Feeding the World with Big Data: Uncovering Spectral Characteristics and Dynamics of Stressed Plants

Kristian Kersting[1], Christian Bauckhage[2,3], Mirwaes Wahabzada[4], Anne-Kathrin Mahlein[4], Ulrike Steiner[4], Erich-Christian Oerke[4], Christoph Römer[5], Lutz Plümer[5]

[1] Computer Science Department, TU Dortmund University, Dortmund, Germany
[2] B-IT, University of Bonn, Bonn, Germany
[3] Fraunhofer IAIS, Sankt Augustin, Germany
[4] INRES-Phytomedicine, University of Bonn, Bonn, Germany
[5] Institute of Geodesy and Geoinformation, University of Bonn, Germany

Abstract. Modern communication, sensing, and actuator technologies as well as methods from signal processing, pattern recognition, and data mining are increasingly applied in agriculture, ultimately helping to meet the challenge of "How to feed a hungry world?". Developments such as increased mobility, wireless networks, new environmental sensors, robots, and the computational cloud put the vision of a sustainable agriculture for anybody, anytime, and anywhere within reach. Unfortunately, data-driven agriculture also presents unique computational problems in scale and interpretability: (1) Data is gathered often at massive scale, and (2) researchers and experts of complementary skills have to cooperate in order to develop models and tools for data intensive discovery that yield easy-to-interpret insights for users that are not necessarily trained computer scientists. On the problem of mining hyperspectral images to uncover spectral characteristic and dynamics of drought stressed plants, we showcase that both challenges can be met and that big data mining can — and should — play a key role for feeding the world, while enriching and transforming data mining.

1 Introduction

Facing a rapidly growing world population, answers to the daunting question of "How to feed a hungry world?" are in dire need. Challenges include climate change, water scarcity, labor shortage due to aging populations, as well as concerns as to animal welfare, food safety, changes in food consumption behavior, and environmental impact. Water scarcity is a principle global problem that causes aridity and serious crop losses in agriculture. It has been estimated that drought can cause a depreciation of crop yield up to 70% in conjunction with other abiotic stresses [11, 49]. Climate changes and a growing human population in parallel thus call for a sincere attention to advance research on understanding of plant adaptation under drought. A deep knowledge of the adaptation process is essential in improving management practices, breeding strategies as well as engineering viable crops for a sustainable agriculture in the coming decades. Accordingly, there is a dire need for crop cultivars with high yield and strong resistance against biotic and abiotic stresses. Addressing this issue and the other ones mentioned above, agriculture –arguably the oldest economic endeavor of humankind– is receiving a technological makeover and information technology makes its appearance in the fields.

Agricultural information is gathered and distributed by means of smartphones, portable computers, GPS devices, RFID tags, and other environmental sensors. Farming companies are working on automation technologies such as GPS steering to operate tractors and other agricultural

machines[5]. Aiming at increased food safety, RFID technologies are used to track animals in livestock; for example, since 2010, European sheep farmers are required to tag their flocks and the European Commission has suggested to extend this to cattle. RFID technologies also provide new possibilities for harvest asset management. For instance, by adding RFID tags, bales can be associated with measured properties such as weight and moisture level [5]. In general, mobile communication networks and technologies which are now commonly deployed in many areas around the world have become a backbone of pervasive computing in agriculture. Researchers and practitioners apply them to gather and disseminate information as well as to market products or to do business [12, 70]. As Farmers need to obtain and process financial, climatic, technical and regulatory information to manage their businesses, public and private institutions cater to their needs and provide corresponding data. For example, the U.S. Department of Agriculture, supplies information as to prices, market conditions, or newest production practices. Internet communities such as e-Agriculture allow users to exchange information, ideas, or procedures related to communication technologies in sustainable agriculture and rural development [5].

Agriculture is thus rapidly becoming a knowledge and data intensive industry. So far, however, much of the research and development in this regard has focused on sensing and networking rather than on computation. In this chapter, we survey our recent efforts on big data mining in agriculture [66, 31, 33, 56]. We point out specific research challenges and opportunities — big data and reification — and hope to increase awareness of this new and exciting application domain.

2 Computational Sustainability in Agriculture

Looking at the scientific literature on precision farming, it appears that, most efforts so far were focused on the development and deployment of sensor technologies rather than on methods for data analysis tailored to agricultural measurements. In other words, up to now, contributions to computational intelligence in agriculture mainly applied off-the-shelf techniques available in software packages or libraries but did not develop specific frameworks or algorithms. Yet, efforts in this direction are noticeably increasing and in this section we survey some recent work on data mining and pattern recognition in agriculture.

Computational sustainability in agriculture involves different areas of computer- and information science. Here, we focus on key areas such as knowledge and information management, geo-information systems, and signal processing. Vernon et al.[68] highlight the importance of information systems for sustainable agriculture. While early work in this direction was focused on the design of (relational) databases, more recent approaches consider semantic web technologies for instance for pest control [43], farm management [62], or the integration of molecular and phenotypic information for breeding [7]. Others consider recommender systems and collaborative filtering to retrieve personalized agricultural information from the web [35] or the use of web mining, for instance, in localized climate prediction [15]. Geo-information processing plays a particular role in computational agriculture and precision farming. Research in this area considers mobile access to geographically aggregated crop information [36], region specific yield prediction [58], or environmental impact analysis [22]. It is clear that, in addition to information infrastructures, applications like these require advanced remote sensing or modern sensor networks. Distributed networks of temperature and moisture sensors are deployed in fields, orchards, and grazing land to monitor growth conditions or the state of pasture [12, 70]. Space- or airborne solutions make use of technologies such as Thermal Emission and Reflection Radiometers or Advanced Synthetic Aperture Radar to track land degradation [8] or to measure and predict levels of soil moisture [40]. Oth er agricultural applications include plant growth monitoring [39] and automated map building [60]. A particularly interesting sensing modality consists

in airborne or tractor-mounted hyper-spectral imaging which records spectra of several hundred wavelengths per pixel. With respect to plant monitoring this allows, for instance, for assessing changes of pigment and chemical composition (water, starch, ligning et ist auch dabei) and information about plant architecture and leaf structure. This in turn allows for remotely measuring phenotypic and physiological reactions of plants due to biotic or abiotic stress [44]. Recently, hyper-spectral imaging is being increasingly used for near range plant monitoring in agricultural research. It enables basic research, for example regarding the molecular mechanisms of photosynthesis [51, 52], but is also used in plant phenotyping, for instance as an approach towards understanding phenotypic expressions of drought stress [4, 34, 56]. Classical image analysis and computer vision techniques are being used in agriculture, too. Examples include automated inspection and sorting in agricultural production facilities [37, 54], the detection of the activity of pests in greenhouses [6], or the recognition of plant diseases [59, 46]. Finally, artificial intelligence techniques are increasingly applied to address questions of *computational sustainability* [24]. Work in this area considers algorithmic approaches towards maximizing the utility of land [23], enabling sustainable water resource management [48], or learning of timber harvesting policies [17].

3 Plant Phenotyping: A Big Data and Reification Challenge

A common theme of the work just reviewed is that they require algorithms and architectures that can cope with massive amounts of data. Owing to the increased use of modern sensors, corresponding solutions have to cope with exploding amounts data recorded in dynamic and uncertain environments where there typically are many interacting components [24]. However, it appears that most work in this area so far did not involve specifically trained data scientists and that, from the point of view of computational intelligence, more efficient and accurate methods seem available. Yet, computer scientists entering the field must be aware that methods they bring have to benefit researchers and practitioners in agriculture. Practitioners "out in the fields" are in need of methods and tools that yield results with concrete connection to their cropping system, ideally runnung on mobile devices, i.e. under constrained computational resources, and in real time in order to help them in their daily work. From the perspective of farming professionals, problems are natural and real phenomena that may be addressed using scientific methods and advanced computing. To them purely theoretic concepts or mathematical abstractions are of little use. The world's food producers are highly technology-oriented people but with a purpose. Even if they may not be adequately trained in information technology, they actually do not need to be. They know their business and if a new technology does not fit into their work flows they will either ignore it or wait until it meets their needs. In the next sections, we present our approaches to meet these big data and reification challenges in plant phenotyping. We shall use these examples to underline the above challenges and to illustrate practical solutions for large-scale data. More precisely, we present results from an ongoing efforts on recognizing and predicting levels of drought stress in plants based on the analysis of hyper-spectral images. There are estimates that drought in conjunction with other abiotic stresses causes a depreciation of crop yields of up to 70% [11, 49]. Because of global warming this trend is expected to increase, so that an improved understanding of how plants adapt to drought is called for to be able to breed more resistant varieties. Yet, mechanisms of stress resistance are characterized by complex interactions between the genotype and the environment which lead to different phenotypic expressions [47]. Progress has been made towards the genetic basis of drought related traits [38, 42] and modern data analysis has lead to molecular insights into drought tolerance [1, 27, 50]. However, as genetic and biochemical research are time consuming and only moderately successful in predicting the performance of new lines in the field, there are increased efforts on phenomic approaches.

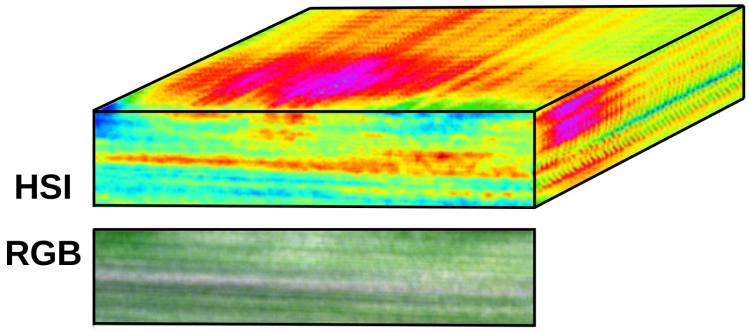


Fig. 1. While conventional RGB images record only three color values (red, green, and blue) per pixel, each pixel of a hyper-spectral image records how a whole spectrum of visible or invisible light waves is reflected from a scene (for technical details, see section A; best viewed in color).

Hyper-spectral imaging provides an auspicious approach to plant phenotyping [51, 52]. In contrast to conventional cameras, which record only 3 wavelengths per pixel, hyper-spectral cameras record a spectrum of several hundred wavelengths ranging from approximately 400nm to 2500nm (see Fig. 1). These spectra contain information as to changes of the pigment composition of leafs which are the result of metabolic processes involved in plant responses to stress. Supervised classification of hyper-spectral signatures can thus be used to predict biotic stress before symptoms become visible to the human eye [55, 57].

However, scale poses a significant challenge in hyper-spectral image analysis, since the amount of phenotyping data easily grows into TeraBytes if several plants are monitored over time. For instance, each individual hypers-spectral recording considered below consists of a total of about 2 (resp. 5.8) Billion matrix entries. Manually labeling such data as well as running established supervised classification algorithms therefore quickly becomes infeasible. Thus, the main goal of plant phenotyping — *the identification of phenotypic features and complex traits which are relevant for stress resistance and to understand the underlying causal networks in the interaction between genotype and environment* — poses important and challenging problems for big data mining: *easy-to-interpret, (un)supervised data mining solutions for massive and high dimensional data over time that scale at most linearly with the amount of data*. This requirement makes it difficult — if not impossible — to use prominent statistical classification or clustering techniques such as SVD, kMeans, (convex) NMF, NMF with volume constraints (see e.g. [45, 61, 3] and references in there), and SVMs, which typically scale at least quadratically with the amount of data if no form of approximation is used that is often accompanied by information loss, may require label information, and/or do not provide easy-to-interpret features/models; they typically produce features that are mathematical abstractions computable for any data matrix. As Mahoney and Drineas argue, "they are not 'things' with a 'physical' reality" [41] and consequently it might be difficult — if not impossible — to provide the plant physiological meaning of, say, an eigenvector.

Addressing these challenges, we have developed novel data mining methods for plant phenotyping that we will review in this chapter. They make only weak assumption on the generating distribution of observed signatures. For instance, one key ingredient is a recent linear time, data-driven matrix factorization approach to represent hyper-spectral signatures by means of convex combinations of only few extreme data samples. Practical results show that the resulting pipeline can predict the level of drought stress of plants well and in turn stress before it even becomes

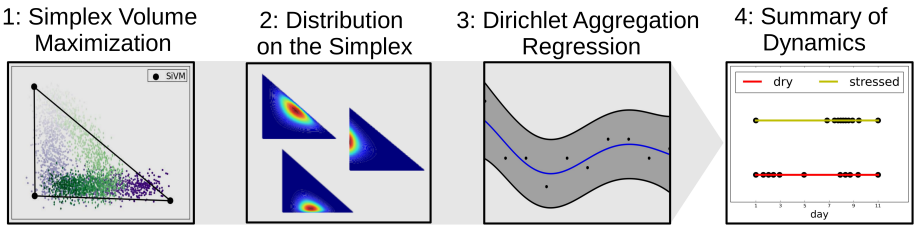


Fig. 2. Pipeline for generating interpretable summaries from hyper-spectral imaging data. (Best viewed in color)

visible to the human eye. Moreover, it can provide an abstract and interpretable view on drought stress progression. In the following we will go through the main ingredients of our pipeline as illustrated in Fig. 2: (1) detecting informative hyper-spectral signatures, (2) modeling distributions over these signatures, (3) smoothing and predicting the evolution of these distributions over time, and (4) summarizing these dynamics using graphical sketches.

4 Interpretable Factorization of Hyper-Spectral Images

Scientists working on plant phenotyping regularly need to find meaningful patterns in massive, high dimensional and temporally diverse observations. For instance, in one of our projects hyper-spectral data of resolution $640 \times 640 \times 69$ were taken of 10 (resp. 12) plants l at 7 (resp. 20) days t. Each record can thus be viewed as a data matrix $\mathbf{X}^{t,l} \in \mathbb{R}^{m \times n}$ with $m = 640 \times 640$ and $n = 69$. Horizontally stacking the data matrices recorded in all experiment then results in a single matrix $\mathbf{X}$ with about 2 (resp. 5.8) Billion entries. Matrix factorization is commonly used to analyze such data. As illustrated in Fig. 3, it factorizes of a matrix $\mathbf{X}$ into a product of (usually) two matrices. That is, $\mathbf{X}$ is approximated as $\mathbf{X} \approx \mathbf{WH}$ where the matrix of basis elements $\mathbf{W} \in \mathbb{R}^{m \times k}$, the coefficient matrix $\mathbf{H} \in \mathbb{R}^{k \times n}$, and $k \ll \min\{m,n\}$. It's useful to think of each column vector in $\mathbf{W}$ as a kind of basis vector or latent factor discoverd in the original data matrix $\mathbf{X}$. A column in $\mathbf{H}$ represents an original data point in terms of the discovered features. This allows for mapping high dimension al data $\mathbf{X}$ to a lower dimensional representation $\mathbf{H}$ and can thus mitigate effects due to noise, uncover latent relations, or facilitate further processing and ultimately help finding patterns in data.

A well known low-rank approximation approach consists in truncating the Singular Value Decomposition (SVD), which expresses the data in terms of linear combinations of the top sin-

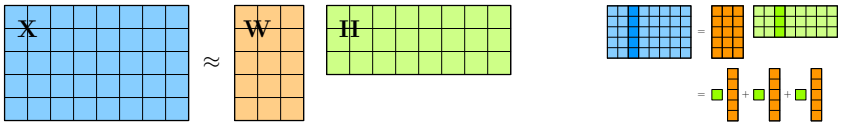


Fig. 3. Data analysis using matrix factorization. (Left) Given an integer $k \leq \min\{m,n\}$, two factor matrices $\mathbf{W} \in \mathbb{R}^{m \times k}$ and $\mathbf{H} \in \mathbb{R}^{k \times n}$ are determined such that $\mathbf{X} \approx \mathbf{WH}$. (Right) Doing so can be viewed as latent factor analysis respectively dimensionality reduction. For each $\mathbf{x}_j \in \mathbb{R}^m$, there is a $\mathbf{h}_j \in \mathbb{R}^k$ expressing $\mathbf{x}_j$ in terms of the found latent factors $\mathbf{W} = [\mathbf{w}_1, \mathbf{x}_2, \ldots, \mathbf{w}_k]$. (Best viewed in color)

Algorithm 1: Interpretable matrix factorization

Input: Matrix $\mathbf{X} \in \mathrm{R}^{m \times n}$, integer c

1 Select c columns from $\mathbf{X}$ and construct $\mathbf{W} \in \mathbb{R}^{m \times c}$;

2 Compute reconstruction matrix $\mathbf{H} \in \mathbb{R}^{c \times n}$ such that the Frobenius norm $||\mathbf{X} - \mathbf{HW}||$ is minimized with respect to $\mathbf{1}^t \mathbf{h}_j = 1$, i.e., all rows of $\mathbf{H}$ sum to one;

3 **Return** $\mathbf{W}$ and $\mathbf{H}$ with $\mathbf{X} \approx \mathbf{WH}$

gular vectors. While these basis vectors are optimal in a statistical sense, the SVD has been criticized for it less faithful to the nature of the data at hand. For instance, if the data are sparse the compressed representations are usually dense which leads to inefficient representations. Or, if the data consists entirely of non-negative vectors, there is no guarantee for an SVD-based low-dimensional embedding to maintain non-negativity. However, the data mining practitioner — as in our application — often tends to assign a "physical" meaning to the resulting singular components.

A way of circumventing these problems, which also hold for other classical techniques such as NMF, kMeans, and sub-space clustering, consists in computing low-rank approximations from selected columns of a data matrix [26] as sketched in the vanilla "interpretable" matrix factorization Alg. 1. Corresponding approaches yield naturally interpretable results, since they embed the data in lower dimensional spaces whose basis vectors correspond to actual data points. They are guaranteed to preserve properties such as sparseness or non-negativity and enjoy increasing popularity in the data mining community [19, 20, 29, 41, 64, 67] with important applications to fraud detection, fMRI segmentation, collaborative filtering, and co-clustering.

But how do we select columns in Line 1? A prominent approach is based on the statistical leverage score [64, 41]. We first compute the top-k right/left singular vectors $\mathbf{V}^{k \times n}$ of $\mathbf{X}$. Then, the statistical leverage score π_i for a particular column i is computed by summing over the rows of the singular vectors, i.e. $\pi_i = \frac{1}{k} \sum_{j=1}^{k} v_{j,i}^2$. The scores π_i form a probability distribution over the columns, and we essentially select columns using that score as an importance sampling probability distribution. Thereby, columns which capture the dominant part of the spectrum of $\mathbf{X}$ are preferred and assigned a higher importance score/probability, cf. [41]. As an alternative, it was shown that a *good* subset of columns maximize their volume [13, 25]. That is we maximize the volume of the parallelepiped (the value of the determinant $\det \mathbf{W}$) spanned by the columns of $\mathbf{W}$. Given a matrix $\mathbf{X}^{m \times n}$, we select c of its columns s.t. the volume $Vol(\mathbf{W}^{m \times c}) = |\det \mathbf{W}|$ is maximized, where $\mathbf{W}^{m \times c}$ contains the selected columns. The criterion, however, is provably NP-hard [13]. Thurau *et al.* [67] introduced recently an approximation, called Simplex Volume Maximization and illustrated in Fig. 4, that was empirically proven to be quite successful. For a subset $\mathbf{W}$ of c columns from $\mathbf{X}$, let $\Delta(\mathbf{W})$ denote the $c-1$-dimensional simplex formed by the columns in $\mathbf{W}$. Now, the volume of the c-simplex $\mathrm{Vol}(\Delta(\mathbf{W}))$ is

$\mathrm{Vol}(\Delta(\mathbf{W}))_c^2 = \theta \det \mathbf{A}$ where $\theta = \frac{-1^{c+1}}{2^c (c!)^2}$ and $\det \mathbf{A}$ is the so-called *Cayley-Menger* determinant [9] that essentially only involves the squared distance $\mathrm{d}_{i,j}^2$ between the vertices i and j (or columns i and j of $\mathbf{W}$), see [9, 67] for details. As Thurau *et al.* have shown, since the distance geometric formulations is entirely based on vector norms and edge lengths, it allows for the development of an efficient greedy algorithm.

Specifically, finding a globally optimal subset $\mathbf{W}$ that maximizes the volume requires the computation of all pairwise distances among the columns in $\mathbf{W}$. For large data sets, this is ill-advised as it scales quadratically with the number n of data points. To arrive at an iterative, approximative $O(cn)$ procedure, Thurau *et al.* proceed greedily: Given a simplex S consisting of $k-1$ vertices, we seek a new vertex $\mathbf{x}_\pi \in \mathbf{X}$ such that $\mathbf{x}_\pi = \arg\max_k \mathrm{Vol}(S \cup \mathbf{x}_k)^2$. Thurau *et al.*

Algorithm 2: Simplex Volume Maximization (SiVM) as introduced in [66].

Input: Matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, integer c, interger l

1 Randomly select r from $1, 2, \ldots, n$;
2 $\mathbf{z} = \arg\max_j d(\mathbf{X}_{*,r}, \mathbf{X}_{*,j})$;
3 **for** $j = 1 \ldots n$ **do**
4 $\quad p_j \leftarrow \log(d(\mathbf{z}, \mathbf{X}_{*,j}))$; $\Phi_{0,j} \leftarrow n_j$; $\Lambda_{0,j} \leftarrow n_j^2$; $\Psi_{0,j} \leftarrow 0$;
5 a $= \max_j(p_j)$;
6 **for** $i = 2 \ldots c$ **do**
7 $\quad$ **for** $j = 1 \ldots n$ **do**
8 $\quad\quad p_j \leftarrow \log(d(\mathbf{w}_{i-1}, \mathbf{X}_{*,j}))$; $\Phi_{i,j} \leftarrow \Phi_{i-1,j} + p_j$; $\Lambda_{i,j} \leftarrow \Lambda_{i-1,j} + p_j^2$;
9 $\quad\quad \Psi_{i,j} \leftarrow \Psi_{i-1,j} + p_j * \Phi_{i-1}$; $p_j \leftarrow$ a $* \Phi_{i,j} + \Psi_{i,j} - \frac{(i-1)}{2}\Lambda_{i,j}$;
10 $\quad$ select $= \arg\max_j\{p_j\}$;
11 $\quad \mathbf{w}_i = \mathbf{X}_{*,\text{select}}$;
12 $\quad \mathbf{W}_{*,i} = \mathbf{X}_{*,\text{select}}$;
13 **Return** $\mathbf{W} \in \mathbb{R}^{m \times k}$

have shown that this leads to the following heuristic

$$\mathbf{v}_\pi = \arg\max_k \left(\log(a) \sum_{i=1}^{n} \log(\mathrm{d}_{i,k}) + \sum_{\substack{i=1 \\ j=i+1}}^{n} \log(\mathrm{d}_{i,k}) \log(\mathrm{d}_{j,k}) - \frac{n-1}{2} \sum_{i=1}^{n} \log^2(\mathrm{d}_{i,k}) \right).$$

that locally increases the volume of the simplex in each iteration. Simplex Volume Maximization (SiVM) is illustrated in Fig. 4 and summarized in Alg. 2. For the first data point to select, we simply take the two points, which are most likely furthest away from each other. In later iterations, we select points in lines 9-18. Pairwise distances computed in one iterations can be reused in later iterations so that, for retrieving c columns, we need to compute distances from the last selected column to all other data points exactly $c+1$ times. As c is constant, we have an overall running time of $\mathcal{O}(n)$. Finally, we note that SiVM is more efficient than other deterministic methods as it supersedes the need for expensive projections of the data. Nevertheless, it aims for solutions that are similar to a greedy algorithm due to Civril and Magdon-Ismail [14] as the projection and orthogonality constraint is implicitly part of the distance geometric objective function.

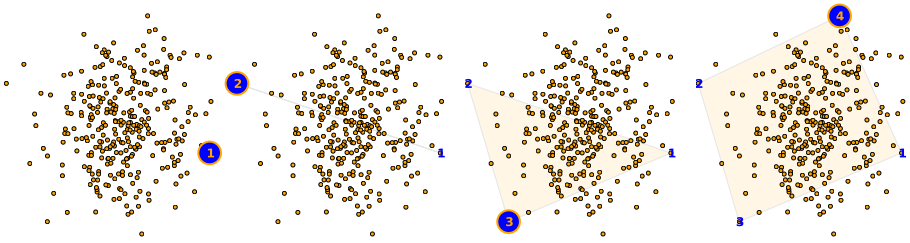
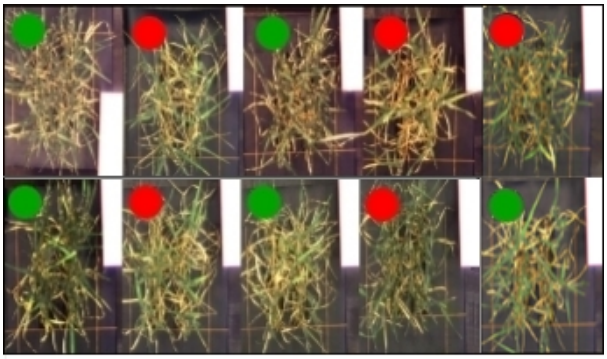
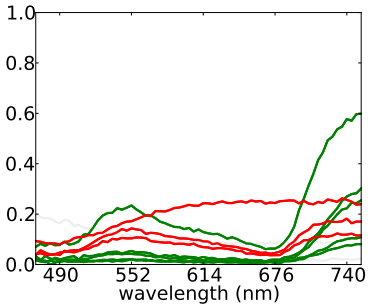


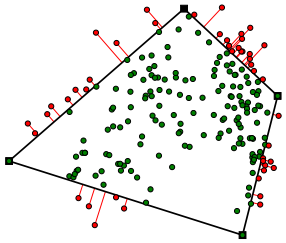
Fig. 4. (From left to right) Didactic example of how Simplex Volume Maximization iteratively determines basis vectors for representation of a data sample by means of convex combinations. (Best viewed in color)



(a) RGB images of plants



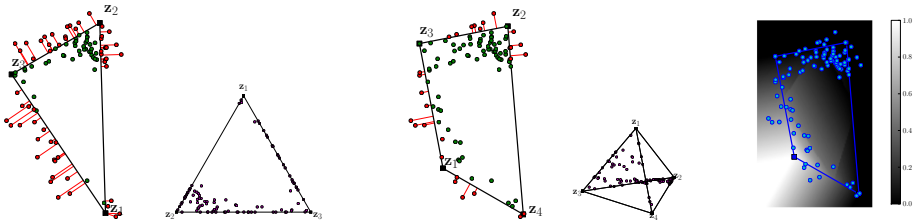
(b) extreme spectra



(c) data reconstruction

Fig. 5. Fast plant phenotyping using SiVM. From left to right: (a) actual examples of RGB images of plants on the fourth measurement day; corresponding hyper-spectral images were recorded from the same point of view and under the same conditions; (b) actual examples of different extreme high-dimensional spectra determined within the hyper-spectral recordings; each of these spectra corresponds to a hyper-spectral pixel and shows the fraction of light reflected at different wavelength; the automatically determined extreme spectra belong to images of “dry" (red) and “healthy” (green) leafs; it is noticeable that dry and healthy plants are not necessarily distinguishable from looking at the RGB images in (a); (c) didactic example of how any sample point can be expressed as a convex combination of selected extremes (see previous figure); while points inside of the convex hull of selected basis elements can be reconstructed exactly, points on the outside are approximated by their projection onto its closest facet; (best viewed in color)

To summarize, SiVM can be used for selecting columns in Line 1 of Alg. 1 in $\mathcal{O}(cn)$ time. Then we compute the coefficients $\mathbf{H}$ in Line 2 by solving constrained quadratic programs [10]. This is $\mathcal{O}(cn) = \mathcal{O}(n)$ since c is a constant. Thus, applied to a data matrix $\mathbf{X}$ resulting from a stacked set of hyper-spectral images, SiVM can be used to detect few representative hyper-spectral signatures $\mathbf{W}$ in time linear of the number of hyper-spectral signatures. This fast plant phenotyping using SiVM is illustrated in Fig. 5. Next to exhaustive lab experiments, field experiments have shown that it can distinguish subtle differences of crop traits in the field [56]. The key to this success is that SiVM paves the way to statistical machine learning.



(a) $k = 3$ (b) $k = 4$

Fig. 6. Bridging geometry and probability. The coefficient vectors $\mathbf{h}_j$ are stochastic, i.e., they sum to one. Hence the coefficient h_{ij} can be thought of as $p(\mathbf{x}_j|\mathbf{w}_i)$ as shown for $k = 4$ (the darker the less probable). (best viewed in color)

5 From Geometry to Probability: Densities over Signatures

From a geometric point of view, as illustrated in Fig. 6, the columns $\mathbf{h}_1, \ldots, \mathbf{h}_n$ of $\mathbf{H}$ are data points (signatures) that reside in a simplex spanned by the extreme elements in $\mathbf{W}$. On this simplex spanned by the extremes, there are natural parametric distributions to characterize the density of the $\mathbf{h}_i$. The best known one is the Dirichlet

$$\mathscr{D}(\mathbf{h}_i|\alpha) = B(\alpha)\prod_{j=1}^{c} h_{ij}^{\alpha_j - 1} \tag{1}$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_c)$. The normalization constant $B(\alpha) = \Gamma(S(\alpha))/\prod_{j=1}^{c}\Gamma(\alpha_j)$ where $\Gamma(\cdot)$ is the gamma function and $S(\alpha) = \sum_{j=1}^{c}\alpha_j$. This distributional view on hyper-spectral data provides an intuitive measure of e.g. the drought stress: the expected probability of observing a healthy spot, which we call the "drought stress level" of a plant. This "drought stress level" is a distribution. To see this, given α, we note that the marginal distribution of the j-th reconstruction dimension follows a Beta distribution $\mathscr{D}(\alpha_j, S(\alpha) - \alpha_j)$ and the expected value of the j-th reconstruction dimension is $\mu_j = \alpha_j/S(\alpha)$. Thus, each α_j controls "aggregation" of mass of reconstructions near the corresponding column c_j which explains the term *Dirichlet aggregation*. Now assume that each dimension was labeled either "background", "healthy", or "dry". Averaging the expected values of "healthy" or "dry" dimensions and treating them as parameters of a Beta distribution yields the drought stress level of a plant. As shown in [56], this can be used to detect drought stress up to 1.5 weeks earlier than by the naked eye.

An alternative to the Dirichlet is the log-normal distribution as e.g. proposed by Aitchison [2] or so-called logratio transformations also due to Aitchison. The latter transform the reconstructions from the simplex sample space to the Euclidean space. In the transformed space, we can then use any standard multivariate method such as estimating multivariate Gaussian distributions Normal(μ, Σ) with mean μ and covariance matrix Σ. Doing so can be a valid altenative to Dirichlets. Under a Dirichlet, the components of the proportions vector are nearly independent. This leads to the strong and unrealistic modeling assumption that the presence of one extreme point is not correlated with the presence of another. A logratio transformation together with Gaussian distributions overcome this problem and may provide a richer view on the interactions of selected columns.

In any case, what do we gain by having distributions on the simplex induced by SiVM? In general, it opens the door to statistical data mining at massive scale. For instance, one could embed the hyper-spectral images into a low-dimensional Euclidean space. This yields easy-to-interpret representations of the relationships among the images and in turn among the plants

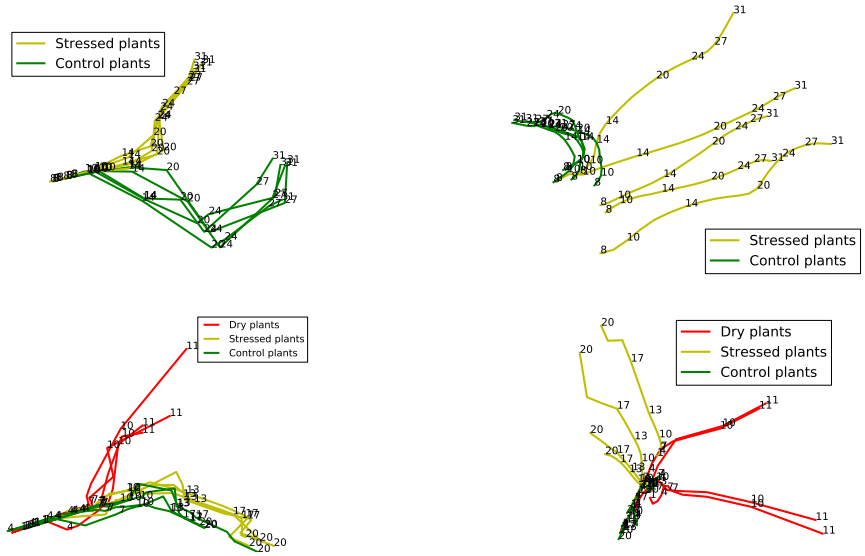


Fig. 7. Improved drought stress detection using DAR. From left to right: (1) Dirichlet traces for 2010 (two groups of measurements) without and (2) with DAR smoothing. (3) Dirichlet traces for 2011 (three groups of measurements) without and (4) with DAR smoothing. Colors indicate controlled/stressed plants; numbers denote the measurement days.

(over time). Probably the most classical examples are multidimensional scaling (MDS) [16] and IsoMap [65]. Whereas MDS uses pairwise distances $D_{i,j}$ only to find an embedding that preserves the interpoint distances, IsoMap first creates a graph G by connecting each object to l of its neighbors, and then uses distances of paths in the graph for embedding using MDS. For plant phenotyping with images over time, one can strike a middle ground taking the temporal relations among plant images into account. The main step, however, is to compute distance metric among the densities Dirichlets. To do so, one could for instance employ the Bhattacharyya distance that is commonly used in data mining to measure the similarity of two probability distributions [30]. It is computed by integrating the square root of the product of two distributions. Fig. 7 shows Euclidean embeddings of hyper-spectral images computed using the Bhattacharyya distance between the induced Dirichlet distributions. Since the images were taken over time, we call this Dirichlet traces. Moreover, as we will show next, we can use more advanced machine learning methods to smooth and even predict the drought levels over time.

6 Pre-symptomatic Prediction of Plant Drought Stress

In order to track drought levels over time, we apply Dirichlet-aggregation regression (DAR) as proposed in [32]. We first select extreme columns from the overall data matrix $\mathbf{X}$. This captures global dependencies as we represent the complete data by means of convex combinations extreme data points selected across all time steps. Then, on the simplex spanned by the extreme points, we estimate Dirichlet distributions specified by $\alpha^{t,l}$ over all reconstructions per day t and plant l. This

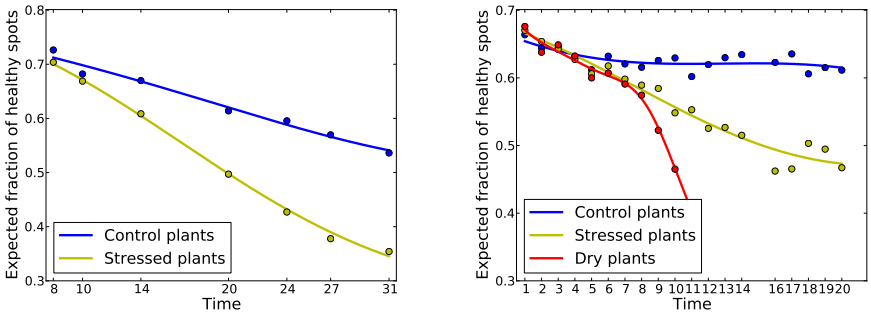


Fig. 8. Dirichlet-aggregation regression (DAR) of drought levels over several days in 2010 (left) and in 2011 (right) using all hyper-spectral images available. Colors indicate controlled/stressed plants. While the x-axis indicates measurement days, the y-axis indicates the fraction of pixels in the analyzed hyper-spectral images that show healthy parts or spots of a plant predicted from our DAR model. Note that experiments in agricultural research cannot seamlessly be repeated at any time but have to adhere to seasonal growth cycles of plants. Accordingly, data not recorded in an experiment may not be available until a year later. In this example, in the experiments in 2010, plants were watered or stressed but not deliberately dried out. In the experiments in 2011, a third set of data was recorded from dry plants (cf. the experimental procedure in section 3.3).

captures local dependencies. Finally, DAR puts a Gaussian process prior on these local Dirichlet distributions. The prior can be a function of any arbitrary types of observed continuous, discrete and categorical features such as time, location, fertilization, and plant species with no additional coding, yet inference remains relatively simple. More precisely, DAR iterates the following steps until convergence:

1. Optimize the logarithm of the complete likelihood w.r.t. the hidden Dirichlet aggregations $\alpha^{t,l}$ for each plant l.
2. Optimize the log-likelihood of all plants w.r.t. the hyper-parameters ϑ of a common Gaussian process prior.

For more technical details, we refer to [33].

This non-parametric Bayesian approach can be used for smoothing the estimated drought level. Fig. 8 shows drought levels estimated by DAR averaged over groups of plants in two data sets considered in our project. As one can see, DAR nicely smoothes SiVM's "hard" drought level (shown as dots). Having a Bayesian regression model at hand, however, we can also move on to make predictions. To do so, we iteratively obtain predictions by making repeated one-step ahead predictions, up to the desired horizon. For the one-step ahead prediction at time t^*, we apply standard Gaussian process regression [53]. For the multiple-step ahead prediction task we follow the method proposed in [21]. That is, we predict the next time step using the estimate of the output of the current prediction as well as previous outputs (up to some lag U) as input, until the prediction k steps ahead is made. Thus, the prediction k steps ahead is a random vector with mean formed by the predicted means of the lagged outputs.

To summarize, based on hyper-spectral images, drought stress levels of plants can be predicted as follows: (1) Using SiVM, we compute few extreme signatures, say 50, and label them

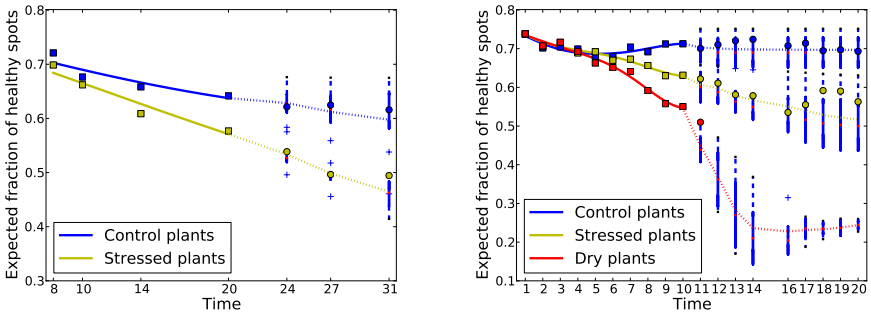


Fig. 9. Bayesian drought level predictions (over time indicated in days) for 2010 (left) and 2011 (right). While the *x*-axis indicates measurement days, the *y*-axis indicates the fraction of pixels in the analyzed hyper-spectral images that show healthy parts or spots of a plant. In both experiments, the drought levels of the second half of measurement days were predicted based on a DAR model (including the extraction of extreme spectra) obtained from the data gathered in the first half of measurement days. Colors indicate controlled/stressed plants. Again, for the measurements carried out in 2010, a control group of dry plants was not available.

accordingly. (2) On the simplex spanned by these extremes, we estimate the latent Dirichlet aggregation values per plant and time step using DAR. (3) Using the Gaussian process over the latent Dirichlet aggregation values, we compute the drought levels of each plant and time step using the labels of extreme spectra, i.e. “background”, “healthy”, and “dry”. (4) Finally, we predict drought levels multiple steps ahead in time using the above Gaussian process approach. Fig. 9 illustrates that this prediction can work well.

7 Sketching Drought Stress Progression

Indeed, one may argue that using non-parametric Bayesian machine learning contradicts the reification challenge. Practitioners are not necessarily trained statisticians or data scientists and hence may not be comfortable dealing them. However, as we will show now, one can compute easy-to-interpret summaries of them.

To create a single sketch describing the hyperspectral dynamics of stressed plants, we advocate the "Equally-Variance Bin Packing" (EBP) decomposition, where we are essentially motivated by the sequential bin-packing problem, a version of the classical bin-packing problem in which the objects are received one by one [28]. However, since we are in a batch[6] setting, we actually face a much simpler instance of the problem, actually with a linear time complexity: *we are looking for a segmentation of ordered objects in B equally weighted bins which preserves the original ordering of the objects.*

Given a matrix $\mathbf{X} \in \mathbb{R}^{K \times N}$ where the columns denote the hyperspectral signatures representing different stages of diseases progression, we can achieve and "Equally Variance Bin Packing" decomposition in B bins as follows: First, we compute the distances of consecutive spectra

[6] In the long run, when plants are monitored over month, the online setting will be relevant.

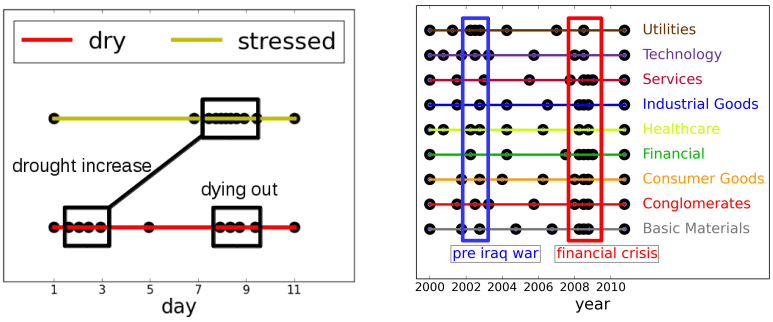


Fig. 10. (Left) Sketching the progression of drought stress over time. Each sketch highlight the interesting periods, where a small edge (between two points) denotes a period of high impact (change in hyperspectral signature). (Right) Financial sketches: during the financial crises in 2008/2009 we have small periods of high impact on many sectors indicating a huge change in stock market prices. (Best viewed in color)

(columns) using Euclidean distance and compute the average bin size as

$$\delta = \frac{1}{B}\sum_{i=1}^{N-1} d_{i(i+1)}(\mathbf{X}) \text{ where } d_{ij}(\mathbf{X}) = \sqrt{\sum_{k=1}^{K}(x_i^k - x_j^k)^2} \tag{2}$$

is the Euclidean distance. Then we fill the $B-1$ bins successively with the objects according to the bin size δ. The last bin is filled with the remaining objects.

This decomposition can be used to draw a single sketch, where the nodes denote the begin (resp. end) of an period and the length of the edge e_b between two consecutive nodes v_b and v_{b+1} is set relatively to the length of the period covered by the objects in bin b. An example of the resulting single sketches are shown in Fig. 10 (left) for drought stressed and actually drying out plants. Each sketch highlight the interesting periods, where a small edge (between two points) denotes a period of high impact (change in hyperspectral signature).

To illustrate the generality of this sketching approach, we additionally computed sketches on a financial dataset. More precisely, inspired by Doyle and Elkan [18], we applied our approach to financial data to obtain a alternative view of economic networks than that supplied by traditional economic statistics. The financial crisis 2008/2009 illustrates a critical need for new and fundamental understandings of the structure and dynamics of economic networks [63]. We computed sketches for the stock price changes of industrial sectors from the S&P 500 as listed on Yahoo! Financial. Specifically, our dataset consists of about 10 years worth of trading data from January 2000 to January 2011. The price of a stock may rise or fall by some percent on each day. We recorded the daily ups and downs for about 3 consecutive months (columns) for all stocks (rows) into a single data matrix per sector. Then we proceed as for the drought stress data. The sketches produced are shown in Fig. 10 (right). As one can see, for the financial crises in 2008/2009 we have small periods of high impact for many of the sectors, indicating a huge change in stock market prices.

Finally, as demonstrated in [69], the sketches can be extended to compute structured summaries of collective phenomena that are inspired by metro maps, i.e. schematic diagrams of public transport networks. Applied on a data set of barley leaves (Hordeum vulgare) diseased with foliar plant pathogens Pyrenophora teres, Puccinia hordei and Blumeria graminis hordei, the resulting metro maps of plant disease dynamics as shown in Fig. 11 conform to plant physiological knowl-

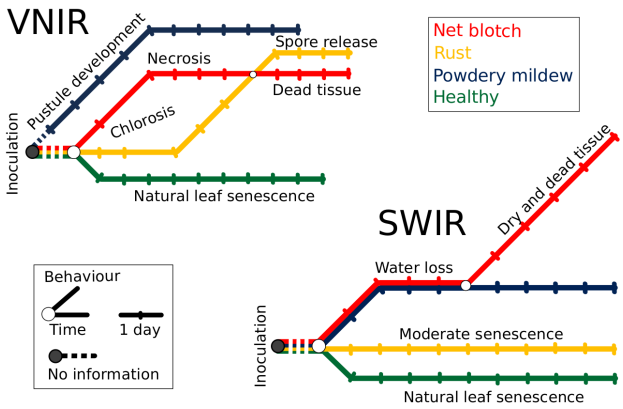


Fig. 11. Collective disease progression via Metro Maps of hyperspectral dynamics of diseased plants for visible-near infrared (VNIR) (top) and short-wave infrared (SWIR) wavelengths (bottom) taken from [69]. Each disease track from hyperspectral images exhibits a specific route in the metro map, the direction and the dynamic steps are in correspondence to biophysical and biochemical processes during disease development. The beginning of all routes is at the same time point/train station (day of inoculation, gray circle). (Best viewed in color)

edge and explicitly illustrate the interaction between diseases and plants. Most importantly, they provide an abstract and interpretable view on plant disease progression.

8 Conclusion

Agriculture, the oldest economic venture in the history of humankind, is currently undergoing yet another technological revolution. Sparked by issues pertaining to sustainability, climate change, and growing populations, solutions for precision farming are increasingly sought for and deployed in agricultural research and practice. From the point of view of pattern recognition and data mining, the major challenges in agricultural applications appear to be the following:

1. The widespread deployment and ease of use of modern, (mobile) sensor technologies leads to exploding amounts of data. This poses problems of big data and high-troughput computation. Algorithms and frameworks for data management and analysis need to be developed that can easily cope with TeraBytes of data.
2. Since agriculture is a truly interdisciplinary venture whose practitioners are not necessarily trained statisticians or data scientists, techniques for data analysis need to deliver interpretable and understandable results.
3. Mobile computing for applications “out in the fields” has to cope with resource constraints such as restricted battery life, low computational power, or limited bandwidths for data transfer. Algorithms intended for mobile signal processing and analysis need to address these constraints.

In this chapter, we illustrated the first two challenges. More precisely, we considered the problems of drought stress recognition, prediction and summarization for plant phenotyping from hyper-spectral imaging. We presented algorithmic solutions that cope with TeraBytes of sensor

recording and deliver useful, i.e. biologically plausible, and interpretable results. In particular, our approach was based on a distributional view of hyper-spectral signatures which we used for Bayesian prediction of the development of drought stress levels. Prediction models of this kind have great potential as they provide better insights into early stress reactions and to identify the most relevant moment when biologists or farmers have to gather samples for invasive, molecular examinations. Moreover, as we have illustrated, even the complex statistical machine learning models can be summarized into easy-to-understand sketches.

In conclusion, the problem of high-throughput phenotyping shows that methods from the broad field of artificial intelligence, in particular from data mining and pattern recognition, can contribute to solving problems due to water shortage or pests. Together with other contributions in the growing field of computational sustainability [24], it thus appears that developments such as mobile, wireless and positioning networks, new environmental sensors, and novel computational intelligence methods do have the potential of contributing to the vision of a sustainable agriculture for the 21st century: the dream of big data feeding a hungry world seems not to be insurmountable.

Acknowledgements:

Parts of this work were supported by the Fraunhofer ATTRACT fellowship "Statistical Relational Activity Mining" and by the German Federal Ministry of Education and Research (BMBF) within the scope of the competitive grants program "Networks of excellence in agricultural and nutrition research - CROP.SENSe.net", funding code: 0315529).

A Experimental Setup

Data sets: In the experimental results reported here (that are actually taken from the corresponding references), we considered two sets of hyper-spectral images. Both data sets were recorded under semi-natural conditions in rain-out shelters at the experimental station of the University of Bonn. For the controlled water stress, three barley summer cultivars Scarlett, Wiebke, and Barke were chosen. The seeds were sown in 11.5 liter pots filled with 17.5 kg of substrate Terrasoil. In 2010 (first data set) the genotype Scarlett was used in *two* treatments (well-watered and with reduced water) with 6 pots per treatment. In 2011 (second data set) the genotypes Wiebke and Barke were used in pot experiments arranged in a randomized complete block design with *three* treatments (well-watered and two drought stressed) with 4 pots per genotype and treatment. The drought stress was induced either by reducing the total amount of water or by completely withholding water. In both cases, the stress was started at developmental stage BBCH31. By reducing the irrigation, the water potential of the substrate remained at the same level as in the well-watered pots for the first seven days but decreased rapidly in the following 10 days down to 40% of the control. For the measurements, the plants were transferred to the laboratory and illumination was provided by 6 halogen lamps fixed at a distance of 1.6 meters from the support where the pots where placed to record hyper-spectral pictures. These were obtained using the Surface Optics Corp. SOC-700 which records images of 640 pixels x 640 pixels with a spectral resolution of approximately 4 nm with up to 120 equally distributed bands in the range between 400 and 900 nm. In 2010, images were taken at 10 time-points, twice per week starting from day four of water-stress. This provided 70 data cubes of resolution $640 \times 640 \times 69$. We transformed each cube into a dense pixel by spectrum matrix. Stacking them horizontally resulted in a dense data matrix with about 2 Billion entries. In 2011 images were taken every consecutive day starting at the second day of watering reduction. Images were taken at 11 time-points for the non-irrigated plants and at 20 time-points for plants with reduced water amount. Applying the same procedure as for the data from 2010 resulted in a matrix of about 5.8 Billion entries.

Analysis Setup: Where required, we split the data from 2011 (resp. 2010) into a first half, denoted 2011.A (resp. 2010.A) and a second half, denoted as 2011.B (resp. 2010.B). Then, we extracted 50 extreme signatures from 2011.A (resp. 2010.A) and determined a DAR regression model on 2011.A (resp. 2010.A). We labeled the extreme signatures as "healthy", "dry", and "background", computed drought levels for 2011.A (resp. 2010.A) based on the DAR model, and used them to predict the drought levels for 2011.B (resp. 2010.B). We also considered the complete 2010 (resp. 2011) data to determine a corresponding DAR model and computed Euclidean embeddings as described in [32] using the smoothed αs. Finally, we note that SiVM can be parallelized so that plant phenotyping from the given data required only about 30 minutes. Estimating DAR models and making predictions happened within minutes.

References

1. A. Abdeen, J. Schnell, and B. Miki. Transcriptome analysis reveals absence of unintended effects in drought-tolerant transgenic plants overexpressing the transcription factor abf3. *BMC Genomics*, 11(69), 2010.
2. J. Aitchison. *The statistical analysis of compositional data*. Chapman and Hall, London, 1986.
3. M. Arngren, M.N. Schmidt, and J. Larsen. Bayesian nonnegative matrix factorization with volume prior for unmixing of hyperspectral images. In *Proc. of the IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2009.
4. A. Ballvora, C. Römer, M. Wahabzada, U. Rascher, C. Thurau, C. Bauckhage, K. Kersting, L. Plümer, and J. Leon. Deep phenotyping of early plant response to abiotic stress using non-invasive approaches in barley. In G. Zhang, C. Li, and X. Liu, editors, *Advance in Barley Sciences*, chapter 26, pages 301–316. Springer, 2013.
5. C. Bauckhage, K. Kersting, and A. Schmidt. Agriculture's technological makeover. *IEEE Pervasive Computing*, 11(2):4–7, 2012.
6. I. Bechar, S. Moisan, M.Thonnat, and F. Bremond. On-line video recognition and counting of harmful insects. In *Proc. ICPR*, 2010.
7. S. Bergamaschi and A. Sala. Creating and querying an integrated ontology for molecular and phenotypic cereals data. In M.A. Sicilia and M.D. Lytras, editors, *Metadata and Semantics*, pages 445–445. Springer, 2009.
8. P.D. Blanco, G.I. Metternicht, and H.F. Del Valle. Improving the discrimination of vegetation and landform patterns in sandy rangelands: a synergistic approach. *Int. J. of Remote Sensing*, 30(10):2579–2605, 2009.
9. L. M. Blumenthal. *Theory and Applications of Distance Geometry*. Oxford University Press, 1953.
10. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
11. J.S. Boyer. Plant productivity and environment. *Science*, 218:443–448, 1982.
12. J. Burrell, T. Brooke, and R. Beckwith. Vineyard computing: Sensor networks in agricultural production. *IEEE Pervasive Computing*, 3(1):38–45, 2004.
13. A. Çivril and M. Magdon-Ismail. On Selecting A Maximum Volume Sub-matrix of a Matrix and Related Problems. *Theoretical Computer Science*, 410(47–49):4801–4811, 2009.
14. A. Çivril and M. Magdon-Ismail. Column subset selection via sparse approximation of svd. *Theoretical Computer Science*, 2011. (In Press) http://dx.doi.org/10.1016/j.tcs.2011.11.019.
15. S. Chakraborty and L. Subramanian. Location specific summarization of climatic and agricultural trends. In *Proc. WWW*, 2011.
16. T. Cox and M. Cox. *Multidimensional Scaling*. Chapman and Hall, London, 1984.
17. M. Crowley and D. Poole. Policy gradient planning for environmental decision making with existing simulators. In *Proc. AAAI*, 2011.

18. G. Doyle and C. Elkan. Financial topic models. In *Working Notes of the NIPS-2009 Workshop on "Applications for Topic Models: Text and Beyond Workshop"*, 2009.
19. P. Feng, Z. Xiang, and W. Wei. CRD: fast co-clustering on large datasets utilizing sampling based matrix decomposition. In *Proc. ACM SIGMOD*, 2008.
20. A. Frieze, R. Kannan, and S. Vempala. Fast monte-carlo algorithms for finding lowrank approximations. *Journal of the ACM*, 51(6):1025–1041, 2004.
21. A. Girard, C.E. Rasmussen, J. Quinonero Candela, and R. Murray-Smith. Gaussian process priors with uncertain inputs – application to multiple-step ahead time series forecasting. In *Proc. NIPS*, 2002.
22. A. Gocht and N. Roder. Salvage the treasure of geographic information in farm census data. In *Proc. Int. Congress European Association of Agricultural Economists*, 2011.
23. D. Golovin, A. Krause, B. Gardner, S.J. Converse, and S. Morey. Dynamic resource allocation in conservation planning. In *Proc. AAAI*, 2011.
24. C.P. Gomes. Computational sustainability: Computational methods for a sustainable environment, economy, and society. *The Bridge*, 39(4):5–13, 2009.
25. S. A. Goreinov and E. E. Tyrtyshnikov. The maximum-volume concept in approximation by low-rank matrices. In D. DeTurck, A. Blass, A.R. Magid, and M. Vogelius, editors, *Contemporary Mathematics*, volume 280, pages 47–51. AMS, 2001.
26. S. A. Goreinov, E. E. Tyrtyshnikov, and N. L. Zamarashkin. A theory of pseudoskeleton approximations. *Linear Algebra and its Applications*, 261(1-3):1 – 21, 1997.
27. P. Guo, M. Baum, S. Grando, S. Ceccarelli, G. Bai, R. Li, M. von Korff, R.K. Varshney, A. Graner, and J. Valkoun. Differentially expressed genes between drought-tolerant and drought-sensitive barley genotypes in response to drought stress during the reproductive stage. *J. Experimental Botany*, 60(12):3531–3544, 2010.
28. A. György, G. Lugosi, and G. Ottucsák. On-line sequential bin packing. *Journal of Machine Learning Research*, 11:89–109, 2010.
29. S. Hyvönen, P. Miettinen, and E. Terzi. Interpretable nonnegative matrix decompositions. In *ACM SIGKDD*, 2008.
30. T. Kailath. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Transactions on Communications*, 15(1):52–60, 1967.
31. K. Kersting, M. Wahabzada, C. Römer, C. Thurau, A. Ballvora, U. Rascher, J. Leon, C. Bauckhage, and L. Plümer. Simplex distributions for embedding data matrices over time. In *SDM*, 2012.
32. K. Kersting, M. Wahabzada, C. Römer, C. Thurau, A. Ballvora, U. Rascher, J. Leon, C. Bauckhage, and L. Plümer. Simplex distributions for embedding data matrices over time. In *Proc. SDM*, 2012.
33. K. Kersting, Z. Xu, M. Wahabzada, C. Bauckhage, C. Thurau, C. Römer, A. Ballvora, U. Rascher, J. Leon, and L. Plümer. Pre–symptomatic prediction of plant drought stress using dirichlet–aggregation regression on hyperspectral images. In *AAAI — Computational Sustainability and AI Track*, 2012.
34. K. Kersting, Z. Xu, M. Wahabzada, C. Bauckhage, C. Thurau, C. Römer, A. Ballvora, U. Rascher, J. Leon, and L. Plümer. Pre-symptomatic prediction of plant drought stress using dirichlet-aggregation regression on hyperspectral images. In *Proc. AAAI*, 2012.
35. F. Kui, W. Juan, and B. Weiqiong. Research of optimized agricultural information collaborative filtering recommendation systems. In *Proc. ICICIS*, 2011.
36. V. Kumar, V. Dave, R. Bhadauriya, and S. Chaudhary. Krishimantra: Agricultural recommendation system. In *Proc. ACM Symp. on Computing for Development*, 2013.
37. S. Laykin, V. Alchanatis, and Y. Edan. On-line multi-sateg sorting algorithm for agriculture products. *Pattern Recognition*, 45(7):2843–2853, 2012.

38. C. Lebreton, V. Lazic-Jancic, A. Steed, S. Pekic, and S.A. Quarrie. Identification of qtl for drought responses in maize and their use in testing causal relationships between traits. *J. Experimental Botanic*, 46(7):853–865, 1995.
39. H. Lin, J. Cheng, Z. Pei, S. Zhang, and Z. Hu. Monitoring sugarcane growth using envisat asar data. *IEEE Trans. Geoscience and Remote Sensing*, 47(8):2572–899, 2009.
40. A. Loew, R. Ludwig, and W. Mauser. Derivation of surface soil moisture from envisat asar wide swath and image mode data in agricultural areas. *IEEE Trans. Geoscience and Remote Sensing*, 44(4):889–899, 2006.
41. M.W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *PNAS*, 106(3):697–702, 2009.
42. J.K. McKay, J.H. Richards, S. Sen, T. Mitchell-Olds, S. Boles, E.A. Stahl, T. Wayne, and T.E. Juenger. Genetics of drought adaptation in arabidopsis thaliana ii. qtl analysis of a new mapping population, kas-1 x tsu-1. *Evolution*, 62(12):3014–3026, 2008.
43. B. Medjahed and W. Gosky. A notification infrastructure for semantic agricultural web services. In M.A. Sicilia and M.D. Lytras, editors, *Metadata and Semantics*, pages 455–462. Springer, 2009.
44. T. Mewes, J. Franke, and G. Menz. Data reduction of hyperspectral remote sensing data for crop stress detection using different band selection methods. In *Proc. IEEE Int. Geoscience and Remote Sensing Symp.*, 2009.
45. L. Miao and H. Qi. Endmember Extraction From Highly Mixed Data Using Minimum Volume Constrained Nonnegative Matrix Factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 45(3):765–777, 2007.
46. M. Neumann, L. Hallau, B. Klatt, K. Kersting, and C. Bauckhage. Erosion band features for cell phone image based plant disease classification. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR–2014)*, pages 3315–3320, 2014.
47. J.B. Passioura. Environmental biology and crop improvement. *Functional Plant Biology*, 29:537–554, 2002.
48. M. Petrik and S. Zilberstein. Linear dynamic programs for resource management. In *Proc. AAAI*, 2011.
49. E. Pinnisi. The blue revolution, drop by drop, gene by gene. *Science*, 320(5873):171–173, 2008.
50. M.A. Rabbani, K. Maruyama, H. Abe, M.A. Khan, K. Katsura, Y. Ito, K. Yoshiwara, M. Seki, K. Shinozaki, and K. Yamaguchi-Shinozaki. Monitoring expression profiles of rice genes under cold, drought, and high-salinity stresses and abscisic acid application using cdna microarray and rna gel-blot analyses. *Plant Physiology*, 133(4):1755–1767, 2010.
51. U. Rascher, C. Nichol, C. Small, and L. Hendricks. Monitoring spatio-temporal dynamics of photosynthesis with a portable hyperspectral imaging system. *Photogrammetric Engineering and Remote Sensing*, 73(1):45–56, 2007.
52. U. Rascher and R. Pieruschka. Spatio-temporal variations of photosynthesis: The potential of optical remote sensing to better understand and scale light use efficiency and stresses of plant ecosystems. *Precision Agriculture*, 9(6):355–366, 2008.
53. C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.
54. A. Rocha, D.C. Hauagge, J. Wainer, and S. Goldenstein. Automatic fruit and vegetable classification from images. *Computers and Electronics in Agriculture*, 70(1):96–104, 2010.
55. C. Römer, K. Bürling, T. Rumpf, M. Hunsche, G. Noga, and L . Plümer. Robust fitting of fluorescence sprectra for presymptomatic wheat leaf rust detection with support vector machines. *Computers and Electronics in Agriculture*, 74(1):180–188, 2010.
56. C. Römer, M. Wahabzada, A. Ballvora, F. Pinto, M. Rossini, C. Panigada, J. Behmann, J. Leon, C. Thurau, C. Bauckhage, K. Kersting, U. Rascher, and L. Plümer. Early drought

stress detection in cereals: Simplex volume maximization for hyperspectral image analysis. *Functional Plant Biology*, 39:878—890, 2012.
57. T. Rumpf, A.-K. Mahlein, U. Steiner, E.-C. Oerke, and L. Plümer. Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture*, 74(1):91–99, 2010.
58. G. Ruß and A. Brenning. Data mining in precision agriculture: Management of spatial information. In *Proc. IPMU*, 2010.
59. S. Sankaran, A. Mishra, R. Ehsani, and C. Davis. A review of advanced techniques for detecting plant diseases. *Computers and Electronics in Agriculture*, 72(1):1–13, 2010.
60. G. Satalino, F. Mattia, T. Le Toan, and M. Rinaldi. Wheat crop mapping by using asar ap data. *IEEE Trans. Geoscience and Remote Sensing*, 47(2):527–530, 2009.
61. R. Schachtner, G. Pöppel, A.M. Tome, and E.W. Lang. Minimum Determinant Constraint for Non-negative Matrix Factorization. In *ICA*, pages 106–113, 2009.
62. M. Schmitz, D. Martini, M. Kunisch, and H.-J. Mosinger. agroxml: Enabling standardized, platform-independent internet data exchange in farm management information systems. In M.A. Sicilia and M.D. Lytras, editors, *Metadata and Semantics*, pages 463–467. Springer, 2009.
63. F. Schweitzer, G. Fagiolo, D. Sornette, F. Vega-Redondo, A. Vespignani, and D.R. White. Economic networks: The new challenges. *Science*, 5939(325):422–425, 2009.
64. J. Sun, Y. Xie, H. Zhang, and C. Faloutsos. Less is More: Compact Matrix Decomposition for Large Sparse Graphs. In *SDM*, 2007.
65. J.B. Tenenbaum, V. De Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 5500(390):2319–2323, 2000.
66. C. Thurau, K. Kersting, M. Wahabzada, and C. Bauckhage. Descriptive matrix factorization for sustainability: Adopting the principle of opposites. *DAMI*, 24(2):325—354, 2012.
67. C. Thurau, K. Kersting, M. Wahabzada, and C. Bauckhage. Descriptive matrix factorization for sustainability: Adopting the principle of opposites. *Journal of Data Mining and Knowledge Discovery*, 24(2):325—354, 2012.
68. R. Vernon, editor. *Knowing Where You're Going: Information Systems for Agricultural Research Management*. International Service for Agricultural Research (ISNAR), 2001.
69. M. Wahabzada, A.-K. Mahlein, C. Bauckhage, U. Steiner, E.-C. Oerke, and K. Kersting. Metro maps of plant disease dynamics—Automated mining of differences using hyperspectral images. *PLoS ONE*, 10(1), 2015.
70. T. Wark, P. Corke, L. Klingbeil, Y. Guo, C. Crossman, P. Valencia, D. Swain, and G. Bishop-Hurley. Transforming agriculture through pervasive wireless sensor networks. *IEEE Pervasive Computing*, 6(2):50–57, 2007.