Some Tools, Techniques, and Resources for Quality Federal Education Data Reporting

Quansheng Shen, Ph.D.

Institute of Education Sciences
National Center for Education Statistics

2007 NCES Winter MIS Conference
February 28 – March 2, 2007
Atlanta, Georgia

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

Outline for discussion

- Data Quality and Your Institution
- Data Reporting Process
- Data Reporting Requirements and Specifications
- Tools
  - Data viewing tools: text editors
  - Data processing tools: SQL, SPSS, SAS
- Techniques – input/output, calculation, analysis, verification
- Resources for data validation/ verification

Data Quality and Your Institution

Education data to your institution is as pixels to the picture.

Complete and accurate data

Incomplete and inaccurate data

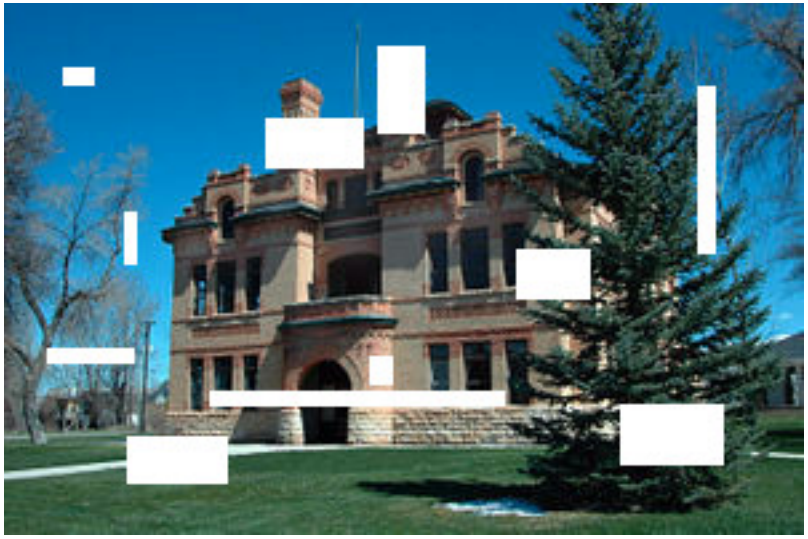
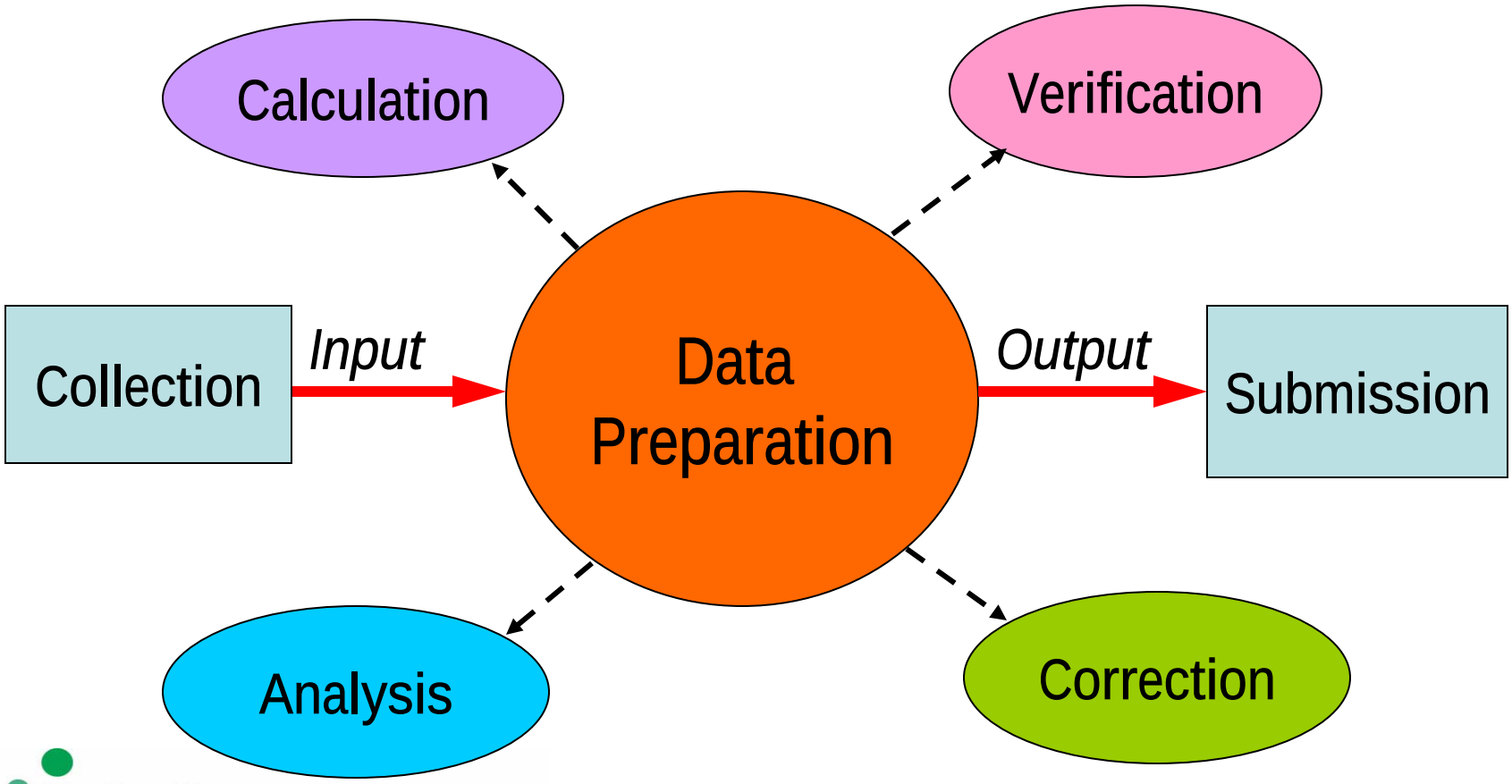


Photo by Cory Maylett

Source of the photo: http://en.wikipedia.org/wiki/Spring_City,_Utah

Process of Data Reporting



Requirements and Specifications for Data Reporting

CCD Instruction Manual – data reporting book

Examples:

- Specific data items for each of the surveys: school, agency, and state
- Enrollment by race/ethnicity by grade by gender
- STAFF FTE counts with one explicit decimal point
- Consistency in CY and PY operational statuses
- Consistency in agency and school grade spans
- Avoid identical data reporting
- Avoid invalid address change
- No word 'COUNTY' in the County Name field
- Mailing or Location City are required to be reported

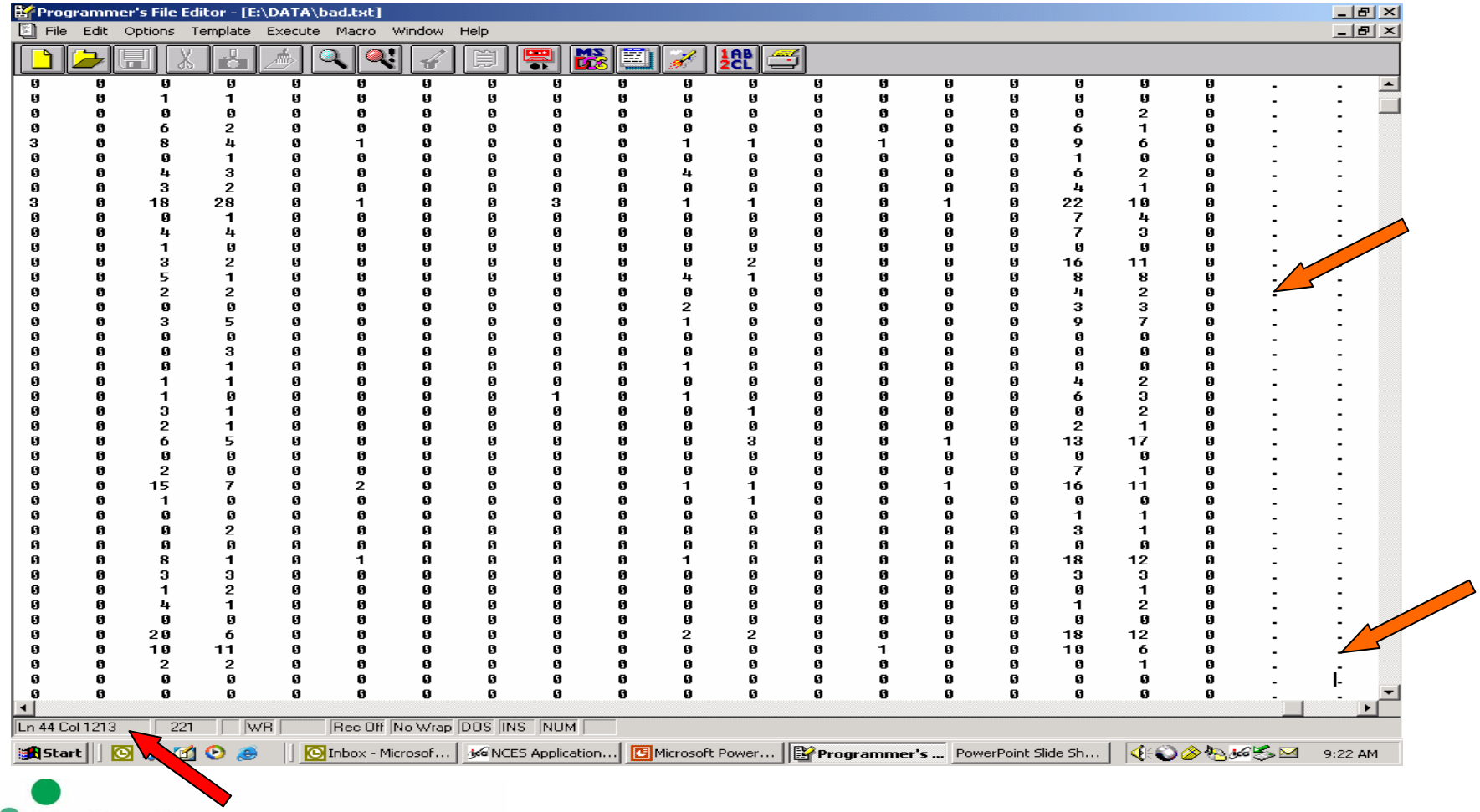
Some data preparation tools

1. Text data editing/viewing tools

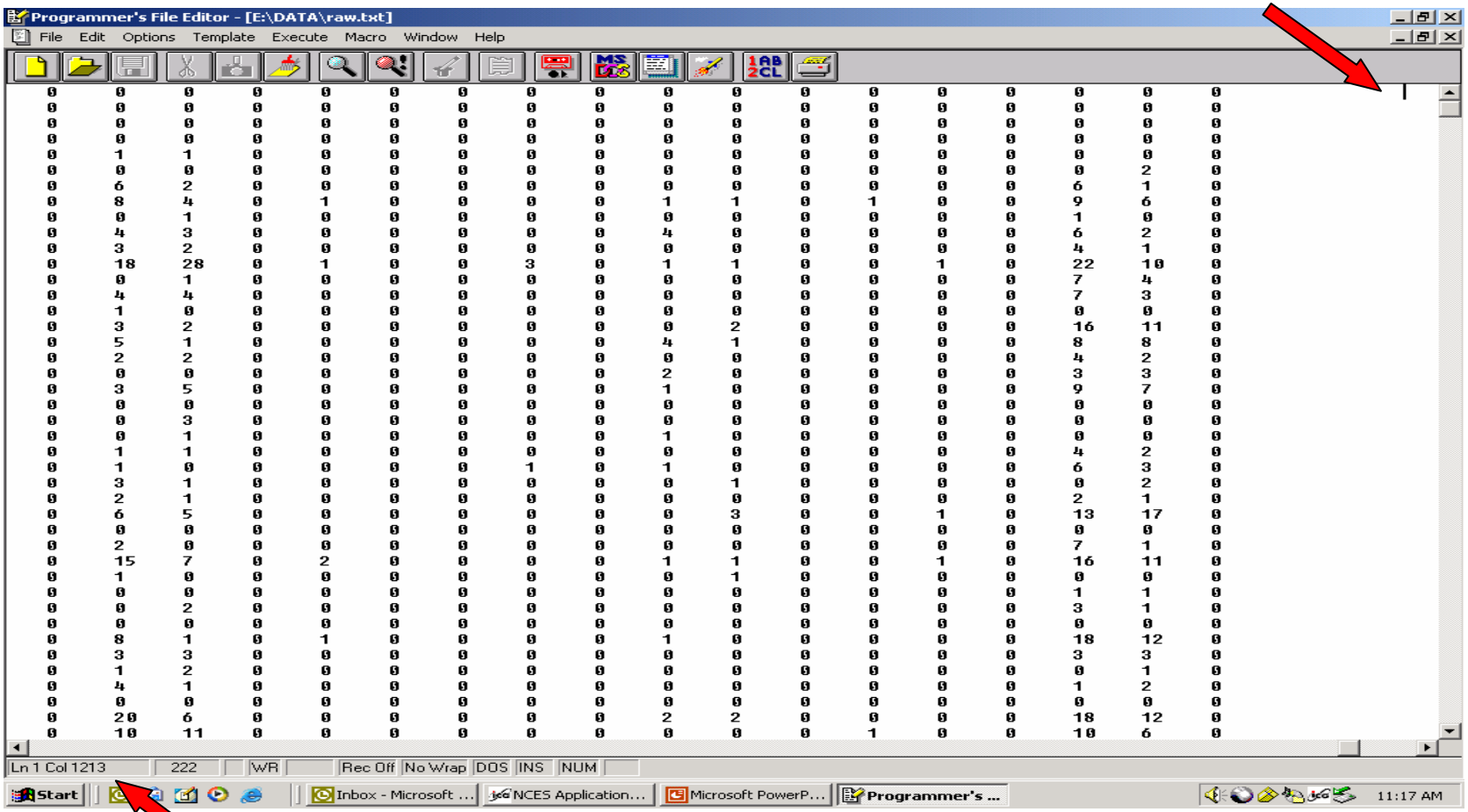
Why do we need a text editor/viewer?
Certain data anomalies can be identified just by eye-balling.

Example: detecting error in output,
examining raw data file, and
making corrections

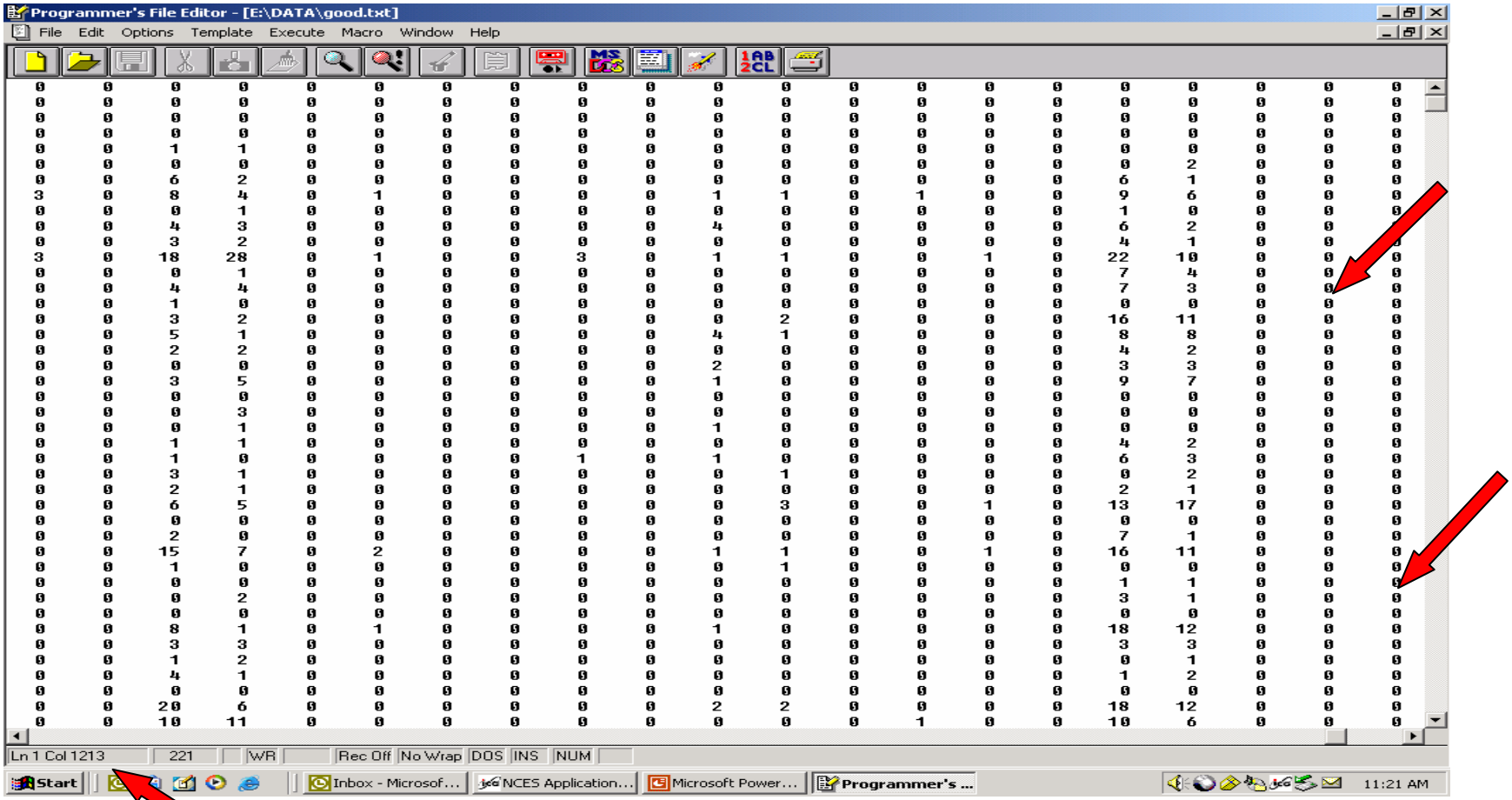
Data editing/viewing tools – an output file with error



Data editing/viewing tools – the original raw data file



Data editing/viewing tools – output with correction



Data Editing/Viewing tools

Choices of text editors:

1. Built-in text editors: DOS, Notepad, WordPad
   Limits: small files, single editor window

2. Programmer's File Editor (PFE)
   Advantages: freely available, multiple file windows, opening large files, cursor position indicator, etc.

   No need for installation – a demonstration.
   Available here: http://www.lancs.ac.uk/people/steveb/cpaap/pfe/default.htm

Data processing tools

SQL – manipulate data within a database management system. It's best to retrieve data from the database systems with such data manipulation command as SELECT, INSERT, UPDATE, DELETE, etc.

SPSS – powerful, user friendly statistical analysis tool. One can write sophisticated programs using SPSS syntax.

SAS – very powerful tool in terms of data manipulation and analysis. Its data step, numerous procedures, and functions make it one of the best tool for data preparation.

Data processing techniques

Examples:

- Input – SAS formats and informats
- Calculations – SAS system return values, etc.
- Analysis and verification – SAS By-group processing / automatic variables
- Correction - SAS functions and procedures
- Output – data formatting

Formatted Input – following CCD specifications

Ways to get the data into shape with SAS informat:

1) Fixed length text file - SAS list input:

```
data school;
  infile ccd_sch;
  input (leaid stid  schno seasch leanm schnam phone)
        ($7.   $14. $5.    $20.   $60.   $50.     $10.  )
        ...
        (status gslo gshi titlei stitli magnet chartr shared fte)
        ($1.    $2.  $2.  $1.   $1.  $1.     $1.     $1.     5.1);
```

Formatted Input *(continued)*

Ways to get the data into shape with SAS informat:

2) Delimited file (csv, tab, etc.) – SAS colon (:) format modifier
   ** Simple list input won't work for the delimited files.

```
data school;
   infile ccd_sch 'x:\school.csv';
   input leaid :$7. stid :$14. schno :$5. seasch :$20. leanm :$60.
         schnam :$50. phone :$10.  status :$1.  gslo :$2. gshi :$2.
         titlel :$1. stitll :$1. magnet :$1. chartr :$1. fte :5.1;
```

Examples: a) output without using the format modifier
b) output using the format modifer
c) output using the format modifer and upcase function

Formatted Input *(continued)*

Ways to get the data into shape with SAS informat:

3) Excel, DBF, MDF, etc. – can be imported/converted into SAS
   ** To conform to the CCD data specifications, need to use SAS LENGTH statement to format data.

   SAS LENGTH statement declares variables and specifies their length and whether their data type is numeric or character

```
data ccd_sch;
  length leaid $7. Stid $20. Schno $5. Seach $20. ...;
  set the_data_file_imported;
run;
```

Formatted Input *(continued)*

SAS proc contents output – formats carried over from Excel file

#	Variable	Type	Len
1	sch_yr	Num	8
2	schnam	Char	60
3	leanm	Char	35
4	coname	Char	15

NOTE: SCH_YR is printed out as 16618 rather than 07/01/2005.

SAS proc contents output – formats using SAS LENGTH and FORMAT statement

```
length schnam $50. leanm $60. coname $30.;
format Sch_yr mmddyy10.;
```

#	Variable	Type	Len	Format
1	schnam	Char	50	
2	leanm	Char	60	
3	coname	Char	30	
4	Sch_yr	Num	8	MMDDYY10.

Data Preparation - Calculation

CCD requirements: Enrollment by grade by race/ethnicity by gender
Techniques: SAS system return value and proc means

```
STUDENT COUNTS BY RACE/ETHNICITY BY GRADE BY GENDER:
AMPKM = (GENDER = 'M' and RACE = ‘A' and GRADE = 'PK');
WHKGF = (GENDER = 'F' and RACE = 'W' and GRADE = 'KG');
AS03M = (GENDER = 'M' and RACE = ‘S' and GRADE = '03');
BL01F = (GENDER = 'F' and RACE = 'B' and GRADE = '01');
HI03M = (GENDER = 'M' and RACE = 'H' and GRADE = '03');
WH06U = (GENDER = ‘U' and RACE = 'W' and GRADE = '06');

STUDENT COUNTS BY RACE/ETHINICITY BY GENDER:
AMALM = (RACE = ‘A’ and GENDER = ‘M’);

STUDENT COUNTS BY GRADE:
TOTPK = (GRADE = ‘PK’);
TOTKG = (GRADE = ‘KG’);

GRAND TOTAL:
TOTAL = 1;
```

Every student is counted !

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

Data Preparation – calculation *(continued)*

Aggregate the individual level data by school:

A. Sort the data by the unique school Ids
B. Use SAS means procedure

Example:

```
proc sort data=your school data file; by leaid schno; run;
proc means SUM data=your sorted school data file;
  id leaid schno;
  var AMPKM ASKGF BLALM TOTKG TOTAL;
  output out=your output file name;
run;
```

Data Preparation – calculation *(continued)*

Calculated Student Counts by Race/Ethnicity by Grade by Gender

LEAID	SCHID	GENDER	RACE	GRADE	AM04F	AM05F	WHPKM	WHKGF	AM03M	BL01F	HI03M	WH06F
8800001	001	M	W	PK	0	0	1	0	0	0	0	0
8800001	001	F	B	01	0	0	0	0	0	1	0	0
8800002	101	F	W	KG	0	0	0	1	0	0	0	0
8800002	101	M	A	03	0	0	0	0	1	0	0	0
8800003	202	M	H	03	0	0	0	0	0	0	1	0
8800003	202	F	N	04	1	0	0	0	0	0	0	0
8800003	202	F	N	05	0	1	0	0	0	0	0	0
8800003	451	F	W	06	0	0	0	0	0	0	0	1
8800003	451	F	W	06	0	0	0	0	0	0	0	1

Raw data fields

Calculated data fields

Data Preparation – calculation *(continued)*

Enrollment data summed by school

LEAID	SCHID	AM04F	AM05F	WHPKM	WHKGF	AM03M	BL01F	HI03M	WH06F
8800001	001	0	0	1	0	0	1	0	0
8800002	101	0	0	0	1	1	0	0	0
8800003	202	1	1	0	0	0	0	1	0
8800003	451	0	0	0	0	0	0	0	2

Calculated data fields

Data Preparation – analysis, verification

Examples: 1. Verifying the school universes – operational statuses

CCD requirements: new, closed, added, changed agency, etc.
Techniques: merge CY data with PY data and cross-tabulate CY status by PY status:

```
proc freq data=your_PY_CY_merged_data;
  tables status_py*status_cy;
run;
```

Data Preparation – analysis, verification *(continued)*

Examples: 1. Verifying your school universes – operational statuses

b)

CY STATUS

a)

PY STATUS

	Open	Closed	New	Added	Changed	Inactive	Future	Reopen
Open	1865	15			3	1		2
Closed	1	8						1
New	20	2						
Added	1							
Changed	2							
Inactive						5		
Future	25							
Reopen	1							

c)

Conclusion: a) PY closed school cannot be reported open in CY.
b) PY closed schools should not be reported in CY.
c) PY open schools cannot be reported reopen in CY.

Data Preparation – analysis, verification *(continued)*

Examples: 2. Verifying school and agency grade span

CCD requirements: agency grade span should cover the associated school grade spans.

Techniques: Use SAS by group processing and automatic variables.

a) Find out the lowest and highest grade spans offered by the associated schools
b) Merge the associated school grade span with the agency file
c) Cross-tabulate the agency grade span by school grade span
d) Analyze, verify, and adjust if necessary.

Data Preparation – analysis, verification *(continued)*

Examples: 2. Verifying school and agency grade span

Step 1. Create a numeric variable corresponding to GSLO, GSHI respectively
e.g. gslon=0 if gslo='PK'; gslon=1 if gslo='KG' … gslon=2 if gslo='01'; etc., and
gshin=0 if gshi='PK'; gshin=1 if gshi='KG' … gshin=4 if gshi='03'; etc.

Step 2. Sort data by leaid schno gslon for the lowest grade offered by the associated schools,
Sort data by leaid schno gshin for the highest grade offered by the associated schools.

Step 3. Use SAS by group processing and the automatic variables first. and last.

a)
```
data oneslo multislo;
    set your_data;
    by leaid schno;
    rename gslo_s=gslo;
    if first.leaid and last.leaid then output oneslo;
    else output multislo;
```

Data Preparation – analysis, verification *(continued)*

Examples: 2. Verifying school and agency grade span

Step 3. Use SAS by group processing and the automatic variables first. and last.

b) proc sort data=multislo; by leaid schno gslon; run;

```
data multislo2 (keep=leaid gslo);
    set multislo;
    by leaid schno;
    if gslon=. then delete;   /* gslo='00' or 'NN' */
    if first.leaid;
```

c) combine oneslo with multislo2 so that we have the lowest grade offered for each LEA.

e.g. data sch_gslo; set oneslo multislo2;

Data Preparation – analysis, verification *(continued)*

Examples: 2. Verifying school and agency grade span

Step 4. Following the same procedures for the highest grade offered except

e) for multishi file, use if last.leaid

f) combine gshi files: data sch_gshi; set oneshi multishi2;

DATA FIELDS						SAS Automatic Variables	
LEAID	SCHNO	GSLO	GSHI	GSLO_N	GSHI_N	FIRST OBS	LAST OBS
9900001	90001	PK	06	0	8	1	1
9900002	90011	PK	07	0	9	1	0
9900002	90012	05	07	7	9	0	0
9900002	90013	09	12	11	14	0	1

Single school district

Multiple schools district

Data Preparation – analysis, verification *(continued)*

Examples: 2. Verifying school and agency grade span

Step 5. Merge sch_gslo with sch_gshi, and then merge the resultant file with the LEA file. The table below shows what the merged file looks like:

LEAID	GSLO	GSHI	GSLO_S	GSHI_S
1000000	01	12	01	12
2000000	05	08	03	09
3000000	09	12	08	12

Resultant File with Above Manipulation

Merged File without Manipulation

LEAID	GSLO	GSHI	GSLO_S	GSHI_S
1000000	01	12	01	06
1000000	01	12	09	12
2000000	05	08	03	05
2000000	05	08	04	09
3000000	09	12	08	10
3000000	09	12	09	11
3000000	09	12	09	12

District
Associated Schools

Data Preparation – analysis, verification *(continued)*

Examples: 2. Verifying school and agency grade span

Step 6. Cross-tabulate LEA GSLO by school GSLO and LEA GSHI by school GSHI
The results allow you to pin-point the changes that need to be made.

School GSLO

a)

b)

Agency GSLO

	01	03	08
01	1	0	0
05	0	1	0
09	0	0	1

c)

School GSHI

Agency GSHI

	09	12
08	1	0
12	0	2

Conclusion: a) The agency GSLO='05' needs to be adjusted to '03',
b) The agency GSLO='09' needs to be adjusted to '08',
c) The agency GSHI='08' needs to be adjusted to '09'.

Data Preparation – analysis, verification *(continued)*

Verify the data so that GENERALLY

1) what should be the same should remain the same and

2) what should be different is different

Technique: SAS compare procedure (proc compare).

Example 1. Verifying the school address information;

Example 2. Verifying the school enrollment information;

Data Preparation – analysis, verification *(continued)*

Bad Addresses Reported – what should be same is not.

Comparison of WORK.NN03 with WORK.NN04
PY = Prior Year CY = Current Year

leaid	SCHNO	Mailing Address Base Value MSTREE04 (PY)	Compare Value MSTREE04 (CY)
9900030	00001	1222 SOUTH PARK ST	250 SOUTH CHERRY STRE
9900030	00003	250 SOUTH CHERRY STRE	777 DISCOVERY ST
9900030	00004	999 EAST JOHNSONS	888 SOUTH RUSSELL ST
9900030	00005	444 VENTURE PLACE	1201 CORPORATE DR. S
9900030	00006	888 SOUTH BRUSSELL	444 N.VENTURE LAN
9900030	00021	1099 MERT DRIVE	999 EAST JOHNSONS AV
9900030	00431	777 DISCOVERY ST	1099 MERT DRIVE
9900060	00007	400 PAL PLAZA	400 PAL PLAZA
9900060	00008	5701 SHACK AVENUE	3335 W. UMAUL AVENUE
9900060	00009	5445 SOUTH OAK	5445 S. OAKLEY
9900060	00010	912 VILLAGE LANE	PO BOX 160
9900060	00011	6669 REY AV	6669 REY AVE
9900060	00012	900 NORTH AVENUE	7777 LYNN ST
9900060	00013	11 SIXTH AVENUE	101 5TH ST
9900060	00014	1100 JOHN STREET	950 N. TON DR.
9900060	00015	1312 EAST TON AV	1100 JOHN ST
9900060	00016	2101 WEST CART AV	130 PAUL WAY
9900060	00017	3875 E AVE	3875 E AVE

Key Points:

Make the corresponding field names in two years files the same to trick PROC COMPARE as if we compare the different versions of the same File, or use VAR ... WITH specification.

Note: The data errors are real, but all addresses in the example are fictitious

Data Preparation – analysis, verification *(continued)*

Bad Enrollment Data Reported – what should be different is not.

When 150 observations out of 151 with ALL compared variables from two years equal, that suggests a problem!

```
                    The COMPARE Procedure
          Comparison of WORK.ENR04 with WORK.ENR05
                      (Method=EXACT)

                    Observation Summary

   Observation      Base   Compare   ID

   First Obs           1         1   leaid2=9900003 SCHNO=00015
   First Unequal      94        93   leaid2=9900007 SCHNO=00206
   Last  Unequal      94        93   leaid2=9900007 SCHNO=00206
   Last  Match       153       151   leaid2=9900011 SCHNO=00189
   Last  Obs           .       153   leaid2=9900013 SCHNO=00071

Number of Observations in Common: 151.
Number of Observations in WORK.ENR04 but not in WORK.ENR05: 2.
Number of Observations in WORK.ENR05 but not in WORK.ENR04: 2.
Total Number of Observations Read from WORK.ENR04: 153.
Total Number of Observations Read from WORK.ENR05: 153.

Number of Observations with Some Compared Variables Unequal: 1.
Number of Observations with All Compared Variables Equal: 150.
```

Data Preparation – analysis and correction

Data Requirement: No word 'COUNTY' in the County Name field

Techniques: SAS function – TRANWRD

TRANWRD replaces or removes all occurrences of a word in a character string

raw data:

```
LEAID   SCHNO CONAME
1000000 00101 APPLE COUNTY
1000000 00102 ORANGE COUNTY
2000000 00201 BANANA COUNTY
2000000 00202 CANTELOP COUNTY
3000000 00301 PEAR
3000000 00302 APRICOT COUNTY
3000000 00303 PEACH COUNTY
```

code snippet:

```
data coname2;
    set coname;
    coname2=tranwrd(coname,"COUNTY","");
run;
```

SAS OUTPUT

OBS	LEAID	SCHNO	CONAME	CONAME2
1	1000000	00101	APPLE COUNTY	APPLE
2	1000000	00102	ORANGE COUNTY	ORANGE
3	2000000	00201	BANANA COUNTY	BANANA
4	2000000	00202	CANTELOP COUNTY	CANTELOP
5	3000000	00301	PEAR	PEAR
6	3000000	00302	APRICOT COUNTY	APRICOT
7	3000000	00303	PEACH COUNTY	PEACH

Data Preparation - Resources

1. Internet

1) State, agency, and school web sites – for various data
2) Google, Yahoo search engines – directory data
3) Yahoo, Google maps – for address information
4) FIPS http://www.itl.nist.gov/fipspubs/index.htm
5) USPS http://www.usps.com/

2. Telephone

Talk to the school or district and usually you will get the data you won't be able to otherwise.

Data Preparation – Resources *(continued)*

Example:

Data requirement: a record must have a CITY address, and
a record should have zip code.

If the city address information is missing, you may enter the street address, state abbreviation, and zip code into Yahoo Map or Google Map. It should show the city information.

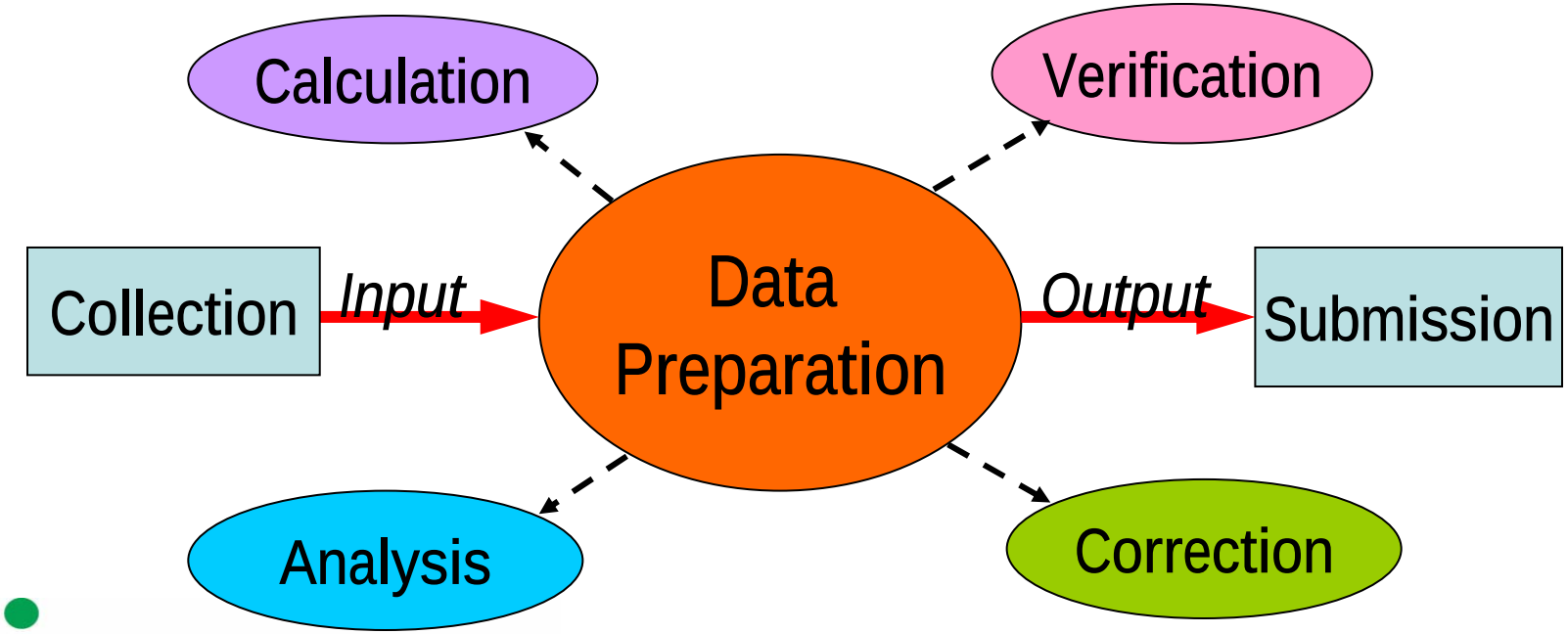
If the zip code information is missing, you may do the same as above, and the zip code information should show up.

Data Preparation – Resources *(continued)*

Phone calls can be very helpful!

Certain information such as directory that is missing may be just a phone call away.

Example: new schools, future schools

Call the districts.

Recap: Quality federal data reporting depends on the reporter's ability to conform to the data requirements and specifications, proper tools and techniques used and, of course data reporters' dedication.