Current and future directions for animal breeding software

Paul VanRaden, retired

Previously: USDA, Agricultural Research Service, Animal Genomics and Improvement Laboratory, Beltsville, MD 20705

vanraden@aol.com

Animal breeders need 7 steps

- Step 1 in making genetic progress is to understand genetics
- Step 2 is to uniquely identify the animals, ancestors, clones
- Step 3 is to know what traits are important
- Steps 4 and 5 are to collect and analyze data
- Step 6 is to write programs that can keep up with data growth
- Step 7 is to help owners improve their animals

35 years of USA data growth 1990 vs 2025

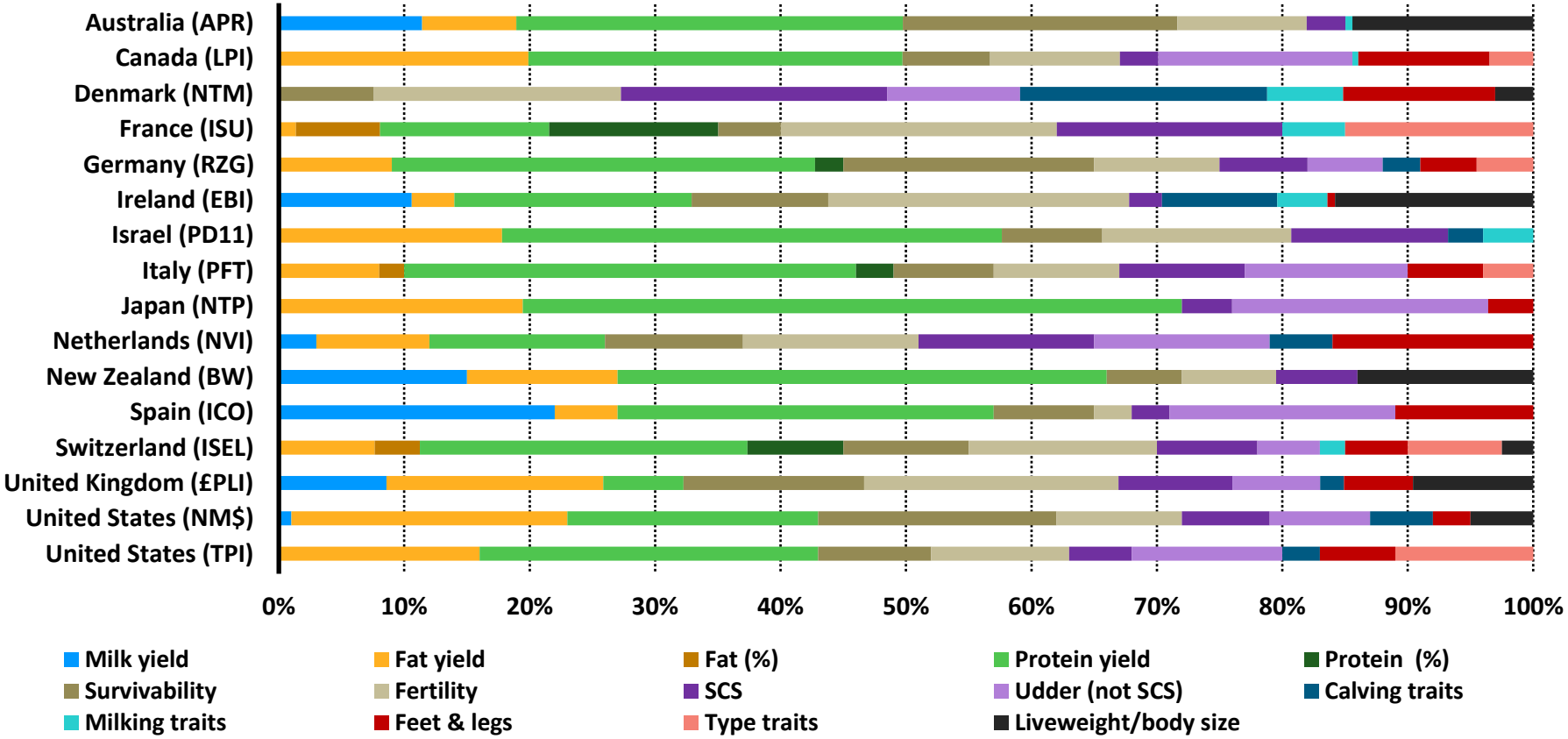
- Growth of data
  - 3x more animals
  - 1000x more foreign IDs
  - Crossbred animals
- New data sources
  - Genotypes
  - Low heritability traits
  - Expensive traits (RFI)

Statistic	1990	2025
Pedigree	30 M	102 M
Foreign IDs	1,500	1.5 M
Crossbreds	0	6 M
Type traits	15	21
Other traits	3	26
Milk lactations	35 M	109 M
Type records *(HO)*	1 M	11 M
DNA marker loci	0	69k
Animals genotyped	0	11 M

Foreign phenotypes since 1995

- Bull geneFiles provided by Interbull, Uppsala, Sweden
- tic evaluations for 35 traits
  - Combined data from 30 countries on 5 continents
  - Deregress (remove PA) to get daughter averages from EBVs
- Merged with domestic data if reliability is higher
  - Benefits small if domestic dataset is already large
  - But helps breeders discover and select foreign bulls

What do other countries include in indexes? 2021



Statistical methods used by other countries in 2019

- Many models to choose from and several currently in use
- 17 countries filled out Interbull Form GENO as of 2019:
  - 10 used multi-step GBLUP linear model
  - 4 used multi-step Bayesian model (non-normal SNP effect distributions)
  - 2 used single-step GBLUP linear model
  - 1 used linear haplotype model after choosing subset by Bayesian model
  - 9 of the 17 countries include a polygenic term (non-SNP genetic variance)
- Where did these methods come from?

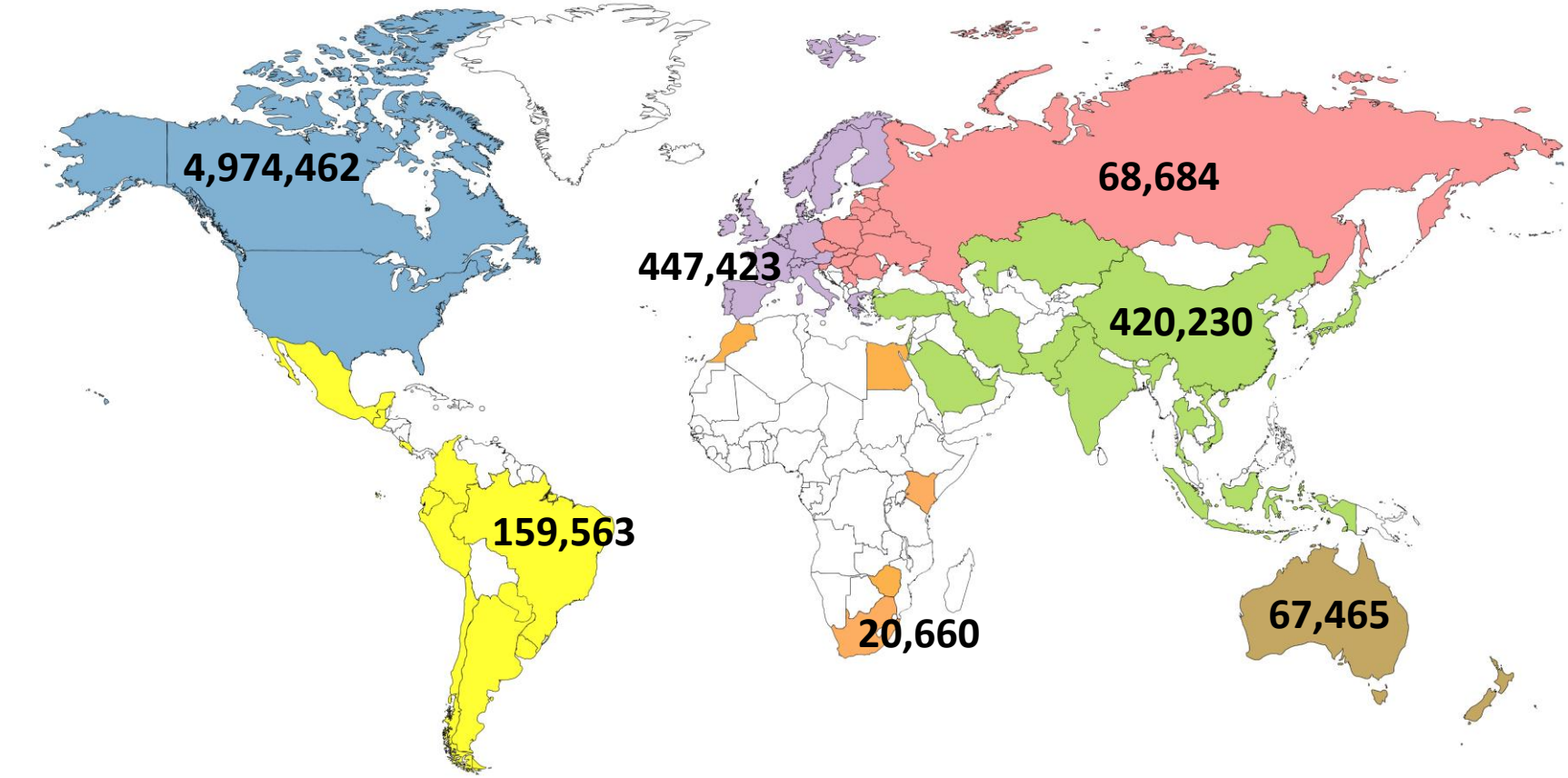
Linear model using genomic relationships

- Nejati-Javaremi, Smith, Gibson (1997) derived GBLUP methods
- First author is a professor at U. Tehran, Iran
- Compared G matrix to A matrix
  - Used individual SNP genotypes (not haplotypes)
  - Simulated 1,000 phenotypes with 30% heritability
  - Selected directly on 100 QTLs (not markers, no polygenic effect)
  - Obtained 71% REL with G matrix vs. 42% REL with A matrix
- Research was from U. Guelph, Canada, but was not cited by Schaeffer (2006)

Bayesian haplotype models using Gibbs sampling

- Meuwissen, Hayes, Goddard (2001) compared Bayesian to linear models
  - Used haplotypes (not individual SNP effects) from 1,000 markers
  - Simulated 2,200 phenotypes with 50% heritability and 100 QTLs with some very large effects
  - Obtained REL = 20% pedigree, 54% GBLUP, 64% BayesA, 72% BayesB
  - Did not cite Nejati-Javaremi et al (1997 JAS) in their 2001 or later papers
- Calus et al (2008) showed that individual SNP effects are better than haplotypes at higher SNP densities; polygenic effects should be included
- BayesA and BayesB had much smaller gains in later studies of real traits

Genotype counts by region (June 2022)



Countries sending most genotypes to CDCB - 2024

- Numbers of females born in latest 5 years
- Countries with advanced breeding programs each have large databases
- Many other countries do not have historical data to compute predictions

Country	Cows genotyped
United States	3,571,054
Canada	320,350
Saudi Arabia	186,499
China	160,558
Japan	135,971
Italy	114,501
Brazil	102,000

USDA traditional (pedigree model) program series

Program	Purpose
Extract programs (mostly sas)	Create y, X, Z variables in model for trait group from DB2
ldxbred.sqc	Compute breed fractions and heterosis for crossbreds
inbreed2.f90	Compute pedigree inbreeding and relationships
traitvar.f90, traitpheno.f90	Apply variance adjustments to phenotype data
1step.f90	Compute traditional EBVs (not genomic yet)
combine_ebv, setbaseall (sas)	Combine EBVs across traits, get base adjustment factors
netmerit.f90	Add foreign and genomic EBVs, get PA, MT and NM$
towithin.sas	Adjust all-breed EBVs to within-breed PTAs
make038.sqc, mk105.c	Format bull and cow data for final distribution

USDA genomic program series

Program	Purpose
extr_geno.py	Extract genotypes from DB2, make chromosome.data
conflict.f90	Remove Mendelian conflicts and merge clones
findhap.f90	Impute genotypes
fixped.f90	Discover ancestors using haplotypes
hapcheck.f90	List carriers of recessive haplotypes (genetic defects)
densemap.f90	Deregress PTAs, estimate SNP effects and GPTAs
Gmatrix.f90	Compute genomic inbreeding and relationships
combineG1, G2, and G3	Combine genomic and trad EBVs within and across traits
deliverG.sas	Make file subsets for each sending organization

Software packages by Paul (from USDA)

- Free software and code
- Taxpayer funded
- Several packages public
- Others upon request
  - Densemap from 2008
- Poor user friendliness
- US data was top priority
- Users may provide ideas

Package	Purpose
Bestpred	Lactation yield from test day
Genosim	Simulate genotypes & traits
Findhap	Impute genotypes
Findmap	Align and call sequence data
Fixped	Ancestor discovery
REMLM	Sire model variances
Nonadd	D and AA pedigree inverses for dominance and epistasis

SNP effect estimation (in densemap.f90)

- Estimate additive genetic merit as sum of SNP effects
  - Assume 90% of variance from SNPs and 10% polygenic
  - BayesA (heavy-tailed) distribution of SNP effects
- Reduce computation by using:
  - Prior estimates from previous month when adding data
  - Only reference animal genotypes during iteration, then apply to young animals or newly received genotypes
  - Many traits in parallel with 1 set of genotypes in memory

Parallel processing

- Most programs use parallel Fortran since 2008
  - Automatic parallel did not work well (defaulted to single)
  - Mostly parallel do loops using OMP directives
  - Outer loop can do batching if few operations per processor
- Match subscript order and loop order
  - Each processor uses data stored together on same page
  - Example: phenotype (animal, trait), process parallel by trait
  - Other languages (or compilers?) store in different order

Fortran coding, compiling, and file naming

- Standard fortran coding
  - Code to the left
  - Indent 2 spaces each level
  - Comments right justified and on separate lines
  - Useful if editor does not color code different text
- Subscript checking on for testing, off for routine use

- File naming options:
  - Let user choose the name
  - Use same name in each date/trait group directory
- Second option seems better for routine processing
  - Can link specific file name
  - Use several simple input files instead of 1 complex

Convergence of EBVs

- Slowest convergence may be for the best new bulls
  - They have thousands of descendants without records
  - Use parent average as starting value instead of 0 priors
  - Automatically bypass unneeded pedigree (like Reduced Animal Model)
  - Genotypes of young animals are also bypassed
- After detecting convergence, process pedigrees and genotypes of descendants without records in one final iteration

Missing traits

- Some ads claim animals not worth using if they are “missing” traits
- Selection index (NM$) requires PTA for each trait
- Options for filling missing PTAs
  - Breed average for birth year
  - Parent average
  - Correlations with other evaluated traits
  - Genomic prediction
- Since 1995, USDA and CDCB ranked all bulls (domestic and foreign) by filling missing traits using correlations in program netmerit.f90

2000 Approximate multi-trait (VanRaden, 2001)

- **Estimated breeding values:**
  - Selection index methods
  - Combine ST EBV into MT
- Approximate productive life EBV, fill missing traits
- Combine single-country GEBV into multi-country GEBV at Interbull (GMACE)

- **Reliability:**
  - Estimate ST REL (using same model as for EBV)
  - Estimate covariances by measuring dataset overlap
  - Apply selection index to combine ST REL into MT
- Computed for all MT models

Within-breed genetic bases

- Most computations are now done on all-breed base with inbreeding and heterosis effects removed
- All-breed EBVs and GEBVs are then converted to PTAs on within-breed bases before publication
  - Crossbred animals also published on purebred bases
  - PTAs include their expected nonadditive effects (Expected Future Inbreeding and heterosis) when mated to a purebred population
  - Automatic penalty for causing inbreeding

Delivery and timing

- All domestic and foreign PTAs are updated 3 times / year
  - In /eval1/2404, 2408, 2412 (Apr, Aug, Dec of year 2024)
- Newly genotyped animals receive:
  - Official GPTAs monthly, reprocessing all genotyped animals
    - In /eval1/2502/ for example
  - Unofficial GPTAs weekly with faster, less precise math
    - In /eval1/20250128/ for example
- Quick turnaround even less accurate, only to owner

Validate that predictions work well

- Regress future on past EBV
- Regress true on estimated
  - Requires good simulation
  - Answer questions not possible with real data
- Cross-validation (subsets)
- Convince users to use EBV
  - May be most important

Example studies
Animal model (4 paths of selection, 1989)
Include foreign data (1995)
Genomics – simulated (true BV, 2008)
Genomics – actual (EBV, 2009)
Compare model options
Heifers instead of bulls (Toghiani, 2024)

Single-step experiences with my own code

- **1step.f90 used algorithm of Legarra-Ducrocq (2012) but would not converge for >100,000 genotyped animals**
- VanRaden, 2012a. Iterative combination of national phenotype, genotype, pedigree, and foreign information. J. Dairy Sci. 95(Suppl. 2):446 Slides
- VanRaden and Tooker, 2012b. Methods to include foreign information in national evaluations. J. Dairy Sci. 95(Suppl. 2):449. Slides
- VanRaden, 2012c. Avoiding bias from genomic pre-selection in converting daughter information across countries. Interbull Bull. 45, 5 pp. Slides

My single-step conclusions slide from 2012c

- 1-step genomic evaluations were tested:
  - Inversion avoided using extra equations
  - Converged well for Jerseys but not for Holsteins
  - Same accuracy, less bias than multi-step
  - Foreign data from MACE included
- Further work is needed on algorithms:
  - Including genomic information
  - Extending to all-breed evaluation

Recent experiences with BLUPF90 code

- **Comparisons with multi-step, crossbreds, UPGs, and fertility**
- Mota, Cesarani, and VanRaden. 2022. Comparison of single-step and multi-step evaluations for U.S. milk, fat, and protein. Proc. 12th World Congr. Genet. Appl. Livest. Prod.
- Cesarani, Lourenco, VanRaden, Nicolazzi, Legarra, Tsuruta, and Misztal. 2022. Options for evaluating multiple breeds in a single-step GBLUP for US dairy population. Proc. 12th World Congr. Genet. Appl. Livest. Prod.
- Tabet, Legarra et al. 2025. All-breed single-step genomic best linear unbiased predictor evaluations for fertility traits in US dairy cattle. J. Dairy Sci. 108:694

Results from 2022 (Mota et al)

Breed	Trait	R^2		R^2		B_1			
		MS	SS	MS+	SS+	MS	SS	MS+	SS+
Holstein	Milk	74	76	80	79	0.95	0.90	1.41	1.00
	Fat	74	81	83	85	0.99	0.96	1.58	1.15
	Protein	60	63	69	68	0.93	0.90	1.52	1.08
Jersey	Milk	78	80	80	81	1.05	0.96	1.47	0.96
	Fat	74	83	79	84	0.99	0.95	1.58	1.09
	Protein	66	72	71	73	0.98	0.93	1.51	1.01

*MS+ and SS+ were multi-step and single-step GEBV with extra regressions on birth year and pedigree EBV
*All tests predicted the final MS GEBV

Single-step conclusions from 2022 (Mota et al)

- Model changes to reduce over-prediction of genetic trend in young bulls GEBVs.
- Intercepts had negative values (over-estimation).
- Future comparisons with foreign data and more traits.
- Results should lead to better understanding and improvements to both SS and MS systems.
- The SS software could simplify and improve multi-trait modeling.
- Switching the current run flow to SS for all traits will take effort and more computation but could be worthwhile.

Include dominance in single-step?

- Pedigree D^{-1} sparse and affordable (Hoeschele and VanRaden, 1991)
  - 9 x 9 matrix links each full sib family to sire-dam interactions
  - Like 3 x 3 matrix linking additive effects to parents
- Calves genotyped early in life
  - Can predict additive + dominant + inbreeding effects
  - Select calves for own merit, not just progeny merit
  - Only doubles number of genomic effects to estimate

Potential improvements in animal models - 1990

- Paul's talk from 4th WCGALP in Scotland 36 years ago
- 1 year after animal model replaced sire model at USDA
- Potential improvements needed (from 1990):
  - Data quantity, quality, and diversity (more traits)
  - Better modeling the means, (co)variances, and distributions
  - Algorithms able to process more data
  - Distributing timely, well-explained rankings and reliabilities

Improve the models:

- Multi-trait with missing data
- Variance adjustments
- Within-herd environmental factors
- Genetic by environmental interactions
- Inbreeding, dominance, epistasis
- Nonlinear models for non-normal traits
- Longitudinal (test day, daily, or continuous) models
- Flexible, user-friendly program packages

Improve the algorithms:

- Specialized code for very large datasets
- Fewer researchers now know how to code
- Packages require larger teams of programmers
- Adapt to new hardware, parallel processing, etc.
- Process new data without reprocessing all previous
  - Weekly updates
  - Use prior estimates
- (Co)Variance estimation

Improve the delivery:

- Easy to understand, well documented evaluations
- Supporting statistics and data summaries
- Convince breeders that EBVs are accurate (validation)
- Selection indexes and selection goals
- Reduce calculation time from data cutoff to delivery
  - Generation intervals are short
  - Analogy: Hurricane forecasting

Conclusions slide from my talk in 1990

- Two basic techniques are useful for improving predictions:
  - Improve the model where it differs from reality
  - Improve reality where it differs from the model
- Reality can be recorded differently, transformed, or adjusted
- Models can account more fully for G, E, and GxE
- Test genes in balanced or randomized trials across environments
- Improved statistical methods cannot replace large amounts of high-quality data for all traits of economic importance

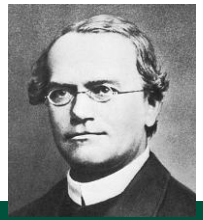
Human genotyping and sequencing (2024)

Technology	Database	People genotyped	Variants genotyped
High-density (HD) SNP chip	Ancestry.com	~25 million	700,000
	23andMe	~15 million	700,000
Whole genome sequence	TopMed (U. Michigan)	160,000	840 million
	Illumina	600,000	500 million
HD imputed to sequence	NIH, Bethesda, MD	58 million	300 million

Total raw data = 600,000 people × 40× coverage × 2.7 billion genome length) = 65 million GB

Summary

- The DNA predictions implemented in 2009 have improved rapidly
  - $4 billion benefit to U.S. dairy in first 10 years (Rexroad, 2019)
  - Accuracy improved as datasets expanded, more traits were included
  - More countries are switching from multi-step to single-step software
- Much cooperation between U GA and USDA
- Methods developed for dairy cattle now used for beef, swine, poultry, fish, honeybees, humans, rice, corn, alfalfa, soybeans, and even peas
- Mendel would be pleased



Acknowledgments

- Taxpayers for funding USDA-ARS-AGIL project 8042-31000-002-00, “Improving dairy animals by increasing accuracy of genomic prediction, evaluating new traits, and redefining selection goals”

- AGIL staff for doing this research

- Council on Dairy Cattle Breeding (CDCB) and its industry suppliers for data

Some of my fellow grad students from 1982-87

Student	Later position	Country	Years
Bassam al-Safadi	Head, Biotechnology Research (govt)	Syria	2018-
Sawsan Magid (al-Sharifi)	Minister of Agriculture (govt)	Iraq	2004-5
M'Naouer Djemali	Director of National Gene Bank	Tunisia	2006-12
Hiroshi Takahashi	Manager, Global Pig Farms	Japan	2000-
Sompop Kassumma	Senior Officer, Livestock Dvlp (govt)	Thailand	2002-
Jay Mattison	CEO of NDHIA and Chair of ICAR	USA	2017-21
Jean Bertrand	Dean, Undergrad Studies, Clemson U	USA	2021-23
Mark Boggess	Director, US Meat Animal Res. Center	USA	2018-
Georgios Banos	1st Director, Interbull Centre	Sweden	1992-99
Jack Dekkers	Distinguished Professor, Iowa State	USA	2013-
Chuck Sattler	Vice President-Genetics, Select Sires	USA	2007-

ISU

UWI

Retirement work: Solutions to World Problems

- Went to university libraries a lot on weekends 1978-94
- Picked a subject, studied until no material left to read
- Got ideas beyond theirs
- Hoped someday to write a book “Solutions to World Problems”
- In 1999 started a web site: paulvanraden.com

Thinking, Computing, and Improving Both

https://www.paulvanraden.com/ThinkingAndComputing.htm

Topics
Thinking
Computing
Artificial intelligence
Connections and ideas
Sorting
Paging and memory management
Linked lists

Topics, continued
Parallel processing
Multitasking
Large datasets
Data storage, access, and cost
Probability and Bayes theorem
Interacting with AI
Cause and effect