Green Eggs And SAS®

Presented To The
Edmonton SAS User Group

October 24, 2017

By John Fleming

How To Merge SAS Programming With Dr. Seuss

```
Data new_ds;
     Merge sas.programming (in=in1)
           dr.seuss (in=in2);
     By green eggs and ham;

     If in1 and in2;

Run;
```

Every So Often, Someone Comes By With A Question

Do you like
green eggs and SAS?

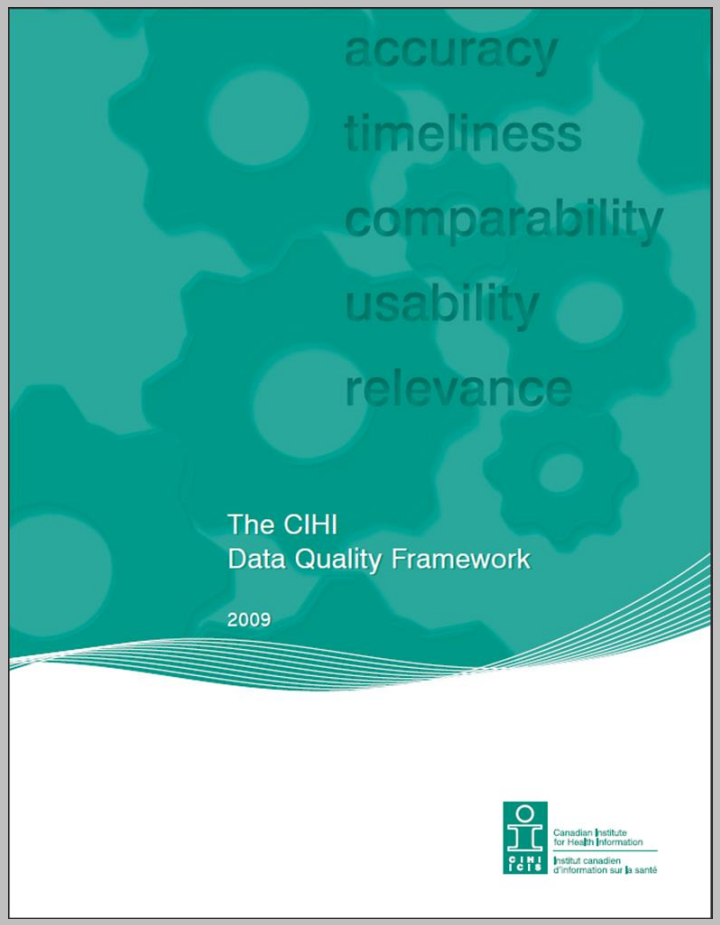
The Questions Are More Likely To Be Something Like

Is this quality data?

How do you know?

What Is Quality Data?

What Is Quality Data?



Our understanding of what constitutes quality data is informed by standards such as "The CIHI Data Quality Framework."

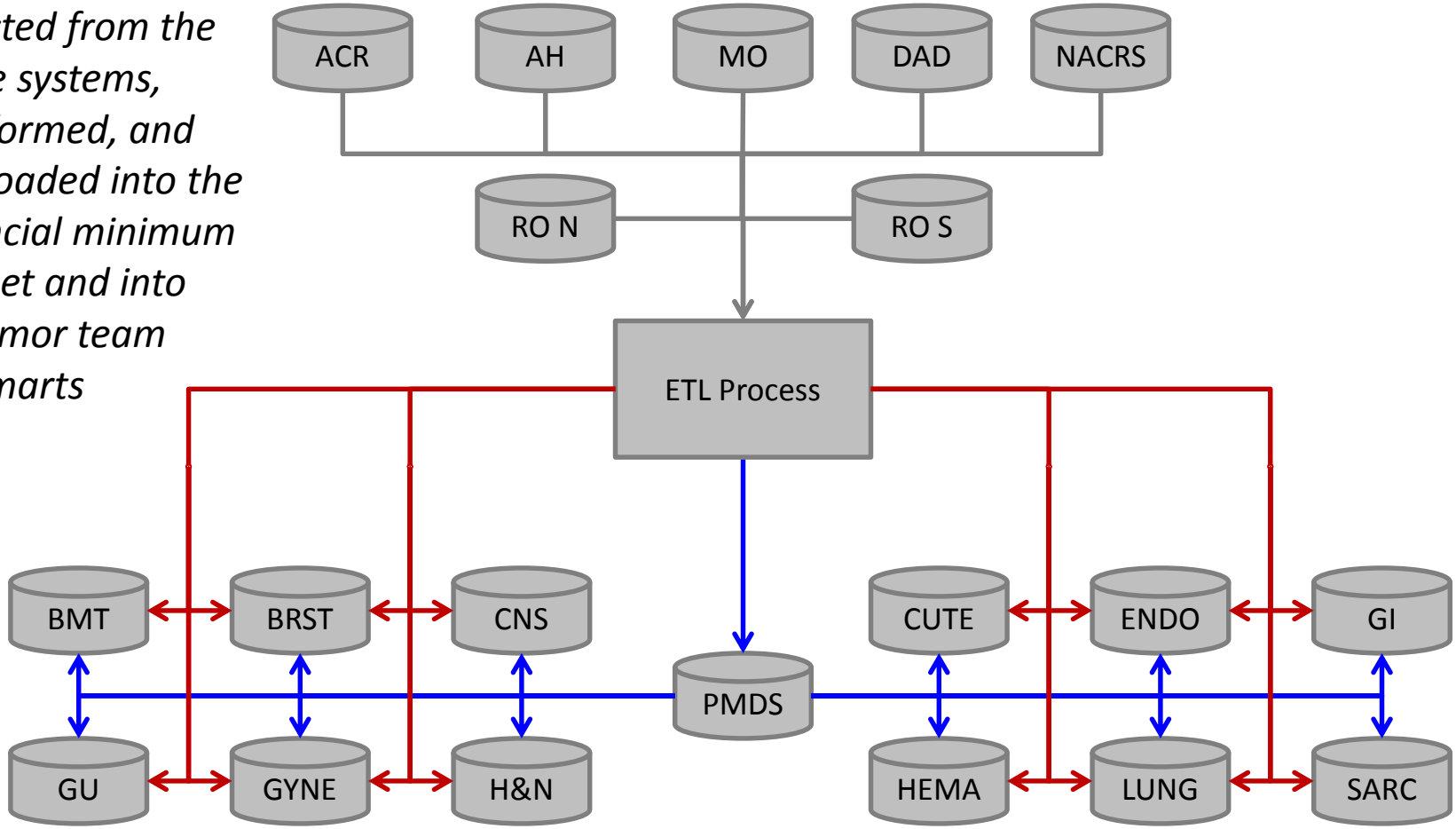
According To “The CIHI Data Quality Framework”

Quality data is:

- Accurate
- Timely
- Comparable
- Usable
- Relevant

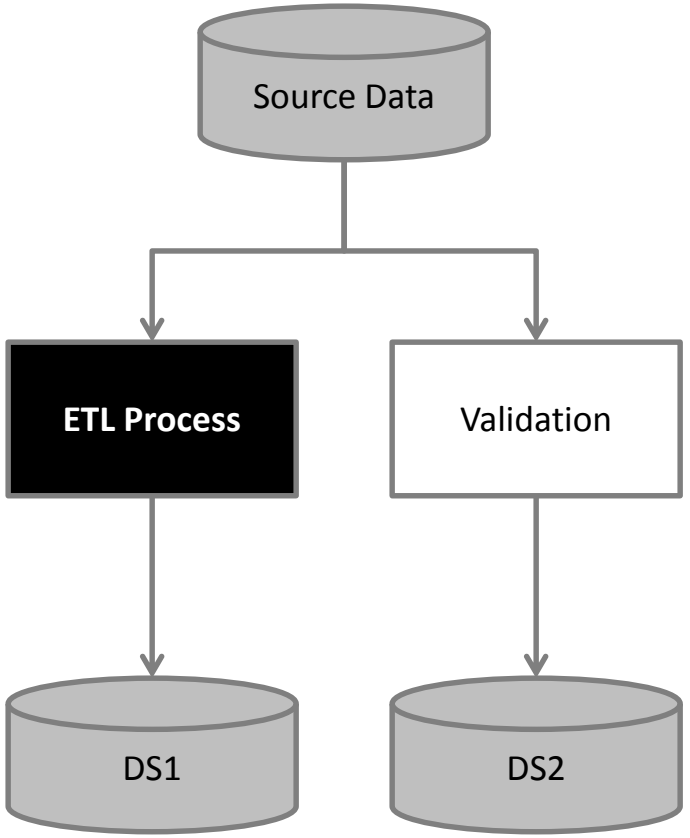
Overview Of The Data Warehouse Building Process

A subset of data is extracted from the source systems, transformed, and then loaded into the provincial minimum data set and into the tumor team data marts



The Data Validation Challenge

By applying the same business rules to data from the same sources, can we produce data that compares positively to the data in the data mart? That is, do we get something similar when we compare ds1 and ds2?



We Won't Produce Identical Results Because:

- *Data warehouse data mart data is refreshed overnight.*
- *We do our data validation using real time data from live systems.*

Validation Raises Some Questions

Would you program in a box?

Would you program with a fox?

Actually, What We Really Want To Know

Are the differences we see in the data:

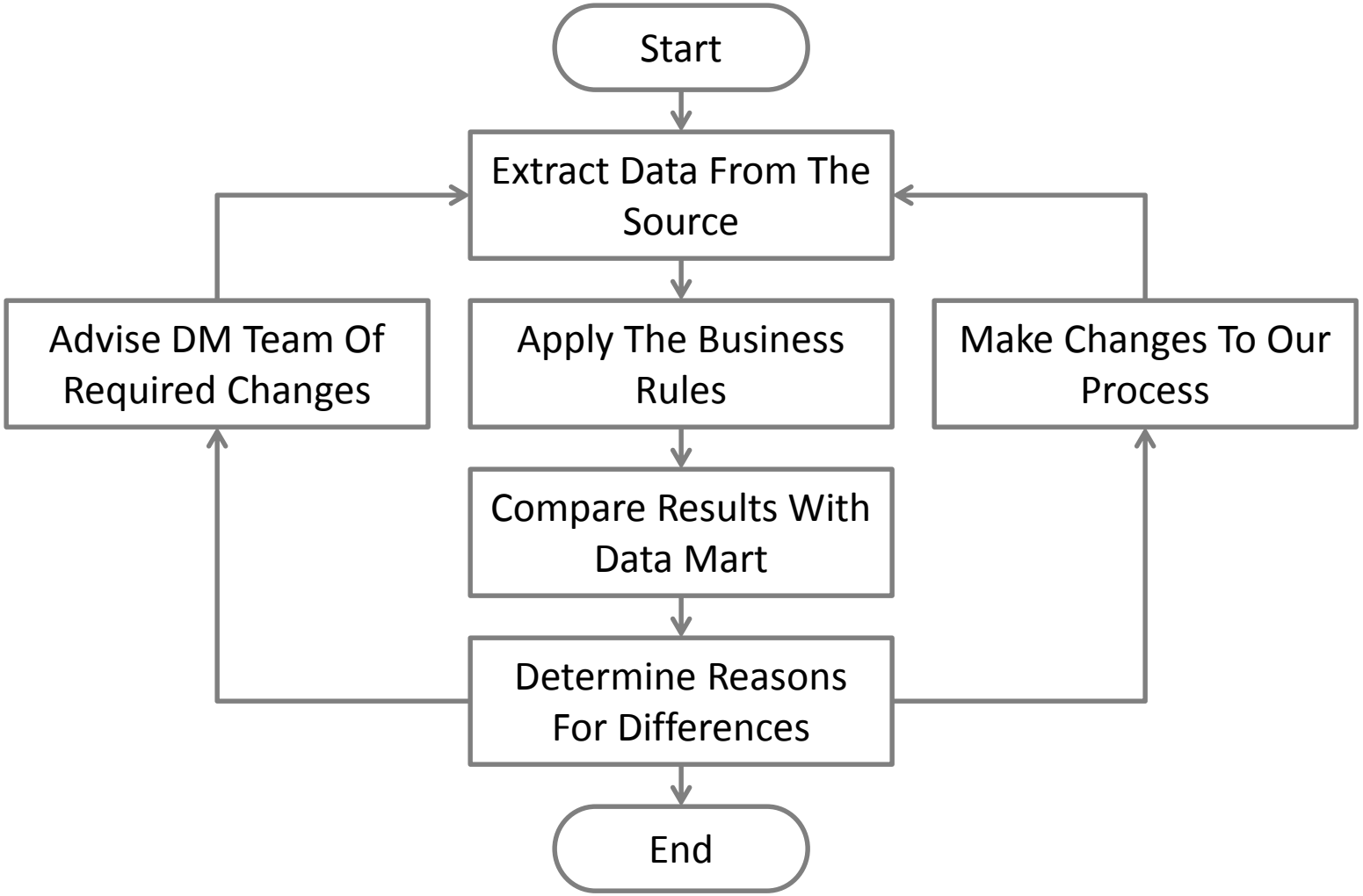
- *The result of issues in the process that loads data into the data warehouse data mart?*
- *The result of issues in our own SAS validation code?*
- *Or the result of changes to the data that occur in the normal day-to-day course of business?*

For Example

The data mart says that, between date 1 and date 2, we had 504 cases of breast cancer diagnosed.

Our validation results say that, between date 1 and date 2, we had 502 cases of breast cancer diagnosed.

The Basic Validation Process



The Rules For Validation

For variables that are not derived, we must not modify the underlying value of the variable in any data record.

For variables that are derived, we must use the same derivation rules with data from the same source tables as defined in the design of the data mart.

The data warehouse has variables for indexing, for example, personseq to uniquely identify each patient. We are not required to recreate these variables.

Comparing The Two Datasets

```
Proc compare base=ds1
             comp=ds2 (drop=personseq)
             out=ds3
             outnoequal
             noprint;
     By var1 var2 . . . Varn;

Run;
```

What We Want to See

```
NOTE: There were 98 observations read from the data set WORK.QSTR_FLIP_PRE_MYELOMA_2.
NOTE: There were 98 observations read from the data set WORK.QSTR_BMT_MMP_VW.
NOTE: The data set WORK.ZNEQSTR_FLIP_PRE_MYELOMA_2 has 0 observations and 23 variables.
NOTE: PROCEDURE COMPARE used (Total process time):
      real time           0.01 seconds
      user cpu time       0.00 seconds
      system cpu time     0.01 seconds
      memory              483.37k
      OS Memory           19644.00k
      Timestamp           08/15/2017 02:03:42 PM
```

What We Often Get

```
NOTE: There were 101 observations read from the data set WORK.DS1.
NOTE: There were 101 observations read from the data set WORK.DS2.
NOTE: The data set WORK.DS3 has 20 observations and 35 variables.
NOTE: PROCEDURE COMPARE used (Total process time):
      real time           0.66 seconds
      user cpu time       0.66 seconds
      system cpu time     0.01 seconds
      memory              25461.18k
      OS Memory           58268.00k
      Timestamp           10/19/2017 01:06:30 PM
      Step Count                       72  Switch Count  10
      Page Faults                      0
      Page Reclaims                    6021
      Page Swaps                       0
      Voluntary Context Switches       34
      Involuntary Context Switches     5
      Block Input Operations           0
      Block Output Operations          960
```

Use Outnoequal Output To Get Records Of Interest

```
Proc sql;

Create table ds4 as

      Select ds1.* from ds1, ds3

      Where ds1.var1 = ds3.var1
            and Ds1.var2 = ds3.var2
            and . . .

Quit;
```

Make Cosmetic Changes To Variable Attributes

```
Proc datasets library=work noprint;

     Modify ds1;

          Rename var11=var1a;
          Format var1a $25.;
          Informat var1a $25.;
          Label var1a="Hello World!";

Quit;
```

Why Use Proc Datasets?

When we use Proc Datasets to make cosmetic changes, we don't need to open the data set and read the data.

This is an advantage when we are dealing with data sets with hundreds of thousands of records and dozens of variables.

How To Get The Variable Attributes

```
Proc contents data=ds1 out=ds2
              noprint;

Run;
```

I Need A Dataset With Exactly The Same Structure

```
Proc sql;

     Create table ds2 like ds1;

Quit;
```

Use Macros And Macro Functions To Automate

```
%macro ima_macro(ds1=, ds2=, ds3=,
                 vars=);

      Proc compare base=&ds1
                   comp=&ds2
                   out=&ds3
                   outnoequal
                   nopring;
           by &vars;

%mend ima_macro;
```

Sample Macro Call

```
%ima_macro (ds1 = x,
            Ds2 = y,
            Ds3 = z,
            Vars = var1 var2 var3 . . . varn);
```

After Making Our Changes, What We Hope to See

```
NOTE: There were 98 observations read from the data set WORK.QSTR_FLIP_PRE_MYELOMA_2.
NOTE: There were 98 observations read from the data set WORK.QSTR_BMT_MMP_VW.
NOTE: The data set WORK.ZNEQSTR_FLIP_PRE_MYELOMA_2 has 0 observations and 23 variables.
NOTE: PROCEDURE COMPARE used (Total process time):
      real time           0.01 seconds
      user cpu time       0.00 seconds
      system cpu time     0.01 seconds
      memory              483.37k
      OS Memory           19644.00k
      Timestamp           08/15/2017 02:03:42 PM
```

The Answer We Want To Give

Is this quality data?

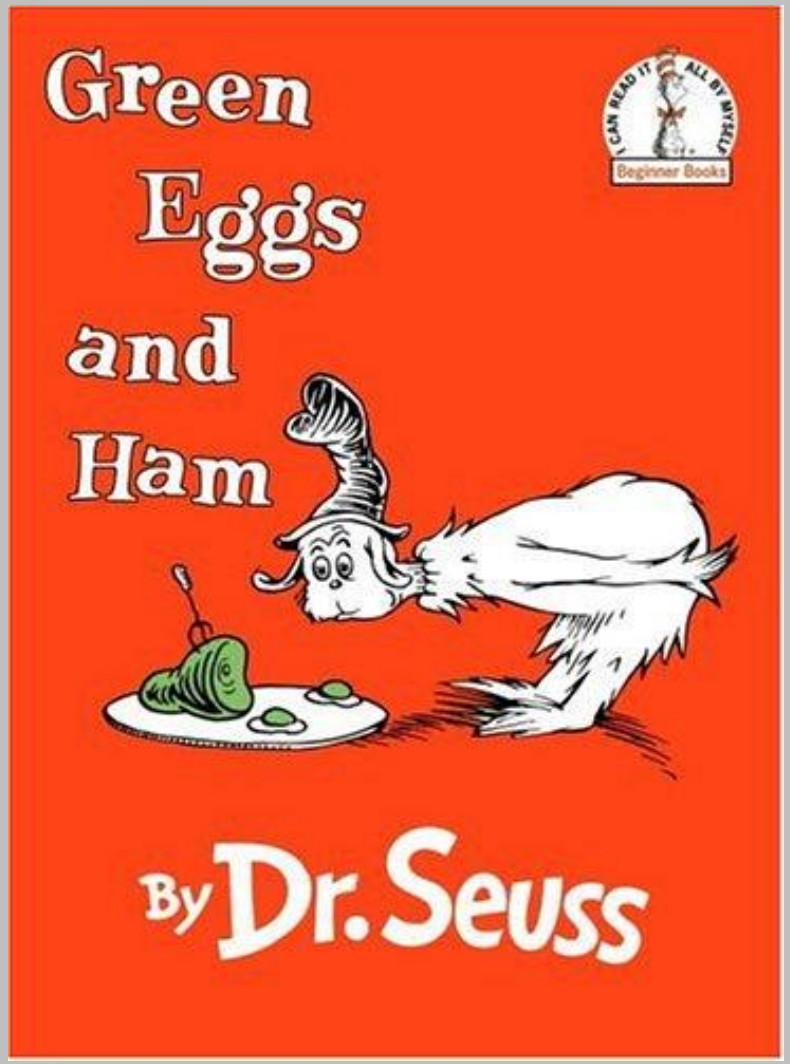
How do you know?

I have the validation right here.

As To Those Other Questions

Do you like
green eggs and SAS?

I do so like
green eggs and SAS.



Green Eggs and Ham is copyright Dr. Seuss Enterprises. Any material borrowed and adapted from this source for use in this presentation is used here for non-commercial and educational purposes only.

Questions?

John Fleming
Alberta Health Services

(780) 643 – 4341

John.fleming@albertahealthservices.ca