SmartFarm: Improving Agriculture Sustainability Using Modern Information Technology

Chandra Krintz, Rich
Wolski, Nevena
Golubovic, Benji Lampel,
Varun Kulkarni
Dept. of Computer Science
UC Santa Barbara

Balaji
Sethuramasamyraja*,
Bruce Roberts**
*Dept. of Industrial Technology
**Dept. of Plant Sciences
Fresno State University

Bo Liu
Dept. of BioResource and
Agricultural Engineering
Cal Poly University, San Luis
Obispo

ABSTRACT

In this paper, we overview our work on an open source, hybrid cloud approach to agriculture analytics for enabling sustainable farming practices. SMARTFARM integrates disparate environmental sensor technologies into an on-farm, private cloud software infrastructure that provides farmers with a secure, easy to use, low-cost data analysis system. SMARTFARM couples data from external cloud sources (weather predictions, satellite imagery, state and national datasets, etc) with farm-local statistics, provides an interface into which custom analytics apps can be plugged, and ensures that all private data remain under grower control.

1. INTRODUCTION

Ecological sustainability depends critically on the ability of world food production to manage increasingly limited natural resources (e.g. arable land and water) with new techniques that both enhance environmental stewardship and increase farm productivity. We believe that the key to future food security, food safety, and ecological sustainability lies in the use of customized data analytics by individual growers, agricultural workers, and food producers. For analytics to become *the farm implements of the twenty-first century*, new Information Technology (IT) developments that are accelerating e-commerce (e.g. cloud computing, machine learning, mobile client-server systems) must (i) be made inexpensive, accessible to, and beneficial for, a vast diversity of the population, (ii) address a number of disparate problems including increasing yields, conserving water, and ensuring soil and plant health, and (iii) facilitate new sustainable agriculture science. Today smallholder agriculture and their communities (as opposed to the industrial-scale farming concerns) are strikingly underserved by modern IT.

The goal of our research is to achieve these goals via a new open source, IT system called UCSB SMARTFARM. SMARTFARM is hybrid cloud technology designed to enable small-holder growers and other agricultural (ag) professionals, researchers, and students, to use analytics to improve environmental sustainability and efficiencies in food production. SMARTFARM combines data aggregation for disparate ag-related data sensors and sources, integrates multiple analytics technologies to facilitate a wide range of data inference, prediction, and visualization, and gives growers full control over the privacy, security, and sharing of their most valuable farm asset: their data.

Although many precision farming and ag decision support systems have emerged recently e.g. [8, 12, 18, 19, 21, 29], extant solutions have not received wide spread use to date. The reasons for this, which we have discovered by working with multiple growers and ag technology vendors over the past year, center around concerns related to data privacy, security, and control. First, extant solutions employ a cloud model to facilitate scalable customer acquisition and software maintenance. To use these solutions, growers must transmit potentially vast amounts of data over expensive and low bandwidth network links (for which the growers pay) to a remote cloud application owned and controlled by the vendor. Such connectivity can be very costly and is infeasible in many rural communities.

Second, this model requires that growers reveal private information and relinquish ownership and control of their data to vendors that supply their sensors, equipment, or other ag products. Data from these technologies reveals private and personal information about grower practices, crop input (chemicals, fertilizers, pesticides), and farm implement use, purchasing and sales details, water use, disease problems, etc., that define a grower's business and competitiveness.

The next issue concerns lock-in – vendors make it very difficult and in some cases impossible for growers to get data *out* of these systems, e.g. to stop using a vendor's service or to move data to another vendor. In addition, vendors increasingly charge a recurring fee to host and export access to the data to growers (via their web browsers), and in many cases, sell the data or advertising opportunities to other companies and/or use the data to target marketing and sales campaigns. Finally, the use of centralized, networked systems for deployment of technologies makes growers vulnerable to security breaches, data loss, and interruption since they expose a large attack surface, introduce single points of failure, and require continuous, high quality Internet connectivity.
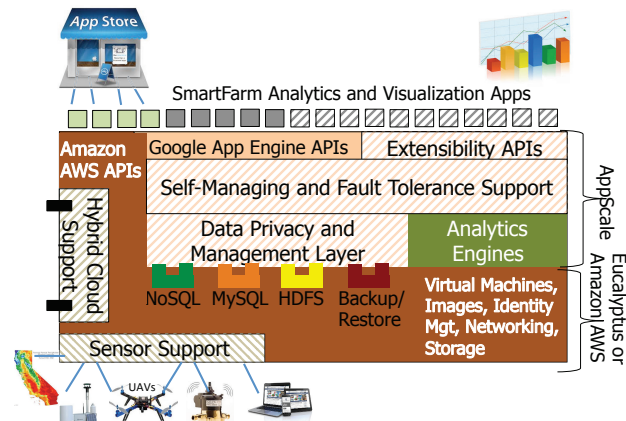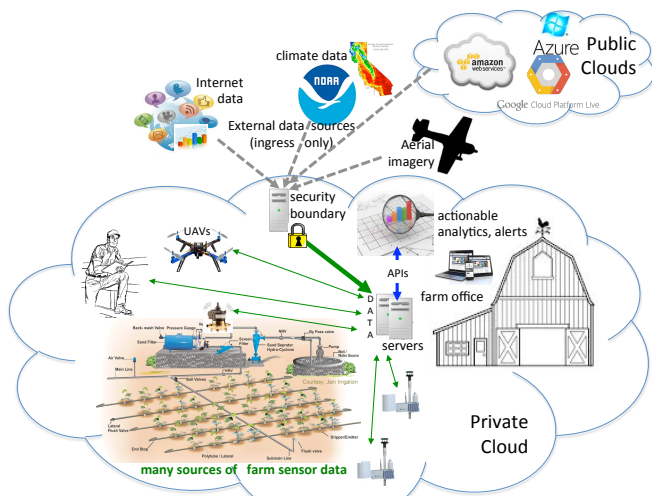
SMARTFARM presents an alternative approach to ag an-

Figure 1: SMARTFARM Vision (left) and the SMARTFARM Software Architecture (right)

alytics and decision support. It provides growers with a hybrid, distributed architecture (hardware and software system) that provides growers with an *on-farm* "analytics appliance". SMARTFARM aggregates data for growers and automatically applies analysis algorithms to implement decision support. If Internet connectivity is available, SMARTFARM can download data analytics and visualization applications (apps) from an *Ag App Store* (otherwise apps can be integrated via DVD mailers). Thus, SMARTFARM "moves the code to the data" rather than "moving the data to the code", thereby significantly reducing network/mailer use.

Because SMARTFARM runs on-farm (but can interoperate with remote data sources if necessary), we have designed it to be very easy to use and to operate autonomously with or without Internet connectivity. Moreover, the system ensures that no private data is leaked and provides robustness and security via its distributed black-box design. In the sections that follow, we overview the design and implementation of SMARTFARM. We then overview our multiple design choices and their implications. Finally, we show how SMARTFARM can be used to extract actionable insights for precision harvesting and root cause analysis.

2. UCSB SMARTFARM

A depiction of our overarching vision for SMARTFARM is shown in Figure 1 (left). SMARTFARM integrates disparate environmental sensor technologies into an on-premise, private cloud software infrastructure that provides farmers with a secure, easy to use, low-cost data analysis system. SMARTFARM couples data from external sources (if/when available) with farm-local statistics, provides an interface into which custom analytics tools can be plugged and automatically deployed, and ensures that all data and analyses remain under the control of farmers. SMARTFARM enables farmers to extract actionable insights from their data, to quantify the impact of their decisions and environmental changes, and to identify opportunities for increasing farm productivity.

Figure 1 (right) presents the SMARTFARM software architecture. SMARTFARM extends open source technologies from the domains of cloud computing, big data analytics, and IoT. While these freely available software technologies are

powerful, they are designed for use primarily in commercial e-commerce and Information Technology (IT) settings. As a result, they are robust but cannot be easily used off-the-shelf because they require significant expertise, IT support, and expensive staffing to setup, manage, and maintain – which are not available to smallholder farming concerns. SMARTFARM integrates and leverages these technologies to construct an open source *data appliance* specifically designed to enable individual growers to practice precision agriculture.

To enable this, SMARTFARM leverages the AppScale cloud platform (PaaS) and the Eucalyptus cloud infrastructure (IaaS) to provide portability across clouds and autonomous IT management. Both of these systems are commercial-grade platforms in wide-spread enterprise use today. AppScale is an open source distributed runtime system that makes it very easy to write and deploy network-accessible applications (including mobile device back-ends) in high-level languages. AppScale is commercially supported, freely available, and is API-compatible with Google's cloud platform App Engine. This means that applications that run on App Engine, also run on AppScale and thus SMARTFARM, without modification. AppScale runs on public cloud infrastructures (Amazon Elastic Compute Cloud (EC2), Google Compute Engine, Microsoft Azure) or on local cluster resources within virtual machine instances over Eucalyptus. Eucalyptus is an open source cloud infrastructure that runs on local cluster resources and provides highly available, fault tolerant virtual machine management. Eucalyptus is API-compatible with Amazon EC2 and Simple Storage Service (S3) so that any EC2 instance/service can run on Eucalyptus without modification. AppScale and Eucalyptus together in SMARTFARM provide us with a production-ready, highly available private cloud system and our users with a vast ecosystem of freely available cloud software from Google and Amazon.

Our software architecture for SMARTFARM extends this private cloud system in multiple ways. First it integrates popular, open source analytics engines from the "Big Data" community. In particular, the SMARTFARM platform supports Spark [5], Hadoop [13], Cassandra [6], postgreSQL [20], MatLab [17], and R [9]. We integrate these disparate technologies into a single distributed system in order to permit

Figure 2: Prototype SMARTFARM appliance for on-farm deployment. Top half contains compute elements running SMARTFARM software stack. Bottom half contains network switch and uninterruptable power supply (UPS).

a wide variety of applications to be executed over SMART-FARM. That is, we envision a model in which developers, researchers, and industry vendors develop analytics and data visualization "apps" that rely on these technologies for their implementation. Growers then download the apps from an "App Store" (for free or for a fee) and execute them on the SMARTFARM appliance using data owned and controlled by the grower. There are numerous such apps available today for general purpose (non-ag) data analysis, visualization, and machine learning [4, 14–16, 23, 26].

Figure 2 shows a picture of an on-farm prototype appliance. The SMARTFARM software implementation (the software "stack") runs on a set of small, durable computing "bricks." Each brick, in this example, contains an 3.1 GHz Intel i7 dual core CPU, 16GB of memory, and 250GB of SSD storage, an 802.11 wireless interface, and a 1 Gb CAT-V ethernet interface. The bricks are "headless" meaning that they can operate without an attached console or keyboard. They communicate with each other using the wired network via a switch (located in the bottom half of the appliance) and all components are plugged into an uninterruptable power supply (UPS) also located in the bottow half.

Because these components can operate at relatively high temperatures (120 degree F or more) and they contain no moving parts, they are well suited to deployment on-farm in an out building or office where there is 120V AC and wireless connectivity. Its firewall enables growers to control fully what data, if any, they share with remote entities. Moreover, SMARTFARM apps can be used to anonymize shared data according to the growers' needs and interests. Finally, SMARTFARM employs a private wired or wireless network for connectivity to on-farm sensors, for full isolation from web accessibility if required.

Our experience with this early SMARTFARM prototype shows that it is feasible to integrate these technologies efficiently and effectively and that it is possible to operate this combination effectively both at the large scale for which it was defined *and, more interestingly, at the small scale which will be typical of on-farm deployments.* Thus SMARTFARM essentially provides a universal, user-controlled cloud platform capable of leveraging the predominant public cloud and popular analytics technologies and applications.

3. DATA INGRESS

We use the term *sensor* herein to describe any device or system that serves as a data source for the analytics or visualization apps executing in a SMARTFARM deployment. Sensors under this definition include physical devices that measure environmental phenomena or record events and activities of farm agents (human, animal, computer, or mechanical) that are germane to precision farming. Additionally, "files" that have been gathered by the grower, extant databases, and Internet accessible web services (e.g. weather data sites) that can provide measurements are modeled and treated as "sensors" by the SMARTFARM system.

Our goal with SMARTFARM is to support as "farm sensors" a variety of data sources including historical records (hand written and digital), sensing devices, the Internet (if/when available), and imagery data. Historical records include those for input (pesticides, water, fertilizers) applications, disease, crop variability and yields, animal health and welfare, and manual information extraction (BRIX tests – sugar content in solution), Anthocyanin (pigment), pressure chambers (plant water status), among others. The sensing devices include those that measure water pump performance/pressure, irrigation processes, and soil moisture/quality as well as weather stations, energy meters (*e.g.* GreenButton [10]), and devices that collect farm implement (planters, harvesters, etc.) activity and animal behavior. Internet data is data that is made available via web APIs and include those from social networks, area weather stations (e.g. CIMIS [7], WeatherUnderground), and web services such as those for geo-location, news, and stock reports. Finally, imagery data include multi-spectral (that reveal plant health, disease, animal activity, water use, etc.) video and still images collected on-farm manually or via UAVs, e.g. [3, 22, 28], and off-farm via fixed-wing aircraft [27] and satellites [11, 24].

No extant system today provides such unified data aggregation. The challenges in doing so include configuring and deploying (i.e. interoperating with) such a diversity of options. Moreover, many vendors hold their sensor device API as proprietary for commercial purposes. In addition, individual vendors are not commercially incentivized or interested in providing a unifying data integration service that works with the devices of their competitors. Thus, such an infrastructure and feature set must come from the research and open source communities.

The most mature step in this direction is with the an open source middleware for the Internet-of-Things (IoT) called Global Sensor Network [1, 2] (GSN). As such, we have integrated GSN into SMARTFARM and have extended it with customized support for agriculture settings. GSN is a robust, widely used technology for deploying sensor networks and processing data produced by them in a distributed setting. Because GSN is a general purpose system, it provides mechanisms for integrating arbitrary sensors. It does so by defining a "virtual sensor" in software that represents and interoprates with one or more physical sensing (data producing) devices. The GSN community has contributed a number of different virtual sensor specifications and implementations for popular IoT devices (home monitoring devices, photographs, weather stations, and others).

Each GSN virtual sensor consists of a component that interfaces to the physical sensor or group of sensors and a component that collects and persists the data from the sensor in the data management layer of SMARTFARM . These components are implemented as part of the GSN virtual sensor software stack consists of a sensor descriptor, data processing logic, and a software wrapper that communicates with one or more instances of a physical sensor. This interface is sufficiently abstract to support a wide range of different

sensor types (including GSN itself to link multiple GSN deployments if necessary).

To date we have developed GSN virtual sensors for local multiple soil moisture sensors, local temperature sensors, and local records including multi-spectral imagery and comma separated value (CSV) files. We store images in a bucket store and their metadata in GSN. In addition we have developed virtual sensors for remote (web/cloud application programming interfaces (APIs)) data sources (when available). The technologies that we have integrated include those from GreenButton (electricity use), CIMIS (weather and irrigation-related data), WeatherUnderground (community weather stations), PowWow Energy (imagery), Tule (evapotranspiration), Irrometer (soil moisture and temperature), as well as DropBox and Box (generic image/CSV file storage). We have also extended GSN with support for direct JSON queries (only vanilla SQL was supported previously). In addition, we have developed multiple tools that automate virtual sensor specification and GSN deployment. Our extensions simplify sensor development and GSN use to make GSN more accessible to non-experts. In addition, they enable GSN users to ingress a wide variety of disparate datasets from a vast diversity of physical sensors and data sources. All of our extensions and tools are available via GitHub at https://github.com/UCSB-CS-RACELab/gsn.

In all settings, deployed sensor data is automatically ingressed by GSN and stored in SQL database tables. GSN provides visualization tools to visualize the data over time and SMARTFARM apps access the data via SQL queries or by transforming the data to HDFS for access via Hadoop and Spark analytics and NoSQL for large scale distributed range queries.

4. DATA EGRESS AND ANALYSIS

Our goal with SMARTFARM is to help farmers and ranchers improve environmental stewardship, while increasing yields, profitability, and animal welfare. To enable this, we make SMARTFARM available as open source to researchers and software developers to investigate new sustainability science and engineering in the areas of precision agriculture, agronomics, and bio-resource and agriculture engineering in the form of SMARTFARM applications (apps). The SMARTFARM platform exports a set of APIs and tools that simplify app deployment and once running, automatically manages their execution. The platform is also able to communicate and control farm-local sensors and equipment based on the outcome of app analytics.

As an example, we have developed a SMARTFARM app for wine grape growers called RootRApp. Our foundation RootRApp employs anthocyanin sampling via Near Infrared (NIR) sensing in combination with Ordinary Kriging to implement automated differential harvesting [25]. This use of sampling and Kriging constitutes an example of statistical analytics used as a *control mechanism* since the harvester uses the Kriging interpolations to separate grapes as it picks them into two different quality batches automatically.

SMARTFARM makes it possible to use the *same data set with different analytics also to implement decision support* for the grower. RootRApp uses samples of anthocyanin (collected for the original study) and attempts to find a "point of origin" for a potential root cause associated with the grapes that had fruit readings lower than $0.87mg/g$ (low quality).

Figure 3 (a) reproduces the map in our original work with respect to anthocyanin readings. In the original study, the grower selected a threshold value of 0.87 to differentiate between grape qualities. The figure shows regions where the harvester used a value interpolated from the sample to decide whether the grapes were Quality A (dark red regions) or Quality B (yellow regions).

Using the same dataset, RootRApp also computes the $\alpha = 0.01$ confidence region for the Quality B grapes based on an assumption of bivariate normality. If the *reason* for quality difference has a point of origin then a bivariate Normal distribution (as a function of distance) is a reasonable first-order model to investigate as a dispersion model. Figure 3 (b) shows the results of this preliminary investigation. Observing the data, it is possible to discern (based on the proximity to the tree and water derrick marked in each) where the likely point-of-origin is located. From these two figures it appears that such a point would lie in the interior of the yellow region shown in the center of Figure 3 (a).

SMARTFARM makes it easy to extract or egress data from the system so that it can be used by online tools and hybrid cloud services. In this example, we use Google Earth to enable the growers to zoom in and out of SMARTFARM data (i.e. "fly" through the annotated image) in real time. Multiple levels of detail combined with analytics calculation allow the grower to pinpoint the exact region in which a point-of-origin is likely to occur. By storing the data and performing the analysis locally, SMARTFARM provides growers with the same outcome that public cloud solutions provide while conserving both bandwidth and Internet data plan charges and preserving grower privacy.

In particular, no information about the farm (other than the GPS locations of the samples) is shared with Google, i.e. no information about anthocyanin or BRIX (which indicate potential productivity for the plot) is shared. RootRApp precludes the need to upload these latter measurements to Google when producing the map for the grower. In this way, an adversary gaining access to Google's infrastructure might learn that a grower is interested in a particular plot of land (based on latitude and longitude) but could not discern what information beyond that were part of the analysis. Because the app is customized to a specific growing region, it is also possible to determine whether a datum falls within the latitude and longitude under study and if it does not, block it from being egressed.

Notice also that RootRApp is purely an analytics calculation. The root cause could be lack of plant stress, disease, insect infestation, soil contamination, etc. That is, we can apply it to a wide range of problem sets and sensor data types. The analytics combined with high-quality and interactive imaging allow the grower to focus his or her efforts to investigate specific plants and areas in the vineyard. For example, if the grower wishes to investigate whether the reason for the lower levels of anthocyanin were due to a water leak that reduced plant stress, the level of detail provided by the interactive capability provided by Google Earth *combined* with the analytics calculation showing the confidence region provides a useful guide on where to look for the leak.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we overview the vision, software architecture, and our on-going work surrounding UCSB SMARTFARM. In addition, we describe a new SMARTFARM application (app), called RootRApp, that makes use of farm sensor
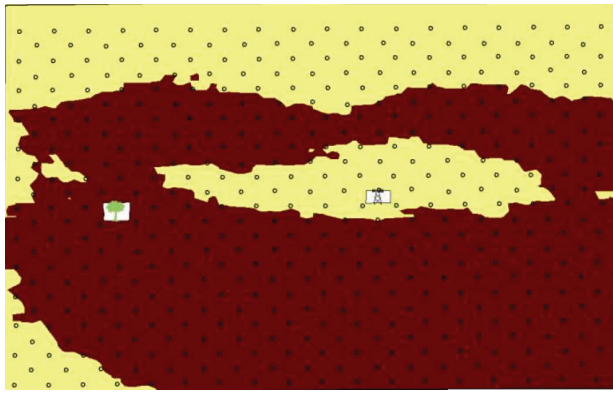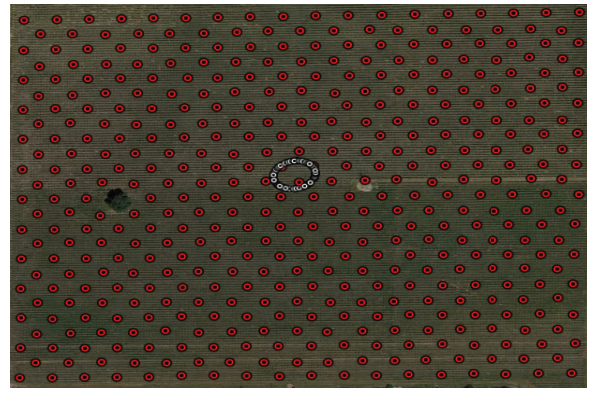
Figure 3: Example SMARTFARM app: Harvester control and origin analysis using the same data and hybrid cloud (local analysis and Google Earth). Figure (a) is an interpolated anthocyanin map from [25]. Points indicate vines that were sampled. Dark red regions indicate anthocyanin readings > 0.87 (Quality A grapes) and yellow regions corresponds to readings <= 0.87 (Quality B grapes). Figure (b) shows the anthocyanin samples (red points) from [25] and point-of-origin using the bivariate Normal for probability 0.99. The white ellipse delineates the $\alpha = 0.01$ confidence region.

data to perform root cause analysis for wine grape quality differences. RootRApp, summarizes and anonymizes data so that it can be shared and visualized with a cloud geospatial mapping service (Google Earth) running locally or at Google.

As part of future work, we are developing techniques that will enable autonomous management of the SMARTFARM private cloud (greater than 1 year) by leveraging the production-quality mechanisms for high availability that Eucalyptus provides. We are also working on SMARTFARM data models and HDFS interoperation to facilitate efficient data query and processing (via the analytics engines). Finally, we are investigating a wide range of SMARTFARM apps for frost damage mitigation, irrigation scheduling, and yield prediction as well as a number of new farm sensor technologies. We hope that SMARTFARM enables growers to ask "what if" questions like the point-of-origin question dynamically, in real time, using the data they have harvested locally – at a much greater frequency and efficacy than is possible today. Closing the loop between grower experience, data gathering, and data analysis will enable growers to leverage their own data assets to improve yields more sustainably.

6. REFERENCES

[1] K. Aberer, M. Hauswirth, and A. Salehi. A middleware for fast and flexible sensor network deployment. Technical Report 2006-006, LSIR - EPFL, 2006.

[2] K. Aberer, M. Hauswirth, and A. Salehi. A middleware for fast and flexible sensor network deployment. In *International Conference on Very Large Data Bases*, 2006.

[3] AgEagle. http://www.ageagle.com [Online; accessed 2-April-2014].

[4] Apache Mahout. http://mahout.apache.org/ [Online; accessed 14-Feb-2015].

[5] Apache Spark. https://spark.apache.org/ [Online; accessed 14-Feb-2015].

[6] Cassandra. http://cassandra.apache.org/.

[7] CIMIS. California Irrigation Management Information System. http://wwwcimis.water.ca.gov/cimis/welcome.jsp [Online; accessed 2-April-2014].

[8] Climate Corporation (A Monsanto Company), 2014. http://www.climate.com/ [Online; accessed 2-April-2014].

[9] CRAN R Project. http://cran.r-project.org.

[10] Department of Energy Green Button Intiative. http://energy.gov/data/green-button. [Online; accessed 25-March-2015].

[11] Google earth. http://www.google.com/earth/ [Online; accessed 14-Feb-2015].

[12] gThrive, 2014. http://www.gthrive.com/ [Online; accessed 2-April-2014].

[13] Hadoop MapReduce. http://hadoop.apache.org/ [Online; accessed 14-Feb-2015].

[14] Lumify: open source big data analysis and visualization platform. http://lumify.io/ [Online; accessed 14-Feb-2015].

[15] Machine Learning Library. http://spark.apache.org/mllib/ [Online; accessed 14-Feb-2015].

[16] Machine Learning Open Source Software. http://jmlr.org/mloss/ [Online; accessed 14-Feb-2015].

[17] MatLab. http://www.mathworks.com/products/matlab/ [Online; accessed 14-Feb-2015].

[18] MyAgCentral. MyAgCentral Launches Cloud-Based Solution For UAV Imagery, Feb. 2014. http://www.myagcentral.com/ http://www.precisionag.com/equipment/ myagcentral-launches-cloud-based-solution-for-uav-imagery/ [Online; accessed 2-April-2014].

[19] OnFarm: SaaS-based Data Visualization Product. www.onfarm.com. [Online; accessed 24-March-2014].

[20] PostgreSQL. http://www.postgresql.org/. [Online; accessed 25-March-2015].

[21] Pow Wow Energy: SaaS-based Smart Leak Detection. https://www.powwowenergy.com/learnMore/. [Online; accessed 24-March-2014].

[22] PrecisionHawk. http://www.precisionhawk.com [Online; accessed 2-April-2014].

[23] Prediction IO. http://prediction.io/ [Online; accessed 14-Feb-2015].

[24] Satellite imaging corp. http://www.satimagingcorp.com/ applications/natural-resources/agriculture/ [Online; accessed 14-Feb-2015].

[25] B. Sethuramasamyraja, H. Singh, and G. Mazhuvancheriparambath. Geospatial modeling of wine grape quality (anthocyanin) for optimum sampling strategy in mechanized on-the-go differential harvesting programs. *International Journal of Engineering Science and Technology*, 2(10), 2010.

[26] Spark Applications. https://github.com/databricks/reference-apps [Online; accessed 14-Feb-2015].

[27] Terravion. http://www.terravion.com/ [Online; accessed 14-Feb-2015].

[28] VineRangers. http://www.vinerangers.com/ [Online; accessed 14-Feb-2015].

[29] WatrHub, 2014. http://www.watrhub.com/ [Online; accessed 2-April-2014].