# Machine learning for Big Data analytics in plants

Chuang Ma[1*], Hao Helen Zhang[2], and Xiangfeng Wang[1,3]

[1] School of Plant Sciences, University of Arizona, 1140 E. South Campus Drive, Tucson, AZ 85721, USA
[2] Department of Mathematics, University of Arizona, 617 North Santa Rita Ave, Tucson, AZ 85721, USA
[3] Department of Plant Genetics and Breeding, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China

**Rapid advances in high-throughput genomic technology have enabled biology to enter the era of 'Big Data' (large datasets). The plant science community not only needs to build its own Big-Data-compatible parallel computing and data management infrastructures, but also to seek novel analytical paradigms to extract information from the overwhelming amounts of data. Machine learning offers promising computational and analytical solutions for the integrative analysis of large, heterogeneous and unstructured datasets on the Big-Data scale, and is gradually gaining popularity in biology. This review introduces the basic concepts and procedures of machine-learning applications and envisages how machine learning could interface with Big Data technology to facilitate basic research and biotechnology in the plant sciences.**

## Big Data technology and machine learning

'Big Data' (large datasets) are frequently encountered in the modern biological sciences (see Glossary). The three 'V' features of Big Data – velocity, volume, and variety – have catalyzed the development of innovative technical and analytical strategies to cope with the data [1]. As a result of rapid advances in high-throughput data generation technologies, biologists have entered the Big Data era [2]. Although the cost of data generation is no longer a major concern for genome-wide research, the computational efficiency of analyzing terabytes or even petabytes of data has become a bottleneck [3,4]. The plant science community is seeking novel solutions to the three grand challenges of Big Data: scalable infrastructure for parallel computation; management schemes for large-scale datasets; and intelligent data-mining analytics, which are similar to the challenges faced by any research field producing Big Data (Box 1). The Apache Hadoop ecosystem, which offers a suite of libraries and tools for data storage, access and automated parallel processing, is considered a promising platform that solves the first two infrastructural problems of Big Data [5–9]. 'Machine learning', an emerging multidisciplinary field of computer science, statistics, artificial intelligence, and information theory, is particularly favored by data scientists for exploiting information hidden in Big Data [10]. The unique features of Big Data – massive, high dimensional, heterogeneous, complex or unstructured, incomplete, noisy, and erroneous – have seriously challenged traditional statistical approaches, which are mainly designed for analyzing relatively smaller samples [1]. Furthermore, large biological systems can be so complex that they cannot be adequately described by traditional statistical methods (e.g., classical linear regression analysis, and correlation-based statistical analysis) developed based on hypothesized or prespecified distribution of data, whereas many modern learning techniques are data-driven and able to provide more feasible solutions. For example, popular machine learning approaches such as support vector machines and classification trees do not presume the distribution for data [11,12].

In biology, most of the methods used in genome-wide research are based on statistical testing and designed for analyzing a single experimental dataset. The data explosion introduced by modern genomics technologies requires biologists to rethink data analysis strategies and to create powerful new tools to analyze the data. In recent decades, machine learning has been envisaged by life scientists as a high-performance scalable learning system for data-driven discovery. The effective performance of machine learning has been demonstrated by the Big-Data-scale exploration of an aggregation of various data sources from the encyclopedia of DNA elements (ENCODE) and model organism encyclopedia of DNA elements (modENCODE) projects in animals [13–15]. However, machine learning has not been widely used for analyzing large datasets in plants [12,16,17]. With the success of the iPlant Collaborative (http://www.iplantcollaborative.org) in building a central supercomputing infrastructure and a data consortium for the plant science community [18], this is now an opportune time for plant scientists to take advantage of Big Data technology to address plant-specific problems in their basic research.

Despite the promising potential of machine learning, it is often misunderstood or misused by biologists, mainly owing to their insufficient knowledge of machine learning and the complexity of the biological systems under study. Therefore, the primary goals of this review are to introduce the basic concepts and procedures of machine learning in biology and to envisage how machine learning could

*Corresponding author:* Wang, X. (xwang1@cals.arizona.edu).

*Keywords:* Big Data; machine learning; large-scale datasets; plants.

[*] *Current address:* College of Life Sciences, Northwest Agriculture and Forest University, Yangling, Shaanxi 712100, China.

1360-1385/

http://dx.doi.org/10.1016/j.tplants.2014.08.004

## Glossary

**Active learning:** a machine-learning approach that iteratively update training dataset by strategically selecting informative data for obtaining a classifier with high prediction performance.
**Adaptive boosting (AdaBoost):** a machine-learning approach that iteratively increases the weight of misclassified samples for boosting weak classifiers to be a stronger classifier.
**Apache Hadoop:** a framework that allows the automated parallel storing and processing of data on a large cluster of computing nodes.
**Apache Mahout:** a project that aims to build a scalable machine-learning library running on Hadoop for Big Data analysis.
**Attributes (or features, inputs, independent variables, predictors):** a set of numerical or categorical quantities used to describe an example.
**Big Data:** a popular term describing large datasets with the features of high velocity, volume, and variety, which are difficult to process using traditional database management and analytical methods.
**Big Data scale:** 1 Exabyte (EB) = 1000 Petabytes (PB) = 1 000 000 Terabytes (TB) = 1 000 000 000 Gigabytes (GB).
**Cloud service:** a new type of use-on-demand data computing and storage paradigm that enables users to build time-consuming applications and manage large datasets on many commodity computing nodes.
**Evaluation metric:** a criterion used to measure the performance of a learned model.
**Examples (or instances, samples):** the objects from which a model is learned or on which a model will be applied for prediction.
**F-score (or F-measure):** a measure that can be used to find the optimized threshold of machine-learning models with both high precision and recall.
**Hadoop ecosystem:** a group of Hadoop-related Big Data storage, access, processing and analysis utilities, including HBase, Spark, Hive, Pig, Sqoop, and Mahout.
**Hadoop Distributed File System (HDFS):** a distributed file system developed for accessing and processing the data stored with Hadoop in a parallel manner.
**Hot deck and cold deck imputation:** two techniques for handling missing data. Hot deck imputation replaces missing data with substituted values randomly selected from similar samples in the same dataset. By contrast, cold deck imputation selects values from other datasets.
**Kernel methods:** machine-learning algorithms that transform the features of samples into a higher-dimensional space using kernel functions, such as polynomial function and radial basis function.
**MapReduce:** a programming model that enables users to easily write programs supporting automated parallel processing distributed on multiple computing nodes.
**Matthews correlation coefficient (MCC):** a measure used in machine learning to evaluate the prediction performance of two-class classifiers by taking into account true positives, true negatives, false positives and false negatives.
**Model (or learner, learning model):** a machine-learning algorithm that assigns an output to an example described with a set of attributes.
**Next-generation sequencing technology:** a technology that produces DNA or RNA fragments with the capacity of high throughput, scalability, speed, and resolution.
**Not Only Structured Query Language (NoSQL):** a new data management system that enables the storage and manipulation of data through the construction of highly reliable, scalable and distributed databases.
**Output (or response, outcome, dependent variable):** the outcome of a learning problem. The output can be a categorical label (qualitative) or a continuous value (quantitative).
**Principle component analysis (PCA):** a statistical technique that eliminates redundancy by converting data into a set of linearly uncorrected variables (i.e., principle components).
**Random forest (RF):** a modern machine learning algorithm that constructs with an ensemble of decision trees for classification and regression problems.
**Rhadoop:** an R package that provides an application programming interface (API) for running R scripts with Hadoop.
**Support vector machine (SVM) classifiers:** models that built with support vector machine algorithm to perform the classification of positive and negative samples in a high-dimensional space.
**Training dataset:** a set of examples used to learn the model.
**Tuning dataset:** a set of examples used to select and validate the model.
**Testing dataset:** a set of examples used to assess the generalization performance of a learned model.

interface with Big Data technology to facilitate basic research and applied biotechnology in plants.

## Basics for building a machine-learning system

Machine learning refers to the process of teaching computers the ability to automatically extract important information from examples to achieve improved prediction or search capabilities for associations and/or patterns in data [19]. Machine learning is a multidisciplinary field incorporating computer science, statistics, artificial intelligence, and information theory. The basic definitions and concepts of machine learning (attributes, evaluation metric, examples, model, output, training dataset, testing dataset, and tuning dataset) are defined in the Glossary.

Based on the goal of learning tasks, machine-learning algorithms are organized taxonomically. Two major algorithms are 'supervised learning' and 'unsupervised learning' [20–22]. Supervised learning takes place when the training examples are labeled with their known outputs. For the example of identifying salt-responsive genes in *Arabidopsis thaliana* shown in Box 2, the label of each example (i.e., gene) in the training dataset is +1 or –1, which is known to the learning model for indicating stress-responsive or non-stress-responsive genes, respectively. Unsupervised learning is used when we observe only attributes for the training examples but not their outputs (i.e., examples are unlabeled, or are labeled but unknown to the learning model). There are other types of machine-learning algorithms that are more complex or a hybrid of different algorithms. For example, semisupervised learning handles both labeled and unlabeled data (i.e., only partial examples in training dataset are labeled) [22], with online learning the model learns sequentially from infinite data streams [23], and active learning is designed to strategically select the most representative examples to be manually labeled [23].

Supervised learning can be further divided into classification and regression based on whether the output is qualitative (categorical) or quantitative (continuous) [21]. The goal of classification is to predict labels of examples, whereas regression involves the estimation of a trend and the prediction of real-valued outputs. In binary classification problems, we typically use the labels +1 or –1 to denote the membership of an example: examples from the two classes are referred to as positive samples or negative samples, respectively.

Commonly encountered machine-learning problems in the real world include:

- Classification (also known as pattern recognition): the problem of learning a classifier that assigns labels (or membership) to new unlabeled examples.
- Regression: the problem of estimating the relation between real-valued outputs and attributes to make predictions.
- Clustering: the task of grouping data such that examples in the same group (called a cluster) are more similar to each other than to those of other groups.
- Recommendation: the task of prioritizing examples based on the attributes of interest.
- Dimensionality reduction: the problem of transforming attributes in a high-dimensional space to a space of fewer dimensions.
- Network analysis: the study of exploring associations between systems components for understanding the biological function of individual components and elucidating the behaviors of biological systems.
- Density estimation: the problem of estimating the probability density function for a population based on the observed data.