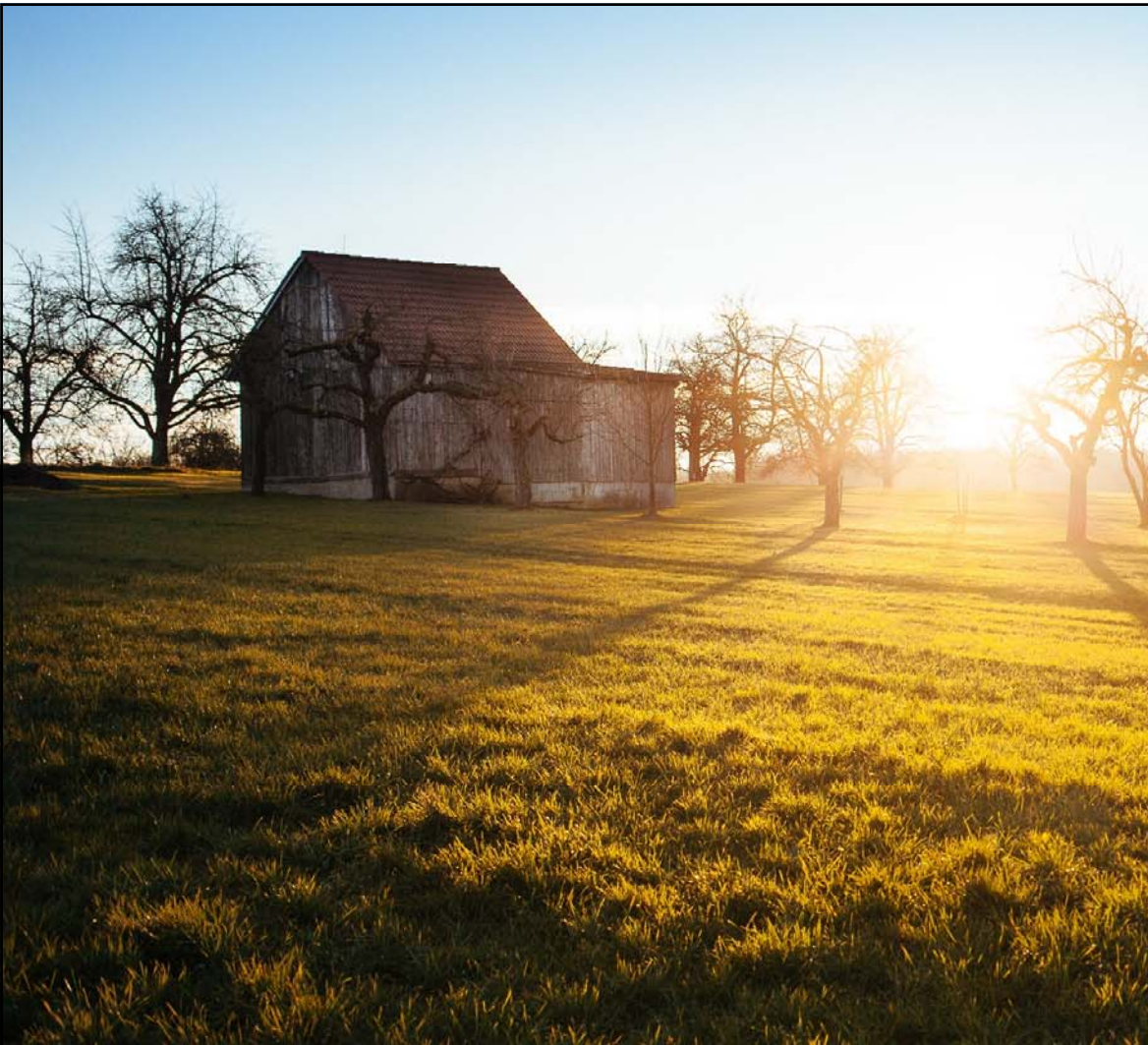


Using Data from Multiple Sources – the Reality of Genetic Evaluations

João Dürr, CDCB CEO
ICAR 2016 – Challenges and Opportunities
Puerto Varas, Chile, October 26, 2016





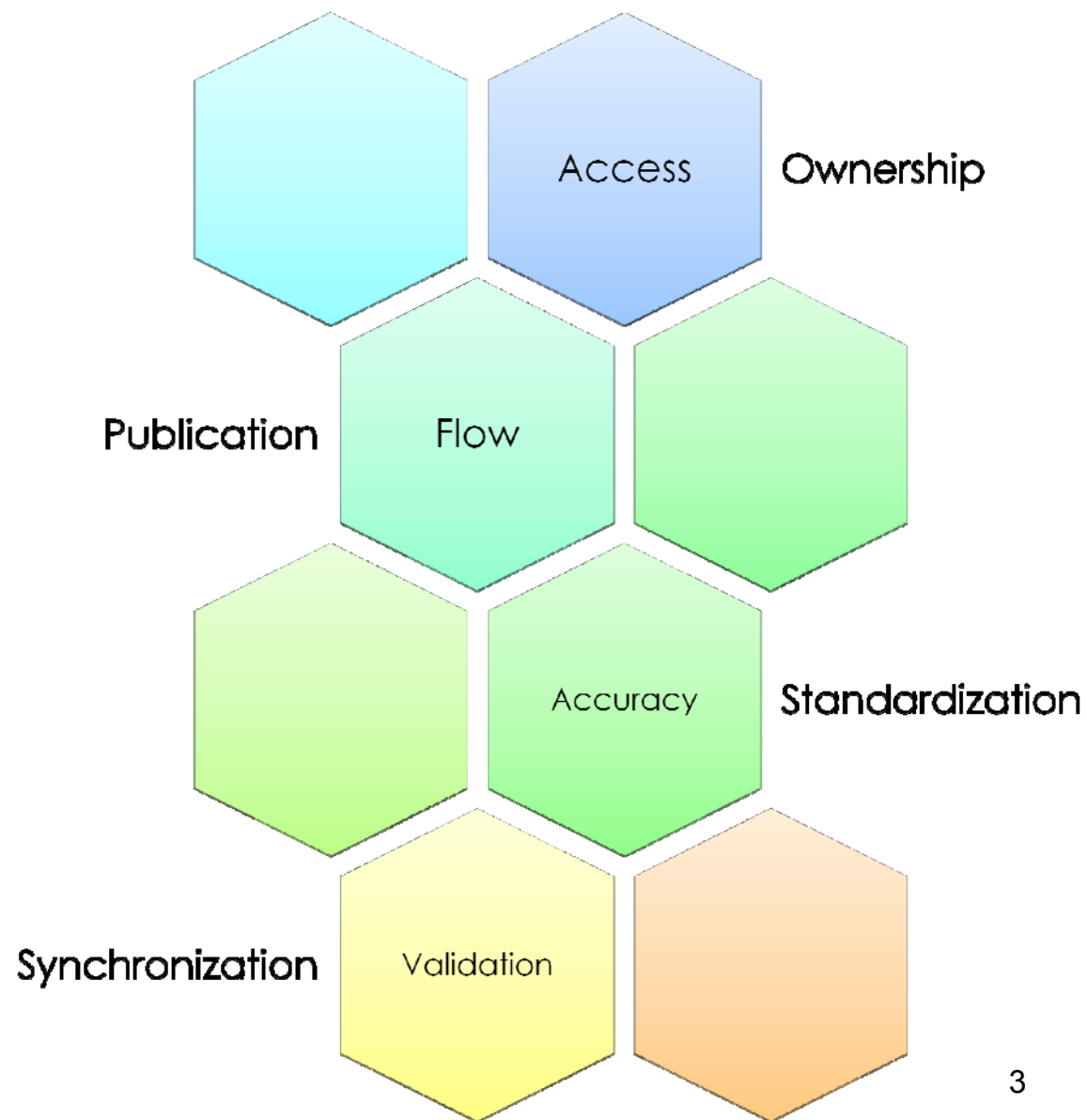
Data management

Case 1: Interbull Centre (2008-2014)

Case 2: CDCB (2014-Present)

Take home messages

Data Policies



Wikipedia: Data validation

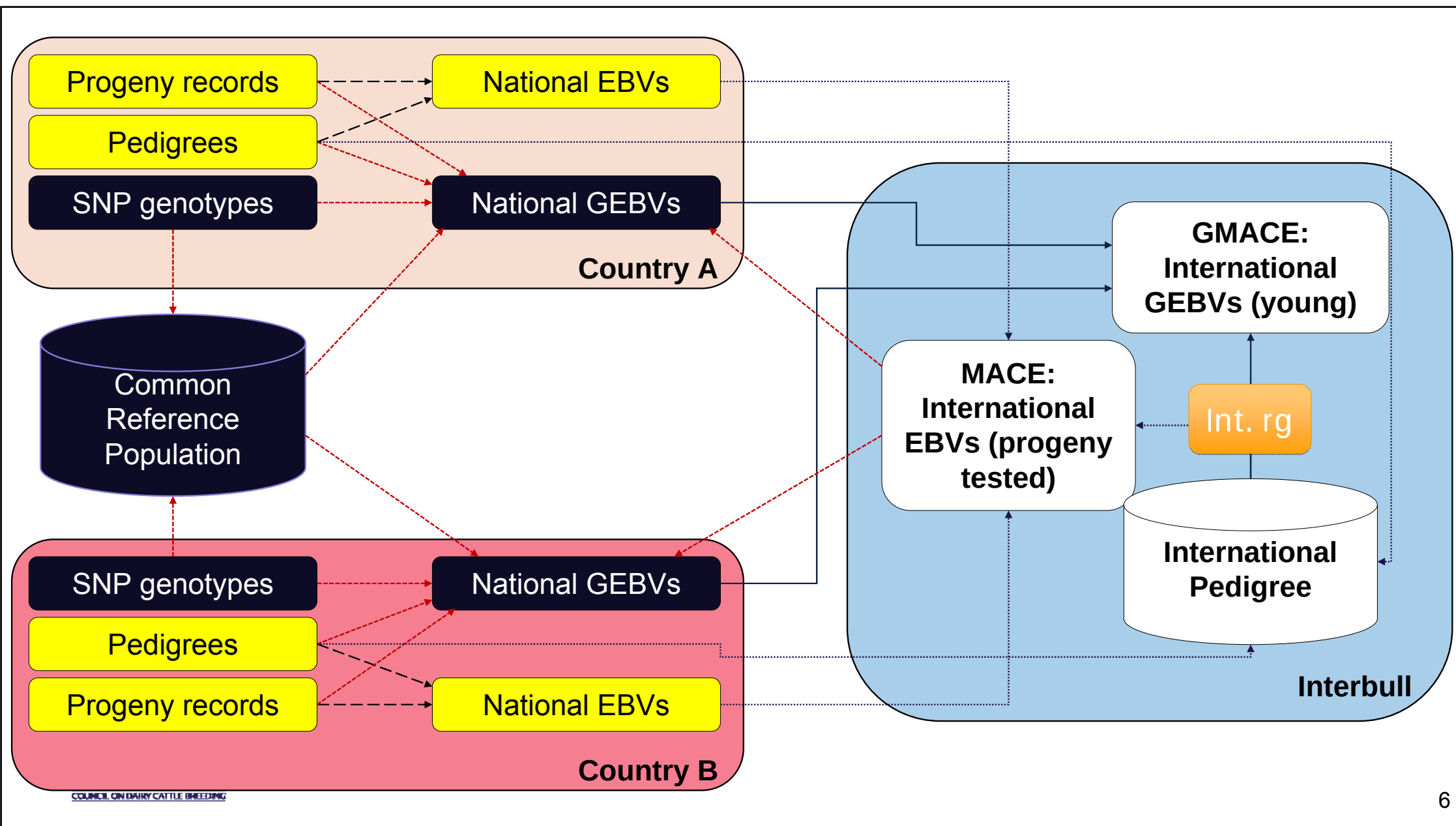
(https://en.wikipedia.org/wiki/Data_validation)

Data validation is the process of ensuring that a program operates on clean, correct and useful data.

It uses routines, often called "validation rules" "validation constraints" or "check routines", that check for correctness, meaningfulness, and security of data that are input to the system.

2008-2014

CASE 1: INTERBULL CENTRE



Features of the Interbull Data Pipeline

- Data suppliers (April 2016)
 - 391 dairy cattle populations, from 34 countries
- Evaluations calendar
 - 3 Annual official evaluations
 - 2 Test runs
 - 5 different national evaluation validation methods
- Data types
 - National genetic merit data (EBV, PTA)
 - 1825 country-breed-trait combinations
 - Pedigrees
 - Population parameters
 - National evaluation validation tests
 - Genotypes (Intergenomics - BSW)

Interbull Centre - 2008 Opportunities

- No database, only flat files
- Each trait group developed separately
 - Independent file formats
 - Duplication - inconsistencies
 - Separate procedures
 - Different edits/checks
 - Separate processing, different levels of automation
 - Analyst-dependent
- Pedigree re-built from scratch every evaluation
- Limited documentation
- Validation of national evaluations not synchronized with users

The joy of developing a database...

Test if you are ready to start developing a DB by answering these very simple questions:

- Why do you need a database?
- ***Which are the business rules?***
- Are those effectively using the DB involved in validating the business rules?
- ***Would a person that knows nothing about your business (the DB developer, for instance) be able to follow the business rules?***
- Have you identified a driver for the project?
- ***Do you have a DB administrator since the beginning of the process?***
- Is your DB Admin happy with the choice of tools?
- ***Is your budget for the project realistic?***

IF YOUR ANSWER FOR ANY OF THE ABOVE IS “NOT SURE”, “NOT YET” OR “ALMOST THERE”

YOU ARE NOT READY TO START!!!

Standardizing data ingestion

- Interbull Centre solution: IDEA
 - Data type and range validation performed locally prior to upload
 - Cross-reference validation performed at the Interbull servers during upload
 - Interactive interface with users to intermediate data acceptance
 - **Golden rule: only data suppliers can modify input data**
- IDEA for pedigrees
 - Principle of “Authoritative Organization”
 - Data flow independent from evaluation deadlines
- IDEA for genetic merit
 - Same file format for all traits
 - “Verify” checks summarized by well established indicators

Interbull Validation of National Evaluation Estimates

- Opportunities
 - Tests applied with subtle differences in implementation yielded different results for users when compared to the Interbull Centre results
 - Much time spent on communication to find out why results were not identical
- Interbull Centre solution
 - Software supplied by the Interbull Centre is run locally
 - Test results and implementation details are recorded
 - Users and the Interbull Centre have access to the same figures

Interbull Centre ISO 9000 Certification

- Write what you do, do what you write
- Good documentation makes your life better
- Comprehensive business rules define your system's credibility
- Version control is much easier when there is only one shared version of the document (Wiki)
- Quality is not an achievement, it is a life style

Lessons from the Case 1

- Databases: be sure you have a plan
- Standardizing data ingestion improves consistency through the use of efficient validation tools
- Keep comprehensive business rules and consistent documentation to stay in business
- Make sure your data suppliers see the same data quality indicators that you see
- Define clear roles and responsibilities between you and your data suppliers

2014 – Present (Discovery phase)

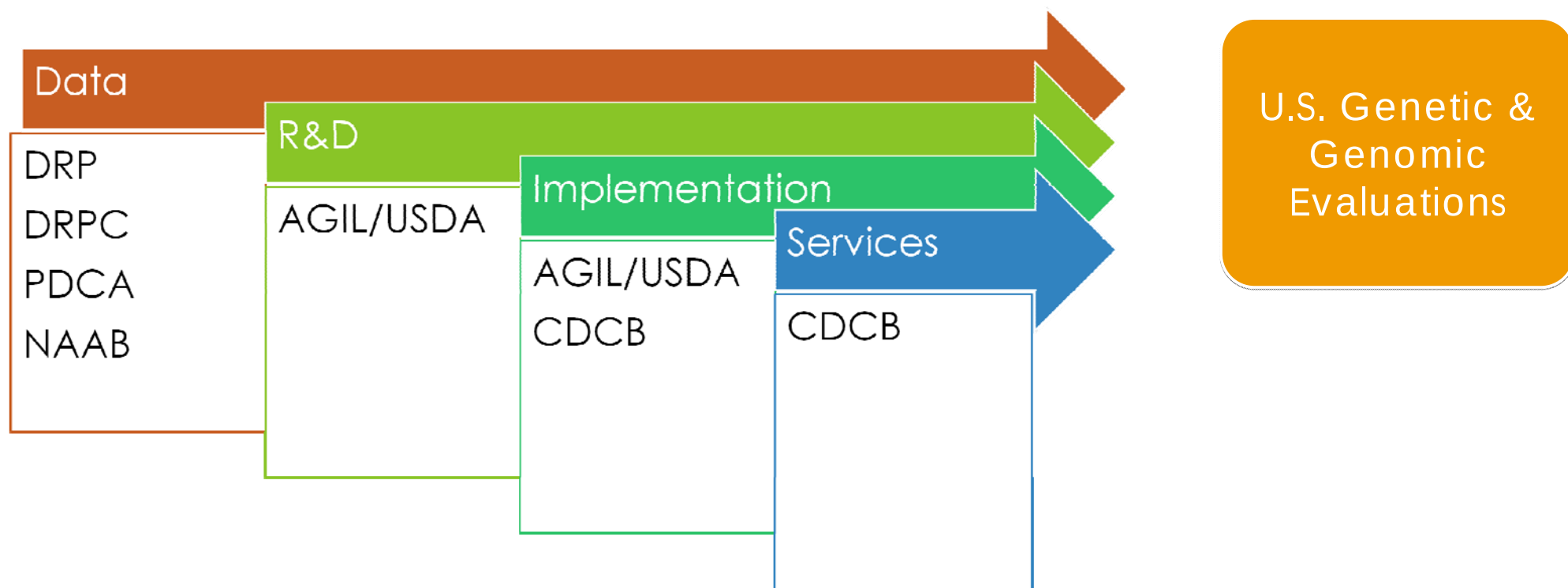
CASE 2: CDCB

Organization



- 12 voting members (3 from each sector)
- 2 nonvoting industry members

US Genetic Evaluation Process



AgSource Cooperative Services
 Arizona DHIA
 Dairy Lab Services
 Dairy One Cooperative Inc.
 DHI Cooperative Inc.
 DHIA West
 Gallenberger Dairy Records
 Heart of America DHIA
 Idaho DHIA
 Indiana State Dairy Association
 Integrated Dairy Herd Improvement
 Jim Sousa Testing
 Lancaster DHIA
 Mid-South Dairy Records
 Minnesota DHIA
 Northstar Cooperative DHI Services
 Puerto Rico DHIA
 Rocky Mountain DHIA
 San Joaquin DHIA
 Southern DHIA Affiliates
 Tennessee DHIA
 Texas DHIA
 Tulare DHIA
 United Federation of DHIA's

Dairy Records Providers (25)

ABS Global. Inc.
 Alta Genetics USA
 American Jersey Cattle Association
 Brown Swiss Cattle Breeders' Association
 Genetic Visions-ST LLC
 Genex Cooperative. Inc.
 Holstein Association USA. Inc.
 Holstein Canada
 National Association of Animal Breeders, Inc.
 Neogen Corporation dba GeneSeek
 New Generation Genetics. Inc.
 Select Sires Inc.
 Semex Alliance
 Tri-State Breeders Cooperative dba Accelerated Genetics
 VHL Genetics
 Zoetis Genetics

Genomic Nominators (16)

American Guernsey Association
 American Jersey Cattle Association
 American Milking Shorthorn Society
 Brown Swiss Cattle Breeders' Association
 Holstein Association USA. Inc.
 Red and White Dairy Cattle Association
 U.S. Arshire Breeders' Association

Purebred Dairy Cattle Association (7)

Agriculture and Horticulture Development Board
 ANAFI
 CDN
 Interbull Centre (34)
 Intergenomics (8)
 Qualitas
 Vit

International Partners (7+)

AgriTech Analytics
 AgSource Cooperative Services
 Dairy Records Management Systems
 DH I-Provo

Dairy Records Processing Centers (4)

Bio-Genesys Ltd.
 Genetic Visions-ST LLC
 Neogen Corporation dba GeneSeek
 VHL Genetics
 Weatherbys Ireland DNA Laboratory
 Zoetis Genetics

Genomic Laboratories (6)

Washington State DHIA





Quality Certification Services Inc.

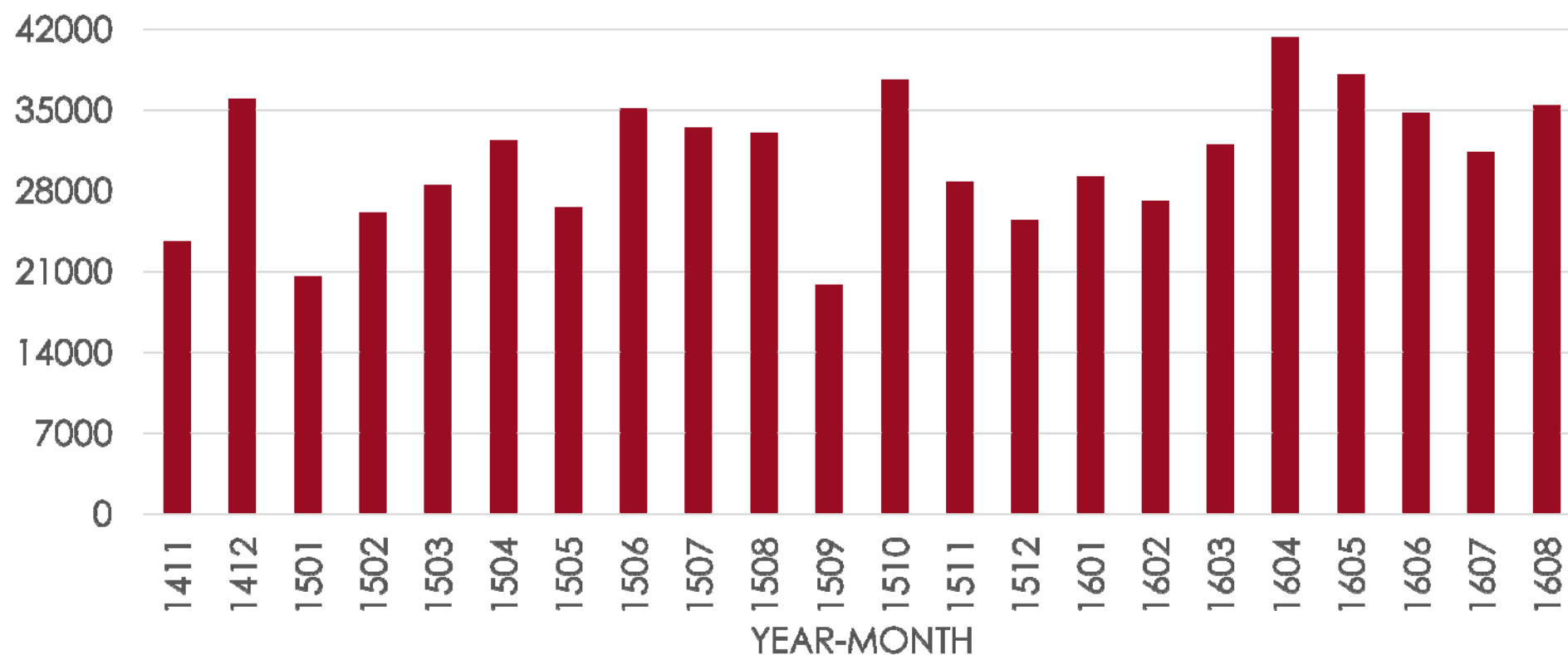
Mission Statement

Providing a reliable source of information to people
interested in the US dairy records industry.

official evaluation runs since December 2015

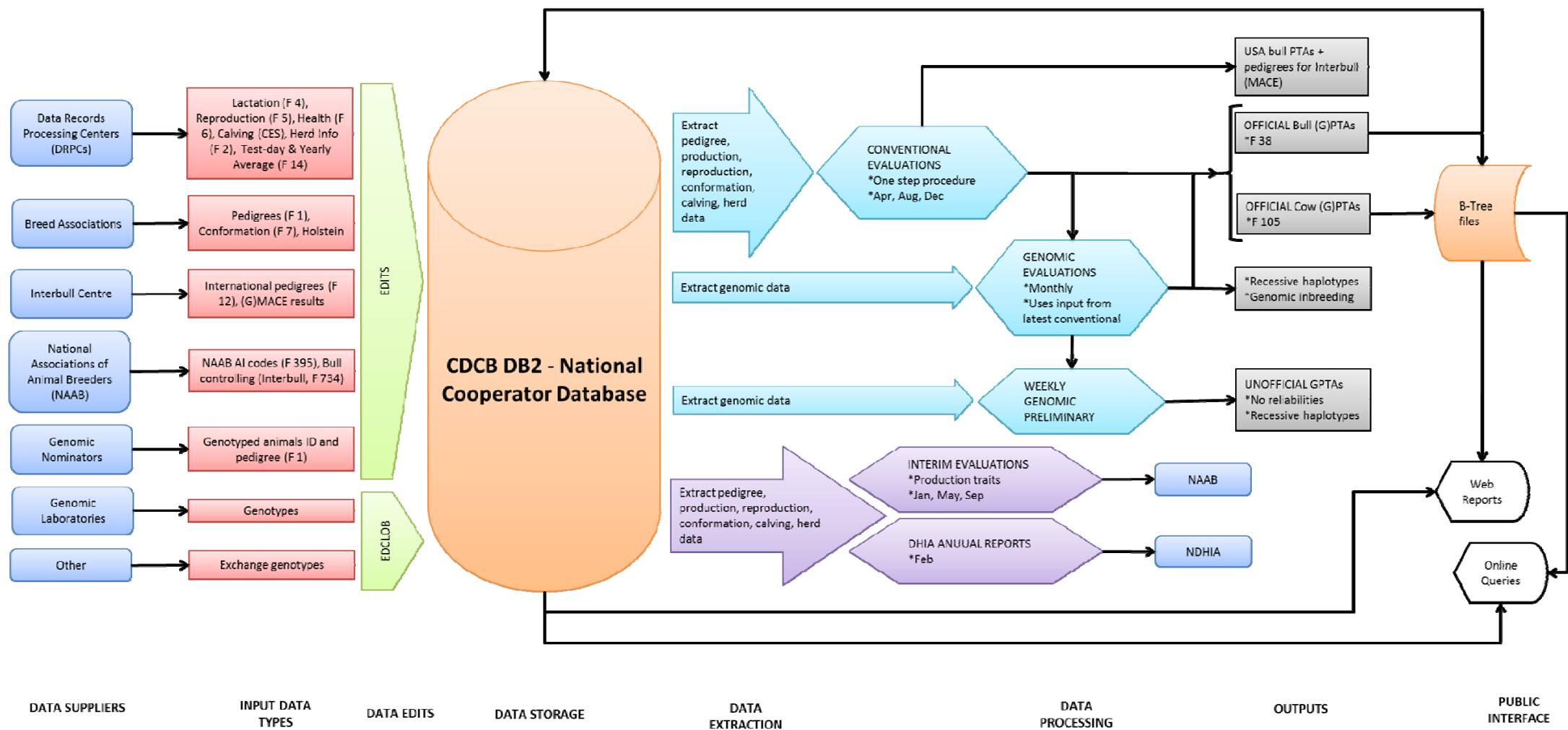
Record type	New records added between December 2015 and April 2016	New records added between April 2016 and August 2016
First lactation test day records	3,012,084	3,061,753
Later lactation test day records	4,578,898	4,752,008
Heifer breeding records	963,249	918,528
Cow breeding records	5,164,212	4,833,899
Calving ease records	401,247	458,785
Stillbirth records	332,704	381,462

Number of genotypes received by CDCB



Number of genotypes stored in the CDCB database by continent of origin, sex and availability of phenotypic information (September 2016)

Continent	Predictor		Predicted		Total
	Females	Males	Females	Males	
Africa	6	-	374	48	428
Asia	15	1,826	2,101	883	4,825
Eastern Europe	24	425	2,120	591	3,160
West and Central Europe	226	15,250	57,113	45,886	118,475
Latin America	343	2	11,983	752	13,080
North America	324,437	29,240	772,096	133,902	1,259,675
Oceania		439			8,785



Dairy Herd
Improvement
Associations



Pedigree, performance and
management records → CDCB



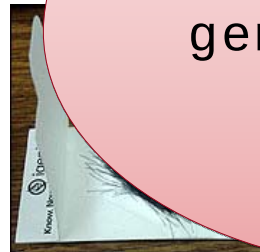
Dairy Records
Processing Centers

Pedigree and
conformation
records →
CDCB



Genomic
Nominators

Pedigree and
genomic records →
CDCB



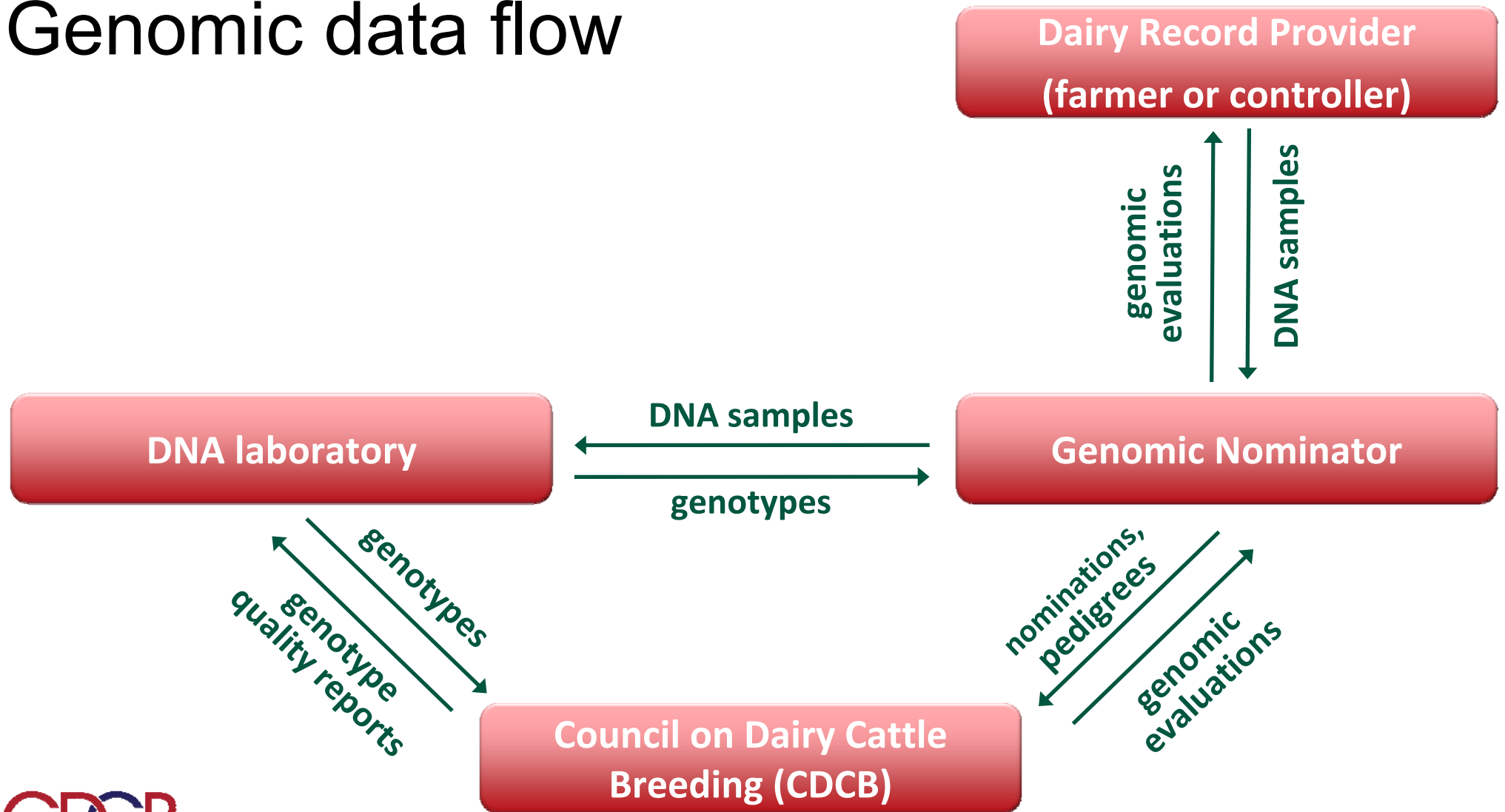
Genomic
Laboratories



Breed Associations



Genomic data flow

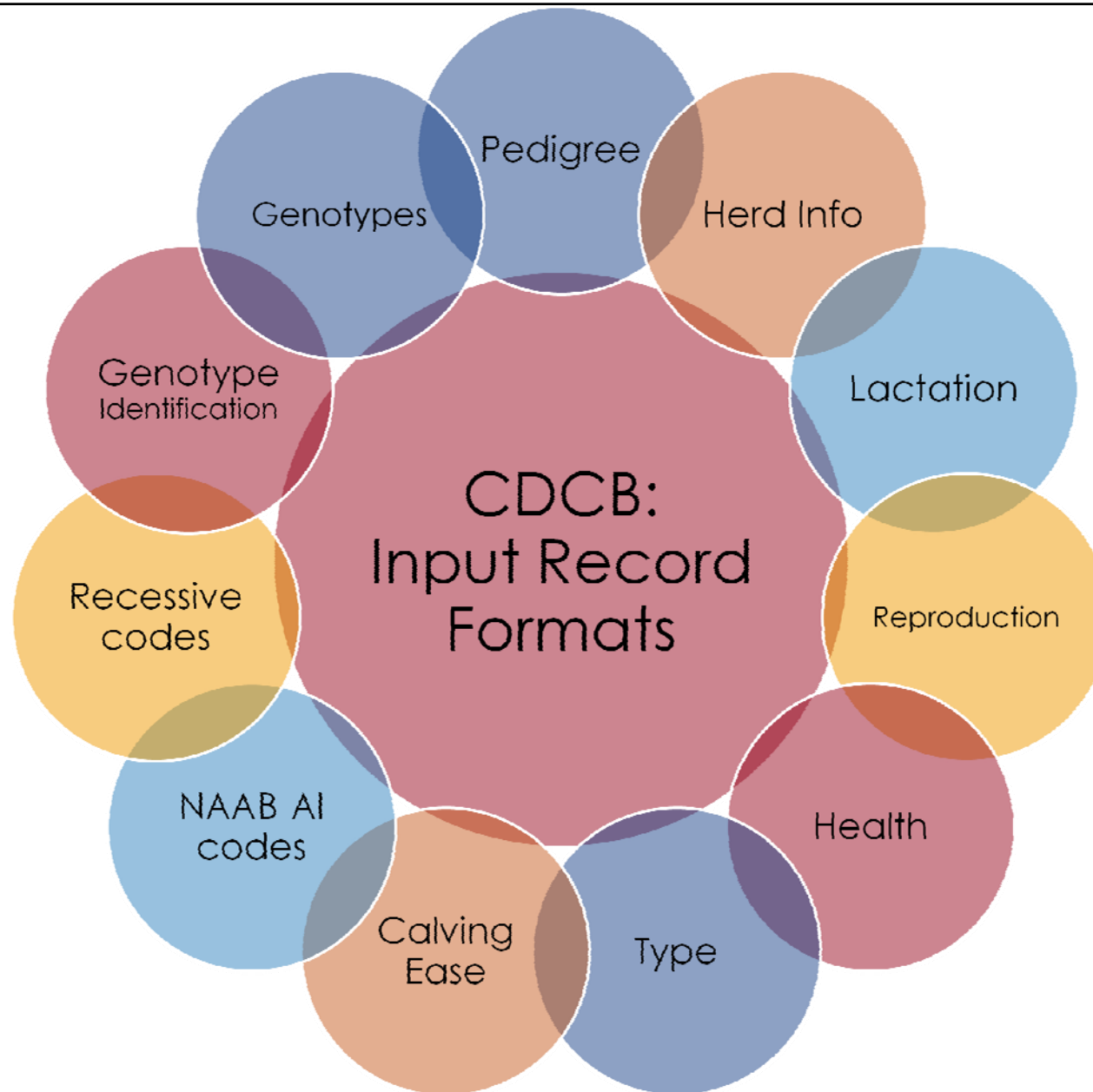


CDCB Fee Schedule

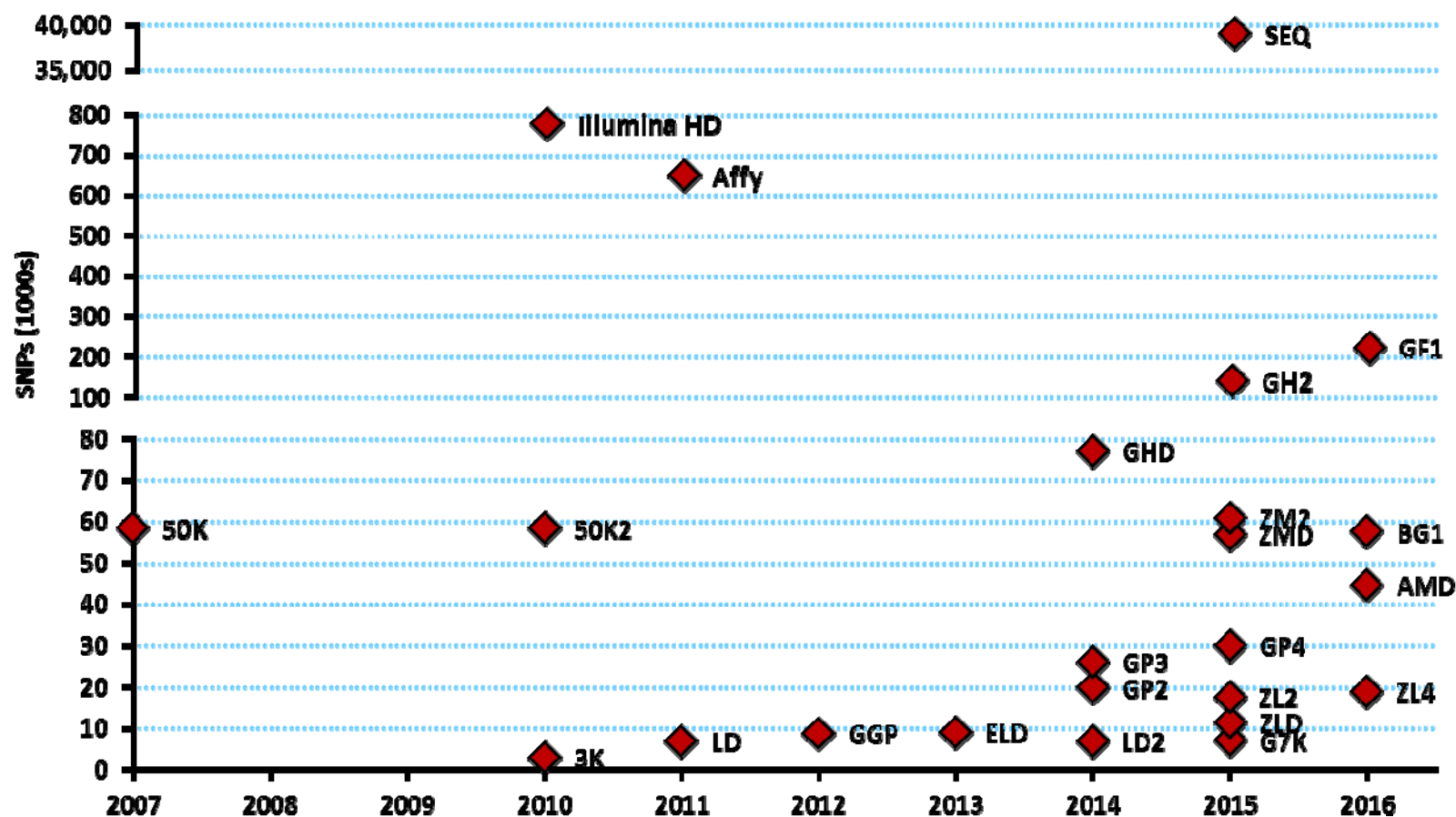
(Updated March 2, 2015)

Rate Code	Participation type	Female fee (\$)	Initial male fee (\$)		AI service fee for males (\$)
1	Total program	0.00	15.00		575.00
2	Member	1.00	22.00		575.00
3	Non-member	3.00	150.00		575.00
			<15 mo	> 15 mo	
4	Canada	6.00	150.00	575.00	575.00
5	Approved partners	7.00	15.00	575.00	575.00
6	All others	7.00	150.00	1200.00	1200.00





Bovine SNP chips processed by the CDCB



Error-Codes for CDCB Data Checks (832)

Error Codes
Complete Error Lists
CSV/Excel
Tab Separated
0 General Record
1 Animal Identification
2 Sire Identification
3 Dam Identification
4 Cross-Reference Identification
5 Birth Date
6 Nontest-Day Production
7 Test Day
8 Reproductive Event
9 Health Event
Genomic Error Documentation

Example: Gender Change Errors

Code	Description	Action	Returned Data	Updated
↑↓	↑↓	↑↓	↑↓	↑↓
OPa	Format 4 can not change gender of animal.	Reject		09/26/2000
OPb	Animal not found under opposite gender. Record type code is changed to 'P'.	Change		11/03/2000
OPc	Change of gender for animal with different master file pedigree.	Reject		09/26/2000
OPd	Change of gender for animal with master file lactations.	Reject		09/26/2000
OPe	Change of gender for animal with master file progeny.	Reject		09/26/2000
OPf	Change of gender for animal with multiple identifications.	Notify	Cross-reference identification and pedigree source	09/26/2000
OPg	Change of gender for animal with homozygous row.	Reject		04/08/2009
OPh	Change of gender for animal with confirmed genotype.	Reject		09/09/2010

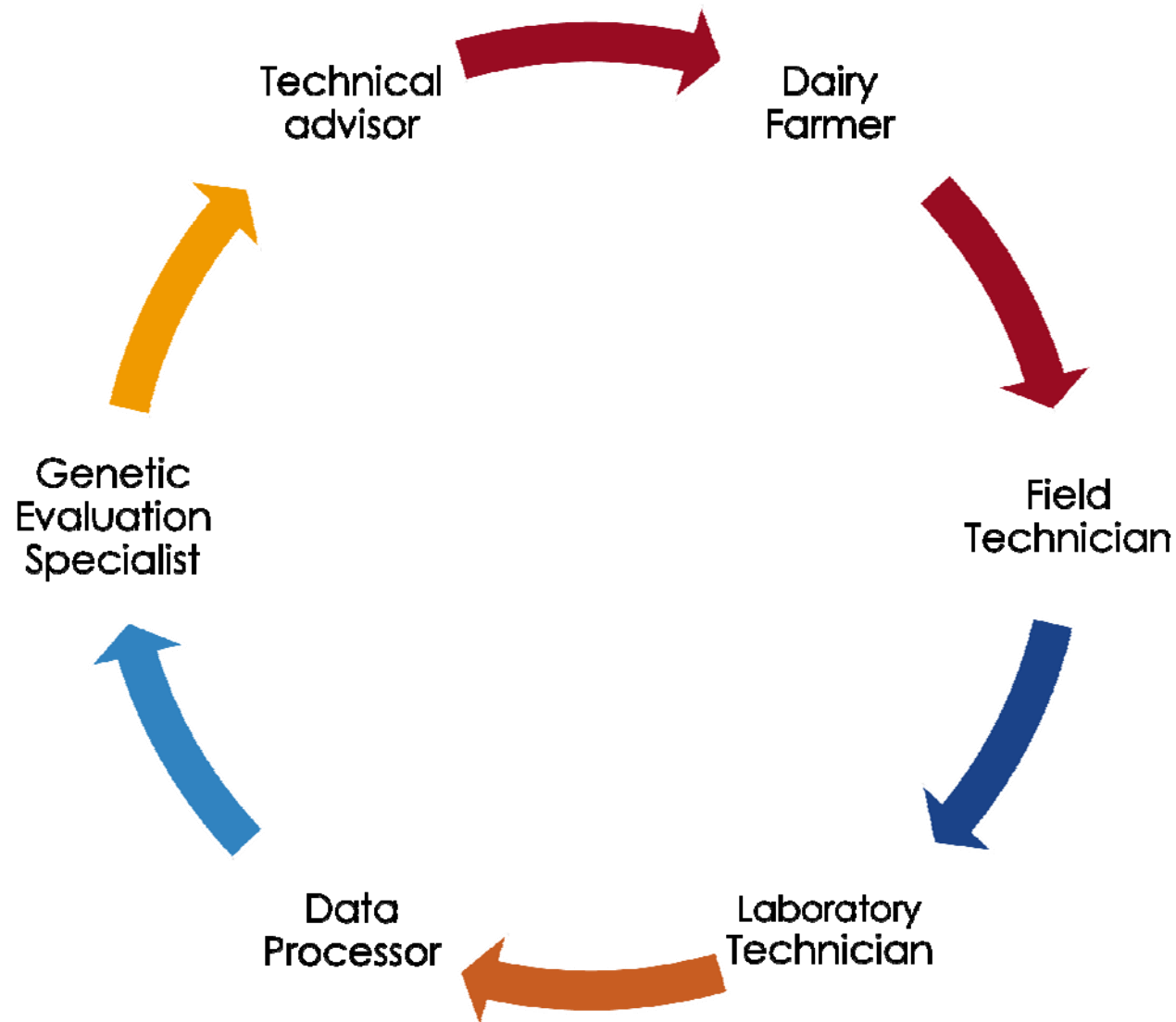
CDCB Evaluation Calendar

- 3 Annual Official Evaluations
 - Conventional
 - Genomic
 - Interbull files
- Monthly Genomic Evaluations
- Weekly Genomic Predictions
- 3 Annual Interim Evaluations

CDCB - Opportunities

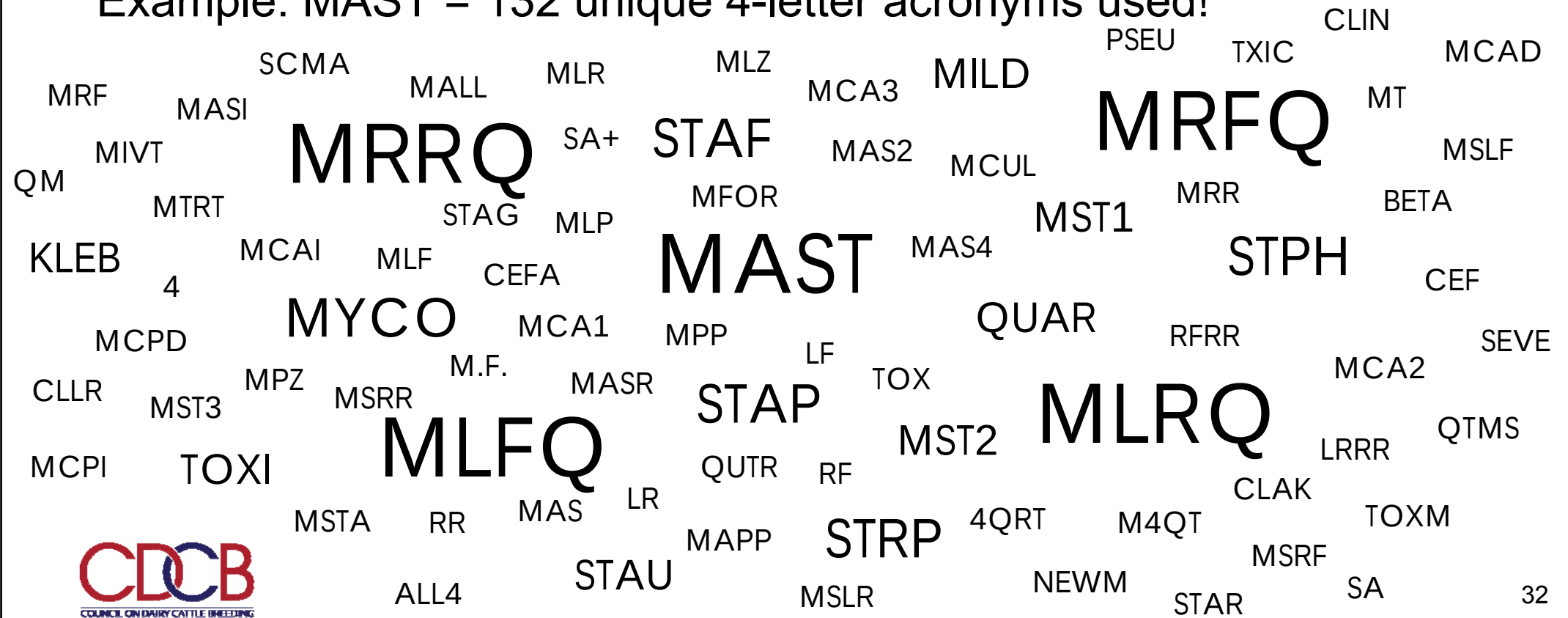
- Transition from USDA to CDCB
 - Recruiting
 - Transfer DB, web applications, directory/files structures, programs
 - Knowledge transfer
 - Roles & responsibilities between AGIL and CDCB
 - Communication
- Multiple file formats
- Web applications developed in several platforms
- Heavy use of SAS in data processing
- Documentation
 - Not consolidated into a unique platform
 - More oriented to operations
 - Limited on business rules

Agents involved in the data pipeline



Standardization of New Data Types

Example: MAST = 132 unique 4-letter acronyms used!



CDCB – First steps

- No changes to the legacy before transition was complete
- Keeping the “old pals” around
- Documenting the legacy
- Strengthening AGIL
- Establishing a policy to compensate phenotypic data suppliers
- Reviewing data access policy
- Developing a new web portal
- Standardizing file formats
- Refining genomic data flow

Lessons from Case 2

- Dairy data awareness has changed the business
 - Control, roles and responsibilities need to be redefined
 - Business rules need to adapt
 - Data access needs to be adjusted
 - Data flow needs to be renegotiated
- Data quality
 - Every link in the chain has to participate
 - Acquiring and validating new data types requires a new mentality

Take Home Message

- Dairy data recording services need to remain relevant for dairy farmers in this fast changing industry.
- Data for genetic evaluations are a by product, not the main goal.
- Making data ingestion more efficient is an effort that involves all agents in the dairy chain.
- Access to dairy data will define the future of dairy genetics.
- Increasing awareness about data quality is the best protection against opportunistic new products in the market.



Acknowledgements:

- Interbull Centre
- AGIL
- CDCB staff
- Data Providers

Thank You!
www.cdcb.us