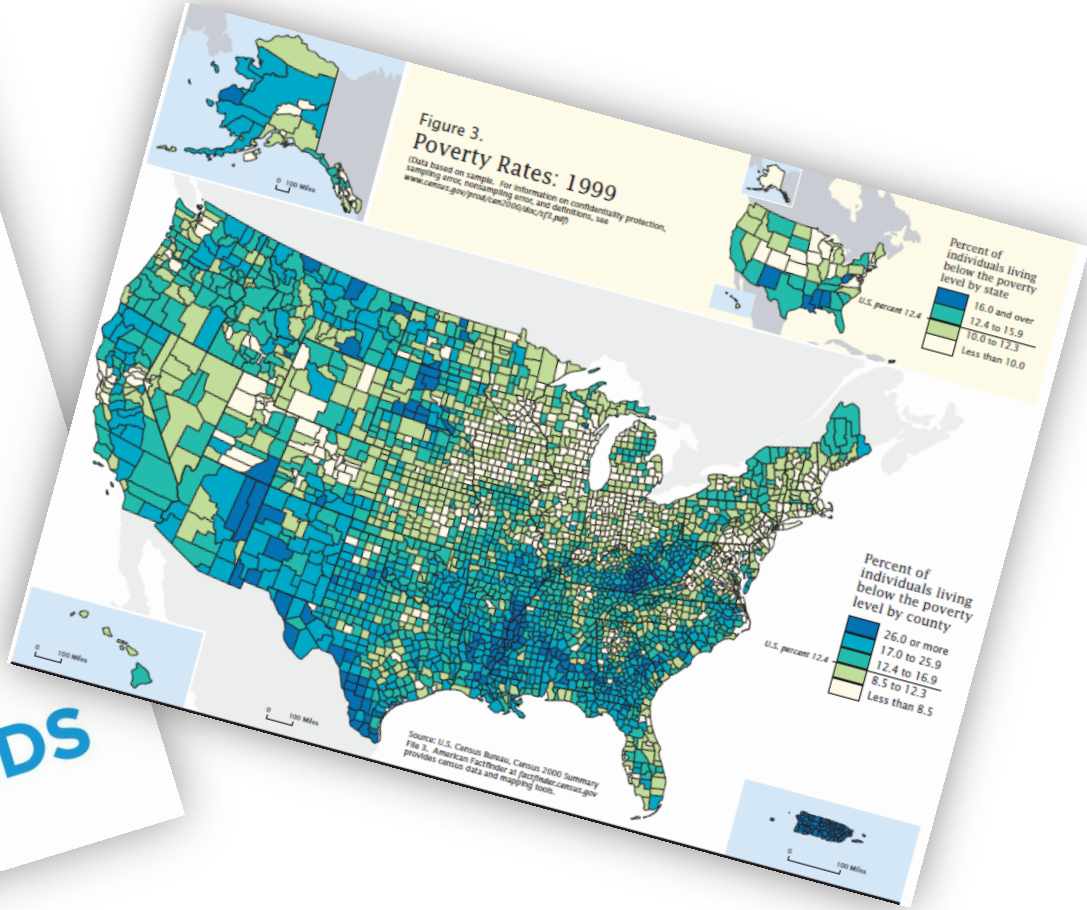
11.520: A Workshop on Geographic Information Systems
11.188: Urban Planning and Social Science Laboratory



Making Sense of the Census

September 29, 2010

Census Discussion Overview

Utilizing Large Tabular Datasets (US Census)

- Understand key aspects of US Census Data
  - What is it and why do we care?
  - How are the data collected?
  - What data are available?
  - Introduction to Census geography and summary levels
  - A quick look at the Census documentation
  - A quick look at some sample data
- Understand nature and use of large, highly structured, public datasets
  - Examine primary US Census Data (SF3=Summary File 3) at the block group level
  - Appreciate differences between Census SF3 CDs and third-party census extracts
  - Learn how to manipulate census data in MS-Access and ArcGIS

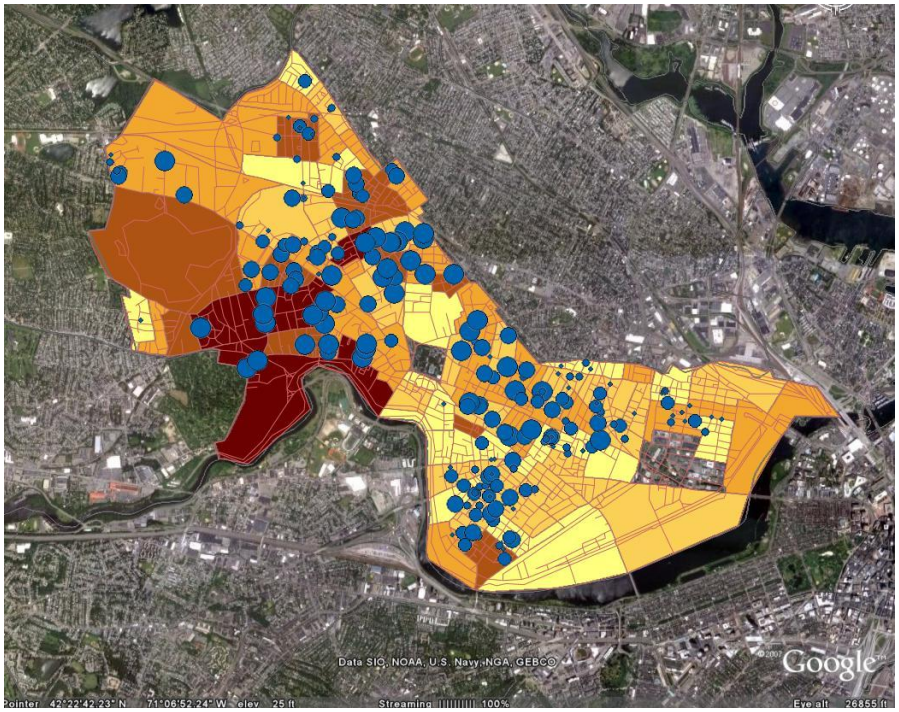
Example - Thematic map of Income (viz., median earnings)

- **Isn't this easy?**

  We did thematic map of income on day 1!

  - Yes, if desired variable is already in attribute table of map
  - There are thousands of variable in the 'long form' census
    - Which combinations zero in on **useful indicator**
    - Drilling down and combining data are often needed
  - What do we mean by 'income'
    - Household, personal??
    - Earned income, all income??
    - At what scale? state, county, city, tract, block group??

- For what time period? weekly, yearly, part-time??



Example - Thematic map of Income (viz., median earnings)

- Example: median 1999 personal earnings from the 2000 US Census
  - **Variable P85 (among the hundreds of census variables and thousands of columns)**
  - P85 records "Median earnings in 1999 dollars by sex for the population 16 years and over with earnings"
    - 'earnings' includes wages, salaries, and net self-employment income (but not entitlements)
  - The P85 table has three columns:
    - P085001 = total (for **universe** of population 16+ years old with earnings)
    - P085002 = male
    - P085003 = female

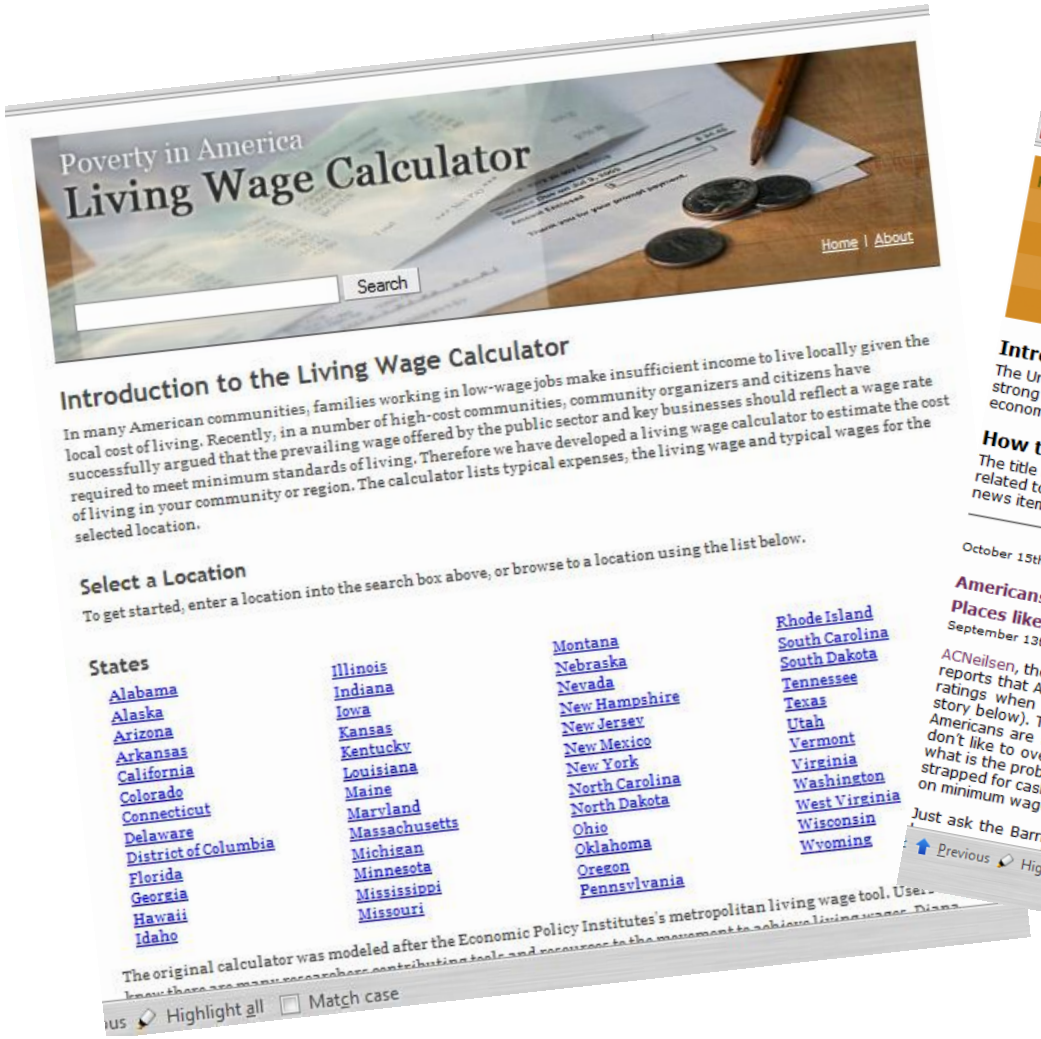
Example - Thematic map of Income (viz., median earnings)

- Use MS-Access database in class locker: M:\data\census2k\hw2_sf3_lite.mdb
  - Two of the 70+ raw US census files (for Massachusetts) have already been loaded
  - Determine the median earnings for Cambridge block groups
    - Understand census data structure and use of raw data
    - Examine ER diagram of relationships among the tables used in this query
    - Illustrate SQL query development in MS-Access
- Save query and 'make table' to have results available in both forms

Example - Thematic map of Income (viz., median earnings)

- Map median earnings for Cambridge block groups
  - Bring MS-Access query results into ArcMap
  - Create thematic map
- Examine Technical Documentation for the SF3 Census data
  - Online site at US Census: http://www.census.gov/prod/cen2000/doc/sf3.pdf
  - Copy in class locker (for faster access): http://mit.edu/11.520/data/census2k/sf3.pdf
- Learn how to identify variables of interest and find them in the data tables

Examples: Research Projects based on Census



Sources: http://livingwage.mit.edu/; http://povertyinamerica.mit.edu/

What Is the Census

- **Mandated by the Constitution of the United States**
- **Census History**
  - The **first census** took place in **1790** to determine the number of seats each state would have in the U.S. House of Representatives.
  - The census was created to gain a better understanding of **where people lived** and to **establish patterns of settlement** as the nation grew.
  - The **modern census** of **population and housing** was established in **1940** with the incorporation of the housing component and the introduction of sampling techniques for the **long form**
- **Census Bureau**
  - The Census Bureau was established in 1902.
  - Today, in addition to administering the census of population and housing, the Census Bureau conducts more than 200 annual surveys, including the **American Community Survey**, the **Current Population Survey** and **economic censuses** every five years.

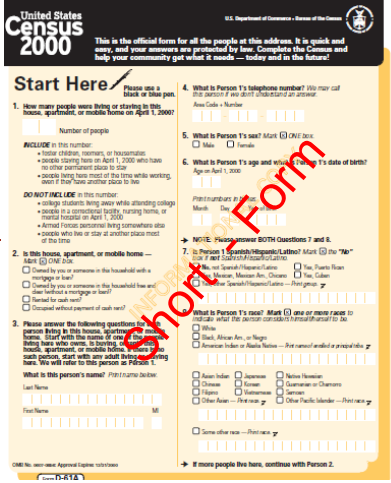
Source: http://2010.census.gov/2010census/

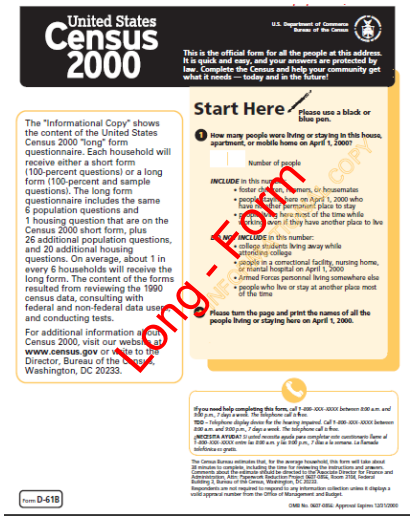
Why Do We Care?

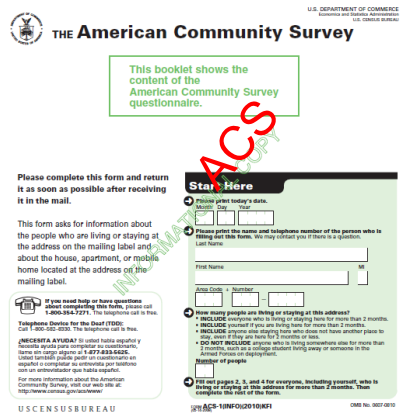
- **U.S. Congressional representatives are apportioned based on census counts.**
- **Federal dollars are distributed based on census**
  - for schools, employment services, highway assistance, housing construction, hospital services, programs for the elderly, etc.
- **Conducted every 10 years** (now → a rolling census)
- **An actual count of entire population** categorized by various criteria
- **The only source for spatially detailed demographic data**
  - with a consistent coast-to-coast data structure
- **The most reliable, detailed, and consistent source**
  - for describing local areas: neighborhoods, cities, counties
  - of time series demographic data available

How the Data Are Collected

- **Collected from households through a mail survey** conducted every decade
- **For the 2000 Census** (2000 Census Home Page)
  - more than 285,000 census takers and support personnel
  - accounted for 118 million households and 275 million persons in the U.S.
- **Two different census questionnaires** are distributed:
  - short-form questionnaire contains questions asked of everyone (summarized in Summary Tape File 1 (STF 1) for 1980 and 1990, **Summary File (SF 1) for 2000**)
  - long-form questionnaire contains questions asked of a population sample (**1/6 households**) (summarized in Summary Tape File 3 (STF 3) for 1980 and 1990, **Summary File 3 (SF 3) for 2000**)
- **The long form is being replaced in the 2010 Census** by the American Community Survey (ACS).
  - This program will survey homes every month and provide updated statistics every year instead of every 10 years. The program begins in 2003.







What's Included:

- **Information on Population, Employment & Housing Characteristics**
- **Long Form: Sample Counts (STF 3/SF 3)**

Population Characteristics	Housing Characteristics
Social Characteristics	*Age of Housing*
Education	*Heating Fuel*
Citizenship	*Facilities*
Ancestry	*Vehicles*
Language	*Mortgage Status*
Disability	
Children	
Place of Birth	
Economic Characteristics	
Income	
Labor Force Status	
Employment	
Place of Work	
Public Assistance	
Retirement Income	

What's Included:

- **Information on Population, Employment & Housing Characteristics**
- **Short Form: 100% Count (STF 1/SF 1)**

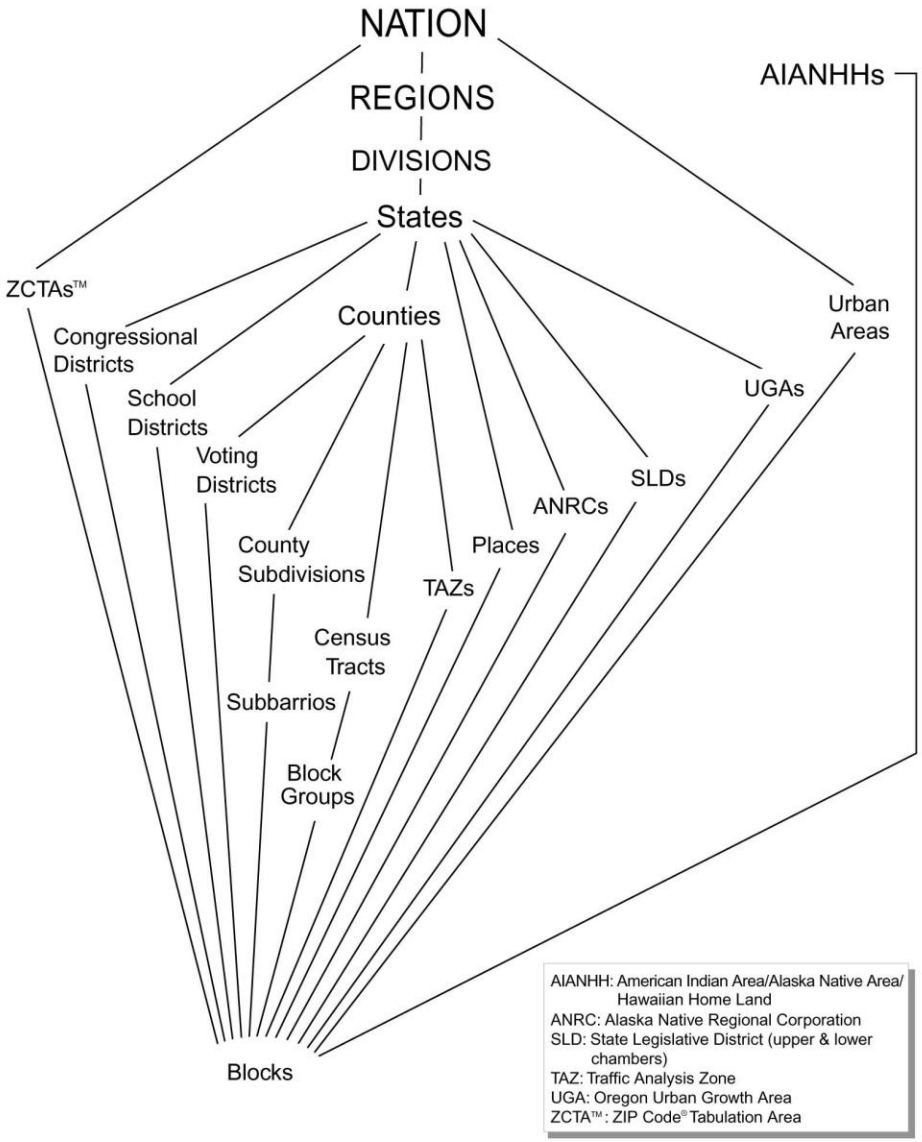
Population Characteristics	Housing Characteristics
Age	*Tenure*
Gender	*Value or Contract Rent*
Race	*Vacancy Status*
Hispanic Origin	*Number of Rooms*
Marital Status	*Units in Structure*
Household Type	*Congregate Housing*
Household Relationship	

- **Why We Need to Know the Two Components**
  - Accuracy of the data varies and counts differ (Why?)
  - It helps us to understand how the data are organized in Summary Files (SFs)

Census Geography and Summary Levels

- The Census organizes and aggregates data into a series of geographic hierarchies

- Standard Hierarchy of Census Geographic Entities (from *Census 2000 Summary File 1 Technical Documentation*, prepared by the U.S. Census Bureau, 2001, p. A-25)



Census Geography and Summary Levels

- **If no line joins 2 geographic types, then an absolute and predictable relationship does not exist between them.**
  - For example, many places are confined to one county. However, some places extend over more than one county, such as New York City. Therefore, an absolute hierarchical relationship does not exist between *counties* and *places*, and any tabulation involving both these geographic types may represent only a part of one county or one place.

- **With connecting lines, the diagram also shows the hierarchical relationships between geographic types.**

 Source: http://factfinder.census.gov/home/en/epss/census_geography.html

Census Geography and Summary Levels

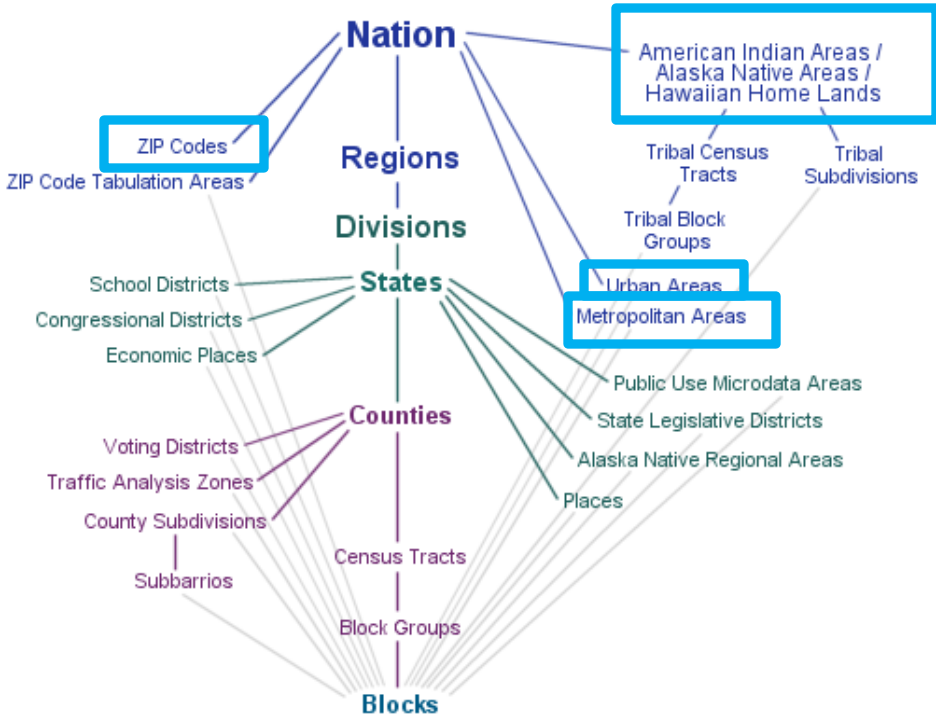
- **State-County-PLACE-Tract-Block Group Nesting**

Summary Level	Geographic Unit
010	United States
020	Region: Northeast (NE), Midwest (MW), South (S) and West (W) Regions
030	Division:
	Northeast Region: New England, Mid Atlantic
	Midwest Region: East North Central, West North Central
	South Region: South Atlantic, East South Central, West South Central
	West Region: Mountain, Pacific
040	State (includes Washington, D.C. & Puerto Rico)
050	County
060	**County Subdivision**
070	**Place**
080	Census Tract / Block Numbering Area (average 4,000 persons)
090	Block Group (average 1,000 persons)
100	Block (average 85 persons)

Census Geography and Summary Levels

- **State-County-Tract-Block Group Nesting**

Summary Level	Geographic Unit
040	State (includes Washington, D.C. & Puerto Rico)
050	County
140	Census Tract
150	Block Group

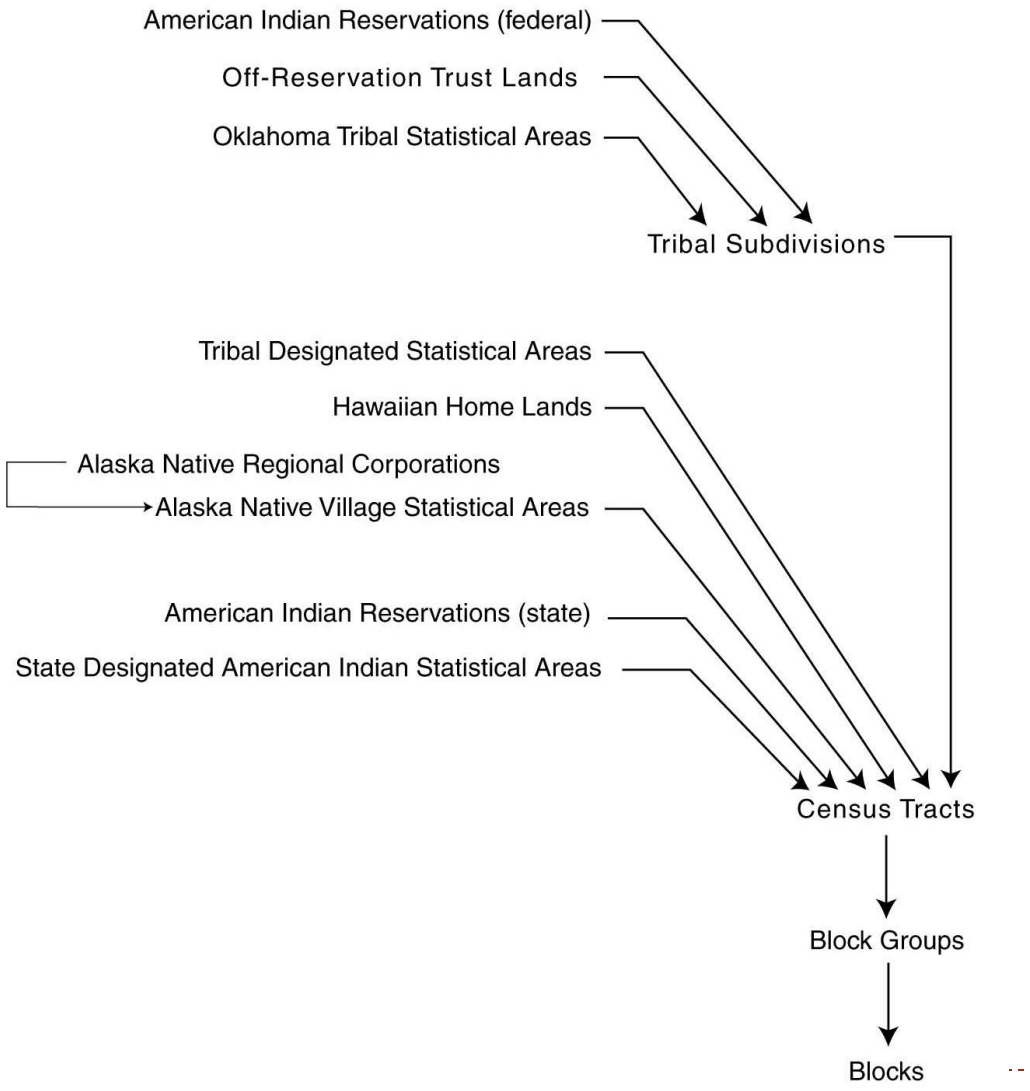


- **Supplemental Geographic Areas**

Summary Level	Geographic Unit
400	Urbanized Areas
300	Metropolitan Areas (MSAs, CMSAs)
200	American Indian and Alaska Native areas
800	ZIP codes

Census Geography and Summary Levels

- Hierarchy of American Indian, Alaska Native, and Native Hawaiian Entities (from *Census 2000 Summary File 1 Technical Documentation*, prepared by the U.S. Census Bureau, 2001, p. A-26)



A Visual Look at Census Geography

- Continental United States (Regions in blue; Divisions in green; States in brown)

Figure A–3. **Census Regions, Census Divisions, and Their Constituent States**

(Source: the U.S. Census Bureau, 2001, p. A-27)

Northeast Region

New England Division:
Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut

Middle Atlantic Division:
New York, New Jersey, Pennsylvania

Midwest Region

East North Central Division:
Ohio, Indiana, Illinois, Michigan, Wisconsin

West North Central Division:
Minnesota, Iowa, Missouri, North Dakota, South Dakota, Nebraska, Kansas

South Region

South Atlantic Division:
Delaware, Maryland, District of Columbia, Virginia, West Virginia, North Carolina, South Carolina, Georgia, Florida

East South Central Division:
Kentucky, Tennessee, Alabama, Mississippi

West South Central Division:
Arkansas, Louisiana, Oklahoma, Texas

West Region

Mountain Division:
Montana, Idaho, Wyoming, Colorado, New Mexico, Arizona, Utah, Nevada

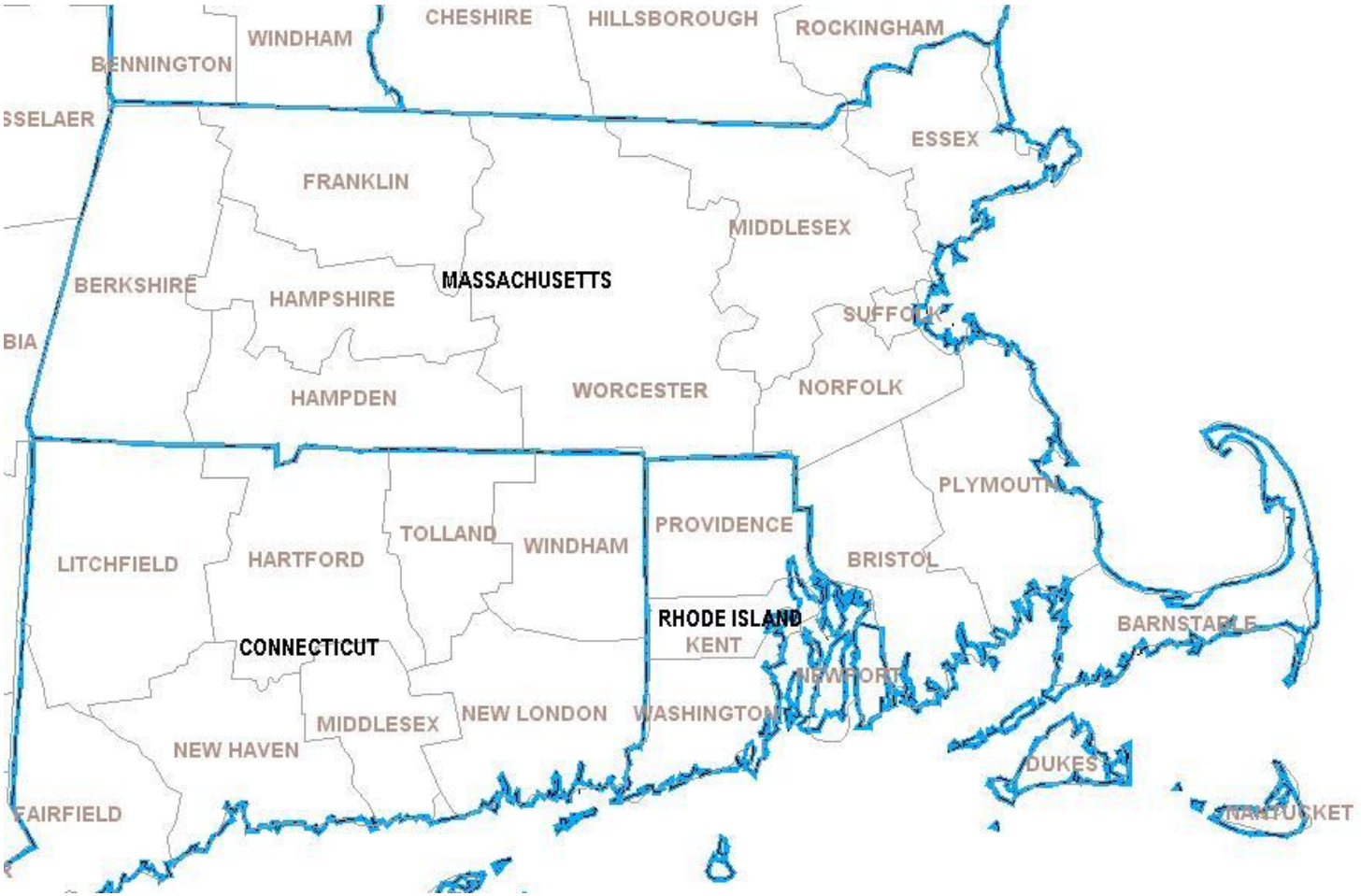
Pacific Division:
Washington, Oregon, California, Alaska, Hawaii

A Visual Look at Census Geography

- Counties

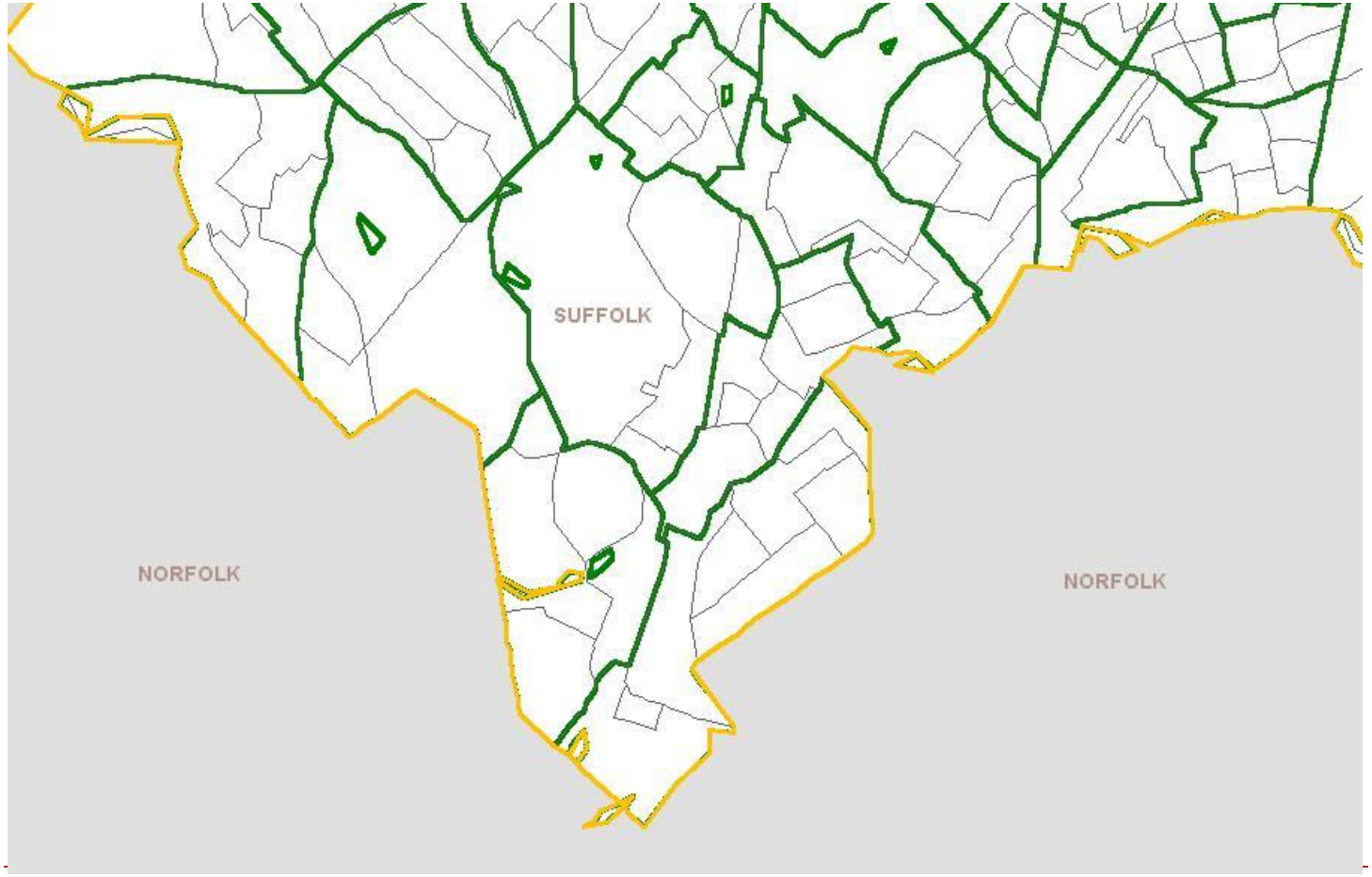
A Visual Look at Census Geography

- A Closer Look at Southern New England Counties



A Visual Look at Census Geography

- Tracts (green lines) and Block Groups (gray lines) in Suffolk County, MA



Census Geography Concepts

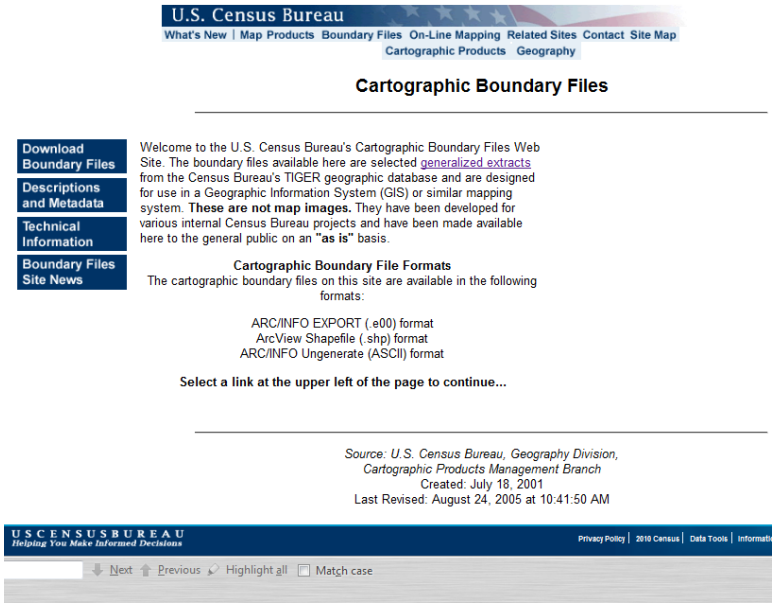
- The **Census Block** is the basic level
- **Confidentiality** must be maintained,
  - data about individual persons and households are not revealed
- **More detailed data are provided for higher levels of geography** (Why?)
- **Many, but not all, items are available at multiple summary levels**

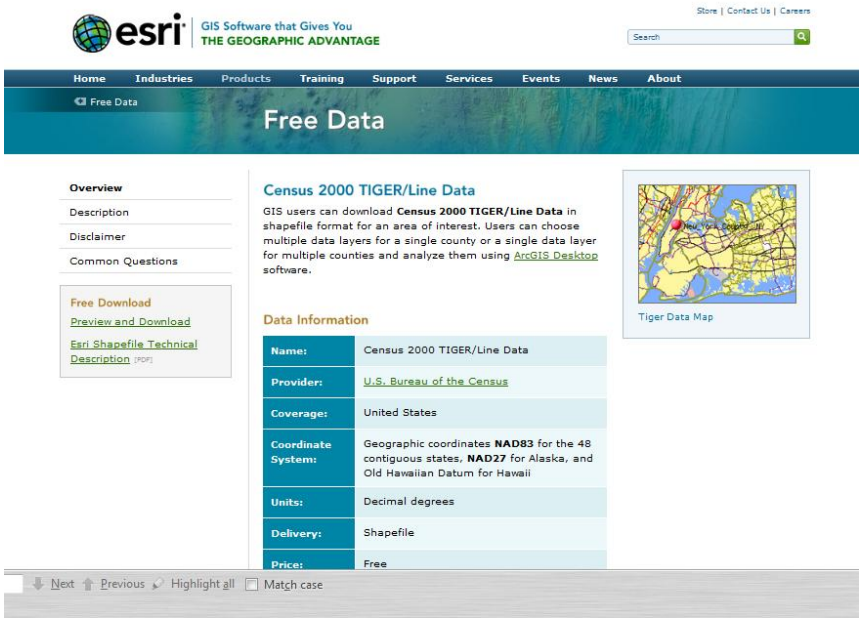
Potential Problems

- **The same geographic name**
  - is used for summary levels corresponding to **different aggregations**
- **Geographic areas at lower levels**
  - may be **subdivided by higher levels of geographic units**
  - E.g., a census tract may be split by town boundaries
- **The same variable names**
  - are used for different variables in the STF/SF 1 and STF/SF 3
  - E.g., in SF1 P0020001= FAMILIES but in SF3: P0020001 = 100-PERCENT COUNT OF PERSONS
  - The way variable values are encoded makes identifying the meaning of variables difficult
- **ZIP codes do not overlay other units cleanly**
- **Geographic boundaries**
  - change with time, making time-series analysis difficult.

Obtaining Census Geographic Boundary Files for Use in a GIS

- ArcView shapefiles and ArcInfo coverage formats are readily available for 1990 and 2000 Census geography boundaries
  - Boundary files from the U.S. Census Bureau
  - Census TIGER 2000 Linefiles from ESRI's Geography Network





Census Summary Files

- **1980** STF 1 and STF 3 are now online
  - You can obtain the 1980 STF 1 via HTTP or FTP and the 1980 STF 3 via HTTP or FTP.
  - Documentation is available from the Odum Institute for Research in Social Science.
- **1990** Census files is distributed by the Census Bureau
  - as DBF files on CD-ROMs.
  - The Census Bureau has posted the contents of many 1990 CD-ROMs online. These are available via HTTP and FTP.
- **2000** Census files is distributed by the Census Bureau
  - on CD-ROMs, DVD-ROMs in a proprietary format, and
  - online in flat ASCII format via HTTP and FTP.

STF/SF 1: 100% count data from the short form

- For the 2000 Census, the SF 1 files encompass **all summary levels.**
- For the 1990 Census, the STF 1 files came in four varieties:
  - **A: States and subdivisions to the block group level**
  - **B: Block level**
  - **C: Entire U.S. and major subdivisions**
  - **D: Congressional Districts**

STF/SF 3: Sample data from the long form

- For the 2000 Census, the SF 3 files encompass **all summary levels.**
- For the 1990 Census, the STF 3 files came in four varieties:
  - **A: States and subdivisions to the block group level**
  - **B: 5-digit ZIP codes**
  - **C: Entire U.S. and major subdivisions**
  - **D: Congressional Districts**
- The 1980 STF 1 and STF 3 files had varieties similar to those of the 1990 Census.

A Quick Look at the Census Data & Documentation

- **1980 Census**
  - Overview from SUNY Albany's Center for Social and Demographic Analysis
  - Data sets available from IPCSR
- **1990**
  - STF 3A Variable Locator
  - State/County FIPS Codes
    - FIPS: Federal Information Processing Standard
    - (e.g., 4472| 4480| 06037| 2 | Los Angeles, CA)
  - Census Data at the Center for Disease Control and Prevention
  - Note that 1990 Census CDs are also available for borrowing from the MIT Rotch Library.

(MSA = metropolitan statistical area;
CMSA = consolidated MSA;
NECMA = New England county MA;
PMSA = primary MSA)

A Quick Look at the Census Data & Documentation

- **2000 Census**
  - American FactFinder
  - Public Law 94-171 (PL 94-171)
    - Documentation
    - Help on Using Browser Software on the CD-ROM
    - Data
  - Summary File 1 (SF 1)
    - Home Page
    - Documentation
    - Help on Processing Data Files in ASCII Format
    - Data
  - Summary File 2 (SF 2)
    - Documentation
    - Help on Processing Data Files in ASCII Format
    - Data
  - Summary File 3 (SF 3)
    - Documentation
  - Summary File 4 (SF 4)
    - Documentation

More Information About the 2000 Census

- **Commercial firms often repackage US census data**
  - **ESRI sample data** (and online geography network) contain common census variables
  - **Rotch Library has Geolytics CDs** with convenient census datasets including 1970-2000 data that has been adjusted to reflect 2000 census tract boundary files
- Data Release Dates
- Subjects Areas of Questions Asked (lists first US census in which subject areas were first included - helpful when contemplating longitudinal studies)

- **Censuses in Other Countries**
  - International Statistics Agencies

Example:
Let's find unemployment rates for Cambridge area block groups

- **How should we measure unemployment rate**:
  - Census definition is: " the fraction of **adults aged 16 or over** who are in **the labor force** and are **unemployed** (during the **sample week in April 1999**)"
- **Find the relevant SF3 census 2000 variables**:
  - we use the SF3 technical documentation (Ch. 3)
  - to find variable P43:
    - employment status by sex, and the name of the text file that includes the raw data for this variable (ma00004.uf3)

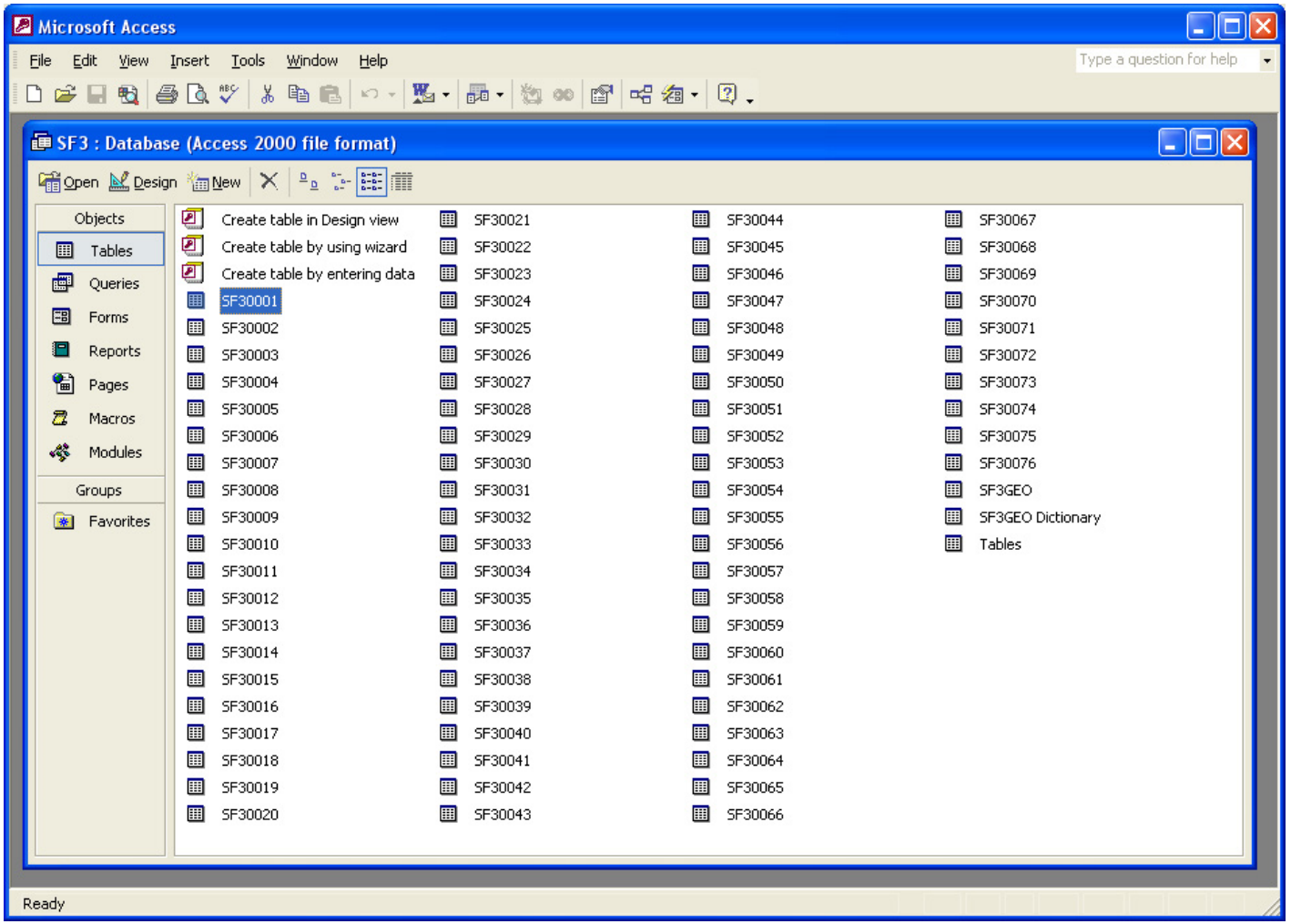
Example:
Let's find unemployment rates for Cambridge area block groups

- **### Find and download the zipped datafile** that contains P43 for Massachusetts as an ASCII 'flat file' - this file is called: ma00004.uf3
- **### Find and download the zipped datafile** that contains the geographic identifiers for Massachusetts - this file is called: mageo.uf3
- **### Find and download the MS-Access templates**
  - The zipped template for MS-Access 2000 is here: http://www.census.gov/support/2000/SF3/Acc2000.zip
- **### Import the relevant Mass data into Access tables** rename the unzipped text files to end in 'txt'
- In the dialogue box that lets you tell MS-Access how to parse the text file, click 'Advanced' and choose the 'specs' that apply to the particular data file (for example, ma000043)

Example:

Let's find unemployment rates for Cambridge area block groups

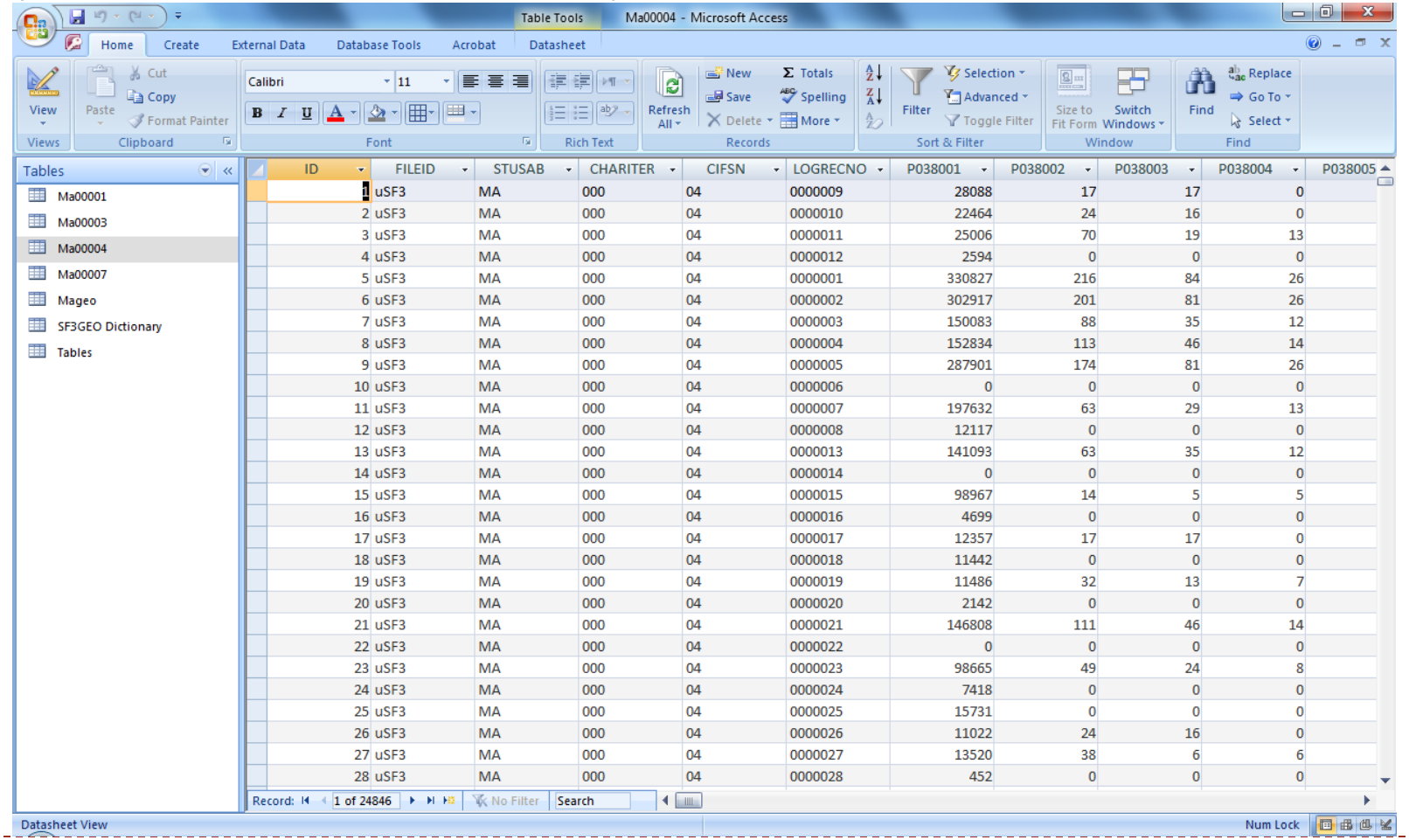
- **Import the relevant Mass data into Access tables**



Example:

Let's find unemployment rates for Cambridge area block groups

- **Import the relevant Mass data into Access tables** (M:\data\census2k\lab5_ma.mdb)



ID	FILEID	STUSAB	CHARITER	CIFSN	LOGRECNO	P038001	P038002	P038003	P038004	P038005
1	uSF3	MA	000	04	0000009	28088	17	17	0	
2	uSF3	MA	000	04	0000010	22464	24	16	0	
3	uSF3	MA	000	04	0000011	25006	70	19	13	
4	uSF3	MA	000	04	0000012	2594	0	0	0	
5	uSF3	MA	000	04	0000001	330827	216	84	26	
6	uSF3	MA	000	04	0000002	302917	201	81	26	
7	uSF3	MA	000	04	0000003	150083	88	35	12	
8	uSF3	MA	000	04	0000004	152834	113	46	14	
9	uSF3	MA	000	04	0000005	287901	174	81	26	
10	uSF3	MA	000	04	0000006	0	0	0	0	
11	uSF3	MA	000	04	0000007	197632	63	29	13	
12	uSF3	MA	000	04	0000008	12117	0	0	0	
13	uSF3	MA	000	04	0000013	141093	63	35	12	
14	uSF3	MA	000	04	0000014	0	0	0	0	
15	uSF3	MA	000	04	0000015	98967	14	5	5	
16	uSF3	MA	000	04	0000016	4699	0	0	0	
17	uSF3	MA	000	04	0000017	12357	17	17	0	
18	uSF3	MA	000	04	0000018	11442	0	0	0	
19	uSF3	MA	000	04	0000019	11486	32	13	7	
20	uSF3	MA	000	04	0000020	2142	0	0	0	
21	uSF3	MA	000	04	0000021	146808	111	46	14	
22	uSF3	MA	000	04	0000022	0	0	0	0	
23	uSF3	MA	000	04	0000023	98665	49	24	8	
24	uSF3	MA	000	04	0000024	7418	0	0	0	
25	uSF3	MA	000	04	0000025	15731	0	0	0	
26	uSF3	MA	000	04	0000026	11022	24	16	0	
27	uSF3	MA	000	04	0000027	13520	38	6	6	
28	uSF3	MA	000	04	0000028	452	0	0	0	

Example:

Let's find unemployment rates for Cambridge area block groups

- **Develop MS-Access query to join the geography and P43 tables**
- **Choose appropriate summary level (150)** in order to get right counts for block groups
- **Refine and use query to pull relevant rows and columns** for block groups in all of Mass (or just for Middlesex County if we only want Cambridge and its neighbors north of the Charles River (all of which are in Middlesex County).
- **Join tabular data to map** of block groups for Middlesex County (obtained use MIT geo-data tool from Library SDE server)

Develop MS-Access query

to join the geography and P43 tables

Table number	Table contents	Data dictionary reference name	Seg-ment	Max size
P43.	**SEX BY EMPLOYMENT STATUS FOR THE POPULATION 16 YEARS AND OVER [15]**			
	Universe: Population 16 years and over			
	Total:	P043001	04	9
	Male:	P043002	04	9
	In labor force:	P043003	04	9
	In Armed Forces	P043004	04	9
	Civilian:	P043005	04	9
	Employed	P043006	04	9
	Unemployed	P043007	04	9
	Not in labor force	P043008	04	9
	Female:	P043009	04	9
	In labor force:	P043010	04	9
	In Armed Forces	P043011	04	9
	Civilian:	P043012	04	9
	Employed	P043013	04	9
	Unemployed	P043014	04	9
	Not in labor force	P043015	04	9

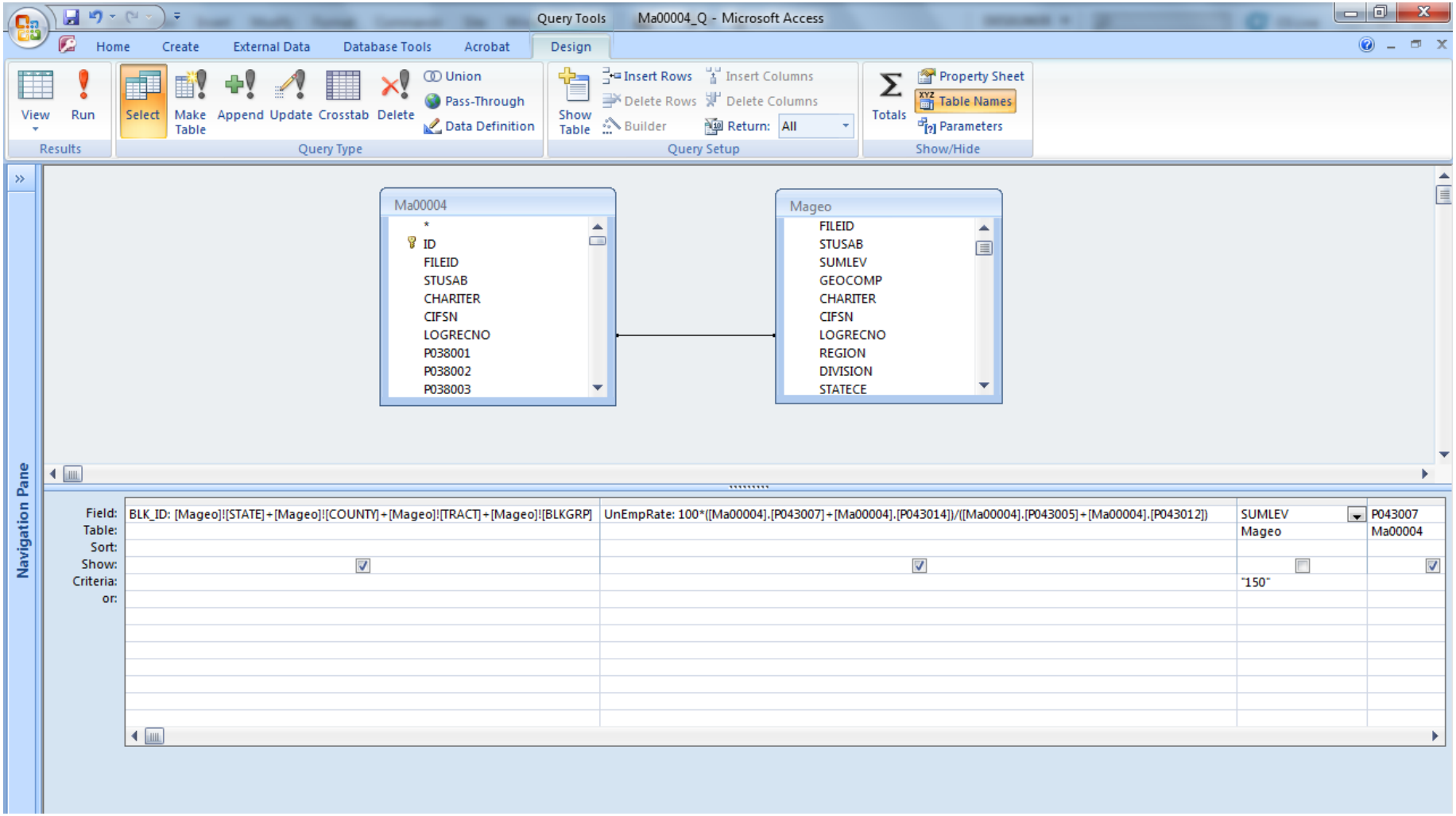
Develop MS-Access query
to join the geography and P43 tables

- P43. SEX BY EMPLOYMENT STATUS FOR THE POPULATION 16 YEARS AND OVER [15]
- Universe: Population 16 years and over
- P043001: Total:
- P043002: Male:
- P043003: In labor force:
- P043004: In Armed Forces
- P043005: Civilian:
- P043006: Employed
- P043007: Unemployed
- P043008: Not in labor force
- P043009: Female:
- P043010: In labor force:
- P043011: In Armed Forces
- P043012: Civilian:
- P043013: Employed
- P043014: Unemployed
- P043015: Not in labor force

- **Join the tables using the 'logrecno' column**
- Build a **state+county+tract+blockgroup** 12-digit **block group identifier** so you can join to the block group map
- Compute the percent unemployed = 100*(P043007+P0430014)/(P043005+P043012)

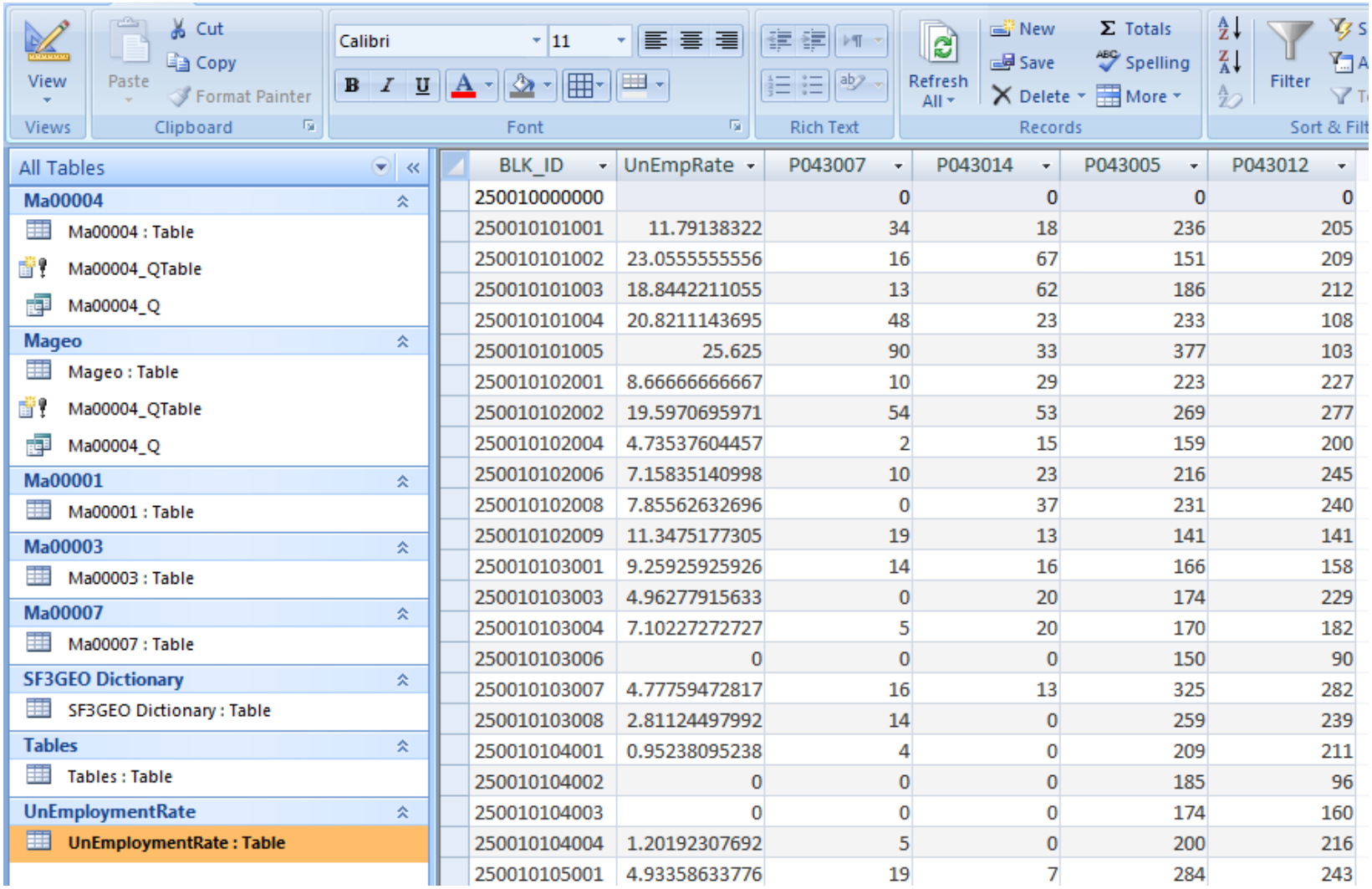
Develop MS-Access query

to join the geography and P43 tables



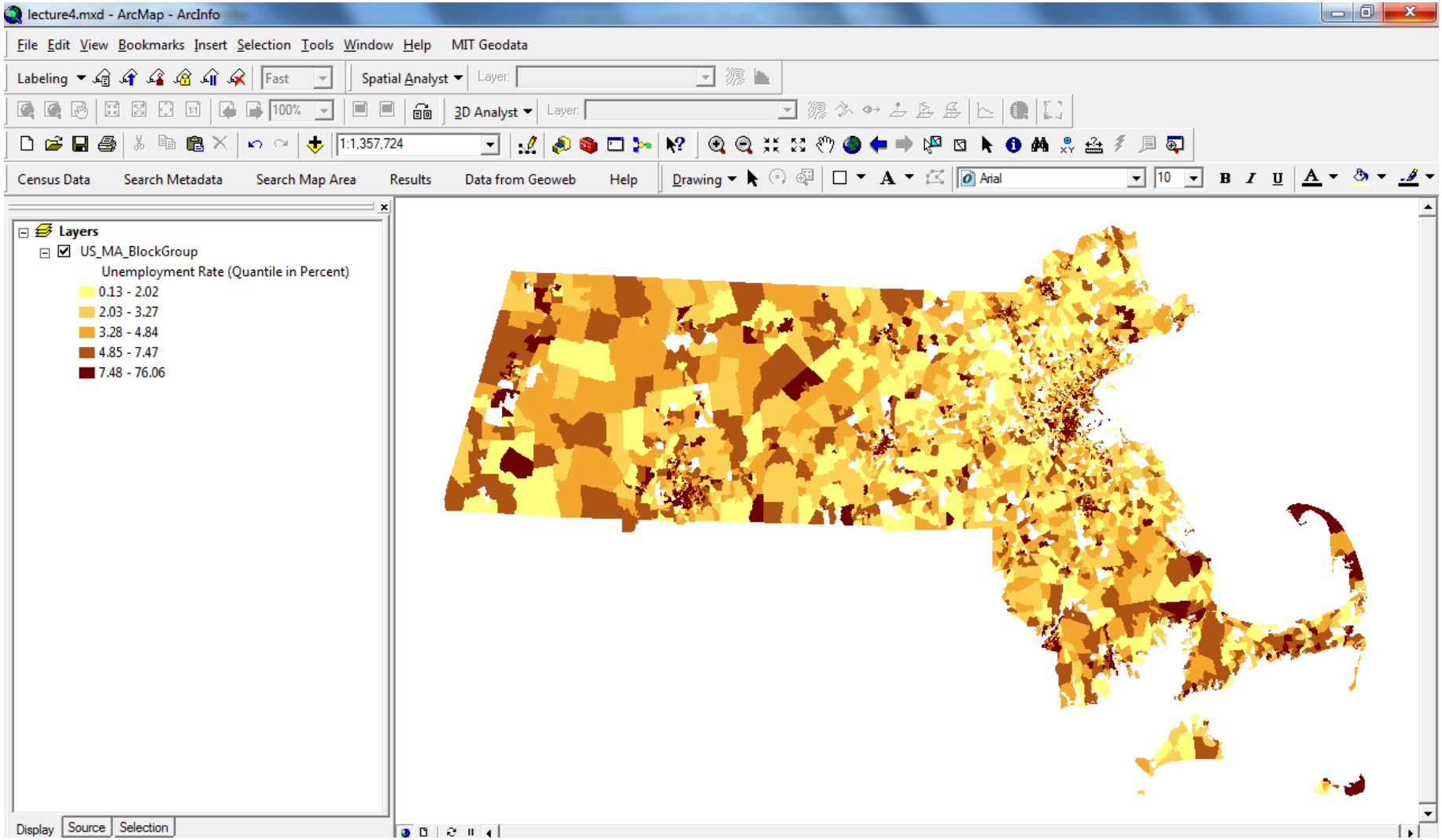
Develop MS-Access query

to join the geography and P43 tables



BLK_ID	UnEmpRate	P043007	P043014	P043005	P043012
250010000000		0	0	0	0
250010101001	11.79138322	34	18	236	205
250010101002	23.0555555556	16	67	151	209
250010101003	18.8442211055	13	62	186	212
250010101004	20.8211143695	48	23	233	108
250010101005	25.625	90	33	377	103
250010102001	8.66666666667	10	29	223	227
250010102002	19.5970695971	54	53	269	277
250010102004	4.73537604457	2	15	159	200
250010102006	7.15835140998	10	23	216	245
250010102008	7.85562632696	0	37	231	240
250010102009	11.3475177305	19	13	141	141
250010103001	9.25925925926	14	16	166	158
250010103003	4.96277915633	0	20	174	229
250010103004	7.10227272727	5	20	170	182
250010103006	0	0	0	150	90
250010103007	4.77759472817	16	13	325	282
250010103008	2.81124497992	14	0	259	239
250010104001	0.95238095238	4	0	209	211
250010104002	0	0	0	185	96
250010104003	0	0	0	174	160
250010104004	1.20192307692	5	0	200	216
250010105001	4.93358633776	19	7	284	243

Develop MS-Access query
to join the geography and P43 tables



Final Notes:

- US Census provides many online tools to obtain census data.
- There are many third-party tools and CDs that repackage the data in smaller chunks, with or without maps, and sometimes in pre-processed forms
  - e.g., after normalizing to percent owner-occupied rather than just as the raw counts.
  - These assorted tools fill many niche markets.
- Relatively few census data users understand the data structure and raw files at the level described in these lecture notes
  - i.e., at the level needed to find and use any of the thousands of columns of data that are available at each level of geography.

Questions? Ask Us

11.520staff@mit.edu

.

This PowerPoint was prepared by Shan Jiang, September 29, 2010
Based on the Lecture Notes on the 11.520 Class Website
Augmented and modified 1999-2010 by Thomas H. Grayson, Anne Kinsella Thompson, Sarah Williams, Xiongjiu Liao, Joe Ferreira, and Shan Jiang.