IT-Solutions for
Animal Production

Imputation of genetic characteristics using deep learning methods

D. Segelke, Lilian Gehrke & Jan Wabbersen

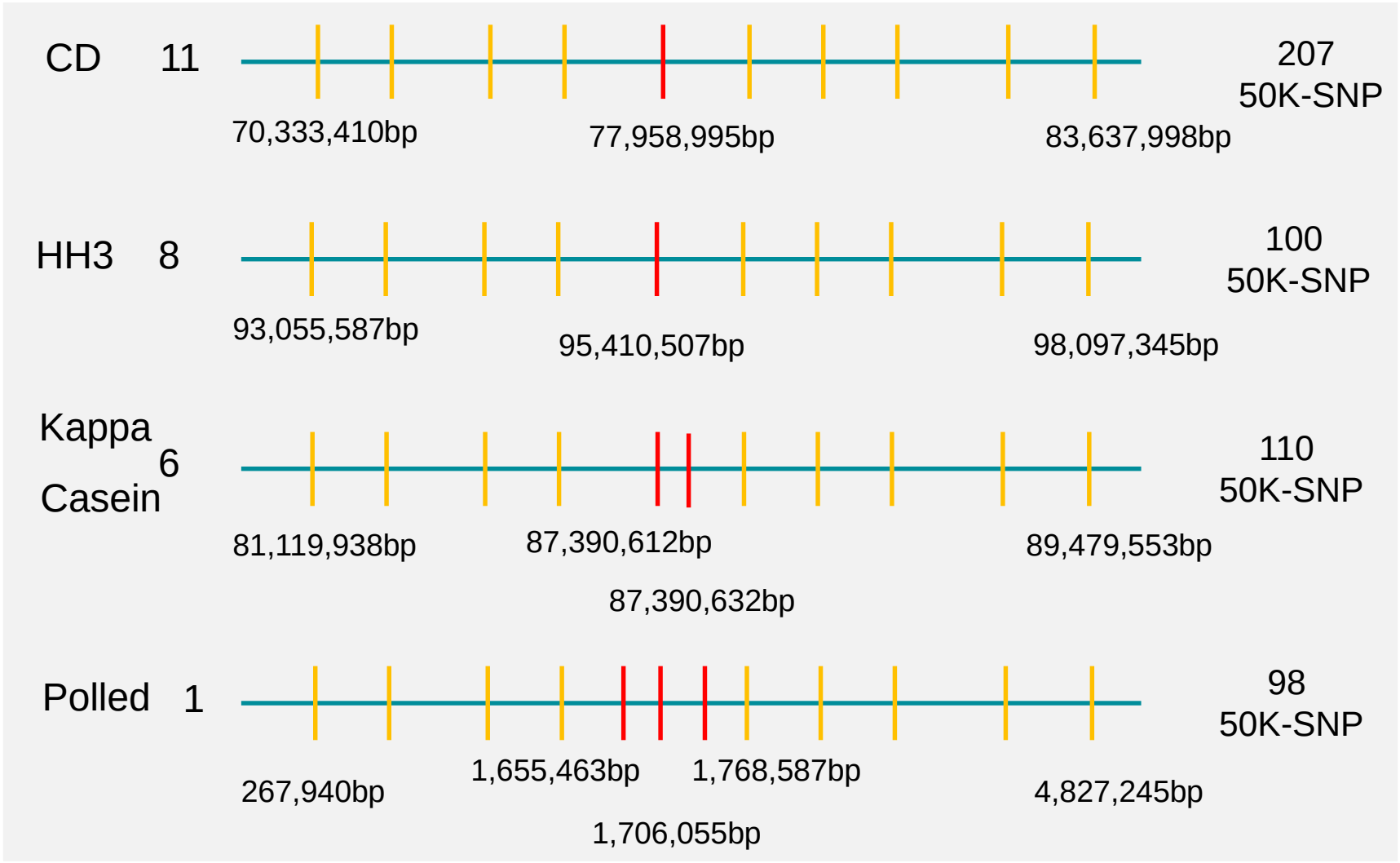
Dierck.Segelke@vit.de

Vereinigte Informationssysteme Tierhaltung w.V. (**vit**), Verden/Germany

Background

- Imputation is well understood & is used in routine evaluation
  - Imputation of different chip densities to a common size
  - Imputation of genetic traits
- Two different approaches are currently used
  - Pedigree based imputation (e.g., Findhap or FImpute)
  - Population based imputation (e.g, Beagle)
- In times of artificial intelligence, deep learning & machine learning methods are becoming more and more popular
  - Sometimes give outstanding results, in contrast to “traditional models”
- Aim of the study: Investigate the imputation accuracy for genetic characteristics using deep/machine learning methods

Materials & Methods
Genetic characteristics



Materials & Methods
SNP data

- Number of animals per chip (polled)

Chip	# SNPs on chip	# animals for training	# animals for validation
EuroG10K V4	11,490	34,632	-
EuroG10K V5	13,787	130,637	-
EuroG10K V7	13,329	164,418	-
EuroG10K V8	13,674	76,126	16,739
EuroG MD	49,331	5,259	16,588

- Imputation from LD to 50K done with FImpute

Materials & Methods
Datasets

	CD	Kappa Casein	HH3	Polled
No. of training animals	242,600	428,974	406,867	406,250
No. of validation animals (born 2019)	33,292	33,873	33,275	33,289
Minor allele frequency training (%)	2.40	34.84	2.52	4.88
Minor allele frequency validation (%)	1.80	39.89	1.91	7.07

Materials & Methods

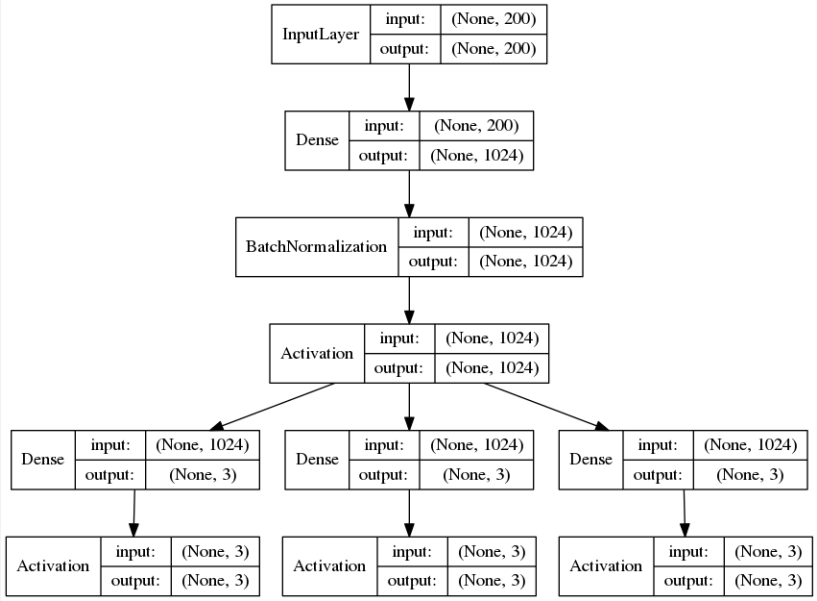
Development of allele frequencies for the different traits

Materials & Methods
Frameworks

- Beagle
  - Genetic optimized algorithm using population based information
  - Version 1398
  - 20 imputation & phase-Iterations
  - 20 threads

- Keras
  - Tensorflow backend
  - Callbacks:
    - Early stopping
    - Checkpoint
  - 20 threads

Keras model plot for polled



Frameworks

- LightGBM
  - fast gradient boosting decision tree algorithm
  - max. 20,000 weak learners (boosted trees) with early stopping
  - learning_rate=0.02
  - num_leaves=8 (max. number of leaves for a tree)
  - colsample_bytree=0.3: ratio of used features for each tree, e.g., to reduce overfitting
  - 20 threads
- Ensemble
  - y_weighted_pred = (0.5 * y_pred_lgbm) + (0.5 * y_pred_keras)
- **Measure of Accuracy:** Correlation between imputed and true genotypes

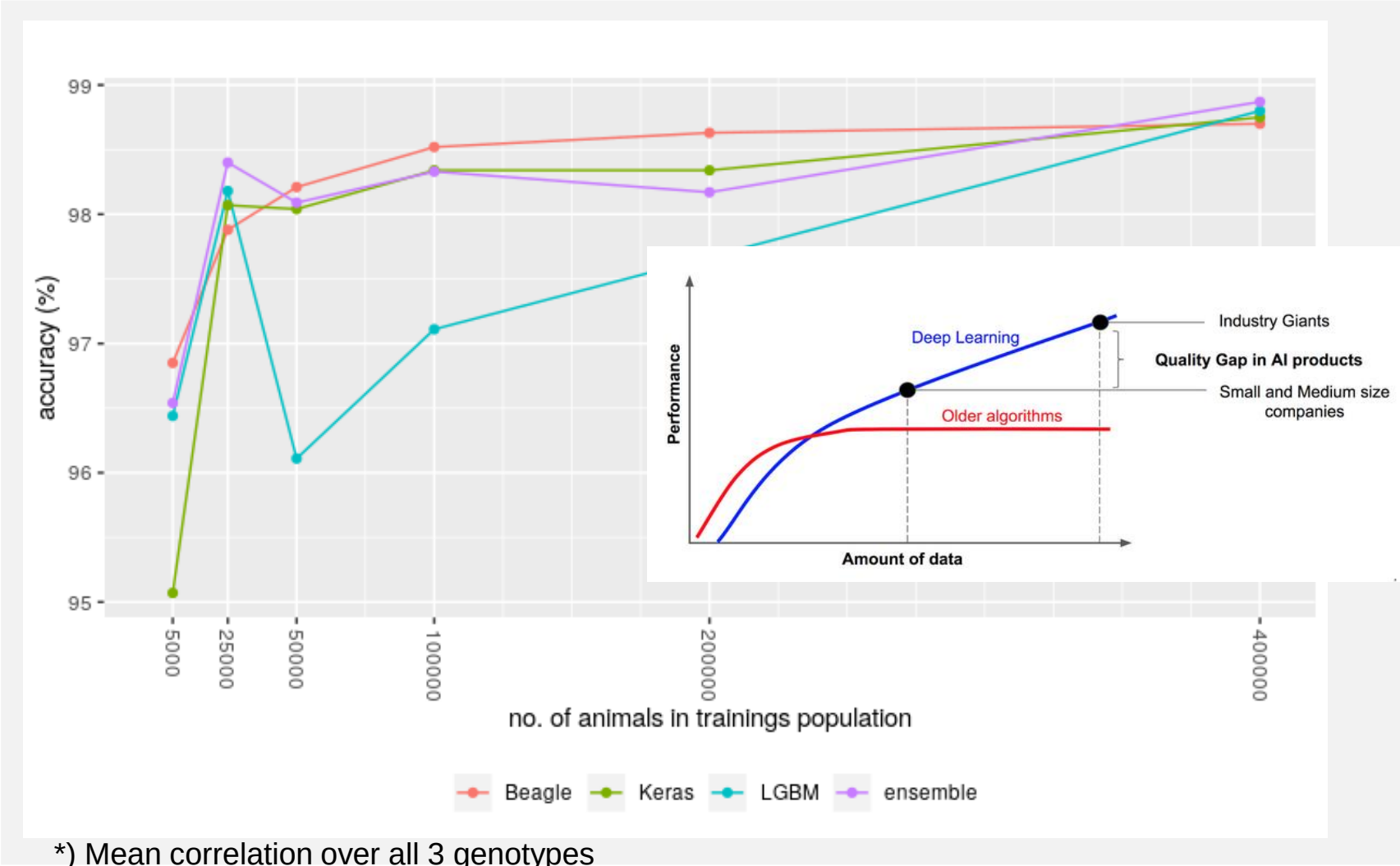
Results

Computation time and accuracy for different traits and methods

Trait	Computation time (h)				Accuracy (%)			
	Beagle	Keras	LGBM	Ens.	Beagle	Keras	LGBM	Ens.
Polled	3:39	0:03	0:02	0:05	98.70	98.75	98.80	98.87
CD	8:41	0:03	0:02	0:05	94.90	96.51	96.73	97.14
HH3	4:05	0:02	0:01	0:03	98.86	99.10	99.35	99.47
Kappa Casein	6:15	0:03	0:08	0:11	99.58	99.55	99.58	99.60

Results

Relationship between accuracy & size of the training dataset (polled)



*) Mean correlation over all 3 genotypes

Results

Accuracy of validation by their relationship to the reference population (polled)

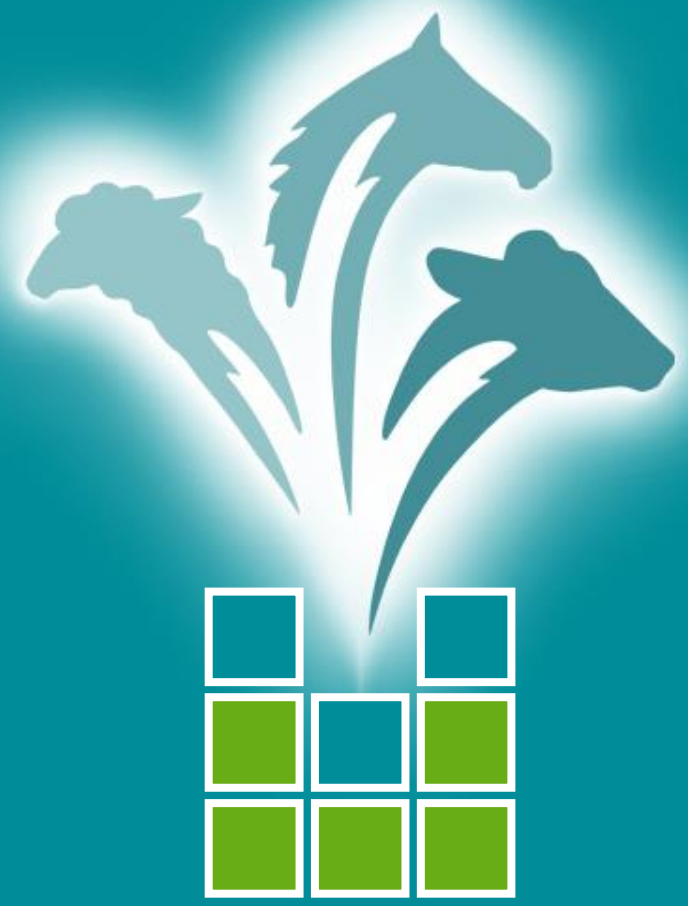
Presence of relatives in training population	Beagle	Keras	LGBM
Sire & dam (n=10,035)	99.28	99.58	99.00
Only dam (n=16,764)	99.22	99.32	99.24
Only sire (n=19,136)	99.10	99.44	99.02
Neiter sire nor dam (n=16,525)	98.36	98.30	98.64
All (n= 33,289)	98.70	98.75	98.80

Conclusion

- Accuracy improved by using deep or machine learning algorithms instead of Beagle
- Computation time decreased drastically
- Combination of lightGBM & keras had the highest accuracy
- Large data sizes are needed to outperform existing methods
- Close relatives in training population is important for all frameworks

Outlook

- Deep and machine learning frameworks have great potential for animal breeding
  - Imputation
  - New phenotypes
    - Sensor data
    - Images
  - MIR Spectra analysis
  - Data anomaly detection (plausability data checks)
- Limit potential for breeding value estimation
  - Linearity



Thank you for attendance!