Information learned from the IGS Genetic and Genomic Evaluation

Dorian Garrick

Department of Animal Science

Iowa State University

dorian@iastate.edu

The current and pre-genomic system

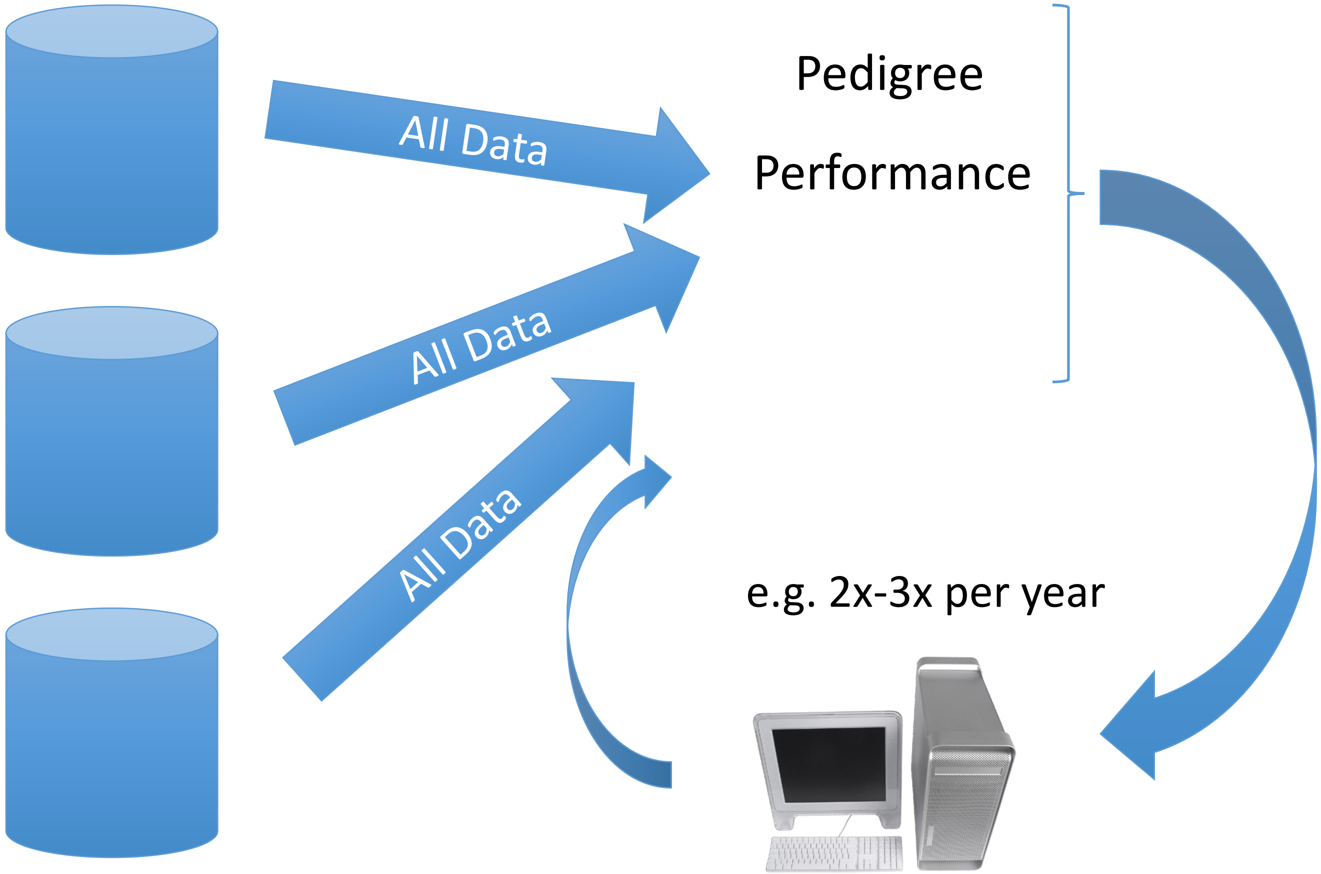
National Cattle Evaluation

- Uses pedigree and performance information to predict the likely outcome of particular matings in terms of progeny performance for particular traits

- Fundamental concept is the Expected Progeny Difference (EPD) for a particular trait or an index of EPD designed to provide balanced improvement over a range of traits

Breed Associations
(IT systems for pedigree
& performance recording)

Merged Data



Pedigree

Performance

e.g. 2x-3x per year

Cornell
Software

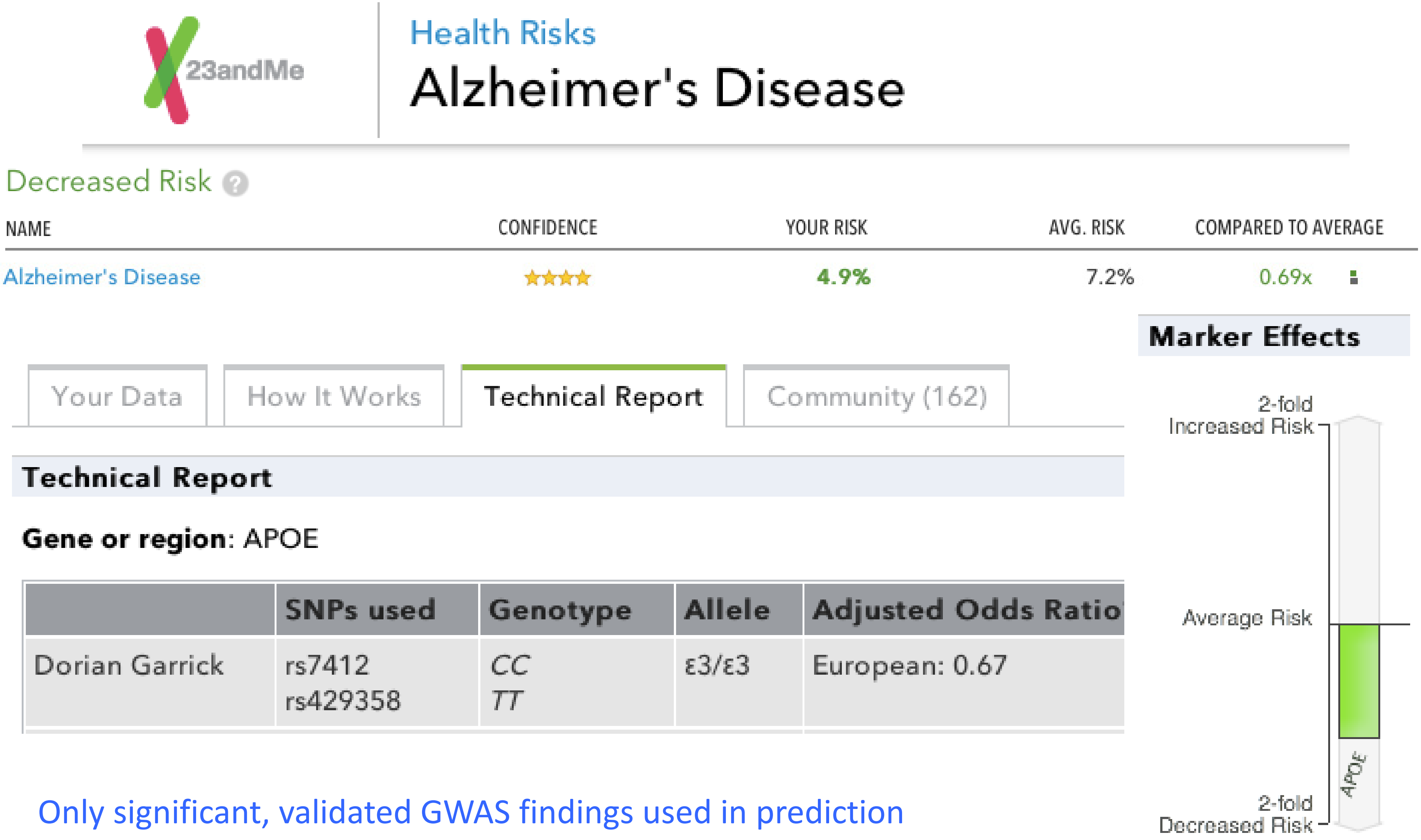
(eg ASA & other IGS partners)

The genomic promise

EPDs are determined by gene effects

- The EPDs for most traits are determined by the collective action of many genes
  - An animal with a favorable EPD must have more gene variants with positive effects than those with negative effects
  - If we could estimate the gene effects we could predict the EPD without pedigree or performance data
    - Or we could combine estimated gene effects with pedigree and performance data to improve the accuracy of EPD for young animals with low accuracy EPD
      - WE cant improve the accuracy of animals with accurate EPD!

www.23andme.com



Health Risks

Alzheimer's Disease

Decreased Risk

NAME	CONFIDENCE	YOUR RISK	AVG. RISK	COMPARED TO AVERAGE
Alzheimer's Disease	★★★★	4.9%	7.2%	0.69x

Technical Report

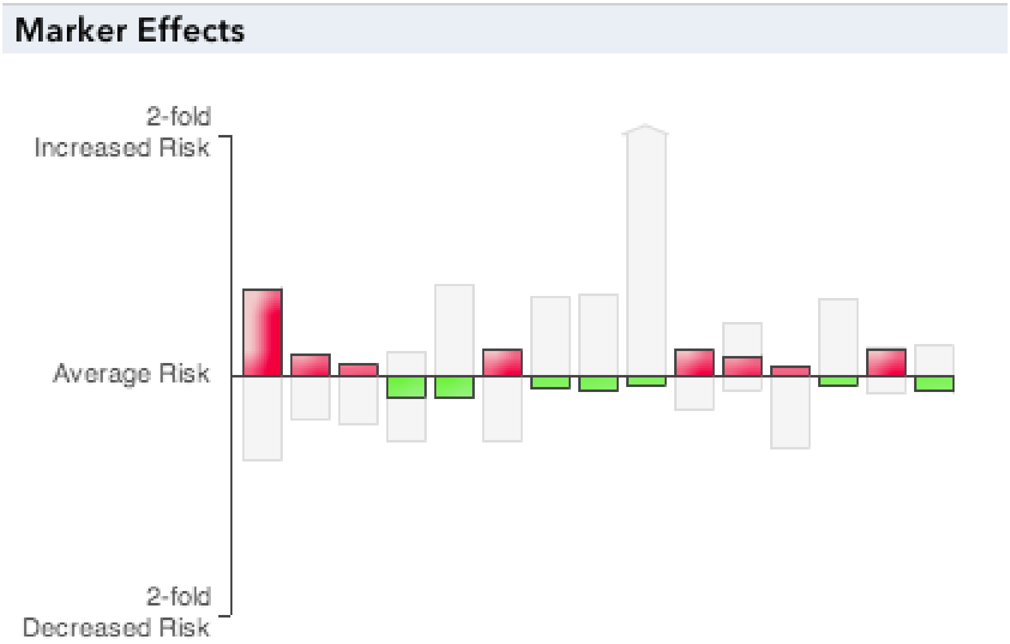
Gene or region: APOE

	SNPs used	Genotype	Allele	Adjusted Odds Ratio
Dorian Garrick	rs7412 rs429358	*CC* *TT*	ε3/ε3	European: 0.67

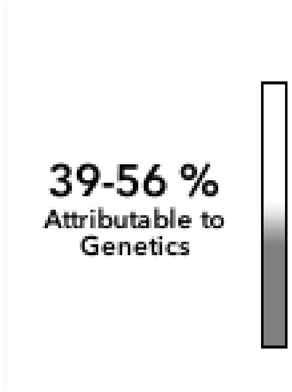
Only significant, validated GWAS findings used in prediction

www.23andme.com

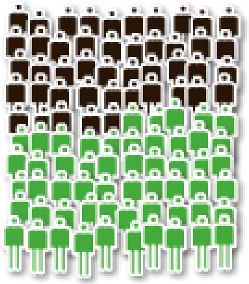
- Coronary Heart Disease



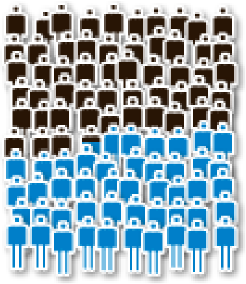
Each bar represents a different risk QTL allele
(mouseover shows the allele and links to the research publications)
QTL=Quantitative Trait Locus



39-56 %
Attributable to Genetics



Dorian Garrick
55.0 out of 100
men of European ethnicity who share Dorian Garrick's genotype will develop Coronary Heart Disease between the ages of 45 and 79.



Average
46.8 out of 100
men of European ethnicity will develop Coronary Heart Disease between the ages of 45 and 79.

Only significant, validated GWAS findings used in prediction

Including genomic information

- Requires collection and storage of genotypes
- Requires new systems and computational approaches for producing EPDs
- Since producers often send samples for genotyping immediately before wanting the results, this necessitates more frequent evaluations

Vision for a turnkey system

just one (authoritative) data system

Data

Ranchers etc

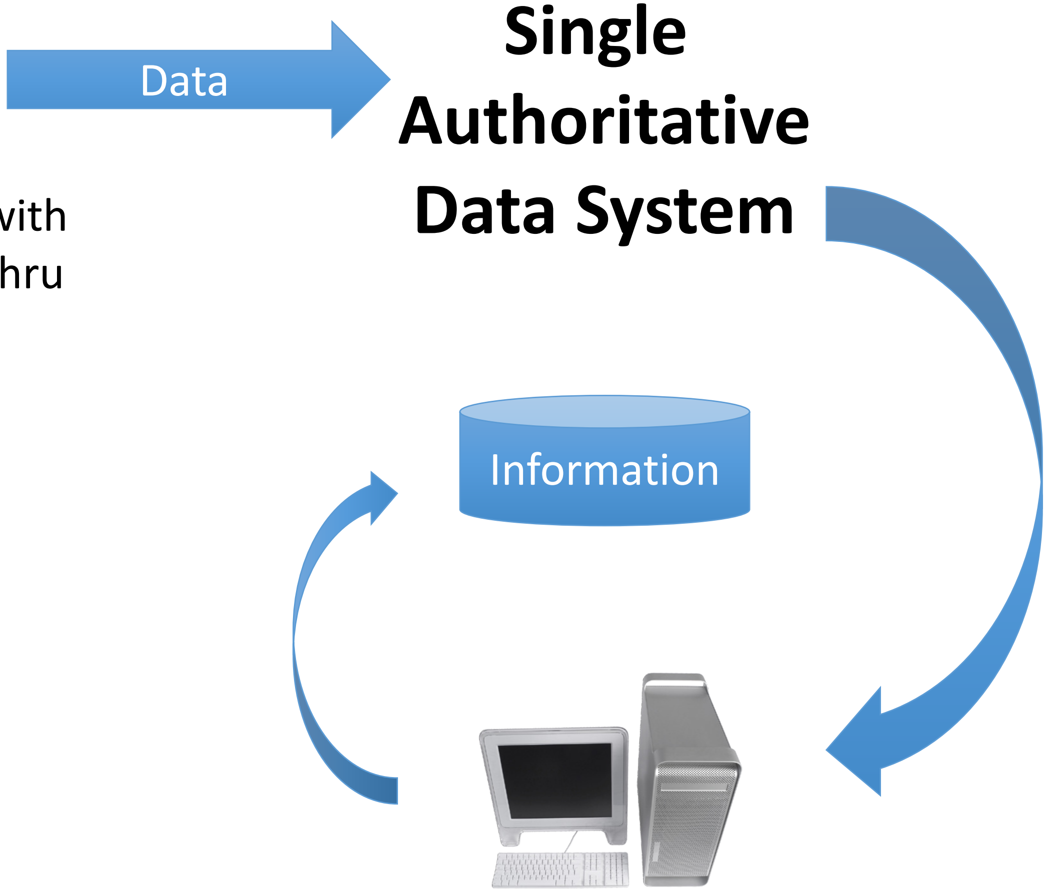
Single Authoritative Data System

Direct interaction with decision support thru

- Cell phones
- Web

BOLT CUDA
Evaluation System

Data ➜ Information



Ranchers etc

Data

Single Authoritative Data System

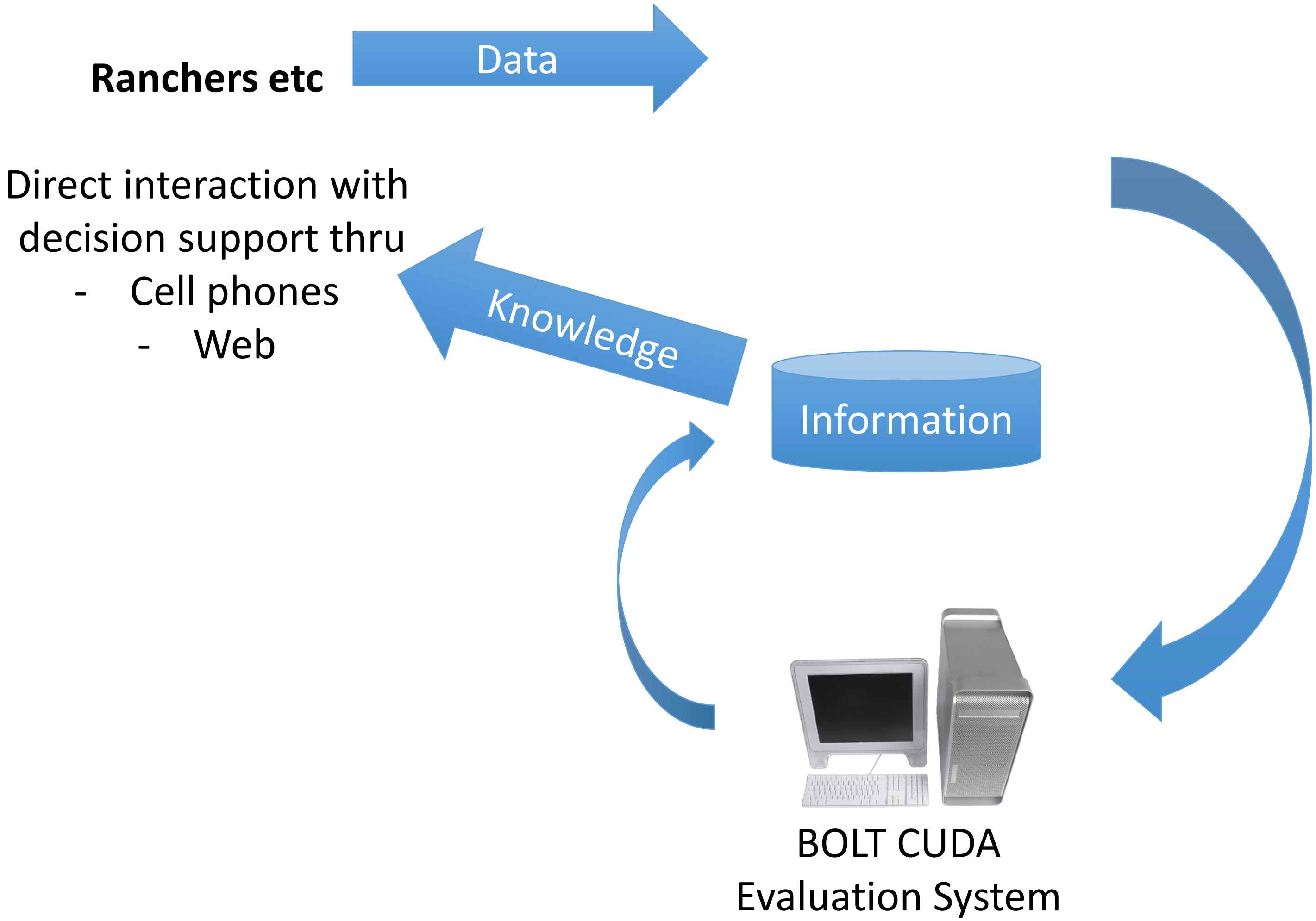
Information

BOLT CUDA
Evaluation System

Direct interaction with decision support thru

- Cell phones
- Web

Data ➔ Information ➔ Knowledge



Ranchers etc

Data

Direct interaction with decision support thru

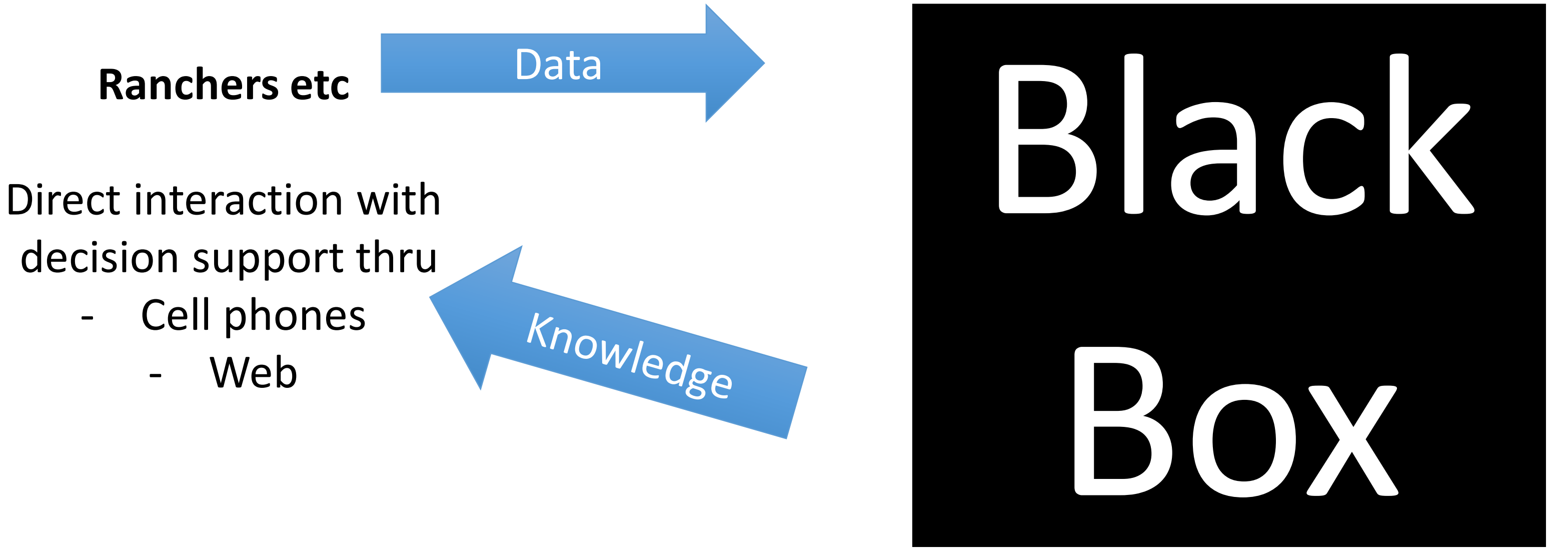
- Cell phones
- Web

Knowledge

Information

BOLT CUDA
Evaluation System

Data ➔ Information ➔ Knowledge

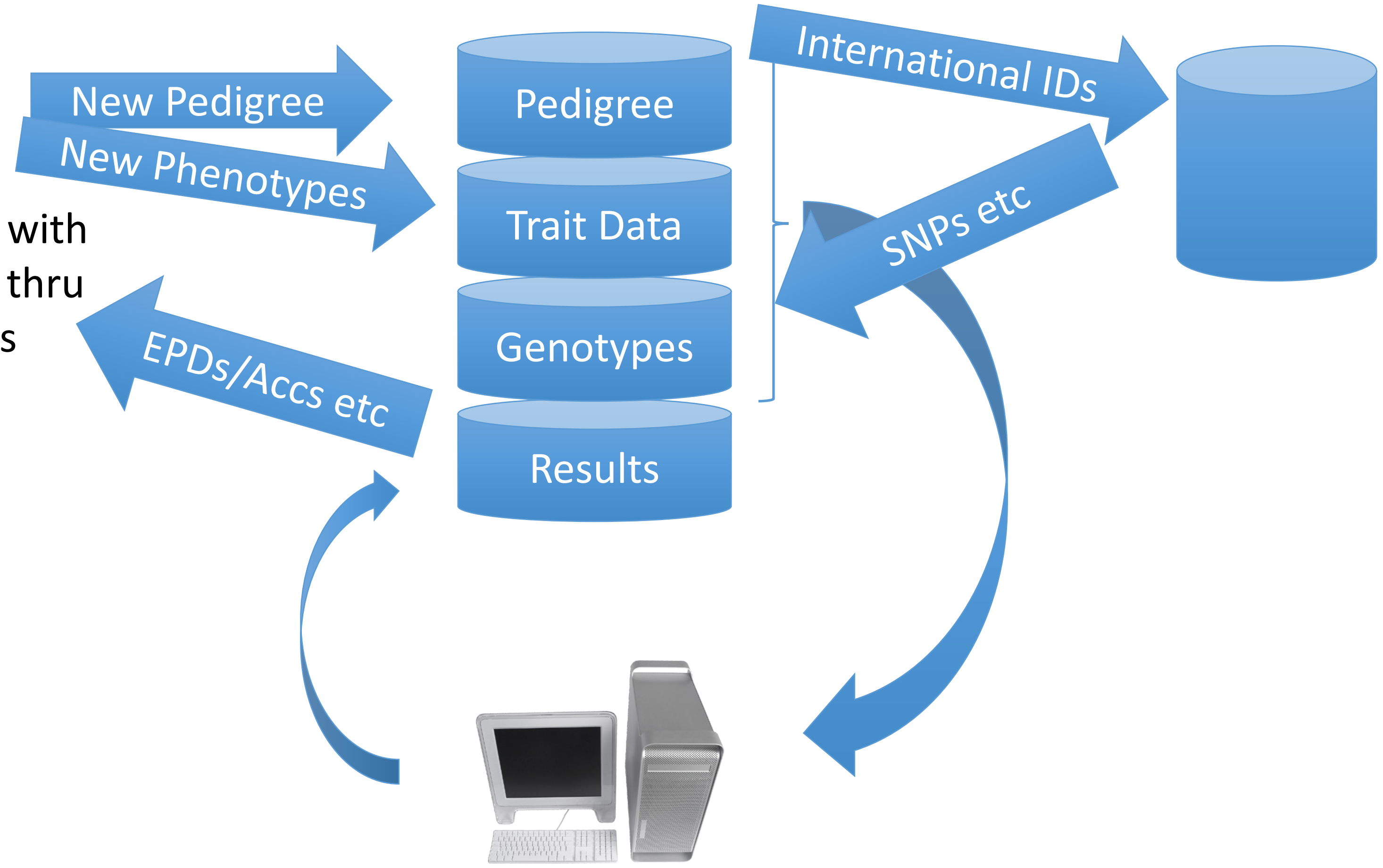


Ranchers etc

Direct interaction with decision support thru
- Cell phones
- Web

Better Decisions!

Single Authoritative Data System



Ranchers etc

New Pedigree

New Phenotypes

Direct interaction with decision support thru

- Cell phones
  - Web

EPDs/Accs etc

Pedigree

Trait Data

Genotypes

Results

International IDs

GeneSeek

SNPs etc

BOLT CUDA
Evaluation System

If we can’t have that – Vision for a turnkey system

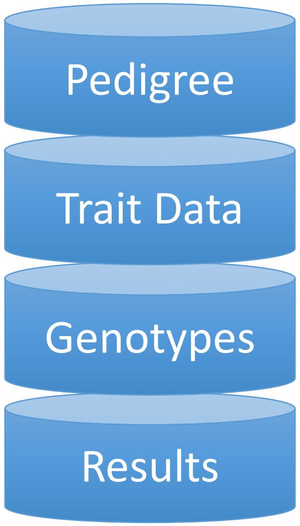
overlapping databases but one authoritative system

Database Duplications

Breed Associations
(IT systems for pedigree
& performance recording)

(eg ASA & other IGS partners)

Authoritative DB
(IT system(s) to facilitate
routine BOLT evaluations)



Pedigree

Trait Data

Genotypes

Results

BOLT CUDA
Evaluation System

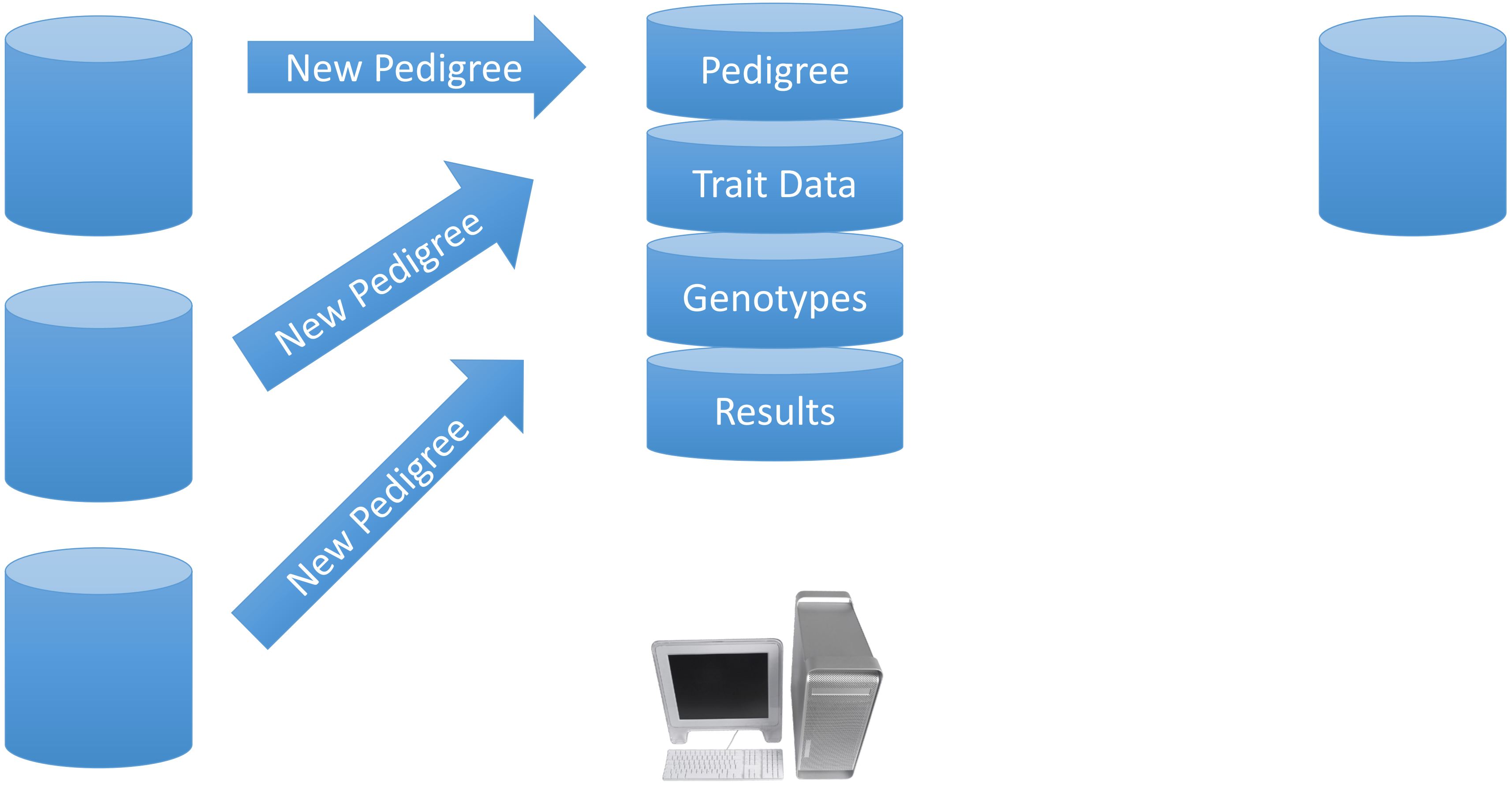
GeneSeek
(IT systems for
LIMS and genotyping)

Database Duplications

Breed Associations
(IT systems for pedigree
& performance recording)

Authoritative DB
(IT system(s) to facilitate
routine BOLT evaluations)

GeneSeek
(IT systems for
LIMS and genotyping)



(eg ASA & other IGS partners)

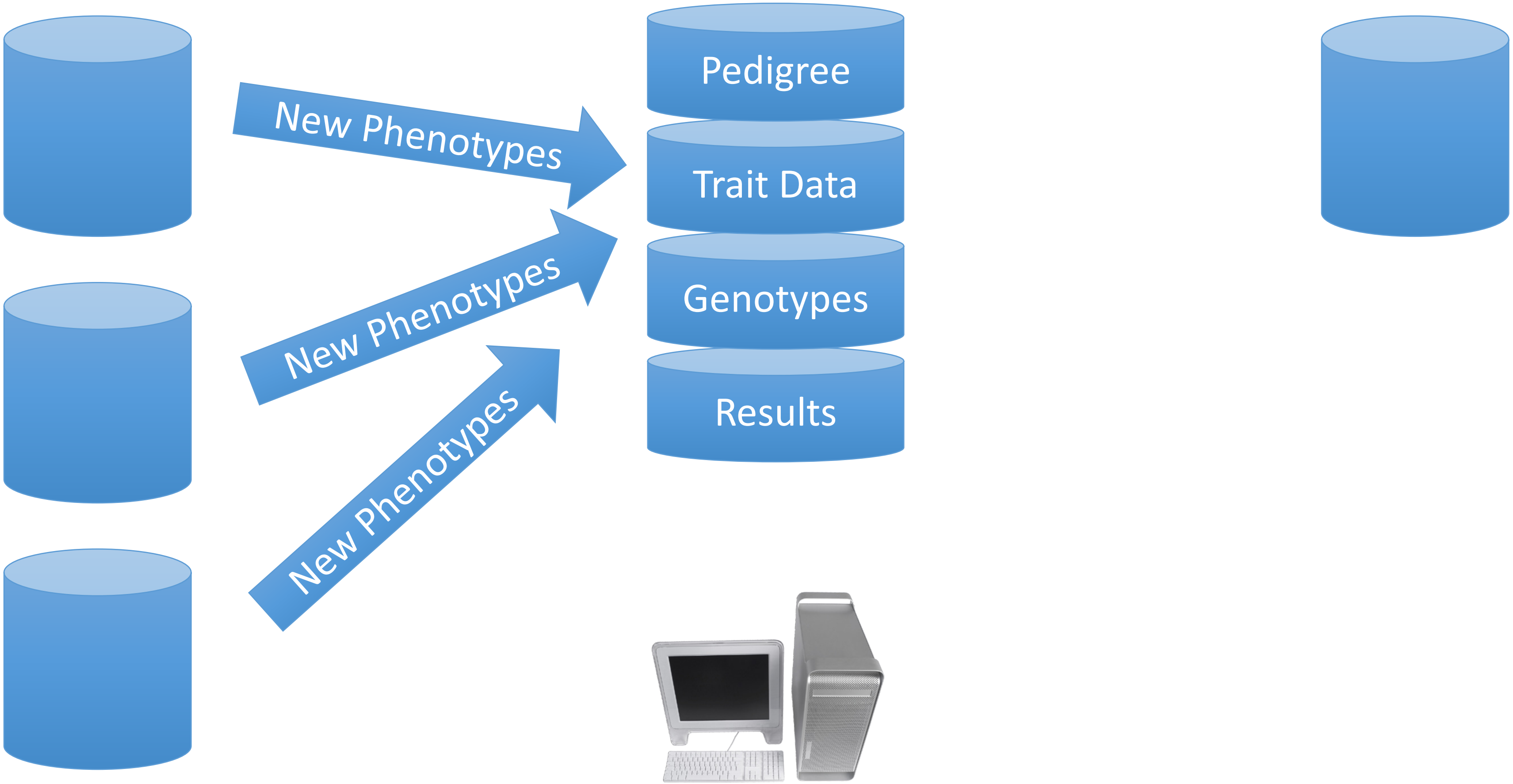
BOLT CUDA
Evaluation System

Database Duplications

Breed Associations
(IT systems for pedigree
& performance recording)

Authoritative DB
(IT system(s) to facilitate
routine BOLT evaluations)

GeneSeek
(IT systems for
LIMS and genotyping)



BOLT CUDA
Evaluation System

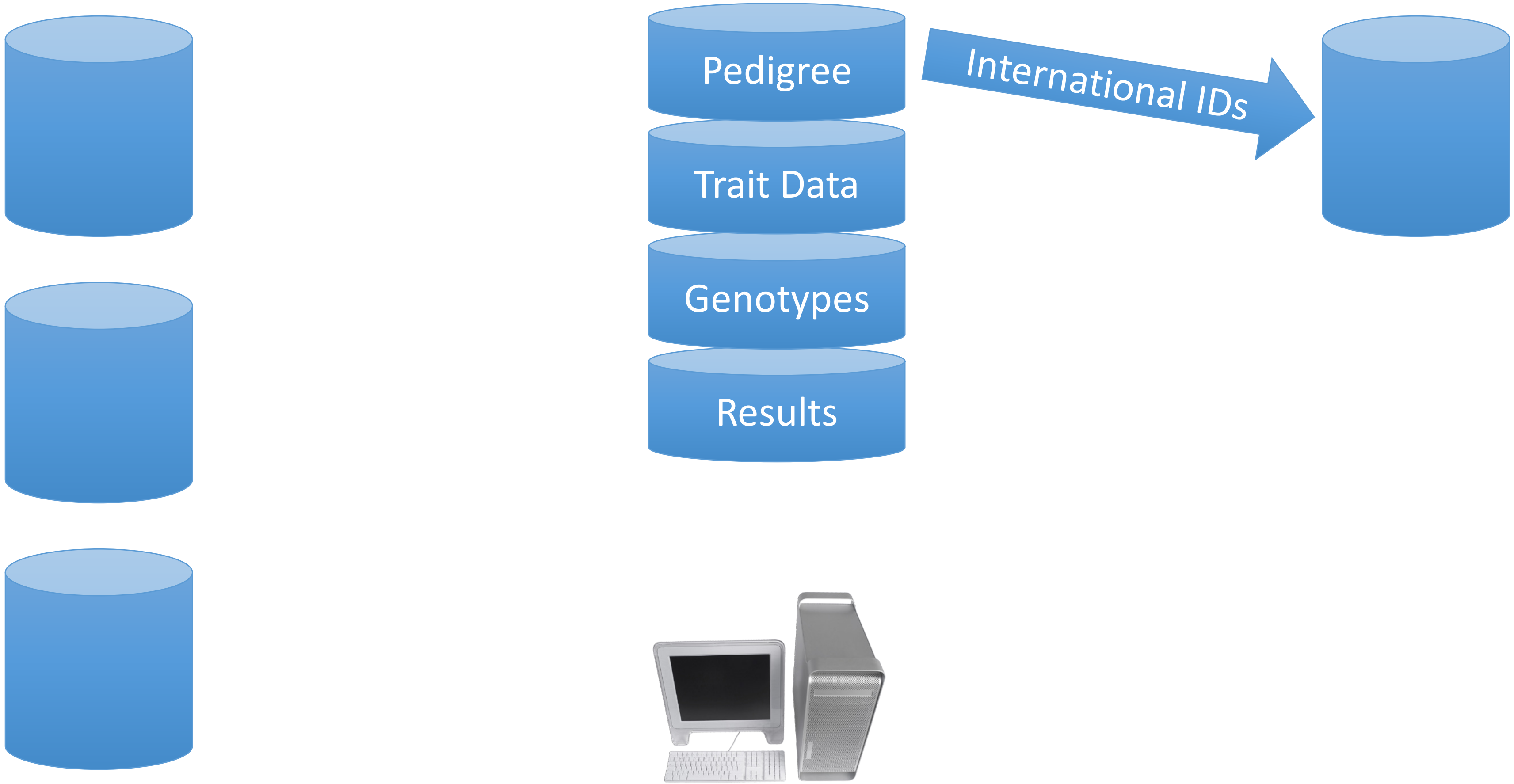
(eg ASA & other IGS partners)

Database Duplications

Breed Associations
(IT systems for pedigree
& performance recording)

Authoritative DB
(IT system(s) to facilitate
routine BOLT evaluations)

GeneSeek
(IT systems for
LIMS and genotyping)



(eg ASA & other IGS partners)

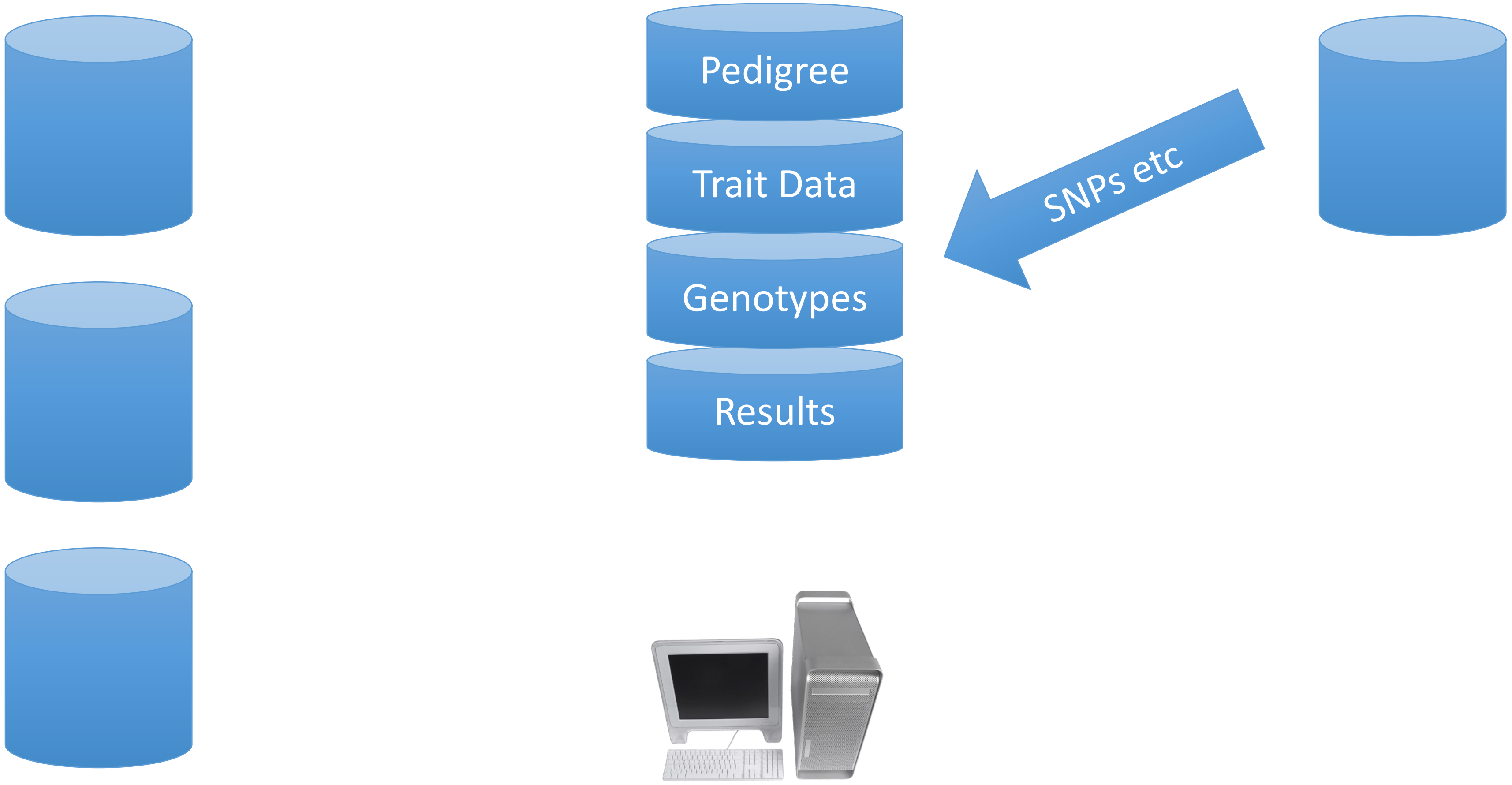
BOLT CUDA
Evaluation System

Database Duplications

Breed Associations
(IT systems for pedigree
& performance recording)

Authoritative DB
(IT system(s) to facilitate
routine BOLT evaluations)

GeneSeek
(IT systems for
LIMS and genotyping)



Pedigree

Trait Data

Genotypes

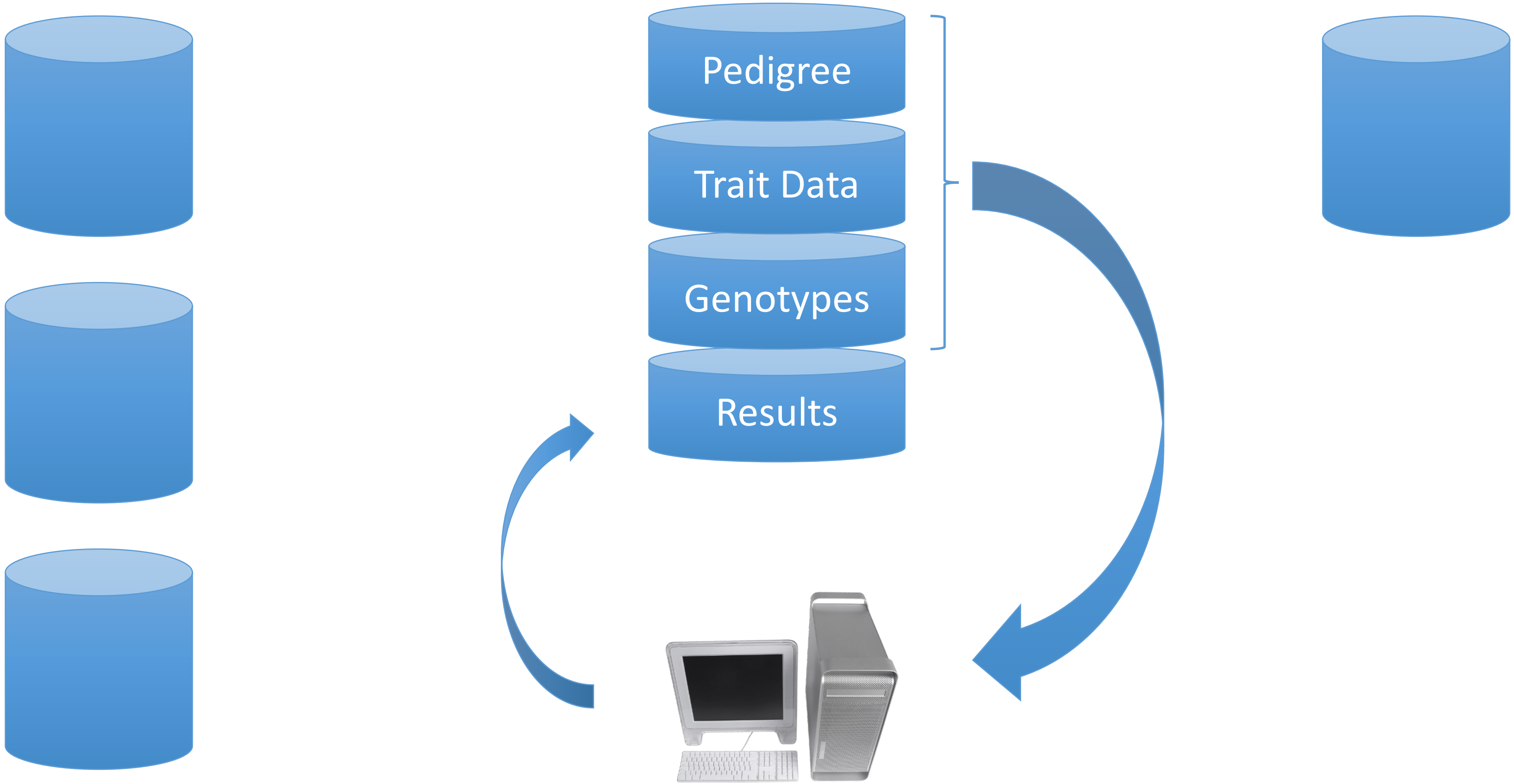
Results

SNPs etc

(eg ASA & other IGS partners)

BOLT CUDA
Evaluation System

Database Duplications



Breed Associations
(IT systems for pedigree
& performance recording)

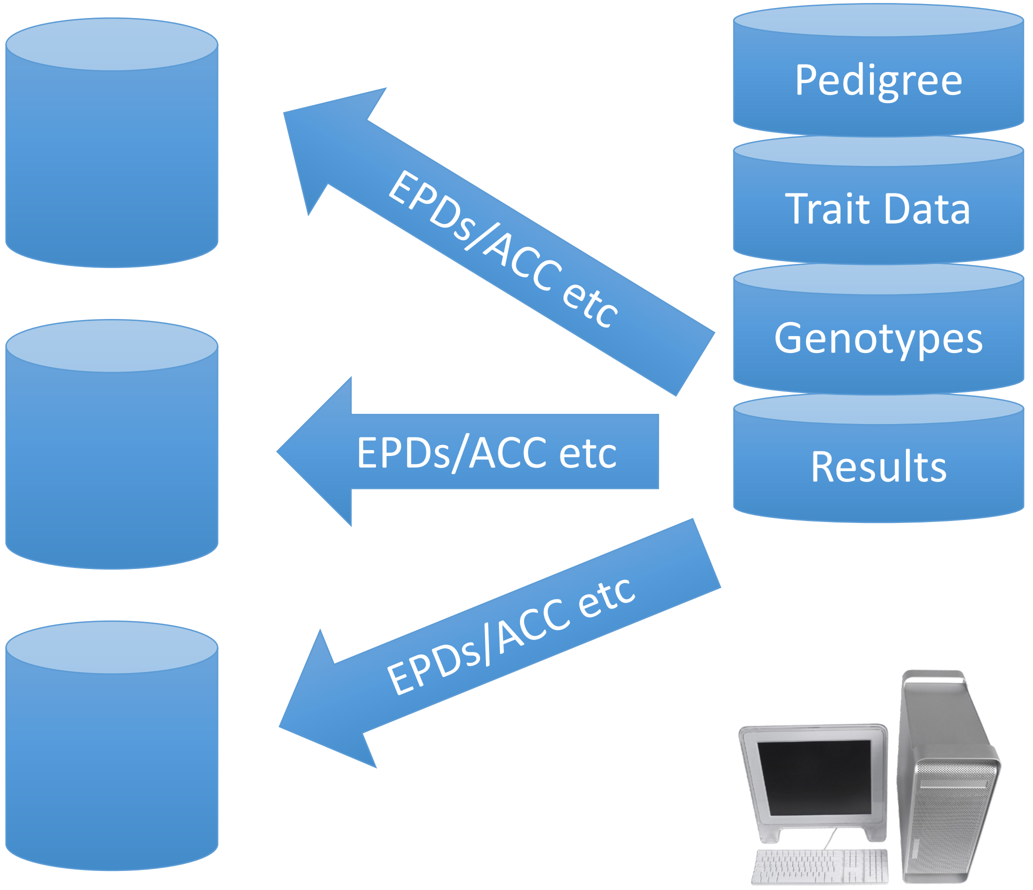
Authoritative DB
(IT system(s) to facilitate
routine BOLT evaluations)

GeneSeek
(IT systems for
LIMS and genotyping)

BOLT CUDA
Evaluation System

(eg ASA & other IGS partners)

Database Duplications



Breed Associations
(IT systems for pedigree
& performance recording)

Authoritative DB
(IT system(s) to facilitate
routine BOLT evaluations)

GeneSeek
(IT systems for
LIMS and genotyping)

Pedigree

Trait Data

Genotypes

Results

EPDs/ACC etc

EPDs/ACC etc

EPDs/ACC etc

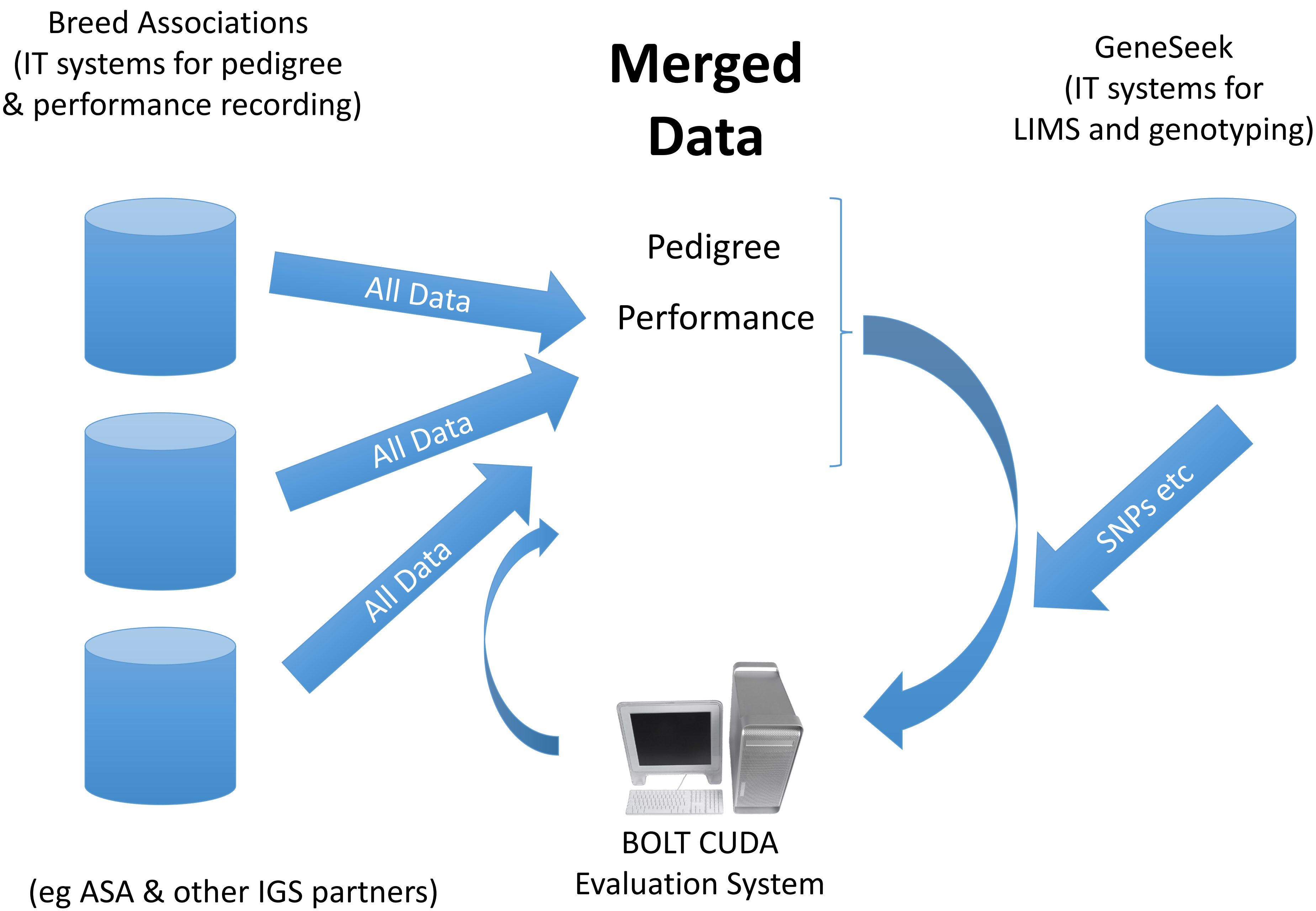
BOLT CUDA
Evaluation System

(eg ASA & other IGS partners)

If we can’t have that – Vision for a turnkey system

repeatedly merge overlapping databases

Merged Data



Breed Associations
(IT systems for pedigree
& performance recording)

All Data

All Data

All Data

(eg ASA & other IGS partners)

Pedigree

Performance

GeneSeek
(IT systems for
LIMS and genotyping)

SNPs etc

BOLT CUDA
Evaluation System

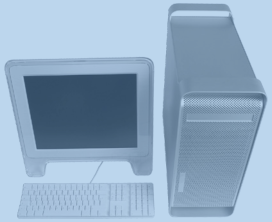
Merged Data

Pedigree

Performance

These systems
to repeatably produce
clean data
are still being
developed and tested

ThetaSolutionsLLC



BOLT CUDA
Evaluation System

These systems
are ready to go
turnkey
when clean data
is available

Merged Data

Monumental effort
for a
“Small step forward”

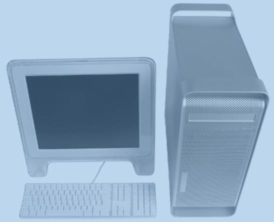
Pedigree

Performance

These systems
to repeatably produce
clean data
are still being
developed and tested

“Quantum leap”

ThetaSolutionsLLC



BOLT CUDA
Evaluation System

These systems
are ready to go
turnkey
when clean data
is available

Genetic Correlations c/c and u/s

		Ribeye Area	Fat	IMF/Marbling
Simmental	**new**	**0.56**	**0.38**	**0.73**
	Old (B, H)	0.8, 0.54	0.79, 0.83	0.74, 0.69
Hereford	**new**	**0.81**	**0.75**	**0.54**
	old	0.75	0.85	0.70

Simmental Old genetic correlations from Crews et al JAS

Genetic Correlations c/c and u/s

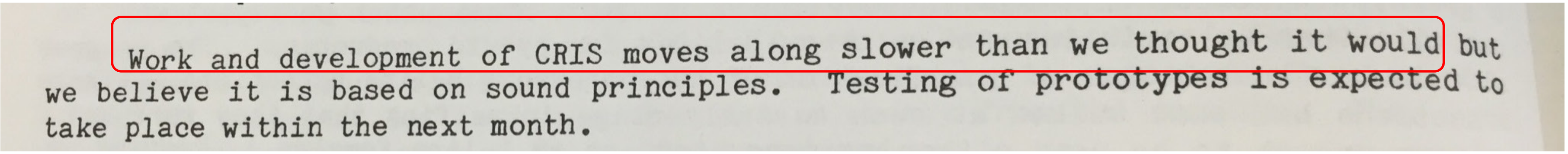
		Ribeye Area	Fat	IMF/Marbling
Simmental	**new**	**0.56**	**0.38**	**0.73**
	Old (B, H)	0.8, 0.54	0.79, 0.83	0.74, 0.69
Hereford	**new**	**0.81**	**0.75**	**0.54**
	old	0.75	0.85	0.70

Simmental Old genetic correlations from Crews et al JAS

Information learned from Pioneer Genetic and Genomic Evaluation

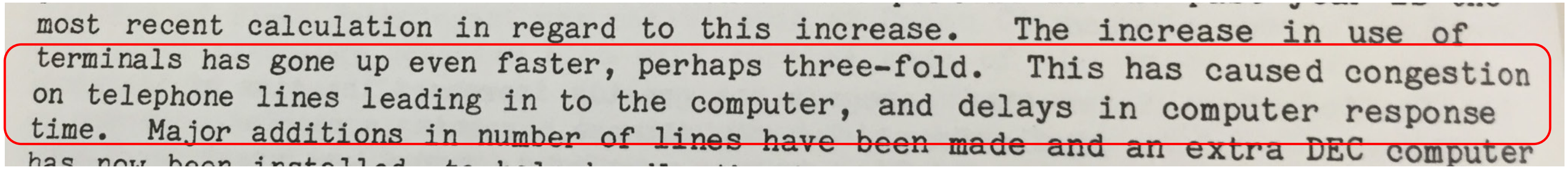
System Development

Software systems: some things are universal



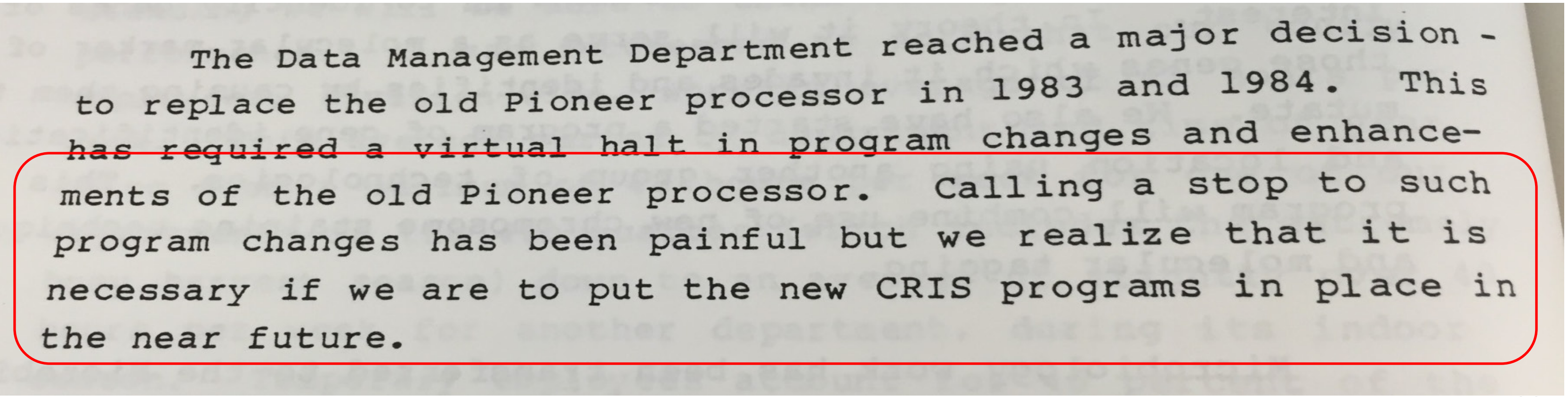
Work and development of CRIS moves along slower than we thought it would but we believe it is based on sound principles. Testing of prototypes is expected to take place within the next month.

Pioneer Annual Research report, 1981. Don Duvick



most recent calculation in regard to this increase. The increase in use of terminals has gone up even faster, perhaps three-fold. This has caused congestion on telephone lines leading in to the computer, and delays in computer response time. Major additions in number of lines have been made and an extra DEC computer

Pioneer Annual Research report, 1982. Don Duvick



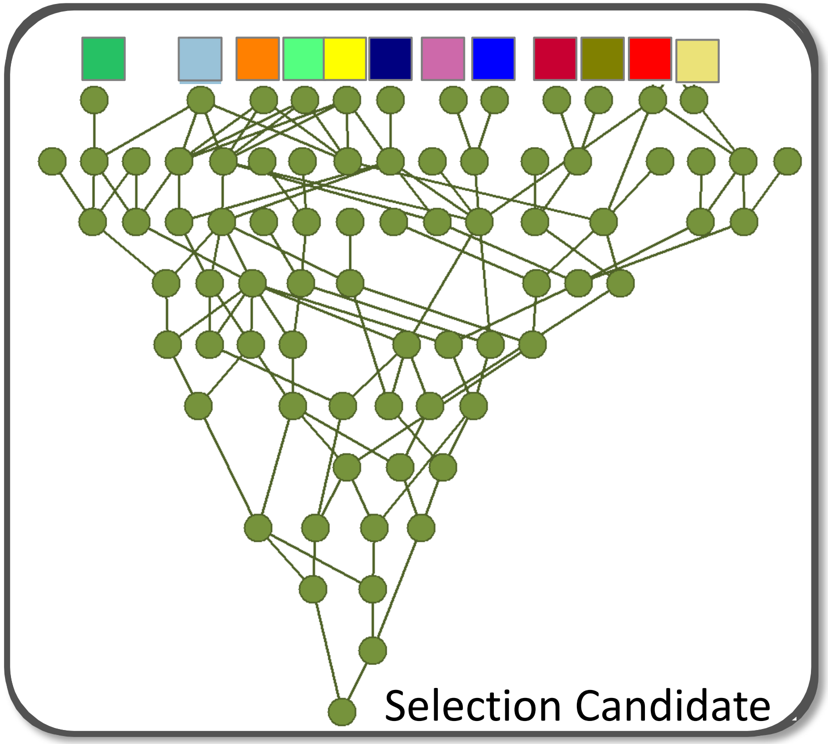
The Data Management Department reached a major decision - to replace the old Pioneer processor in 1983 and 1984. This has required a virtual halt in program changes and enhancements of the old Pioneer processor. Calling a stop to such program changes has been painful but we realize that it is necessary if we are to put the new CRIS programs in place in the near future.

Pioneer Annual Research report, 1984. Don Duvick

Visualization

Information Visualization

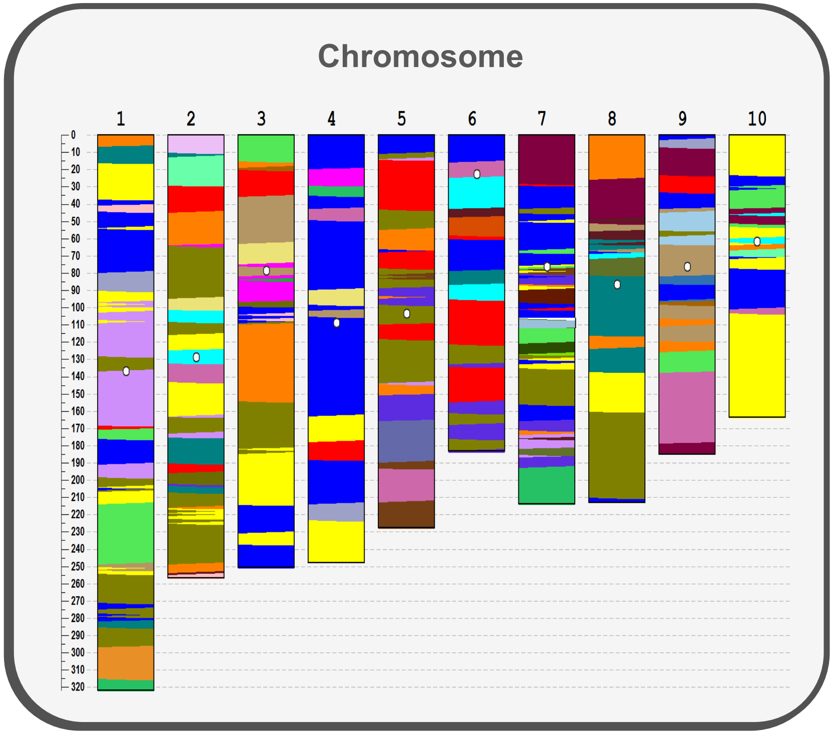
Founder



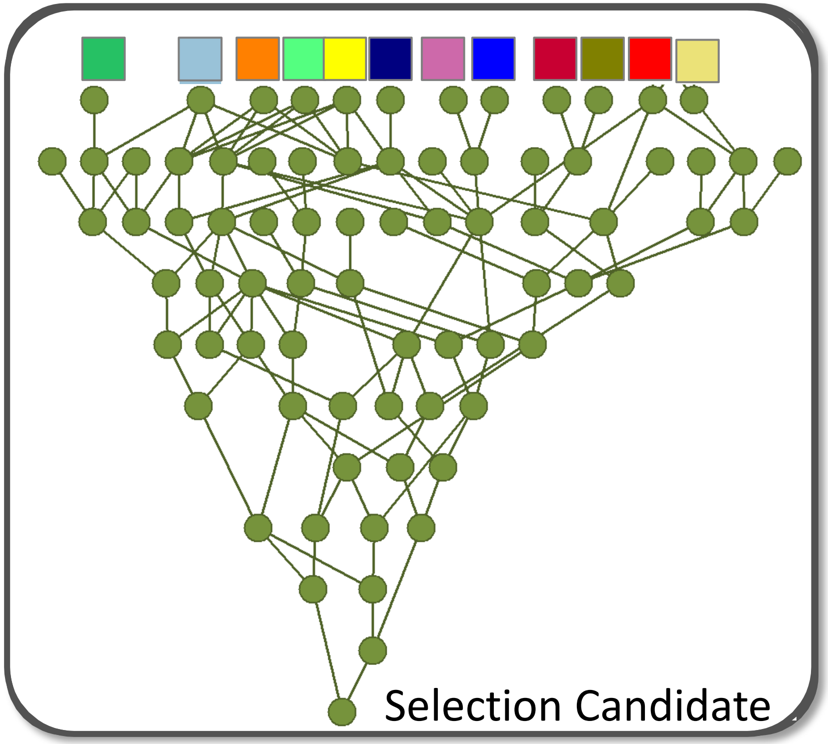
Selection Candidate

Information Visualization

Inheritance Summary

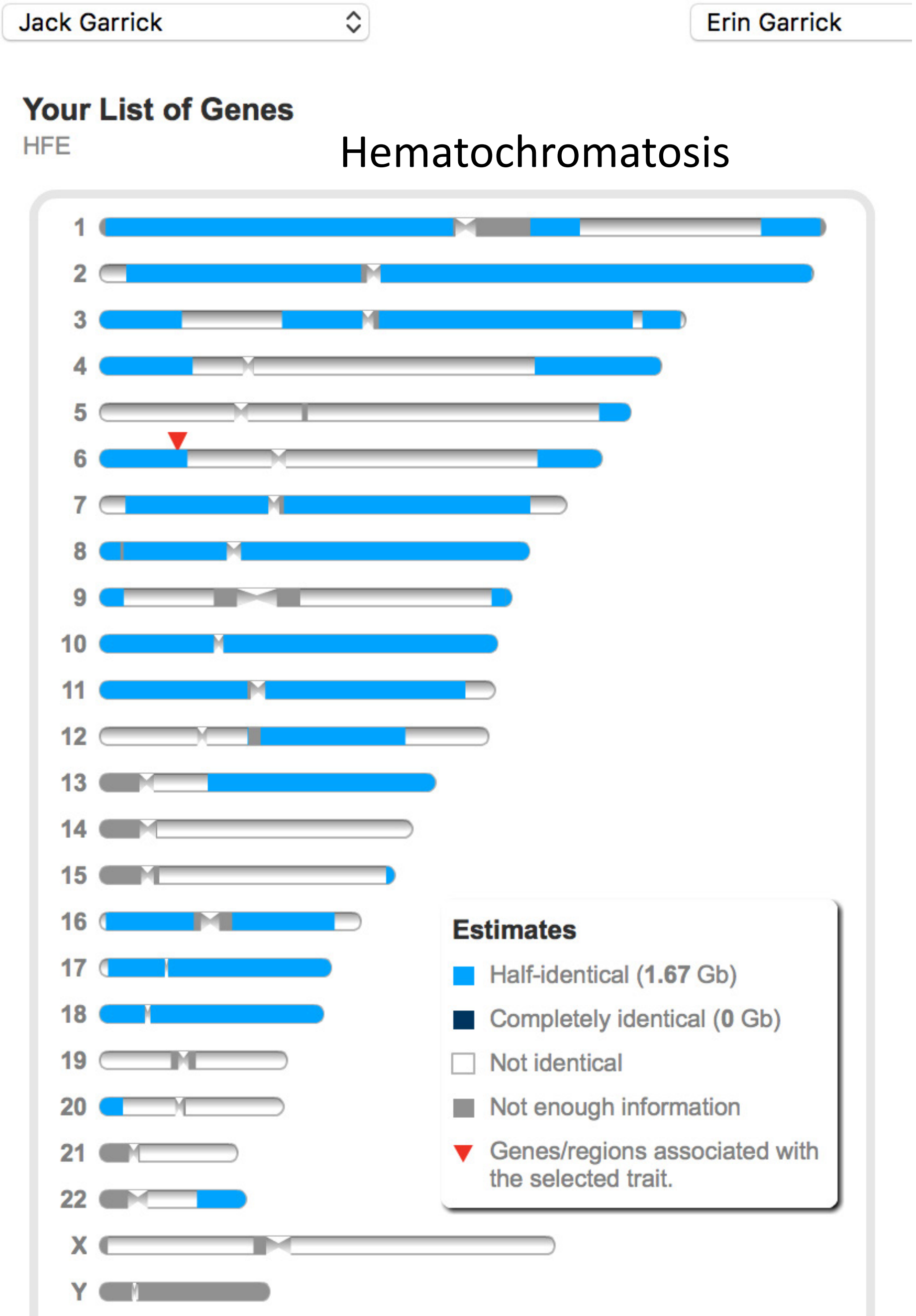


Founder



Paternal Grandfather

Hematochromatosis


Jack Garrick

Erin Garrick

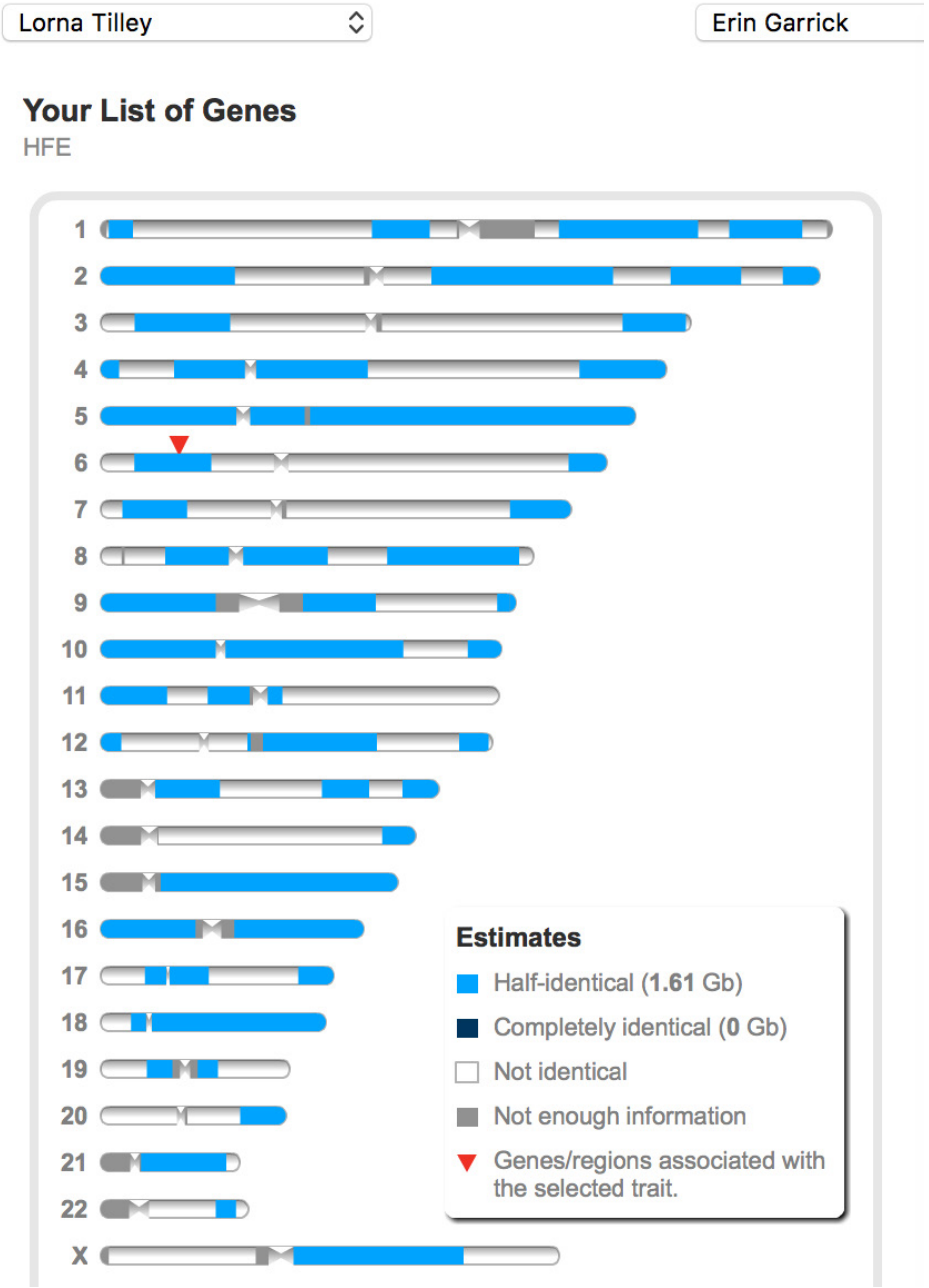
Your List of Genes

HFE

Estimates

- 🟦 Half-identical (**1.67** Gb)
- ■ Completely identical (**0** Gb)
- ☐ Not identical
- ▨ Not enough information
- ▼ Genes/regions associated with the selected trait.

Maternal Grandmother


Lorna Tilley

Erin Garrick

Your List of Genes

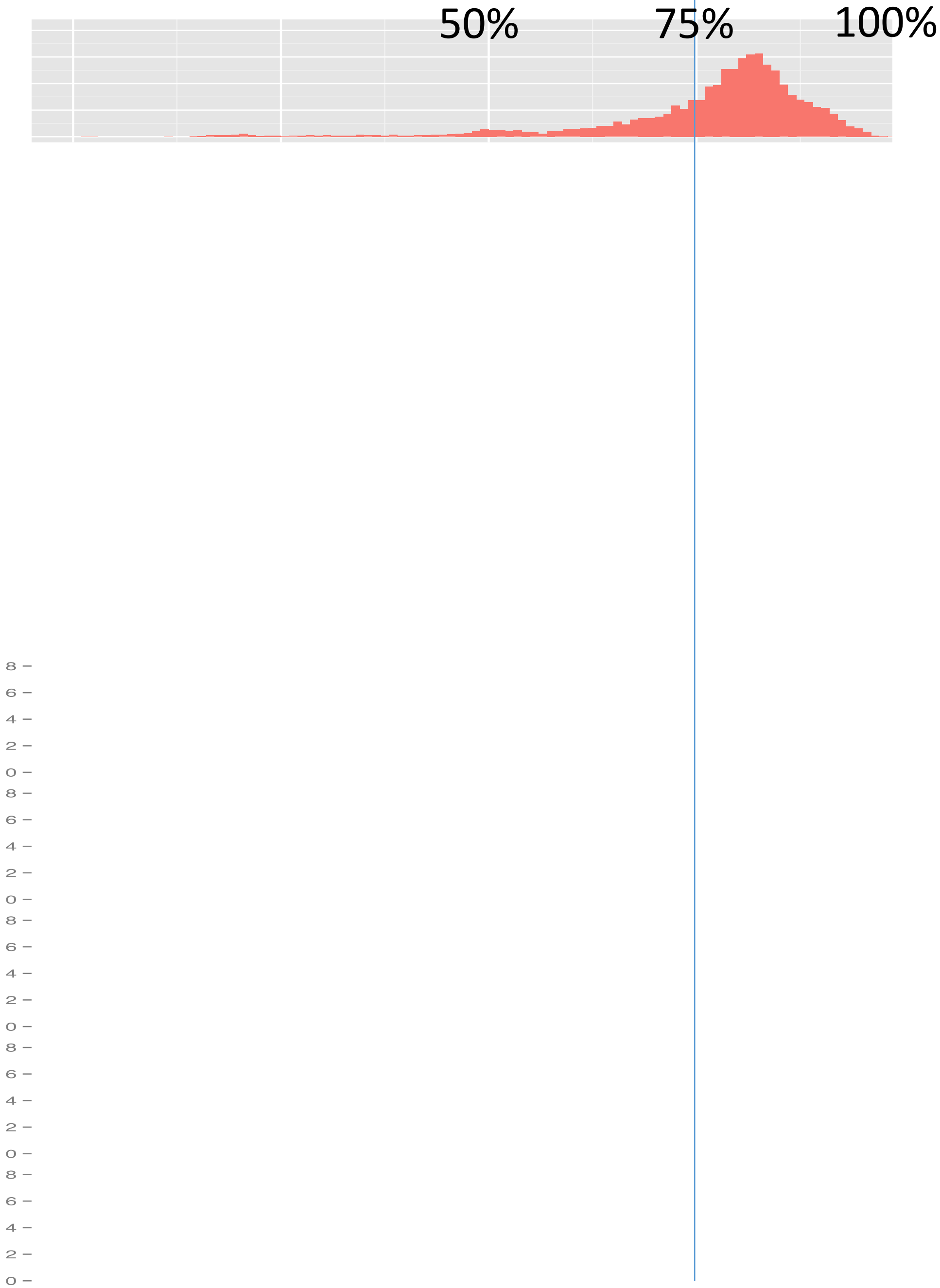
HFE

Estimates

- 🟦 Half-identical (**1.61** Gb)
- ■ Completely identical (**0** Gb)
- ☐ Not identical
- ▨ Not enough information
- ▼ Genes/regions associated with the selected trait.

www.23andme.com

IGS data



Angus

Brangus

Shorthorn

Charolais

Gelbvieh

Hereford

Limousin

Red Angus

Maine Anjou

Simmental

Proportion of genome that comprises common haplotypes

Significant fractions of genotyped animals comprise rare haplotypes (seen <1% time) in >25% their genomes

Its impossible to separately estimate effects of multiple rare haplotype alleles observed only once in the same individual

Information to use in evaluation

Most Accurate Prediction

- The most accurate predictions don’t come about from using ALL the data
- The most accurate predictions come about from using the MOST INFORMATIVE data

- We need to test this using IGS data
  - when we have access to a suitable dataset
  - Regional data from related breeds to the selection candidates may be more accurate than using data from all breeds and all regions

Information learned from Irish Cattle Breeding Federation Genetic and Genomic Evaluation

Genotyping Costs are Declining

- Bulk deals committing to large volumes of samples have been able to enjoy 50K SNP chip prices of $20 per sample
    - including DNA extraction, genotyping and reporting
- Should all parents be required to be genotyped?

Basic Issues Need Attention

Animal Identifiers

- We use a variant of the Interbull ID system

SIMUSAM000000123456

19-digit international ID

Breed Code
AAN=Angus
BRG=Brangus
BSH=Shorthorn
CHA=Charolais
HER=Hereford
LIM=Limousin
NEL=Nellore
RAN=Red Angus
RDP=Maine-Anjou
SIM=Simmental

Country Code
ARG
AUS
CAN
URG
USA

Sex Code
M=bull
F=cow
(U=unknown)

Registration Number
Left-padded with 0
Can include alphanumerics

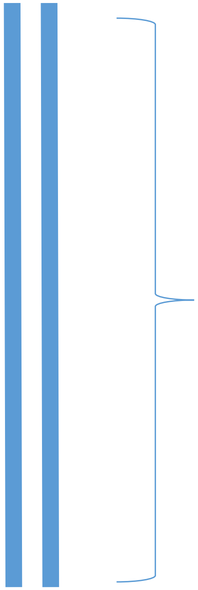
We use Breed Association rather than Breed
(unless animals are not registered)
Prefer to use country/breed of first registration

It would be helpful if all the IGS breed associations fully adopted this approach

Genotype Quality Control

- Genotyped sex must agree with the pedigree-recorded sex
  - Many samples fail this test

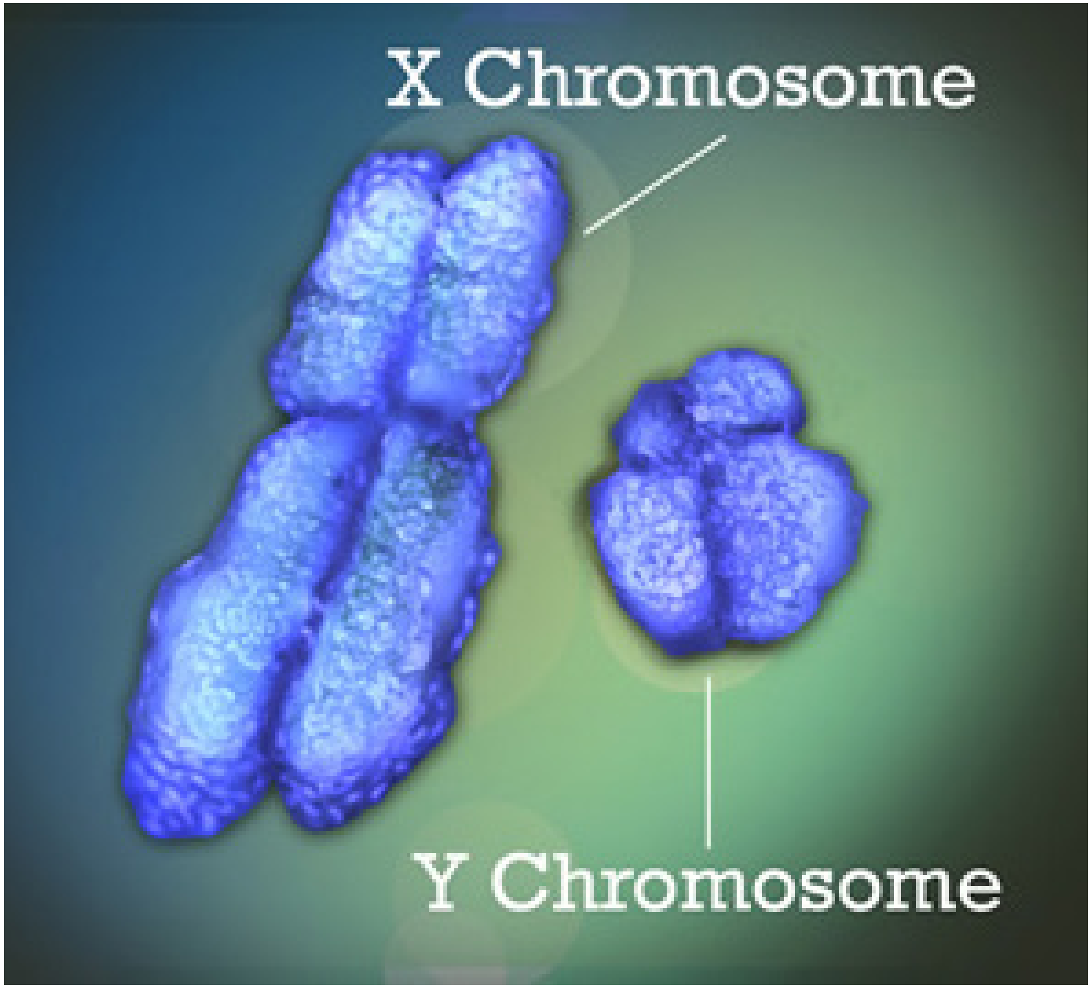
Autosomes vs Sex Chromosomes



A true "pair" of chromosomes
are about the same length
contain the same genes in the same order
have minor variants (eg SNPs)
in the version of the
gene inherited from the sire vs the dam

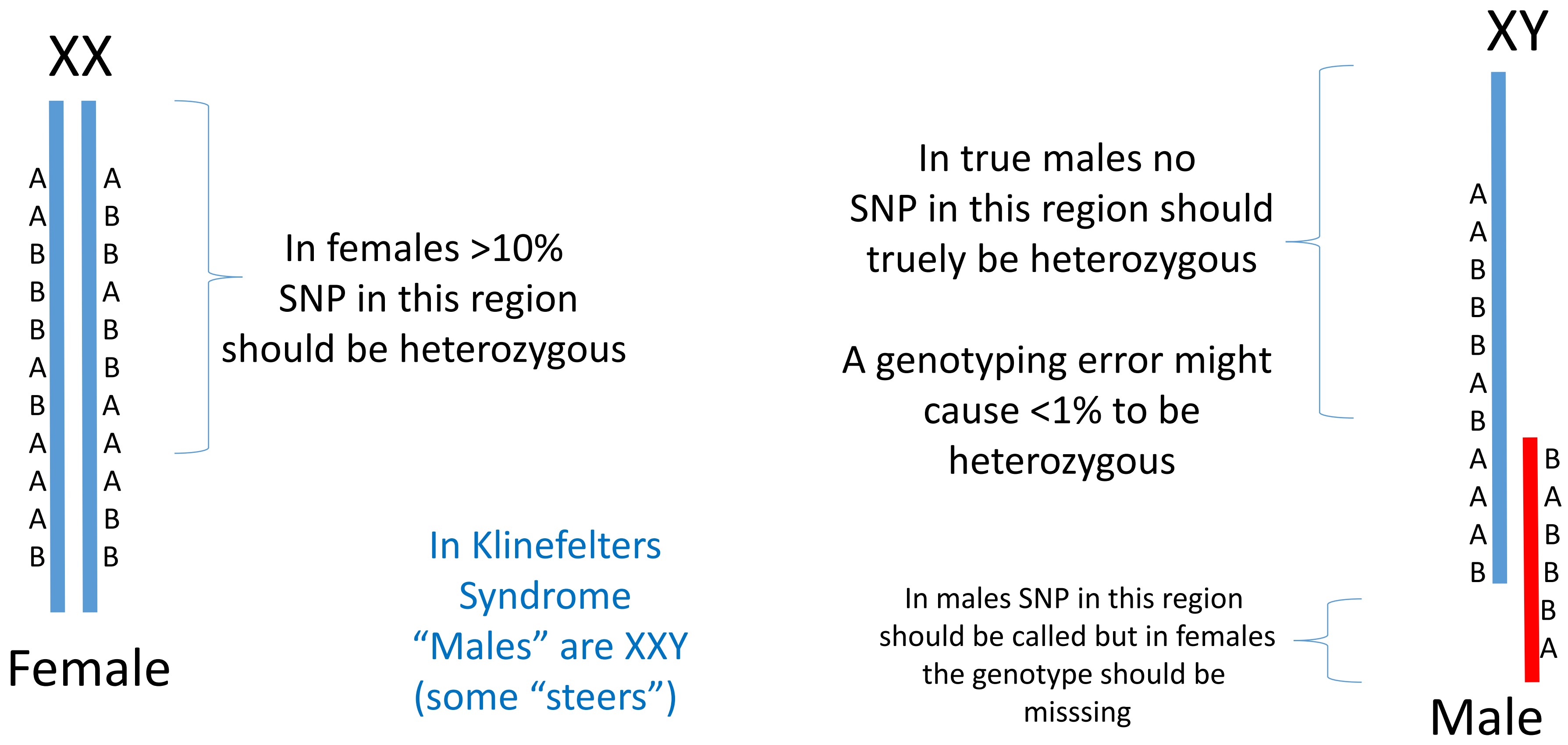
In contrast, sex chromosomes are not proper "pairs"........

Autosomes vs Sex Chromosomes



In contrast, sex chromosomes are not proper “pairs”……..

Autosomes vs Sex Chromosomes



In females >10% SNP in this region should be heterozygous

In true males no SNP in this region should truely be heterozygous

A genotyping error might cause <1% to be heterozygous

In Klinefelters Syndrome "Males" are XXY (some "steers")

In males SNP in this region should be called but in females the genotype should be misssing

Genotype Quality Control

- Genotyped sex must agree with the pedigree-recorded sex
- Genotypes should not exhibit parent-offspring conflicts (IGS failure rate > 6% fail vs USMARC < 3%)
  - Many samples fail this test
  - This test becomes easier to do as more animals have one or both parents genotyped
  - With all animals genotyped, parentage conflicts can be resolved from the genotype panels

Genotype Quality Control

- Genotyped sex must agree with the pedigree-recorded sex
- Genotypes should not exhibit parent-offspring conflicts
- Genotyped breed (or breed composition) should agree with pedigree
  - Only relevant when the parent-offspring tests cant be done

Information learned from Various Genetic and Genomic Evaluations

Predictive Ability

- We need research focus on improving the accuracy with which we can predict animal performance
- Many options are available for improving predictions
  - Better marker panels – fewer better features used
  - More animals genotyped
  - More phenotypes collected
    - particularly for carcass, reproduction and disease
  - Improved quality control of all data
  - Better models and analytical methods

Summary

- The purpose of collecting pedigree, performance and genomic data is to make better selection decisions

Summary

- The purpose of collecting pedigree, performance and genomic data is to make better selection decisions
- The information systems used to input, store and analyze that data need ongoing development

Summary

- The purpose of collecting pedigree, performance and genomic data is to make better selection decisions
- The information systems used to input, store and analyze that data need ongoing development
- Current systems used by most breed associations in most parts of the world are well short of the visions we have for modern information systems

Summary

- The purpose of collecting pedigree, performance and genomic data is to make better selection decisions
- The information systems used to input, store and analyze that data need ongoing development
- Current systems used by most breed associations in most parts of the world are well short of the visions we have for modern information systems
- Implementation of new and improved analytical systems are currently being held back by lack of best practice in data systems (fit for purpose data)

el Abed (2009) Data Governance – a business value-driven approach

There is also bad news

- No one has even the vaguest idea what software *really* costs over time. No one.

There is also bad news

- The unfortunate notion of “software sustainability” has become popular in the grant writing world.
  - No one wants to hear that “sustaining” means a budget that is the same annual budget as development, likely forever, or at least as long as this complex formula:

Sustaining time =
(How long do you want it to actually work) – (About 3 weeks).

There is also bad news

- It used to be said “Open source isn’t like free beer, it’s like a free puppy”.
- It’s really more like a free house, with a mortgage. Only a mortgage that doesn’t end in 30 years.
  - The only reasonable notions of “sustainable” in a house with an endless mortgage:
    - Have an annuity bigger than maintenance costs.
    - Sell the liability to some other poor sucker.