doi:10.20944/preprints201903.0115.v1

Review

On to the next chapter for crop breeding: Convergence with data science

Elhan S. Ersoz [1,2,*], Nicolas F. Martin [2], and Ann E. Stapleton [3]

[1] Umbrella Genetics, 2501 Woodridge Road, Champaign, IL 61822; elhan@umbrellagenetics.org
[2] Department of Crop Sciences, University of Illinois, Urbana-Champaign, 61820; nfmartin@illinois.edu
[3] Department of Biology and Marine Biology, University of North Carolina Wilmington,601 S. College, Wilmington, NC 28401; stapletona@uncw.edu
* Correspondence: elhan@umbrellagenetics.org; Tel.: +1-515-346-2259

Version March 6, 2019 submitted to Crop Science

Abstract: Crop breeding is as ancient as the invention of cultivation. In essence, the objective of crop breeding is to improve plant fitness under human cultivation conditions, making crops more productive while maintaining consistency in life cycle and quality. The applications of predictive breeding has been gaining momentum in agricultural industry and public breeding programs for the last decade, in the aftermath of genomic selection being recognized and widely applied for accelerating genetic gain in breeding programs. The massive amounts of data that has been generated by industry and farmers year after year through several decades has finally been recognized as an asset. A wide range of analytical methods such as machine learning, deep learning and artificial intelligence that were initially developed for diverse quantitative disciplines are now being adopted to crop breeding decision making processes. New technologies are currently being developed that would enable integration of data from various domains such as geospatial variables and a multitude of phenotypic responses as well as genetic information, in order to identify, develop and improve crop faster via partial or full automation of the decisions that pertain to variety development. Here we will discuss and summarize efforts from public and private domains for predictive analytics, and its applications to crop breeding and agricultural product development, and provide suggestions for future research.

Keywords: machine learning; agroclimactic modelling; crop breeding and genetics; GxE

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
DL	Deep Learning
AI	Artificial Intelligence
GS	Genomic Selection
GEBV	Genomic Estimated Breeding Value
GWAS	Genome-wide Association Study
MET	Multi-environment trials
TPE	Target Population of Environments
QTL	Quantitative Trait Locus
MAS	Marker Assisted Selection
MARS	Marker Assisted recurrent Selection
MAB	Marker Assisted Backcross Introgression
GM	Genetically Modified
GxE	Genetics-by-Environment interactions
MLM	Mixed-Linear Model
AMMI	Additive Main effects and Multiplicative Interaction Model
RM	Relative Maturity
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Networks

1. Introduction: How did we get here?

Although the common wisdom states that the only objective of crop breeding is to improve productivity, i.e. yield – this is clearly an oversimplification. The actual root objective of breeding is to improve cultivate-ability, while increasing productivity (Borlaug, 2007).

Cropping systems vary widely- and are conditional on the available resources and constraints at the time of cultivation. Some of the general features that define the cultivation conditions are geography and climate of the farm; economic, social and political pressures; available technologies and instruments as well as the philosophy and culture of the farmer and the society (Borlaug, 2007). All these constraining features of cultivation are very temporally dynamic, with the exception of geography and climate. Geography and climate are the only features of cropping systems that are stable over time and thus are long-term influencers of crop improvement practices through-out the ages (Bullock, 1992).

Crop improvement practices in reality, require a breeder to address multiples of breeding goals some of which are antagonistic (Kwon and Torrie, 1964; Meredith and Bridge, 1971; Kato and Takeda, 1996; Triboi *et al.*, 2006; Erskine *et al.*, 1985). These goals are frequently related to productivity and cropping systems, as well as post-harvest characteristics and cost-of-goods and economics of seed production systems (Guanming *et al.*, 2010). For instance, for seed crops, there is a known negative genetic correlation between protein and oil content of the seed that is postulated to be rooted in C-N partitioning biochemistry and physiology of the seed(Reekie and Bazzaz, 1987; Moose *et al.*, 2004; Guo *et al.*, 2013; Li *et al.*, 2010). Likewise, negative correlations were reported for traits like fruit size and number of inflorescences in various plants(Reekie and Bazzaz, 1987; Schoen and Dubuc, 1990), or between overall seed yield and protein content in some crops such as wheat and soybeans (Oury and Godin, 2007; Rotundo *et al.*, 2009) and between plant biomass and seed yield in grasses like sorghum (Piper and Kulakow, 1994).

Crop-wildrelatives studies comparing the degree of antagonism between these correlated traits for elite and native versions of crop species revealed that the extreme antagonisms observed in elite crop varieties are often variety specific, and are likely to be driven by the genotype x environment x cropping systems interactions and genetic drift- due to the bottlenecks the breeding populations experience under extreme selection (Ledford, 2017; Greene *et al.*, 2018; von Wettberg *et al.*, 2018), a concept referred to as genetic erosion (Van de Wouw *et al.*, 2010).

Modern studies with mutants and gene editing successfully validated that these antagonisms are often undesirable by-products of domestication and breeding practices (Ledford, 2017), and indeed some can be overcome by genetic modifications (Li *et al.*, 2015; Soyk *et al.*, 2017) albeit at the cost of adding an additional trait to select into the breeders' tally of breeding goals.

These and other constraints described previously lead to breeding selections and advancement decisions in a breeding program becoming a complicated balancing act; approaches for balancing the various priorities are still heavily debated(Batista *et al.*, 2018). Some favor straight truncation selection, others suggest applying index selection, with variable indices at each advancement stage gates, with no consensus over a standard process.

The changing economics of the seed industry starting around the late 1980's set the breeders up for yet another incredibly hard task: Starting with the existing elite varieties-those that are indeed regionally adapted and genetically differentiated, create varieties that are adapted to large cropping areas that demonstrate stable performance across environments [1] Following the development of the transformed strain, the transgene is then conventionally back-cross introgressed into the rest of the elite germplasm.

Since this is a costly process, economic and practical constraints have driven the number of varieties that should be designated elite for marketing to a fraction of what it

[1] This objective was driven by the emergence of GM technologies. In most crops, there is only a single transformable lab strain/ variety, and that strain is often the only strain that is genetically modified.

was before. Regional adaptation concept started to be frowned upon, and the quest for finding that elusive elite line that works well everywhere began(van Eeuwijk *et al.*, 2016).

Since all the starting material was elite, and all had acceptable productivity traits, the logical reasoning for how to achieve this task was through modification of flowering time characteristics of an existing variety so that it could be grown in additional relative maturity zones (Doubler, 2016). The proposed path to doing so was through backcross introgression and directional selection. The idea was if a cross between a high-line and a low-line for the favorable trait is available then a cross between the two would be segregating for the desirable trait. The breeder then could start back-crossing selected progeny to either parent to create recombinant-lines that would only be different from the target parent in its flowering time characteristics, but would otherwise be identical to parental phenotypes.

Although the logic of this process- a natural extension to decades long QTL mapping and marker assisted breeding studies- was correct and applicable to some crops, i.e. soybeans (Doubler, 2016), the complex nature of the flowering time as a trait in many other crops, influenced by hundreds of QTL with numerous alleles was not amenable to this approach(Buckler *et al.*, 2009). The genetic architecture of flowering time is highly complex in many crop species and thus involves multitudes of loci scattered across the genome. An attempt to co-introgress more than a handful of donor loci alleles to a recipient background with only limited number of recombination possibilities in a biparental cross creates a requirement for hundreds of thousands of progeny per cross to be generated at each back-cross generation to find the elusive recombinants with the right combination of alleles across their genome.

One recent example that demonstrates this process clearly is the Tenallion (2018) study, where the authors employed 13 generations of divergent selection to an initial biparental cross, and obtained five distinct populations with a time-lag of roughly two weeks between Early- and Late/VeryLate-flowering populations (Tenaillon *et al.*, 2018). Although this may attest to the feasibility of such a process, 13 generations roughly translates to 6 years of breeding, minimum. Therefore, it is clearly not-optimal for commercial product development, where a product life cycle for seed varieties are often estimated to be at 6 to 7 years.

After about three decades of trying to full fill these objectives with a combination of marker assisted breeding(Ribaut and Hoisington, 1998) and ideotype selection approaches(Gauffreteau, 2018) breeders finally realized that for majority of breeding objectives (e.g. improve yield, change RM and introgress transgenes and disease resistance alleles at the same time) this was a logistically challenging task and rarely produced the fully intended product. The number of loci that required concurrent selection was increasing exponentially, as the number of disease resistance traits and

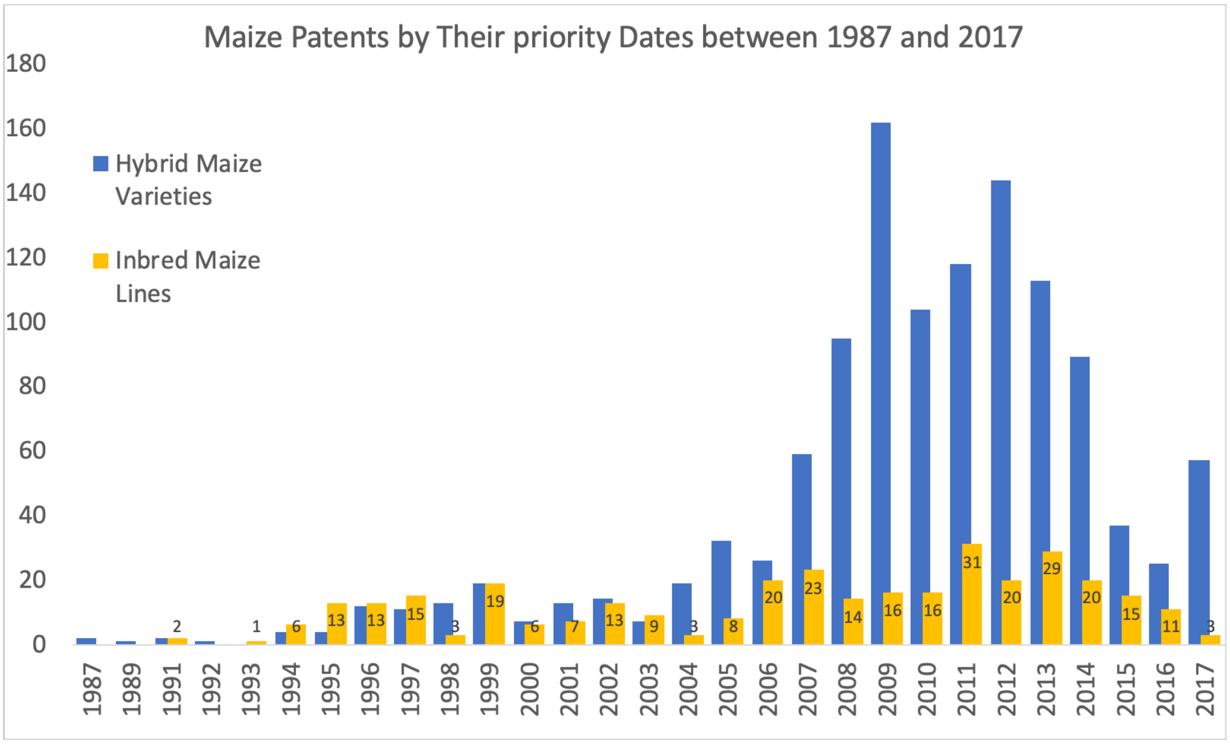

Figure 1. Number of maize patents filed between 1987 and 2017, ordered by their priority dates, and broken down as inbred(gold) and hybrid(blue). Data extracted from patents.google.com, and is available as supplementary material. The number of inbred lines developed is more uniform over time, while the number of hybrids show a big jump, starting in 2007, marking the onset of priority shifts from inbred line development to hybrid testing driven by hopes of generating hybrids of existing elite parents that can work over larger geographies. The dip observed in hybrid numbers in 2015 marks the onset of the products developed through predictive modelling, whereby cutting back the numbers for expensive advanced stage testing,trialing and registration.

desired properties increased. As a result, the process quickly became intractable with existing methods, and far exceeded the desired delivery times and created undue stress on logistics and operations (Tutino, 2016).

Rate of genetic gain in breeding programs stayed stagnant, and ambitious product development deadlines set due to increasing economic pressures in the competitive markets passed unmet which in turn effected the economics of the organizations and the research funds available for breeding. These trends are reflected in the number of inbred and hybrid variety patents filed for maize and soybean varieties, comparing the historical trends with those after 2010 (Figure 1). These economic limits were one of the significant drivers that lead to the adoption of genomic selection as a concept, and the breeding community embracing predictive modelling as the method of choice for rapid genetic gain in breeding programs.

Genomic selection offers the unique perspective of selecting the best individuals out of the available population pool, in contrast to trying to create the ideal individual based on an ideotype, hence removing the constraint for selecting the best alleles at each loci in a trade-off for selecting the best combination of alleles across the genome that work well together and is already available within the set of existing varieties.

2. Scribes, Prophets and Match Makers: Clustering, Classification and Combinatorics for Breeding

"Predictive algorithms start out as historians: they study historical data to detect patterns. Then they become prophets: they devise mathematical formulas that explain the pattern, test the formulas against historical data withheld for the purpose, and use the formulas to make predictions about the future." Lepore (2018)

2.1. Prediction of Unobserved Phenotypes by their genomic similarity to known samples

One of the earliest and most widely used applications of machine learning in crop breeding is its utilization for prediction of unobserved phenotypes, based on genetic markers. This process often referred to as GEBV estimation (for Genomic Assisted Breeding Value estimation) utilizes the resemblance between relatives principles as measured by the genomic similarity between individuals in a population. The initial applications leveraged pedigreed populations, where the parents of a bi-parental cross and their progeny were used to develop models that could predict un-observed progeny phenotypes from the same cross (Meuwissen *et al.*, 2001; Meuwissen and Goddard, 2001). Later on, methods were developed to apply similar principles to predict phenotypes for un-pedigreed, diverse line populations leveraging identical by state based resemblence between individuals as opposed to identical by descent(Peiffer *et al.*, 2014; Ersoz ES, 2017).

Statistically this is a very straightforward process, wherein the genome is characterized by assayable markers, where the markers are assayed in both parents and in a fraction of the progeny. In a sense, this step leverages the same principles as QTL mapping approaches - where the objective is assignment of a phenotypic value to each segment of the genome. The main operational difference from QTL-mapping lies in the fractionation of phenotypic variation that is attributed to each marker. While in QTL mapping this objective is achieved through a direct test of correlation between the allelic state at a genomic position vs. the phenotype, in many GS applications the first step is inference of individuals' similarities to each other via calculation of a G- matrix, and then fractionating the G-explained phenotype. The more conceptual, theoretic distinction between QTL mapping and GS is in how the genetic differences are modeled to contribute to the phenotype, with the QTL models primarily targeting relatively large-effect markers/intervals and the GS models aiming to capture polygenic, small-effects as a composite value across all intervals in the genome (Meuwissen and Goddard, 2001). More recent models allow searching and reporting both types of effects and use flexible distributions (Moser *et al.*, 2015).

Naturally, this approach was first applied to fully characterized populations where predicted and observed phenotypic values have been compared and contrasted. The level of correlation between the observed versus expected phenotypes define the level of *accuracy-of-prediction* that can be achieved by this approach. Since 2010, the research community has focused on ways to incorporate flexibility and additional features into these models, including environmental parameters, as well as genotype-environment interaction (g x e) terms.

Nonlinear components can also be added into the linear predictions (Voss-Fels *et al.*, 2018). Nonlinear (epistatic or conditional) predictions are especially important in elite breeding populations; the form of g x e selection will vary between optimization for wide adaption varieties and specific small-holder or regionally-branded varieties – but consideration of the environment is now an integral part of genomic selection algorithms (Voss-Fels *et al.*, 2018).

2.2. Paradox of Choice: Many methods one GEBV

There are many recent published examples of selection on relatively small scale data sets in crop and animal breeding (Gorjanc *et al.*, 2018; Dimitrijevic and Horn, 2018; Ozimati *et al.*, 2019; Rio *et al.*, 2019; Sweeney *et al.*, 2019). Linear methods such as GBLUP typically work quite well for building predictions for almost all applications. Heslot et al.(Heslot *et al.*, 2012) compared and contrasted performance metrics from several methods (Heslot *et al.*, 2012) and concluded that when multiple methods have been applied to the same data set, prediction quality is often similar for GBLUP, bayesian models, and recursive partitioning methods such as random forest (Heslot *et al.*, 2012) – however the exact set of SNP alleles varies with the specific method. These comparisons reported were based on predictions in biparental populations.

When various methods and models were compared for their predictive accuracy in un-pedigreed populations, however, the observation was that optimization of the training population composition based on relatedness to a target population can potentially increase the prediction accuracy and alter the algorithmic efficacy of various methods and models applied for within and cross-population for prediction accuracies and GEBV calculations (Ersoz ES, 2017).

Multi-environment studies for variety testing and development are often statistically under powered for detecting small effect sizes, due to the economic trade-off created by the total number of plots that can be included in an experiment. Breeders are often required to choose between small number of samples in many sites vs. large number of samples in few sites while designing these experiments. In the absence of preliminary

doi:10.20944/preprints201903.0115.v1

power studies to guide such choices, these decisions are often driven by operational optimization concerns based on advancement stage of the material tested. In early stages of line development, few sites with many samples are favored, while in later stages many sites with few samples are favored.

This logistic limitation of multienvironment testing, often disregards the variability expected in QTL effects across environments by virtue of GxE, as well as penetrance, as nuisance parameters [2] that are anticipated but are not expected to effect the estimates of "real effectors".

Since the underlying objective in commercial breeding is often *adaptation to large geographies and cropping systems*, only the effects that can be detected despite these limitations are considered "real" and are leveraged for MAS applications. This process inadvertently creates an implicit selection bias ***against*** GxE responsive QTL in elite variety testing and material development.

Another concern in QTL mapping and genomic selection is perceived pleitrophy [3]. As was described previously, both QTL mapping and GS approaches work through fragmental evaluation of the genome for the correlated change in allele states and the phenotype of interest. The larger the size of each fragment evaluated, the higher the likelihood that it will contain QTL for more then one phenotype, especially if the phenotypes of interest are complex in their genetic architecture (large number of small effect QTLs). We are calling this phenomenon *perceived pleitrophy* as opposed to genetic pleitrophy, since it is in fact an artifact of the analysis methods and does not actually offer any insights into the molecular nature of co-regulation of multiple phenotypes.

Kemper *et al.* (2018) has shown with simulations that the assumption of pleitrophy across multiple environments, generates significantly different results in contrast to analysis of each trait-environment combination individually. The difference was often not a significant change in predictive efficiency, but rather in the distribution of significant effects and the markers implicated to capture that variation.

This and the observation that any random set of independent markers are as predictive as known-allele/effector networks in explaining observed phenotypic variation (Buckler *et al.*, 2009) casts doubt over the utility of GS and by proxy GWAS in its ability to identify functional variation (Moore *et al.*, 2018). More work is needed to understand such observations as to their cause, whether they are artifacts of the methods used, or if they are biologically based.

[2] See Nuisance Parameter https://en.wikipedia.org/w/index.php?title=Nuisance_parameter&oldid=875520938

[3] See Pleiotropy https://en.wikipedia.org/w/index.php?title=Pleiotropy&oldid=883927922

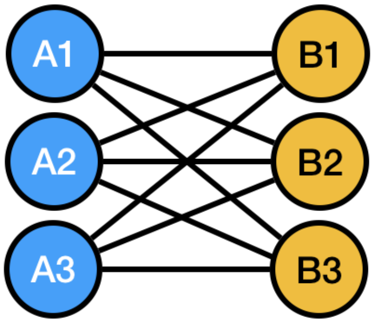

Figure 2. Graph network of potential breeding crosses considered

2.3. *Black sheep to Black Swans : Parent, Cross and Progeny Selection*

The first practical question any breeder needs to repeatedly address at the beginning of each and every breeding cycle is which crosses to make, or in other words ***parent & cross selection***. This problem is a variation of a very well studied mathematical problem, formally known as combinatorics. It is defined as dealing with combinations of objects belonging to a finite set in accordance with certain constraints. One of the oldest and most accessible parts of combinatorics is graph theory and set theory.

Assume you have lines $A1, A2, A3$ and $B1, B2, B3$ available to make crosses with Figure 2. You start with the constraint that the A types should not be crossed with each other and neither does B types. Of the remaining nine combinations, you need to pick one. An additional constraint is that you would like to maximize the transgressive segregation potential of the cross, with generating the maximum number of progeny that perform better than either parent.

This problem can be visualized as a network graph (Figure 2). The graph consists of six nodes and nine vertices and is not directed- meaning the direction of the vertices from node to node is not defined. If the directionality of a cross is important for the process (e.g. Although monoecious, the lines will be treated like they are diocious: A lines will be designated females, and B lines will be designated males, and reciprocals are not considered equal), the direction can also be defined.

Let's assume that each of the vertices carry a weight. This weight can be the genetic distance or similarity between the nodes (Figure 3a), the average yield observed in the F1 when the two parents at the nodes were crossed (Figure 3b), or the fraction of transgressive segregants observed in the past experiments when these crosses were made(Figure 3c), or a vector or scalar combination of all of the above weights. For visualization purposes, the lengths of the vertices can be scaled to represent their weights, which would present an immediate visual representation of all of the above mentioned relationships between nine individual lines.

The intuitive visuals is just one advantage of using graph models to visualize breeding crosses. Analytically, graph modelling of breeding/crossing networks as

doi:10.20944/preprints201903.0115.v1

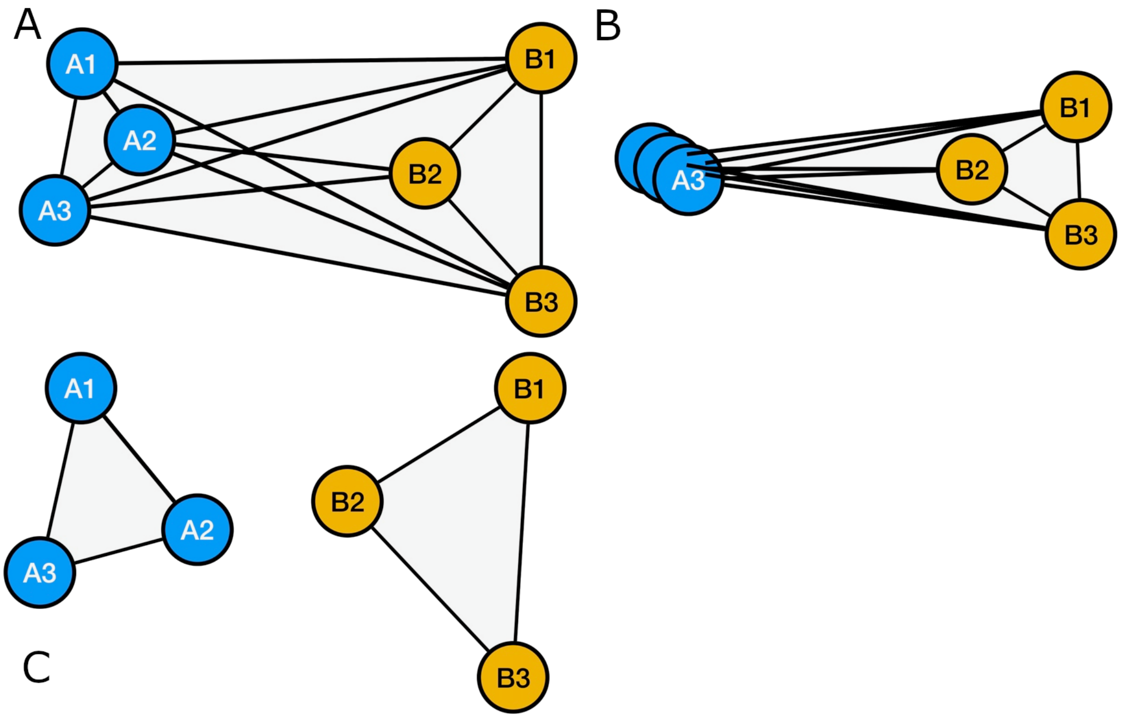

Figure 3. Example cross network graphs with the vertices scaled according to A. Genetic distance, B. average F1 yield, C.Fraction of transgressive segregants. If no transgressive segregation was observed, the vertex was deleted.

described allows applications of graph mining methods and kernel methods for ranking, clustering and classification of nodes(*parents*) and vertices(*crosses*) in a graph model. Furthermore, graph models allow for application of methods such as link analysis (Harper and Harris, 1975) and algorithms such as PageRank (Page *et al.*, 1999) that were originally developed for Social Network analysis. These methods would provide new methods for predicting the weight/value of an unobserved vertex(*cross*) from existing cross networks (Park and Yook, 2014) (Figure 4A, 4B), with or without leveraging genetic marker information.

This type of analytics opens up extensive possibilities for breeders, and will also allow integration of data across different experiments- that are partially overlapping across time and geographies(Simko and Pechenick, 2010).

Furthermore, these algorithmic applications allow development of learning-systems through iterative cycles of prediction& observation during breeding.

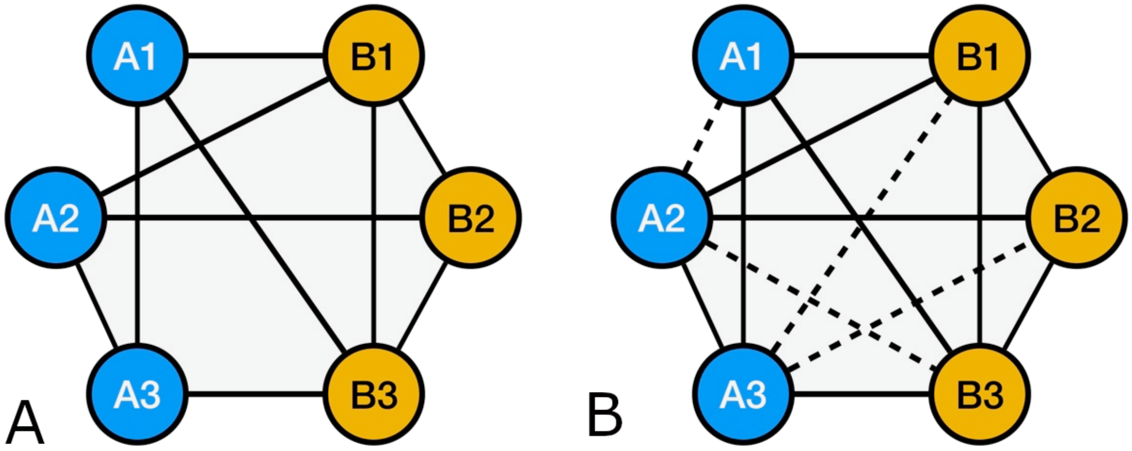

Figure 4. A. Incomplete network graph representation of observed breeding crosses B. The complete network with the unobserved vertices shown as dashed lines, representing predictions

3. Introgression in Breeding : Applications of Mathematical Optimization & Shortest Path Algorithms for machine learning

Marker assisted selection applications (MAS - MARS) have a long history in crop and animal improvement (Ribaut and Hoisington, 1998; Lande and Thompson, 1990; Haley and Visscher, 1998; Ribaut and Ragot, 2006). Despite the high hopes of the breeding community to apply MARS for multiple traits simultaneously for multivariate index selection, the method's dependency on very large sample sizes to achieve sufficient efficiency (Lande and Thompson, 1990) have restricted its use and limited its success. In fact, literature reports on theoretical efficiency of multi-trait MAS are scarce and the theoretical limits of efficiency were not empirically tested and published until 2010 (Togashi and Lin, 2010). Togashi and Lin examined the effectiveness of five different selection methods under 72 scenarios. They reported that an index consisting of three traits which are controlled by 40, 50 and 60 segregating loci, respectively, each with two alleles would require 3^{40} x 3^{50} x 3^{60} $=$ 3.7 x 10^{71} different kinds of aggregate genotypes- which arguably can not all be available in the standing population under selection- and led to the acknowledgement that MAS applications for low heritability, high complexity phenotypes are practically not feasible.

The most important and historically most successful application of MAS in breeding is for backcross-introgression of a single high value locus- such as a transgene, into new genetic backgrounds, that is also known as MAB for Marker Assisted Backcrossing. In MAB, the objective is to introgress the target region to a new genetic background with the minimum amount of *linkage drag*(Hospital, 2005). The very first and most well known application of MAB is for the introgression of *Bt* gene into elite breeding material(Ragot *et al.*, 1995). In MAB, *foreground selection* and *background selection* are applied simultaneously controlling for the size of the introgression at the target as well as other potential introgressions from the donor to the recipient by maximizing the recipient allele recovery across the background, respectively at each generation of backcrossing. For a review of early applications and successful examples of MAS and MAB see (Collard and Mackill, 2008) and the references therein.

The practical problems that needs to be addressed for improvement of preexisting varieties is based on the idea of routinely generating rationally driven combinations of functional genetic variants (Wallace *et al.*, 2018) in a target background. The task is the same as it was defined in MAS several decades ago, however the scale and efficiency needed to increase. In natural evolutionary processes, there are two complimentary processes recognized for their efficacy in rapid creation of genetic novelties: hybridization and introgression. In crop breeding however, due to its often unpredictable outcomes, hybridization is often avoided, and introgression is preferred.

Here we suggest combining controlled hybridization with subsequent recurrent selection to increase the speed and scale of the process, leveraging mathematical optimization(Evans, 2017) and shortest path algorithms (Festa, 2019). Assume that we are staring with a collection of 5 individual elite varieties, each of which posses 10 favorable haplotypes for a total of 50 loci to be combined in one background. Assume that the ranking of each haplotype for each locus in the recipient background is also available. The task is to maximize the individual locus haplotype ranks while minimizing the number of crosses to be made and the number of progeny required per cross.

We would start with evaluating the pairwise genetic relatedness between the recipient and the donors, which can be visualized as a network graph, where the nodes are the individuals and the vertices are a measure of genetic relatedness between the individuals. Each node has 10 unique alleles/loci to contribute, including the 10 that are unique to the recipient.

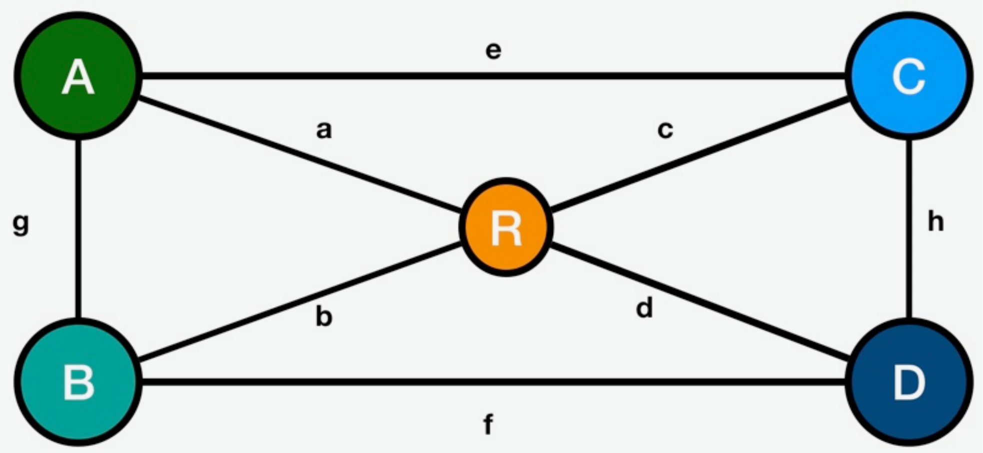

Figure 5. The graph model network of genetic relatedness between the donors and the recipient described in Section 3

To avoid the unpredictable outcomes of hybridization (non-linear responses), the first step would be the identification of the closest pairs, which would maximize the linear response outcomes in the hybrid. In this example, it will be g, the distance between $A - B$ pair and h, the distance between $C - D$ pair. An AB F1 will have all the favorable alleles from both A and B as heterozygotes in the cross. Similarly, a CD F1 will have all the favorable alleles available. That is a total of two crosses. Now lets cross ABxR and CDxR . Next we cross ABRxCDR allowing all 50 alleles to segregate in the $ABRCD$ population. $ABRCD$ population will contain any number of combinations of alleles at the 50 loci of interest would be equivalent to an F2 population, in relation to any one of the parents.

We can then evaluate the available progeny for their allelic composition, and identify a number of progeny to be back-crossed to the R parent. We continue backcrossing hybrids to the R parent, until the introgression size at each of the 50 loci is reduced to under 1 cM. In other words, the genome is all R minus 40 cM of introgressions.

Note that we have not yet made a selection based on ranks at each locus, just created a diverse set of introgression lines, that can realize any number of possible combinations of alleles across the 50 targets. Next, we rank each of the available progeny for their predicted values across each of the loci, and pick the ones that scores the max.

That is 4-5 generations of crosses, for population development (two for hybridization, 2-3 for backcross introgression).The algorithm described above can easily be applied to larger number of donors and larger number of loci. The time to complete hybridization in number of generations will be x , where number of donors $N = 2^x$ Number of back crosses required will be proportional to the total size of the introgressions measured as percentage of the genome. The larger the sum of all introgressions, shorter the number of generations to completion.

doi:10.20944/preprints201903.0115.v1

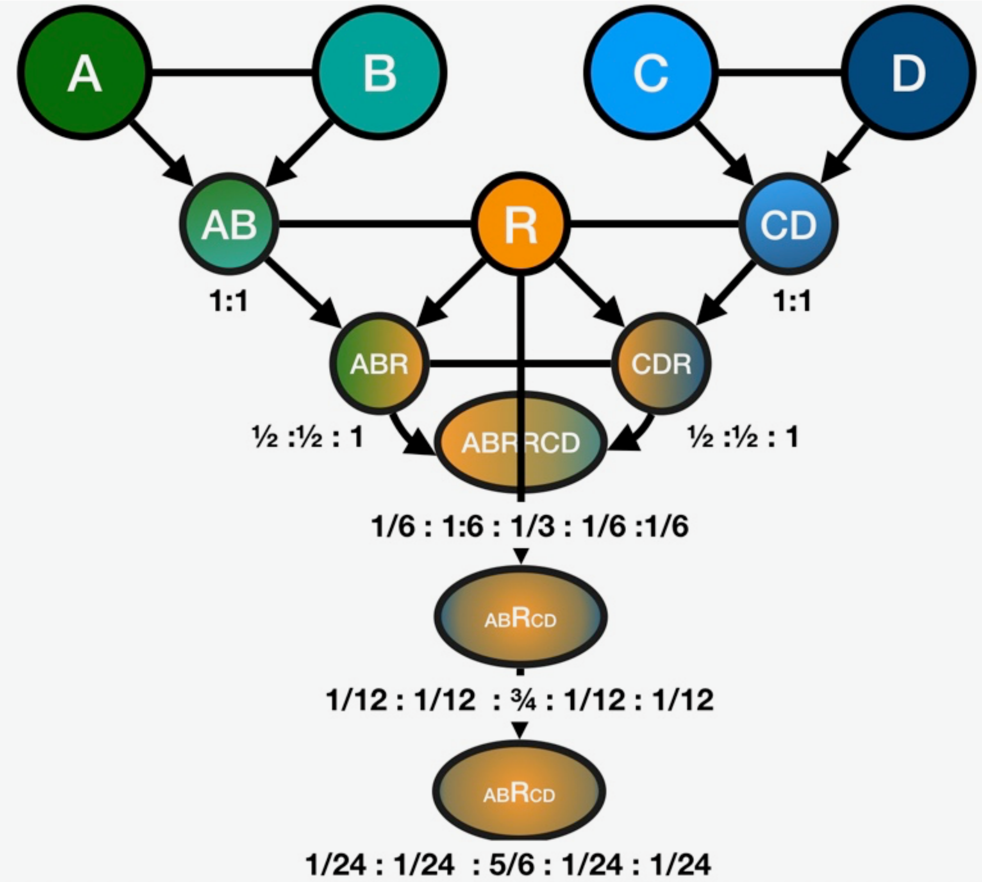

Figure 6. Each generation of crosses described in the scenario in section 3 of the manuscript. At the end of each generation, the expected genomic contribution of each parent is listed under the nodes. A total of 5 generations of crosses shown, at the end of which the expected value of genomic contribution ratio for each of A, B, C, D each are $1/24th$ while R is $5/6th$. The variance around these expectation can be estimated by simulations

When a large collection of individuals(hundreds of thousands) in a breeding population are considered, the path defined above to a fully stacked line, can be evaluated by using algorithms such as gradient descent in neural networks. The algorithm will need to go through iterative cycles of identifying crosses that carry favorable alleles to be combined but are otherwise most similar to each other in overall genetic composition, followed by GEBV calculations for the resulting progeny, and ranking of the progeny for their potential to carry maximum number of favorable alleles going into a back-cross. It will then evaluate every progeny generated as a result of the back-cross process for its GEBV and provide rankings to be advanced to the next generation of back-crosses.

In simpler terms, this is an iterative decision process, where in each step, most closely placed pair of lines are selected, and used for creation of cross, and the F1 of that cross is crossed again to the next closest line that is available in the network, until sufficient hybridization is achieved, and then the resulting lines are cleaned up and ranked for most favorable combinations.

This approached was utilized in an unusual product development project by Syngenta (Ritchie *et al.*, 2017). Briefly, one of the coveted maize products developed that carry an over-expressed transgene for Vip3A protein that was shown to confer insecticidal effects resulting in chewing insect resistance with no negative impact on the hybrid productivity phenotypes. However, later on it was revealed that Vip3A can cause decreased male fertility in certain inbred maize plants under normal growing conditions, creating issues around parent line maintenance, and hybrid seed production. In a multi-year, multi-environment study, nine QTL across various locations across the genome that create this male sterility phenotype has been identified. The research team then utilized the hybridization:backcrossing approach described here to develop an introgression plan for introgressing these QTL back into a large set of commercial line parents in development to mitigate any potential fertility issues (Ritchie *et al.*, 2017) .

This type of iterative decision making tasks are what machine learning and deep learning algorithms excel at, and there are a variety of algorithms that can be used for identifying the shortest path that will traverse through a predefined landscape. A few examples of such machine learning algorithms are gradient descent(Fedoryuk, 1989; Smirnov, 2010; Ruder, 2016) and Voronoi diagram methods with applications to 1-NN clustering (Sebastiani, 2002).

4. Variety testing trial design and advancement decisions

Successful breeding for new cultivars involves understanding the expected conditions in growing regions and the crop adaptations needed to meet these expectations. To ensure the alignment of cultivar characteristics and expected growing conditions breeding efforts seek information from both to guide breeding decisions about trialing, parent selection and progeny selection. The standing variation in existing domesticated cultivars drive the decisions around where to get the useful genetics and climactic adaptation information. For instance, the effects of geographic variation and isolation on cultivar genetic variability can be illustrated by the variation in Peruvian highland maize cultivars resulted from cultivation by Andean farmers in a wide range of conditions (Ortiz *et al.*, 2008; Grobman, 1961). The management practices unique to the Moray plateau region in Peru, i.e. terraces designed as con-circular depressions that vary in depth and orientation with respect to wind and sun resulting in temperature adaptive differences of as much as 15°C (Wright *et al.*, 2010) are considered one of the marvels of ancient agricultural engineering, and is postulated to have contributed to the high levels of genetic variation observed in Peruvian landraces (Ortiz *et al.*, 2008; Grobman, 1961).

Modern breeding programs for field crops conduct field trials along the different breeding stages increasing the number of sites as they progress in the cultivar selection pipeline (Cooper *et al.*, 2014; Byrum *et al.*, 2016). Ideally, these multi-environment trials (METs) are in the regions where the cultivar is expected to be adapted by matching climactic properties of a trial site with the location-of-origin climactic properties of a new variety (Kang, 1997).This process seeks to ensure that when new cultivars are made available to farmers, they are adapted to a region often defined as the target population of environments (TPE) (Löffler *et al.*, 2005).

4.1. MET inference

Information collected with MET is the result of statistical analysis of agronomic field experiments following well-established principles of experimental design in the form of statistical inference, to develop descriptive statistics. Over the years better access to computing power and statistical methods enhanced the information obtained from field experimentation. For instance, with the widespread introduction of mixed linear models(MLM) in the 1990s it became possible to consider cultivar performances as random effects, widening the inference space of BLUPs and making possible to model spatial trends within the field that often challenge implicit assumptions of the trial designs related to residual independence.

More importantly, MLMs makes it possible to dissect genotype by environment interactions (in form of matrices) after considering the main effects of the genotypes *and* environment. For example, Additive Main effects and Multiplicative Interaction (AMMI) model is an analytical tool to interpret G-by-E based on the spectral decomposition of GxE effects captured. However, using these descriptive results as the main basis to make decisions on newly developed material and data is a risk due to sampling bias (Crossa *et al.*, 1991) leading to important interactions from a single environment with unique conditions being overlooked. Also, these methods are often based on classification of labels in a discrete fashion where each trial is treated as a combination of factors that can be summarized by a combination of location name and year. This crude classification approach makes it difficult to make comparisons between trialing networks where regions from multiple season from different countries are included, and complicate the inferences for understanding the biological basis of cultivar adaptation.

4.2. Mechanistic/ Deterministic modeling

Some of these limitations addressed by using crop growth and development models structured on crop ecophysiology principles quantifying phenology, biomass accumulation, and partition. This framework allows the integration of agroclimatic conditions, crop management practices and genotypic information (Messina *et al.*, 2010). In the context of breeding programs, phenotypic differences between genotypes can represent genotypic coefficients(Lecoeur *et al.*, 2011). These coefficients can be derived from phenotypic measurements on cultivars and traits, enabling generation of cultivar specific predictions in the context of alternative management practices and seasonal conditions. Sinclair et al. 2010 (Sinclair *et al.*, 2010), used crop simulations to evaluate the impact modifying soybean cultivar characteristics associated with drought tolerance in different US soybean producing regions at 30 km resolution. Cooper *et al.* (2014), illustrates an example of ten corn hybrids for the 2012 season based on the characterization of drought tolerance traits according to the methodology described by Messina *et al.* (2010). One of the main limitations to this approach is the throughput and accuracy of phenomics: collecting precise phenotypic information to estimate these model parameters require specialized training and is resource intensive. This makes it difficult to generate extensive training datasets for subtle variations. In addition, a deterministic model might fail to capture hidden interactions between environmental factors, or sufficiently compensate for innate probabilistic nature of the biological processes.

4.3. Predictive Analytics framework

In the last decade, one field after another has been influenced by the emergence of *Predictive Analytics*. At its core, predictive analytics involve using datasets generated from activities such as product development or a research program in fields ranging from marketing to medicine to gain new insights to improve them. It uses advanced mathematics and statistical computing algorithms to represent trends, identify complex patterns and forecast trends (Dinov, 2018).

This transformation is largely driven by access to computing resources and more importantly by access to machine learning and optimization techniques in a wide variety of platforms. Among the predictive analytics approaches *supervised learning systems* use a training dataset to find patterns and trends in complex datasets (Jordan and Mitchell, 2015).

Predictive analytics approaches in contrast to traditional research programming approaches could have a critical role in improving the understanding of genotype by environment interactions, gaining insights into cultivar stability across environments and using this information to guide decisions in breeding programs. Traditionally a data set is collected from MET, and a generic statistical model is used to generate an mainly descriptive inference about the cultivars under evaluation. Alternatively, data collected from these trials can be used as input in the biophysical crop model along quantitative seasonal conditions information, and after calibration predictions can be generated on expected performance under untested scenarios (Cooper *et al.*, 2014). The predictions can further be extended to untested varieties, via applications of Bayesian models that would leverage allelic composition similarity as opposed to line-id in such a model.

This proposed approach uses cultivar phenotypic information from MET trialing networks in combination with environmental conditions to train new classification rules.

doi:10.20944/preprints201903.0115.v1

These new classification rules can then be applied to generate predictions for tested varieties' expected performance in TPEs. This approach has been widely used in studies of ecology and biodiversity and was used to predict species diversity distributions based on bioclimatic variables.

Bioclimatic variables characterized at a resolution of 1 km^2 global raster have been cited or utilized in more in more than 14,364 studies according to citation analysis from Google scholar for Hijmans *et al.* (2005). These datasets are often used to extrapolate from a site where a species of interest has been observed, to where else it may be found, using a machine learning framework (Ferrier and Guisan, 2006). There are a few examples of how this framework may be applied in breeding programs. Using datasets from multiple years of field trials, Zhong *et al.* (2018) combined machine learning techniques and robust optimization methods to identify soybean cultivars for a candidate environment.

Expanding on this approach Marko *et al.* (2017) compared multiple machine learning methods in combination with local and global optimization techniques to identify portfolios of cultivars adapted to target production environments.

To use this framework in breeding programs implies making changes in how cultivar metrics are evaluated. One is to consider adaptation as a probabilistic forecast. So instead of asking for a measure of central tendency often in the form of yield estimates with a confidence interval for an environment, it would be possible to ask what are the odds that a certain cultivar is adapted to a certain environment. This modification facilitates using new methods of analysis and performance metrics (Hofman *et al.*, 2017).

Then the challenge is how adaptation is defined if we consider it as a binomial outcome. In species distribution research, adaptation is often defined in binary terms,as presence or absence. In a breeding program, that question can be restated as whether a cultivar is more productive than another cultivar defined as a standard check. This framework allows integrating cultivar performance with specific environmental variables and creates training and validation datasets linking MET and TPEs, facilitating the introduction of predictive analytics techniques.

There are two main drivers for development and application of these methodologies in crop improvement. The first one is its application to selection of variety testing networks, and the second is the identification of new geographies where existing material can be suitable.

Global climate classification schemes aim to identify distinct climate types and map their geographical extents. By discretizing a multitude of local climates into a manageable number of climate types, we can simplify comparisons and allow ease of interpretation. For these purposes, BIOCLIM[4] was created (Fick and Hijmans, 2017); this resource is a

[4] See http://worldclim.org/version2

currated set of spatially interpolated monthly climate data for global land areas at a very high spatial resolution.

Commercial varieties are often developed with the directive for maximizing profitability. Therefore, it is often a commonly requested task to evaluate suitability of an existing variety to new geographies that define new markets. These experiments are often carried out by leveraging geoclimactic condition matching classifications between METS and TPEs as was described previously.

With the advent of climate change creating potential disturbances and redefining boundaries of the relative maturity zones for many crops, this process has become relatively unreliable if based on short-term environmental trends. It has become crucial to collect and leverage high resolution long term geoclimactic-records to drive this process(Atlin *et al.*, 2017). More than ever, it is now crucial to characterize and re-characterize maturity zones for their compatibility with existing varieties.

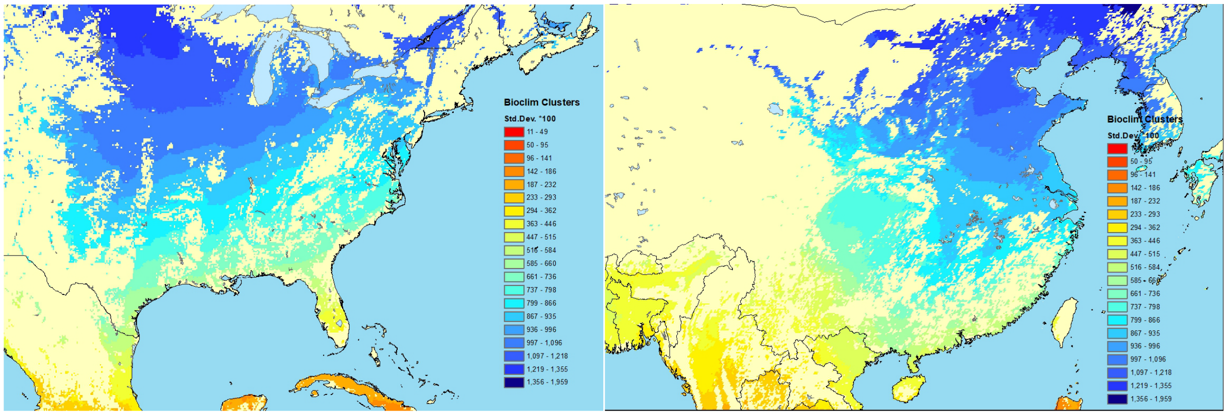

Figure 7. Classification of North America and East Asia cropland according to temperature seasonality variability (BIO4 variable from BIOCLIM). http://worldclim.org/bioclim

Testing effects of inclusion of geoclimactic features in predictive modelling of phenotypic performance across geographies, bioclicmactic parameters from bioclim have been used as features in machine learning algorithms across geographies (Martin unpublished, Figure 7). In an *Arabidopsis thaliana* case study (Ersoz, 2019) it was reported that inclusion of the environmental features into a Random Forest prediction model together with genetic features found no conclusive trends, since for some phenotypes there was increase in prediction accuracy and at yet for others, a decrease was observed. This was postulated to be due to the direction of the GxE parameters (un-adapted versus well-adapted GxE combinations) that was not explicitly controlled in the model. More work is needed to develop these models and evaluate their efficacy in crops.

Advancements in this exciting area requires not only private and public investments but also opportunities to access datasets from large MET network breeding and farmers organizations for public research and training. In recognition of this necessity, Syngenta has been coordinating an annual crowd-sourcing analytics challange in partnership with the Analytics Society of INFORMS since 2015. Examples of methods developed in these contests, known as Syngenta-Informs Challanges are Marko *et al.* (2017) and Zhong *et al.* (2018). More funding, public and private, and more data would enable future innovations in this domain, as well as training of domain experts that can apply these methods.

5. Machine Learning for Sustainable Agriculture

In addition to the geoclicmactic variables mentioned in the previous section (Figure 7), there are many other publicly available data resources that can be leveraged to evaluate various characteristic of cropping systems versus geographies.

For example, USDA- Economic Research Service [5] compiles and distributes annual data for land use and farm land utilization and management practices across the US such as crop rotation and average reported yield per crop. USDA Pest Information Platform for Extension and Education (PIPE)[6] collects and distributes pest and disease data from across north america. National Centers of Environmental Information division of National Oceanic and Atmospheric Administration (NOAA), collects and distributes drought monitor and extreme climate event data[7] and there are many others, that are too numerous to cite here. We will refer to this collection of repositories and data as *biogeoclimactic data* for ease of reference.

Biogeoclimactic data can be compiled in such a way that each data layer, anchored to geographic coordinates, may be treated as different layers of an image. Such restructuring and compilation of this data would enable use of Convolutional Neural Networks (CNNs), a tool designed for analysis of such layered data by developing ML models. These models can then be leveraged not only to fine tune classification of adaptive region definitions for existing varieties, but can also be leveraged to forecast suitability of existing varieties based on changing bioclimactic conditions (e.g. What is the expected performance for variety X in geography Y?, or what is the expected performance of variety X with climactic parameter set Z?). These models can ultimately be used for making management recommendations for optimizing on-farm management practices, for increasing sustainability of agriculture and to improve margins for the growers.

Another potential approach to modelling biogeoclicmactic effects for crop productivity is using the data without the discretization, on a continuous scale.

[5] https://www.ers.usda.gov/data-products/data-visualizations/

[6] http://sbr.ipmpipe.org/cgi-bin/sbr/public.cgi

[7] https://www.climate.gov/maps-data/data-snapshots/data-source-drought-monitor

Continuous scale modelling would allow capturing influences of extreme-climate events, i.e. droughts, floods, locust years, etc. without smoothing them over as regional averages or discarding them as outliers. This can be achieved by Recurrent Neural Network (RNN) Modelling, as these models seem to capture properties of such extreme weather events (Saha and Mitra, 2016).

6. Discussion

In this review, our objective was to present an overview of the current and potential applications of predictive modelling with machine learning for its applications to contemporary crop breeding and improvement. We started by emphasizing the fact that crop improvement is inseparably tangled with geoclimactic adaptation, followed by a discussion on available tools and strategies for predicting unobserved phenotypes under various conditions to accelerate and inform variety development efforts.

We have also discussed leveraging bioclimactic classification methods and long term geo-climactic data to inform variety development and placement as well as generating forecasts for crop productivity under today's rapidly changing environmental conditions.

We did not discuss genomics and biotechnology applications of ML & AI relevant to breeding such as gene editing target identification, gene and allele discovery, or leveraging crop-wild relatives for variety development. To do these topics justice, a separate article would be warranted.

We also have not discussed the fine details of algorithmic and methodological complexity of applications of machine learning approaches, since we are of the opinion that such discussions would best be included in an actual investigation study rather then a review ms.

Our intention was to keep this review focused on providing examples for how ML and AI can help improve breeding logistics and operational challenges, thereby rate of genetic gain in breeding programs for accelerated variety development. We have also presented a case for leveraging this data and methods for enabling sustainable agricultural practices by redefining and forecasting adaptive variety and management recommendations for farmers.

We hope that the examples we provided here would serve and inspire the community to develop and apply these novel models and methods in crop breeding in the near future.

Funding: This research, and preparation of the manuscript was fully voluntary on the authors' parts and no external funding was received for NFM and ESE. This project was partially supported by National Research Initiative Competitive Grant no. 2017-67013-26188 from the USDA National Institute of Food and Agriculture to AES.

Acknowledgments: We are grateful to Mark Cooper for asking us to contribute an ms to this special issue of Crop Science, and providing recommendations as to its topic. The funders had no role in study design, data collection and analysis,decision to publish, or preparation of the manuscript.

Conflicts of Interest: EE is the founder and Research Director for Umbrella Genetics, which is a for profit organization that offers big data analytics services for agricultural product research and development. NFM was the Chair of the Challenge Review Committee for Syngenta- Informs Challenge starting 2017 up to 2019. These connections had no role in the collection, analyses, or interpretation of data presented here, other then as cited, or in the writing of the manuscript, or in the decision to publish the ms.

References

Borlaug, N. Sixty-two years of fighting hunger: personal recollections. *Euphytica* **2007**, *157*, 287 – 297.

Bullock, D. Crop rotation. *Critical Reviews in Plant Sciences* **1992**, *11*, 309–326, [https://doi.org/10.1080/07352689209382349]. doi:10.1080/07352689209382349.

Kwon, S.; Torrie, J. Heritability of and Interrelationships Among Traits of Two Soybean Populations 1. *Crop science* **1964**, *4*, 196–198.

Meredith, W.R.; Bridge, R. Breakup of Linkage Blocks in Cotton, Gossypium hirsutum L. 1. *Crop science* **1971**, *11*, 695–698.

Kato, T.; Takeda, K. Associations among characters related to yield sink capacity in space-planted rice. *Crop science* **1996**, *36*, 1135–1139.

Triboi, E.; Martre, P.; Girousse, C.; Ravel, C.; Triboi-Blondel, A.M. Unravelling environmental and genetic relationships between grain yield and nitrogen concentration for wheat. *European Journal of Agronomy* **2006**, *25*, 108–118.

Erskine, W.; Williams, P.C.; Nakkoul, H. Genetic and environmental variation in the seed size, protein, yield, and cooking quality of lentils. *Field Crops Research* **1985**, *12*, 153–161. doi:10.1016/0378-4290(85)90061-9.

Guanming, S.; Jean-paul, C.; Kyle, S. An Analysis of the Pricing of Traits in the U.S. Corn Seed Market. *American Journal of Agricultural Economics* **2010**, p. 1324.

Reekie, E.; Bazzaz, F. Reproductive effort in plants. 2. Does carbon reflect the allocation of other resources? *The American Naturalist* **1987**, *129*, 897–906.

Moose, S.P.; Dudley, J.W.; Rocheford, T.R. Maize selection passes the century mark: a unique resource for 21st century genomics. *Trends in Plant Science* **2004**, *9*, 358–364. doi:10.1016/j.tplants.2004.05.005.

Guo, Y.; Yang, X.; Chander, S.; Yan, J.; Zhang, J.; Song, T.; Li, J. Identification of unconditional and conditional QTL for oil, protein and starch content in maize. *The Crop Journal* **2013**, *1*, 34–42. doi:10.1016/j.cj.2013.07.010.

Li, Y.H.; Li, W.; Zhang, C.; Yang, L.; Chang, R.Z.; Gaut, B.S.; Qiu, L.J. Genetic diversity in domesticated soybean (Glycine max) and its wild progenitor (Glycine soja) for simple sequence repeat and single-nucleotide polymorphism loci. *New Phytologist* **2010**, *188*, 242–253. doi:10.1111/j.1469-8137.2010.03344.x.

Reekie, E.; Bazzaz, F. Reproductive effort in plants. 1. Carbon allocation to reproduction. *The American Naturalist* **1987**, *129*, 876–896.

Schoen, D.J.; Dubuc, M. The Evolution of Inflorescence Size and Number: A Gamete-Packaging Strategy in Plants. *The American Naturalist* **1990**, *135*, 841–857. doi:10.1086/285077.

doi:10.20944/preprints201903.0115.v1

Oury, F.X.; Godin, C. Yield and grain protein concentration in bread wheat: how to use the negative relationship between the two characters to identify favourable genotypes? *Euphytica* **2007**, *157*, 45–57. doi:10.1007/s10681-007-9395-5.

Rotundo, J.L.; Borrás, L.; Westgate, M.E.; Orf, J.H. Relationship between assimilate supply per seed during seed filling and soybean seed composition. *Field Crops Research* **2009**, *112*, 90–96. doi:10.1016/j.fcr.2009.02.004.

Piper, J.K.; Kulakow, P.A. Seed yield and biomass allocation in Sorghum bicolor and F1 and backcross generations of S. bicolor× S. halepense hybrids. *Canadian Journal of Botany* **1994**, *72*, 468–474.

Ledford, H. Fixing the tomato: CRISPR edits correct plant-breeding snafu. *Nature News* **2017**, *545*, 394. doi:10.1038/nature.2017.22018.

Greene, S.L.; Khoury, C.K.; Williams, K.A., Wild plant genetic resources in North America: an overview. In *North American Crop Wild Relatives, Volume 1*; Springer, 2018; p. 3–31.

von Wettberg, E.J.; Chang, P.L.; Başdemir, F.; Carrasquila-Garcia, N.; Korbu, L.B.; Moenga, S.M.; Bedada, G.; Greenlon, A.; Moriuchi, K.S.; Singh, V.; others. Ecology and genomics of an important crop wild relative as a prelude to agricultural innovation. *Nature communications* **2018**, *9*, 649.

Van de Wouw, M.; Kik, C.; van Hintum, T.; van Treuren, R.; Visser, B. Genetic erosion in crops: concept, research results and challenges. *Plant Genetic Resources* **2010**, *8*, 1–15.

Li, L.; Zheng, W.; Zhu, Y.; Ye, H.; Tang, B.; Arendsee, Z.W.; Jones, D.; Li, R.; Ortiz, D.; Zhao, X.; et al.. QQS orphan gene regulates carbon and nitrogen partitioning across species via NF-YC interactions. *Proceedings of the National Academy of Sciences* **2015**, *112*, 14734–14739. doi:10.1073/pnas.1514670112.

Soyk, S.; Lemmon, Z.H.; Oved, M.; Fisher, J.; Liberatore, K.L.; Park, S.J.; Goren, A.; Jiang, K.; Ramos, A.; Knaap, E.v.d.; et al.. Bypassing Negative Epistasis on Yield in Tomato Imposed by a Domestication Gene. *Cell* **2017**, *169*, 1142–1155.e12. doi:10.1016/j.cell.2017.04.032.

Batista, L.; Gaynor, R.C.; Margarido, G.R.; Byrne, T.; Amer, P.; Gorjanc, G.; Hickey, J.M. Plant breeders should be determining economic weights for a selection index instead of using independent culling for choosing parents in breeding programs with genomic selection. *bioRxiv* **2018**, p. 500652.

van Eeuwijk, F.A.; Bustos-Korts, D.V.; Malosetti, M. What Should Students in Plant Breeding Know About the Statistical Aspects of Genotype Environment Interactions? *Crop Science* **2016**, *56*, 2119–2140. doi:10.2135/cropsci2015.06.0375.

Doubler, T.W. The use of genetic information to predict the relative maturity of soybeans. PhD thesis, IAState, Iowa State University Digital Repository, 2016.

Buckler, E.S.; Holland, J.B.; Bradbury, P.J.; Acharya, C.B.; Brown, P.J.; Browne, C.; Ersoz, E.; Flint-Garcia, S.; Garcia, A.; Glaubitz, J.C.; et al.. The Genetic Architecture of Maize Flowering Time. *Science* **2009**, *325*, 714–718. doi:10.1126/science.1174276.

Tenaillon, M.I.; Seddiki, K.; Mollion, M.; Guilloux, M.L.; Marchadier, E.; Ressayre, A.; Dillmann, C. Transcriptomic response to divergent selection for flowering time in maize reveals convergence and key players of the underlying gene regulatory network. bioarxiv. https://www.biorxiv.org/content/early/2018/11/23/461947, 2018. doi:10.1101/461947.

Ribaut, J.M.; Hoisington, D. Marker-assisted selection: new tools and strategies. *Trends in Plant Science* **1998**, *3*, 236–239. doi:10.1016/S1360-1385(98)01240-0.

Gauffreteau, A. Using ideotypes to support selection and recommendation of varieties. *OCL* **2018**, *25*, D602. doi:10.1051/ocl/2018042.

Tutino, C. Optimizing Breeding Process Builds Yield. Syngenta Thrive Magazine, 2016.

Lepore, J. What 2018 Looked Like Fifty Years Ago **2018**.

Meuwissen, T.H.E.; Hayes, B.J.; Goddard, M.E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* **2001**, *157*, 1819–1829.

Meuwissen, T.H.; Goddard, M.E. Prediction of identity by descent probabilities from marker-haplotypes. *Genetics, Selection, Evolution: GSE* **2001**, *33*, 605–634. doi:10.1186/1297-9686-33-6-605.

Peiffer, J.A.; Romay, M.C.; Gore, M.A.; Flint-Garcia, S.A.; Zhang, Z.; Millard, M.J.; Gardner, C.A.; McMullen, M.D.; Holland, J.B.; Bradbury, P.J.; others. The genetic architecture of maize height. *Genetics* **2014**, *196*, 1337–1356.

Ersoz ES, Myers C, K.D.e.a. Harnessing global genetic potential with CropOS. F1000Research 2019, 8:202 (poster) https://doi.org/10.7490/f1000research.1116445.1, 2017.

Moser, G.; Lee, S.H.; Hayes, B.J.; Goddard, M.E.; Wray, N.R.; Visscher, P.M. Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *PLOS Genetics* **2015**, *11*, e1004969. doi:10.1371/journal.pgen.1004969.

Voss-Fels, K.P.; Cooper, M.; Hayes, B.J. Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics* **2018**. doi:10.1007/s00122-018-3270-8.

Gorjanc, G.; Gaynor, R.C.; Hickey, J.M. Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theoretical and Applied Genetics* **2018**, *131*, 1953–1966.

Dimitrijevic, A.; Horn, R. Sunflower hybrid breeding: from markers to genomic selection. *Frontiers in plant science* **2018**, *8*, 2238.

Ozimati, A.; Kawuki, R.; Esuma, W.; Kayondo, S.I.; Pariyo, A.; Wolfe, M.; Jannink, J.L. Genetic Variation and Trait Correlations in an East African Cassava Breeding Population for Genomic Selection. *Crop Science* **2019**.

Rio, S.; Mary-Huard, T.; Moreau, L.; Charcosset, A. Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. *Theoretical and Applied Genetics* **2019**, *132*, 81–96.

Sweeney, D.W.; Sun, J.; Taagen, E.; Sorrells, M.E. Genomic Selection in Wheat. In *Applications of Genetic and Genomic Research in Cereals*; Elsevier, 2019; pp. 273–302.

Heslot, N.; Yang, H.P.; Sorrells, M.E.; Jannink, J.L. Genomic Selection in Plant Breeding: A Comparison of Models. *Crop Science* **2012**, *52*, 146. doi:10.2135/cropsci2011.06.0297.

Kemper, K.E.; Bowman, P.J.; Hayes, B.J.; Visscher, P.M.; Goddard, M.E. A multi-trait Bayesian method for mapping QTL and genomic prediction. *Genetics Selection Evolution* **2018**, *50*, 10. doi:10.1186/s12711-018-0377-y.

Moore, J.H.; Olson, R.S.; Schmitt, P.; Chen, Y.; Manduchi, E. How computational thought experiments can improve our understanding of the genetic architecture of common human diseases. *The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE)* **2018**, p. 23–30. doi:10.1162/isal_a_00012.

Harper, W.R.; Harris, D.H. The Application of Link Analysis to Police Intelligence. *Human Factors* **1975**, *17*, 157–164, [https://doi.org/10.1177/001872087501700206]. doi:10.1177/001872087501700206.

Page, L.; Brin, S.; Motwani, R.; Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab, 1999. Previous number = SIDL-WP-1999-0120.

Park, J.; Yook, S.H. Bayesian Inference of Natural Rankings in Incomplete Competition Networks. *Scientific Reports* **2014**, *4*, 6212. doi:10.1038/srep06212.

Simko, I.; Pechenick, D.A. Combining partially ranked data in plant breeding and biology: I. Rank aggregating methods. *Communications in Biometry Crop Science* **2010**, *5*, 41 – 55.

Lande, R.; Thompson, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **1990**, *124*, 743–756.

Haley, C.S.; Visscher, P.M. Strategies to Utilize Marker-Quantitative Trait Loci Associations. *Journal of Dairy Science* **1998**, *81*, 85–97. doi:10.3168/jds.S0022-0302(98)70157-2.

Ribaut, J.M.; Ragot, M. Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations, and alternatives. *Journal of Experimental Botany* **2006**, *58*, 351–360. doi:10.1093/jxb/erl214.

Togashi, K.; Lin, C.Y. Theoretical efficiency of multiple-trait quantitative trait loci-assisted selection. *Journal of Animal Breeding and Genetics* **2010**, *127*, 53–63. doi:10.1111/j.1439-0388.2009.00817.x.

Hospital, F. Selection in backcross programmes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2005**, *360*, 1503–1511. doi:10.1098/rstb.2005.1670.

Ragot, M.; Biasiolli, M.; Delbut, M.; Dell'Orco, A.; Malgarini, L.; Thevenin, P.; Vernoy, J.; Vivant, J.; Zimmermann, R.; Gay, G. Marker-assisted backcrossing: a practical example. *COLLOQUES-INRA* **1995**, pp. 45–45.

Collard, B.C.; Mackill, D.J. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2008**, *363*, 557–572. doi:10.1098/rstb.2007.2170.

Wallace, J.G.; Rodgers-Melnick, E.; Buckler, E.S. On the Road to Breeding 4.0: Unraveling the Good, the Bad, and the Boring of Crop Quantitative Genomics. *Annual review of genetics* **2018**, *52*, 421–444.

Evans, J. *Optimization algorithms for networks and graphs*; Routledge, 2017.

Festa, P. The shortest path tour problem: problem definition, modeling, and optimization. e-book semanticscholar.org https://pdfs.semanticscholar.org/d2f6/39c285fe6f1f549b403aa52f9c22ecc5e37e.pdf, 2019.

Ritchie, S.W.; Chintamanani, S.P.; Dunn, M.; Ersöz, E.S.; Foster, D.J.; Martin, N.F.; Skibbe, D.S.; Tucker, D.M. Genetic Markers Associated with Increased Fertility in Maize, 2017. US Patent App. 15/117,491.

Fedoryuk, M. Asymptotic methods in analysis. In *Analysis I*; Springer, 1989; pp. 83–191.

Smirnov, V. On the estimation of a path integral by means of the saddle point method. *Journal of Physics A: Mathematical and Theoretical* **2010**, *43*, 465303.

Ruder, S. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747* **2016**.

Sebastiani, F. Machine learning in automated text categorization. *ACM computing surveys (CSUR)* **2002**, *34*, 1–47.

Ortiz, R.; Crossa, J.; Franco, J.; Sevilla, R.; Burgueño, J. Classification of Peruvian highland maize races using plant traits. *Genetic Resources and Crop Evolution* **2008**, *55*, 151–162.

Grobman, A. *Races of maize in Peru: their origins, evolution and classification*; Vol. 915, National Academies, 1961.

Wright, K.R.; Wright, R.M.; Valencia, Z.; McEwan, G.F.; others. *Moray: Inca engineering mystery.*; American Society of Civil Engineers (ASCE), 2010.

Cooper, M.; Messina, C.D.; Podlich, D.; Totir, L.R.; Baumgarten, A.; Hausmann, N.J.; Wright, D.; Graham, G. Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. *Crop and Pasture Science* **2014**, *65*, 311–336.

Byrum, J.; Davis, C.; Doonan, G.; Doubler, T.; Foster, D.; Luzzi, B.; Mowers, R.; Zinselmeier, C.; Kloeber, J.; Culhane, D.; others. Advanced analytics for agricultural product development. *Interfaces* **2016**, *46*, 5–17.

Kang, M.S. Using genotype-by-environment interaction for crop cultivar development. In *Advances in agronomy*; Elsevier, 1997; Vol. 62, pp. 199–252.

Löffler, C.M.; Wei, J.; Fast, T.; Gogerty, J.; Langton, S.; Bergman, M.; Merrill, B.; Cooper, M. Classification of maize environments using crop simulation and geographic information systems. *Crop Science* **2005**, *45*, 1708–1716.

Crossa, J.; Fox, P.; Pfeiffer, W.; Rajaram, S.; Gauch, H. AMMI adjustment for statistical analysis of an international wheat yield trial. *Theoretical and Applied Genetics* **1991**, *81*, 27–37.

Messina, C.D.; Podlich, D.; Dong, Z.; Samples, M.; Cooper, M. Yield–trait performance landscapes: from theory to application in breeding maize for drought tolerance. *Journal of experimental botany* **2010**, *62*, 855–868.

Lecoeur, J.; Poiré-Lassus, R.; Christophe, A.; Pallas, B.; Casadebaig, P.; Debaeke, P.; Vear, F.; Guilioni, L. Quantifying physiological determinants of genetic variation for yield potential in sunflower. SUNFLO: a model-based analysis. *Functional plant biology* **2011**, *38*, 246–259.

Sinclair, T.R.; Messina, C.D.; Beatty, A.; Samples, M. Assessment across the United States of the benefits of altered soybean drought traits. *Agronomy Journal* **2010**, *102*, 475–482.

Dinov, I.D. *Data science and predictive analytics: Biomedical and health applications using R*; Springer, 2018.

Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260.

Cooper, M.; Messina, C.D.; Podlich, D.; Totir, L.R.; Baumgarten, A.; Hausmann, N.J.; Wright, D.; Graham, G. Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. *Crop and Pasture Science* **2014**, *65*, 311–336.

Hijmans, R.J.; Cameron, S.E.; Parra, J.L.; Jones, P.G.; Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology* **2005**, *25*, 1965–1978.

Ferrier, S.; Guisan, A. Spatial modelling of biodiversity at the community level. *Journal of applied ecology* **2006**, *43*, 393–404.

Zhong, H.; Li, X.; Lobell, D.; Ermon, S.; Brandeau, M.L. Hierarchical modeling of seed variety yields and decision making for future planting plans. *Environment Systems and Decisions* **2018**, *38*, 458–470.

Marko, O.; Brdar, S.; Panić, M.; Šašić, I.; Despotović, D.; Knežević, M.; Crnojević, V. Portfolio optimization for seed selection in diverse weather scenarios. *PloS one* **2017**, *12*, e0184198.

Hofman, J.M.; Sharma, A.; Watts, D.J. Prediction and explanation in social systems. *Science* **2017**, *355*, 486–488.

Fick, S.E.; Hijmans, R.J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology* **2017**, *37*, 4302–4315.

Atlin, G.N.; Cairns, J.E.; Das, B. Rapid breeding and varietal replacement are critical to adaptation of cropping systems in the developing world to climate change. *Global Food Security* **2017**, *12*, 31–37. doi:10.1016/j.gfs.2017.01.008.

Ersoz, E. Evaluation of geo-climactic location-of-origin features for effects on phenotype prediction accuracy with ML models: A case study in *Arabidopsis thaliana*. F1000Research 8:203 (poster) https://doi.org/10.7490/f1000research.1116446.1, 2019.

Zhong, H.; Li, X.; Lobell, D.; Ermon, S.; Brandeau, M.L. Hierarchical modeling of seed variety yields and decision making for future planting plans. *Environment Systems and Decisions* **2018**, *38*, 458–470.

Saha, M.; Mitra, P. Recurrent neural network based prediction of Indian summer monsoon using global climatic predictors. 2016 International Joint Conference on Neural Networks (IJCNN). IEEE, 2016, pp. 1523–1529.