Capstone Faculty-Sponsored Project (FS#1)-- Fall 2019

Faculty Sponsor(s):

- Marianthi-Anna Kioumourtzoglou, Assistant Professor, Environmental Health SciencesEarth Institute | Lamont Doherty, Fu Foundation School of Engineering and Applied Science, Mailman School of Public Health;*Email:mk3961@cumc.columbia.edu*; Lab: https://www.mailman.columbia.edu/people/our-faculty/mk3961
- John Paisley, Associate Professor, Electrical Engineering; *Email: jpaisley@columbia.edu*

Project Description: Exposure to air pollution is a major global public health concern. One of the main limitations of air pollution epidemiologic studies is exposure assessment. Although multiple prediction models have been developed to predict pollutant concentrations in high spatio-temporal resolution, the predictive accuracy of these models greatly varies in space and time. Furthermore, these predictions are subsequently included in the health models as "true" exposures, ignoring any uncertainty related to them. To address these two critical limitations, we have developed a novel method, called Bayesian Nonparametric Ensemble (BNE), that integrates information across multiple existing prediction models, using adaptive weights to weigh each model by its predictive accuracy in each space and time point yielding substantially higher predictive accuracy than any single prediction model and other commonly used ensembles. Importantly, BNE fully characterizes the intra- and inter-model uncertainty, a highly desirable feature as it can both allow propagation of uncertainty into health models and help identify those areas with highest uncertainty to guide placement of monitoring stations. We have applied this method in a small area around the city of Boston, MA, using only three existing prediction models. Our goal is to apply this method nationwide using information on > 25 existing prediction models. The students involved in this capstone project can run this research project to apply the developed method nationwide. Specifically, this is a highly interdisciplinary project that is clearly data science driven. The predictions of the existing models to be used as inputs in BNE need to be downloaded and harmonized in a computationally scalable manner (these datasets are expected to massive; these will be daily predictions of pollutants for ~15 years at a 1km by 1 km grid resolution), BNE is a bayesian nonparemetric ensemble that requires understanding of machine learning allowing the students to acquire familiarity with these highly advanced methods, it is a highly interdisciplinary work in collaboration with machine learning experts, biostatisticians, environmental epidemiologists and atmospheric chemists, and the output of this work will be highly impactful for nationwide air pollution exposure assessment that can subsequently inform both regulatory action and the placement of additional monitoring stations at those locations where air pollution prediction is most uncertain. Finally, an additional aim of this project will be to create an interactive visualization tool to present the results of the BNE predictions and their uncertainties, and also create a website that will hold these predictions and the interactive visuals, so researchers around the US can benefit from our analyses. Therefore, this project is ideal for a data science capstone project.

Dataset Description: The data for this project will be predictions from existing spatio-temporal prediction models. Most of these predictions are publicly available and the students will need to download these datasets. There are multiple other prediction models that are not publicly available, but our collaborators have agreed to provide all predictions to us. These datasets have already been identified and are clearly defined. The datasets are definitely big (> 25 models with daily predictions of three pollutants across the US for ~15

years starting at a 1km by 1km grid resolution) and complex since they are available at different grid networks and resolutions and will need to be harmonized for use. Although all datasets are ready for use, there is a need for processing as these will need to be harmonized to be used as inputs in BNE. If the project requires extensive computing resources, I will be able to provide additional funding for this.

Data format: The final dataset of the project (BNE predictions and uncertainties) will be made publicly available at a website both as interactive maps and in a downloadable format appropriate for further use (eg could be csv files, or shapefiles)

Project Deliverables: The ultimate goal of this project is to make publicly available the best possible predictions for three air pollutants in the United States and the spatio-temporal uncertainties of these predictions. So far, my team has developed the method and has performed a small application using data from around the city of Boston, MA, using predictions from three existing models. The goal is to obtain nationwide predictions using information on more than 25 existing prediction models as inputs in BNE.

Possible deliverables:Model, Software, Interactive Visualization tool and website

Required background and skill sets: machine learning, bayesian machine learning, statistical inference. the students should be able to apply the methods described in this paper: https://drive.google.com/file/d/1lk--tA9gptWrwaIgRCDkwnemvAIbWqsq/view

Learning opportunities: Data versioning and management, Data cleaning, Combining data sources, Exploratory data analysis and visualization, Supervised modeling, Working with tabular data, Working with geospatial data

Capstone Faculty-Sponsored Project (FS#2)-- Fall 2019

Faculty Sponsor(s):

- Henning Schulzrinne, Levi Professor of Computer Science, Computer ScienceFu Foundation School of Engineering and Applied Science;*Email:hgs@cs.columbia.edu*; **Project Description**: Providing Internet access to all US residents has been a public policy goal for about twenty years. Private investment, new models such as community networks and networks run by electric utilities, have emerged, but we have a limited understanding of how effective these policies have been. Using very large datasets gathered by FCC documenting broadband availability and performance (speed) as well as financial market data, the project will investigate questions such as whether the network neutrality rules or their absence have affected build-out, whether the presence of certain types of carriers have accelerated build-out and if we can measure the importance of community engagement and media.

Dataset Description: We will be drawing on three very large (multi-GB) national datasets published by the Federal Communications Commission (FCC), namely the Form 477 national broadband availability data and the Measuring Broadband America (MBA) data. The Form 477 data sets have previously been used by other DSI capstone projects to address other questions, but new data is published every 6-12 months. In addition, we will draw on census data (such as ACS) and public financial data (e.g., SEC 10-K data in XML format).

Data format: Downloadable CSV and XML files.

Project Deliverables: The goal is to produce a report suitable for submission to the annual Technology Policy Research Conference (TPRC), an annual meeting of researchers working on communication policy. Two earlier student projects achieved that goal. The overall objective is to further the understanding of national broadband policy and its effectiveness and understand the influence of various related treatment factors, including public subsidies, presence of certain organizations and historical factors such as the nature of telephone service.

Possible deliverables:Report, Paper

Required background and skill sets: Since the project involves very large datasets, ability to work with Google Cloud or similar is helpful. Use of traditional and ML statistical techniques is required. Programming likely in Python.

Learning opportunities: Combining data sources, Exploratory data analysis and visualization, Supervised modeling, Establishing evaluation metrics, Working with time series data, Working with geospatial data

Capstone Faculty-Sponsored Project (FS#3)-- Fall 2019

Faculty Sponsor(s):

- Pierre Gentine, Professor, Earth and Environmental EngineeringFu Foundation School of Engineering and Applied Science;*Email:pg2328@columbia.edu*; Lab: http://www.gentine.com

Project Description: Stochastic climate physics with machine learning -- This project will be dedicated to using machine learning to better represent clouds in climate models. Climate models are currently widely used to make predictions of the future of the Earth's climate such as the evolution of global temperature or changes in precipitation. Yet, those models suffer from important biases, many of them attributable to clouds. Clouds are small-scale processes which cannot easily be represented in coarse-scale resolution (100km or so) climate models. To tackle this we will use machine learning techniques harvesting high-resolution simulations (so called cloud resolving models) to better represent clouds in climate models. The new representation of clouds will then be tested operationally in a climate model. Some of the challenges that we will have to tackle are: 1. Good generalization and out-of-sample generalization, 2. Inherent stochasticity of clouds and rainfall using probabilistic models such as Generative Adversial Networks or variational autoencoders, 3. Respecting statistical and physical invariances (mass and energy conservation). This project requires previous background in machine learning but no previous knowledge of climate science or physics is needed.

Dataset Description: High-resolution climate model outputs available on the Habanero server

Data format: netcdf files

Project Deliverables: The objective of this project is to develop a new representation of clouds in climate models based on machine learning. Specifically, we want to introduce the stochasticity (randomness) of clouds an precipitation using GANs and variational autoencoders.

Possible deliverables:Model, Report, Paper

Required background and skill sets: Python and tensorflow/Keras

Learning opportunities: Supervised modeling, Establishing evaluation metrics

Capstone Faculty-Sponsored Project (FS#4)-- Fall 2019

Faculty Sponsor(s):

- Maura Boldrini, Research Scientist, PsychiatryColumbia College, Graduate School of Arts and Sciences, Vagelos College of Physicians and Surgeons;*Email:mb928@columbia.edu*; Lab: https://datascience.columbia.edu/maura-boldrini
- Hanga Galfalvy, PhD, Associate Professor of Biostatistics (in Psychiatry) at Columbia University Medical Center, Division of Biostatistics at the Department of Psychiatry; *Email: Hanga.Galfalvy@nyspi.columbia.edu*

Project Description: Using autopsy human brain tissue and shotgun proteomics liquid chromatography-tandem mass spectrometry (LC-MS/MS), we are collecting data on the proteomic profile of the human hippocampus in subjects with major depression and suicide and non-psychiatric controls. The project is a collaboration between the Medical Center and the Columbia College main campus. The assays generate huge amounts of data and we are in need of students willing to do data mining to identify peptides and proteins with altered expression between groups, and identify association networks between protein that have a functional relationship with each other. This data-driven approach employing data mining strategies may help discover biosignatures of suicide and depression and new molecular treatment targets for these devastating illnesses that impose a high disease burden on society.

Dataset Description: These are medical data. The database is very large. We will provide the data. All the data are available. The students do not need to gather data. We provide data in excel form and they are ready. The data set is defined by thousands of peptides per subject and is big enough for creating learning opportunities. The data set is ready. We do pre-processing and at the Proteomic Center and CUMC we have all the computing resources needed.

Data format: XLS file

Project Deliverables: The goal of this project is to identify the proteome profile of the hippocampus in depressed suicide individuals and controls. We also aim to identify functional networks of the identified proteins that will reveal molecular pathways altered in these diseases. We have run the proteomic assays and have the data in a set of subjects. We have pre-analyzed the data and need further data mining. We are collecting data on more subjects under a currently funded grant.

Possible deliverables:Model, Report, Paper, Software

Required background and skill sets: We routinely quantify more than 4,600 protein group in digests of brain tissue. With spectral libraries,10,000 protein groups have been quantified (Bruderer et al., 2016) with the Q Exactive HF mass spectrometer that we use. Database searches will be done with Mascot v. 2.5.1 (Matrix Science Ltd.) combined with post-processing with the powerful Elucidator Protein Expression Data Analysis Software (supported by PerkinElmer, Boston MA under an ongoing software maintenance service contract). This system has been used very effectively in previous work (Shimada et al., 2016; Stephens et al., 2018; Wobma et al., 2018a; Wobma et al., 2018b; Yang et al., 2014b). This software, developed by Merck & Co./Rosetta (Hendrickson et al., 2015; McAvoy et al., 2014; Paweletz et al., 2010) extends the dynamic range

of fold-change data, especially for low abundance proteins. This uniquely powerful system matches accurate mass and retention time across all liquid chromatography/mass spectrometry (LC/MS/MS) chromatograms in an experiment, multiplying the effectiveness of our leading edge mass spectrometry platform. The software includes a full slate of statistical tools including principal component analysis, ANOVA and false discovery rate corrections. The groups will be compared against each other and differentially expressed proteins will be identified using standard criteria and false discovery rate calculations incorporated into the software. Protein ratio P-values for differential expression will be calculated by the program using the xdev parameter (Dai et al., 2002; Weng et al., 2006). Multivariate statistical analysis tools including false discovery control (correction for multiple comparisons), principal component analysis, multidimensional scaling, ANOVA and cluster analysis are part of the Elucidator program. Elucidator runs on our in-house Oracle Linux v. 6.4 server with 128 GB RAM and 32 Tb of RAID network attached storage.

Learning opportunities: Data acquisition and scraping, Data versioning and management, Data cleaning, Exploratory data analysis and visualization, Supervised modeling, Unsupervised modeling, Working with text data

Capstone Faculty-Sponsored Project (FS#5)-- Fall 2019

Faculty Sponsor(s):

- Jeffrey Shaman, Professor, Environmental Health SciencesMailman School of Public Health;*Email:jls106@cumc.columbia.edu*; Lab: https://www.columbia.edu/~jls106/
- Sasikiran Kandula, Senior Staff Associate, Environmental Health Sciences; *Email: sk3542@cumc.columbia.edu*

Project Description: The proposed project would estimate and disseminate situational awareness of deaths resulting from suicides and drug overdoses in the US using publicly available, near real-time data from the National Emergency Medical Services Information System (NEMSIS) (1), a national database of EMS patient care information collected in response to emergency 911 calls.

Although mortality data in the US are available through systems such as CDC's WONDER (2), due to established reporting protocols and quality checks, it can take 6-18 months for information to be publicly available. While the need for such robust processes and checks are indisputable, more timely, albeit less accurate, data could aid public health action and is particularly relevant in the context of the increasing mortality from suicides and drug poisonings in the US. As events resulting from these two causes are very likely to result in EMS calls, the NEMSIS system which records, standardizes, aggregates and makes accessible EMS data (3) in a shorter time frame (estimated to be less than a week) is worth exploring as a secondary source to track trends in mortality from these two causes.

The initial aim of the project would be to build a web interface for geographically resolved (census region level) mortality data from suicides and drug poisonings, with potential additional stratification by age, race/ethnicity and gender. We envision an interface with basic functionality - visual components that effectively communicate trends, allow for comparison across different subject groups, support for data/chart downloads etc.

More importantly, while some of the EMS events would be explicitly tagged with ICD codes specific to suicides/drug poisoning in NEMSIS, we hypothesize that better estimates are possible through supervised learning models that utilize other elements in the dataset (for example, medications given, procedures performed during EMS response, etc.). An additional extension could be to develop methods for short-term forecasts of mortality trends, of which our group has considerable experience in the context of infectious diseases.

Taken together, we believe these objectives would allow DSI students to translate data-science techniques from classroom to more applied settings and yield a demonstrable application that would be of value as they pursue employment opportunities, while creating a resource with significant public health utility.

References 1. National Emergency Medical Services Information System. https://nemsis.org/what-is-nemsis/ 2. Centers for Disease Control and Prevention, National Center for Health Statistics (2017) Underlying Cause of Death 1999-2016 on CDC WONDER Online Database. http://wonder.cdc.gov/ucd-icd10.html. 3. https://nemsis.org/view-reports/public-reports/ems-data-cube/

Dataset Description: The National Emergency Medical Services Information System (NEMSIS), a national database of EMS patient care information collected in response to emergency 911 calls. The data is hosted online by a federal agency and is open to the general public. It contains information on most EMS calls reported to the NEMSIS system from 2017 onwards which can be filtered to events of interest. We do not anticipate students would be required to gather additional data or perform complex cleaning/processing. Similarly, we do not believe extensive computing resources would be required. A server to host the application would be provided.

Data format: Downloadable as CSV/XLS.

Project Deliverables: The initial aim of the project would be to build a web interface for geographically resolved (census region level) mortality data from suicides and drug poisonings, with potential additional stratification by age, race/ethnicity and gender. We envision an interface with basic functionality - visual components that effectively communicate trends, allow for comparison across different subject groups, support for data/chart downloads etc.

While some of the EMS events would be explicitly tagged with ICD codes specific to suicides/drug poisoning in NEMSIS, we hypothesize that better estimates are possible through supervised learning models that utilize other elements in the dataset (for example, medications given, procedures performed during EMS response, etc.). An additional extension could be to develop methods for short-term forecasts of mortality trends, of which our group has considerable experience in the context of infectious diseases.

Possible deliverables:Model, Report, Software

Required background and skill sets: CSOR W4246 Algorithms for Data Science, STAT GR5702 Exploratory Data Analysis & Visualization COMS W4721 Machine Learning for Data Science

Learning opportunities: Project planning and scoping, Combining data sources, Exploratory data analysis and visualization, Supervised modeling, Establishing evaluation metrics, Working with time series data, Working with tabular data, Working with geospatial data

Capstone Faculty-Sponsored Project (FS#6)-- Fall 2019

Faculty Sponsor(s):

- Ryan Abernathey, Associate Professor, Earth and Environmental ScienceEarth Institute | Lamont Doherty, Fu Foundation School of Engineering and Applied Science;*Email:*; Lab: https://rabernat.github.io
- Carl Vondrick, Assistant Professor, Computer Science; *Email: vondrick@cs.columbia.edu*

Project Description: Superresolution and Prediction of Ocean Sea-Surface Temperature

Ocean sea surface temperature (SST) is one of the fundamental controls over our weather and climate. SST anomalies can lead to weather anomalies that cause droughts, floods, and other emergencies which strongly impact our society.

This project will explore the application of machine learning to augment the SST observed by satellite. The satellite data have two main limitations: 1 - limited spatial resolution 2 - no observations of the future!

Emerging techniques from deep learning have the potential to help on both counts. We will adopt the technique of "image superresolution" via generative adversarial neural networks from computer vision to the problem of SST. A training dataset of high-resolution SST images will be used to train a model which can effectively enhance the resolution of coarse-resolution satellite images. This problem is very clearly posed and should be a straightforward application of existing techniques. As such, it provides an ideal entry point into oceanography for a data science student.

The second topic involved prediction of SST. This is more challenging. In collaboration with Prof. Carl Vondrick of Computer Science, we will adopt a self-supervised learning technique used to predict video to the problem of SST. The goal is to learn how forecast the SST for days and weeks into the future given observations of the past and present state. This is a more challenging project, because it requires more domain-specific knowledge and requires a deeper understanding of machine learning algorithms.

Dataset Description: As leader of the Pangeo project (http://pangeo.io), I have extensive experience in cloud-based management of large, complex datasets. All work for this project will be conducted in a cloud environment, where the large datasets can be effortlessly processes.

Two datasets will be used:

NASA JPL Multi-scale Ultra-high Resolution Sea Surface Temperature https://mur.jpl.nasa.gov/ The data are distributed in netCDF format, but we will generate a cloud-native copy in Zarr format which is optimized for machine learning pipelines.

MITgcm LLC4320 SST These are simulated SSTs which can be used as a testbed for training and prediction. https://medium.com/pangeo/petabytes-of-ocean-data-part-1-nasa-ecco-data-portal-81e3c5e077be These data have already been converted to a cloud-native analysis-ready format and are stored on google cloud: https://pangeo-data.github.io/pangeo-datastore/master/ocean/llc4320.html

Both datasets will use the new cloud-native Zarr format: https://zarr.readthedocs.io

Data format: Cloud Native Zarr Format

Project Deliverables: Success on topic 1 (SST superresolution) would mean the ability to skillfully predict a high resolution (e.g. 1km pixel size) image from a medium resolution one (e.g. 20km pixel size).

Success on topic 2 (SST prediction) would mean the ability to forecast SST a week in advance based on the previous week of data.

In both cases, it is also important to understand which deep learning architectures are effective for this type of oceanographic application and how model design differs from classic computer vision applications.

Possible deliverables:Report, Software

Required background and skill sets: Students should be experience with deep learning in the context of images, particularly convolutional neural networks. Domain knowledge of oceanography is not expected.

Learning opportunities: Supervised modeling, Unsupervised modeling, Working with image data

Capstone Faculty-Sponsored Project (FS#7)-- Fall 2019

Faculty Sponsor(s):

- Tal Danino, Assistant Professor, Biomedical EngineeringFu Foundation School of Engineering and Applied Science;*Email:td2506@columbia.edu*; Lab: http://daninolab.nyc/
- Svebor Karaman, Postdoc, Electrical Engineering ; *Email: sk4089@columbia.edu*

Project Description: 2. MOTIVATION, BACKGROUND AND OVERVIEW

Bacteria-related illnesses – caused by strains of Escherichia coli, Salmonella enterica, Listeria monocytogenes, Vibrio cholerae, Corynebacterium diptheriae, methicillin-resistant Staphylococcus aureus (MRSA), and Pseudomonas aeruginosa—are responsible for approximately 5 million deaths per year worldwide. Many of these infectious diseases are not only prevalent in the developing world, but are becoming more common in the developed world with increasing instances of food contamination, hospital-acquired infections, and bioterrorism. Identification and monitoring of bacteria outbreaks and antimicrobial resistance is critical for tracking health developments and is recommended by the WHO. Current microbial identification approaches can be extraordinarily time consuming, laborious, and expensive. Specifically, in the case of gold-standard clinical microbiology assays, such as staining and microscopy following culture on solid media in Petri dishes, the assistance of experts for morphological identification is required. Furthermore, these identification methods are often sample destructive. Thus, the development of a rapid, automated process to identify and characterize bacterial species from environmental and clinical samples would be a major health innovation. The aim of this proposal is to develop a machine-learning based approach for the identification and characterization of bacteria via analysis of photographs of bacteria colonies from environmental samples grown on solid media in Petri dishes. The aim would be to build towards machine learning algorithm which could classify colonies on a plate into species, based on aspects such as colony morphology, color, and other features.

Dataset Description: Name: Images of bacterial colonies from soil samples Description: Scanned images of Petri dishes with well-separated colonies from a single species per plate (for training) and scanned images of Petri dishes, captured by Epson scanner at resolution ranging from 400 to 3200 dpi. All jpeg format. Additional Petri dishes from different species on the same plate (goal is to classify these).

Data format: Raw images

Project Deliverables: The goal of this project is to build an algorithm that analyzes an image of a Petri dish with colonies from different bacterial species and classifies them by species. One such deep learning classifier has already been developed to classify colonies from two species, K. pneumoniae and E. coli, which could be extended for the more varied dataset. However, there are a variety of approaches and the research goal of this project is to determine which types of algorithms for this task can perform best.

Possible deliverables:Software

Required background and skill sets: See below

Learning opportunities: Project planning and scoping, Data cleaning, Exploratory data analysis and visualization, Supervised modeling, Unsupervised modeling, Establishing evaluation metrics, Working with image data, Working with time series data

Capstone Faculty-Sponsored Project (FS#8)-- Fall 2019

Faculty Sponsor(s):

- Galen McKinley, Professor, Earth and Environmental ScienceEarth Institute | Lamont Doherty;*Email:mckinley@ldeo.columbia.edu*; Lab: https://galenmckinley.github.io/assets/doc/McKinley-CV-ALL-Apr2019.pdf
- Lucas Gloege, PhD student, Earth and Environmental Science; *Email: gloege@ldeo.columbia.edu*

Project Description: The ocean significantly mitigates climate change by absorbing fossil fuel carbon from the atmosphere. Cumulatively since the preindustrial times, the ocean has absorbed at least 40% of emissions. To understand past changes, diagnose ongoing changes, and to predict the future behavior of the ocean carbon sink, we must understand its spatial and temporal variability. However, the ocean is poorly sampled and thus we cannot do this from direct measurements. In this project, students will build on existing efforts to develop novel neural network and random forest approaches to interpolating sparse ocean data with which the sink can be estimated. Specifically, we will work on developing explicit point-by-point uncertainty estimates to accompany the results of mapped products. We will develop an adaptive Bayesian regression model that uses a Gaussian random field as prior (Dimatteo et al. 2001; De Oliveira 2005) to capture spatial smoothness and allow for explicit point-wise uncertainty estimation.

Dataset Description: SOCAT surface ocean pCO2 are the primary dataset. This is freely available: https://www.socat.info

Other data are satellite sea surface temperature and chlorophyll, observed sea surface salinity, mixed layer depth, and atmospheric pCO2.

All of these dataset are already in use by McKinley and graduate student Lucas Gloege. Data preparation will not be a significant component of this project.

Data format: NetCDF is the typical format

Project Deliverables: I would like to learn what we can do with Gaussian Random Field. Can we use it everywhere in the global ocean? Or perhaps only in some regions? I do not know how challenging this will be, so it is an exploration and so as long as we learn and keep working to make progress, that will be success.

Possible deliverables:Model, Report, Software

Required background and skill sets: In addition to the data science skills and python, students should have an interest in the oceans and climate. It would be great if students have had some introductory earth / ocean science classes in the past, but this is not required.

Learning opportunities: Combining data sources, Exploratory data analysis and visualization, Supervised modeling, Establishing evaluation metrics, Working with geospatial data

Capstone Faculty-Sponsored Project (FS#9)-- Fall 2019

Faculty Sponsor(s):

- Jiook cha, Assistant professor of neurobiology, PsychiatryVagelos College of Physicians and Surgeons;*Email:jc4248@cumc.columbia.edu*; Lab: https://datascience.columbia.edu/jiook-cha

Project Description: Human cognitive neuroscience is the field where rigorous data driven science--powered by artificial intelligence--could have a major impact. Our research group studies genetic and brain underpinnings of human mind and behaviors in normal or pathological conditions. We are interested in applying scalable analytics, i.e., deep neural networks, to big human neuroscience data (3d and 4d brain MRI, genome seq data, cognitive and behavioral data) in supervised, unsupervised, or semi-supervised learning. We have multiple tasks and data more than enough for tens of DSI students' semester research projects. Participating DSI students will get access to the larger research group outside of Columbia. I.e., our team closely collaborates with AI and big data science experts (high energy physics, material science, climate science, or energy science) at Brookhaven, Argonne, and Lawrenece Berkeley national laboratories. We hold regular GPU hackathon at Columbia or National Laboratories, where DSI students can closely and intensively work with AI and HPC experts and GPU programming mentors from NVIDIA. Two to three research teams, each consisting of three to four DSI students, can be organized for distinct research projects/tasks.

Dataset Description: The pi will provide access to the data and computing infrastructure. The data includes 3d and 4d brain MRI, genome sequencing data, and cognitive behavioral data from tens of thousands of human subjects. The full datasets are about hundredala terabytes; students will have access to the partial or full data. The computing infrastructure includes CPU or GPU base supercomputers at national laboratories. The PI and his research team has access to generous allocations multiple supercomputers. Students will have access to the partial or full infrastructure. Nevertheless, the PI request to DSI an access to computing resources for DSI students, e.g., GCP credits, because the academic supercomputers have a high demand.

Data format: Csv xls files, raw and pre-processed images are available on remote database (clouds and supercomputers).

Project Deliverables: 1. Training deep neural networks to learn associations among genomes, brain connectomes, and phenomes. 2. Predicting an abnormal brain and cognitive developmental and aging process including current status and future trajectory. 3. Training deep neural networks to learn domain invariant brain signals linked to human mind and behavior in normal or pathological conditions.

Possible deliverables:Model, Paper, Software

Required background and skill sets: Strong interests or sufficient hands on experience in deep neural networks. Students should have successfully completed the following courses: a DSI core course, a graduate level deep learning course or alternatively either an undergraduate-level or certified online deep learning course with demonstrated sufficient hands-on experience in deep neural networks. Exceptions: Even if students have no prior experiences in deep neural network, if they have either (1) exceptionally strong interests and motivation in deep learning and *neuroscience/psychology/medicine/or health sciences* (a must), (2) strong backgrounds in statistics (not biostat), (applied) math, or (statistical) physics, (3) hands on

experience in high performance computing, or (4) exceptional programming skills, the pi will financially support accelerated online training (e.g., Coursera courses) of deep learning. In this case, contact the pi to schedule a in person meeting.

Learning opportunities: Project planning and scoping, Data cleaning, Combining data sources, Exploratory data analysis and visualization, Supervised modeling, Unsupervised modeling, Working with text data, Working with image data, Working with time series data, Working with tabular data, Distributed neural network. High performance data analysis. High performance computing

Capstone Faculty-Sponsored Project (FS#10)-- Fall 2019

Faculty Sponsor(s):

- Colin Wayne Leach, Professor, Psychology, Africana Studies, IRAASBarnard College, Graduate School of Arts and Sciences, School of Social Work;*Email:cwl2140@columbia.edu*; Lab: https://colinwayneleach.weebly.com [colinwayneleach. weebly. com]
- Courtney Cogburn, Assistant Professor, Social Work; *Email: cc3803@columbia.edu*

Project Description: A growing number of behavioral and data scientists are aiming to leverage the best available statistical techniques to analyze the temporal sequence of activity in multiple forms of media so as to trace influence of activity from one form to another. In this case, we are interested in tracing over time the influence of activity in news media and social media (i.e., Twitter) centered around incidents of police use of force against unarmed Black victims. Key indicators of important activity are sentiment, syntax (e.g., active vs. passive verbs -- killed vs. shot). In addition to the important and interesting social issues addressed, the analytic issues include integrated analysis of two multi-level datasets each embedded in its own network and with its own temporal dynamics.

Dataset Description: I have obtained all necessary data. The database of online news articles on Ferguson immediately after the killing of Michael Brown was collected by former Columbia Psych grad students Turetsky and Riddle (2018) who have published a first paper with it and are now collaborators. "Articles about the Ferguson shooting were collected from the top overall 51 online media sources and top 18 African American–oriented online media sources, identified by Pew Research Center (2015) based on number of unique visitors in January, 2015. Using the search engines of sources' websites and Google search functions, we obtained all articles containing the key word "Ferguson" published by these sources in the 10 days after Michael Brown was shot, August 9–19, 2014. After accounting for overlap in these sources and sites without any relevant articles, our final sample included 3,284 articles from 66 online news sources (16 of which are oriented to African Americans)." The Twitter data was purchased by a collaborator in communications at NYU (see Freelon et al. 2016). These Tweet IDs and can be turned ("hydrated") into tweets in a JSON file with Hydrator software: https://github.com/DocNow/hydrator. "We purchased directly from Twitter all public tweets posted during the yearlong period between June 1, 2014 and May 31, 2015 containing at least one of 45 keywords related to BLM and police killings of Black people under questionable circumstances (see Table 3). The keywords consist mostly of the full and hashtagged names of 20 Black individuals killed by police in 2014 and 2015. The resulting dataset contains 40,815,975 tweets contributed by 4,435,217 unique users." The subset of this data that overlaps with the period of the news articles in August 2015 is most relevant: Aug 9-31, 12,589,057 Tweets from 1,734,541 unique users. Other periods during the year may also be relevant for comparison (e.g., Nov 24-Dec 2, the period when it was announced that the office who killed Michael Brown would not one indicted).

Data format: CSV/XLS file of news article database, Tweet IDs into Twitter JSON with Hydrator software: https://github.com/DocNow/hydrator

Project Deliverables: The overarching goal is to compare the sentiment and other qualities of linguistic and visual representations of a serious news event (the killing of Mike Brown and the subsequent protests in

Ferguson) in news media and social media. This should enable an assessment of influence from one form of media to the other. The analyses should enable a publication (or two) in behavioral science journals, and perhaps related media and other accessible reports/summaries to increase dissemination. The results may also be used to inform journalistic practice, which is a possibility I am discussing with colleagues at the School of Journalism.

Possible deliverables:Model, Report

Required background and skill sets: JSON data file creation and management for Twitter data: https://github.com/DocNow/hydrator Text analysis/sentiment analysis Time series analysis Possibly, if there is interest and expertise: social network analysis; machine learning for image categorization

Learning opportunities: Data versioning and management, Data cleaning, Combining data sources, Exploratory data analysis and visualization, Supervised modeling, Establishing evaluation metrics, Working with text data, Working with image data, Working with time series data