Processing of data discrepancies for U.S. dairy cattle and effect on genetic evaluations

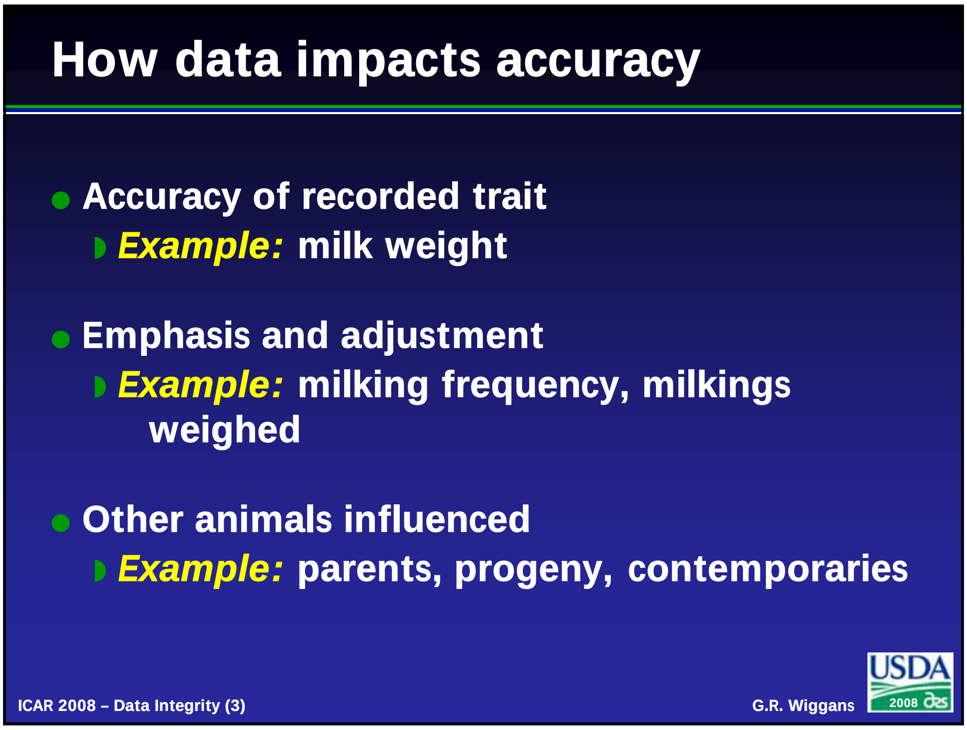
G.R. Wiggans and L.L.M. Thornton

Animal Improvement Programs Laboratory
Agricultural Research Service, USDA, Beltsville, MD, USA

george.wiggans@ars.usda.gov

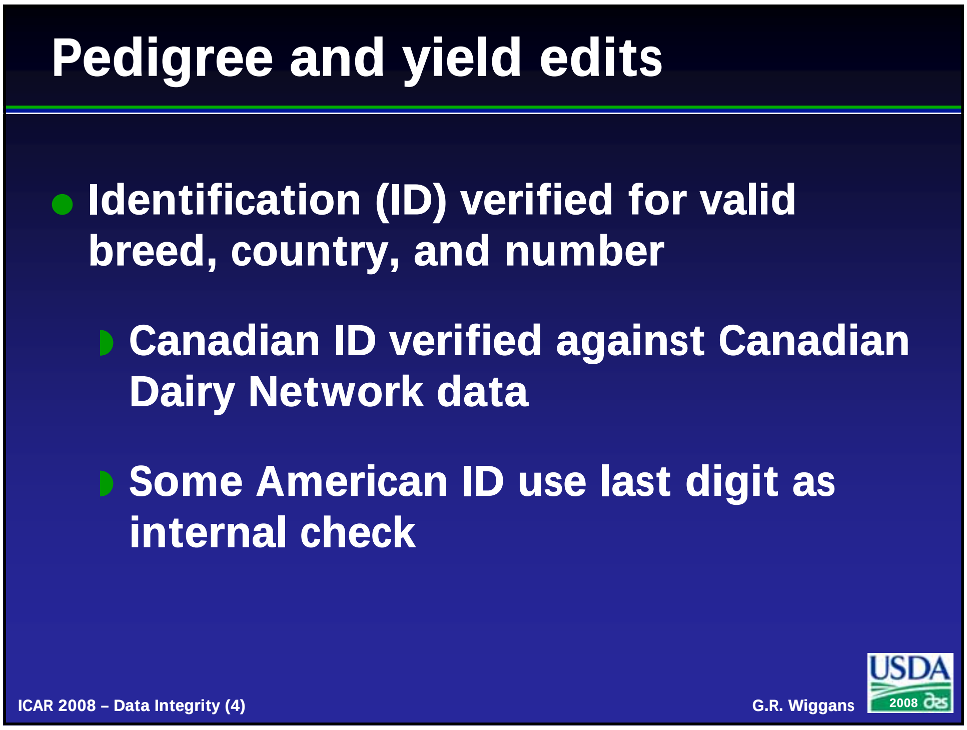
Data sources

- Dairy records processing centers – **milk recording**
- Breed registry societies – **pedigree and conformation (type)**
- National Association of Animal Breeders – **calving traits and bull status**



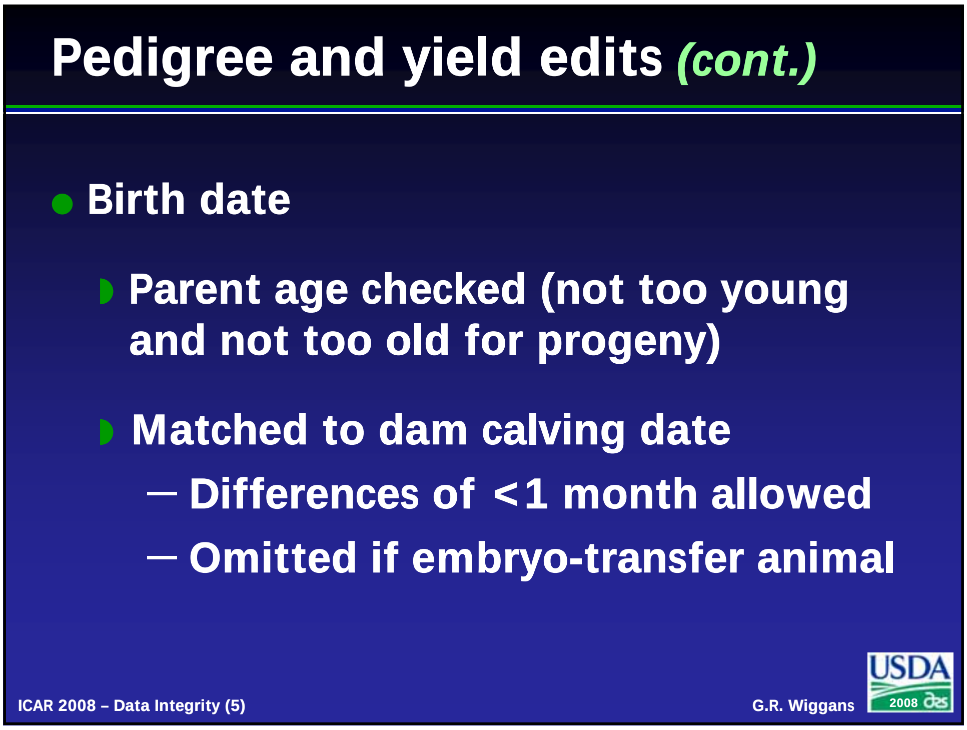
How data impacts accuracy

- Accuracy of recorded trait
  - *Example:* milk weight
- Emphasis and adjustment
  - *Example:* milking frequency, milkings weighed
- Other animals influenced
  - *Example:* parents, progeny, contemporaries



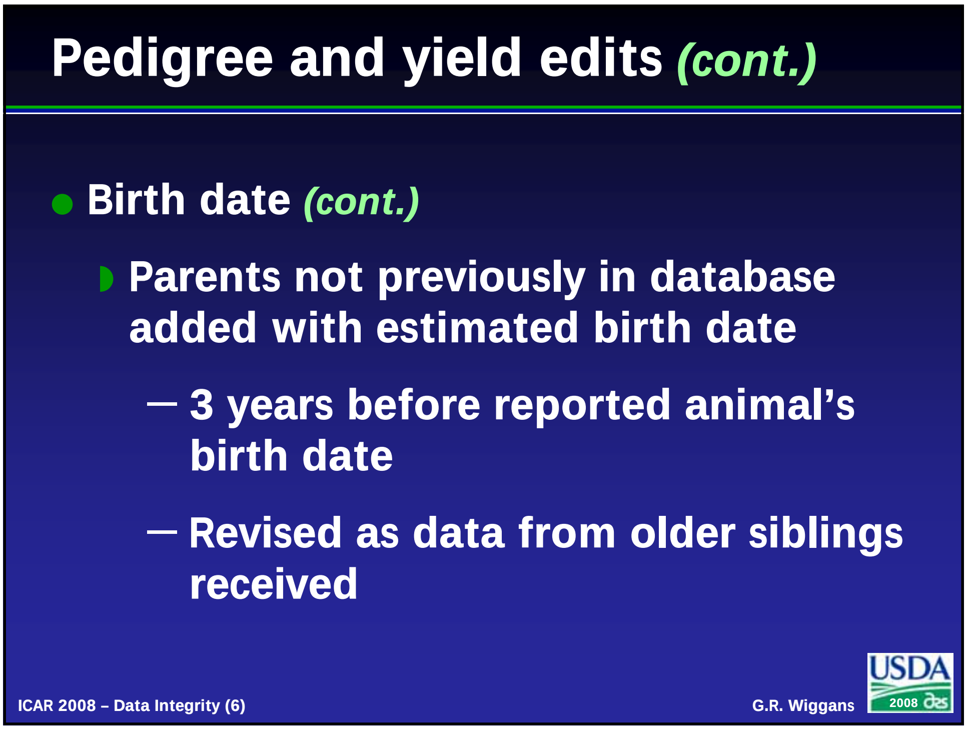
Pedigree and yield edits

- Identification (ID) verified for valid breed, country, and number
  - Canadian ID verified against Canadian Dairy Network data
  - Some American ID use last digit as internal check



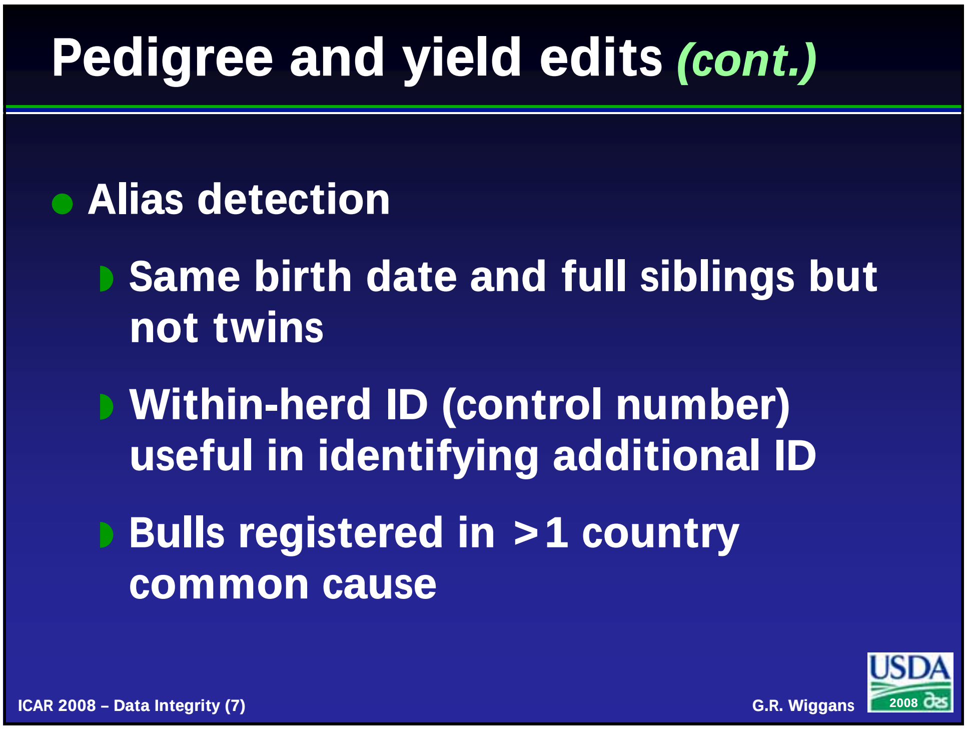
Pedigree and yield edits *(cont.)*

- Birth date
  - Parent age checked (not too young and not too old for progeny)
  - Matched to dam calving date
    - Differences of <1 month allowed
    - Omitted if embryo-transfer animal



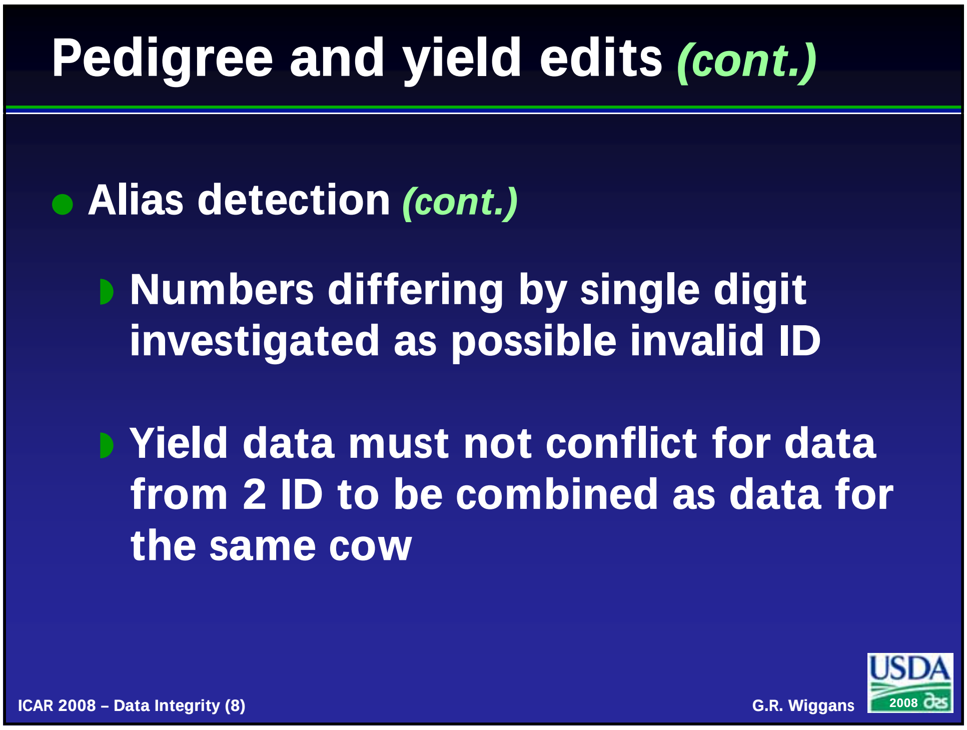
Pedigree and yield edits *(cont.)*

- Birth date *(cont.)*
  - Parents not previously in database added with estimated birth date
    - 3 years before reported animal's birth date
    - Revised as data from older siblings received



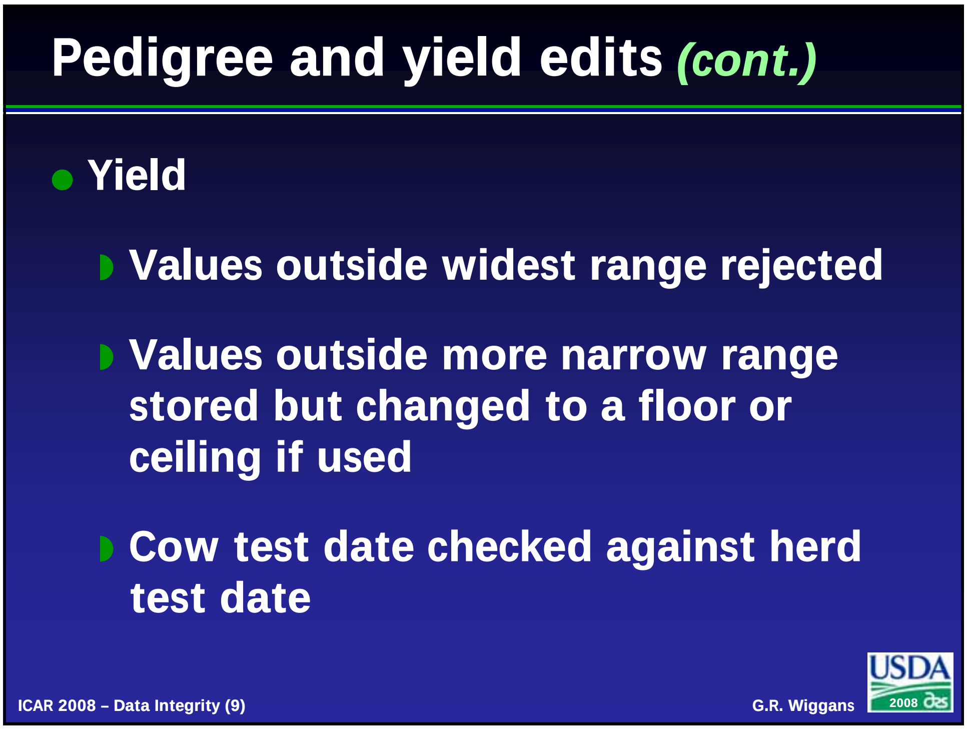
Pedigree and yield edits *(cont.)*

- Alias detection
  - Same birth date and full siblings but not twins
  - Within-herd ID (control number) useful in identifying additional ID
  - Bulls registered in >1 country common cause



Pedigree and yield edits *(cont.)*

- Alias detection *(cont.)*
  - Numbers differing by single digit investigated as possible invalid ID
  - Yield data must not conflict for data from 2 ID to be combined as data for the same cow

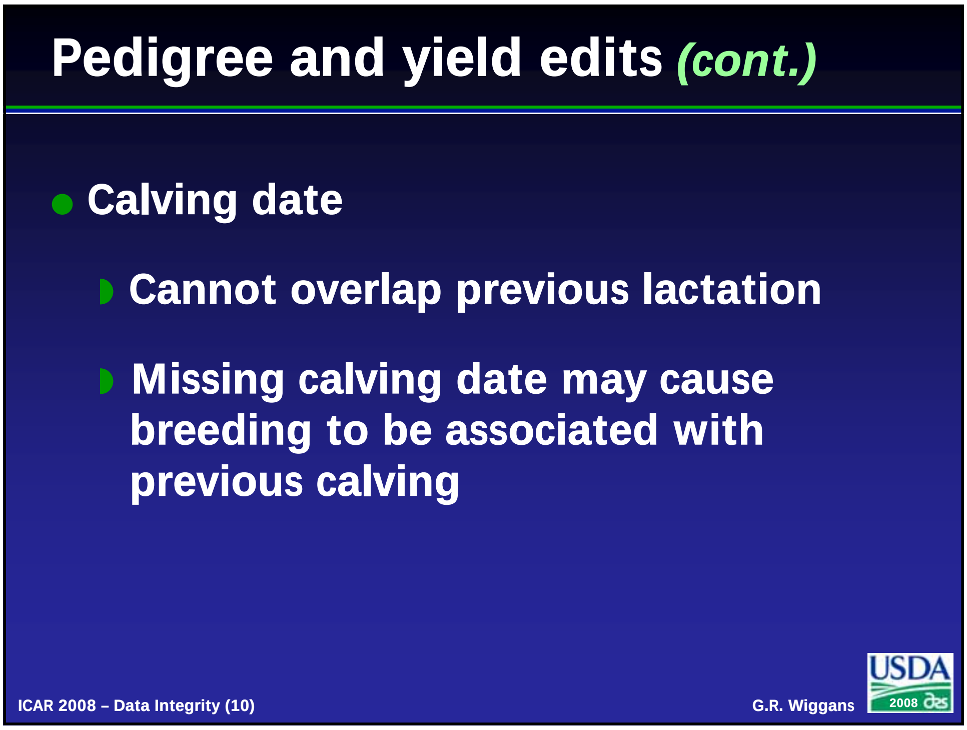


Pedigree and yield edits *(cont.)*

- Yield
  - Values outside widest range rejected
  - Values outside more narrow range stored but changed to a floor or ceiling if used
  - Cow test date checked against herd test date

ICAR 2008 – Data Integrity (9)

G.R. Wiggans

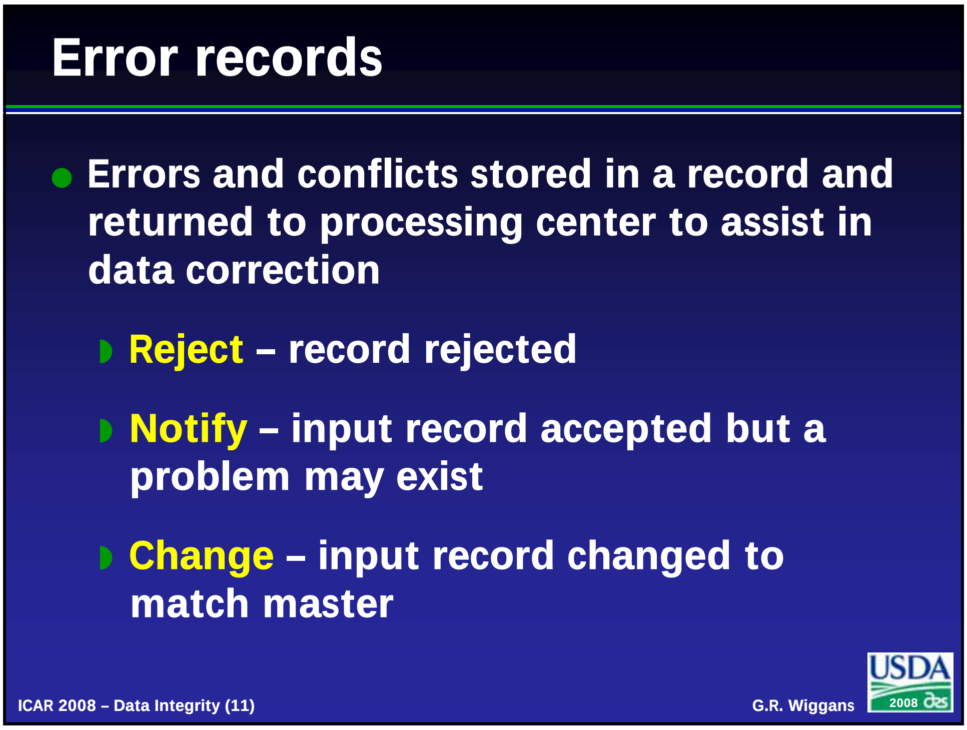


Pedigree and yield edits *(cont.)*

- Calving date
  - Cannot overlap previous lactation
  - Missing calving date may cause breeding to be associated with previous calving

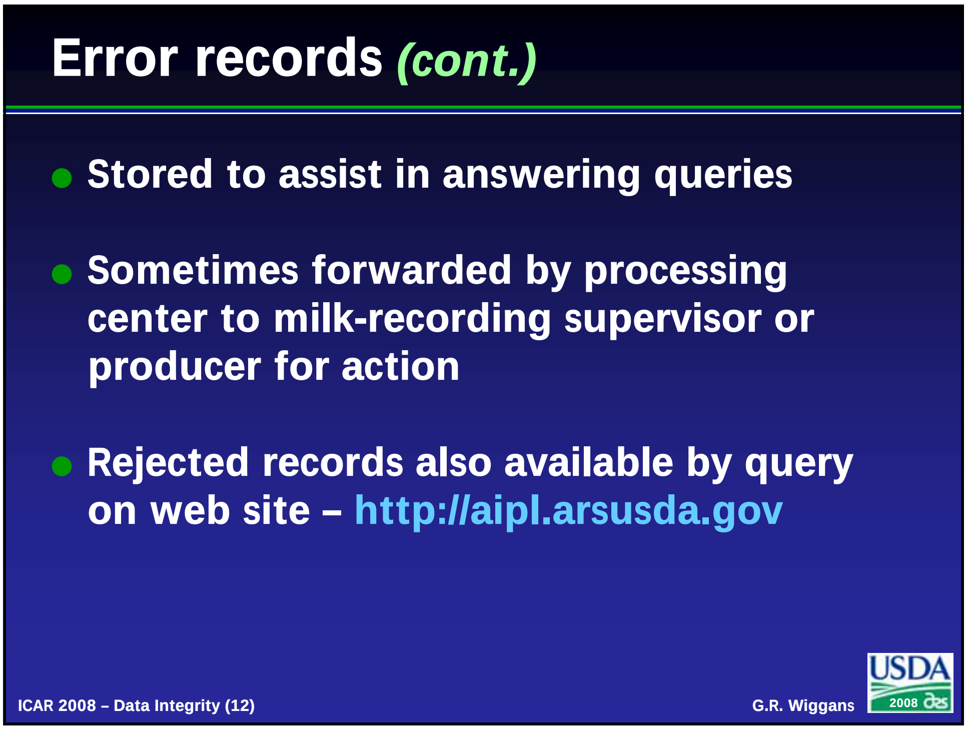
ICAR 2008 – Data Integrity (10)

G.R. Wiggans



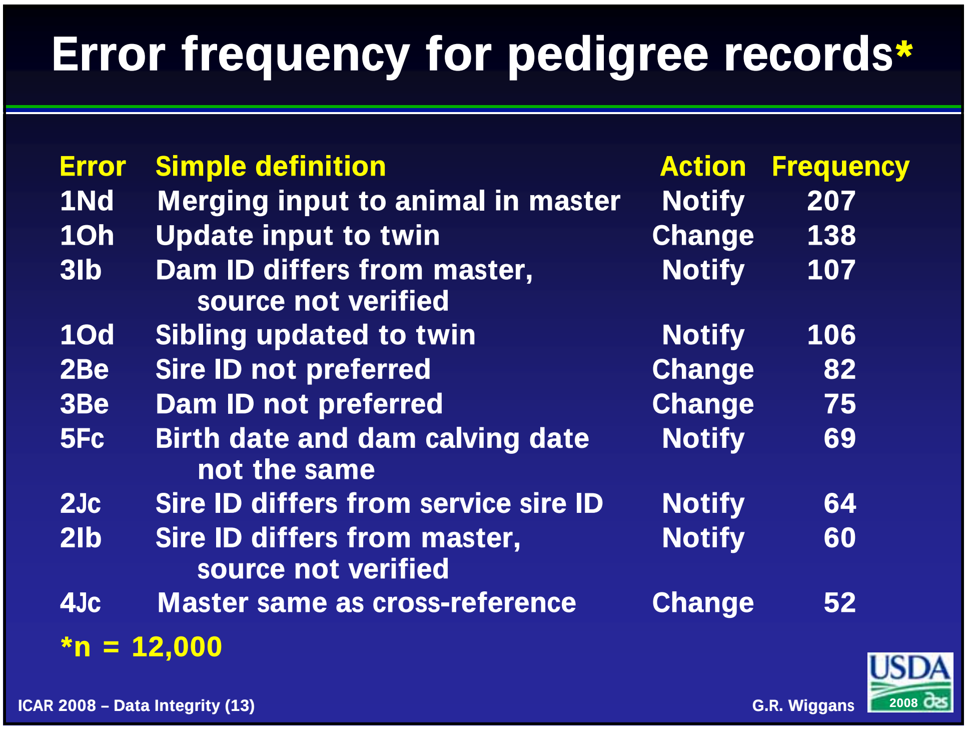
Error records

- Errors and conflicts stored in a record and returned to processing center to assist in data correction
  - Reject – record rejected
  - Notify – input record accepted but a problem may exist
  - Change – input record changed to match master



Error records *(cont.)*

- Stored to assist in answering queries
- Sometimes forwarded by processing center to milk-recording supervisor or producer for action
- Rejected records also available by query on web site – http://aipl.arsusda.gov

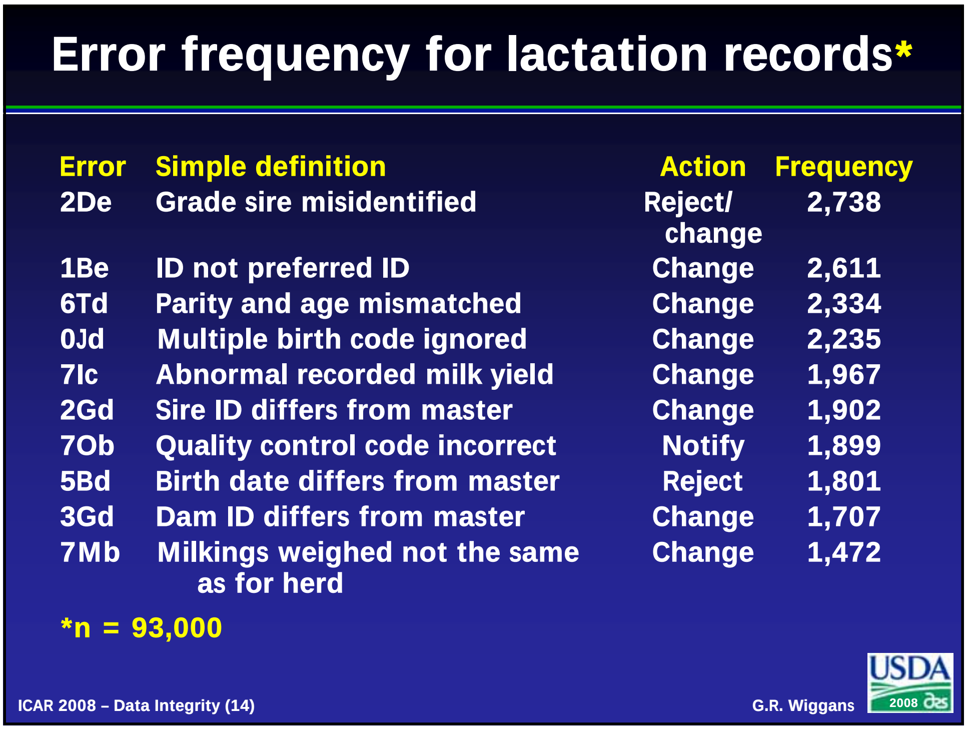


Error frequency for pedigree records*

Error	Simple definition	Action	Frequency
1Nd	Merging input to animal in master	Notify	207
1Oh	Update input to twin	Change	138
3Ib	Dam ID differs from master, source not verified	Notify	107
1Od	Sibling updated to twin	Notify	106
2Be	Sire ID not preferred	Change	82
3Be	Dam ID not preferred	Change	75
5Fc	Birth date and dam calving date not the same	Notify	69
2Jc	Sire ID differs from service sire ID	Notify	64
2Ib	Sire ID differs from master, source not verified	Notify	60
4Jc	Master same as cross-reference	Change	52

*n = 12,000

ICAR 2008 – Data Integrity (13) G.R. Wiggans



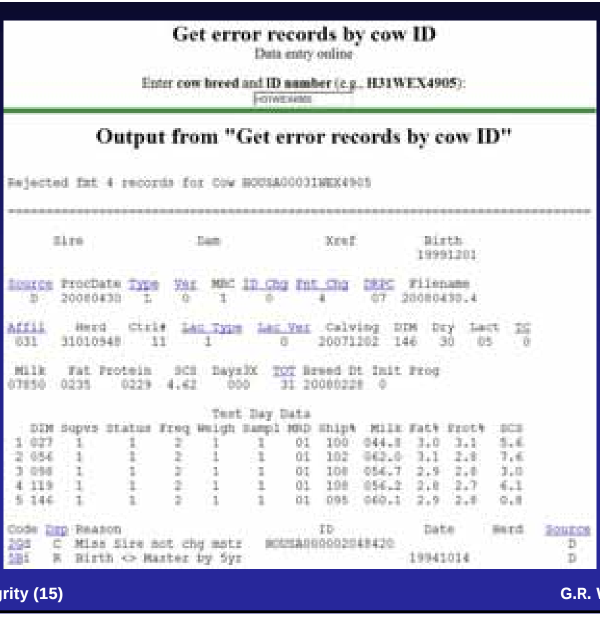
Error frequency for lactation records*

Error	Simple definition	Action	Frequency
2De	Grade sire misidentified	Reject/ change	2,738
1Be	ID not preferred ID	Change	2,611
6Td	Parity and age mismatched	Change	2,334
0Jd	Multiple birth code ignored	Change	2,235
7Ic	Abnormal recorded milk yield	Change	1,967
2Gd	Sire ID differs from master	Change	1,902
7Ob	Quality control code incorrect	Notify	1,899
5Bd	Birth date differs from master	Reject	1,801
3Gd	Dam ID differs from master	Change	1,707
7Mb	Milkings weighed not the same as for herd	Change	1,472

*n = 93,000

ICAR 2008 – Data Integrity (14) G.R. Wiggans

Error records query



Get error records by cow ID

Data entry online

Enter cow breed and ID number (e.g., H31WEX4905):

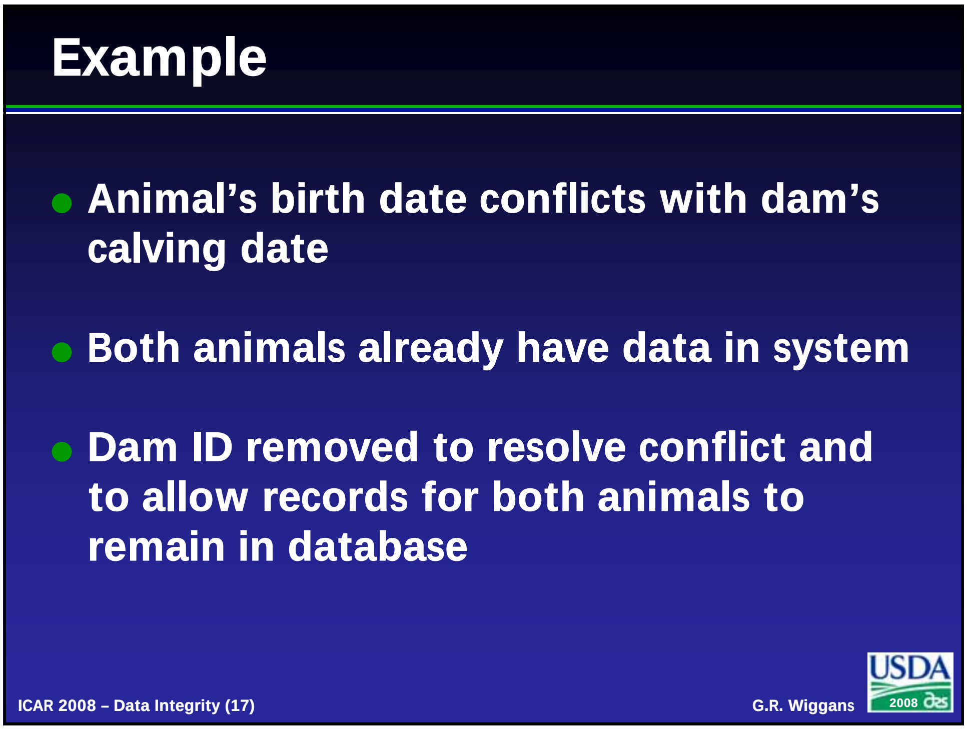
HO1WEX4905

Output from "Get error records by cow ID"

```
Rejected fmt 4 records for Cow HOUSA00031WEX4905

================================================================================

     Sire              Dam                Xref          Birth
                                                        19991201

Source ProcDate Type  Ver  MRC ID Chg Fmt Chg  DRPC  Filename
   D   20080430  L     0    1    0       4       07  20080430.4

Affil   Herd    Ctrl#  Lac Type  Lac Ver  Calving  DIM  Dry  Lact  TC
 031   31010948   11      1          0    20071202  146   30   05    0

 Milk   Fat Protein   SCS  Days3X  TOT Breed Dt Init Prog
07850  0235    0229  4.42    000    31 20080228  0

                           Test Day Data
   DIM Supvs Status Freq Weigh Sampl MRD Ship%  Milk Fat% Prot%  SCS
 1 027   1      1     2     1     1   01  100  044.8  3.0  3.1   5.6
 2 056   1      1     2     1     1   01  102  062.0  3.1  2.8   7.6
 3 098   1      1     2     1     1   01  108  056.7  2.9  2.8   3.0
 4 119   1      1     2     1     1   01  108  056.2  2.8  2.7   4.1
 5 146   1      1     2     1     1   01  095  060.1  2.9  2.8   0.8

Code Dsp Reason                      ID                 Date     Herd   Source
208   C  Miss Sire not chg mstr   HOUSA000002048420                        D
5B1   R  Birth <> Master by Syr                        19941014             D
```

Editing principles

- Data either rejected or modified when errors encountered
- Effect of rejection
  - Loss of possibly valuable information
  - No genetic evaluation for animals of interest
- System designed to retain data whenever possible
- Data elimination preferred to retention of conflicting data

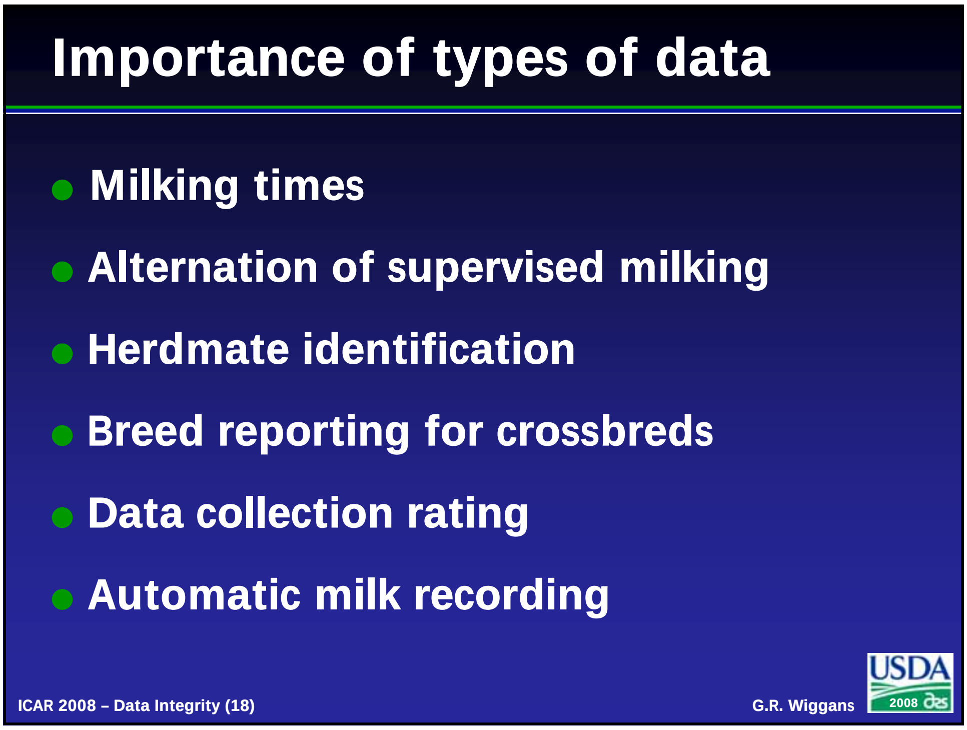


Example

- Animal’s birth date conflicts with dam’s calving date
- Both animals already have data in system
- Dam ID removed to resolve conflict and to allow records for both animals to remain in database

ICAR 2008 – Data Integrity (17) G.R. Wiggans

USDA
2008 aes

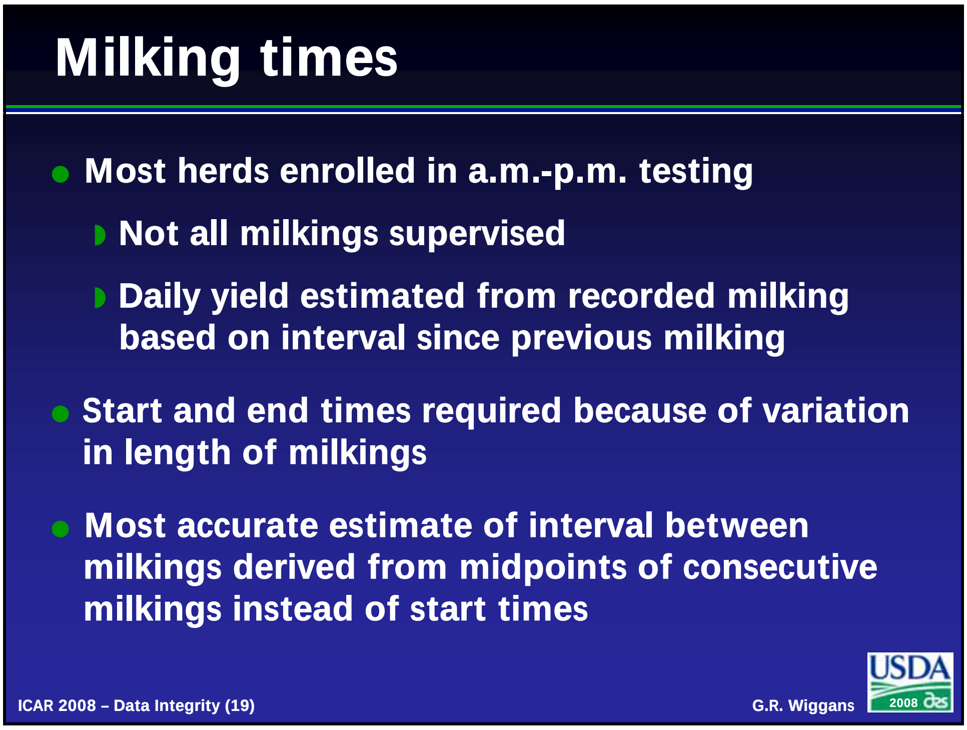


Importance of types of data

- Milking times
- Alternation of supervised milking
- Herdmate identification
- Breed reporting for crossbreds
- Data collection rating
- Automatic milk recording

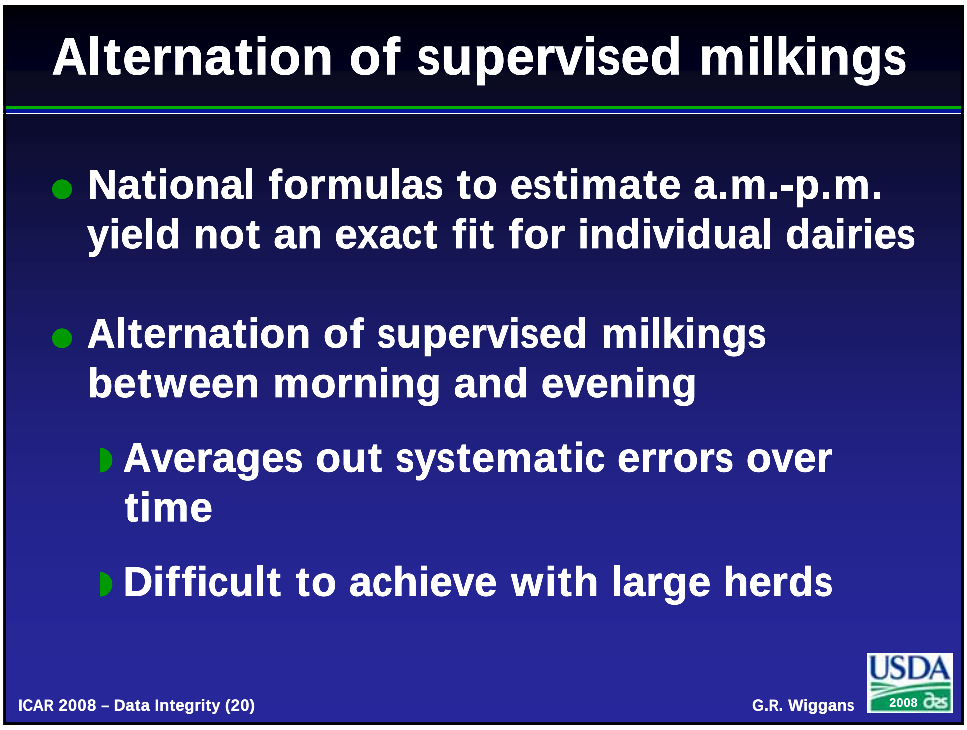
ICAR 2008 – Data Integrity (18) G.R. Wiggans

USDA
2008 aes



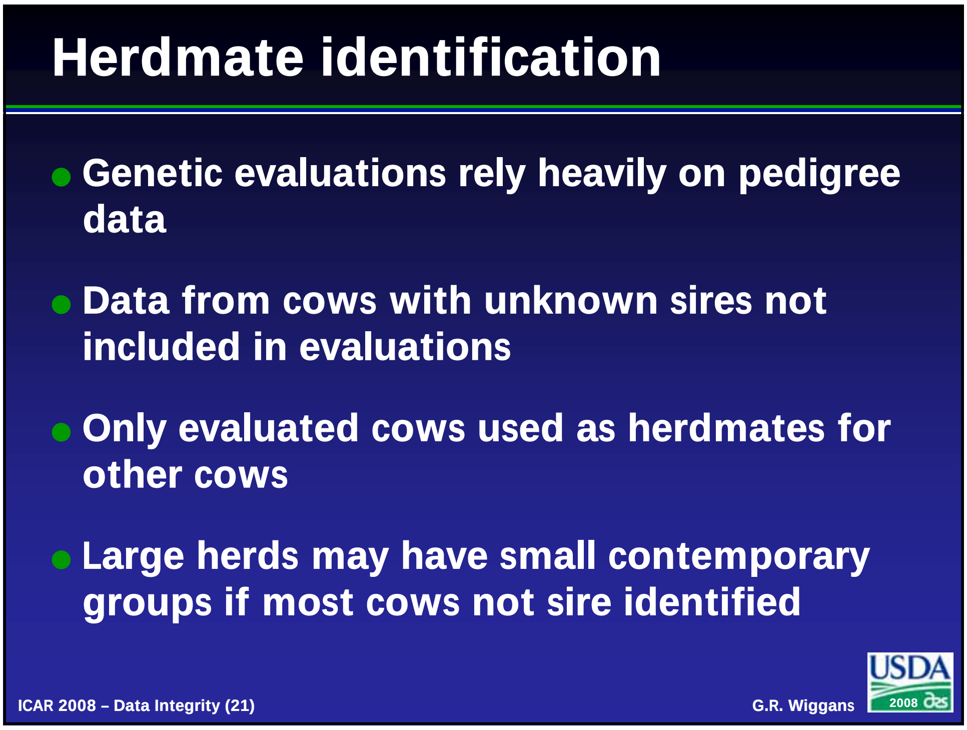
Milking times

- Most herds enrolled in a.m.-p.m. testing
  - Not all milkings supervised
  - Daily yield estimated from recorded milking based on interval since previous milking
- Start and end times required because of variation in length of milkings
- Most accurate estimate of interval between milkings derived from midpoints of consecutive milkings instead of start times



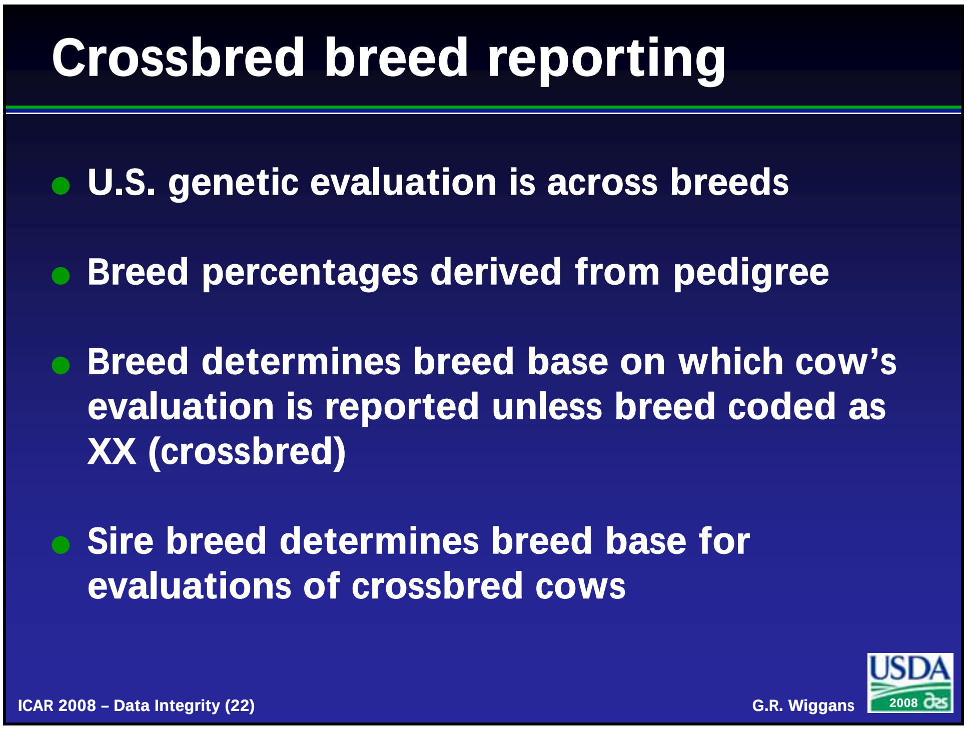
Alternation of supervised milkings

- National formulas to estimate a.m.-p.m. yield not an exact fit for individual dairies
- Alternation of supervised milkings between morning and evening
  - Averages out systematic errors over time
  - Difficult to achieve with large herds



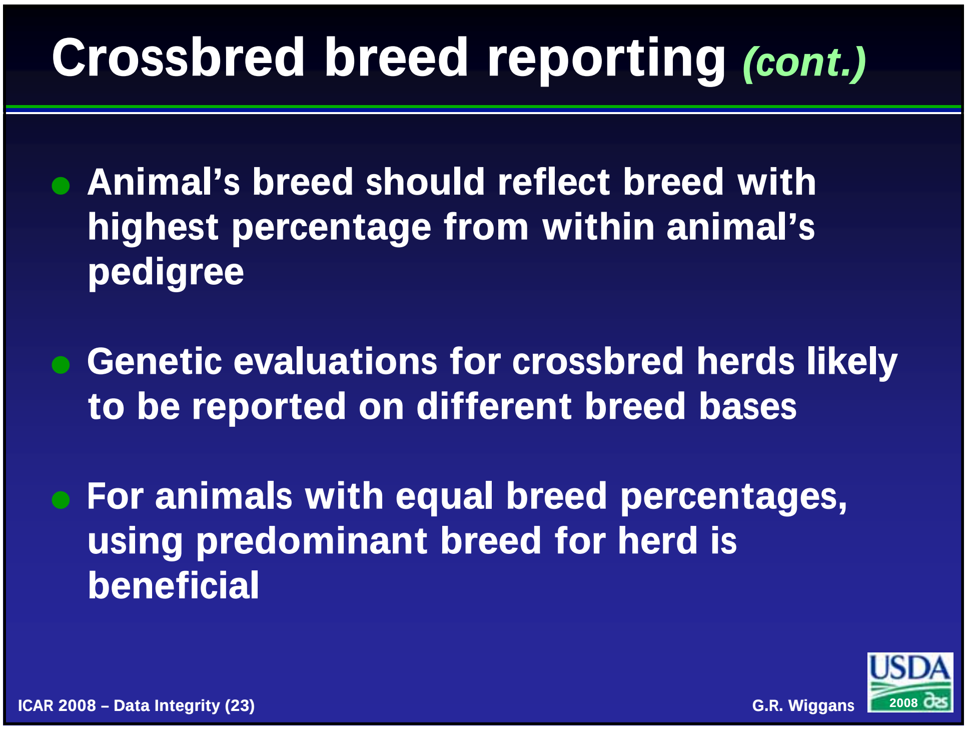
Herdmate identification

- Genetic evaluations rely heavily on pedigree data
- Data from cows with unknown sires not included in evaluations
- Only evaluated cows used as herdmates for other cows
- Large herds may have small contemporary groups if most cows not sire identified



Crossbred breed reporting

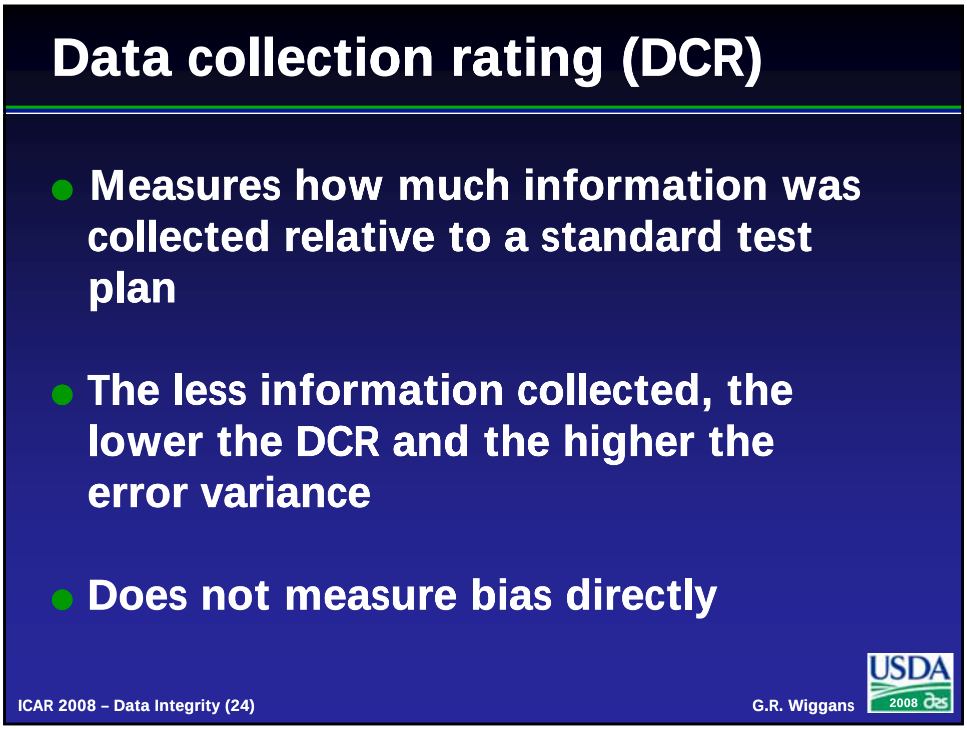
- U.S. genetic evaluation is across breeds
- Breed percentages derived from pedigree
- Breed determines breed base on which cow’s evaluation is reported unless breed coded as XX (crossbred)
- Sire breed determines breed base for evaluations of crossbred cows



Crossbred breed reporting *(cont.)*

- Animal’s breed should reflect breed with highest percentage from within animal’s pedigree
- Genetic evaluations for crossbred herds likely to be reported on different breed bases
- For animals with equal breed percentages, using predominant breed for herd is beneficial

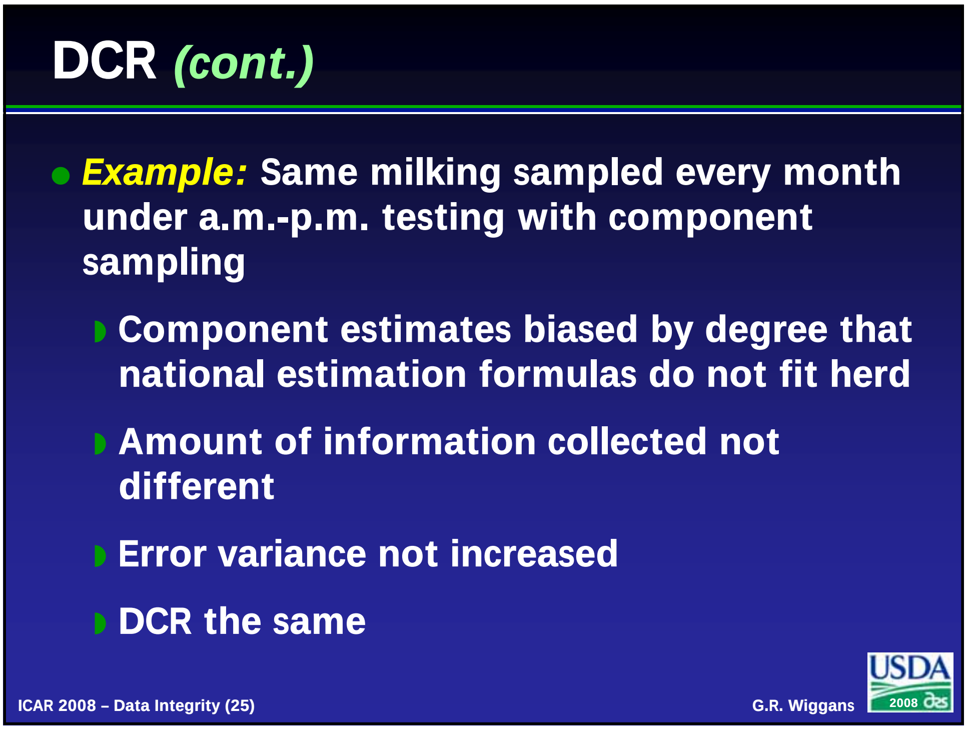
ICAR 2008 – Data Integrity (23) G.R. Wiggans



Data collection rating (DCR)

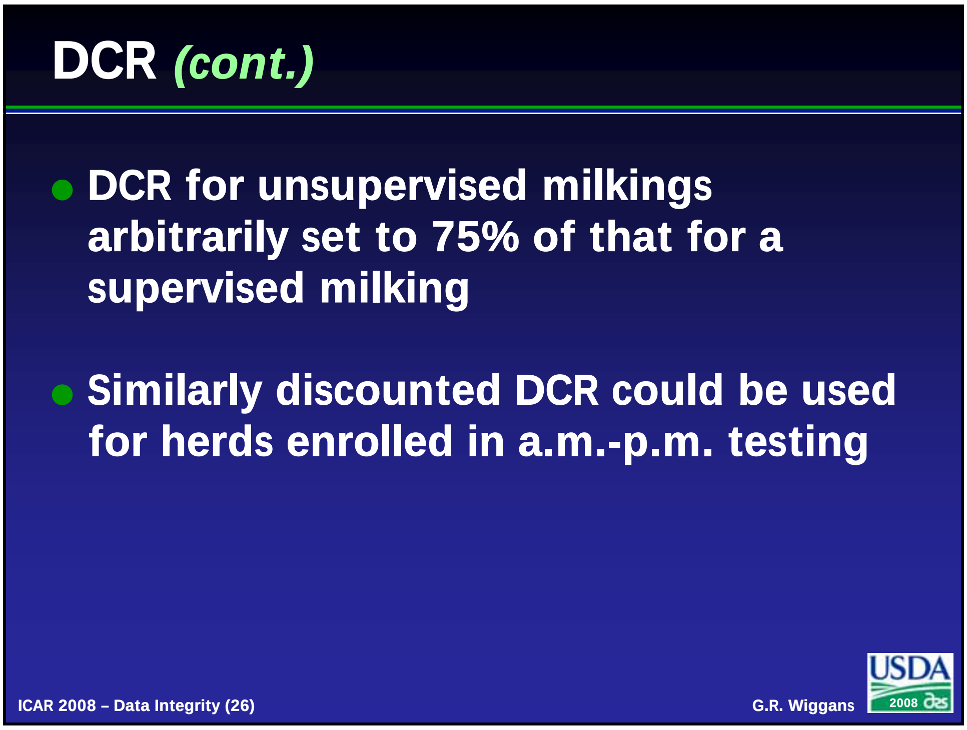
- Measures how much information was collected relative to a standard test plan
- The less information collected, the lower the DCR and the higher the error variance
- Does not measure bias directly

ICAR 2008 – Data Integrity (24) G.R. Wiggans



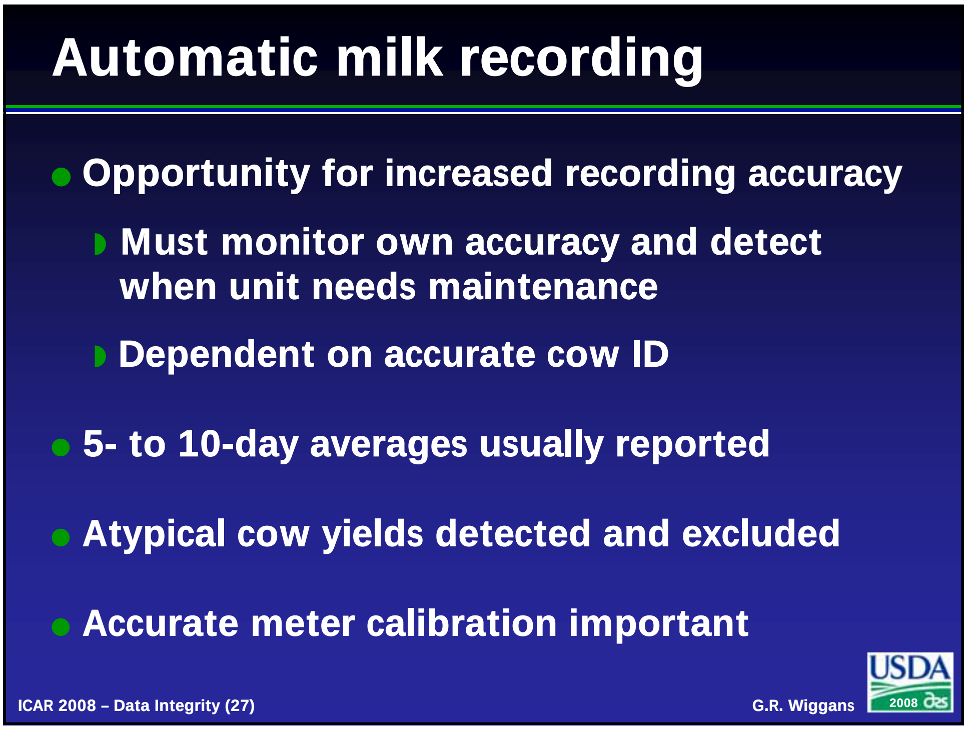
DCR *(cont.)*

- ***Example:*** Same milking sampled every month under a.m.-p.m. testing with component sampling
  - Component estimates biased by degree that national estimation formulas do not fit herd
  - Amount of information collected not different
  - Error variance not increased
  - DCR the same



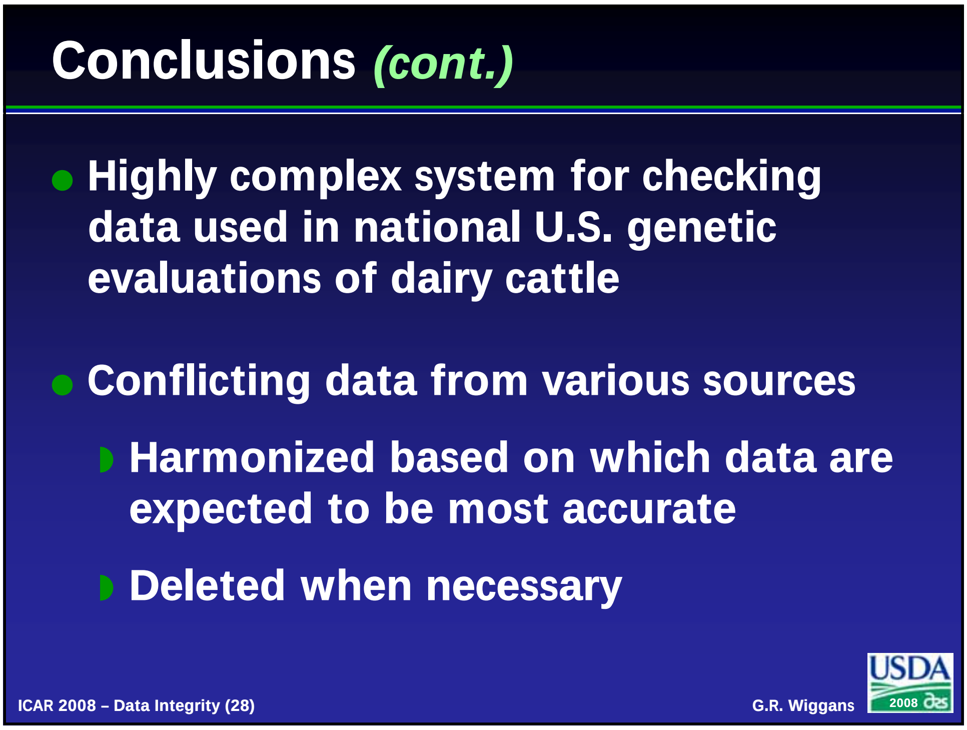
DCR *(cont.)*

- DCR for unsupervised milkings arbitrarily set to 75% of that for a supervised milking
- Similarly discounted DCR could be used for herds enrolled in a.m.-p.m. testing



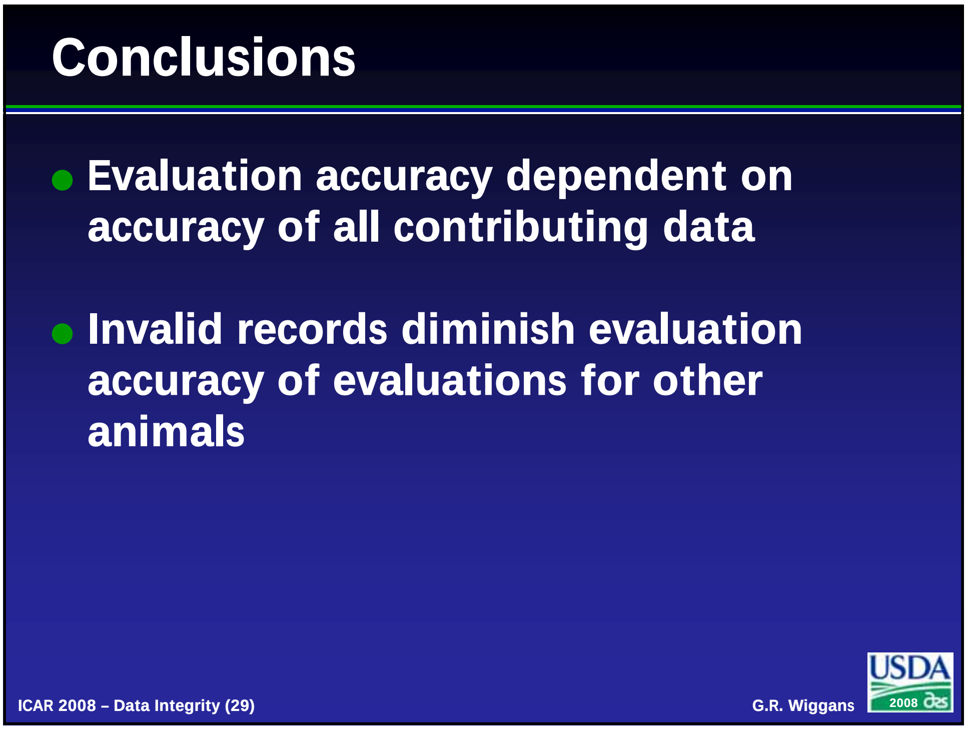
Automatic milk recording

- Opportunity for increased recording accuracy
  - Must monitor own accuracy and detect when unit needs maintenance
  - Dependent on accurate cow ID
- 5- to 10-day averages usually reported
- Atypical cow yields detected and excluded
- Accurate meter calibration important



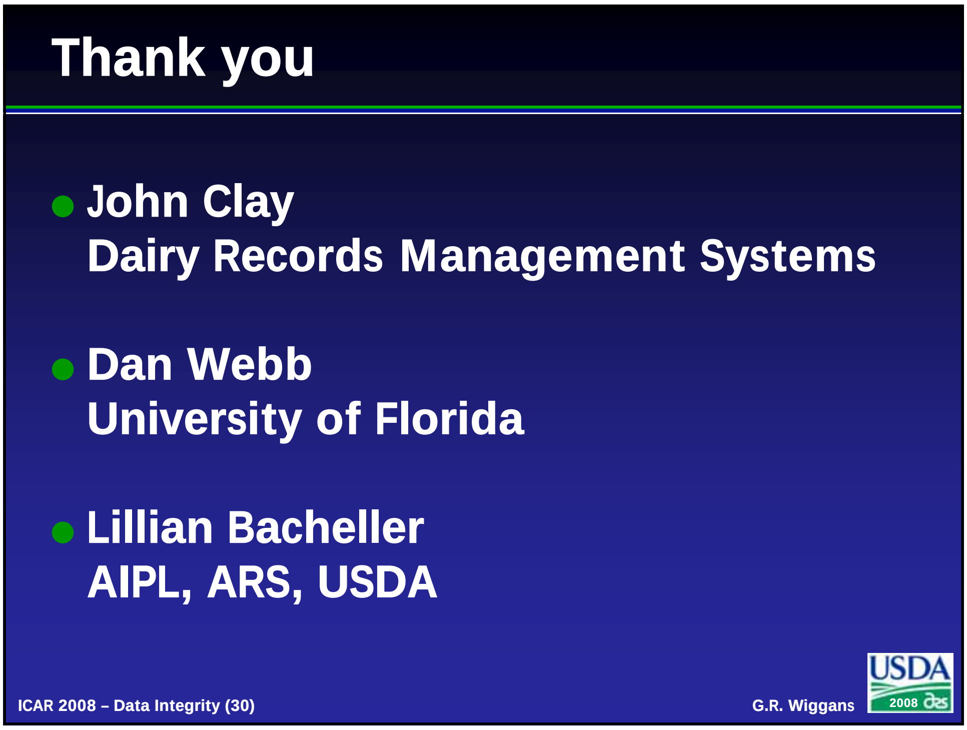
Conclusions *(cont.)*

- Highly complex system for checking data used in national U.S. genetic evaluations of dairy cattle
- Conflicting data from various sources
  - Harmonized based on which data are expected to be most accurate
  - Deleted when necessary



Conclusions

- Evaluation accuracy dependent on accuracy of all contributing data
- Invalid records diminish evaluation accuracy of evaluations for other animals



Thank you

- John Clay
  Dairy Records Management Systems
- Dan Webb
  University of Florida
- Lillian Bacheller
  AIPL, ARS, USDA