

Machine Learning in Genetic Evaluations

Nathan Blair
nathan.blair@uscdcb.com



Hi! I'm Nathan.



- Proud husband and girl Dad
- Live in the Denver area
- Computer Nerd 🤪
- Pretty new to dairy, not a geneticist
- CTO at CDCB

What we're going to talk about

- How we're using AI at CDCB
 - AI in this context stands for "Artificial Intelligence."
- How AI differs from traditional computer programming

How did we get here?

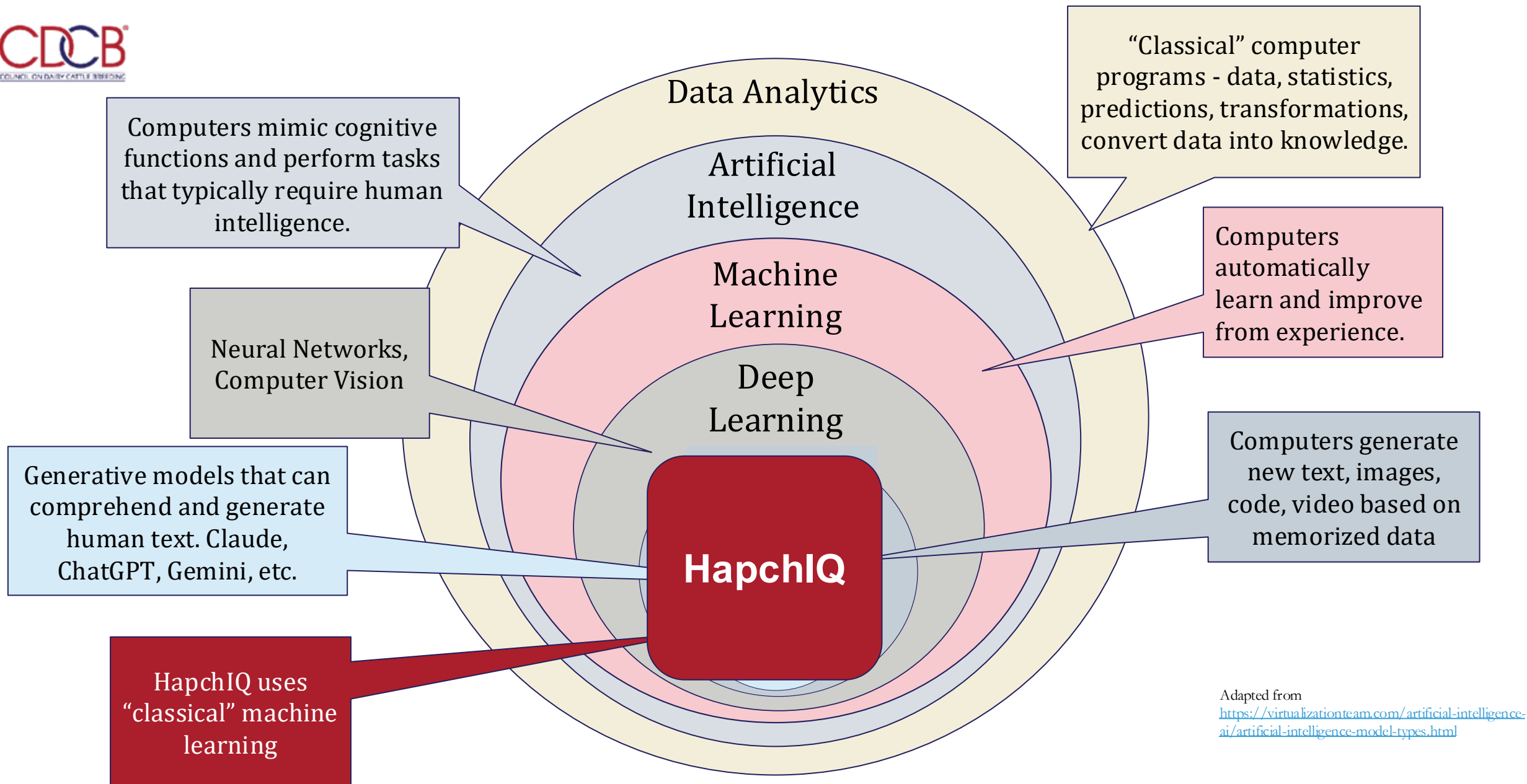
- CDCB has a program, originally from USDA/AGIL, called **hapcheck**. Hapcheck takes SNP genotypes and finds animals that are carriers of particular haplotypes.
 - Holstein Haplotype 1 (HH1), Jersey Neuropathy with Splayed Forelimbs (JNS), etc.
 - Uses the methods described in VanRaden et al. (2011).
- We wanted to incorporate newer haplotypes, but data and programming problems made that difficult in hapcheck.

How did we get here?

- So, we built a program called “**HapchIQ**” (pronounced *hap-SHEEK*) to replace **hapcheck**.
- **HapchIQ** uses the same input data and produces the same output data as hapcheck, but does it in an entirely different way.
 - Not sensitive to the genotype sequence
 - Not specific to any haplotypes

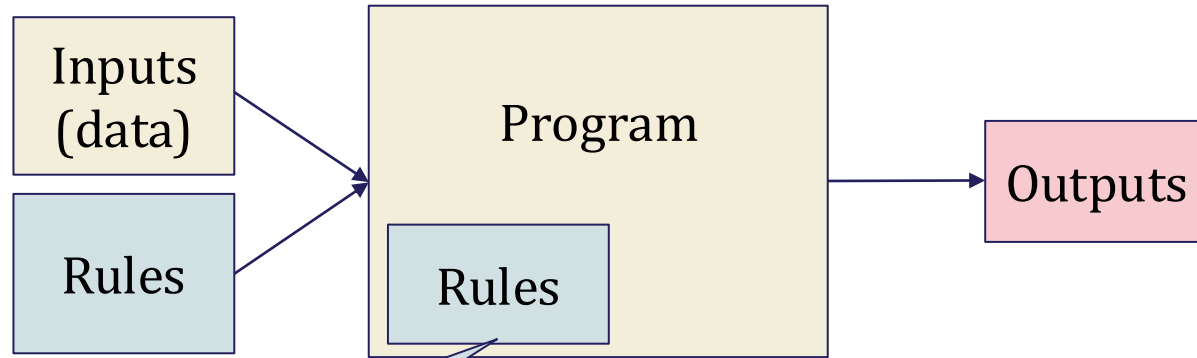
AI Landscape

AI is a very big topic, so let's level set on where CDCB is currently having success with it in our evaluation system.



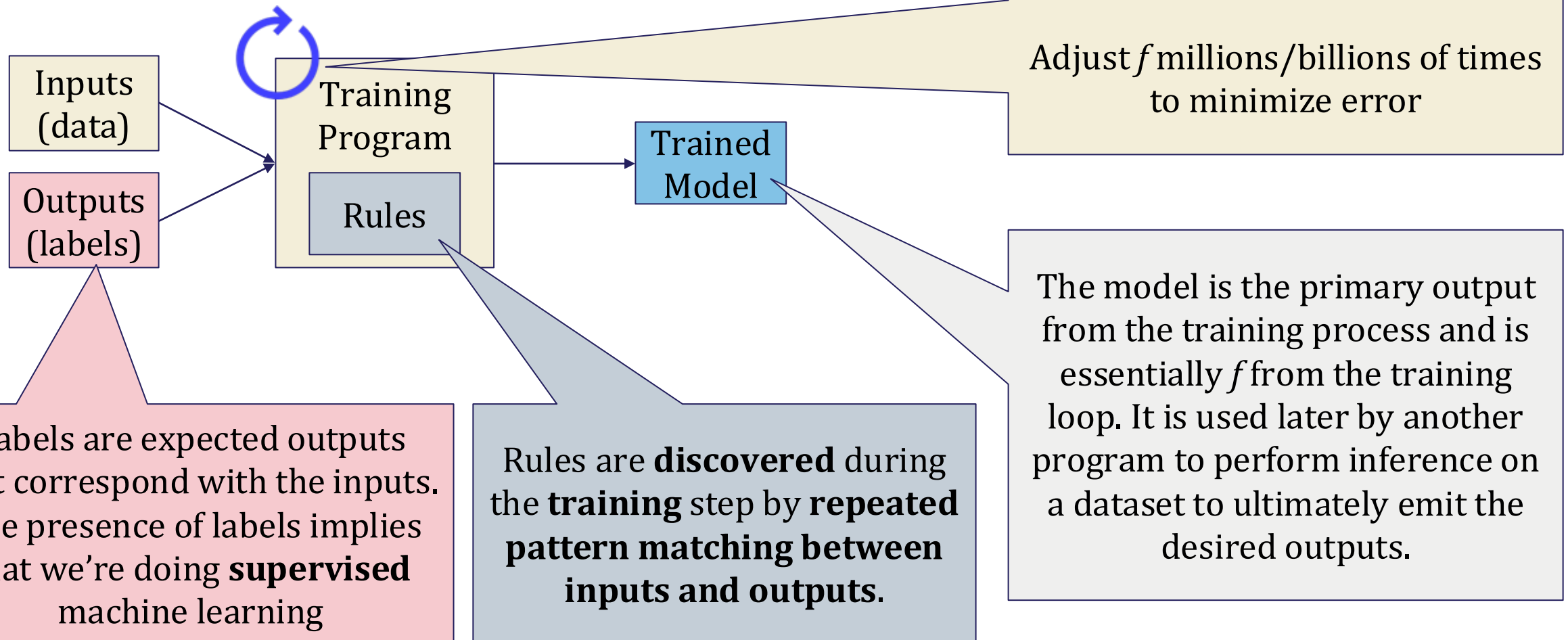
Adapted from
<https://virtualizationteam.com/artificial-intelligence-ai/artificial-intelligence-model-types.html>

Traditional Data Processing & Analytics

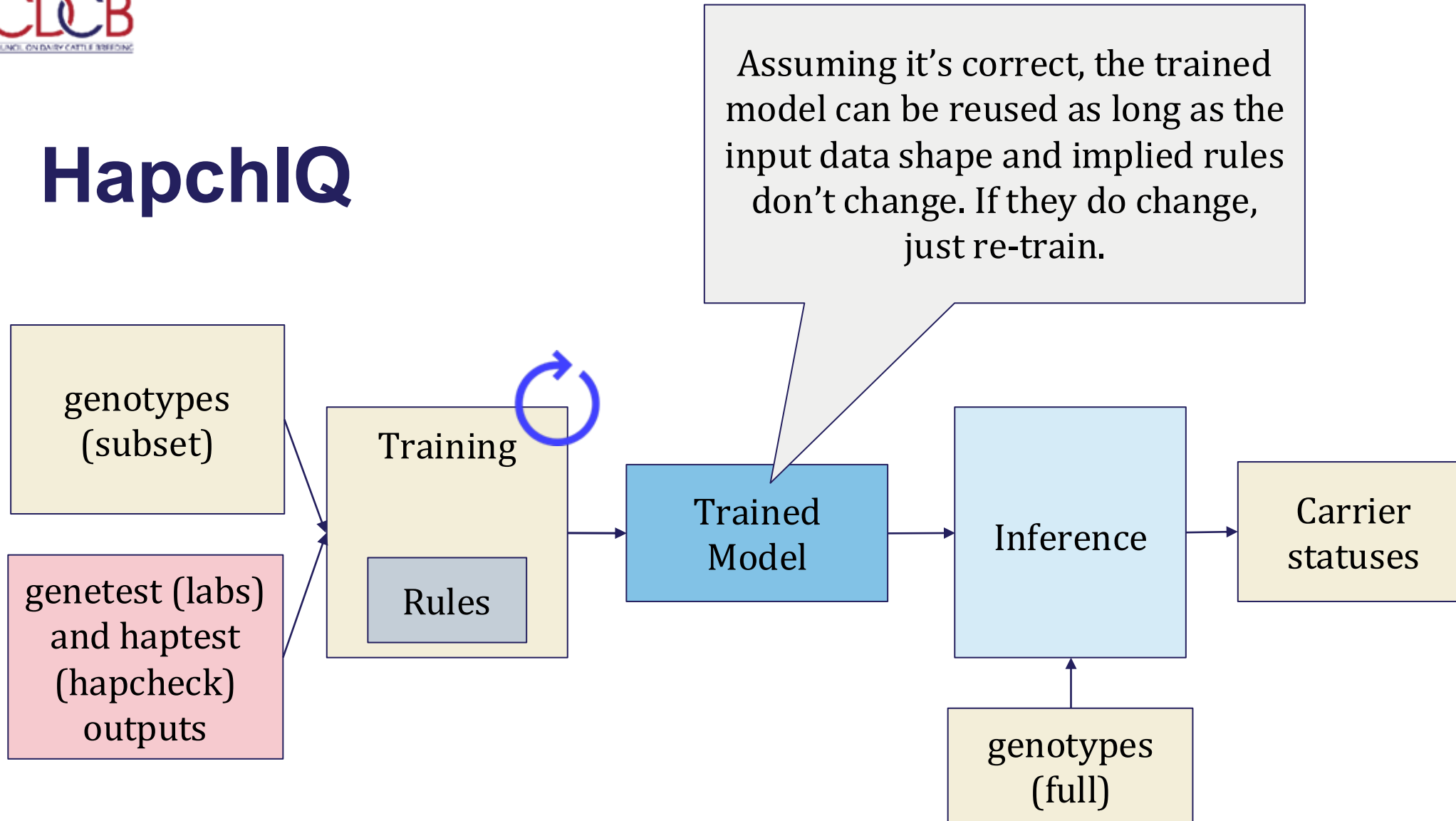


Rules are added, often over time, by software developers, and become part of the program.

Machine Learning



HapchIQ



How well did **HapchIQ** work?

- High (99.5+%) concordance with hapcheck on most haplotypes.
- Let us use our existing data for a new haplotype (BH6) that's not directly represented in the SNP list.
- **Better than hapcheck for polled haplotypes.**
- Some difficulties with Holstein Muscle Weakness (HMW)
 - May not work as well for haplotypes with incomplete penetrance.
 - We're exploring multi-chromosome inputs and hyperparameter tuning for this.

First application: Brown Swiss Haplotype 6 (BH6)

- Affected embryos die, causing fertility reductions.
- The causal allele is located on chromosome 2.
- It's not in our SNP list, but we can still detect it via linkage disequilibrium (LD) because it falls between 2 SNPs we do track.



5365	BTA-45265-no-rs	86065338
	BH6	86191230
5366	Hapmap49925-BTA-24427	86305851

Concordance of BH6 calls from Hapcheck and HapchIQ with gene test results

Program	SNP list	Concordance
Hapcheck	Official	46.5%
Hapcheck	Modified ¹	92.2%
HapchIQ	Official	90.7%

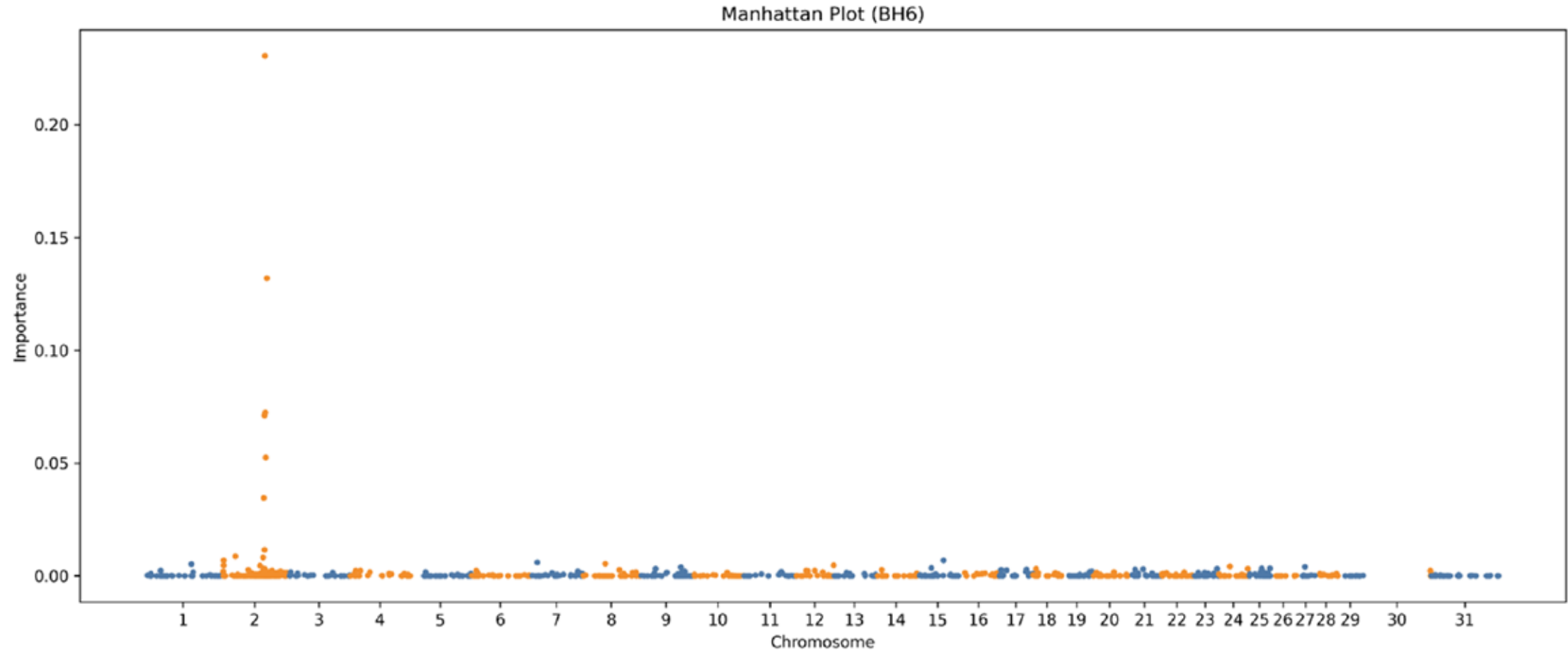
¹ Addition of new DNA markers to the official SNP list requires that all genotypes be re-imputed. This was done only for Brown Swiss animals for internal testing

Run-to-run variation versus training set-to-training set variation

					Changes in Status	
Training Set 1	Evaluation Run 1	Training Set 2	Evaluation Run 2	Number of Animals	Non-carrier to carrier	Carrier to non-carrier
December 2024	December 2024	—	April 2025	82,553	3	0
December 2024	April 2025	—	August 2025	83,829	1	1
December 2024	August 2025	December 2025	December 2025	90,062	78	46

This shows that a pre-trained model seems to have good longevity

There's no SNP for BH6 - what DID it find?



Polled haplotypes

- Multiple mutations produce the polled haplotype.
- Hapcheck worked well for Holstein but not Brown Swiss or Jersey
- **HapchIQ** does much better with the official SNP list.
- When additional polled SNPs are added to the data, the methods are comparable.
 - But we don't want to do that! Modifying the SNP list is difficult.

Breed	SNP list	Program	Matches	Mismatches	Check	Concordance
BS	Official	Hapcheck	17,679	3,047	20,726	85.3%
BS	Official	SNP	16,531	2,949	19,480	84.9%
BS	Official	Hapchiq	20,697	29	20,726	99.9%
BS	SNP name fix	Hapcheck	20,661	65	20,726	99.7%
BS	SNP name fix	SNP	20,721	5	20,726	100.0%
BS	SNP name fix	Hapchiq	20,723	3	20,726	100.0%
JE	Official	Hapcheck	17,559	9,928	27,487	63.9%
JE	Official	SNP	17,643	9,671	27,314	64.6%
JE	Official	Hapchiq	27,231	256	27,487	99.1%
JE	SNP name fix	Hapcheck	26,784	671	27,455	97.6%
JE	SNP name fix	SNP	27,452	33	27,485	99.9%
JE	SNP name fix	Hapchiq	27,470	15	27,485	99.9%
HO	Official	Hapcheck	55,943	2,714	58,657	95.4%
HO	Official	SNP	55,929	2,667	58,596	95.4%
HO	Official	hapchiq	58,248	409	58,657	99.3%
HO	SNP name fix	Hapcheck	58,079	577	58,656	99.0%
HO	SNP name fix	SNP	58,632	24	58,656	100.0%
HO	SNP name fix	Hapchiq	58,590	66	58,656	99.9%

Other Haplotypes

Comparison of hapcheck and HapchIQ results for Holstein from August 2025

Haplotype	Mismatches	Genotypes	Concordance
Holstein Haplotype 1	1,233	9,088,709	99.99%
Holstein Haplotype 2	5,963	9,089,160	99.93%
Holstein Haplotype 5	2,170	9,088,709	99.98%
Holstein Cholesterol Deficiency	44,410	8,914,901	99.50%
Holstein Muscle Weakness	67,024	8,173,668	99.18%

Comparison of hapcheck and HapchIQ results for other breeds from August 2025

Haplotype	Mismatches	Genotypes	Concordance
Ayrshire Haplotype 2	109	20,448	99.47%
Ayrshire Arthrogryposis Multiplex Congenita	48	20,448	99.97%
Brown Swiss Weaver Syndrome	101	90,081	99.89%
Brown Swiss Spinal Dysmyelination	136	90,081	99.85%
Brown Swiss Haplotype 14	104	90,081	99.88%
Jersey Neuropathy with Splayed Forelimbs	479	930,718	99.95%

Maintainability and Runtime performance

- **HapchIQ** is about half as many lines of code as hapcheck.
 - No changes to the code expected for new haplotypes or SNPs.
- Training performance on traditional CPU hardware is on par with hapcheck runtimes; inference performance is significantly faster.

Dangers with this approach

- Overfitting - the model memorizes the data instead of learning the patterns.
- The trained model is fairly opaque & not easily understood.
- The data has to be good. Most of the effort is spent on getting the data up to the right quality and format.
 - Consistent
 - No major class imbalances
- You have to be very specific with the questions you're asking of the data with this type of AI.
- Training a model can be computationally intense.

Future plans

- Incorporate pedigrees in addition to genotypes.
- Using more of this type of AI for additional data analysis (not just haplotype calling).
- Using more advanced AI algorithms/approaches for computer vision, code analysis and generation.

Acknowledgements

- Thank you to Dan Null, John Cole, Ezequiel Nicolazzi, and João Dürr for all the help and support on this project!

Thank you!

Questions?

Please feel free to reach out if you have any follow up questions.

nathan.blair@uscdcb.com

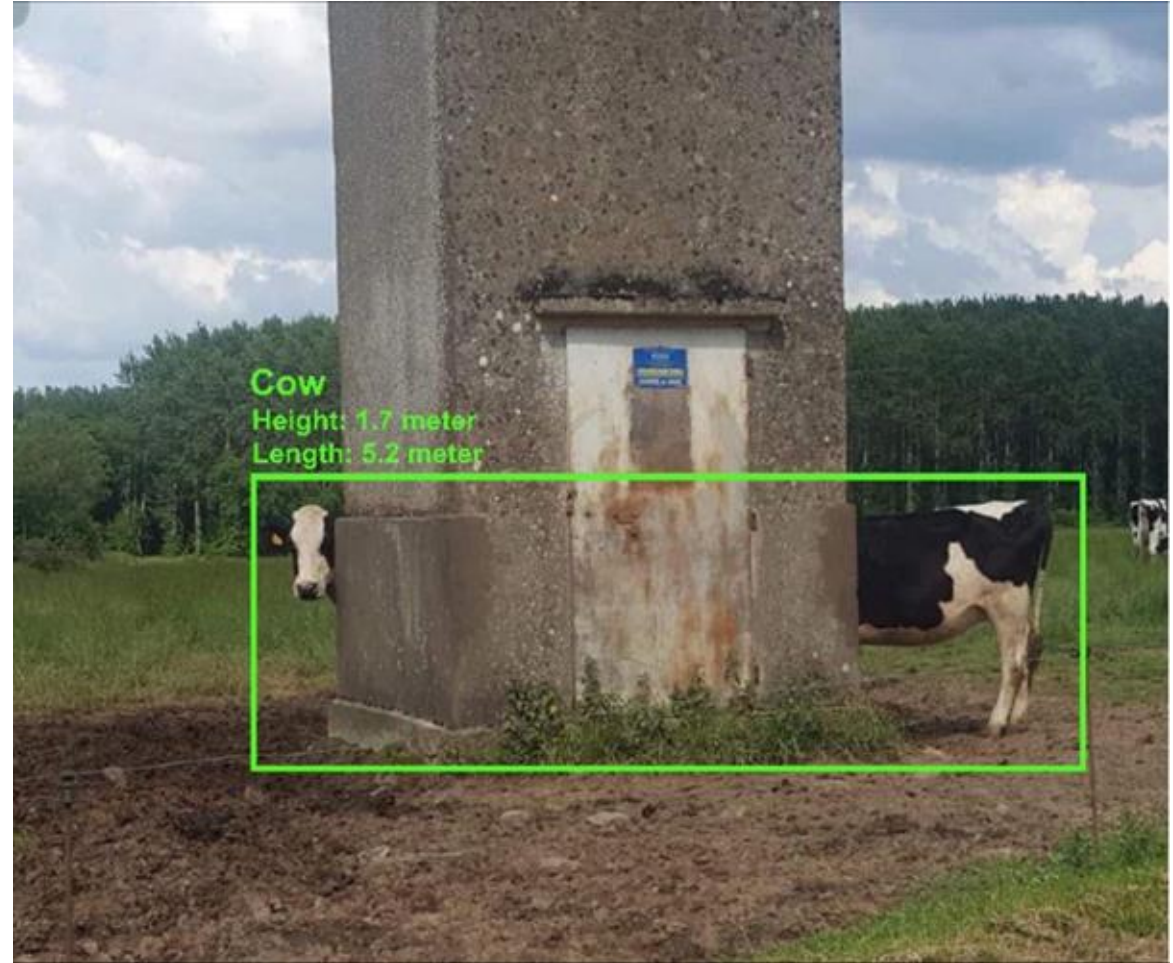


Image credit to [Ralph Aboujaoude Diaz](#)