DCP4300/URP6931. AI&BE

# Lecture 03: Urban Data Description and Visualization

Instructor: Shenhao Wang
Assistant Professor, Director of Urban AI Lab
Department of Urban and Regional Planning
University of Florida

Reminder: Use Zoom Video Recording

# Announcements

1. Voting for office hour.
2. Assignments (Heads-up: lab sessions are important.)
3. Practicum AI.

# Outline

1. Data landscape: an example in Chicago
2. Census/ACS and Shapefiles
3. Python Lab 3.1: Descriptive data analysis
4. Python Lab 3.2: Spatial data visualization
5. Python Lab 3.3: Census data collection & processing [May be skipped]

# Part 1. Data landscape: an example in Chicago

Github repository: https://github.com/sunnyqywang/Chicago-Integrated-Data-Repo.git

Motivation and Background

1. Data is the fuel to research. However, researchers are increasingly limited by data availability, since private companies hold more data than researchers.
2. Summarizing a “complete” list of data sets for Chicago, with an emphasis on urban mobility. Such a list is generally applied to other places.

# Ten data sources

1. Meta data sets.
2. Socio-demographics from Census or American Community Surveys (ACS)
3. GIS shapefiles
4. Mobility flow
5. Spatiotemporal ridership
6. Travel surveys
7. Urban imagery
8. Point-of-Interests
9. Other mobility-related data in public health, air pollution, and energy.
10. Web scraping.

**Notes** – some redundancy exists in the ten categories. But I target the completeness for this introduction.

# 1. Meta data sets

Most useful but commonly ignored: universities, research centers, governments, and individual researchers have collected useful meta information about data sources.

**Northwestern University**

https://libguides.northwestern.edu/c.php?g=114808&p=748275

**University of Illinois at Chicago**

http://csun.uic.edu/datasets.html

**City of Chicago Data Portal**

https://data.cityofchicago.org/

**NREL Transport Data Center**

https://www.nrel.gov/transportation/secure-transportation-data/

# Northwestern university library

Link
https://libguides.northwestern.edu/c.php?g=114808&p=748275



## Chicago Area Transportation

Search this Guide Search

- Getting Started
- Local Transportation Agencies and Operators
- Books and Articles
- Statistics and Data
  - General Sources for Transportation Statistics and Data
  - Local Data
- Historical Information
- Financial Data

### General Sources for Transportation Statistics and Data

- Bureau of Transportation Statistics
  BTS data collections include traffic, passenger flow, employment, financial condition, and on-time performance of commercial aviation; the Commodity Flow Survey; transborder movement of freight by mode of transportation; a census of ferry operations, precursor safety data for transit operations, and data on near misses and equipment failures in offshore operations.
- TranStats
  A searchable index of over 100 transportation-related data bases across every mode of transportation — with many social and demographic data sets that are commonly used in transportation analysis.
- ProQuest Statistical Insight
  ProQuest Statistical Insight provides fast and easy access to statistical information produced by U.S. Federal agencies, States, private organizations, and major intergovernmental organizations. PQSI brings together a massive collection of statistical data in a single easy-to-use search interface, with additional features such as descriptive abstracts, detailed indexing, full-text PDFs of source documents, tables, and downloadable spreadsheets containing table data.
- Statistical Abstract of the United States
  The Statistical Abstract of the United States, published since 1878, is the authoritative and comprehensive summary of statistics on the social, political, and economic organization of the United States.
- Centers for Disease Control and Prevention
  Provides statistics on transportation related deaths and injuries

### Local Data

- Air Traffic Data
  The Chicago Department of Aviation (CDA) compiles monthly statistics, including flight operations, passenger totals and cargo tonnage. The data is provided by the airlines and summarized by the CDA. Use the drop downs below to view all air traffic data from O'Hare and Midway International Airports from as far back as the year 2000.
- Average Daily Traffic Counts
  Average Daily Traffic (ADT) volumes, truck volumes, road construction locations, and more for specific locations in Illinois

# 2. Socio-demographics data from Census & ACS

**Census (American Community Survey):** most high-quality socio-demographic and socioeconomic data sets in USA

https://www.census.gov/programs-surveys/acs

But to download the ACS data easily, you should use Python packages

**e.g. Python CensusData**

https://github.com/jtleider/censusdata

https://jtleider.github.io/censusdata/

# Census (ACS) data example

**Rows**: spatial units (e.g., census tracts)

**Columns**: variables

| | pop_total | sex_total | sex_male | sex_female | age_median | households | race_total | race_white | race_black | race_native | ... | travel_walk_ratio | travel_work_home_ratio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2812.0 | 2812.0 | 1383.0 | 1429.0 | 39.4 | 931.0 | 2812.0 | 2086.0 | 517.0 | 0.0 | ... | 0.014815 | 0.024242 |
| **1** | 4709.0 | 4709.0 | 2272.0 | 2437.0 | 34.2 | 1668.0 | 4709.0 | 2382.0 | 1953.0 | 0.0 | ... | 0.022150 | 0.004615 |
| **2** | 5005.0 | 5005.0 | 2444.0 | 2561.0 | 34.1 | 1379.0 | 5005.0 | 2334.0 | 2206.0 | 224.0 | ... | 0.026141 | 0.027913 |
| **3** | 6754.0 | 6754.0 | 2934.0 | 3820.0 | 31.3 | 2238.0 | 6754.0 | 4052.0 | 1671.0 | 326.0 | ... | 0.052697 | 0.004054 |
| **4** | 3021.0 | 3021.0 | 1695.0 | 1326.0 | 44.1 | 1364.0 | 3021.0 | 2861.0 | 121.0 | 0.0 | ... | 0.003014 | 0.013059 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **4162** | 15742.0 | 15742.0 | 7957.0 | 7785.0 | 41.0 | 5517.0 | 15742.0 | 13894.0 | 1128.0 | 64.0 | ... | 0.000000 | 0.062212 |
| **4163** | 5723.0 | 5723.0 | 2914.0 | 2809.0 | 43.0 | 2001.0 | 5723.0 | 4664.0 | 482.0 | 0.0 | ... | 0.017050 | 0.047581 |
| **4164** | 10342.0 | 10342.0 | 4657.0 | 5685.0 | 37.6 | 3746.0 | 10342.0 | 7956.0 | 1351.0 | 13.0 | ... | 0.000000 | 0.038862 |
| **4165** | 8960.0 | 8960.0 | 4166.0 | 4794.0 | 37.2 | 3324.0 | 8960.0 | 6286.0 | 1831.0 | 0.0 | ... | 0.024021 | 0.064132 |
| **4166** | 5083.0 | 5083.0 | 2573.0 | 2510.0 | 39.2 | 1755.0 | 5083.0 | 3753.0 | 987.0 | 0.0 | ... | 0.045983 | 0.046488 |

4167 rows × 88 columns

# 3. Geographic Information System (GIS) Shapefiles

**Topologically Integrated Geographic Encoding and Referencing (TIGER)** from Census: including the hierarchy of census tracts, block groups, and blocks as the shapefile formats.

https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-data.html

**Local GIS information** is also provided by **universities and governments**

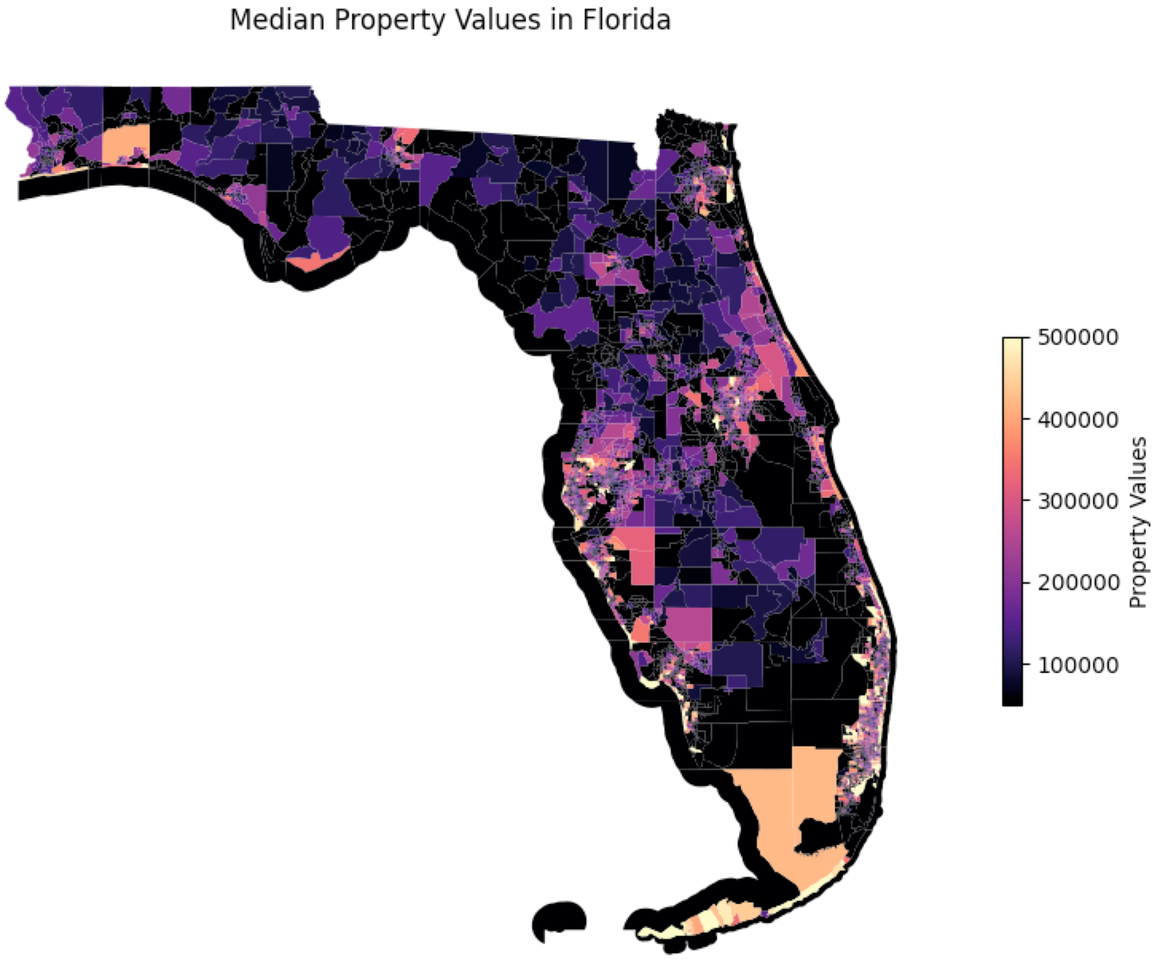
https://www.uis.edu/gis/projects/data/

Including the spatial information for transport networks (e.g. bus, transit, and Metra stations and lines, bike routes and racks, and many others

https://data.cityofchicago.org/Transportation/Bike-Racks-Map/4ywc-hr3a

**OpenStreetMap** also provides GIS shapefiles for buildings, land use, water, natural resources, and others. (See POI section)

# Shapefile examples



# 4. Mobility flow data

**Mobility flow data** refer to the network data sets that include the origin, destination, and the flow counts between origin-destination (OD) pairs.

**ACS Commuting** provided by census transportation planning products (CTPP), which is funded by USDOT.

https://www.fhwa.dot.gov/planning/census_issues/ctpp/

**Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES)**: it provides the home-work locations and the traffic flows between the ODs

https://lehd.ces.census.gov/data/#top

# 5. Spatiotemporal ridership

**Spatiotemporal ridership** refers to the real-time (individualized) mobility information, including vehicle ID, departure and arrival time, OD geospatial information with **high spatiotemporal resolution** (e.g., seconds; longitude & latitude).

**City of Chicago data portal**: providing the spatiotemporal travel flow data for a variety of travel modes, including buses, subway lines, Metra, TNCs, etc.

Links:

https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p

https://data.cityofchicago.org/Transportation/E-Scooter-Trips-2019-Pilot/2kfw-zvte

https://data.cityofchicago.org/browse?q=scooter&sortBy=relevance

# 6. Travel Survey

Travel surveys are typically conducted at the individual-level, and provide trip-level information, such as travel time, travel costs, and modes.

**National Household Travel Survey (2017)**: nation-wide travel survey data conducted by USDOT. Once every ten years.

https://nhts.ornl.gov/

**State Add-Ons to NHTS**: states can opt in to conduct refined travel survey. In 2017, six states developed their state add-on's, including Arizona, California, Wisconsin, Georgia, Iowa, etc.

https://www.nrel.gov/transportation/secure-transportation-data/tsdc-cleansed-data.html

**Chicago Metropolitan Agency for Planning (CMAP)** – typically local planning agencies also conduct such surveys regularly.

https://datahub.cmap.illinois.gov/dataset/mydailytravel-2018-2019-public

**Google Distance Matrix API** – it helps to fill in the missing travel information, such as travel distance, time, etc.

https://datahub.cmap.illinois.gov/dataset/mydailytravel-2018-2019-public

# 7. Urban Imagery

Typically the urban imagery include satellite, street-view, and night-light imagery.

**Google Static Street View API** – helps to collect the street-view images
https://developers.google.com/maps/documentation/streetview/overview

**Google Maps Static API** – helps to collect the satellite images
https://developers.google.com/maps/documentation/maps-static/start

**Others: U.S. Geological Survey (USGS) Earth Explorer, Landsat-8, etc.**
https://earthexplorer.usgs.gov/
https://eos.com/landsat-8/

# 8. Point-of-Interests (POI)

POIs include the information needed to identify particular **venues and activities** with street addresses and GPS coordinates.

**OpenStreetMap (OSM) data source**

https://wiki.openstreetmap.org/wiki/Downloading_data

**Google Place API**

https://developers.google.com/maps/documentation/places/web-service/overview

# 9. Public health, energy, pollution, etc.

Other mobility-related topics.

**Electricity sales**
https://www.eia.gov/electricity/data/eia861m/

**Array of Things (AoT)**: air quality and vehicle-pedestrian counts.
https://arrayofthings.github.io/

**CDC 500 Cities Project**: public health outcomes at census tract level
https://www.cdc.gov/places/about/500-cities-2016-2019/index.html

# 10. Web scraping data

**Zillow**: rent and property information.
https://www.zillow.com/howto/api/GetSearchResults.htm

**Twitter**: comments and geo tag information.
https://developer.twitter.com/en/docs/twitter-api

**Yelp**: comments, rating, spatial and temporal information.
https://www.yelp.com/dataset

**Other web scraping tools**
https://www.webharvy.com/articles/zillow-extraction.html

Question

# How to choose from so many data sets?

Two challenges:

1. Hard to choose without knowing all the specifics.
2. You need to know your research question before choosing data sets.

In today's class, we focus on the most common and important data

**(1) Census + (2) Spatial shapefiles**

# Part 2. Census data and shapefiles

# Census, a.k.a., American Community Survey (ACS)

- American Community Survey (ACS) replaces the traditional census since 2005.
- ACS has the yearly sampling in the form of a huge and rolling survey that is continually refreshed.
- ACS includes nearly 3 million addresses per year. It is around 2.5% sample of the total US households. In a five-year cycle, it accounts for about 12.5% of the total households.
- ACS is a nationwide and continuous survey.

# Census data is of high quality

- Census is the only source for demographic data with a wide geographic scope.
- Census is the most reliable and detailed information for describing local areas, e.g., neighborhoods, cities, and counties.
- Census is the most consistent source of time series demographic data available.
- Basically, census is the starting point of our urban data analysis.

# Variables in ACS (Census)

**Social characteristics**

- Marital status
- Place of birth
- Education
- Ancestry
- Residence
- Language
- etc.

**Economic characteristics**

- Income
- Labor force status
- Place of work
- Journey to work
- etc.

**Housing unit characteristics**

- Units in structure
- Number of rooms
- Number of bedrooms
- Year structure built
- Year moved into unit
- Vehicles
- Etc.

**Financial characteristics of housing**

- Value of home
- Monthly rent
- Shelter costs

**Transport characteristics**

- Travel modes

# ACS data is released every year as 1-year, 3-year, and 5-year estimate.

ACS provides single-year and multiyear estimates:

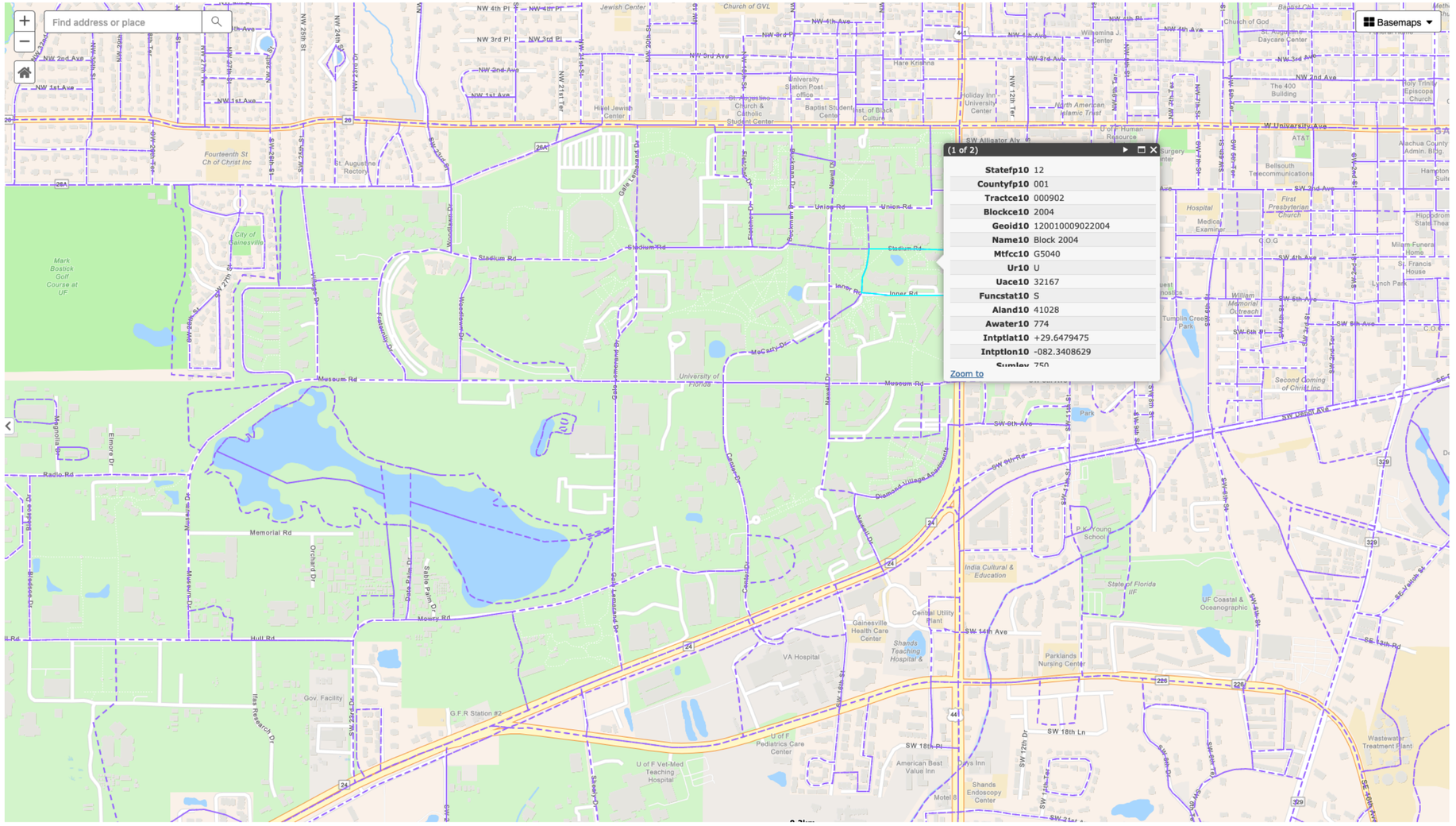
- Single-year = information is collected over 12-month period
- 3-year = information is collected over 36-month period
- 5-year = information is collected of a 60-month period

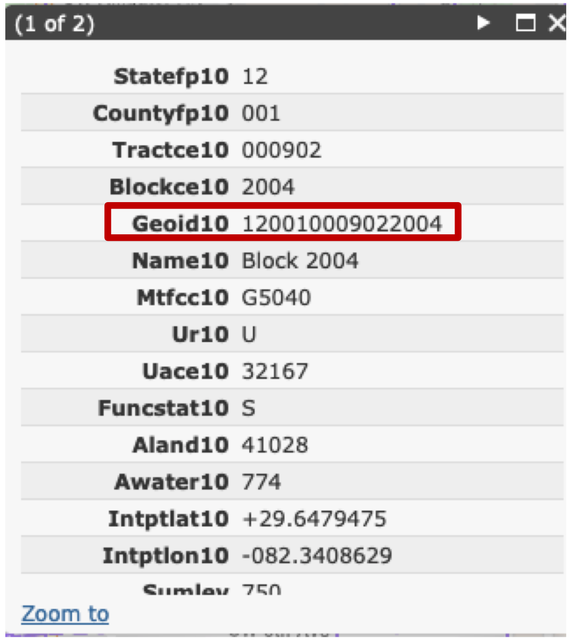
**Distinguishing features of ACS 1-year, 3-year, and 5-year estimates**

| 1-year estimates | 3-year estimates | 5-year estimates |
|---|---|---|
| 12 months of collected data | 36 months of collected data | 60 months of collected data |
| Data for areas with populations of 65,000+ | Data for areas with populations of 20,000+ | Data for all areas |
| Smallest sample size | Larger sample size than 1-year | Largest sample size |
| Less reliable than 3-year or 5-year | More reliable than 1-year; less reliable than 5-year | Most reliable |
| Most current data | Less current than 1-year estimates; more current than 5-year | Least current |
| **Best used when** | **Best used when** | **Best used when** |
| Currency is more important than precision<br><br>Analyzing large populations | More precise than 1-year, more current than 5-year<br><br>Analyzing smaller populations<br><br>Examining smaller geographies because 1-year estimates are not available | Precision is more important than currency<br><br>Analyzing very small populations<br><br>Examining tracts and other smaller geographies because 1-year estimates are not available |

# Federal Information Processing Series (FIPS) codes represent the hierarchical spatial structure

## What is the FIPS code for our department on the UF campus?

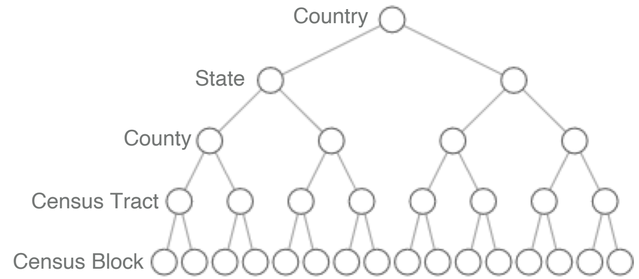


The hierarchical spatial structure is captured in this GEOID

Our GEOID: 12 001 000902 2004

State County Census Tract Census Block




**Notes**

- Implicitly this GEOID indicates a tree structure.
- The tree structure has multiple levels, and each level uses consistent spatial units.
- The GEOID is the identity of the spatial units, and we will use it for data integration and visualization.