CSE6250 Projects: Big Data Analytics for Healthcare

Jimeng Sun

***Abstract*—CSE6250 Big Data Analytics for Healthcare is a graduate level course focusing on practical big data technology for health analytic applications. One big part of this course is to conduct an individual/group project that addresses a real-world data science problem in healthcare. The project should provide an end-to-end coverage of data science activities in addressing a real healthcare problem. The project should utilize big data systems such as Hadoop and Spark, machine learning algorithms that are covered in this class and real-world health related data. I hope that the best projects (with some additional effort) can lead to publications at the best medical informatics venues such as Journal of the American Medical Informatics Association (JAMIA), Journal of Biomedical Informatics (JBI), Journal of Medical Internet Research (JMIR), Artificial Intelligence in Medicine, IEEE Journal of Biomedical and Health Informatics (JBHI).**

This document provides the project guideline such as expectation, timeline, deliverables. We also introduce recommended project topics for selection but you are welcome to propose your own project as long as they are related to big data technology covered in this course and addressing healthcare problems.

***Index Terms*—Big data, Health analytics, Data mining, Machine learning**

I. Introduction

BIG data and healthcare applications interact closely nowadays thanks to the advancement in electronic data capturing technology such as electronic health records, on-body sensors and genome sequencing. This course is about learning practical skills on big data systems, scalable machine learning algorithms and their applications to healthcare. Through (painful) homework exercises, all the students should have by now learned big data systems and acquired sufficient knowledge about healthcare data. We believe you are ready to take on the next level of challenges as a data scientist in healthcare. That is, you are going to propose, execute and report an awesome data science project. The final results of this project includes **1) a publishable report and 2) a convincing presentation, and 3) reusable software and sufficient documentation from your project.**

Next we will cover the project life cycle, timeline, deliverables, grading scheme and project topics.

II. Project Life Cycle

As a data scientist working on a real-world project, you have to be able to conduct all aspects of the big data project independently in a timely manner. In particular, here are some tasks that a data scientist will have to conduct in a big data project: project initiation, project execution and project report.

J. Sun is with School of Computational Science and Engineering Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.sunlab.org).

A. Project initiation

As a data scientist, projects are not always there for you to work on. You have to create them and convince your boss (e.g., your CEO) to fund that. Before your project is officially launched, you have to conduct many steps to make that happen. Here are the checklist of things that you should do during the project initiation.

1) Identify and motivate the problems that you want to address in your project.
2) Conduct literature search to understand the state of arts and the gap for solving the problem.
3) Formulate the data science problem in details (e.g., classification vs. predictive modeling vs. clustering problem).
4) Identify clearly the success metric that you would like to use (e.g., AUC, accuracy, recall, speedup in running time).
5) Setup the analytic infrastructure for your project (including both hardware and software environment, e.g., AWS or local clusters with Spark, python and all necessary packages).
6) Discover the key data that will be used in your project and make sure an efficient path for obtaining the dataset. This is a crucial step and can be quite time-consuming, so do it on the first day and never stops until the project completion.
7) Generate initial statistics over the raw data to make sure the data quality is good enough and the key assumption about the data are met.
8) Identify the high-level technical approaches for the project (e.g., what algorithms to use or pipelines to use).
9) Prepare a timeline and milestones of deliverables for the entire project.

All the above steps in project initiation should be demonstrated in your proposal.

B. Project execution

Once your project is approved, you should quickly work on getting results and iterate with your sponsors on the progress. Iteration is the key. The first iteration should be fast and positive otherwise you are at risk losing momentum from the sponsors/project owners (e.g., your boss, clinical experts, your partners from another organization). This successful execution will lead to long-term sustainability of your team and will

greatly improve your reputation in the organization, so please focus on that.

1) Gather data that will be used in your project if you haven't already.
2) Design the study (e.g., define cohort, target and features; carefully split data into training, validation to avoid overfitting)
3) Clean and process the data.
4) Develop and implement the modeling pipeline.
5) Evaluate the model candidates on the performance metrics.
6) Interpret the results from your model (e.g., show predictive features, compare to literature in terms of finding, present as cool visualization).

All the steps in project execution should be done by the paper draft due date and iterate at least another time by the final due day.

C. Project report

Finally, you are close to the end of the project. You need to summarize what you have done and learned throughout the project. This will be a comprehensive, concise and well-written report as the foundation for future projects. This can lead to publications and other external communication. You will probably need to give a presentation to your sponsors. So do the best you can in written report and presentation. Bad delivery at this step can overshadow all the great work your team have put in throughout the project, so do spend sufficient time to prepare a slick presentation and write a comprehensive report.

- Your report should consists of the following sections.
  1) Title and abstract
  2) Introduction and background
  3) Problem formulation
  4) Approach and implementation
  5) Experimental evaluation
  6) Conclusion
- Prepare a presentation deck and deliver a convincing and informative presentation.
- Clean up and package your code, and document the necessary steps for future usages by others.

Please use the above process to guide your own project for this semester and possibly your future data science career.

III. Logistics

Next we summarize the timeline and deliverables for your project in this semester.

A. Timeline

Due Date	Task Description
Sep 30	Project group formation
Oct 14	Project proposal submission
Nov 11	Project draft
Dec 9	Final Submission (paper + presentation + code)

B. Deliverables

0) Paper Templates:

- Please use either MS Word or LaTex template from the link provided below for your proposal, draft, and final paper, but you should submit it in PDF format at the end.
- AMIA Templates [Word][LaTex]

1) Project Proposal:

- Min. 1 to Max. 3 pages write-up
- Guide:
  - Explain about the problem/topic you choose and how to solve it
  - It is recommended to try to cover as many aspects as described in project initiation if it is possible.
  - Conduct a literature survey and cite at least 4 papers that are relevant to the project.

2) Project draft:

- Up to 5-page write-up + 1 page of references
- Guide
  - Make sure your write-up cover all aspects described in project execution.
  - Conduct literature search and cite at least 8 papers or more that are relevant to the project.

Please check the course web page for the review format and follow it.

3) Final report:

- up to 7-page write-up + 1 page reference (the same template as project draft), see sample papers for reference.
- 5-min presentation (Youtube (include link in paper) or audio attached slides) + slides.
- software implementation and documentation

C. Grading scheme

Your draft and final paper should be in a form of regular research publication. It means all sections common in typical research publications such as Abstract, Introduction, Method, Experimental Results, Discussion, and Conclusion must be there even with different section names or structures. You should organize well and write clearly each section so that it is easily readable for other readers.

Here are the grading guideline for your project.

- Project 40%
  - 3% proposal
  - 7% paper draft
  - 10% final presentation
  - 20% final paper

IV. Project Topics

We introduce several project topics for your consideration but you can also propose your own project outside this scope as long as your project uses big data tools (e.g., Hadoop and Spark) and is about healthcare applications.

A. Chest X-ray Disease Diagnosis

Mentor: Sungtae An ⟨stan84@gatech.edu⟩

X-rays are the oldest and most frequently used form of medical imaging, but they require significant training for clinicians to read correctly. This makes the analysis of x-rays costly, time consuming, and prone to error. Luckily, the latest big data techniques, especially deep learning, are making automated analysis of x-ray images increasingly more realistic, and groups are publishing large x-ray imaging dataset to help researchers train, test, and improve their approaches. Creating an automated diagnosis system would speed up processing, reduce effort from clinicians, reduce errors, and make x-rays more practical for diagnoses that currently rely on more expensive but easier to analyze technologies like computerized tomography.

- **Dataset**:
  - NIH Chest X-ray Dataset
- **Related Work**:
  - Wang et al. [14] introduce the Chest X-ray dataset (containing over 110K images) with an overview of the data sources, methodology for deriving the labels, and provide some initial benchmark results using different pre-trained convolutional neural networks to classify each disease type. They also use a weakly-supervised localization techniques to understand where in the image the network believes the disease occurs and use a subset of 1,600 images to evaluate the accuracy of this localization approach.
  - Rajpurkar et al. [11] significantly improve on the modeling techniques to achieve state-of-the-art results using a much deeper and more sophisticated 121-layer DenseNet architecture. They evaluate the effectiveness of this network versus radiologists and find their network provides even better results. They also use an alternative localization approach to understand where in the image the network identifies a disease but don't provide specific validation results.

B. NLP for Healthcare

Mentor: Balaji Sundaresan ⟨bsundaresan3@gatech.edu⟩

Unstructured healthcare data like clinical notes usually contain much richer information than structured data such as structured parts of electronic health records (EHR) and insurance claims records. However, it is difficult to manually extract useful information from unstructured data in terms of time and labor cost. Therefore, it is getting important more and more to handle the operations of ETL (Extract, Transform, and Load) using Natural Language Processing (NLP) in healthcare domain.

- **Resources**:
  - Criteria2Query: Automatically Transforming Clinical Research Eligibility Criteria Text to OMOP Common Data Model (CDM)-based Cohort Queries
  - github for EliIE[8]
  - github for dataset and algorithm used in [10]
  - MIMIC III, which is useful in many cases. Please submit your data access request to MIT using your GT-Email and MIMIC III Certificate you have done in HW1 via above link. Request is made per person, not per team
- **Related Work**:
  - Kang et al. [8] presented an open-source information extraction system called Eligibility Criteria Information Extraction (EliIE) for parsing and formalizing free-text clinical research eligibility criteria (EC) following Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) version 5.0. EliIE parses EC in 4 steps: (1) clinical entity and attribute recognition, (2) negation detection, (3) relation extraction, and (4) concept normalization and output structuring. A sequence labelingbased method was developed for automatic entity and attribute recognition. Negation detection was supported by NegEx and a set of predefined rules. Relation extraction was achieved by a support vector machine classifier. They further performed terminology-based concept normalization and output structuring. According to their evaluation, machine learning-based EliIE outperforms existing systems and shows promise to improve.
  - Ma and Weng [10] investigated the correlation between drug safety label changes and study population focus shift patterns for existing interventional drug trials. They defined the Convergent Focus Shift (CFS) pattern for each prescription drug as the converged focus in post-marketing trials compared to that in premarketing trials. They hypothesized that drugs with potential safety warnings have different CFS patterns compared to those without warnings. They demonstrated the added value of linked public data and the feasibility of integrating ClinicalTrials.gov summaries and drug safety labels for post-marketing surveillance.
  - Ma and Weng [9] presented a method for identifying questionable exclusion criteria for 38 mental disorders. They extracted common eligibility features (CEFs) from all trials on these disorders from ClinicalTrials.gov. Network Analysis showed scale-free property of the CEF network, indicating uneven usage frequencies among CEFs. By comparing these CEFs term frequencies in clinical trials exclusion criteria and in the PubMed Medical Encyclopedia for matching conditions, they identified unjustified potential overuse of exclusion CEFs in mental disorder trials. Then they discussed the limitations in current exclusion criteria designs and made recommendations for achieving more patient-centered exclusion criteria definitions.
  - He et al. [5] developed a method for profiling the collective populations targeted for recruitment by multiple clinical studies addressing the same medical condition using one eligibility feature each time. Using a previously published database COMPACT as the backend, they designed a scalable method for visual aggregate analysis of clinical trial eligibility features.

This method consists of four modules for eligibility feature frequency analysis, query builder, distribution analysis, and visualization, respectively. This method is capable of analyzing (1) frequently used qualitative and quantitative features for recruiting subjects for a selected medical condition, (2) distribution of study enrollment on consecutive value points or value intervals of each quantitative feature, and (3) distribution of studies on the boundary values, permissible value ranges, and value range widths of each feature.

C. Mortality Prediction in ICU

Mentor: Ming Liu ⟨mliu302@gatech.edu⟩

Accurate knowledge of a patients disease state and trajectory is critical in a clinical setting. Modern electronic healthcare records contain an increasingly large amount of data, and the ability to automatically identify the factors that influence patient outcomes stand to greatly improve the efficiency and quality of care. The goal of this project might be to repeat and improve previous study or to propose a new study using MIMIC-III data[12] implemented with big data tools (e.g., Hadoop and Spark). You can also compare your model with the benchmark testing model.

You must present detailed steps such as the prediction target, feature selection, feature construction, predictive model and performance evaluation. You may inititally start with a small subset of data as you develop your model locally. However, after fine-tuning it, your final paper must be based on results from the entire data.

Do not forget that you should utilize big data analytics tools for the project

- **Dataset**: MIMIC III (Submit your data access request to MIT via above link using your GT-Email and MIMIC III Certificate you have done in HW1 via above link. Request is made per person, not per team)
- **Resources**: Paper, Presentation, Video.
- **Related Work**:
  - Hrayr et al. [4] introduced four clinical prediction benchmarks (including mortality prediction) using data derived from the publicly available Medical Information Mart for Intensive Care (MIMIC-III) database [7].
  - Xu et al. [15] recently developed an attention-based RNN model for predicting mortality using physiological monitoring data at ICUs: MIMIC-III Waveform Database Matched Subset [1].

D. Sepsis prediction or other MIMIC-III Benchmark Tasks

Mentor: Yanbo Xu ⟨yxu465@gatech.edu⟩

Hrayr et al. [4] recently introduced four clinical prediction benchmarks using data derived from the publicly available Medical Information Mart for Intensive Care (MIMIC-III) database [7]. In one of these projects, or with any other idea you want to propose, you will extend the current benchmark further in many aspects.

[1] https://physionet.org/physiobank/database/mimic3wdb/matched/

1) Creating more benchmark tasks: Currently, there are four types of prediction/classification benchmark: in-hospital mortality, decompensation, length of stay, and phenotype. We can add more benchmark tasks such as sepsis prediction. Sepsis is a leading cause of death in the United States, with mortality highest among patients who develop septic shock. Early aggressive treatment decreases morbidity and mortality. While general-purpose illness severity scoring systems are useful for predicting general deterioration or mortality, they typically cannot distinguish with high sensitivity and specificity which patients are at highest risk of developing specific acute condition. You can try to come up with not only sepsis prediction, but also any other potentially beneficial benchmark tasks. Then, the goal of the project is to construct cohort and build some baseline models for the task which can be a new part of the current benchmark set.

2) Enriching the current benchmark dataset: You can also improve the current dataset by adding more clinical variables (features) that have not been included in the dataset yet such as medications, infusions, and treatments. After you modify the dataset, you will need to show how the performance of baselines for each task is changed also. Therefore, the goal of this project will be fully explored variable study supported by some literature survey and utilizing big data analytics with supporting results from each predictive modeling or classification task.

3) Other ideas: You can also propose any other ideas to extend or to improve the benchmark in terms of the dataset or the performance of methodologies.

Do not forget that you should utilize big data analytics tools for the project

- **Main Reference**: *Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, and Aram Galstyan. Multitask Learning and Benchmarking with Clinical Time Series Data. arXiv:1703.07771*[4](Code)
- **Dataset**: MIMIC III (Submit your data access request to MIT via above link using your GT-Email and MIMIC III Certificate you have done in HW1 via above link. Request is made per person, not per team)
- **Related Work**:
  - Henry et al. [6] analyzed routinely available physiological and laboratory data from intensive care unit patients and developed TREWScore, a targeted real-time early warning score that predicts which patients will develop septic shock.
  - Desautels et al. [2] applied a newly proposed definition for sepsis, Sepsis-3, as a gold standard for the implementation of their predictive algorithm, InSight, a machine learning classification system that uses multivariable combinations of easily obtained patient data (vitals, peripheral capillary oxygen saturation, Glasgow Coma Score, and age), to predict sepsis using MIMIC-III dataset, restricted to intensive care unit (ICU) patients aged 15 years or more. Following the Sepsis-3 definitions of the sepsis syndrome, they compared the classification performance of InSight versus quick sequential organ failure assessment (qSOFA), modified early warning score (MEWS),

systemic inflammatory response syndrome (SIRS), simplified acute physiology score (SAPS) II, and sequential organ failure assessment (SOFA) to determine whether or not patients will become septic at a fixed period of time before onset.

- Ghassemi et al. [3] examined the use of latent variable models (viz. Latent Dirichlet Allocation) to decompose free-text hospital notes into meaningful features, and the predictive power of these features for patient mortality. This work considered three prediction regimes: (1) baseline prediction, (2) dynamic (time-varying) outcome prediction, and (3) retrospective outcome prediction. In each, their prediction task differs from the familiar time-varying situation whereby data accumulates; since fewer patients have long ICU stays, as they move forward in time fewer patients are available and the prediction task becomes increasingly difficult. Note that MIMIC-II (not III) was used in this work.
- Xu et al. [15] explored a richer dataset, in which physiological signals including Electrocardiogram (ECG), Photoplethysmography (PPG), vital signs and so on were continuously recorded along with the discrete clinical data. They proposed an attention-based RNN model that can efficiently encode the long-term multi-channel dense signals and predict mortality and length of stay in real time. The dataset used in the work is the recently released MIMIC-III Waveform Database Matched Subset [2].

E. Learning from Sleep Data

Mentor: Siddharth Biswal⟨sbiswal7@gatech.edu⟩

We can analyze sleep data, especially electroencephalography (EEG), to determine the sleep apnea, cardiovascular disorders, and sleep stage annotations, etc. There are many benefits we can achieved from sleep data. For instance, using single channel sleep stage detector patients can be monitored at home using wearable sleep EEG devices.

- **Dataset**: Sleep Heart Health Study (you have to request data access), PhysioNet: The Sleep-EDF database [Expanded] (smaller, but immediately downloadable), etc.
- **Related Work**:
  - Tsinalis et al. [13] used convolutional neural networks (CNNs) for automatic sleep stage scoring based on single-channel electroencephalography (EEG) to learn task-specific filters for classification without using prior domain knowledge. Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks.
  - Biswal et al. [1] proposes a method to automatically annotate sleep stages from EEG data which is collected from overnight sleep studies. There are 5 different sleep stages N1,N2,N3,REM, Wake and usually trained clinicians annotate sleep EEG to identify these sleep stages. However, this is a very time consuming and labor intensive task. Therefore, the authors describe a system which uses expert defined features with recurrent neural network to annotate an entire sleep study. The results are presented to clinician using web based visualization which can be used to annotate mistakes made the model to further improve the results. They compared the result with other methods such as convolutional neural network, etc.
  - Zhao et al. [16] introduce a predictive model that combines convolutional and recurrent neural networks to extract sleep-specific subjectinvariant features from RF signals and capture the temporal progression of sleep by using radio measurements without any attached sensors on subjects. They applied a modified adversarial training regime that discards extraneous information specific to individuals or measurement conditions, while retaining all information relevant to the predictive task.

[2] https://physionet.org/physiobank/database/mimic3wdb/matched/

F. Other projects

You are welcome to propose your own projects as long as 1) they are health analytic projects and 2) they use big data tools covered in this class (Hadoop, Spark). Note that we may not able to provide much support on those projects. Also, please contact TA Sungtae An ⟨stan84@gatech.edu⟩ and Ming Liu ⟨mliu302@gatech.edu⟩first to verify validity of the topic before you dive into actual work.

V. Conclusion

Best of luck on your project and data science rocks!

Acknowledgment

Thanks all the TAs for their time and effort in creating the course material together. Thank all the students for their dedication and feedback.

References

[1] S. Biswal, J. Kulas, H. Sun, B. Goparaju, M. Brandon Westover, M. T. Bianchi, and J. Sun. SLEEPNET: Automated sleep staging system via deep learning. 26 July 2017.

[2] T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, D. Shimabukuro, U. Chettipally, M. D. Feldman, C. Barton, D. J. Wales, and R. Das. Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. *JMIR Med Inform*, 4(3):e28, 30 Sept. 2016.

[3] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits. Unfolding Physiological State: Mortality Modelling in Intensive Care Units. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 75–84, New York, NY, USA, 2014. ACM.

[4] H. Harutyunyan, H. Khachatrian, D. C. Kale, and A. Galstyan. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*, 22 Mar. 2017.

[5] Z. He, S. Carini, I. Sim, and C. Weng. Visual aggregate analysis of eligibility features of clinical trials. *Journal of biomedical informatics*, 54:241–255, 2015.

[6] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7(299):299ra122–299ra122, 2015.

[7] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Sci Data*, 3:160035, 24 May 2016.

[8] T. Kang, S. Zhang, Y. Tang, G. W. Hruby, A. Rusanov, N. Elhadad, and C. Weng. Eliie: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, 24(6):1062–1071, 2017.

[9] H. Ma and C. Weng. Identification of questionable exclusion criteria in mental disorder clinical trials using a medical encyclopedia. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 219–230. World Scientific, 2016.

[10] H. Ma and C. Weng. Prediction of black box warning by mining patterns of convergent focus shift in clinical trial study populations using linked public data. *Journal of biomedical informatics*, 60:132–144, 2016.

[11] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[12] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. *Critical Care Medicine*, 39(5):952–960, May 2011.

[13] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou. Automatic sleep stage scoring with single-channel eeg using convolutional neural networks. *arXiv preprint arXiv:1610.01683*, 2016.

[14] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471. IEEE, 2017.

[15] Y. Xu, S. Biswal, S. R. Deshpande, K. O. Maher, and J. Sun. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2565–2573. ACM, 2018.

[16] M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International Conference on Machine Learning*, pages 4100–4109, 2017.