HAL
archives-ouvertes.fr

Big Data Architecture for Environmental Analytics

Ritaban Dutta, Cecil Li, Daniel Smith, Aruneema Das, Jagannath Aryal

▶ **To cite this version:**

Ritaban Dutta, Cecil Li, Daniel Smith, Aruneema Das, Jagannath Aryal. Big Data Architecture for Environmental Analytics. 11th International Symposium on Environmental Software Systems (ISESS), Mar 2015, Melbourne, Australia. pp.578-588, 10.1007/978-3-319-15994-2_59 . hal-01328610

HAL Id: hal-01328610
https://hal.inria.fr/hal-01328610

Submitted on 8 Jun 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution| 4.0 International License

Big Data Architecture for Environmental Analytics

Ritaban Dutta[1], Cecil Li[1], Daniel Smith[1], Aruneema Das[2], Jagannath Aryal[2]
[1] CSIRO Digital Productivity Flagship, CSIRO Hobart, Tasmania 7001, Australia
[2]University of Tasmania, CSIRO Hobart, Tasmania 7001, Australia
ritaban.dutta@csiro.au

Abstract. This paper aims to develop big data based knowledge recommendation framework architecture for sustainable precision agricultural decision support system using Computational Intelligence (Machine Learning Analytics) and Semantic Web Technology (Ontological Knowledge Representation). Capturing domain knowledge about agricultural processes, understanding about soil, climatic condition based harvesting optimization and undocumented farmers' valuable experiences are essential requirements to develop a suitable system. Architecture to integrate data and knowledge from various heterogeneous data sources, combined with domain knowledge captured from the agricultural industry has been proposed. The proposed architecture suitability for heterogeneous big data integration has been examined for various environmental analytics based decision support case studies.

Keywords: big data · architecture · machine learning · semantics

1 Introduction

The ultimate challenge in agricultural decision support systems is to overcome the data unavailability and uncertainty to improve the natural resource management efficiency and achieve better business objectives. Uncertainty factors in the agricultural and environmental monitoring processes are more evident than before due to current technological transparency achieved by most recent advanced communication technologies. Poor data quality and uncertainties make most agricultural decision support systems unreliable and inefficient. This inefficiency leads to failure of agricultural and environmental resource management [1-3].

It is evident that there is a serious need to capture and integrate environmental and agricultural knowledge from various heterogeneous data sources including sensor networks, individual sensory systems, large scale simulated models, patched historical weather data, satellite imagery, domain knowledge and contextual user experience for better decision support with high reliability and confidence. This approach would be able to provide a much wider framework for complementary knowledge validation and meaningful utilization of the acquired knowledge. In agricultural domain knowledge is very volatile in nature, which can only be passed from generation to generation, can vary on a daily basis based on requirements. The farmers can understand and only tell the story of the field, just by looking at the colour of the soil or by looking at the sky. One of the key motivations of this research was to develop an architecture, which can incorporate unstructured, undocumented, ad-hoc knowledge

into a structure rule base to be used directly in the big data analytics for better decision support system. Recent development of heterogeneous big data analytics and architectural development for big data integration for agricultural solution was the main motivation behind this work.

2 Big Data Centric Architecture

In the intelligent environmental knowledgebase (i-EKBase) project we aimed to integrate multi scale heterogeneous data from 'Australian Water Availability Project (AWAP)', 'Australian National Cosmic Ray Soil Moisture Monitoring Facility (CosmOz)', 'SILO Data', 'ASRIS Soil data', '250m resolution \ NASA MODIS data', '30m resolution NASA LANDSAT data', 'Australian Digital Elevation Data', and finally 'Domain knowledge is available (i.e. farmer's long experience of daily decision making based on generation wise climate adaptation knowledge). Then on top of this large data integration, our novelty was to apply data driven spatio-temporal artificial intelligence (machine learning) analytics to learn and establish new environmental correlations to make the decision support system more sustainable and scientifically justified to improve business profitability and productivity [4-6].

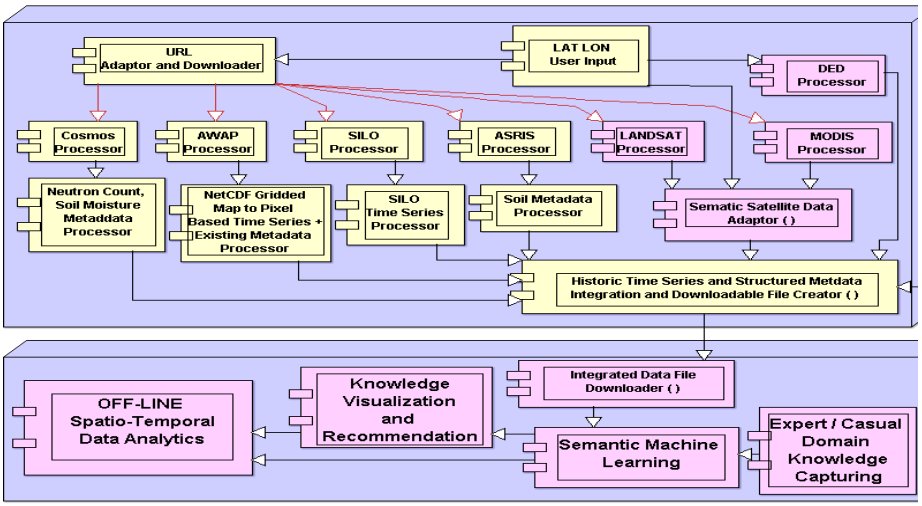

Fig. 1. UML diagram of the proposed architecture.

Fig.1 shows the UML diagram of the system architecture that has been proposed and implemented in this study. In the proposed mechanism we process the available metadata separately from the data observations. Feature extraction process was applied to the metadata to create a unique feature base from the individual data source. This processing was done using data mining and text mining techniques. It also involved unsupervised machine learning techniques (i.e. unsupervised clustering tech-

niques). Main application area of machine learning technique was in the processing of actual observations from various sensors and systems. Various supervised machine-learning techniques (i.e. neural networks) were used to extract feature base from the observation database. Uniqueness of the architecture in this context was in the selection of machine learning methodology to extract targeted feature base from the observations. The targeted feature base was determined on the basis of the actual application and also based on the available domain expert knowledge usually undocumented and belonging to individuals (i.e. farmers) as long-term field experience. We used the domain knowledge to provide guidance to the machine learning processes to extract the desired feature space. This was done to make the feature extraction process more meaningful and unique. This approach was able to make the machine learning process (so called black box approach) more usable in terms of bio-physically explanations for the domain people. Domain guided extraction using machine learning was named as semantic extraction hence it produced semantic feature base. Meta-feature base and semantic feature base were integrated to form an enriched feature space, which was a most significant representation of the heterogeneous big data. The dimension reduction is an important step in any big data related architecture, especially when extremely large number of highly correlated observations is present. In this architecture we show that dimension reduction could be done in a domain specific meaningful way to increase the efficiency of the system, and also to increase the accuracy of the system by enriching the data semantically [1-5].

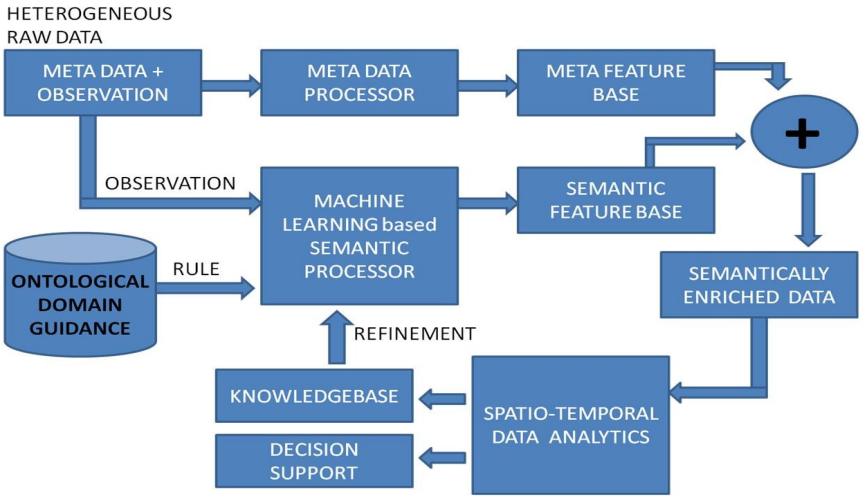

Fig. 2 Concept diagram of the proposed architectural overview to integrate big heterogeneous knowledge.

In the second phase of the architecture, various supervised and unsupervised algorithms were used to analyze the extracted and enriched spatio-temporal feature space. Temporality of the feature space was determined based on the required decision support frequency of the targeted domain and the application. Fig. 2 shows the concept

diagram of the proposed architecture for agricultural decision support system based on true heterogeneous data integration.

3 Architecture for Domain Knowledge Capturing

In this part of the work we investigated how domain knowledge can be efficiently represented in Linked Open Data (LOD) and how this representation can be incorporated for use with machine learning models. Further, we will investigate how to perform learning, inference and prediction tasks with LOD. The work package aims to demonstrate the value of combining semantic and machine learning representations for environmental modeling. Primary objective of this part of the architecture was to establish a standard protocol for capturing essential domain knowledge in a way that could directly be used in automated reasoning and machine learning based processes. Agricultural domain experts and farmers could potentially define the complete feature space required to optimize the harvesting. Often that could be in a casual format or unstructured know-ledge that makes it inaccessible from the system point of view. It is also true that domain knowledge often get lost in translation from generation to generation due to lack of standard mechanism to capture. In all kind of agricultural decision support system there are some essential environ-mental attributes and prior defined thresholds for those variables. The farmers based on their best experience and business profitability in the past usually predefine these thresholds. They also have some special features (i.e. soil colour to determine soil moisture) that they use in the daily decision making process. Decision making process is always a classification problem with two or more possible solutions. It is also possible to find some historical ground truth data (scattered over long time frame) from the farming companies that could be used for cross validations. Fig. 3 shows this part of the architecture.

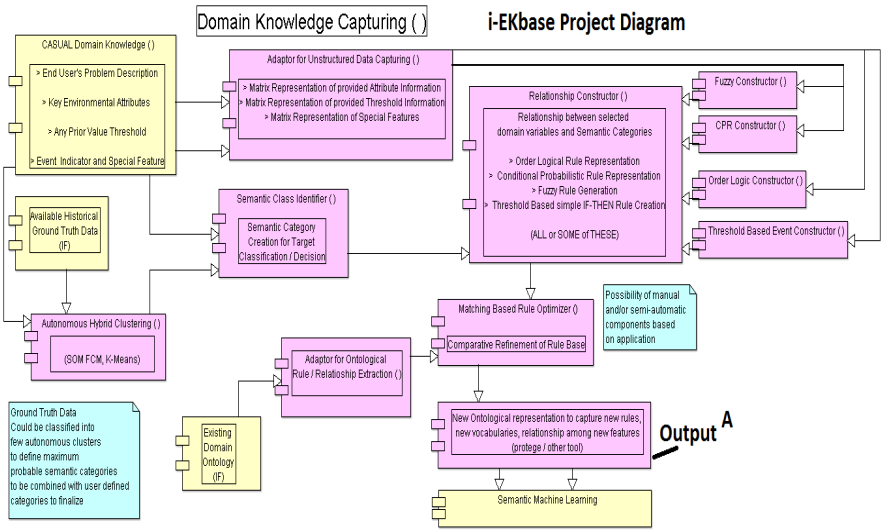

Fig. 3. UML diagram of the domain knowledge capturing architecture.

In this architecture we have applied data driven autonomous hybrid unsupervised clustering techniques (combining Principal Component Analysis (PCA), Self Organizing Map (SOM) and Fuzzy C Means (FCM) to automatically create and match class labels for possible decisions those are defined and stored in farmers mind. Ad-vantage of this autonomous clustering was that this architecture was able to establish few more class definitions. We have defined this part as semantic class identifier. Any of these class labels defining a kind of decision should be related to a set of essential attributes defined by the farmers. We have defined these labels as semantic categories of the decision-making system. Attributes, thresholds and associated decision class could be modeled as a complete association set that was captured as rule in our architecture. This was captured within the relationship constructor module. Four different rule constructors were used to formulate these relationship, namely, fuzzy rule constructor (using fuzzy rule base creation), conditional probabilistic rule constructor (using conditional probability rule generation), order logic constructor (using order logic to formulate rules) and threshold based significant event rule constructor (where threshold of few environmental variables, directly defined by the farmers in conjunction with an event which led to make an unusual decision). In this architecture we proposed to use all or some of these rule constructors based on requirements and availability of ground truth data.

4 Analytics Driven Big Data Architecture

The i-EKbase" knowledgebase architecture represents semantically enriched spatio-temporal features from the environment in a unified manner. The purpose of this work was to learn spatio-temporal patterns acquired from different knowledge sources, translating the learned knowledge into a more efficient decision support system. In this part we describe the detailed architecture behind the semantic analysis. Data driven machine-learning methodologies were employed for this purpose. Machine learning could be used for autonomous knowledge discovery and online inference of knowledge for the intended application scenarios. This approach included “Unsupervised Learning”, “Ensemble Learning”, “Deep Learning”, and "State Based Learning". The aim was to incorporate various learning algorithms on the cloud based computing infrastructure to provide the analytical power required to capture and integrate interesting patterns from the heterogeneous data sources. Captured knowledge from this framework was also integrated with the domain knowledge to be used in future prediction, knowledge recommendation, and decision support systems.

Design of knowledge integration architecture was motivated by the fact that none of the existing data model integration architectures were capable of handling, processing and analyzing multiple large environmental data sources simultaneously. Database on its own does not carry any weightage unless data is converted into knowledge. Based on the domain specific rule or relationship structure an ontology translator was created for automatic reasoning from the dynamic time series data from the environmental sensor and sensor networks. A task of this translator was to convert the domain knowledge into usable format that could be used in functional block called ‘Semantic Signal Translator’ (as shown in the Fig. 4).

The block called ‘Data Conditioner’ was a time series integrator for the semantic signal selector function. Pre-processed time series data were batch processed and represented as daily averaged data for this study. Data from the different sources measuring the same environmental attribute were harmonized. Again different measured attributes from the same node were also harmonized according to the daily average. Simultaneously a generic time window selector function was developed to define the required temporal frequency defined by the domain ontology.

Combination of this kind formed a pool of similar variables, which should be able to validate or complement each other in case of missing values in the time series. Complementary method identified the missing value segments of a time series and replaced those segments with an average segment based on other available time series in the same pool. Next a ‘cross-correlation technique’ was used to measure the similarities between two complemented time series signals representing similar scenarios (in terms of location and time period).

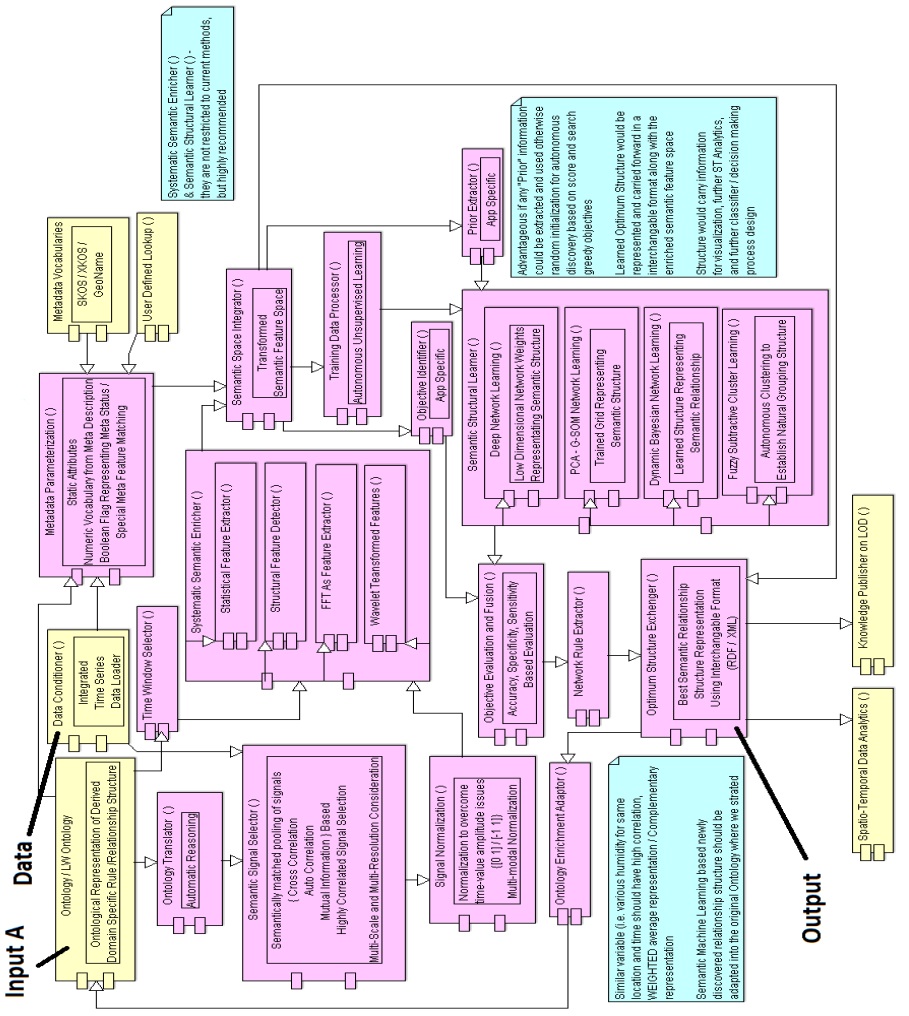

Fig. 4. UML diagram of the semantic machine learning based big data architecture.

Other purpose of this step was to cross validate similar time series data in the same pool to find a representative time series regarding that particular pool. If the two signals being compared were completely identical then the cross-correlation coefficient should be equal to 1 and if there are no similarities between the signals it should be equal to 0. A scoring protocol was designed on cross correlation results. The time series with highest score were selected from each sub group as best representative of the associated environmental variable for that time period. Semantics representations are usually intended as a medium for conveying meaning about some world or environment. One of the significant novelties of this architecture was to implement a generic mechanism to parameterize all possible available metadata into static attributes (i.e. Boolean flag for each of the metadata or a numeric representation with a library of definition). Metadata is "data about the data" and it can provide the description of the what, where, who and how about data. For example, a sensor node metadata could describe when and where the sensor node was deployed, who deployed that node, which environmental attributes are being measured, what are the features or characteristics of that particular sensory system, and finally the valid range of measurement that could be expected. However, metadata are generally used to describe the principal aspect of data with the aim of sharing, reusing, and understanding heterogeneous data sets. In fact, different types of sensor or sensor-simulation model metadata may be considered, namely, static and dynamic sensor metadata and associated sensing information. Other important aspect of the metadata was the inclusion of domain specific decision support class categories and experience based expected outcomes, extracted in the previous domain knowledge extraction block. Parametric representation of metadata provided us with a unique functionality to reason with metadata programmatically in an analytical process. In this process already existing vocabularies (i.e. Geoname, XKOS etc.) and user defined lookup tables were consulted programmatically to refine the parameterization of the metadata [1]. The functional block for this part was called 'Metadata Parameterization'. Data normalization was used in the functional block called 'Signal Normalization' to prepare the integrated sensor response matrix for the subsequent signal pre-processing paradigms on a local or global level. Wide range feature selection using machine learning and signal processing techniques, in a domain specific manner was a key element of this architecture.

The next level of functionality was called 'Semantic Space integrator' where integration of newly extracted features from the semantic enriched block was combined with the Meta features. From the domain perspective this was a pivotal point where wide range of observation-based features was meaningfully integrated with domain specific Meta knowledge. At this point the architecture had enough information to formulate a learning process and design a decision support system. The primary focus moved towards the usage of various supervised machine-learning techniques along with unsupervised techniques to learn about semantically enriched data. Four essential functional blocks were implemented for this purpose, namely, 'Training Data processor', 'Objective Identifier', 'Prior Extractor' and finally 'Semantic Structure Learner'. In the hybrid learning process a training set is always required against a ground truth data set as training target. The 'Semantic Structural Learner' functional block was the core body for learning and spatio-temporal analytics. Four different learners were

used which were based on 'Deep Network Learning (for low dimensional network weights representing semantic structure)', 'Principal Component Analysis (PCA) – Guided Self Organizing Map (G-SOM) network learning (to train a grid structure to represent the semantic structure)', 'Dynamic Bayesian Network Learning (for learning the structure behind the semantic relationship)' and 'Fuzzy Subtractive Cluster Learner (for learning the natural grouping among the feature space using autonomous clustering)'. The parallelization was implemented to establish a wide range of cross-validation and complementary learning. An autonomous unsupervised learner algorithm was implemented to select randomized training sets for the learning phase. The 'Training Data Processor' was responsible for this step. The other two functional blocks, the Objective Identifier' and the 'Prior Extractor' were responsible to generate the ground truth based training target for the training phase and also for the testing phase which is called 'Objective Evaluation and Fusion'. Testing of the learning phase was based on evaluation parameters i.e. accuracy, specificity and sensitivity. To process the evaluation and accuracy estimates from the testing of the trained structure learner models a functional block called 'Network Rule Extractor' was implemented. These rules could be structured representation of some existing rules or they could also be new associations derived as part of the learning. The functionality of this block was to identify new rules those are being generated during semantic structure learning processes. The new rules were represented as new knowledge into the original domain ontology marked as 'Input A'. This was done as part of the implemented enrichment block called 'Ontology Enrichment Adaptor'.

The functional block called 'Optimum Structure Exchanger' was implemented to perform this stage of the architecture. This block was constructed based on Resource description framework (RDF), uniform resource identifier (URI) and triple store technologies. Extracted rules were converted into RDF format to be represented on the LOD. This i-EKbase architecture was implemented on the CSIRO's Bowen research cloud infrastructure. In order to establish the effectiveness of this newly proposed heterogeneous big data analytical framework, this paper presents few real life case studies and associated performance factors to highlight the principal achievements of this architecture [1-8].

5 Big Data Architecture Based Environmental Case Studies

This case study has three main achievements. Firstly a multi-source environmental knowledge framework (i-EKbase) was developed to provide large-scale availability of relevant sensor-model databases for any environmental application. We have developed lightweight ontologies based on extracted metadata from heterogeneous data sources. Next historic surface water balance for one location in Tasmania, Australia was estimated using unsupervised machine learning knowledge recommendation [8].

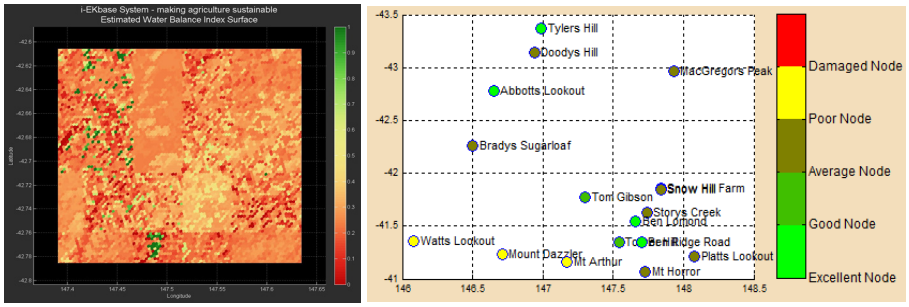

Fig. 5(a). Area wise water balance estimation using multi source heterogeneous data architecture; **(b)** Dynamic Annotation and Recommendation about the sensor network's node based data quality.

The traditional water balance model based method does not include several other environmental parameters (i.e. solar radiation, temperature and humidity), which might have significant influence on the water balance. This issue has been addressed by using multi dimensional and heterogeneous data in i-EKbase system. The machine learning based data driven approach behind i-EKbase system was demonstrated as an effective method to estimate water balance with potentially higher accuracy. Supervised machine learning paradigms were experimented to explore generalization capability and prediction accuracy of this proposed water resource management solution based on multi sensor – model integration [7].

The Radial Basis Function Network based 91.3% accuracy performance proved that newly proposed predictive water resource estimation method based on large multi scale knowledge integration could potentially make the irrigation decision support systems more robust and efficient. Fig. 5(a) shows the area wise water balance estimated using newly developed big data analytical architecture [1, 7-8].

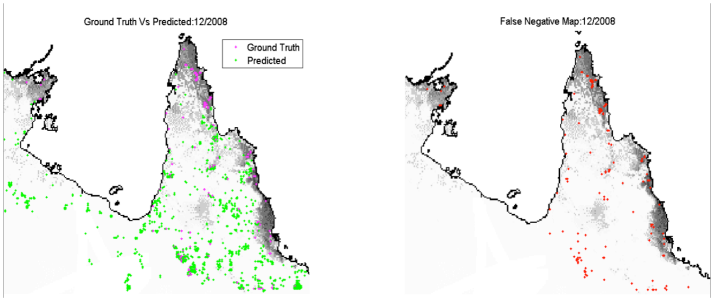

Fig. 6. Bush-fire hot-spot estimation and prediction based on the proposed analytical architecture.

In a different case study we focused on automatic sensor data annotation and visualisation of dynamic weather data acquired from a large sensor network using this newly proposed big data analytics platform. Aim was to develop a data visualisation method for CSIRO's South Esk hydrological sensor web to evaluate the overall network performance and visual data quality assessment. This visual data quality technique developed from this study could be used for quality assurance of any sensor network (See Fig 5(b)). In another case study, Main aspect of this study was to establish a methodology to predict wild fire prone locations as spots on the Australian map based

on publicly available gridded maps of Australian weather variables and hydrological variables. Integrated data sets were created from the gridded maps available from Australian Water Availability Project (AWAP) and Bureau of Meteorology (BOM). On the other hand, NASA-MODIS historical active fire image archives for Australia were used as ground truth. 70% of the data (2008-2009) were used to train all the neural networks where as the monthly image data from 2010 (30% of the data set) were used to test the networks. Independent training and testing were critical for this study to prove the generalization capability of the hybrid architecture based on the neural networks. 94% overall prediction accuracy was achieved from this approach, with 93% sensitivity and 95.3% specificity. Maximum false positive rate was 0.7% whereas overall precision was 96% (Fig. 6) [6].

6 Conclusions

Our understanding of the environment is greatly associated with the interlinked knowledge of the phenomena surrounding to us. Such knowledge is a result of data and extracted information. With the availability of very high and even ultra-high resolution sensor data there is a greater need of managing data, information and essentially the knowledge. With the advent of technological novelties and their wider applications the generated data is surpassing our capacities to store it. There is an urgent need for improved methods and advancement in data-intensive science to retrieve, filter, integrate, and share data.

Data and meaningful information are key for the actors in every walk of life, however, how to conceive, perceive, recognize and interpret such data in space and time is a big question and a big challenge. Taking this challenge into the perspective, we have presented an opportunity of recommending environmental big data using machine learning approaches. We have a firm belief that our simple approach will contribute to the body of knowledge in big data study and big knowledge management in this era of data intensive science.

References

1. Dutta R, et al. Recommending Environmental Big Data Using Semantic Machine Learning, CRC Book on Future Trend on Big Data Analytics, 463-494, CRC Press, Taylor & Francis Group, (2014).
2. Plaisant C, et al. Interface and data architecture for query preview in networked information systems. ACM Transactions on Information Systems (TOIS) 17.3, 320-341 (1999).
3. Marchal S, et al. A Big Data Architecture for Large Scale Security Monitoring. Big Data (BigData Congress), 2014 IEEE International Congress on. IEEE, (2014).
4. Zhong T, et al. On mixing high-speed updates and in-memory queries: big-data architecture for real-time analytics. Big Data, 2013 IEEE International Conference on. IEEE, (2013).

5. Luo, Yi, et al., A Component-based Software Development Method Combined with Enterprise Architecture, 2013 International Conference on Advanced Computer Science and Electronics Information (ICACSEI 2013). Atlantis Press, (2013).
6. Dutta R, et al. Deep cognitive imaging systems enable estimation of continental-scale fire incidence from climate data, Nature Scientific Reports 3, 3188 (10.1038/srep03188), 1-4 (2013).
7. Dutta R, et al., Performance Evaluation of South Esk Hydrological Sensor Web: Using Machine Learning and Semantic Linked Data Approach, IEEE Sensors Journal 13 (10), 3806-3815, (2013).
8. Li Ce, et al., Area Wise High Resolution Water Availability Estimation Using Heterogeneous Remote Sensing and Ensemble Machine Learning, Accepted in 13th IEEE Sensors 2014, Valencia, Spain.