Satisfying energy demands by harnessing sustainable resources is a critical challenge facing our nation and the world. Sustainable resources, such as wind and solar, are fundamentally more variable than traditional fossil fuels and thus incorporating them into the electrical grid creates new issues due to their high variability. The current approach in utilizing sustainable resources requires extra provisioning of slack capacity in order to compensate for this variability which makes sustainable energy deployment costly. Such issues are further complicated by the inherent difficulty in forecasting energy demand, which although well-studied, remains an open problem.

Both energy demand and weather are examples of complex systems in which it is not possible to write down a concise set of equations which accurately describe their behavior. However, we do have a lot of data. In particular, with the advent of the smart grid and smart meters the amount of data being collected continues to increase rapidly. At the same time, over the past decade, machine learning has achieved state-of-the-art results in an ever expanding array of domains – some prominent examples being speech recognition and autonomous vehicle navigation. The data rich environment surrounding the modern energy grid is ideal for the application of modern machine learning methods.

I propose to further the development and deployment of sustainable energy with an approach that combines "big data" with the development and application of state-of-the-art machine learning algorithms. My own "big data" experience stems from previous experience at Google, where I was responsible for building and launching a number of systems operating on web-scale datasets. However, the fundamental question with any such dataset is what to do with it, and in my current role as a Ph.D. student in the Machine Learning Department at CMU, I have the opportunity to study and develop the state-of-the-art in machine learning. It is our belief that by applying these methods to the difficult problems posed by sustainable energy we will advance the state-of-the-art in both fields.

As an example of this approach, we propose sparse Gaussian conditional random fields for forecasting power production at a wind farm. This is the discriminative analogue of the sparse Gaussian Markov random field, a model that has seen a significant amount of work in recent years, particularly using ℓ_1 methods for efficient estimation. Sparse Gaussian Markov random fields model the inverse covariance matrix of a multivariate Gaussian with a sparse graph in which the edges correspond to conditional independencies. However, when we are chiefly concerned with a prediction task, such as forecasting wind power, it has been often observed that discriminative methods can be superior to generative ones. Thus, we propose to extend Gaussian MRFs by combining the benefits of sparse inverse covaraiance selection with ℓ_1-regularized regression.

In particular, we formulate wind forecasting as a multivariate regression in which the output variables correspond to wind power predictions across multiple locations at various future time points. Wind power at any particular location correlates across time and for nearby locations we also have correlations across space. Furthermore, power production is highly dependent on external factors, namely the wind, for which we typically have a forecast. By

jointly modeling both types of dependencies, we hope to develop a model that improves significantly upon the current state-of-the-art. Critically, we also make the additional assumption that the dependencies in this distribution will be *sparse*, which correspond directly to our intuition that conditional dependencies across time and space will be sparse. We encode this assumption using the ℓ_1 penalty which leads to a convex optimization problem.

Formally, letting $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^p$ represent the intput and output variables, we formulate the Gaussian CRF as a log-linear model with

$$p(y|x;\Lambda,\Theta) = \frac{1}{Z(x)} \exp\left\{-\frac{1}{2}y^T\Lambda y - x^T\Theta y\right\} \tag{1}$$

where $Z(x)$ represents the partition function. The maximum likelihood estimator is given by the optimization problem

$$\underset{\Lambda,\Theta}{\text{minimize}}\, f(\Lambda,\Theta) = -\frac{1}{2}\log|\Lambda| + \frac{1}{2}\Lambda S_{yy} + \Theta S_{yx} + \frac{1}{2}\Theta\Lambda^{-1}\Theta^T S_{xx} \tag{2}$$

which is a convex problem, with $np+p(p+1)/2$ parameters. For high-dimensional problems, such as modeling wind forecasts across many locations simultaneously, p grows which makes the maximum likelihood problem a poor estimator. However, we anticipate that conditional dependencies in the wind forecasting problem will be *sparse*, an intuition that we can encode using the ℓ_1 penalty. The resulting optimization problem is more challenging than previous formulations and in our recent work we formulate a new optimization algorithm based upon the Alternating Direction Method of Multipiers (ADMM) technique.

Thus, by utilizing insights from the state-of-the-art in machine learning, we are able to formulate an important problem in sustainable energy in a novel way leading to new insights and results. As a running example, we have considered the problem of forecasting wind power, but its straightforward to see that our explanation generalizes to many scenarios where we are interested in making spatiotemporal predictions. In particular, early experiments in forecasting power demand demonstrate that our model improves over the state-of-the-art solution deployed in the Pennsylvania power grid.

One significant open question that we will address as part of future research is how to scale to this model to "big data" regimes. In order to do this, we must improve the efficiency of our current approach, as well as develop novel ways to decompose the problem across machine boundaries. This is a difficult problem clearly on the boundary of the existing solutions and thus will have impact not just in sustainble energy, but across all domains in which we can apply machine learning with big data. In general, by developing novel machine learning methods to solve the difficult challenges present in the sustainable energy domain, we anticipate that we will have impact on advancing the state-of-the-art in both fields.