# Machine Learning Techniques and Applications for Plant Breeding

University of California, Davis
Aleksandra Taranov
MSc International Agricultural Development

**Executive Summary:**

Given that most of the world's poor work in agriculture, any improvements in crop yield or drought/disease resistance are of special interest to practitioners of international development. As a result, I have been very interested in plant breeding and the technological advances that improve or speed up breeding in order to accomplish these goals in staple crops such as corn, wheat, and rice. Because many of these advances are occurring in the application of machine learning, I have chosen to research machine learning applications in plant breeding for both genomic selection and image classification. Recent trends show that machine learning and especially deep learning are outperforming robust statistical genetics methods but that these older method are still very useful in setting a baseline against which to compare newer methods. Additionally, there is a huge trend towards using mathematical methods to better understand the importance of variables after a machine learning algorithm is used to select a variety. Finally, there is a lot of research being done in modeling multiple traits and multiple environments, which is essential given the difficulties of studying staple crops in different environments. This paper begins with a historical and technical introduction to machine learning algorithms, followed by discussion about how statistical genetics and machine learning are being used in crop selection, and concludes with discussion about how other machine learning applications such as image classification for diseases and drought tolerance could help increase the number of observations in a way that aids scientists working on genomic selection research. Overall, it offers a promising view of the many ways that advanced statistical and computational methods have been transforming the world of plant breeding and with it improving global outcomes in agriculture.

## Introduction

Plant breeding has changed a lot with new technology. Early plant breeding was based on phenotypic selection of certain traits that were desirable. In research centers like CIMMYT in Mexico, pedigrees were used to keep track of the breeding history and experiment with crossing and back-crossing varieties, and molecular markers were used to practice genomic selection as well (Crossa 2010). However, with the explosion of genomic data, statistical and machine learning methods are necessary to make use of the data in order to make selection more accurate or faster by lowering the number of plants or grow cycles needed to keep breeding plants. Moose and Mum argue that although there is "intense interest among plant breeders and crop scientists…[methods of molecular plant breeding] have received relatively little attention from the majority of plant biologists engaged in basic scientific research" (Moose and Mum, 2008). Hopefully, this capstone paper will encourage greater interest in machine learning methods and their applications in both research and practical crop improvement contexts.

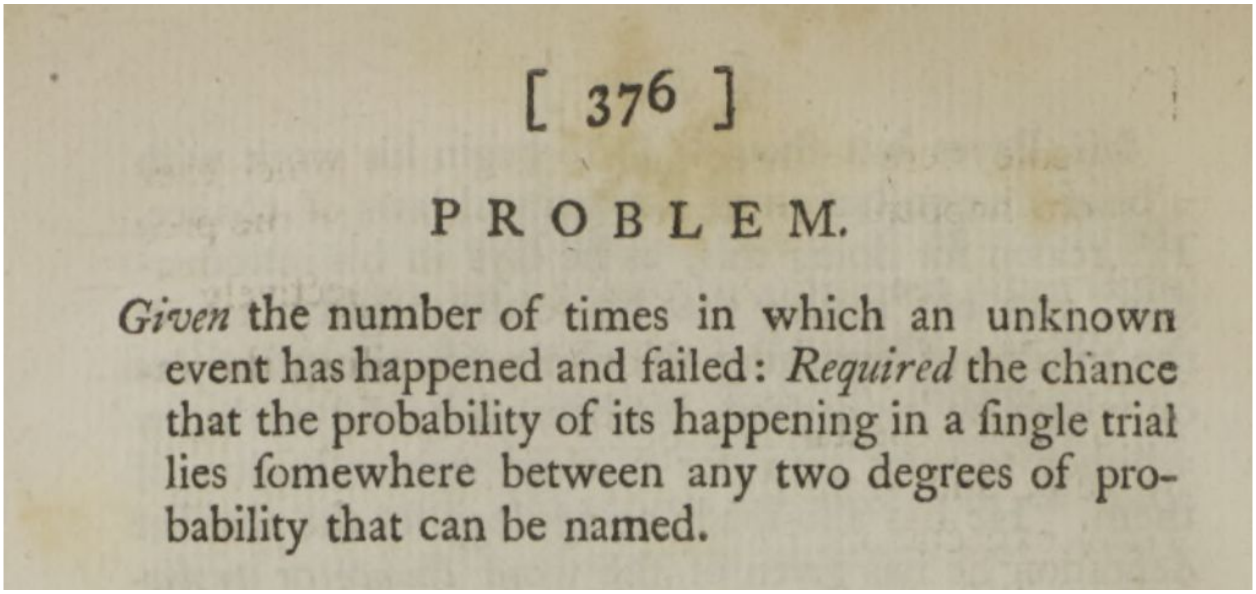
## Part 1: Historical and Technical Introduction to Machine Learning

### A Brief History of Machine Learning

Machine learning is sometimes defined as such: "a field of computer science that evolved from studying pattern recognition and computational learning theory in artificial intelligence...the learning and building of algorithms that can learn from and make predictions on data sets" (Simon et al, 2016). Rather than following exact instructions, this allows for algorithms that use inputs to construct models that can make data-driven predictions. Although machine learning represents a novel and emerging set of techniques and applications, it would not have been possible without several centuries of progress that preceded it, and a firm understanding of its origins is necessary to understand its potential future. Starting in the 18th century, advances in statistical theory and mathematical methods set the stage for machine learning, which then exploded in the 20th century as computing became increasingly powerful.

## Bayes Theorem

One of the earliest developments that strongly contributed to machine learning was bayesian inference, a statistical  method in which the probability of a hypothesis is updated with new information. Bayes theorem is named for the work done by Thomas Bayes and compiled by his friend Richard Price in a 1763 posthumous publication titled “An Essay towards Solving a Problem in the Doctrine of Chances” (Bayes, 1763). It seems that much of the work was also written by Price himself. Statistician and historian Stephen Stigler argues that Price was motivated by wanting to refute David Hume on miracles and prove the existence of god and was actually the first person to apply Bayes’ theory (Stigler, 2018). In any case, Bayes himself describes the problem as such: “Given the number of times in which an unknown event has happened and failed: Required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named” (Bayes, 1763). Although this early work did not include a discrete form of the theorem, it remains the first known work to tackle conditional probability, probability priors, and the question of how one predicts the results of a new observation.



[ 376 ]

PROBLEM.

*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a ſingle trial lies ſomewhere between any two degrees of probability that can be named.

A version of the actual Bayes Theorem was first published in by Pierre-Simon Laplace (1774). Although he is believed to have published this without any prior knowledge of Bayes' work, he later mentioned Bayes in his 1814 Essai Philosophique sur les Probabilites. (Gorroochurn, 2016). In this work he notes that he had expounded on Bayes' theory concerning "the probability of causes and future events, concluded from events observed" (Laplace, 1814). His big contribution was a discrete formulation, which goes like this: "The probability of the existence of any one of these causes is then a fraction whose numerator is the probability of the event resulting from this cause and whose denominator is the sum of the similar probabilities relative to all the causes; if these various causes, considered à priori, are unequally probable, it is necessary, in place of the probability of the event resulting from each cause, to employ the product of this probability by the possibility of the cause itself. This is the fundamental principle of this branch of the analysis of chances which consists in passing from events to causes." (Laplace, 1814). Laplace then uses this principle of inverse probability to solve a problem about drawing black and white tickets from a ballot and asking the probability that the next ticket would be white. This can be written mathematically and compared to the common formula of bayes theorem on the right.

$$P(\theta_i \mid F) = \frac{P(F \mid \theta_i)}{\sum_{j=1}^{n} P(F \mid \theta_j)} \qquad\qquad P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

However, bayesian statistics fell out of favor and was later revived by mathematicians such as Sir Harold Jeffreys, who famously stated that Bayes Theorem “is to the theory of probability what the Pythagorean theorem is to geometry” (Jeffreys, 1973).

### Least Squares

Another notable contribution was the independent discovery of least squares by both Gauss and Legendre in the early 19th century (Stigler, 1981). Least squares is a standard approach to regression and is used in data fitting that involves minimizing the sum of the squares of the residuals.

### Markov Chains

Additionally, in the 20th century, Andrey Markov developed a stochastic model called the markov chain in which there are sequences of events based on probabilities that depend only on the state obtained from the previous event. Markov chains are essential for probabilistic modeling and are the basis for Markov Chain Monte Carlo (MCMC) methods, which are used extensively in bayesian statistics and machine learning.

## Artificial Intelligence (AI) in 1950-1979

Although these developments in mathematics and statistics were crucial, computing is what truly brought machine learning into the spotlight. In 1950, Alan Turing proposed a learning machine that could become intelligent (Turing, 1950). The following year Marvin Minsky and Dean Edmonds built the first neural network, the Stochastic Neural Analog Reinforcement Calculator, otherwise known as SNARC (Russell, 2003). SNARC learned from experience and was used to search a maze. The following year, in 1952, Arthur Samuels created the first machine learning computer program, which could play checkers and learn how to improve its moves (Leonard, 1999). Soon, scientists developed a machine to play tic tac toe and the nearest neighbors algorithm was invented. Although this was a time of staggering progress, there were also huge setbacks. In 1957, Frank Rosenblatt invented the Perceptron at the Cornell Aeronautical Laboratory, which was a simple linear classifier that in a large network could become a powerful model. This was widely celebrated until Marvin Minsky and Seymour Papert published Perceptrons in 1967, claiming that neural networks are fundamentally limited and would not be able to solve problems like the exclusive or (XOR) logical function (Osterland, 2019). Pessimism about the effectiveness of machine learning methods and the slow pace of breakthroughs led to a lack of research in the 1970s sometimes referred to as the AI Winter.

### AI in the 1980s and onwards

In the 1980s, backpropagation is rediscovered and research in machine learning, especially neural networks, resumes. It is revitalized because of the probabilistic approach based on bayesian statistics that allows for uncertainty in parameters to be incorporated into models. This allows for a more data-driven approach and for there to be more layers in neural networks, allow them to learn over time. Machine learning continued to develop after this point in time, with the invention of support vector machines, random decision forest, and ongoing research in areas of image processing and more.

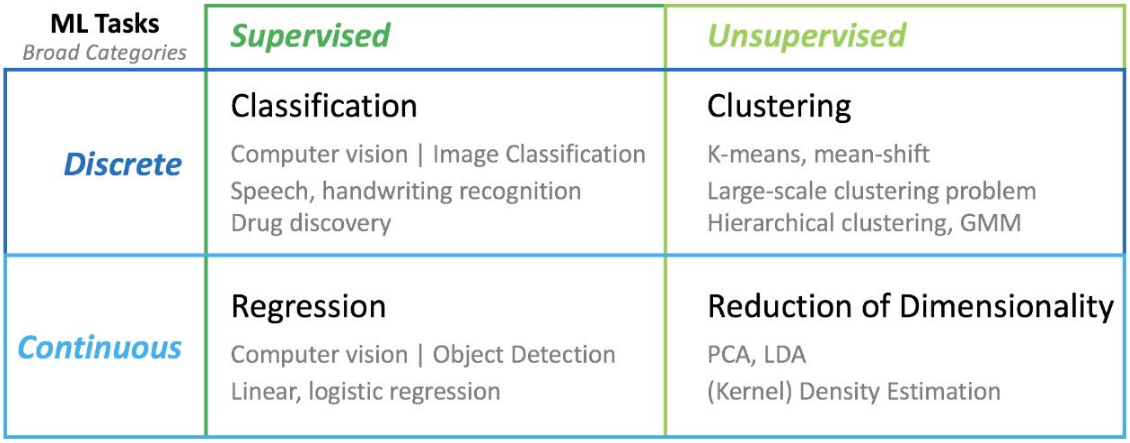
In 1980, we see Kunihiko Fukushima's work on the neocognitron, an artificial neural network (ANN) that was a precursor to convolutional neural networks (CNN) (Fukushima, 1980). In current time, there is an emphasis on deep learning and strong potential for future research in this area. Deep learning is a subset of machine learning focused on deep artificial neural networks, which are a set of newer and more powerful algorithms that have multiple hidden layers. Because these newer methods are very intensive, there is more focus on GPUs. In any case, current machine learning research focuses on both improving algorithms and finding new applications. However, it is critical to see how early statistical methods, specifically bayesian theory, set the stage for the probabilistic thinking that, combined with computational intensity, allowed for the explosion of machine learning.

## Machine Learning Technical Overview

The following offers a quick technical overview of existing machine learning algorithms, which is necessary for the analysis and conclusions of this paper.

### Learning Styles

Machine learning algorithms can be grouped by learning style into the categories of supervised, unsupervised, and semi-supervised learning. This grouping helps a user think through the input data and select a model preparation process that is suitable for the specific problem at hand. In supervised data, the input data is referred to as training data because it trains a model that then allows for accurate conclusions given new data. This works well with pre-labeled or categorized data and can be applied to the problems of classification or regression using algorithms such as logistic regression. On the other hand, in unsupervised learning, data is not labeled, so the machine is given inputs and must find patterns and relationships from them without the same structure. Typical problems that this approach is used for include clustering or association rule learning using algorithms such as k-means.



| **ML Tasks**<br>*Broad Categories* | *Supervised* | *Unsupervised* |
|---|---|---|
| *Discrete* | Classification<br>Computer vision \| Image Classification<br>Speech, handwriting recognition<br>Drug discovery | Clustering<br>K-means, mean-shift<br>Large-scale clustering problem<br>Hierarchical clustering, GMM |
| *Continuous* | Regression<br>Computer vision \| Object Detection<br>Linear, logistic regression | Reduction of Dimensionality<br>PCA, LDA<br>(Kernel) Density Estimation |

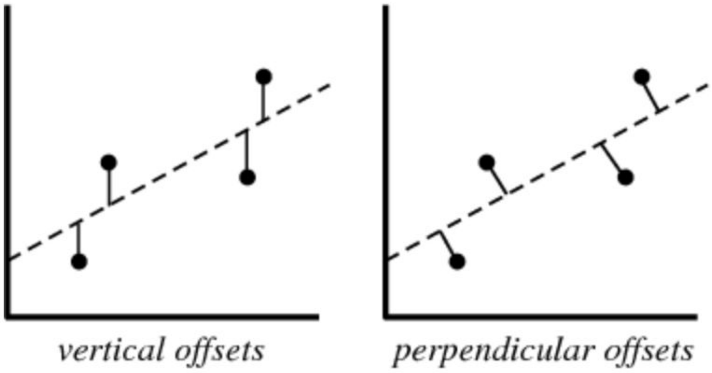
Sometimes we see a mixture of labeled and unlabeled inputs, which is where semi-supervised learning comes in. In this case the model must organize data but also make predictions. There are other instances of mixed models as well and instances of algorithms such as generative adversarial networks (GAN) which are unsupervised but used supervised methods in the training portion. Neural networks can be either supervised or unsupervised based on whether the desired output is known and are critical to understanding the field of machine learning.

## Algorithms by Functions

An alternate way of grouping machine learning algorithms involves looking at their functions. The broad problems which machine learning can tackle are sometimes grouped into regression or classification (which are usually supervised) and clustering, association, or dimension reduction aka generalization (which are often unsupervised.) However, there are also families of algorithms such as decision trees that can build regression or classification models that have a distinctive mechanism and are therefore grouped separately. This technical overview will therefore begin with the broad functions named above and then discuss several distinctive groupings of algorithms.

## Regression

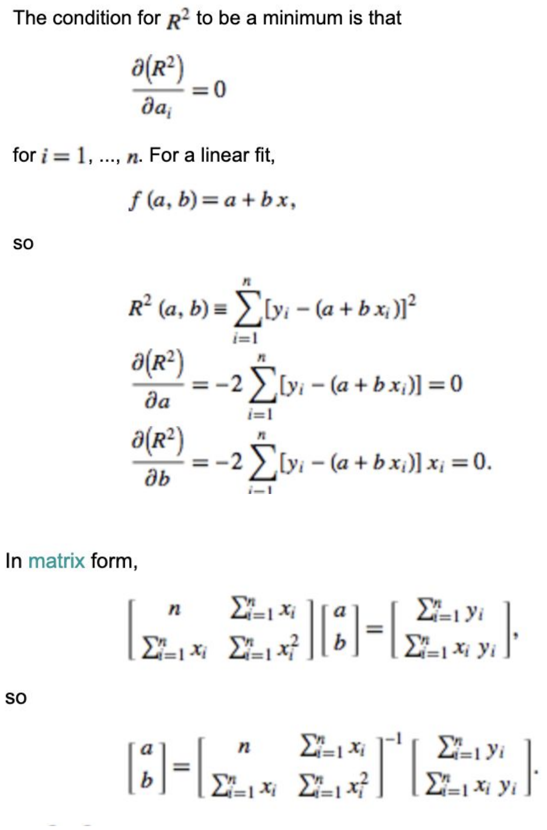
Regression analysis is a statistical tool that models the relationship between quantitative variables using measurements of error made by the model. In its simplest form we have simple linear regression, in which we have one explanatory variable and a dependent variable and a line is drawn between all points. Multiple explanatory variables would result in the term multiple linear regression or just linear regression and multiple correlated dependent variables that are predicted would be multivariate linear regression. However, these all have the same basic underlying principles.



Least Squares Fitting. (n.d.). Retrieved from http://mathworld.wolfram.com/LeastSquaresFitting.html

Least Squares is a common method that is used to find the best fit line in regression by minimizing the sum of squares of the errors. The vertical offsets are minimized because it provides analytically simpler form and squares of the errors are used so that residuals can be treated as differentiable. It can be used in both linear and polynomial regression and works by creating a function in which we find the sum of squares of the vertical deviations and then, using calculus, take the derivative and set it to 0 to minimize it. This can be written in matrix form as well and solved with shortcuts using linear algebra.

$$R^2 \equiv \sum [y_i - f(x_i, a_1, a_2, \ldots, a_n)]^2$$



The condition for $R^2$ to be a minimum is that

$$\frac{\partial\left(R^2\right)}{\partial a_i} = 0$$

for $i = 1, \ldots, n$. For a linear fit,

$$f\,(a, b) = a + b\,x,$$

so

$$R^2\,(a, b) \equiv \sum_{i=1}^{n} [y_i - (a + b\,x_i)]^2$$

$$\frac{\partial\left(R^2\right)}{\partial a} = -2 \sum_{i=1}^{n} [y_i - (a + b\,x_i)] = 0$$

$$\frac{\partial\left(R^2\right)}{\partial b} = -2 \sum_{i=1}^{n} [y_i - (a + b\,x_i)]\,x_i = 0.$$

In matrix form,

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\,y_i \end{bmatrix},$$

so

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\,y_i \end{bmatrix}.$$

Least Squares Fitting. (n.d.). Retrieved from http://mathworld.wolfram.com/LeastSquaresFitting.html

Although least squares is a mathematically elegant solution to finding the best fit line, the computational method called gradient descent that uses it is sometimes faster. Instead of using the formula once, gradient descent is a step by step minimization algorithm in which the formula tells you the next position each time in order to go down to the steepest descent. It picks an initialization point and a learning rate and follows the slope down with each step until it minimizes it. This may seem clunkier but is effective and often gets results faster. This step from mathematical solutions to slightly less elegant but more efficient methods is key in the differentiation between pure statistical methods and evolving machine learning methodology.

Occasionally, instead of fitting a best line or polynomial, it makes more sense to use a nonparametric method and find the actual curve that is between all of the points. This is called

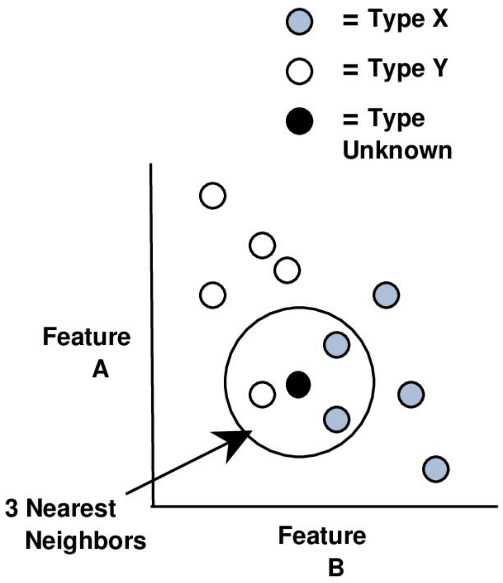
Locally Estimated Scatterplot Smoothing (LOESS) or local regression because fitting as point x is weighted based on the points at that point in the x axis. This can work for fitting a smooth curve between two variables or a smooth surface between an outcome and many predictor variables.

Finally, we also examine logistic regression, in which we analyze how multiple independent variables relate to a categorical dependent variable and predict the probability of an event occuring using a logistic curve. The basic idea is that $p=\alpha+\beta x$ is not a good model because values will fall outside of 0 to 1 when calculating odds. Instead, we transform the odds using the natural logarithm and model that as a linear function (Peng, Lee & Ingersoll, 2002). The parameters are alpha and beta, x is the explanatory variable, and p is the probability of the outcome:

$$logit(y)=ln\ (odds)=ln\ (p/(1-p)\ \ =a+\beta\chi$$

There are other methods such as stepwise regression, ridge regression, elastic net, Multivariate Adaptive Regression Splines (MARS), and Least Absolute Shrinkage and Selection Operator (LASSO),  that are useful for specific situations that rely on the core principles outlined to solve regression problems.

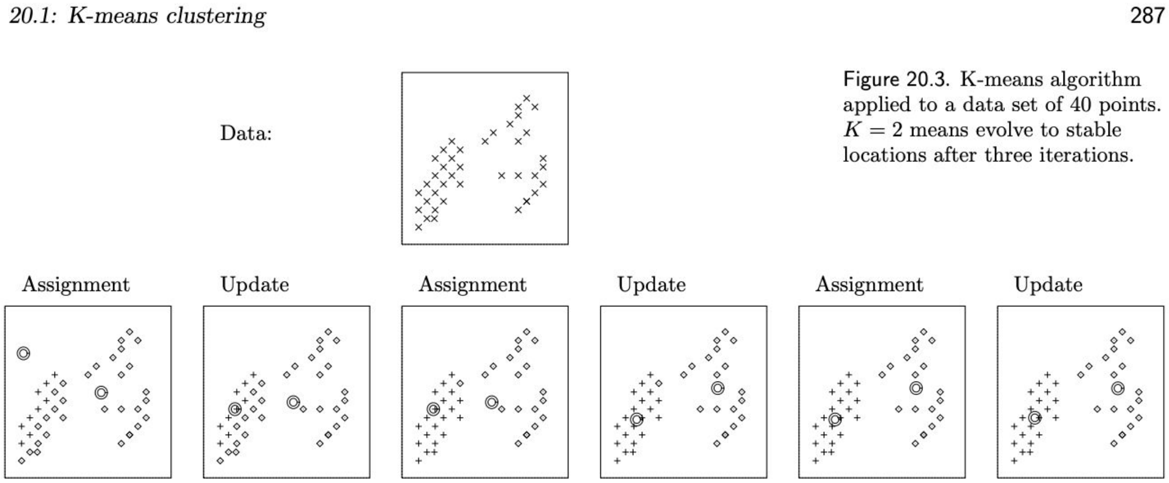
## Classification



Goodman, Erik & Pei, Min & Chia-shun, Lai & Hovland, P. (1994). Further Research on Feature Selection and Classification Using Genetic Algorithms.

Classification algorithms involve splitting objects based on attributes. Examples include the previously discussed logistic regression as well as k-Nearest Neighbor (kNN), Learning Vector Quantization (LVQ), Self-Organizing Map (SOM), Locally Weighted Learning (LWL), and Support Vector Machine (SVM). Let’s closely examine one of these algorithms, the kNN. KNN is a nonparametric method that works by grouping similar items (which can be defined by proximity using the distance function or other measurements such as the Manhattan.) The following diagram depicts a simple example in which an unlabeled data point must be categorized into one of two groups.

## Clustering



20.1: K-means clustering 287

Figure 20.3. K-means algorithm applied to a data set of 40 points. $K = 2$ means evolve to stable locations after three iterations.

MacKay, D. J. (2003). *Information theory, inference, and learning algorithms.* Cambridge: Cambridge University Press.

Clustering involves grouping objects based on unknown features and is similar to classification except that the data is unlabeled and it is unsupervised. One commonly used algorithm is k-Means, which finds the cluster with the nearest mean and then calculates the new means of the centers of the clusters. A modification can be made to use median instead of mean, which is called k-Medians. Expectation Maximization (EM) and Hierarchical Clustering also fall into the same family of algorithms.

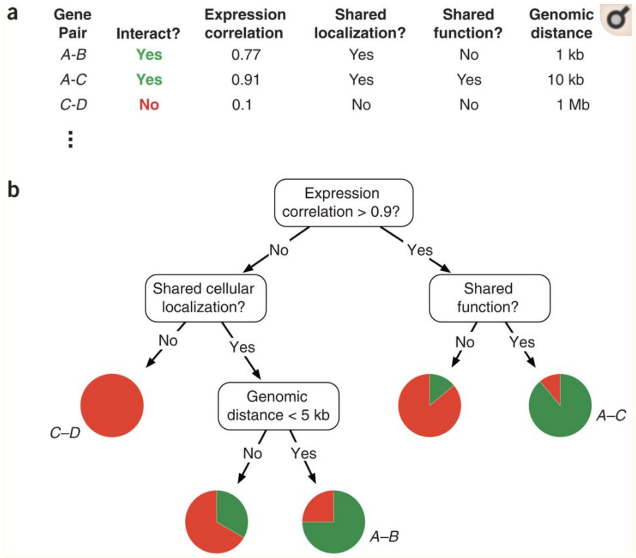
## Dimensionality Reduction

Dimensionality reduction or generalization is an unsupervised approach towards simplifying data which can be used for classification or regression purposes. The most common dimensionality reduction algorithm is Principal Component Analysis (PCA), which was invented in 1901 by statistician Karl Pearson (Pearson, 1901). PCA reveals the internal structure of the data. It works by first standardizing the data, then doing a computation on the symmetric covariance matrix to identify correlations, and finally computing the eigenvectors and eigenvalues of that matrix to identify the principal components. It is sometimes used to pre-process data for neural networks.

### Bayesian

The family of bayesian algorithms explicitly apply bayes theorem to do classification or regression. The most fundamental one is called Naive Bayes, but depending on the assumption of how the features are distributed, you may end up using Gaussian Naive Bayes or Multinomial Naive Bayes.

### Decision Tree



a

| Gene Pair | Interact? | Expression correlation | Shared localization? | Shared function? | Genomic distance |
|---|---|---|---|---|---|
| *A-B* | Yes | 0.77 | Yes | No | 1 kb |
| *A-C* | Yes | 0.91 | Yes | Yes | 10 kb |
| *C-D* | No | 0.1 | No | No | 1 Mb |
| ⋮ | | | | | |

Figure 1
A hypothetical example of how a decision tree might predict protein-protein interactions

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees?. Nature biotechnology, 26(9), 1011–1013. doi:10.1038/nbt0908-1011

Decision tree algorithms depend on choosing tree-like branches to look at decisions and arrive at a certain conclusion (Kingsford, 2008). The following diagram shows how a decision tree can classify items based on a series of questions about features. Decision trees often become more powerful as ensembles, in which you case you may end up seeing boosting or random forest techniques. According to Kinsford, "In the random forests approach, many different decision trees are grown by a randomized tree-building algorithm," whereas boosting involves reweighting training data over and over again to improve results. These techniques are commonly used in computational biology to make good predictions but are also used in finance and other industries.

### Artificial Neural Networks (ANN)

Finally, neural networks are inspired by the structure of the human brain and look at examples in order to perform tasks without explicit instruction.

## Part 2: Genomic Selection and Phenotypic Prediction

Advancements in genetic sequencing in the last fifty years brought about the development of new statistical and computational methods to make use of this data. Medicine, animal breeding, and plant breeding borrow methods from each other, but some methods work better in certain fields than others. Additionally, there have been trends away from traditional statistical and genetic research methods towards machine learning methods. Here we'll outline the new emphasis on genomic prediction over genome-wide association studies (GWAS), the continuing impact of BLUP, and new machine learning methodologies that are being developed in current plant breeding research.

Genome-wide association studies (GWAS) that map alleles using the principle of linkage disequilibrium (LD) have fallen from favor in many medical communities. GWAS uses LD between genotyped single nucleotide polymorphisms (SNPs) and variants, and the strength of association between alleles is affected by allele frequencies. However, Ikegawa describes fading enthusiasm around this method, “The GWAS society is facing severe criticisms. A frequent complaint is that GWAS results mean little to patients due to the small effect of variants on disease risk and their relatively small contribution to common disease etiology. This claim might be partially correct” (Ikegawa, 2012). However, some researchers argue that GWAS is more effective for studying plants than humans. After all, a 2011 paper in Genome Biology makes the following argument: “When human GWAS find associations that have genome-wide significance, the SNPs explain only a tiny fraction of the phenotypic variation revealed by family-based studies. But the results of recent GWAS in plants (in Arabidopsis thaliana, rice, and maize) have explained a much greater proportion of the phenotypic variation than that explained by human GWAS studies. It seems that, in plants at least, the assumption that common genetic variation explains common phenotypic variation holds” (Brachi, 2011). Although some researchers continue to use GWAS methods to study plants, much of the research focus has turned away from mapping specific genes.

Instead, much of the new research focuses on genomic prediction or predicting phenotypes based on genetic and environmental information. Prediction methods themselves go back as far as the 1950s and 1960s, when Henderson developed best linear unbiased prediction (BLUP) in the context of animal breeding (Henderson, 1950). Henderson explains that since selection has occurred in herds from which researchers get data, the usual linear model methods should not be applied and unbiased estimators should be used instead. BLUP is similar to empirical Bayes estimates of random effects apart from the fact that sample-based estimates replace unknown variances. BLUP continues to be a solid method for genomic prediction and for this reason, many researchers compare new methods against BLUP to see whether their method is better. A recent paper evaluating machine learning methods for predicting phenotypes in yeast, rice, and wheat compares machine learning methods against the standard statistical genetic method BLUP. The researchers find that although machine learning methods outperformed BLUP, BLUP still did fairly well in all cases, especially when population structure was present (Grinberg, 2017). Therefore, BLUP remains a good statistical genetic method and is especially useful for testing out newer methods by comparison, but machine learning methods are outperforming it and are the current trend towards improving genomic prediction.

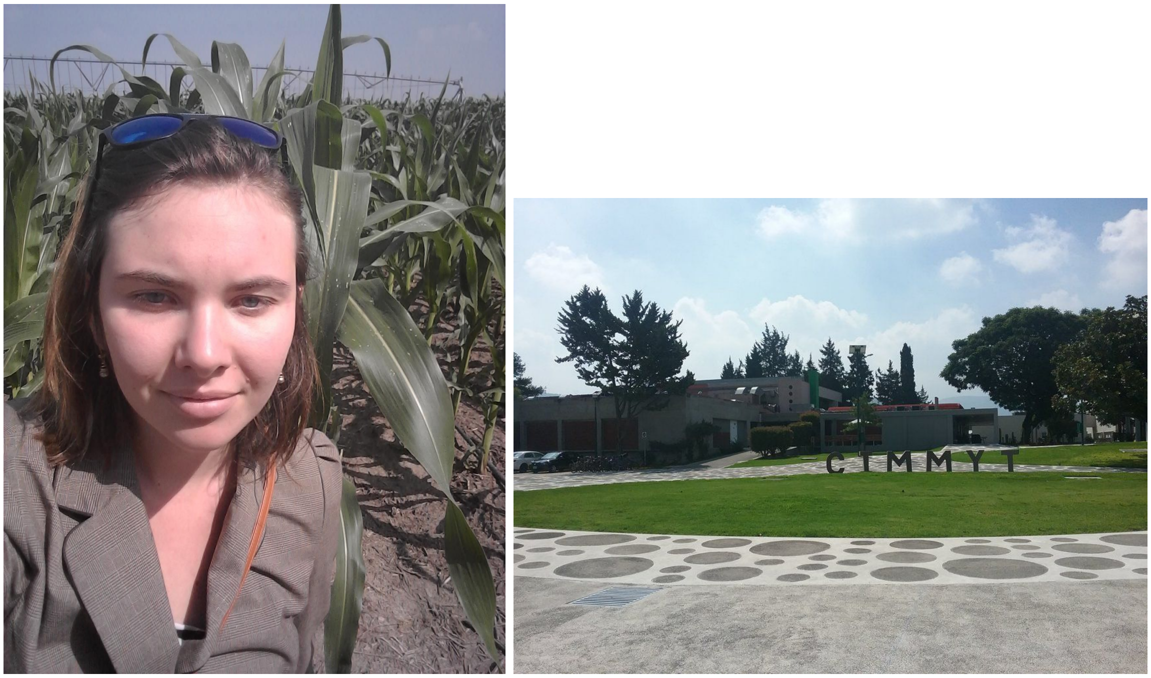
A study by Bloom et al examines 1008 haploid yeast strains using GWAS and Grinberg et al apply BLUP and a variety of machine learning methods to test their efficacy (Bloom 2013). They apply variants of linear regression such as elastic net, ridge regression, and lasso regression as well as the random forest method, gradient boosting machine (GBM), and support vector machines (SVM). According to their results, "at least one standard machine learning approach [sic] outperforms Bloom and BLUP on all but 6 and 5 traits, respectively" (Grinberg et al, 2017). For their data, GBM performed the best for most traits, although LASSO was the strongest for several as well. SVM tended to perform as well as BLUP but was more time consuming to run. These results were done using existing machine learning packages in R and show how promising the methods may be.

However, the yeast had similar lab environments and therefore that study could not take into account environmental influence. For that reason, we also examine several other studies on wheat, rice, and corn. In the messier wheat data, researchers study 254 strains of wheat, but the complete genotypes are not known, the crop is hexaploid, and the environment is not controlled

(Poland, 2012). The results of the Grinberg analysis on this study show that SVM performs the best for all traits, followed by BLUP. It is thought that these differences come from the complexity of the interactions between markers and environmental factors. This remains a huge challenge for genomic prediction in the context of plant breeding.

Despite the good results from many machine learning algorithms, they are sometimes criticized for being black box techniques that do not fully explain what is happening. For this reason, another major trend in genomic selection research is finding techniques to evaluate this. Dr. Runcie's recent paper proposes a RATE measure to better understand variable importance in Bayesian nonlinear methods. This is another promising line of research in this field (Crawford et al, 2018).

### Case Study: Deep Learning and Genomic Prediction at CIMMYT



**My photos: taken at CIMMYT in June 2018 during a summer internship**

A May 2019 paper from CIMMYT highlights the use of a deep learning model for genomic prediction that works for multiple traits (CIMMYT 2019). This can be viewed as an archetypical use of machine learning for plant breeding purposes. Therefore, we elaborate on the details on the model in order to give a more thorough view of how machine learning is actually implemented for genomic prediction.

To do their analysis, the CIMMYT researchers used seven wheat data sets that come from CIMMYT's Global Wheat Program. This same data was used previously by Juliana et al. to evaluate the "potential integration of GS for grain yield (GY) in bread wheat" (Juliana et al 2018). Although there are seven data sets, they come from only four elite yield trial (EYT) nurseries. These EYT nurseries were second-year yield trial nurseries that each had 1092 lines, which were developed using a selected bulk breeding method. First the plants were bulked at early generations until the individual plants were derived from F5, F6, or F7 lines. Then, 70,000 individual plant-derived lines were grown in 1m x 1m plots and visually chosen based on phenotypic characteristics such as disease and spike health. The 9000 lines that were chosen using this method were used for the first year yield trial. These were selected further using the same visual phenotypic selection process to get the 1092 lines that were called the EYT and used in this paper. These EYT nurseries were planted in mid-November at the Norman E. Borlaug Research Station in Sonora, Mexico and were irrigated with 500 mm of water. The nurseries had 39 trials of 28 lines (which results in 1092 total) and 2 high-yielding checks (Kachua and Borlaug) that were in an alpha lattice design with six blocks and three replications. The nurseries were evaluated for number of days from germination to 50% spike emergence, number of days from germination to 50% physiological maturity, grain yield in tons per hectare, and plant height in centimeters. Dataset 1 came from the 2013-2014 season and had 767 lines, dataset 2 came from 2014-2015 and had 775 lines, dataset 3 came from 2015-2016 and had 964 lines, and dataset 4 came from 2016-2017 and had 980 lines (Juliana et al). In each season they also studied six environments based on the level of irrigation and whether they had a flat or bed planting system, but not all six of these environments were studied in each of the 4 datasets. Dataset 5 was part of dataset 3 and was obtained the same way but instead looked at the discrete categories of grain color (yes or no), leaf rust (1 to 5), stripe rust (1 to 3) and the continuous category of yield. Dataset 6 and 7, which had 945 and 1145 lines respectively, were part of the wheat yield trial (YT) nurseries at CIMMYT and both had measurements of the continuous trait of wheat and the discrete trait of lodging. This sums up the phenotypic datasets.

In terms of genotypic data, of the 4368 lines in datasets 1 through 4, all were genotyped using genotyping-by-sequencing at Kansas State University. Researchers were left with 2038 markers after removing markers that had over 60% missing data, less than 5% minor allele frequency, or more than 10% percent heterozygosity. The lines in datasets 5 through 7 were similarly genotyped. All 7 of the datasets, with phenotypic and genotypic data, are freely available for download at the link hdl:11529/10548140.

Let’s take a closer look at each of the 7 datasets. It is worth thinking about which of the measurements in each were binary, ordinal, or continuous because, as discussed earlier, the categories of machine learning and types of algorithms depend on this information. In the supervised context, regression is used for continuous traits, while classification would be used for binary or ordinal traits. Dataset 1 had 1 binary category (height), 2 ordinal (days to heading and days to maturity), and 1 continuous trait (yield). It turned out that across the environments, height was category one 46.4% of the time and it was category two 53.6% of the time. The two ordinal traits were similarly distributed between categories, and the continuous trait of yield was 6 tons per hectare for three of the environments (bed with irrigation 5, flat with irrigation 5, and EHT) and 3 tons per hectare for LHT. In dataset 2, height was 47.2% of cases in category one and 52.8% of cases in category two. The same two ordinal traits of days to heading and days to maturity were again similarly distributed between categories. In terms of yield, it was 6 tons per hectare for bed with irrigation 5, flat with irrigation 5, and EHT and 3 tons per hectare for LHT again, but the bed with irrigation 2 was 4.3 tons per hectare. In dataset 3, the data for height was similar to the previous ones, as was the distribution of the ordinal traits, but the yield was above 6 tons per hectare for bed or flat with irrigation 5 and 4 tons per hectare in the beds with irrigation 2, and was less than 3 tons per hectare in flat with drip irrigation and LHT. Dataset 4 was similar again in terms of height and the two ordinal categories, but in terms of yield, it was 6 tons per hectare in EHT and bed or flat with irrigation 5 and less than 3 tons per hectare in flat beds with drip irrigation. Now, datasets 5 through 7 were different from datasets 1 through 4. Dataset 5 measured grain color and had 79.7% in category 1 and 20.3% in category 2. For leaf rust, category 1 was 9.2%, category 2 was 4.6%, category 3 was 11.7%, category 4 was 10.0%, and category 5 was an overwhelming 64.5% of cases. For stripe rust, category 1 was 90.4% of cases, category 2 was 0.90%, and category 3 was 8.7%. The average yield was 4.4 tons per hectare. In dataset 6, the ordinal trait of lodging was category 1 for 12.2% of cases, category 2 for 7.8% of cases, category 3 for 9.4% of cases, category 3 for 38.2% of cases, and category 5 for 32.4%. The average grain yield was 6.7 tons per hectare. For the final dataset, the average yield was 5.75 tons per hectare, and the ordinal trait of lodging had “50.6%, 14.1%, 16.4%, 10.9% and 8.0% of individuals in categories 1, 2, 3, 4 and 5, respectively” (CIMMYT 2019).

Using these datasets, the researchers set out to test out the prediction performance of multiple-trait deep learning with mixed phenotypes (MTDLMP) models against univariate deep learning (UDL) models. The reason for doing so is that although there is a lot of progress being made in developing genomic prediction approaches with continuous variables, it is harder to do multivariate prediction of mixed phenotypes where you have binary, ordinal, and continuous variables. Typically, in the case of mixed phenotypes, a separate univariate analysis is done for each trait, but this ignores correlations between multiple traits. Therefore, much of this CIMMYT paper is dedicated to comparing the MTDLMP and UDL using five-fold cross validation in which pearson’s correlation is used as a metric of prediction performance for continuous traits and percentage of cases correctly classified is used for the discrete cases.

All in all, researchers at CIMMYT used supervised deep learning in order to predict a number of features such as the continuous trait of yield (which is a regression problem) and discrete traits such as grain color or lodging (which is a classification problem.) In terms of results, MTDLMP was better than UDL in 4 out of 7 cases where G x E interaction was taken into account and 5 out of 7 in which it was not. However, these gains in prediction performance occurred only for the continuous trait of yield and not for the discrete variables. This can in part be attributed to the fact that the phenotypic correlations between traits were not high and that the number of markers was low. The application was novel since there is no existing multi-trait model for simultaneous prediction of mixed phenotypes. The strongest argument for the two deep learning models (UDL and MTDLMP) is that there is currently no other known alternative for this particular task. Therefore, they recommend that plant breeders and plant geneticists have deep learning methods as part of their toolkits.

## Part 3: Phenotypic Classification

Phenotypic classification through image processing is often considered to be a completely separate endeavor from genomic selection. However, I propose that the two inquiries should be examining each other much more closely not only to learn from each others' methods but also because phenotypic classification can solve some of the lingering issues in genomic selection and speed up many processes by providing larger samples of data. A major problem in genomic selection models has to do with mathematical issues that arise from having many more genetic markers than numbers of phenotypic observations. Additionally, researchers are increasingly interested in studying multiple traits and multiple environments and their interactions with the genetic markers. All of these goals are best achieved not only by improving prediction through advanced machine learning methods but also by employing machine learning methods to generate greater amounts of data with which to use these methods.

The field of computer vision for agriculture has been developing but without being applied specifically to these problems. We see research in using machine vision to identify types of plant stress using deep convolutional neural networks (Ghosal et al, 2018), deep learning for classification of tomato diseases (Brahimi et al, 2017), and also applications involving yield assessment. Machine learning has been very successful in quickly and accurately identifying plant diseases based on images of leaves and could possibly be applied towards gathering large amounts of data to be used in genomic selection.

## Conclusion

All in all, traditional statistics and new machine learning algorithms continue to play import roles in plant breeding, through genomic selection as well as image processing and phenotypic classification. These methods are often applicable across many academic and industry problems, although some work better in certain situations than others. The trends are towards developing more machine learning methodologies, using mathematics to more clearly interpret these black box methods, and towards developing models for multiple traits and multiple environments. A major obstacle is the need for more data to avoid problems that occur when there are many more genetic markers than observations and machine learning could be useful in acquiring that data. Even if it is not applied in this context, the many useful applications of machine learning in plant breeding suggest that these techniques will continue to evolve and influence the field of plant breeding.

## References

Akhtar, A. Khanum, S. a. Khan, A. Shaukat, Automated Plant Disease 401 Analysis (APDA): Performance Comparison of Machine Learning Techiques, in: 2013 *11th International Conference on Frontiers of Information 403 Technology*, {IEEE} Computer Society, 2013, pp. 60–65.

Bayes, Thomas. (1763). "An Essay Towards Solving a Problem in the Doctrine of Chances." Philosophical Transactions of the Royal Society of London, 53: 370– 418.

Bloom, J. S., I. M. Ehrenreich, W. T. Loo, T.-L. V. o. Lite, and L. Kruglyak (2013). Finding the sources of missing heritability in a yeast cross. Nature 494(7436), 234–7.

Bock, C. H. Poole, P. E. Parker & T. R. Gottwald (2010) Plant Disease Severity Estimated Visually, by Digital Photography and Image Analysis, and by Hyperspectral Imaging, Critical Reviews in Plant Sciences, 29:2, 59-107,DOI: 10.1080/07352681003617285

Brachi, B., Morris, G. P., & Borevitz, J. O. (2011). Genome-wide association studies in plants: The missing heritability is in the field. *Genome Biology*, 12(10), 232. doi:10.1186/gb-2011-12-10-232

Brahimi, Mohammed, Kamel Boukhalfa & Abdelouahab Moussaoui (2017) Deep Learning for Tomato Diseases: Classification and Symptoms Visualization, Applied Artificial Intelligence, 31:4, 299-315, DOI: 10.1080/08839514.2017.1315516

CIMMYT (Osval A. Montesinos-López, Javier Martín-Vallejo, José Crossa, Daniel Gianola, Carlos M. Hernández-Suárez, Abelardo Montesinos-López, Philomin Juliana and Ravi Singh). (2019). New Deep Learning Genomic-Based Prediction Model for Multiple Traits with Binary, Ordinal, and Continuous Phenotypes. *G3: Genes, Genomes, Genetics*. 9(5) 1545-1556; https://doi.org/10.1534/g3.119.300585

Crawford, Runcie, et al. (2018). Variable Prioritization in Nonlinear Black Box Methods: A Genetic Association Case Study. *ArXiv, Stat.ME*, 1-28. Retrieved from https://arxiv.org/abs/1801.07318v3.

Crossa, Jose et al. (2010). Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. *Genetics.* October 1, 2010 vol. 186 no. 2 713-724; https://doi.org/10.1534/genetics.110.118521

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4), 193-202. doi:10.1007/bf00344251

Ghosal, S., Blystone, D., Singh, A. K., Ganapathysubramanian, B., Singh, A., & Sarkar, S. (2018). An explainable deep machine vision framework for plant stress phenotyping. *Proceedings of the National Academy of Sciences,* 115(18), 4613-4618. doi:10.1073/pnas.1716999115

Goddard, M. (2008). Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica,* 136(2), 245-257. doi:10.1007/s10709-008-9308-0

Goodman, Erik & Pei, Min & Chia-shun, Lai & Hovland, P. (1994). Further Research on Feature Selection and Classification Using Genetic Algorithms.

Gorroochurn, P. (2016). Classic topics on the history of modern mathematical statistics: From Laplace to more recent times. Hoboken, NJ: John Wiley & Sons.

Grinberg, N. F., Orhobor, O. I., & King, R. D. (2017). An Evaluation of Machine-learning for Predicting Phenotype: Studies in Yeast, Rice and Wheat. doi:10.1101/105528

Henderson CR. (1950). Estimation of genetic parameters. Ann Math Stat. (Abstract) 21: 309-310.

Henderson CR. (1973). Sire evaluation and genetic trends. In Proceedings of the Animal Breeding and Genetics Symposium in Honour of Dr.Jay L. Lush 10-41. ASAS and ADSA, Champaign, Ill.

Hiary, H, S. Bani Ahmad, M. Reyalat, M. Braik, Z. ALRahamneh, Fast and Accurate Detection and Classification of Plant Diseases, *International Journal of Computer Applications* 17 (2011) 31–38.

Ikegawa S. (2012). A short history of the genome-wide association study: where we were and where we are going. Genomics & informatics, 10(4), 220–225. doi:10.5808/GI.2012.10.4.220

Jeffreys, H. (1973). Scientific inference. University Press: Cambridge.

Juliana, P., Singh, R.P., Poland, J., Mondal, S., Crossa, J. et al. 2018. Prospects and challenges of applied genomic selection-A new paradigm in breeding for grain yield in bread wheat. *The Plant Genome* DOI:10.3835/plantgenome218.03.0017.

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees?. Nature biotechnology, 26(9), 1011–1013. doi:10.1038/nbt0908-1011

Laplace, P.-S. (1774). "Memoire sur la Probabilite des Causes par les Evenements." Memoires de Mathematique et de Physique Presentes a l'Academie Royale des Sciences, Par Divers Savans, & Lus dans ses Assembl´ees, 6: 621–656.

Least Squares Fitting. (n.d.). Retrieved from http://mathworld.wolfram.com/LeastSquaresFitting.html

Leonard, T., & Hsu, J. S. (1999). *Bayesian methods: An analysis for statisticians and interdisciplinary researchers.* Cambridge, U.K.: Cambridge University Press.

MacKay, D. J. (2003). *Information theory, inference, and learning algorithms.* Cambridge: Cambridge University Press.

Moose, S. P., & Mumm, R. H. (2008). Molecular plant breeding as the foundation for 21st century crop improvement. *Plant physiology*, 147(3), 969–977. doi:10.1104/pp.108.118232

Osterlind, S. J. (2019). *The error of truth: How history and mathematics came together to form our character and shape our worldview.* Oxford, United Kingdom: Oxford University Press.

Pearson, K. (1901). *On lines and planes of closest fit to systems of points in space*. London: University College.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Y. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sanchez-Villeda, M. Sorrells, and J. L. Jannink (2012). Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. Plant Genome 5(3), 103–113.

Runcie, Daniel E and Sayan Mukherjee. (2013). Dissecting High-Dimensional Phenotypes with Bayesian Sparse Factor Analysis of Genetic Covariance Matrices. *Genetics.* vol. 194 no. 3 753-767; https://doi.org/10.1534/genetics.113.151217

Russel, S. (2013). *Artificial intelligence: A modern approach.* Erscheinungsort nicht ermittelbar: Pearson Education Limited.

Simon, Annina & Singh Deo, Mahima & Selvam, Venkatesan & Babu, Ramesh. (2016). An Overview of Machine Learning and its Applications. *International Journal of Electrical Sciences & Engineering*. Volume 1. 22-24.

Stigler, Stephen M. Gauss and the Invention of Least Squares. Ann. Statist. 9 (1981), no. 3, 465--474. doi:10.1214/aos/1176345451. https://projecteuclid.org/euclid.aos/1176345451

Stigler, Stephen M. Richard Price, the First Bayesian. Statist. Sci. 33 (2018), no. 1, 117--125. doi:10.1214/17-STS635. https://projecteuclid.org/euclid.ss/1517562029.

Stigler, Stephen M. The True Title of Bayes's Essay. Statist. Sci. 28 (2013), no. 3, 283--288. doi:10.1214/13-STS438. https://projecteuclid.org/euclid.ss/1377696937.

Turing, A. M. (1950). Computing machinery and intelligence. Oxford: Blackwell for the Mind Association.

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American journal of human genetics*, 90(1), 7–24. doi:10.1016/j.ajhg.2011.11.029