BIG DATA IN ECOLOGY

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

I dedicate this to my parents who selflessly devoted the best of their lives to raise their kids to be hard working, humble, selfless and to take steps towards the greater human wisdom

ACKNOWLEDGMENTS

I would like to thank Dr. Daisy Zhe Wang for believing in me and providing the great opportunity of tackling the whole real of big data. I would also like to thank Dr. Yuguang Fang for his great support and commitment. Dr. Paul Gader and Dr. Stephanie Bohlman were great mentors, without contributions of whom this work would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

Table page

LIST OF FIGURES

Figure page

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

BIG DATA IN ECOLOGY

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 return signals. LiDAR is a popular active remote sensing technology for remote sensing. Below we explain each in more details:

1.1.1 Hyperspectral

1.1.2 LiDAR

1.2 Data Variety

As of

1.2.1 Markov Logic Network

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.

1.3 Proposed Work

1.4 Proposal Structure

CHAPTER 2
REMOTE SENSING

CHAPTER 3
MASSIVE DATA MINING

CHAPTER 4
PRELIMINARY RESULTS

CHAPTER 5
INFORMATION EXTRACTION

CHAPTER 6
CONCLUSION

REFERENCES

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.

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

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.

BIOGRAPHICAL SKETCH

This section is where your biographical sketch is typed in the bio.tex file. It should be in third person, past tense. Do not put personal details such as your birthday in the file. Again, to make a full paragraph you must write at least three sentences.