A User-Centered Approach to Computing Optimization in Ecological Modeling Workflows

Scott Sherwin Farley

Master's Thesis Living Document

Advisor John W. Williams

Introduction

Global environmental change, specifically climate warming and anthropogenic land use change threatens to severely alter biodiversity patterns worldwide. Rates of extinction are increasing and habitat fragmentation and change is likely to be a major factor in determining changes in species occurrence over the foreseeable future. Species ranges are thought to be primarily climate-induced, though other factors, such as other species, may also have significant influence. Using statistical methods, ecologists often forecast the distribution of plant and animal species into the future under different warming scenarios.

Though climate change is threatening to dramatically alter the distribution of species on the earth, scientists are in a good position to forecast and adapt to the coming changes. Environmental monitoring efforts, such as the Long Term Ecological Research Network (LTERN), National Ecological Observatory Network (NEON), and the Paleocological Observatory Network (PalEON), community curated databases, like the Neotoma Paleoecological Database and the Paleobiology Database (PBDB), and modern biodiversity occurrence databases, such as the Global Biodiversity Information Facility (GBIF), are coming to fruition to support global scale environmental change synthesis efforts. New information storage facilities provide a tremendous amount of information to researchers attempting to understand how the earth system will change during the next century. However, as the volume and variety of data increases, so do the challenges associated with dealing with what can now be considered Big Data. While ecological data may in the past have not been considered Big Data, the massive influx of new data clearly requires new techniques to derive insight from the data.

Insight in Big Data is derived from statistical modeling or 'data mining' of the dataset. As datasets grow, the statistical methods used to mine them grow in complexity. Massive datasets, like the popular microblogging service Twitter, require distributed, parallel, streaming models to determine trending topics and other important factors of the real time data stream. Ecologists have begun to apply some sophisticated machine learning techniques to ecological forecasting techniques, and have seen excellent predictive ability in applying these methods. However, traditional methods in ecology, even those methods at

the contemporary cutting edge of the field, are not suited to the large influx of data coming into databases each year. Ecological modelers need to look ahead to a time when there will be over a billion occurrence records in repositories like GBIF, an event that is likely to occur by 2020 [check exact date on this, and cite]. With so many records to work with, ecologists will need to adopt techniques more often associated with fields like geonomics, including distributed processing, ensemble methods, and cloud computing.

Data on paleoenvironmental proxies, including fossil pollen, macrofossils, and freshwater and marine diatoms, add additional information to ecological data collected in the modern era. The addition of paleodata to questions of biogeography and species niches can help researchers come closer to approximating a species' fundamental niche rather than its realized niche [@Veloz:2012jw], which is characterized by the modern data. Furthermore, including paleodata can shed light on species responses to climates that do not currently persist on the globe today. Williams and Jackson [@Williams:2007iwa], not the high probability of encountering novel and no modern analog climates in the near future.

Climate-driven ecological forecasting models, also known as species distribution models (SDMs) have seen extensive use in the ecological discipline, including global change biology, evolutionary biogeography(@Thuiller:2008enb, [@Araujo:2005jy]), reserve selection [@Guisan:2013hqa], and invasive species management [@Ficetola:2007bn]. These models are used by ecologists, land managers, and biologists to characterize a species' biospatial patterns over environmental gradients [@Franklin:2010tn]. These models use model-driven (statistical) or data-driven (machine-learning) techniques to develop a functional approximation of the way in which a species responds to a climatic gradient. A trend towards computationally intensive modeling approaches, including Bayesian methods that rely on repeated sampling of full joint probability distributions [@Dawson:2016wa], is apparent in recent years. These methods utilize occurrence data – places where a species presence was recorded – and the environmental covariates to those places as input into the model, regardless of the algorithm chosen to develop the response surface. As more occurrence data becomes available through portals such as GBIF and Neotoma, these models become increasingly complex. Sequential learning methods, while seeing widespread use in the literature and have demonstrated high predictive accuracy, are not scalable to very large datasets. To date, it has often been acceptable to cut back on amount of modeling, or focus of a study, to comply with computational limitations. However, as more data becomes available to modelers, this will no longer be a viable option [NEED TO CITE].

Cloud computing offers a technological solution to some of the problems posed by the increasing Bigness of ecological data. Cloud computing refers to a broad category of computer architectural design patterns that enable "ubiquitous, convenient, and on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort" ([@Mell:2012jj], [@Hassan:2011uh], [@Anonymous:mc8EfgMa]). With

the rapid commercialization and popularization of cloud computing, scientists have, in practice, an unlimited supply of configurable computing resources at their disposal, with the only practical barrier to their use being the ability to afford to 'rent' the resources. The Cloud has been advertised by many of Silicon Valley's biggest players as the net big thing in the technology industry. It has been credited with Obama's 2012 presidential election win, Netflix's ability to provide streaming entertainment to millions of consumers, and Amazon's massive success in online retailing ([@Mosco:2014cu]). The National Aeronautics and Space Administration (NASA) and the National Science Foundation (NSF) have both officially endorsed the updating of constituent computing system to include Cloud technology [@Mosco:2014cu]. In the geospatial sciences specifically, the cloud has been posited as the future of geospatial computing and modeling [@Yang:2011bd].

Although the cloud seems promising to supports ecology's entry into the Big Data world, it is not a pancea, as there is to-date little guidance on when the benefits, in reduced computing time, outweigh the costs of a cloud-based solution. The Cloud works on an entire different model of computing cost than traditional scientific computing. Transitioning to the cloud comes with a transition away from large, up front captial expenses to a model of monthly usage fees – an operational expense model [@Hassan:2011uh]. In the captial expense model, the computing power must be determined in advance, and users are locked into the level of performance they choose at the time of purchase. Under the cloud model, on the other hand, users may scale up or scale down the number and quality of computing resource they have, or even configure the system to automatically scale the number of resources to the task at hand using a computer algorithm as the load on the server changes. Along with a transition away from traditional desktop computing to cloud solutions comes a marked increase in the complexity of the solutions. Cloud-based solutions are exceptionally complex to set up and maintain, especially for those not experienced in using virtual instances, shell scripting, and IT management. While the complexity costs of engineering and implementing a cloud based solution are difficult to estimate, the computational time gains achieved by running models on faster computers can be measured empirically and combined with estimates of cost per hour to provide guidance on when a cloud based solution would be economically rational.

In this thesis, I develop a theoretical framework to determine the optimal computing solution for a given species distribution modeling workflow. I treat the workflow characteristics as model parameters, and then build a theoretical predictive model that minimizes the time cost of running a computational model while simulaneously minimizes the financial cost of provisioning the computational resources for that run. I gather data on empirical runtimes of four different classes of species distribution models, and then fit a gradient boosted regression tree model to the training data. The fitted model is capable of predicting the execution time of future modeling scenarios, even if the particular combination has not yet been seen. I evaluate the model's predictive skill and evaluate it on a SDM case study.

My findings suggest that if SDMs, and ecology more generally, is to benefit from Cloud computing, future effort must be directed towards developing models that more explicitly take advantage of parallelism and distributed processing frameworks. Currently modeling trends are mostly sequential, and do not leverage more than one computing core. [ADD STUFF ABOUT MEMORY WHEN WE HAVE IT]. The models I have are capable of guiding future modeling efforts in the field.

The remainder of this thesis proceeds as follows: in the following section I introduce my research questions, motivated by the tension between a potential decrease algorithm runtime and an associated increase in computational cost. Section 3 works with the "Four V's Framework" of Big Data to justify ecological datasets as Big Data. Section 4 reviews relevant background literature on system benchmarking, runtime modeling, species distribution modeling, and cloud computing. In Section 5, I present a theoretical framework for predicting the optimal computing solution for a given ecological forecasting modeling. Section 6 describes the methodology used to collect runtime and accuracy data on four different SDM instances. Section 7 presents the results of the experiments and comments on their implications. Conclusions and future work is discussed in Section 8.

2. Research Questions

- Might need to revise?

i. To what degree can the runtime of climate-driven ecological forecasting models be predicted?
    a. Can a predictive model out-predict a null model that suggests that all researchers utilize a single desktop computer for all modeling activities?

ii. Can an optimal solution for a given modeling workflow be predicted using workflow characteristics?
    i. If so, what are the most influential workflow characteristics in making this decision?
    ii. Do contemporary published studies vary in the characteristics that matter most in execution time?

iii. What modeling scenarios are best suited for transition to the cloud, if any?

3. Justify ecological data as big data The vast expansion of data the sciences has necessitated the development of revolutionary measures for data management, analysis, and accessibility [@Schaeffer:2008kl]. Worldwide data volume doubled nine times between 2006 and 2011, and successive doubling has continued into this decade [@Chen:2014fc]. With the large influx of massive geonomic sequences, long term environmental monitoring projects, phylogenetic histories, and biodiversity occurrence data, robust, expressive and quantitative methods are essential to the future the field [@Schaeffer:2008kl]. As datasets become larger, significant challenges are encountered, including inability to move datasets across networks, necessity

> of high performance and high throughput computing techniques, increased metadata requirements for storage and data discovery, and the need for greater agility to respond to new and previously unsupported uses for the data in data access and analysis applications [@Schnase:2014dn].

Ecological occurrence data are records of presence, absence, or abundance or individuals of a species, clade or higher taxonomic grouping that are key to biodiversity analyses, ecological hypothesis testing, and global change research. These data are increasingly being stored in dedicated, large community-curated databases like the Neotoma Paleoecological Database, the Global Biodiversity Information Facility (GBIF), and the Paleobiology Database (PBDB). Since the early 1990s, the internet and associated information technology and an increased willingness to share primary data between scientists precipitated rapid influxes of enormous numbers of digital occurrence records. While there are known problems with the quality and consistency of data records in large occurrence databases [@soberon2002issues], they provide a low-friction way to consume large amounts of data that would otherwise be prohibitively time consuming to derive from the literature or in the field [@Beck:2014ky], [@Grimm:2013uu]]. Entire new fields, namely 'Biodiversity Informatics' [@Soberon:2004jh], 'Ecoinformatics' [@Michener:2012ho], and 'Paleoecoinformatics' [@Brewer:2012bk] have been developed and delineated to address the growing challenges and opporunties presented by the management, exploration, analysis and interpretation of primary data regarding life, particularly at the species level now presented to ecological researchers [@Soberon:2004jh].

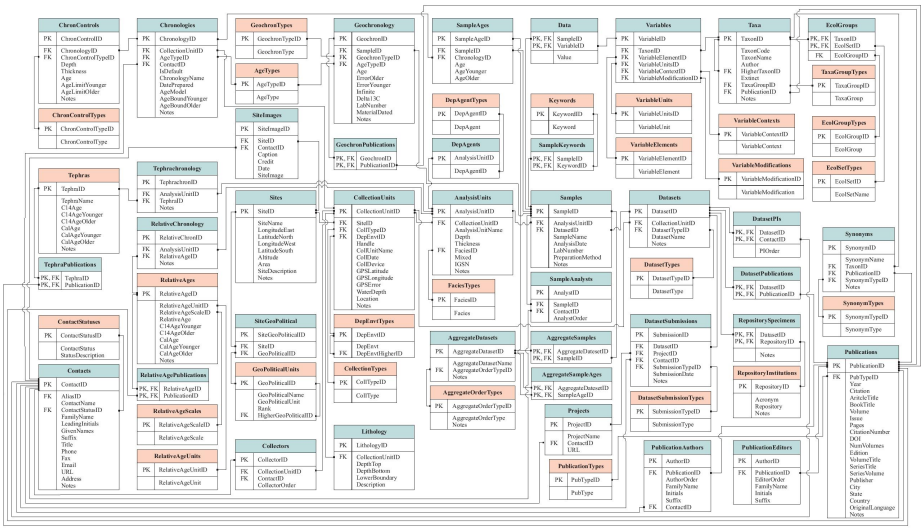
The term Big Data is typically used to describe very large datasets, whose volume is often accompanied by lack of structure and a need for real-time analysis. Big Data, while posing significant management and analysis challenges, brings the opportunity for discovering new insights to difficult problems [@Chen:2014fc]. Though the precise definition of Big Data is loose, there are two prominent frameworks for discriminating Big Data from traditional data. One characterizes Big Data as "a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard statistical software" [@Snijders:2012ww]. This ambiguous delineation is echoed in the advertising and marketing literature that accompanies products like cloud computing that facilitate Big Data analysis. For example, Apache Hadoop, a popular distributed computing framework, has described Big Data as "datasets which could not be captured, managed, and processed by general computers within an acceptable scope" [@Chen:2014fc].

Under this framework, the Bigness of the data is specific to both the time of analysis and the entity attempting to analyze it. [@Manyika:2015vk] suggest that the volume of data required to be Big can change over time, and may grow with time or as technology advances. Furthermore, the criteria for what constitutes Big Data can vary between problem domains [@Chen:2014fc], the size of datasets common in a particularly industry and the kinds of software tools that are commonly used in that industry [@Manyika:2015vk]. The Big Data

label is most often applied to datasets between several terabytes and several petabytes (2^40 to 2^50 bytes). However, because of ecology's lack of experience with massive datasets and limited analysis software common in the discipline, ecological occurrence data can be shown to fall under the pretense of Big Data.

The recent development of complex relational databases that store spatiotemporal occurrence records and their metadata demonstrate that traditional methods of data handling were not sufficient for modern ecological occurrence data. While the datasets are not particularly large in storage volume, they are composed of millions of heterogenous records with complex linkages. Consider the complexity of the relationships between different data records, for example. Figure {X} shows the Neotoma relational table structure, and the complicated web of relationships between each entity. Now, consider keeping track each attribute for hundreds of millions of records, and synchronizing datasets among thousands of independent researchers, and it becomes clear why dedicated databases like GBIF and Neotoma have been developed. Further developments, like application programming interfaces and language specific bindings, supplement the tasks of accessing, filtering and working with the large occurrence datasets [@Goring:2015cr]. While occurrence data does not require the disk space of popular commercial applications like Twitter and Youtube, it has clearly demonstrated a need for new, custom built tools to store, analyze, and use large numbers of records.

NEOTOMA: ENTITY RELATIONSHIP DIAGRAM



A second important framework by which to assess Big Data is the 'Four V's Framework'. First introduced by IBM, it was used by large technological companies in the early 2000's to characterize their data, it is now a popular and flexible framework under which to describe Big Data. Under this framework, a dataset's Bigness is described by its Volume, Variety, Veracity, and Velocity. [@Yang:2013gm] describe this framework, suggesting that "volume refers to the

size of the data; velocity indicates that big data are sensitive to time, variety means big data comprise various types of data with complicated relationships, and veracity indicates the trustworthiness of the data" [@Yang:2013gm] p.276. Using this framework as a rubric, it can be shown that ecological occurrence demonstrates three of the four V's.

Since the late 1990s, the scale of biodiversity information alone has become challenging to manage. Figures [X] and [X] track the growth in collections of Neotoma and GBIF through time. In 1990, only 2 of the records now stored in Neotoma were in digitized collections. Today, there are over 14,000 datasets containing [XXX] individual occurrence records, and associated spatial, temporal, and taxonomic metadata, corresponding to an average growth rate of 1.4 records per day. Nearly all records in Neotoma are derived from sedimentary coring or macrofossil extraction efforts, data gathering techniques that require large expenditures of time and effort [@Davis:1963hk], [@Glew:2002fv]. By number of records, GBIF's collections are far larger than Neotoma's, perhaps reflecting the lower degree of effort required to gather modern ecological occurrence data. GBIF houses digital records of well over 600 million observations, recorded specimens (both fossil and living), and occurrences noted in the scientific literature. Since its first operation in 2001, the facility's holdings have grown nearly 300%, from about 180 million records in 2001 to approximately 614 million records in 2016, demonstrating the data's volume. Note that GBIF's reliance on literature and museum specimens allow its holdings to extend beyond its origin in 2001.

The second characteristic of Big Data in the four V's framework is the Variety of the data, and its 'various types with complicated relationships' [@Yang:2013gm]. Biodiversity data is highly diverse with many very complicated relationships and interrelationships. Neotoma's holdings feature 23 dataset categories, including X-ray fluorescence (XRF) and isotopic measurements, macro fossils or verebrates and plants, modern and fossil pollen records, and freshwater diatom and water chemistry series. Similarly, in GBIF, there are 9 distinct record types, including human observations, living and fossil specimens, literature review, and machine measurements. Though the records coexist in large biodiversity database, they are distinctly different, derived using different protocols by different researchers.

The data's spatial and temporal nature causes complex interrelationships between data entities. All of Neotom's records and 87.6% of GBIF's records are georeferenced to specific places on the earth's surface. The spatial information in these databases is supplemented by other fields that describe the location of the observation, such as depositional setting, lake area, and site altitude, to improve contextual interpretation of occurrence data. Managing data with a spatial component is nearly always more challenging than managing data without it [FIND GOOD QUOTE HERE]. Furthermore, occurrence data represents the work of many dispersed, individual researchers and research teams. The controlled vocabularies and organization of aggregating databases helps to efficiently assimilate large numbers of records, however, nearly every record was collected, analyzed and published by a different scientist. While some scientists have

contributed many datasets to occurrence databases, most have only contributed one or two. The median number of datasets contributed to Neotoma is only 2 and the third quantile value is just 7 datasets. Each researcher is likely to use different equipment, employ different lab procedures, and utilize different documentation practices, contributing to a highly variable dataset.

Biodiversity data also has high levels of uncertainty associated with it – the third V in the Four V's Framework. Some of the sources of uncertainty in the data, like spatial or temporal position uncertainty can be estimated [@Wing:2005wl] or modeled [@Blaauw:2010kg]. Other sources of uncertainty have yet to be quantified, for example inter-researcher identification differences, measurement errors, and data lost in the transition from field to lab to database. A recent paper by the Paleon working group used expert elicitation to quantify the differences between the dates assigned to European settlement horizon, a process they argue varies between sites, and depends on the "temporal density of pollen samples, time-averaging of sediments, the rapidity of forest clearance and landscape transformation, the pollen representation of dominant trees, which can dampen or amplify the ragweed signal, and expert knowledge of the region and the late-Holocene history of the site." The findings of this exercise suggest that paleoenvironmental inference from proxy data is highly variable between researchers. Moreover, some information will undoubtedly be lost in the process of going from a field site through a lab workflow to being aggregated in the dataset. Though some procedural information accompanies the data records, not all process details can be incorporated into database metadata fields, and probably more importantly, contextual details essential to proper interpretation of the data often gets lost on aggregation.

Both Neotoma and GBIF show high levels of quantifiable uncertainty, and are likely to show high levels of unquantifiable uncertainty as well. Of a random sample of 10,000 records of the genus *Picea* from GBIF, over half did not report spatial coordinate uncertainty. Of the 4,519 records that did, the average uncertainty was 305 meters, and the maximum was 1,970 meters. Clearly, such high levels of uncertainty might be problematic for modeling efforts [@Beck:2014ky]. Neotoma records show a similar uncertainty in their temporal information. Neotoma records each have a minimum, maximum, and most likely age for each age control point (e.g., radiocarbon date). Out of a sample of 32,341 age controls in the database, only 5,722 reported any age uncertainty at all. The summary statistics for these age controls suggest that the median age model tie point has a temporal uncertainty of 260.0 years. The 25% percentile is an uncertainty of 137.5 years and the 75% 751.2 years, suggesting that dates are only identifiable down to $\pm$ 130 years of the actual date. [NEOTOMA UNCERTAINTY THROUGH TIME]. Considering sediment mixing, laboratory precision, and other processes at work this is a relatively minor uncertainty, but it certainly contributes to occurrence data's lack of veracity.

The final piece of the Big Data framework is the dataset's velocity, which characterizes the dataset's sensitivity to time. High velocity data must be

analyzed in real time as a stream to produce meaningful insights. Tweets, for example, are analyzed for trends as they are posted. User's are drawn to participation in up-to-the minute discussion, and significant effort has been put towards sophisticated algorithms that can detect clusters and trends in real time [@Kogan:2014hh], [@Bifet:2011wa]. The rate of increase in data volume in both Neotoma and GBIF is not fast enough to invalidate the results from previous analyses, suggesting that it's velocity is not enough to warrant streaming Big Data techniques. Neotoma's growth rate of approximately 1.4 new datasets each day (1990-2016 average) and GBIF's daily growth rate of about 59,000 records (2000-2015 average) are small compared to the total number records in the database. Unlike in many private sector applications, there is little incentive to researchers to immediately analyze new biodiversity records, since all new findings will be reported on in the academic paper cycle, typically several months to years. Moreover, automated analyses of distributional data have been warned against, due to the overall poor data quality [@soberon2002issues] and high levels of uncertainty.

While not time sensitive, ecological occurrence data requires advanced, sophisticated techniques to store and analyze, and demonstrates high volume, low veracity, and significant variety, and should therefore fall under the auspices of Big Data. Remaining traditional techniques of occurrence data storage analysis are likely to be unsuitable in the coming years due to large annual data growth rates. Both GBIF and Neotoma are experiencing sustained and increasing growth since the early 1990s. To fully and accurately derive value from new data being added to distributional databases, new advanced techniques for modeling and analyzing this data are required. Many of the modeling algorithms used in ecological data analysis should be reevaluated to bring them into the world of big data and to take advantage of the advanced computational infrastructure now available.

Species Distribution Models

Species Distribution Models (SDMs) quantify the relationships between a species and its environmental range determinants through statistical methods [@Svenning:2011jq]. While these models sometimes include mechanistic or process components, they most often refer to correlative models, after [@Elith:2009gj]. SDMs rely on ecological occurrence data to provide training data to which statistical learning procedures are applied to estimate the species-specific response to a particular environmental or climatic covariate. With the widespread availability of statistical software and machine learning code libraries, and increased availability of environmental and occurrence data, the utilization of this technique has grown substantially in recent years [@Franklin:2010tn; @Svenning:2011jq]. SDMs are used in a variety of fields related to global change biology and have been shown to provide reliable estimates of climate-driven ecological change when compared to independent datasets. Figure [x] shows the dramatic increase of academic literature focusing on "topic=species distribution models" in Web

of Science.

SDMs work by approximating the functional form of the species niche. Hutchinson ([@Hutchinson:2016tg]) characterized a species' fundamental niche as an n-dimensional hypervolume that defines the environmental spaces where the intrinsic population growth rate of the species is positive [@Williams:2007iwa]. The realized nice describes the subset of environmental space that the species acutally occupies at some point in time, and is smaller than the fundamental niche due to biotic interactions with other species. Most scholars argue that SDMs come close to approximating the species' realized niche [@Guisan:2000tc; @Soberon:2005vt; @Miller:2007br], though the inclusion of fossil data in the model fitting process can increase the likelihood that calibration captures the fundamental niche [@Veloz:2012jw] and improve the assumption of niche conservatism [@Thuiller:2008ena].

While SDMs are often used in the context of forecasting the effects of 21ˆst century ecological change, their calibration and application to paleogeographic problems can provide important commentary on their function and accuracy. The paleorecord provides a well-documented set of species occurrences and community responses to large, rapid, and/or persistent environmental changes and spatial extents ranging from local to global and at temporal resolutions ranging from subannual to millennial [@Maguire:2015ev; @NoguesBravo:2009iv]. While niche models fit with paleodata face a number of additional challenges, often related to the data's veracity, they have the potential to harness information provided by additional training data, mitigate the effect of *a priori* assumptions, and enable ecological hypothesis testing of the drivers of environmental ranges.

SDMs rely on three important assumptions, As a fundamental justification for applying predictions across space and time, all SDMs assume niche conservatism, i.e., that the niche of species remains constant across all spaces and times [@Pearman:2008it]. While the addition of paleodata to model fitting increases enlarges the modeled niche, niche adaptation, evolution and speciation are not modeled. [@Peterson:1999ff] suggests that species typically demonstrate niche conservatism on multi-million year time scales. Second, SDMs rely on the assumption that species are at equilibrium with their environment, a phenomenon that occurs when a species occurs in all environmentally suitable areas while being absent from all unsuitable ones [@NoguesBravo:2009iva]. Given dispersal limitations and biotic interactions between species, this is rarely the case in practice. For example, many European species are still strongly limited by postglacial migrational lag [@Svenning:2008gs].

Finally, SDMs must deal with extrapolation to novel and no-analog climates for which there is no training data. As inductive learning algorithms, SDMs are fitted with a set of labeled target examples to develop a mapping between the features of the examples and the output of the example. In this case, environmental covariates to species presence are used to learn species presence or abundance. Inductive learning is severely impacted when it is used to predict onto future examples that were not included in the set of training examples. Williams et

al 2007 [@Williams:2007iwa] note the high likelihood of encountering novel and no-analog climates in the near future. Fitting the models with data from the paleorecord increases the likelihood that climatic assemblages will have been encountered by the learning algorithm during the fitting process. However, given rapid and highly uncertain climate change, the problem of projecting models onto unseen climates still exists.

Despite the strong assumptions that must be made, SDMs have been used for a wide variety of paleo and contemporary studies of geographic and environmental distribution. In the paleo domain, SDMs have been used to support hypotheses on the extinction of Eurasian megafauna [@Anonymous:2008jc], identifying late-Pleistocene glacial refugia [@Waltari:2007gc; @Keppel:2011ft; @Flojgaard:2009ha], and to assess the effect of post-glacial distributional limitations and biodiversity changes [@Svenning:2008gs]. SDMs are often combined with genetic, phylogeographic, and other methods to develop a complete assessment of a species biogeographical history [@Fritz:2013er].

A Taxonomy of Species Distribution Models

SDMs can range from simple 'boxcar' algorithms that develop a 'climate envelope' for a species to a multivariate bayesian techniques that use Markov Chain Monte Carlo methods to develop probability distributions around projections. While all have the same fundamental goal of characterizing responses to climatic gradients, [@Franklin:2010tn] notes the multiple ways in which SDMs can be categorized. One conceptually meaningful way to group modeling algorithms is into data-driven, model-driven, and stochastic algorithms. The data-driven model-driven dichotomy is introduced in [@Hastie:2009up] and employed by [@Franklin:2010tn] in her text on SDMs. I add the burgeoning methods of stochastic, probability-based Bayesian methods to this taxonomy due to their recent uses and high accuracy.

The goal in SDM is to use a learning algorithm and training examples to approximate the relationship between a set of inputs and outputs. The supervised learning paradigm relies observation of process to assemble labeled training examples or mappings between inputs and outputs where both values are known,

$$T = (x_i, y_i), i = 1, 2, ..., N$$

The learning algorithm approximates the real relationship $\widehat{f}$ by evaluating a loss function based on the difference $y_i - \widehat{f}_i$. The observed inputs may be a p-dimensional vector of observed input features, $X = x_1, x_2, ..., x_p, p = 1, 2, ..., P$. The resulting functional approximation then has a p-dimensional domain. Hastie et al (2009) argues for approaching supervised learning in terms of functional approximation, noting that it encourages the utilization of geometrical concepts of Euclidean spaces and mathematical concepts of probabilistic inference.

Model-driven, parametric, or statistical methods fit parametric statistical models to a dataset. Hastie et al (2009) suggests that these models demonstrate low

bias but high variance, in other words, it relies heavily on *a priori* assumptions about the parametric form of the chosen model. These models were the first to see substantial use in SDM applications and have seen widespread continued use because of their strong statistical foundations and ability to realistically model ecological relationships [@Austin:2002vy]. The earliest models were simple boxcar algorithms, fitting simple multidimensional bounding boxes around species presence in niche space [@Guisan:2000tc]. Other model-driven techniques include variants of logistic regression and generalized linear models on binary outputs [@Vincent:1983uw; @Franklin:2010tn].

In terms of asymptotic complexity, parametric methods tend to be some of the least complex. Consider the generalized linear model, which generalizes simple linear models into their multivariate case, so that

$$\gamma_i = \beta_0 + \beta_1 x_i + ... + \beta_n x_n$$

The user of this model must then specify a link function that describes how the mean of y_i depends on the linear predictor, e.g., $g(\mu_i) = gamma_i$, as well as a variance function that describes how the variance of y_i depends on the mean μ_i. Depending on the matrix decomposition method, the asymptotic runtime complexity of least squares is either $p^3 + Np^2/2$ operations or Np^2 operations. Fitting a model with lasso regression also has this complexity [@Hastie:2009up]. Sure to converge?

The increase in available computing power has spurred the development and application of non-parametric, data-driven, machine learning modeling algorithms. These models, have, in some cases been shown to significantly out perform their model-driven counterparts [@JaneElith:2006vt]. These models demonstrate high bias but low variance, as they do not rely on any stringent assumptions about the underlying data, and can adapt to any situation, though any particular subregion of the model depends on a handful on input points, making them wiggly and highly sensitive to small changes in the input data. Data-driven algorithms include genetic algorithms [@JaneElith:2006vt], regression trees [@Elith:2008el], artificial neural networks [@Hastie:2009up], support vector machines, and maximum entropy techniques [@Elith:2010cea]. Since 2006, MaxEnt, a maximum entropy algorithm has seen widespread use and has demonstrated its ability to perform consistently even on small sample sizes [@Phillips:2006ffa]. A review of recent SDM literature suggests that MaxEnt is the most popular SDM method in use today. Recent evaluations of MaxEnt, however, suggest that its performance, especially on small, presence only datasets, may be questionable when compared with other SDM algorithms [@Fitzpatrick:2013cb].

The asymptotic complexity of data-driven models tends to be larger than that of model-driven algorithms, because more passes over the data are typically required. To fit an additive models with p-dimensional inputs and N training examples, the total number of operations needed to fit the models is $pNlogN + mpN$, where m is the number of applications of a smoothing method, typically less than

20 [@Hastie:2009up]. Support vector machines (SVMs) with m support vectors require $m^3 + mN + mpN$. Multivariate adaptive regression splines (MARS) require $NM^3 + pM^2N$ operations to build an M term model. Building regression trees require $pNlogN + pNlogN$ operations, suggesting that in the worst cases, trees require N^2p operations. Random forests build Q full trees and average the results, making their complexity QN^2P. Boosting, a technique that combines many weak learners into a committee ensemble also increases on the complexity of building a standard tree because it sequentially builds trees stagewise until a loss function has been minimized. Specific implementations of any specific algorithm depends on its implementation and language details. Not sure to converge?

The third and most complex grouping of SDMs involve the application of bayesian methods and stochastic draws from posterior densities in the estimation of species presence. These models are on the cutting edge of application in ecology.

4. Selected literature review

- This will need some serious revision from last spring
- Focus more on the ecological dimensions of why this is important
- Then connect to computing, machine learning, etc
- Finally, review algorithms and optimization techniques

1. Species distribution models
2. What are they? (brief)
3. Ecological foundations, niches, use of paleodata to improve accuracy
    - Data availability
4. Machine learning and species distribution models
    - Models used to be simple (boxcar models)
    - Now they're very complex
    - High variance, low bias
    - Low variance, high bias
    - Look at cited AUC/accuracy metrics
    - No clear winner for all tasks
    - All methods are still widely used
    - Maxent and its popularity
    - Ensemble and parallel methods and their application/accuracy
5. Prediction and hindcasting using models as a key way to understand the past and future
    - Cite land manager uses here (this is more than just hypotheses for ecological testing)
    - These are real issues that need support (invasive species)
6. Meta-analysis/results of targeted reading
    - Other papers commenting on the growth of the field
    - This will flow nicely from the review of what people actually use these models for
7. Cloud computing as a technology to support researchers
8. Support for machine learning

9. Designed for big data and distributed processing
    - We've already clarified that ecological data is Big Data, so this will be easy to reinforce here
10. The cloud as a research tool, rather than a market device
    - Not too much on this, but note the economic underpinnings of the computing as a service
    - Cite NSF/NASA/others that require cloud computing for research
11. Benchmarking, timing, and why it matters
12. Systems evaluation and benchmarking
    - Overview of types of benchmarks
    - Application level benchmarks are the best
    - Need for repeated measurements
    - Point of section: stochastic variance in benchmarks
    - Non-linear, complex, hard to model
    - But it's okay
    - Potentially, consequences of using virtual instances –> few, using monitor scripts
13. Algorithms Optimization
    1. What affect's an empirical/theoretical runtime?
    - Introduce my experimental variables
    - Need to read more on the theoretical underpinnings of memory/paging/CPU/etc
    - Briefly touch on theoretical runtime complexity
    2. Other attempts at empirical runtime modeling
    - Need to read more on this
    - We extend this away from just algorithm inputs to hardware inputs too.
    3. Sensitivity analysis vs. optimization analysis
    - Maybe we need to change some terminology here,
    - I think with the alg. opt. literature I can still call it optimization and prediction.
14. Problem Formulation

- Do I need to update this? Probably more or less close to being done

6. Specific components of the framework to address in the thesis

- The framework introduces six components involved in the optimization
- I just look at one of the central components (time to compute, and address the others tangentially)
- Demonstrate the proof of concept of the framework, leave the other components to other researchers

7. Methods
8. Data collection
    1. Species distribution modeling inputs
    - GBIF and Neotoma
    - Climate model output

- Data preparation and cleaning

2. Simulated data for large memory experiments
- Do I need to do this? Maybe GBIF would let me do a real species.
- Simulated data would make more sense from a computing standpoint
- Real data would make more sense from a user/thesis standpoint

3. Cost model data
- Does this go in data? probably

9. Computing experiments
    1. Computing set up
    - Flowchart framework
    - Google cloud description
    2. Serial SDM experiments
    3. Inter-model differences
    4. Taxonomic differences
    5. Parameter sensitivity
    6. Training example sensitivity
    7. Serial SDMs with large memory requirements
    - I think this will be a nice flow of experiment descriptions

    3. Parallel SDM experiments
    - Need to specifically introduce that these need to be considered separately in my framework, because they respond to differences in cores
    - Might have less accuracy or cost more than methods above,
    - Might have more accuracy than methods above, and can be executed on a single core
    - Just random forests
        - Parallel machine learning methods are a topic of active CS research,
        - This probably needs to go into literature review, or could go into discussion/conclussion

10. Predictive Modeling Building
    1. Runtime prediction
    2. Linear model
        - Do I even need to show results of LM?
        - Ref: comments from CI

    3. GBM
        - Able to capture non-linearities

    4. Accuracy prediction
    - Build one accuracy model for each SDM class
    - Can we test this from the literature too?

    4. Cost optimization model building
11. Discussion and Results
12. Computational runtime prediction accuracy assessment
    - Should formalize this
        - Least squares?
13. Accuracy prediction assessment

    - Parallel methods and their accuracy
14. Cost optimization assessment
    - This will be tricky to assess quantitatively
    - Need to think about this more
    - Qualitatively, we can do this fairly easily
15. Case study
    - Need to find a good case study
    - Illustrate model results and utility
    - Discuss limitations and uncertainties
    - Discuss confidence in results
16. Limitations of current approach
    - How much will the additional components of the framework influence the results?
    - Modeling expertise can do more than predictive modeling
    - Stress uncertainties and lack of predictive skill
    - Scientific realities over modeled optima
    - we should try to find some literature about compromising workflows to meet computational demands.
17. Conclusion
18. Reiterate and answer research questions
19. Next steps to reduce uncertainty remaining in the model
20. Areas where additional research is needed
    - Parallel machine learning methods
21. Bibliography