Research Statement

Neeti Pokhriyal

My primary research interests lie in the area of developing novel methods in data mining and applied machine learning. I am interested in modelling scenarios characterised by noisy, uncertain and high-dimensional data coming from different sources, sensors, modalities or feature spaces. My aim is to **jointly learn** and reason under **uncertainty** from such **heterogeneous** data. I am passionate about problems that have **broader social impact**, are best tackled in an interdisciplinary and collaborative setting and aim to improve the lives of poor.

1. Philosophy: Given the intricate linkages between technology and society, I seek to develop solutions, where the notions of fairness, inclusion and explainability are as important as robustness, technical rigor, and performance accuracy. I believe that transformative solutions that affect people's lives and livelihood can be only be developed by equal ownership and participation from multiple scientific domains and stakeholders who are involved.

I will, now, describe how some these ideas have been the guiding philosophy of my research journey till now; and later, propose my vision for the future.

2. Doctoral Thesis: Many learning problems involve data coming from multiple sources, sensors, modalities or feature spaces, corresponding to different views of some underlying phenomenon and provide both unique and complementary information. The varied data sources are termed as views, and the task of learning from such multi-view data is known as multi-view learning [12]. For example, learning from audio and video, text and images, web-pages and click-through data, spatially and temporally displaced sensors, texts in different languages etc.

My doctoral thesis, **Learning from Disparate Data: Applications in Biometrics and Sustainability**, was done under the supervision of Prof. Venu Govindaraju, Department of Computer Science, University at Buffalo in September 2019. I was on maternity leave during Fall 2017 and Spring 2018. My thesis focused on developing two classes of models for learning from multiple views in supervised setting. The first class of modelling framework is based on Gaussian Processes which employs Bayesian uncertainty to combine the predictions from multiple views. The second class of methods dealt with factorized subspace learning, where the idea is to learn representation from multiple views such that it captures both the shared as well as the per-view components [2].

2.1. Computational Approach to Poverty Mapping: A major component of my thesis dealt with developing computational techniques for improved measurement of country-wide poverty estimates; and was published in the **Proceedings of National Academy of Sciences (PNAS)**, in Oct'17 [6]. Poverty mapping is the task of mapping the spatial distribution of poverty and associated deprivations for a country. These maps are important to get a baseline depiction of poverty and to assess the impact of interventions by governments and developmental agencies.

Current ways to estimate poverty are costly and time consuming; and as a result these estimates are grossly delayed for poor economies. I showed the efficacy of using anonymized mobile data for high resolution country-wide poverty mapping for Senegal [4]. This work won the National Statistics prize and a monetary award of 2.5K in the **Data for Development Challenge** held Netmob, MIT, 2015 [4].

My work received a grant from **Bill and Melinda Gates Foundation** (OPP1114791)[1] to develop a robust methodology for poverty mapping as well as to travel to Senegal and work with the National Statistics Office of Senegal (NSO), a governmental body responsible for policy planning. It was a collaborative grant with different universities and public-private partnership in Senegal. My share was **USD 20K** and I was the **project lead** from University at Buffalo.

I developed a novel computational framework that takes in data from multiple sources, like environmental data (related to food security, economic activity, and accessibility to facilities) and anonymized mobile phone data. I formulated the task of poverty mapping as a regression problem and build a framework that is based on Gaussian Process regression (GPR), which provides uncertainties associated with the predictions and allowed to combine them from multiple sources. This work was done in collaboration with Dr. Damien Jacques (then

[1] https://www.gatesfoundation.org/How-We-Work/Quick-Links/Grants-Database/Grants/2014/10/OPP1114791

a doctoral student at Universite Catholique de Louvain) and showed that combining disparate data sources, like environmental data, and mobile data provides more accurate predictions of poverty and its individual dimensions and using this approach poverty can, now, be estimated more frequently, accurately and for spatially finest micro-regions [6].

A major challenge occurs due to the existence of varying spatial and temporal granularity of different datasets that are used to predict poverty. We employ an aggregation mechanism that brings the different data sets to a common spatial granularity. Via the use of semi-parametric learning methods we provide insights into the important factors from these datasets that our model deems important as well as uncertainties as a measure of trust. This work also received the **Chih Foundation Research and Publication Award**, May 2019, which is a single award of USD 2.5K given each year for doctoral research related to innovation for the betterment of society at University at Buffalo.

This work emphasized that *big data* sources like mobile phone data and satellite imagery can provide intercental estimates of poverty, which is especially important in view of the **Sustainable Development Goals** (SDGs).

2.2. From Research to Practice: As a part of this collaborative effort, I discussed my work with multiple stakeholders (Sonatel, Senegal) and local governmental agencies like the National Statistics Office, UN Development Program, UNICEF etc in 2019, so that the benefit of these technologies can reach to the most vulnerable of human populations.

As a follow-on grant for **USD 15K**, I collaborated with the Oversees Development Institute (ODI) [2], London and Datapop Alliance [3] to implement the poverty prediction module using mobile phone data into a privacy preserving platform (called Open Algorithms [1]).

During 2019-2020, I started a new collaboration (as an independent consultant) with the **Inter-American Development Bank**[4] (IDB), Washington DC, to develop a computational framework to map inequality and poverty for Haiti, the poorest country in the Western hemisphere that is constantly battered by natural disasters and political instability. I showed how publicly available and aggregated satellite and mobile phone data can be used to *nowcast* the inequality and social indicators of poverty. Some of the detailed inferences for Haiti are given in the working paper [11].

As a follow-on work, I am engaged with the bank to do a participatory exercise with Quisqueya University in Port-au-Prince during early 2021 to transfer the algorithms for poverty mapping in Haiti. I am excited about this work, as it involves building capacity of the Haitian people and enabling change at grassroot level.

2.3. Multimodal learning in Biometrics: The second part of my thesis focused on biometrics. I worked in the area of **behavioral biometrics**, that attempts to learn a usage pattern for a person's behavioral and psychological attributes, like swipe statistics, gait analytics, language usage, etc; and can be combined with physical biometrics, like fingerprint to provide better authentication capabilities. I developed a novel cognitive biometric modality [7, 8], that can be used to build more secure and seamless continuous authentication systems via understanding the language usage of an individual from social media data.

I also developed an algorithm to fuse information from multiple views called **Discriminative Factorized Subspace** that learns factorized subspaces from the different views in a supervised setting, by mapping them to a low dimensional single shared subspace (that captures the common information among the different views) and view-specific subspaces (that captures the information private to each view), and showed its efficacy for touchscreen biometrics [5].

3. Postdoctoral Research

: Since Oct 2019, I began my postdoctoral work funded by the **Institute for Security, Technology and Society**, Dartmouth College, NH and jointly done under the supervision of Prof. Soroush Vosoughi, Department of Computer Science and Prof. Benjamin Valentino, Department of Government. Some of the exciting avenues of my research are as follows:

3.1. Building robust and interpretable models for tracking economic indicators from social media data: The motivation of this work arises from the deficiencies of existing ways of measuring these indicators. Sample surveys have been the de facto instrument for understanding the social, political and economic realities of the population. Such measures are vital in creating accurate official statistics needed to design appropriate policy interventions and shape private investment decisions.

However, existing surveys are time and money intensive, and currently suffers from reduced public participation. As a result, these statistics are available at low frequency; suffers from significant time lag and, generally, not available at finer spatial granularity.

[2] https://www.odi.org/
[3] https://datapopalliance.org/
[4] https://www.iadb.org/en

I proposed a computational model called **Group Additive Gaussian Processes** that uses anonymized **social media** data to capture information about public behavior, especially consumer sentiment and labor flows, and produces accurate high frequency estimates. My model encodes interpretability in the modeling framework and inherits the advantages of Gaussian Processes (GPs), which are robust non-parametric Bayesian methods that provide both an estimate and uncertainty associated with it. This model, also, attempts to exploit the underlying structure in data and, thus, improves its generalization ability beyond the training samples, which is an important concern when learning with small data regimes [3, 9]. Prior works in this area used linear models with simplistic assumptions; limited extrapolatory power and no insights as to why these predictive methods work.

Sensing with social media data has some important limitations, primarily the non-representativeness of the data compared to general population. I am working on an exciting idea to quantify this bias, by exploring the similarities between survey data and social media data.

3.2. Mapping country-wide energy access for the majority world: In June, I was **awarded** the seed grant from the **Irving Institute for Energy and Society**, Dartmouth College, as a PI in collaboration with Prof. Soroush Vosoughi (postdoc advisor) and Dr. Emmanuel Letouze (Datapop Alliance) for **USD 31K**.

This work targets the important problem of **energy deficit** in poorer economies. Existing methods that have attempted to map energy access for a country have not translated well to the realities on the ground as they were, mostly, based on single data source, and primarily used linear models, which are not rich enough to capture the complex relationships between the data and energy access targets.

We propose to efficiently combine vast and heterogeneous sources of data like satellite imagery, mobile phone data, electricity infrastructure data, pollution data with traditional data sources, like census and surveys to accurately learn the relationship between the country-wide access of electrification and cooking fuels at micro-regional level [10]. The idea is to use this relationship to generate estimates of energy access that are relevant in 2020-2021. The advantage of learning from multiple sources is that we draw strengths of each source, while preventing from inheriting the biases existing in either one of them. Since the data sets are dynamic, noisy and time sensitive, we propose a Bayesian learning framework, that provides uncertainty associated with predictions, which will assist in decision making by telling where and how much to trust the individual prediction.

4. Future Vision

: I am excited about the broad applicability of my research work, both from the computational and application point of view. As a part of my research, I have identified several key computational challenges that needs to be addressed so that the real benefit of these technologies can be realized.

I want to lead a research group that works on the important problems of **sustainable human development using knowledge inspired and data-driven computational lens**. My aim is to **combine purely data driven learning** with existing **knowledge** from domain scientists. As purely data driven learning is often stymied with the inadequacies of data, which can encode different types of biases both overt and subtle.

4.1. Short term goals: Some of my short term goals are outlined below:

1. I want to address and mitigate several **algorithmic challenges** that arise when learning from diverse data, especially in application areas highlighted in my work. While some of the challenges are specific to the data sets used (e.g. the inherent noise in satellite imagery, non-representativeness in social media data), other challenges arises when learning from multiple datasets (e.g. the disparity in spatial, temporal and semantic granularity - mobile and social media data are associated with individuals and satellite imagery captures space and time).

2. I find the idea of modeling **uncertainty**, i.e. *knowing what the model does not know* highly relevant in real world problems where data is often noisy and decisions have costly and irreversible effects. In my past works, uncertainty has been associated with the predictions, as a measure of model's trust. In my current work on understanding energy accessibility, I am modeling uncertainty at observations (input driven), at targets (heteroscedastic) and studying how these different types of uncertainty propagates at different timesteps and affects model inference.

3. Other challenges include **robust validation** methodologies, which depends on the availability of good ground truth data. In our case, validation data comes from expensive and sparse surveys/census; and getting spatially-detailed census usually involves building relationship with the country representatives.

4. Existence of different types of **biases** in individual datasets remain a concern too, as is the **shifting distribution** of these datasets between the training and testing times. Nevertheless, the use of novel datasets to estimate and track important indicators, be it consumer confidence for the market or poverty index to ascertain human development, remains very promising. These methodologies hold promise to perform **high-frequency, high-resolution and almost real-time tracking** of indicators of interest, and can

bring in considerable cost savings in between the long cycles of surveys; or even help reduce the frequency of some of the surveys.

5. I want to develop methods that can help **delineate** these statistics along important dimensions of gender, race, spatial regions, urban-rural divide e.g. disaggregating country-wide poverty for different gender; and economic indicators for different races to highlight the inequalities that are hidden by national averaged statistics.

4.2. Long term goals: As my long term goals, apart from the significant challenges highlighted above, I plan to bridge the gap from research to practice. I also plan focus on some of the additional areas as mentioned below:

1. **Incorporating structure**: I am interested in modeling paradigms that help incorporate more **structure** from data. Structure inevitably occurs in different types of data – existence of longitudinal sequences in biological domains; molecular structure in computational biology and chemistry; spatial and temporal dependence in economic data are some often encountered examples. One way to incorporate structure is via using **informed priors** with Bayesian methods. My aim is to combine data driven learning with specialized knowledge from domain scientists.

2. **Data and methods for SDGs**: I am really excited about exploring other novel data and develop better computational methods, which can help estimate and track other important indicators of **sustainable development**, like health and education attainment; access to clean water and energy; other infrastructural needs and financial inclusion and well-being.

3. **Interactions**: Since poverty lies at the crux of many social maladies, as an interdisciplinary work, I am excited to study the **complex interplay** between poverty and disease modelling, or energy access and climate vulnerabilities, and build models that can not only predict but also provide delineation and explanation of its predictions based on socioeconomic factors, so that we can know how a particular disease or a climate event will impact the poor population. The aim is to assist policy makers into designing target policies for the poor. Additionally, I want to build computational tools that can capture the **population-wide impact** to the implementation of a specific policy via novel datasets.

4.3. Potential sources of funding: I have identified some potential sources of funding mainly NSF and its early career programs. I will target awards for early career researchers from various government agencies, like DOE. I will also target UN agencies interested in tracking SDGs along with foundations like Gates, MacArthur, Omidyar, Rockefeller etc. who are interested in policy developmental issues. I will also work on collaborative grants for World Bank (WB), Inter-American Development Bank (IADB) etc. which can help in transitioning research to practice.

I believe in the immense power of interdisciplinary collaboration between domain scientists and computer scientists. However, to unlock the true and broader potential of AI as a transformative engine of societal growth, we need to tackle not only the technical challenges, but also broader reflective concerns related to ethics, privacy, fairness etc. of our proposed solutions.

References

[1] OPAL: Open algorithms for better decisions. https://www.opalproject.org/home-en.

[2] N. Pokhriyal. *Learning from disparate data: Applications in Biometrics and Sustainability.* PhD thesis, University at Buffalo, State University of New York, 2019.

[3] N. Pokhriyal, A. Dara, B. Valentino, and S. Vosoughi. Social media data reveals signal for public consumer perceptions. *ACM International Conference on AI in Finance (ICAIF '20)*, 2020.

[4] N. Pokhriyal and W. Dong. Virtual network and poverty analysis in Senegal. *D4D Challenge Senegal Scientific Papers, Netmob, MIT*, 2015.

[5] N. Pokhriyal and V. Govindaraju. Learning discriminative factorized subspaces with application to touch-screen biometrics. *IEEE Access*, 8:152500–152511, 2020.

[6] N. Pokhriyal and D. C. Jacques. Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*, (46), 2017.

[7] N. Pokhriyal, I. Nwogu, and V. Govindaraju. Use of language as a cognitive biometric trait. In *IEEE*.

[8] N. Pokhriyal, K. Tayal, I. Nwogu, and V. Govindaraju. Cognitive-biometric recognition from language usage: A feasibility study. *IEEE Trans. Information Forensics and Security*, 12(1):134–143, 2017.

[9] N. Pokhriyal, B. Valentino, and S. Vosoughi. An interpretable model for real-time tracking of economic indicators. *Submission, ACM Transactions on Data Science*, 2020.

[10] N. Pokhriyal and S. Vosoughi. Assessing countrywide socio-economic deprivations using auxiliary data sets. *AI for Africa for Sustainable Economic Development Workshop, ACM International Conference on AI in Finance (ICAIF '20)*.

[11] N. Pokhriyal, O. Zambrano, J. Linares, and H. Hernandez. Estimating and forecasting income poverty and inequality in haiti using satellite imagery and mobile phone data. *Working Paper, Inter-American Development Bank*, 2020.

[12] C. Xu, D. Tao, and C. Xu. A survey on multi-view learning, 2013.