Predictive Breeding

GHENT UNIVERSITY

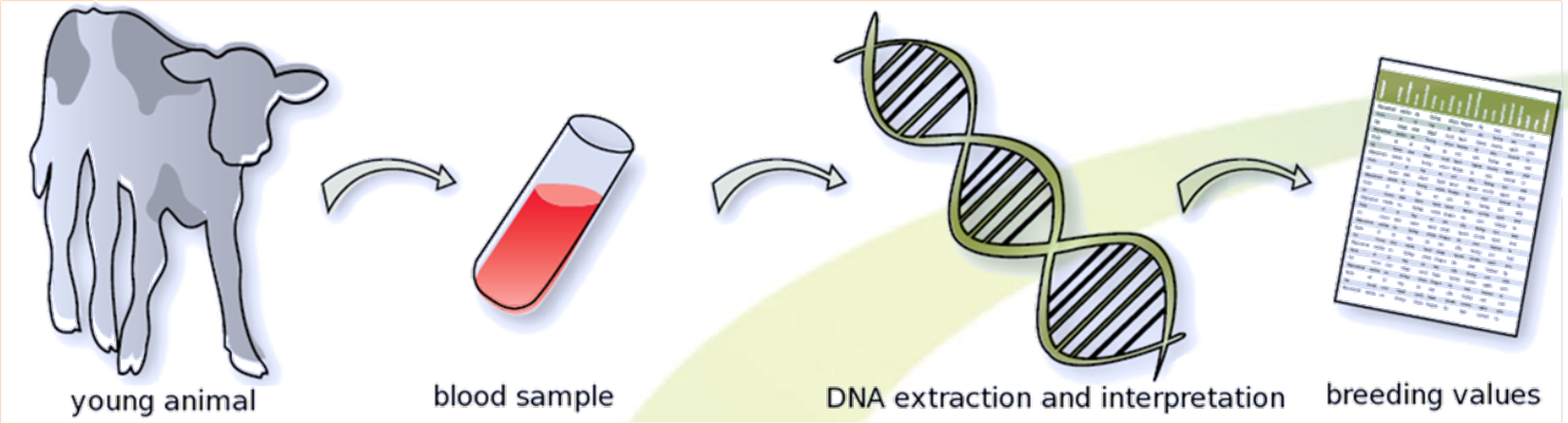
Machine Learning for Genomic Selection,

don't believe everything you think…

Prof. Steven Maenhout

Steven.Maenhout@UGent.be

Genomic selection



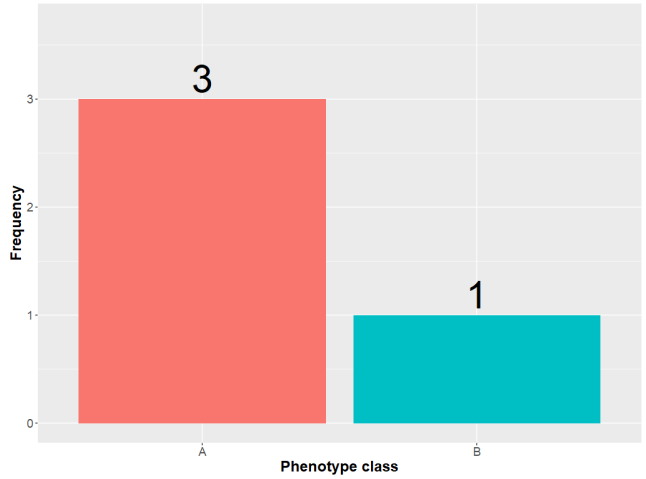
- Prediction of genetic values from numerous variations in the DNA code
- Introduced in landmark paper of Meuwissen et al. in 2001
- Rapid adoption by the animal and plant breeding community. Routinely applied in many plant and animal species

The quantitative geneticists' revenge

GHENT UNIVERSITY

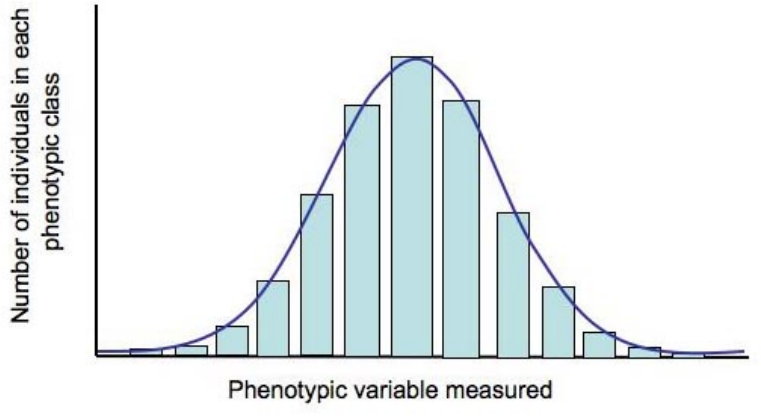
Simple traits:

- small number of genes
- Mendelian inheritance patterns
- limited influence from environment
- QTL mapping, Marker-assisted Selection
- Gene modification / editing

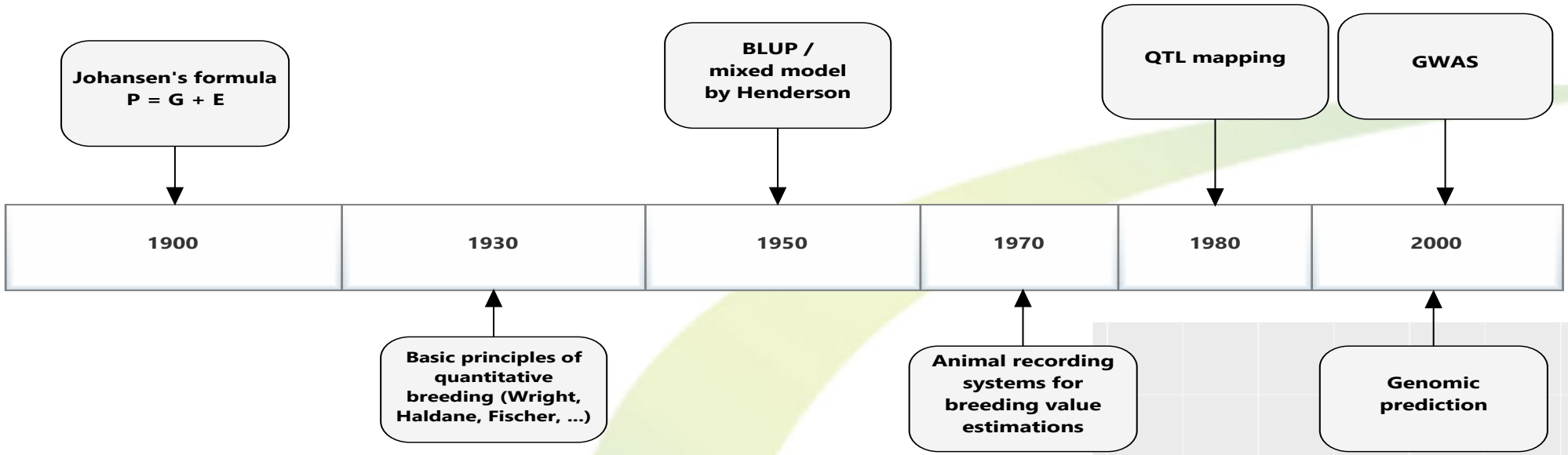


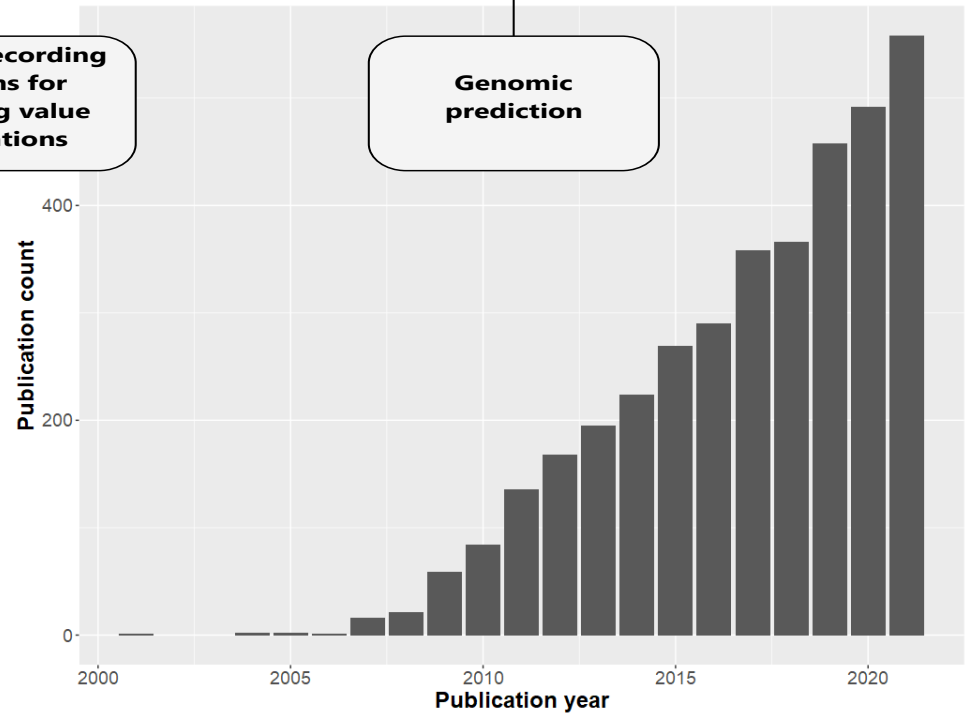
Complex traits:

- large number of genes
- distributional assumptions
- quantifiable influence from environmental
- breeding value estimation
- GWAS, **Genomic Selection**

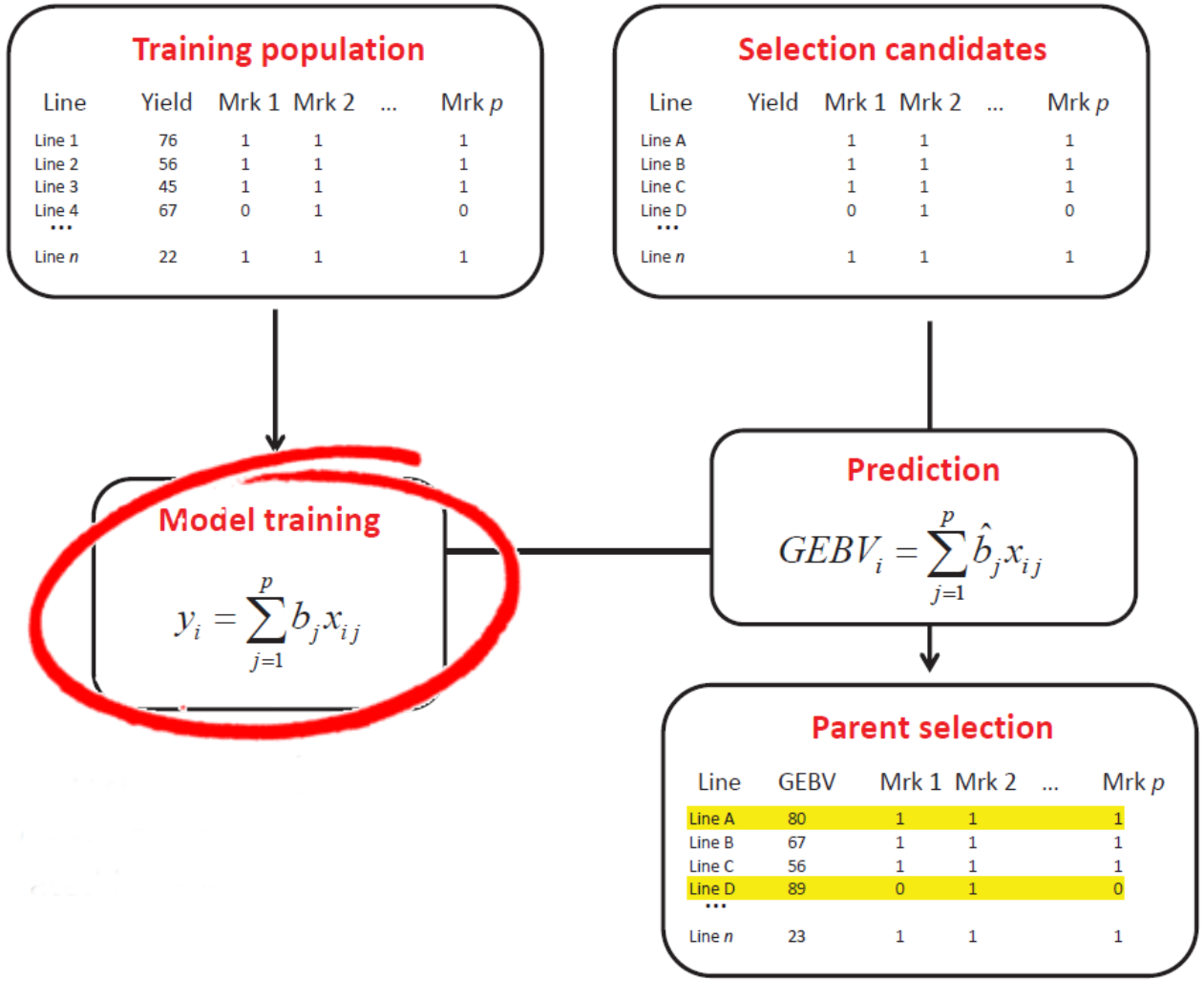


Timeline


Johansen's formula
P = G + E
BLUP /
mixed model
by Henderson
QTL mapping
GWAS
1900
1930
1950
1970
1980
2000
Basic principles of quantitative breeding (Wright, Haldane, Fischer, ...)
Animal recording systems for breeding value estimations
Genomic prediction


400
200
0
Publication count
2000
2005
2010
2015
2020
Publication year

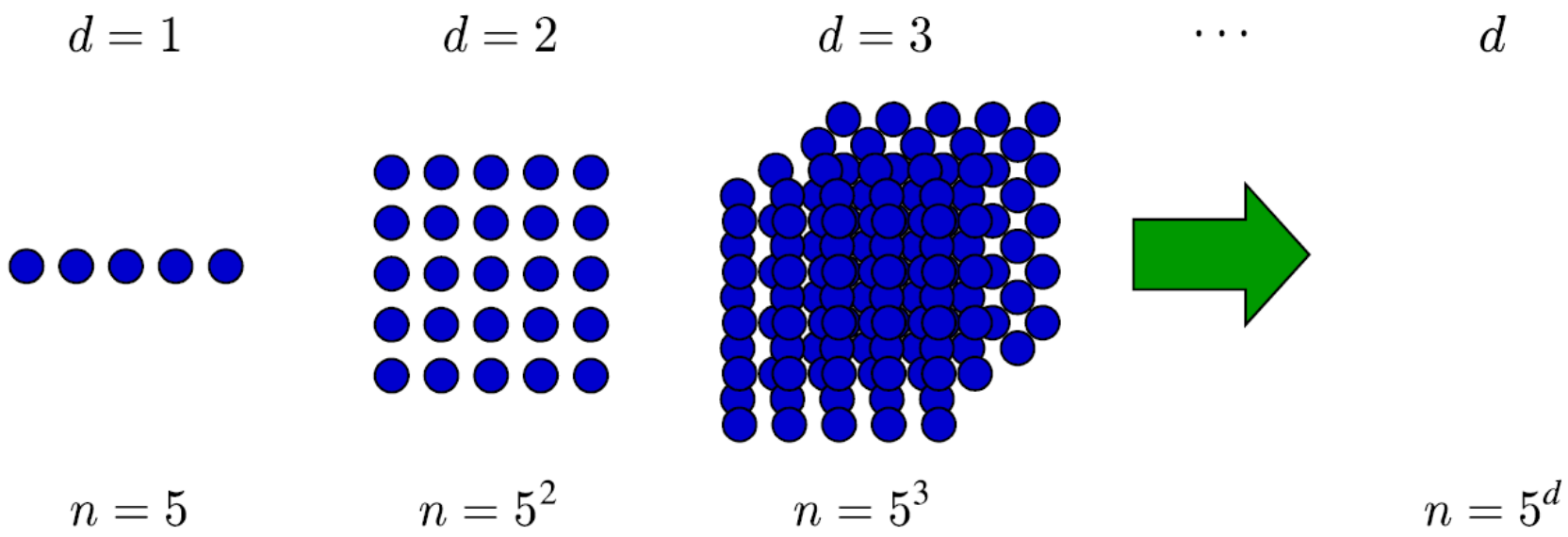
Genomic Selection concept



Escaping the curse of dimensionality

GHENT UNIVERSITY

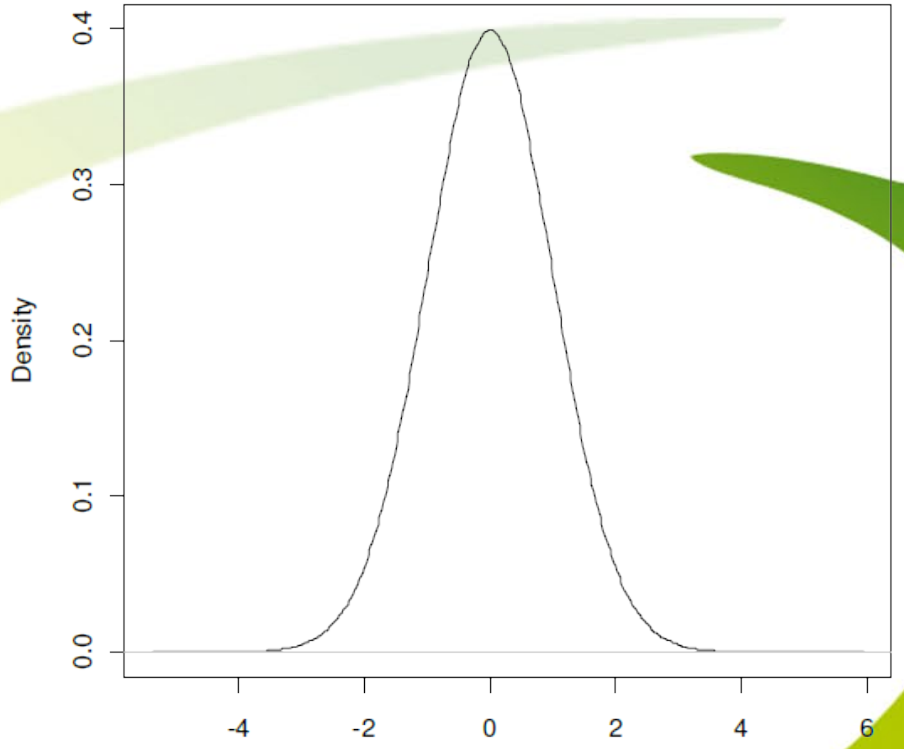
- As the number of dimensions (i.e. molecular markers) grows, the amount of data we need to generalize accurately grows Exponentially



Statistical Modelling: The Two Cultures

GHENT UNIVERSITY

- **1a**) Traditional statistical modelling like the Linear Mixed Model framework: GBLUP, RRBLUP
- data is generated by a stochastic model:
  - additive, linear effect of each marker
  - marker effects adhere to a Gaussian distribution
  - residuals adhere to a Gaussian distribution

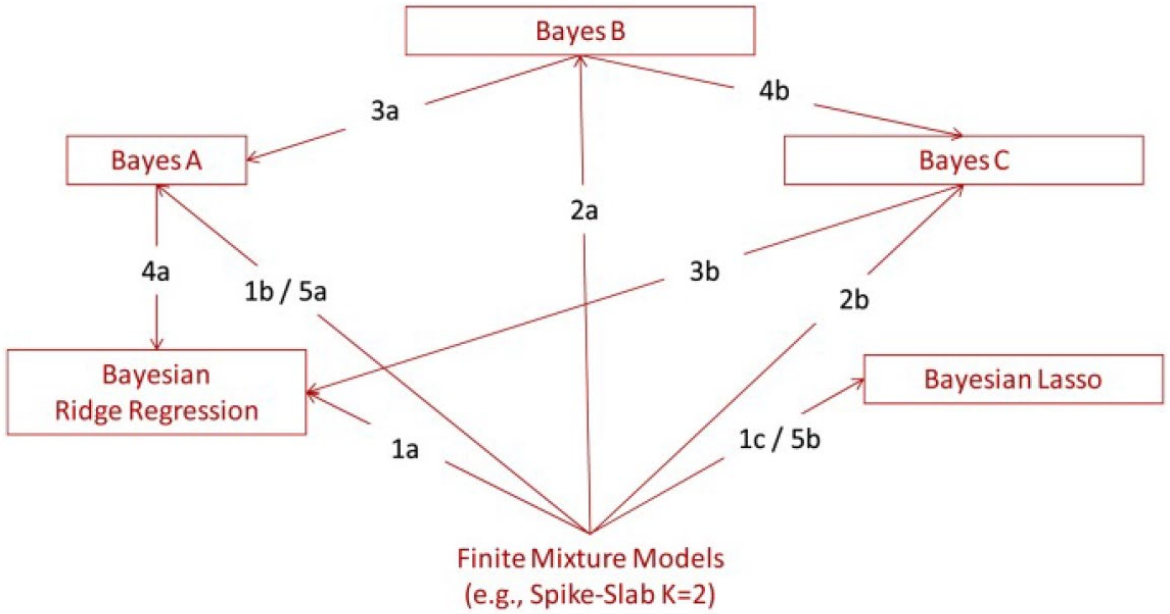


"this commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems.", Breiman 2001

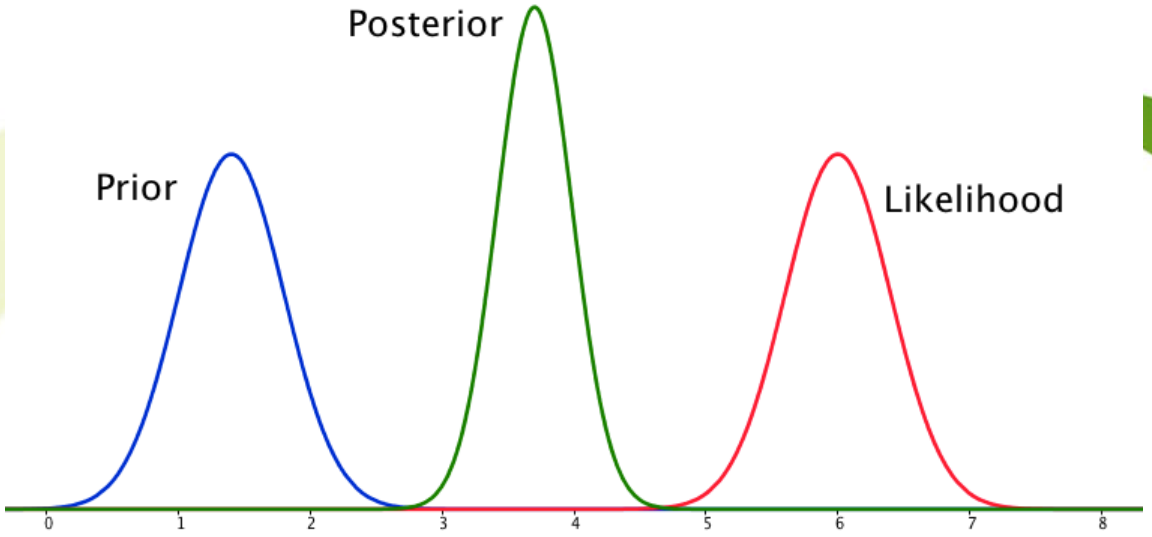
Statistical Modelling: The Two Cultures

GHENT UNIVERSITY

- **1b**) Bayesian models



$$p(\beta_j|\pi_1,...\pi_K,\omega_1,...,\omega_K)=\sum_{k=1}^{K}\pi_j p(\beta_j|\pi_j,\omega_j)$$



"Bayesians address the question everyone is interested in, by using assumptions no-one believes" Lyons, 2007

Statistical Modelling: The Two Cultures

GHENT UNIVERSITY

- 2) Machine learning approaches: uses an algorithm to learn a function, treating the data mechanism as unknown:

$$y = f(X)$$

- fewer assumptions
- Random forests, neural networks, support vector machines, …

ML timeline

1950
Alan Turing created a test to check if a machine could fool a human being into believing it was talking to a machine.

1952
The first computer learning program, a game of checkers, was written by Arthur Samuel.

1957
First neural network for computers (the perceptron) was invented by Frank Rosenblatt, which simulated the thought processes of the human brain.

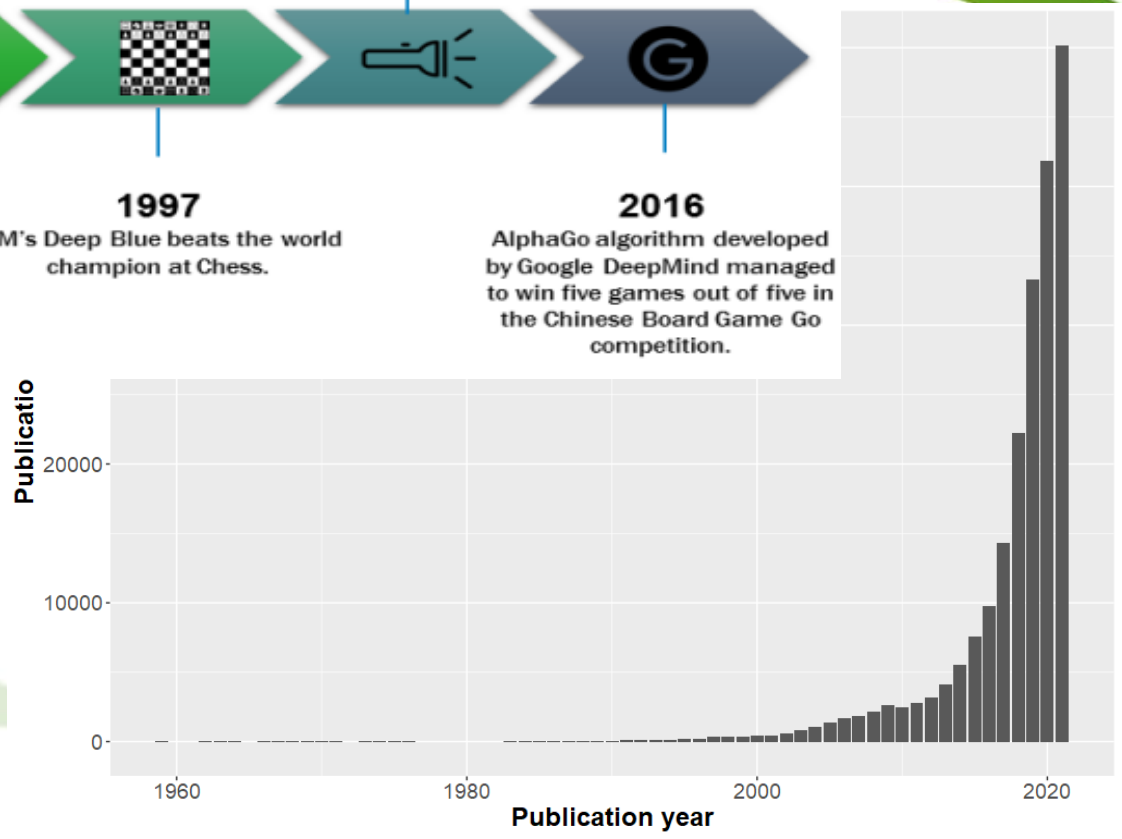
1967
The Nearest Neighbor Algorithm was written.

1979
Students of Stanford University, California, invented the Stanford Cart which could navigate and avoid obstacles on its own.

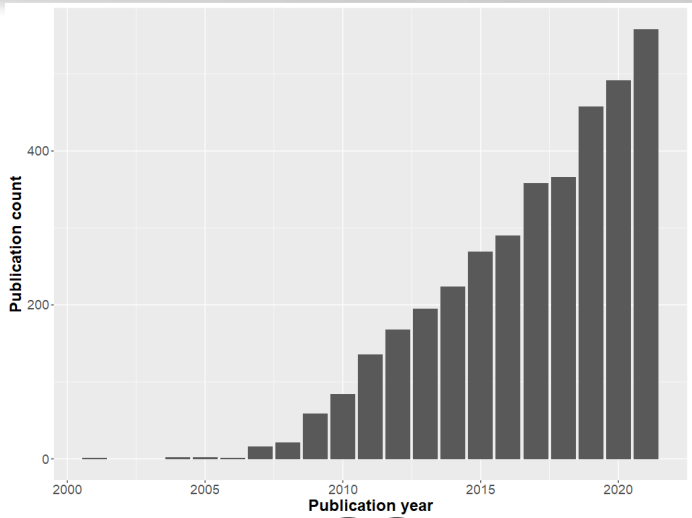
1997
IBM's Deep Blue beats the world champion at Chess.

2002
A software library for Machine Learning, named Torch is first released.

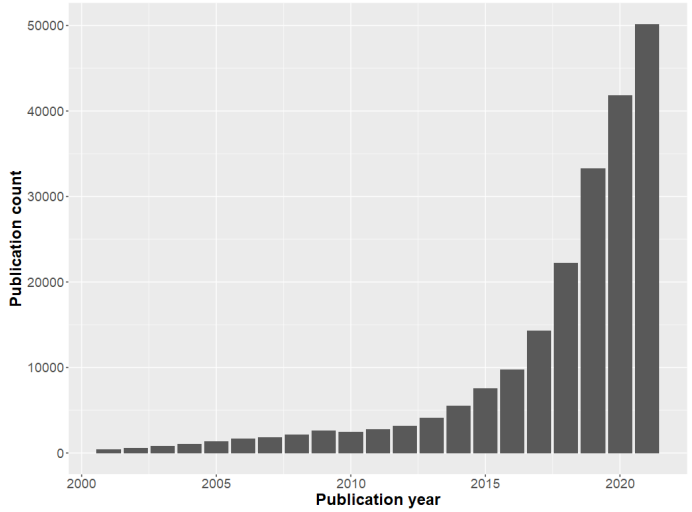
2016
AlphaGo algorithm developed by Google DeepMind managed to win five games out of five in the Chinese Board Game Go competition.



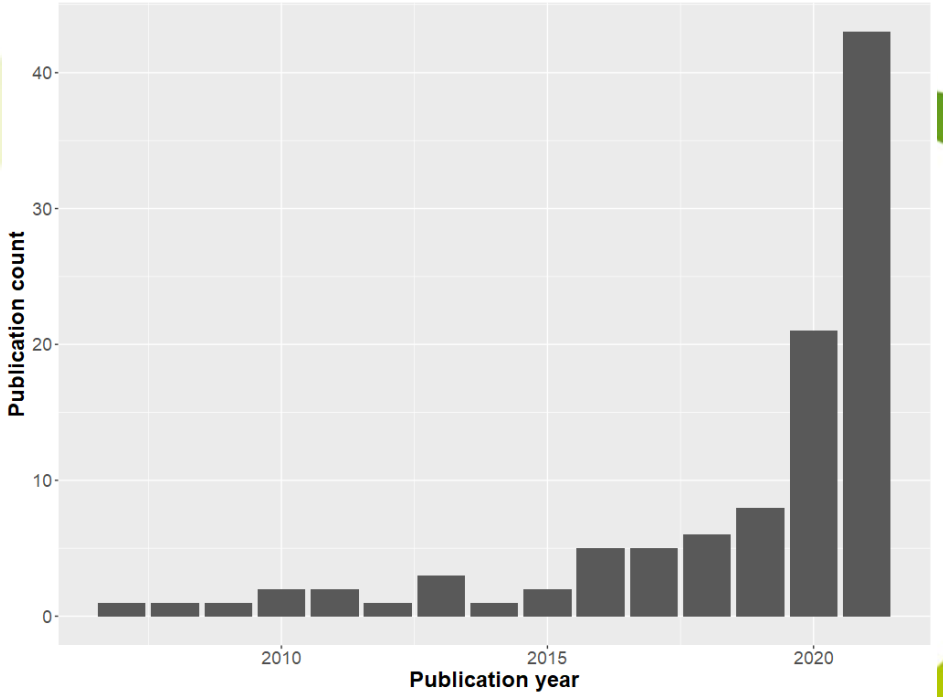
GS + ML: a match made in heaven?



GS



ML



GS + ML

Portrait of a Disappointment

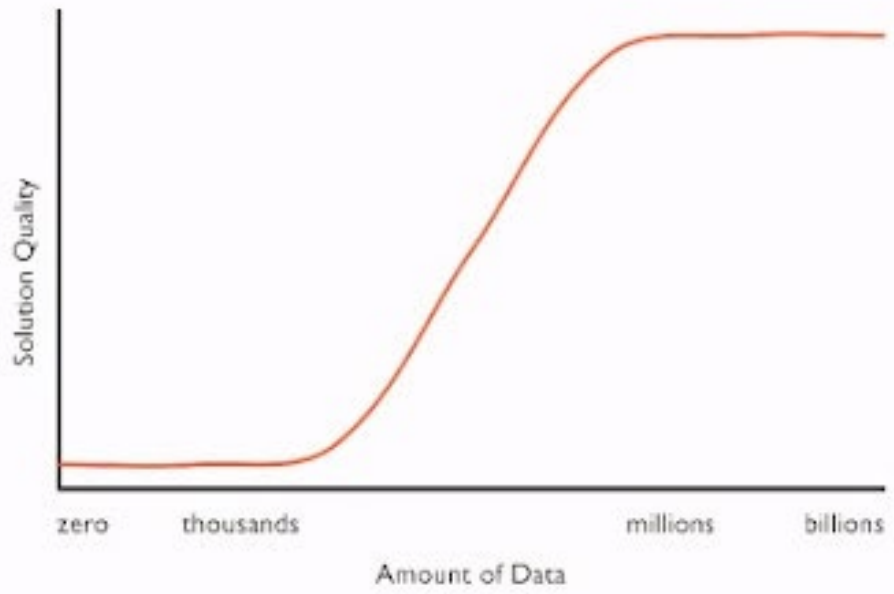
GHENT UNIVERSITY

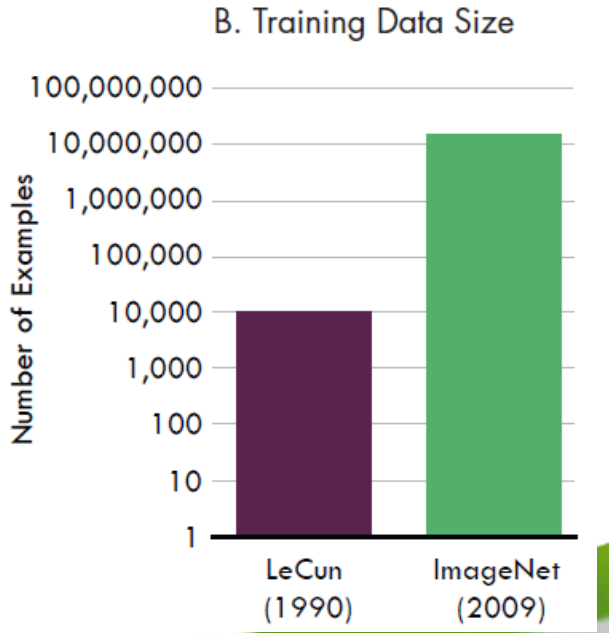
- 2010: Machine learning techniques are generally not able to significantly outperform model-based genomic prediction approaches due to the **limited size of the training populations** in a plant breeding context (Maenhout et al.)
- 2017: Machine learning **slightly outperformed** other methods, but required parameters optimization for GS implementation (Bin Kwong et al.)
- 2018: experimental results indicate that DeepGS **can be used as a complement** to the commonly used RR-BLUP in the prediction of phenotypes from genotypes (Ma et al.)
- 2019: although artificial neural networks **did not perform best for any trait**, we identified strategies that boosted their performance to near the level of other algorithms
- 2020: CNNGWP provides a **promising approach** for GWP, but the magnitude of improvement depends on the genetic architecture and the heritability (Waldmann et al.)
- 2021: DL models gave **0 to 5% higher prediction accuracy** than rrBLUP model under both cross and independent validations for all five traits used in this study (Sandhu et al.)

The Unreasonable Effectiveness of Data

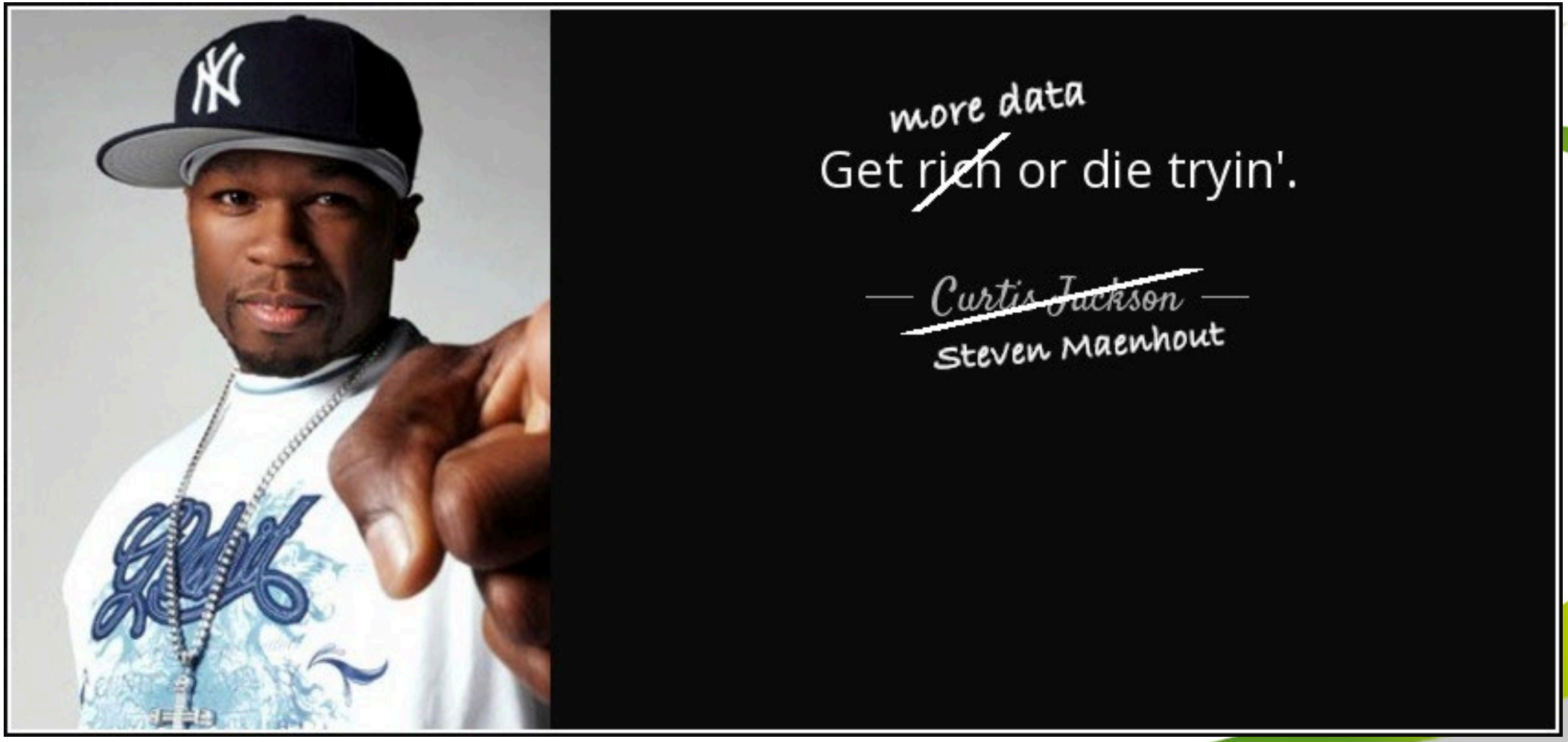
GHENT UNIVERSITY

- *"simple models and a lot of data trump more elaborate models based on less data", Halevy et al., 2009*
- "*performance of the model increases logarithmically as the training dataset increases*", Sun et al. 2017
- "*while a tremendous amount of time is spent on engineering and parameter sweeps; little to no time has been spent collectively on data*", Sun et al. 2017



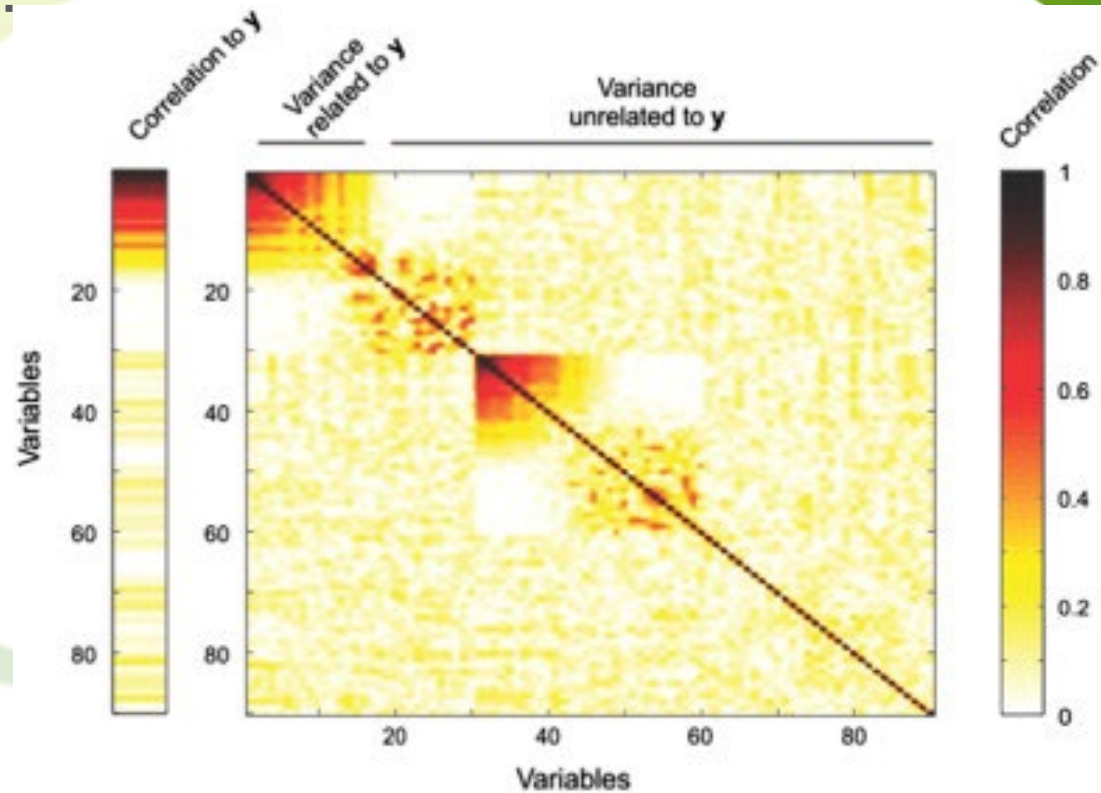


Addressing the elephant in the room



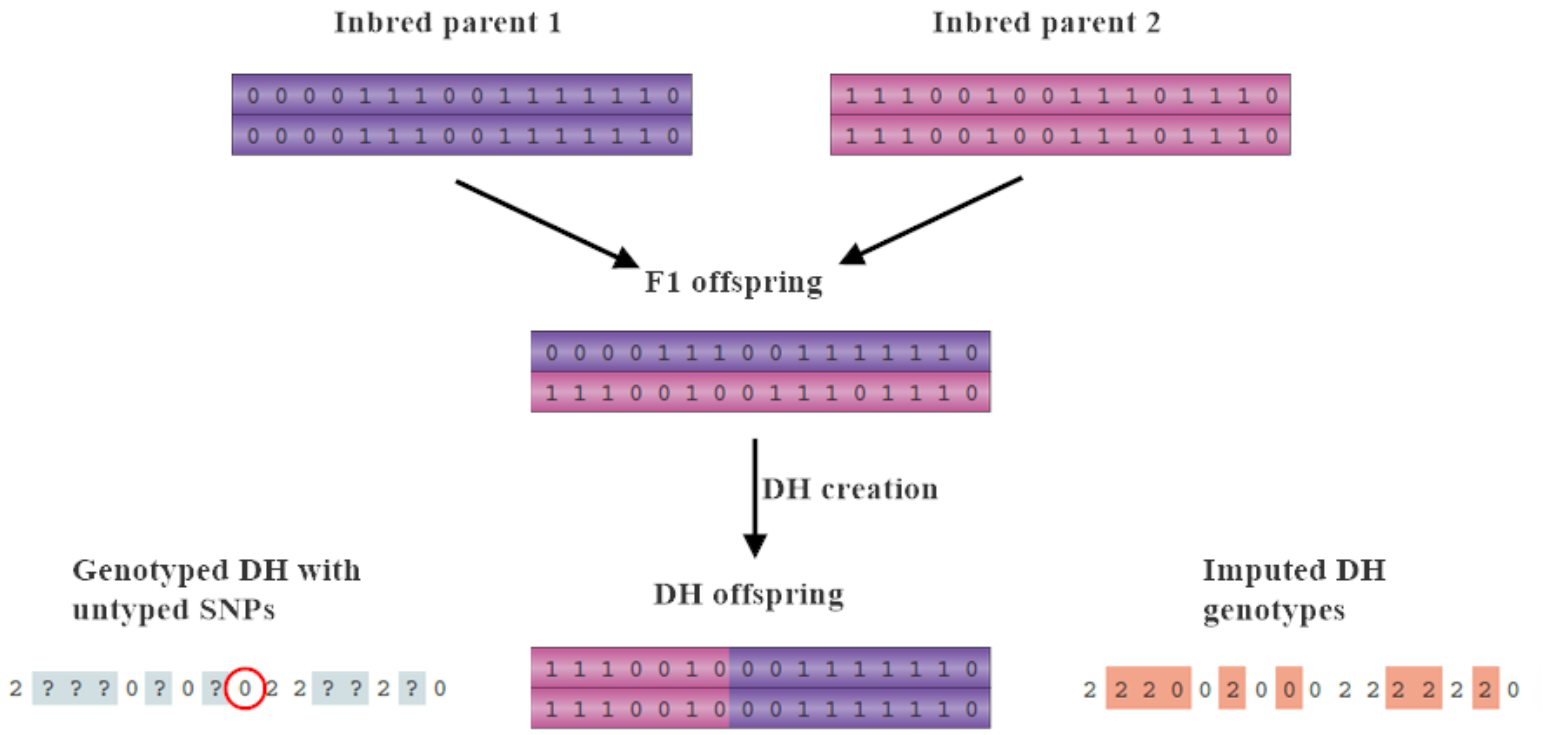
Phenotyping bottleneck

- Remote sensing technologies: UAV, UGV, RGB, NIR, LiDAR, MRI
- Functional structural plant models (FSPMs)
- Estimate selection trait from proxy traits:
  - (Kernel)-PLS
  - Random forests
  - multivariate genomic prediction



Genotyping bottleneck

- Low-cost genotyping:
  - **ultra low-density** genotyping complemented by imputation
  - **skim sequencing** complemented by imputation



Business model

- Genomic Prediction As A Service (GPAAS):
- consortium members provide a combination of
  - plant material
  - genotypic data
  - extracted DNA samples
  - standardized phenotypic data
- on-line portal providing multi-trait genomic predictions of their breeding pool

Predictive Breeding

GHENT UNIVERSITY

Questions, datasets, research topics, copyright infringement claims?

Steven.Maenhout@UGent.be