Other software packages

- Microsoft Excel
  - spreadsheet
  - very convenient for entering data in flat-file format
  - clients very frequently bring data to statisticians in Excel format
  - NOT reliable and accurate for doing statistical analysis

- Microsoft Access
  - relational database management system

Reading data files into SAS from other software packages

- Import Wizard
  - point-and-click interactive reading
  - convenient if file only needs to be read once
  - can write `proc import` code to be copied into programs

- `proc import`
  - can be used instead of data step in SAS programs
  - much more convenient if file needs to be read in multiple programs, or program using file needs to be run repeatedly

Importing from Other Sources

Types of files that the Import Wizard and/or `proc import` can read

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Input Data Source</th>
<th>Extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCESS</td>
<td>Microsoft Access database</td>
<td>.MDB</td>
</tr>
<tr>
<td>DBF</td>
<td>dBASE file</td>
<td>.DBF</td>
</tr>
<tr>
<td>WK1</td>
<td>Lotus 1 spreadsheet</td>
<td>.WK1</td>
</tr>
<tr>
<td>WK3</td>
<td>Lotus 3 spreadsheet</td>
<td>.WK3</td>
</tr>
<tr>
<td>WK4</td>
<td>Lotus 4 spreadsheet</td>
<td>.WK4</td>
</tr>
<tr>
<td>EXCEL</td>
<td>Excel V 4 or 5 spreadsheet</td>
<td>.XLS</td>
</tr>
<tr>
<td>EXCEL4</td>
<td>Excel V 4 spreadsheet</td>
<td>.XLS</td>
</tr>
<tr>
<td>EXCEL5</td>
<td>Excel V 5 spreadsheet</td>
<td>.XLS</td>
</tr>
<tr>
<td>EXCEL97</td>
<td>Excel 97 spreadsheet</td>
<td>.XLS</td>
</tr>
<tr>
<td>DLM</td>
<td>delimited file (default is blank)</td>
<td>*</td>
</tr>
<tr>
<td>CSV</td>
<td>delimited file (comma-sep vals)</td>
<td>.CSV</td>
</tr>
<tr>
<td>TAB</td>
<td>delimited file (tab-delimited )</td>
<td>.TXT</td>
</tr>
</tbody>
</table>

Restriction: The data sources available to you depend on the SAS/ACCESS products that you have licensed. If you do not have any SAS/ACCESS products licensed, then the only types of data source files available to you are .CSV, .TXT, and delimited files.
Example

- from R or Splus
  - use write.table to write data out as a delimited file

Data frame that comes with R

```r
> USArestr
       Murder Assault UrbanPop Rape
Alabama 13.2 236.0    58.5 21.2
Alaska  10.0 263.0    48.4 44.5
Arizona  8.1 294.0    60.3 31.0
Arkansas  8.8 190.0    50.1 19.5
California 9.0 276.0    91.4 40.6
```

R command to write out file as tab-delimited data file

```r
> write.table(USArestr, file="C:\Documents\166\USArestr.txt", sep="\t", quote = FALSE, col.names = TRUE)
```

Now in SAS....

File / Import Data

Import Wizard
Select a data source from the list below
Choose "Delimited File (*.txt)"
Where is the file located?
Give full path name, e.g.
C:\My Documents\166\USArestr.txt

Choose SAS destination:
Library: (defaults to WORK)
Member: (fill in name of your choice; e.g. USArestr)

Question as to whether you want wizard to generate proc import statements so you can just run them next time

What it generated

```sas
PROC IMPORT OUT= WORK.usarestr
    DATASET= "C:\Documents\166\USArestr.txt"
    DBNDS= DUN REPLACE;
    DELIMITER= '0'='; /* needed correction to DELIMITER='0'='; */
    GETNAMES= YES;
    DATANUM= 2;
RUN;
```

Example of reading Access database

```sas
PROC IMPORT OUT= WORK.courses
    DATASET= "Courses"
    DBNDS= ACCESS97 REPLACE;
    DATABASE= "C:\My Documents\166\uniV0_v7";
RUN;
```

More on file merging
Example: Acid rain deposition in Colorado
Looking at multiple records for each site

- Suppose we want to look at the annual sulfate ion deposition at the CO sites for each year from 1991-2000, inclusive
- We want to estimate site-specific random slopes on year, as well as fixed-effects intercept and coefficients of year and elevation
- Which SAS procedure?
- How should input data look?

```sas
Options linesize = 75 pagesize = 60 nodate nound; data depo;
infile 'depo90s90s.asp' firstobs = 8;
input SiteID $ Per $6. Year Crit1 Crit2 Crit3 Crit4 Ca Mg Ba NH4 N2O InorgC Cl SO4 H3ab HField Svol Pct Pct;
ValidF ValidD Days 0196 Date2 mmddy10. 0299 Date2 mmddy10. ;
drop Per Crit-Crit4 Ca Mg Ba NH4 N2O InorgC Cl H3ab HField Svol Pct Pct ValidF ValidD;
daysop = Date2 - Date1;
format Date2 Date1 date8. ;
run;
*proc sort ;  * needed if records are not already in order ;
*by SiteID ;  *
*run ;
data depo ;
set depo ;
by SiteID ;
run ;
proc print data = depo (obs=25);
run;
data sites;
infile '/space/akules/166/lectures/lect1mk/stateCO.asp' firstobs = 19
missover;
input SiteID $ $30 siteName $18. $40 startDate mmddy10. $53 startDate mmddy10. $68 elev;
if startDate ne ;  * subsetting if: exclude obs meeting condition ;
format startDate startDate date8. ;
drop siteName ;
run;
```

- we need a “match merge”
- must process records in both files by a common variable
- then merge them by this variable
The log file

```
options linesize = 75 page size = 60 nodate nume r;
data depo;
  infile 'depo&mp90s.asp' firstobs = 8;
  input SiteID $ Per $8; Year $ Crit2 Crit 4 Ca Mg
   K Na NH4 N03 InorgCl S04 H1b B H1f S0v Pmt Pct
   ! Valid $ Valid $
   days 0196 Date2 mmddyy10. 2009 Date2 mmddyy10. ;
   d o int Per Crit-Crit2 Ca Mg K Na NH4 N03 InorgCl H1b B H1f
   ! H1f0ed
   S0v Pmt Pct Valid $ Valid $;
   days0 = Date2 - Date1;
   format Date2 Date1 date8. ;
   run;
```

NOTES: The infile 'depo&mp90s.asp' is:
File
Name = /tmp_mnt/space/kcowles/666/lectures/le ct2mnc/depo&mp90s.asp.
Owner Name = kcowles Group Name = faculty.
Access Permissions = r------.
File Size (bytes) = 9562

NOTES: 161 records were read from the infile 'depo&mp90s.asp'.
The minimum record length was 218.
The maximum record length was 218.

NOTES: The data set WORK.DePG has 161 observations and 7 variables.

DATA step used:
real time 0.12 seconds
CPU time 0.07 seconds

Skipping stuff about sites file as we have seen it all.

```
data combined;
  merge depo sites;
  by SiteID;
  run;
```

NOTES: There were 161 observations read from the dataset WORK.DePG.
NOTES: There were 18 observations read from the dataset WORK.Sites.
NOTES: The data set WORK.COMBINED has 162 observations and 10 variables.

DATA step used:
real time 0.02 seconds
CPU time 0.02 seconds

The SMS System

```
The combined file

```
```
```
```
```
```
```
```
```
```

The combined file

```
```
```
```
```
```
```
```
```
```
```

```
The combined file

```
```
```
```
```
```
```
```
```
```
```

```
The combined file

```
```
```
```
```
```
```
```
```
```
```

```
Omitting records missing from one file

data combined ;
merge depo(in=ina) sites ;
   * creates variable "ina" — true if record is in depo, o.w. false ;
by SiteID ;
if ina ;   * subsetting if ;
run ;

Proc mixed

proc mixed data = combined ;
class SiteID ;
model so4 = year elev / s ;
random year / subject = SiteID s ;
run ;

Proc transpose: exchanging rows and columns

- Suppose instead we needed to process the data in the following format:
  - a single row (record) for each site
  - a column (variable) for each year’s so4 value

proc transpose data=combined out=Combran ;
by SiteID ;
id year ;
vvar so4 ;
run ;
proc print data=Combran (obs=10) ;
run ;
More on “by” processing

- We can use certain SAS internal variables to extract certain observations within the “by groups.”

```sas
data firsts04;
set depo;
by SiteID;
if first.SiteID;
run;

proc print data = firsts04;
run;
```