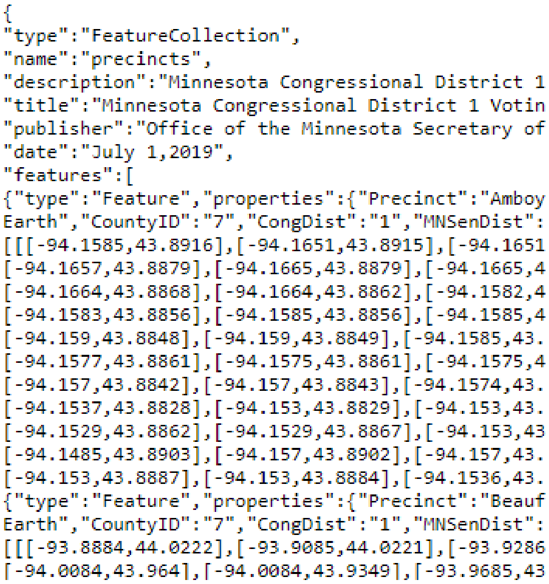
Dominant format for geographic data due to the market dominance of ESRI

GeoJSON

Be alert for MultiPolygon data

- Open standard format for representing simple geometric features
- Based on JSON
- Types – Point, LineString, Polygon, MultiPolygon
- Supported by Leaflet, Google Maps, et al
- Position information expressed as longitude, latitude

Become familiar with conversion SW



```
{
"type":"FeatureCollection",
"name":"precincts",
"description":"Minnesota Congressional District 1
"title":"Minnesota Congressional District 1 Votin
"publisher":"Office of the Minnesota Secretary of
"date":"July 1,2019",
"features":[
{"type":"Feature","properties":{"Precinct":"Amboy
Earth","CountyID":"7","CongDist":"1","MNSenDist":
[[[-94.1585,43.8916],[-94.1651,43.8915],[-94.1651
[-94.1657,43.8879],[-94.1665,43.8879],[-94.1665,4
[-94.1664,43.8868],[-94.1664,43.8862],[-94.1582,4
[-94.1583,43.8856],[-94.1585,43.8856],[-94.1585,4
[-94.159,43.8848],[-94.159,43.8849],[-94.1585,43.
[-94.1577,43.8861],[-94.1575,43.8861],[-94.1575,4
[-94.157,43.8842],[-94.157,43.8843],[-94.1574,43.
[-94.1537,43.8828],[-94.153,43.8829],[-94.153,43.
[-94.1529,43.8862],[-94.1529,43.8867],[-94.153,43
[-94.1485,43.8903],[-94.157,43.8902],[-94.157,43.
[-94.153,43.8887],[-94.153,43.8884],[-94.1536,43.
{"type":"Feature","properties":{"Precinct":"Beauf
Earth","CountyID":"7","CongDist":"1","MNSenDist":
[[[-93.8884,44.0222],[-93.9085,44.0221],[-93.9286
[-94.0084,43.964],[-94.0084,43.9349],[-93.9685,43
```

Non-Geographic Data

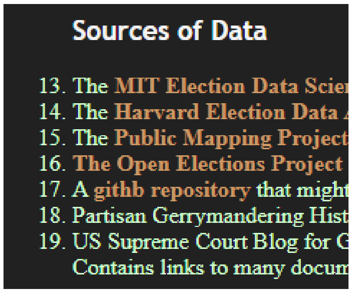
- Election results data
- Population data
  - Total population
  - Voting age population (VAP)
  - Citizen voting age population(CVAP)
- Demographic data
  - Racial/ethnic
  - Income

If you cannot get data by precinct, you may need to sum up contained census blocks

Consistently use one category of population data – VAP is best

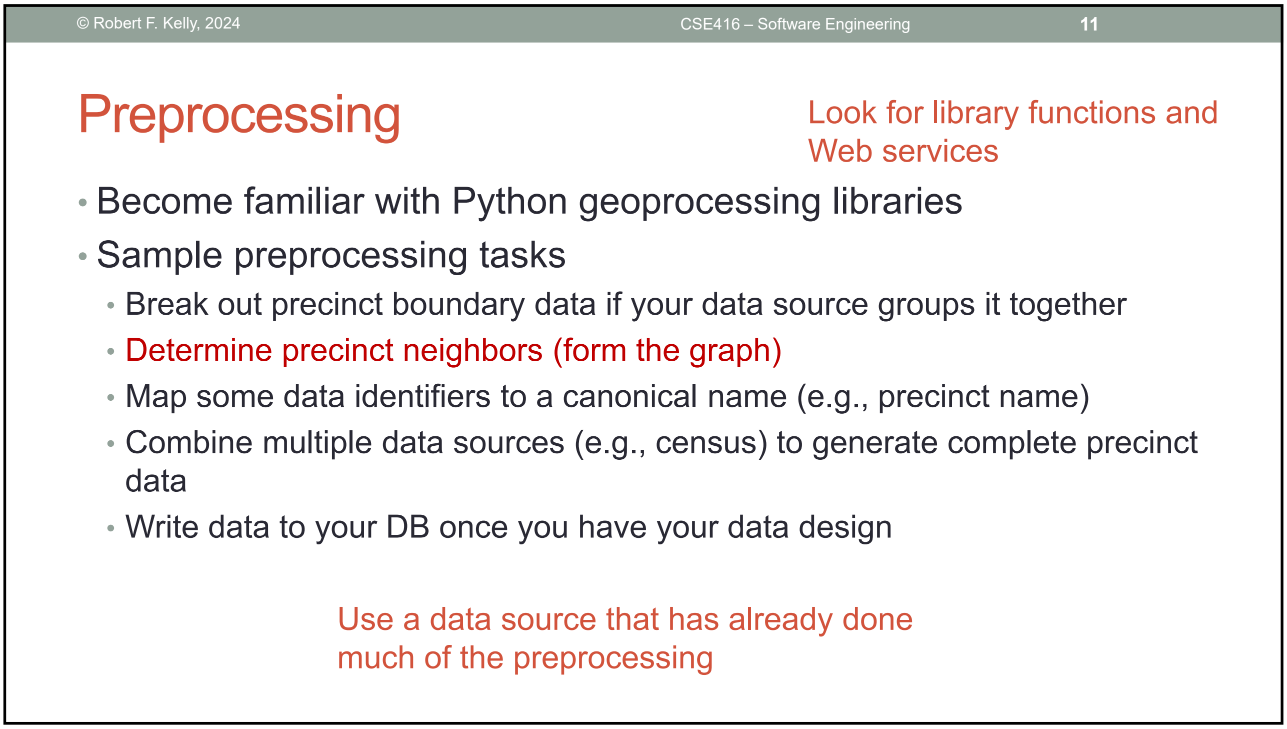
Sources of Data

- Project Web site suggests many sources
- For example
  - Redistricting Data Hub
  - Harvard
  - MIT
  - US Census Bureau
  - Open Elections
- Easier sources of data (including some consolidation) are available
- Choose a data source that provides data at a level you need for your states



Sources of Data

13. The MIT Election Data Scier
14. The Harvard Election Data /
15. The Public Mapping Project
16. The Open Elections Project
17. A githb repository that might
18. Partisan Gerrymandering Hist
19. US Supreme Court Blog for G
    Contains links to many docum



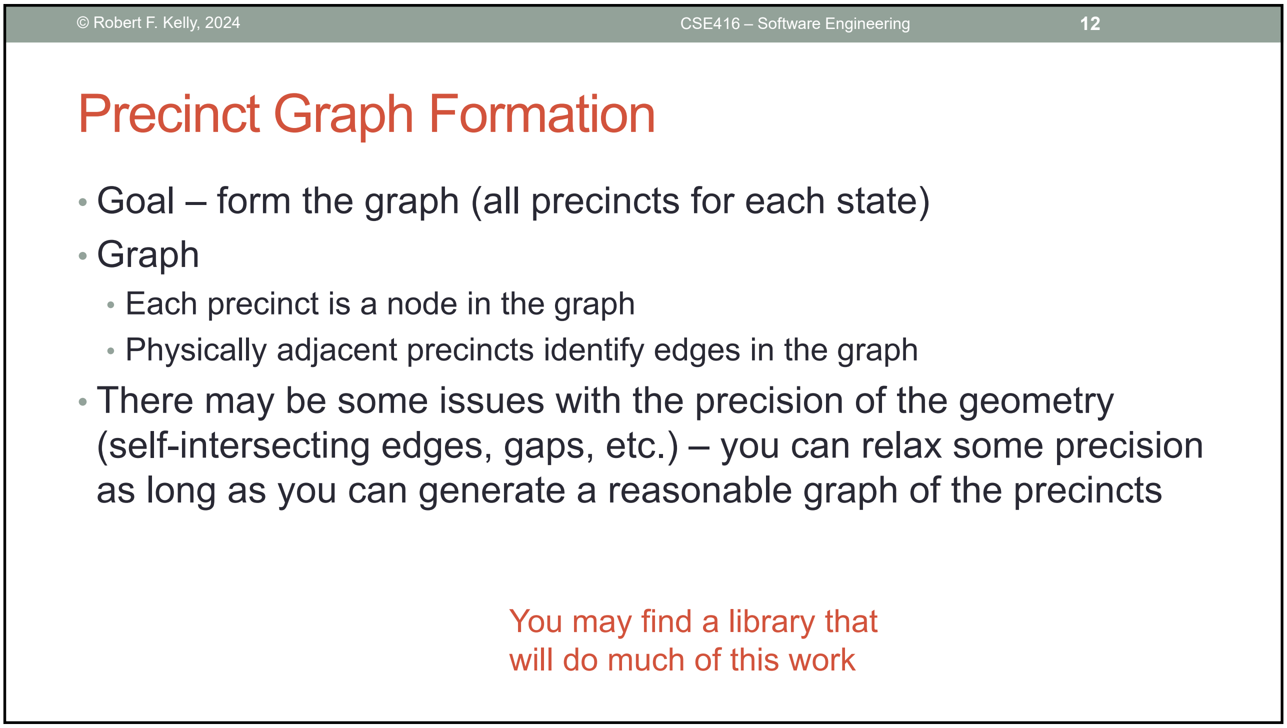
Preprocessing

Look for library functions and Web services

- Become familiar with Python geoprocessing libraries
- Sample preprocessing tasks
  - Break out precinct boundary data if your data source groups it together
  - Determine precinct neighbors (form the graph)
  - Map some data identifiers to a canonical name (e.g., precinct name)
  - Combine multiple data sources (e.g., census) to generate complete precinct data
  - Write data to your DB once you have your data design

Use a data source that has already done much of the preprocessing

11

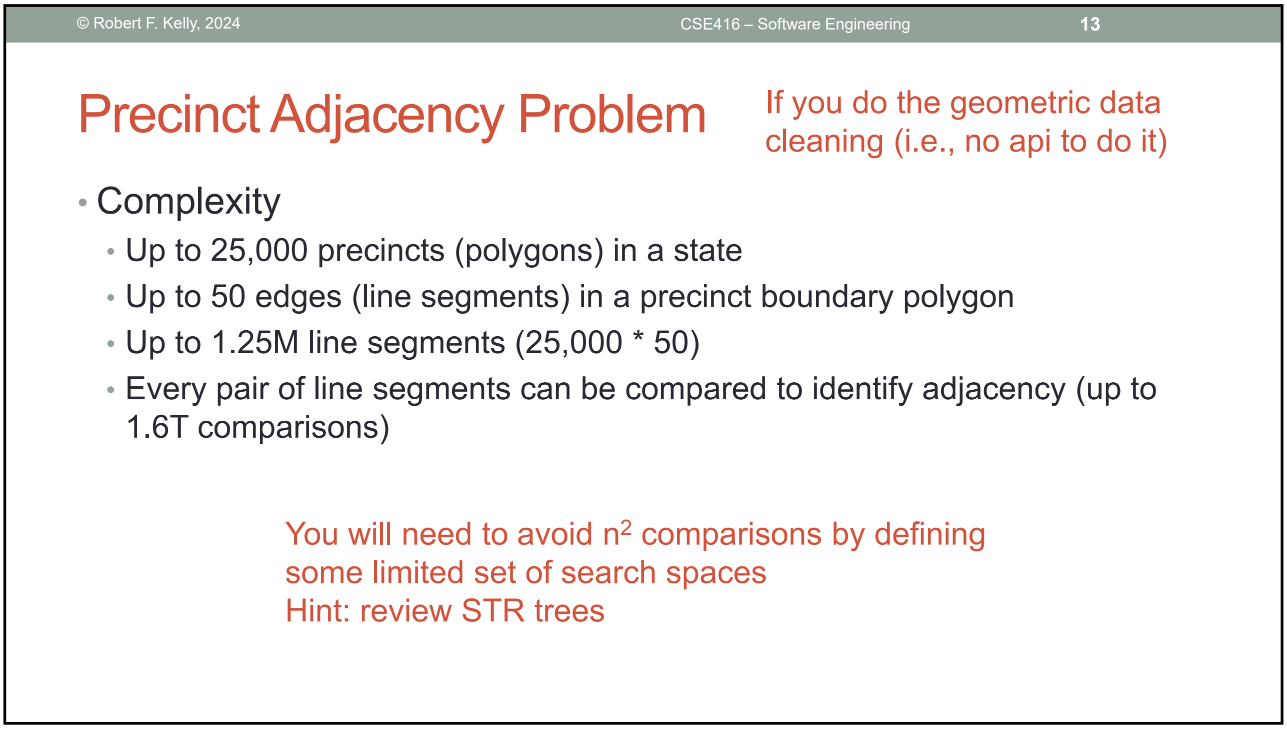


Precinct Graph Formation

- Goal – form the graph (all precincts for each state)
- Graph
  - Each precinct is a node in the graph
  - Physically adjacent precincts identify edges in the graph
- There may be some issues with the precision of the geometry (self-intersecting edges, gaps, etc.) – you can relax some precision as long as you can generate a reasonable graph of the precincts

You may find a library that will do much of this work

12

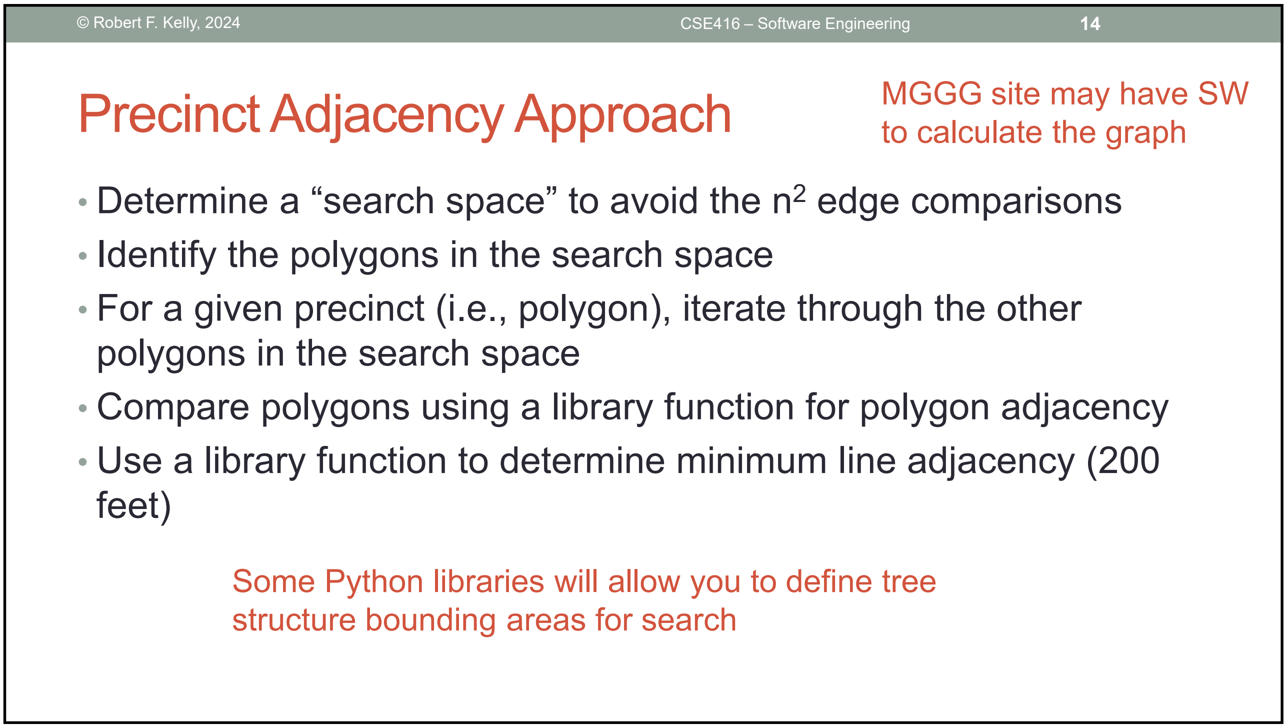


Precinct Adjacency Problem

If you do the geometric data cleaning (i.e., no api to do it)

- Complexity
  - Up to 25,000 precincts (polygons) in a state
  - Up to 50 edges (line segments) in a precinct boundary polygon
  - Up to 1.25M line segments (25,000 * 50)
  - Every pair of line segments can be compared to identify adjacency (up to 1.6T comparisons)

You will need to avoid n^2 comparisons by defining some limited set of search spaces
Hint: review STR trees

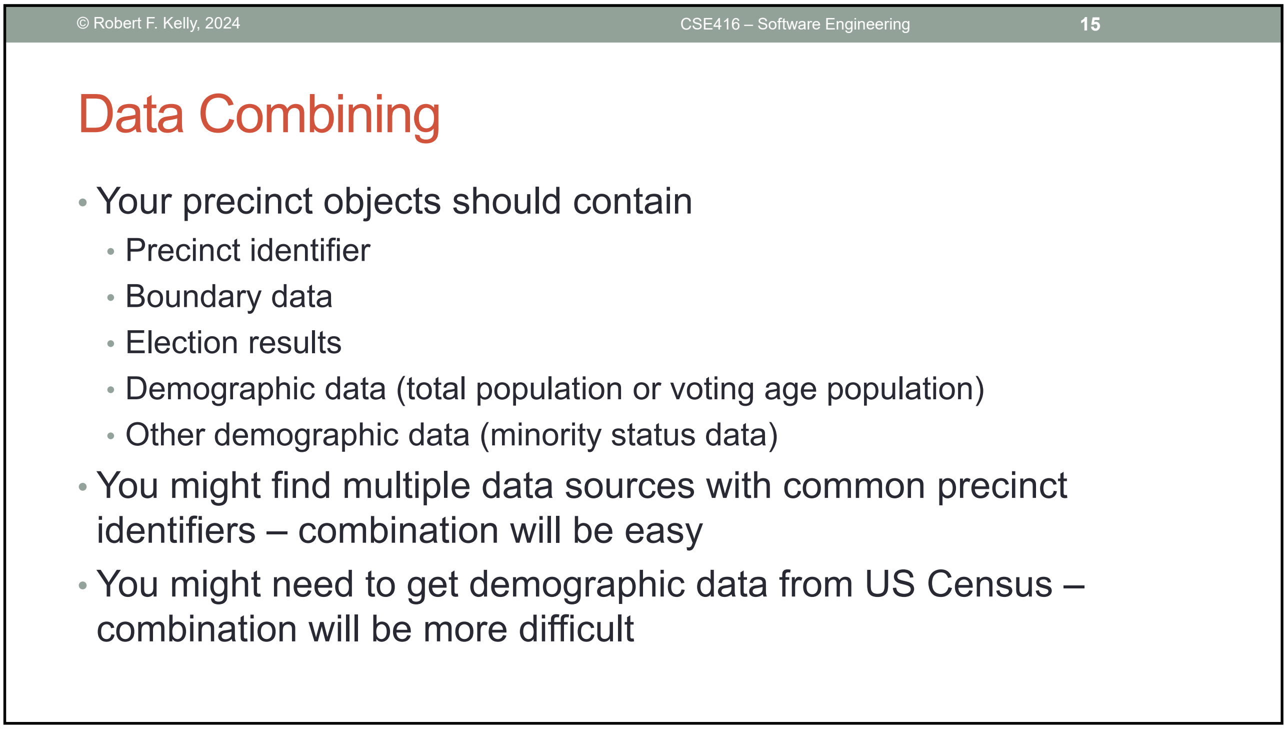


Precinct Adjacency Approach

MGGG site may have SW to calculate the graph

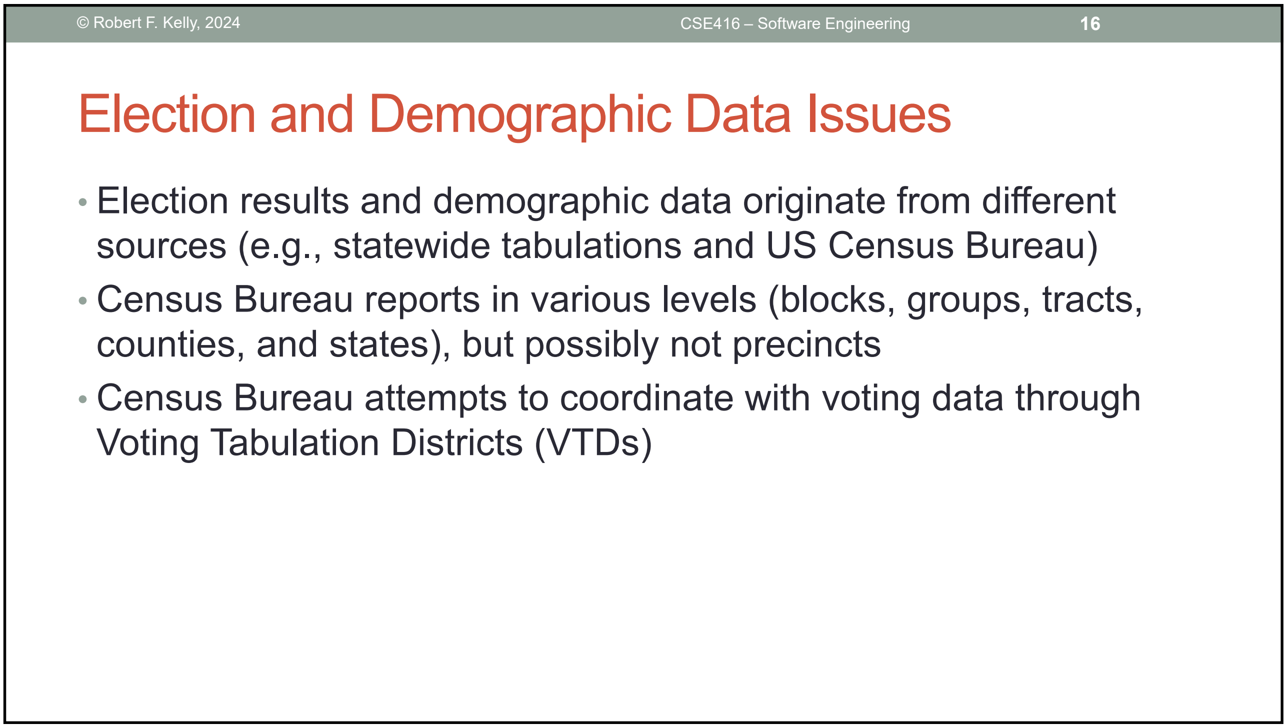
- Determine a “search space” to avoid the n^2 edge comparisons
- Identify the polygons in the search space
- For a given precinct (i.e., polygon), iterate through the other polygons in the search space
- Compare polygons using a library function for polygon adjacency
- Use a library function to determine minimum line adjacency (200 feet)

Some Python libraries will allow you to define tree structure bounding areas for search



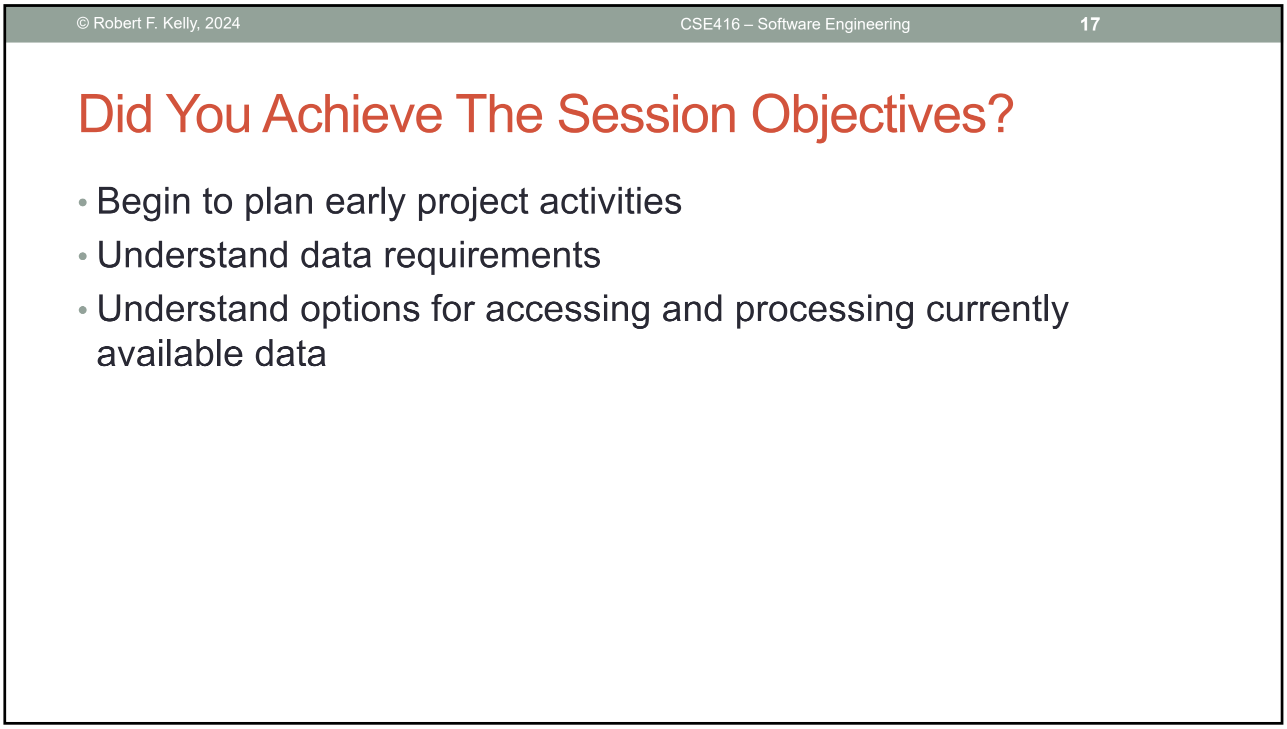
Data Combining

- Your precinct objects should contain
  - Precinct identifier
  - Boundary data
  - Election results
  - Demographic data (total population or voting age population)
  - Other demographic data (minority status data)
- You might find multiple data sources with common precinct identifiers – combination will be easy
- You might need to get demographic data from US Census – combination will be more difficult



Election and Demographic Data Issues

- Election results and demographic data originate from different sources (e.g., statewide tabulations and US Census Bureau)
- Census Bureau reports in various levels (blocks, groups, tracts, counties, and states), but possibly not precincts
- Census Bureau attempts to coordinate with voting data through Voting Tabulation Districts (VTDs)



Did You Achieve The Session Objectives?

- Begin to plan early project activities
- Understand data requirements
- Understand options for accessing and processing currently available data