APPLYING BIG DATA TECHNOLOGY TO REMOTE SENSING FOR SPECIES IDENTIFICATION

By

MORTEZA SHAHRIARI NIA

A DISSERTATION PROPOSAL PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHYLOSOPHY

UNIVERSITY OF FLORIDA

2015

TABLE OF CONTENTS

LIST OF TABLES

Table page

LIST OF FIGURES

Abstract of Dissertation Proposal Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Phylosophy

APPLYING BIG DATA TECHNOLOGY TO REMOTE SENSING FOR SPECIES IDENTIFICATION

By

Morteza Shahriari Nia

April 2015

Chair: Dr. Daisy Zhe Wang
Major: Electronics and Computer Engineering

Ecological sciences benefit from the huge diversity of plant species which play an important role in large scale ecological aspects such as global warming, land cover change, CO^2 emission, invasive species, fire hazard, and etc. State-of-the-art species classification techniques utilize remote sensing data such as hyperspectral and LiDAR, however this task involves plenty of field data collection which is both highly time consuming, costly and can only be accomplished by ecological experts. Among thousands of the most commomly found plant species there is huge similarities between them from a remote sensing point of view which makes the task of species classification very daunting; therefore we see a whole body of literature specifically dedicated to this issue which is yet far from real world scenarios with thousands of possible species. While this is an indicator of the importance and complexity of the issue, little has been done to tackle the problem from a computational point of view harnessing the power of "big data". Periodic airborne campaigns can generate terrabytes of data on vast swaths of land. To tackle these problems we propose to use probabilistic knowledge bases and deep learning both of which work best when there is lots and lots of data. Probabilistic knowledge base captures ecological expert knowledge in terms of probabilistic rules, which will be maped to remote sensing data and used to infer new facts and therefore enhance species classification accuracy. Deep learning on the other hand as a semi-supervised algorithm will benefit

from the vast amounts of data available and capture intrinsic features of data through its layered architecture and thus help in reducing the amount of labeled data required.

CHAPTER 1
INTRODUCTION

Understanding the dynamics of ecological structures is very important in determining how climate, land cover , fire hazards, and biodiversity evolve. Precision study of plant species is of high environmental and economical impacts which is only possible through geo-mapping the distribution of plant species abundances at ecological scale. Large scale study of ecological domains has been made possible through spaceborne or airborne campaigns utilizing remote sensing technologies such as *(multi/hyper)-spectral* and *LiDAR*. In this project we focus on airborne hyperspectral and LiDAR data. Each campaign covering tens of acres of land can generate terra-bytes of data depending on measurement resolution (large volume). On the other hand, apart from state-of-the-art machine learning algorithms, there is a great wealth of expert ecological knowledge covering a whole variety of domains (along with their in-ground associated data) that can be used to enhance species mapping that is not being used and is left for ecological scientists for manual interpretation (data variety). Furthermore, data is being generated at faster pace day after day as technology becomes more afordable. After satellite sensors, airborne sensors came into place and now as airborne is still costly there is a surge of interest towards more affordable drone campaigns (Zhou et al., 2009). So we are facing data being generated at unprecedented rates (data velocity). The final aspect is verasity: imperfect sensors, non-standardized measurements, atmosphere impacts (clouds, humidity, aerosols) and et cetera all create uncertainities that need to be accounted for. Velocity, verasity, volume, and variety are the four V's that indicate ecology is stepping into the realm of "big data"(Hampton et al., 2013; Soranno and Schimel, 2014).

1.1 Remote Sensing

From an ecological point of view, there are two types of remote sensing approaches: active and passive. *Passive* remote sensing uses sunlight as the source of energy and sensors captures the intensity of light being reflected from earth's surface. Light intensity

measurements happens at various wavelengths; if a few (usually 3 to 10) relatively broad wavelength bands are captured it is called multi-spectral. If light intensity at dozens to hundreds of narrow band signals are collected it is called hyperspectral. *Active* remote sensing on the other hand uses laser light emission as its source of energy and captures the intensity of returned signals. LiDAR is a popular active remote sensing technology. Below we explain each in more details:

1.1.1 Hyperspectral

Spectrometers measure the amount of light reflected from surface materials: An optical dispersing element (like a prism) refracts the received light into its constituent spectrums and the energy in each band range is measured by a separate detector. Bands can be as narrow as $0.01\mu m$ over a wide wavelength range of typically $0.4\mu m$ to $2.5\mu m$. Figure 1-1 shows the basic components of an imaging spectrometer.

Raw sensor readings (digital number) can be affected by light source conditions, sensor, atmosphere, and surface material. Raw data which is a unit-less light intensity measure is then calibrated into radiance which has a physically meaningful unit through applying a gain and offset to the pixel values. It essentially means how much light the instrument "sees" from the object being observed. Some reference materials like a pure

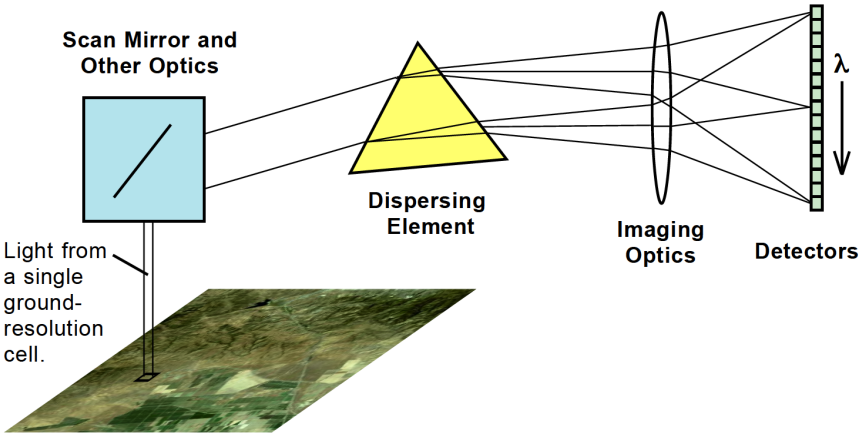


Figure 1-1. Schematic diagram of the basic elements of an imaging spectrometer where λ is the wavelength (Smith, 2006).

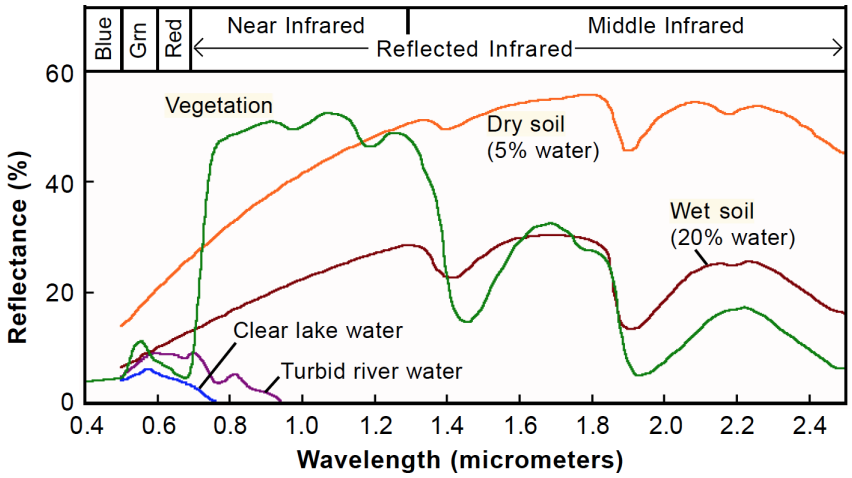


Figure 1-2. Some reflectance examples as how reflectance of different material show different absroption features at differnet bands. (Smith, 2006).

white or pure black sheets can be used in this process. After adjustments for sensor, atmospheric, and terrain effects are applied, pixel reflectance value is calculated which is the proportion of the radiation striking a surface to the radiation reflected off of it. Reflectance demonstrates light abroption features of the surface material and can be compared with field or laboratory reflectance spectra in order to recognize and map surface materials such as particular types of vegetation or diagnostic minerals associated with ore deposits. Reflectance varies with wavelength for most materials because energy at certain wavelengths is scattered or absorbed to different degrees (Smith, 2006). In this project we deal with reflectance values and refer the reader to (Varshney and Arora, 2004) for more details on how to compute reflectance values.

Figure 1-2 shows some example materials when observed through an spectrometer. In vegetations, chlorophyll and some leaf pigments show high absorptions in blue and red ranges and not so much in green; therefore our eyes see vegetation as green. We can see this as a small peak in green compared to other visible wavelength range. From red to near infrared there is a sharp rise known as *red edge* up to a value of about 50% for some plants. High values in the near-infrared region is mainly due to internal cellular structure of leaves which differs significantly across species but can also be different in

a single specie due to plant stress. High reflectance in near-infrared can interact with other leaves in the canopy and therefore its sensor readings can be dependent on canopy structure as well. Beyond $1.3\mu m$ reflectance decreases with increasing wavelength, except for two pronounced water absorption bands near $1.4\mu m$ and $1.9\mu m$. At the end of the growing season leaves lose water and chlorophyll. Near infrared reflectance decreases and red reflectance increases, creating the yellow, brown, and red leaf colors of autumn (Smith, 2006).

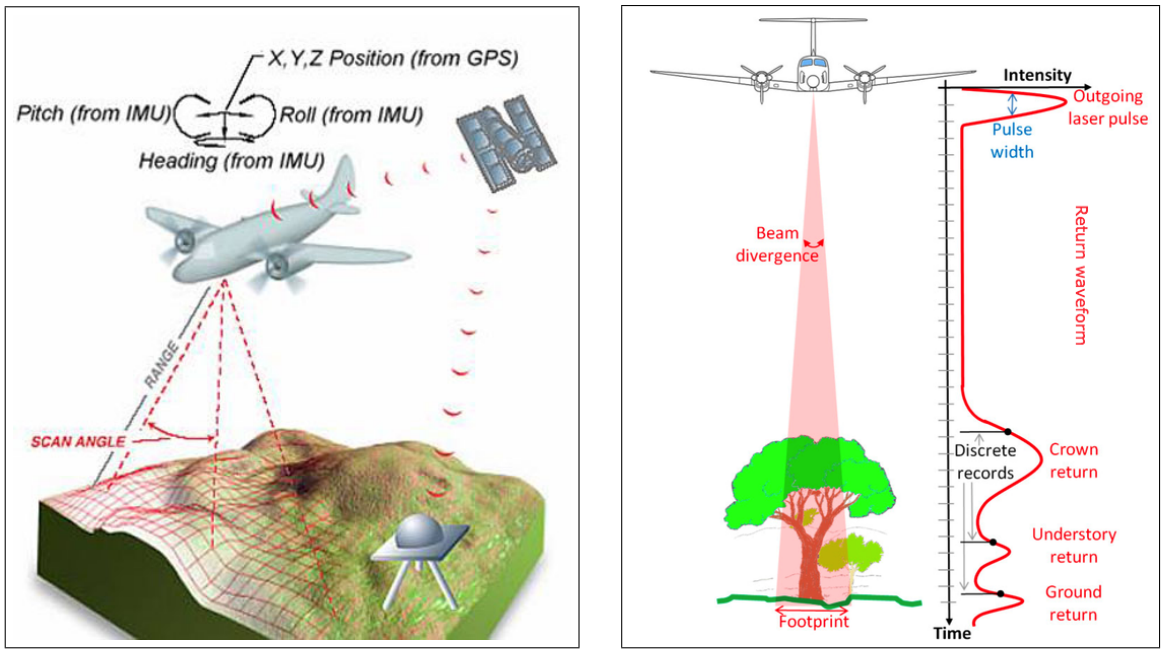
One note worthy concept is the issue of mixing. Pixel size can be large based on the distance between the sensor and target; this makes it likely that more than one material contribute to the signal received by the sensor. The received signal is called a *composite* or *mixed* signal and the "pure" signal that contribute to the mixture are called *endmember* signal. The mixture model can be either linear or non-linear. Linear mixture happens at macro-scale where we have for example a large patch of land beside a large patch of vegetation. If we denote the received signal, land, and vegetation as S,L, and V respectively the mixture model could be $S = 60\%L + 40\%V$. Linear mixture of three endmembers can easily be shown to fall within the triangle where its vertices are endmember spectra. Variations in lighting can be included directly in the mixing model by defining a "shade" endmember (shade, deep water body, dark asphalt or etc) that can approximate light changes and mix with the actual material spectra. Non-linear mixture model on the other hand is more intimate and happens at micro-scale. For example in a microscopic mixture of mineral particles found in soil, a single photon can interacts with more than one material, therefore resulting in a non-linear mixture (Keshava, 2003; Bioucas-Dias et al., 2012; Dobigeon et al., 2014).

1.1.2 LiDAR

Light Detection And Ranging (LiDAR) is a form of active remote sensing where a laser beam is used to detect points on earth. LiDAR sensor is mounted on an airplane and it keeps swiping earth surface as the plane moves forward. By knowing parameters

of the LiDAR sensor itself (tilt angle of the pulse), plane GPS coorinates, and Inertial Measurement Unit (IMU) coordinates (heading, pitch and roll) and the speed of light, the exact coorinate of each point reading is calculated. LiDAR scans at pulse rates of over 300,000 pulses per second (300 kilohertz) depending on sensor technology. By subtracting plane altitude from the distance that laser has traveled we get the ground elevation. Using some filtering techniques points on the ground are classified from points above the ground and this yeilds the height of the objects on the ground. Figure 1-3 (A) demonstrates a schematic view of a LiDAR flightand its parameters (Schmid et al., 2008).

Each LiDAR pulse can have multiple returns depending if it can penetrate through and how much signal intensity is reflected back. Figure 1-3 (B) shows how each LiDAR pulse can have multiple return values. Unlike hyperspectral imaging, LiDAR returns a



A B

Figure 1-3. A) Schematic view of a LiDAR flight: GPS coordinates, IMU coordinates, and tilt angle of pulse together determine the position of the point on the ground (Source: Ohio Department of Transportation). B) Return waveform of a single LiDAR pulse both in its actual continuous form and its 3 discrete records (crown, understory, and ground) each as a point in the point cloud (Fernandez-Diaz et al., 2014).

long list of points <x (latitude), y (longtitude), z (elevation), i (intensity), n (number of this return for given pulse), d (direction of scan), e (edge of flight line), t (time)>. For lidar usually green or near infrared is used as it has high reflectance from vegetation.

1.2 Knowledge Sources

Ecological domain features a large variety of concepts from microbial-scale to soil, plants, animals, geology, climate, and all in between. This shows the wealth of knowledge and inter-related aspects of bio-diversity that needs to be taken into account to be able to have a comprehensive understanding of environment. This information can be found in a variety of sources: Here we would like to introduce some major peer-reviewed data repositories available online: TRY database (datasets covering a wide variety of biomes and geographic areas), TraitNet (Trait databases), USGS Land Cover (planetary diversity, trends, and aquatics), http://ecologicaldata.org and http://datadryad.org (community-driven datasets), DataOne and EarthCube (NSF-funded data archiving resources for ecological, environmental and geosciences data products), iPlant/iAnimal (gene level data portals) and NEON with over 120 high level data products (apart from hundreds of low-level data types) ((Keller et al., 2010), (Lunch et al., 2014)). There are also crowd-sourced datasets available such as UFs iDigBio, BudBurst or Citizen Science initiative platforms.

Besides well-structured data, there is a great wealth of knowledge in textual scientific content that is untapped. Resources such as the Silvics of North America (Burns and Honkala, 1990), local field guides such as (Inventory, 1990; Lillybridge et al., 1995), any of the thousands of peer-reviewed academic journal or conference publications and etc are good sources of knowledge.

There are projects that share the same concept but tackle different domains, the most famous of which is GeoDeepDive from Hazy group at Stanford. GeoDeepDive focuses on geographical domain and collects knowledge from thousands of scientific papers.

1.3 Proposed Work

Based on the domain of information sources available in ecological sciences we propose a system architecture to capture the intrinsic nature of "big data" to enhance species classification in remote sensing applications. We propose to use the concept of probabilistic knowledge bases to contain knowledge sources as a set of probabilistic rules that we can perform inference on. The architecture will be scalable able to handle millions of facts and rules about various concept of ecological sciences. We also propose to use deep learning to act as a semi-supervised machine learning algorithm. We pretrain the network on all the pixels available both labled and unlabled and then fine-tune its weights using the few labeled data that might be available. In the following chapter we describe the details of our architecture in detail.

1.4 Proposal Structure

In chapter 1 we provide an over view of the problem domain. In Chapter 2 we provide our preliminary results on various remote sensing species classification techniques. Chapter 3 provides the details on our information extraction architecture from plain text resources. In Chapter 4 we elaborate state-of-the-art techniques for information extraction and deep learning architectures. Finally we descibe the details of our design for the proposed architecture using both probabilistic knowledge bases and deep learning architectures in Chapter 5.

CHAPTER 2
SPECIES CLASSIFICATION

2.1 Species Classification using SVM

SVM paper

2.2 Unmixing: MESMA, SPICE, PCOMMEND

MESMA results and overview SPICE, PCOMMEND

2.3 LiDAR Stack and Field Statistics

watershed

CHAPTER 3
INFORMATION EXTRACTION FROM TEXT

TREC KBA paper

CHAPTER 4
BIG DATA TECHNIQUES

In this chapter we introduce the tools that we will be using to accomplish the proposal.

4.1 Markov Logic Networks

as first order logic software systems become intractable with not so large set of rules and without even probabilities mln adds probabilities and uses mcmc to tackle scalability.

Markov logic network is a probabilistic logic that applies the concept of Markov network to first order logic. For inference, instead of usning intractable algorithms of prolog or lisp, it uses MCMC sampling.

if initial weights are wrong it is called apriori weights, over time by adding more data, aposteriory weigts and probabilities will be fixed according to data.

there has been plenty of research on first order logic inference and tools such as prolog and lisp have been developed to address this need. The problem with those is that they are highly sensitive to input rule correctness or otherwise they might fall into infinite un-resolutable search spaces. even in the cases of correct rules, search space grows exponentially with the number of parameters and rules and beyond certain thresholdsthey become intractable to compute. Other techniques such as bayesian and graphical models address the rules with a probability assigned to them but they grow exponentially with the number of parameters and variabels in the number for inference purposes. What sampling techniques such as MCMC and graphical models such as MLN do is to use a statistical framwork to derive the probability of queries whether it is MAP or marginal inference. This enables us to achieve as much goodness as we are willing to wait for. Which in reality is pretty good estimates of probabilities in real applications.

4.2 Deep Learning

use large amounts of data at hand.

CHAPTER 5
PROPOSED WORK

In this chapter we demonstrate how we use mln and deep learning for species classification.

5 page proposal

REFERENCES

Bioucas-Dias, José M, Plaza, Antonio, Dobigeon, Nicolas, Parente, Mario, Du, Qian, Gader, Paul, and Chanussot, Jocelyn. "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches." *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of* 5 (2012).2: 354–379.

Burns, Russell M and Honkala, Barbara H. *Silvics of north America*, vol. 2. United States Department of Agriculture, 1990.

Dobigeon, Nicolas, Tourneret, J-Y, Richard, Cédric, Bermudez, José Carlos M, McLaughlin, Stephen, and Hero, Alfred O. "Nonlinear unmixing of hyperspectral images: Models and algorithms." *Signal Processing Magazine, IEEE* 31 (2014).1: 82–94.

Fernandez-Diaz, Juan Carlos, Carter, William E, Shrestha, Ramesh L, and Glennie, Craig L. "Now you see it now you dont: Understanding airborne mapping LiDAR collection and data product generation for archaeological research in Mesoamerica." *Remote Sensing* 6 (2014).10: 9951–10001.

Hampton, Stephanie E, Strasser, Carly A, Tewksbury, Joshua J, Gram, Wendy K, Budden, Amber E, Batcheller, Archer L, Duke, Clifford S, and Porter, John H. "Big data and the future of ecology." *Frontiers in Ecology and the Environment* 11 (2013).3: 156–162.

Inventory, Florida Natural Areas. *Guide to the natural communities of Florida*. Florida Natural Areas Inventory Tallahassee, USA, 1990.

Keller, M, Alves, L, Aulenbach, S, Johnson, B, Kampe, T, Kao, R, Kuester, M, Loescher, H, McKenzie, V, Powell, H, et al. "NEON scientific data products catalog." *WWW document] URL http://www.neoninc.org/science/data [accessed 25 August 2014]* (2010).

Keshava, Nirmal. "A survey of spectral unmixing algorithms." *Lincoln Laboratory Journal* 14 (2003).1: 55–78.

Lillybridge, Terry R, Kovalchik, Bernard L, Williams, Clinton K, Smith, Bradley G, et al. "Field guide for forested plant associations of the Wenatchee National Forest." (1995).

Lunch, C, Kirby, K, and Denicholas, J. "NEON LEVEL 1, LEVEL 2 AND LEVEL 3 DATA PRODUCTS CATALOG." *WWW document] URL http://www.neoninc.org/sites/default/files/basic-page-files/NEON_DOC_002652.pdf [accessed 15 April 2015]* (2014).

Schmid, Keil, Waters, Kirk, Dingerson, Lindy, Hadley, Brian, Mataosky, Rebecca, Carter, Jamie, and Dare, Jennifer. "Lidar 101: An introduction to lidar technology, data, and applications." *National Oceanic and Atmospheric Administration (NOAA) Coastal Services Center* (2008).

Smith, Randall B. "Introduction to hyperspectral imaging." *Microimages. Retrieved on June* 30 (2006): 2008.

Soranno, Patricia A and Schimel, David S. "Macrosystems ecology: big data, big ecology." *Frontiers in Ecology and the Environment* 12 (2014).1: 3–3.

Varshney, Pramod K and Arora, Manoj K. *Advanced image processing techniques for remotely sensed hyperspectral data*. Springer Science & Business Media, 2004.

Zhou, Guoquing, Ambrosia, Vince, Gasiewski, Albin J, and Bland, Geoff. "Foreword to the special issue on Unmanned Airborne Vehicle (UAV) sensing systems for earth observations." *Geoscience and Remote Sensing, IEEE Transactions on* 47 (2009).3: 687–689.