CS 560/460: BIG DATA ENGINEERING
STUDENT PROJECTS & PRESENTATIONS

DATE ANNOUNCED:	**OCT 6, 2022**
PROJECT SELECTION DUE:	**OCT. 13, 2022**
INITIAL PRESENTATION (PROPOSAL):	**OCT 18, AND OCT 20, 2022**
PROGRESS REPORT:	**WEEK OF NOV. 7, 2022**
FINAL PRESENTATION:	**WEEK OF DEC. 5, 2022**
FINAL REPORT DUE:	**DEC 11, 2022, 11:59PM.**

Basic groping: Groups of 2 or 3 students each; (could be 2-, or 3-member groups**)**
There are several projects. Each group will select and work on one project of their choice. Grouping and choice of projects is to be done by the students on their own. Thus, you should choose your project topic as soon as possible. You are advised to select projects based on personal (or group) interest. No more than 2 groups on one given project.

For each group, assessment will be based mainly (**but not only**) on the assessment pointers described below. As a group, you could also suggest your own project if you have a relevant topic.

General Assessment Pointers
In general, you are required to use the project description (when provided), or the paper and its reported results **as a starting point** to develop your own methods. Thus, the first step could be to analyze the paper, and then implement and repeat the reported experiments (where necessary). **[Note: in some cases, it is not necessary to repeat the experiments in the reference paper].** Then, based on these you can design your own **new** algorithms and experiments.
The project will be evaluated as **a mini research project** on the following grounds:

- Performance of the basic function required using graphs, tables, complexity analysis, etc.
- Effect of different parameters in the algorithm or project
- Comparative performance with other methods (use tables, charts, complexity analysis, etc.)
- **Limitations of approach** (as in the relevant paper(s), or based on your own observations)
- **Creativity, Implemented improvements and other suggestions (where possible)**
- Minimal description on how to run/execute your programs or your simulations (where applicable).
- Elucidation on how your problem, data used, and approach fit the characteristics of Big Data, especially the V's of Big Data, and their connection to Big Data Engineering.
- Ability of the students to identify, search, and find on their own, the required Big Data tools, data sets, etc., as needed for their project**.**
- **Use of tools/techniques covered in class, high performance computing platforms, etc.**

Note that the list of considerations above is meant to be indicative only and not necessarily exhaustive. Thus, you do not need to consider each and every one of them in doing the assignment. Similarly, other ideas not necessarily included in the above list will be more than welcome.

For group work, it might be helpful to divide the work such that some (perhaps, with interest in programming) will concentrate on the implementation, and the others will concentrate on the more theoretical issues. Group members will receive the same mark for the project.

The papers referenced are available online, on the general net.

What to submit
1. A report on your project (similar to a technical paper), including details of how the above assessment pointers have been considered. The report should be **no more than 8 pages**, except for appendix, if needed. Descriptions beyond the 8th page will be ignored. You can follow the standard ACM format (http://www.acm.org/sigs/publications/proceedings-templates#aL2), or IEEE Transactions format.

2. Any program source codes and executables that you might have used.
3. Submit your reports (in .pdf) and accompanying codes through eCampus.

Project 0: Your own group's suggested project
If you have a project topic that is relevant to the course, you are free to suggest it, and we can discuss it.

Project 1: Stock Market Prediction via Social Media Interactions
Specific Problems: (1) Predicting general stock market direction and (2) stock market trading ranges, using social media data (for instance, using Twitter data, Google query trends, other social media sources, other big data sources, and/or their combination). (3) Doing the above on different time scales (hour, intra-day, weekly, monthly, yearly).

Relevant materials:
1. Kavyashree Ranawat, Stefano Giani, Artificial intelligence prediction of stock prices using social media, arXiv, 2021. https://arxiv.org/abs/2101.08986
2. Tobias Preis, Helen Susannah Moat and H. Eugene Stanley (2013). "Quantifying Trading Behavior in Financial Markets Using Google Trends". *Scientific Reports* **3**: 1684. doi:10.1038/srep01684.
3. Bollena, J, Maoa,H, Zeng, X (2011), "Twitter mood predicts the stock market", Journal of Computational Science 2, pp 1–8. http://personal.stevens.edu/~rchen/readings/stock.pdf
4. Philip Ball (26 April 2013). "Counting Google searches predicts market movements". *Nature*.
5. Bernhard Warner (25 April 2013). "'Big Data' Researchers Turn to Google to Beat the Markets". *Bloomberg Businessweek*.

Advisor: Prof. Adjeroh

Project 2: Health Risk Predictions via Social Media
Specific Problems: (1) Social media data collection; (2) Preprocessing, data cleaning, and natural language processing; (3) Health risk prediction via social media; (4) Integrating various available datasets, including county-level data on cardiovascular risk and information from social media platforms.

Relevant materials:
Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., ... & Weeg, C. (2015). Psychological language on Twitter predicts county-level heart disease mortality. Psychological science, 26(2), 159-169

Eichstaedt, johannes C., H. A. Schwartz, Salvatore Giorgi, Margaret L. Kern, Gregory Park, Maarten Sap, Darwin R. Labarthe, et al. 2018. "More Evidence That Twitter Language Predicts Heart Disease: A Response and Replication." PsyArXiv. March 15. doi:10.31234/osf.io/p75ku.

Advisor: Prof. Adjeroh

Project 3: Protein Functional Annotation via PageRank
Specific Problems: (1) characterizing uncharacterized proteins; (2) protein network analysis; (3) Integrating various available omic datasets for new insights on protein function annotation.

Relevant material:
José González-Gomariz, Guillermo Serrano, Carlos M. Tilve-Álvarez, Fernando J. Corrales, Elizabeth Guruceaga, and Victor Segura. UPEFinder: A bioinformatic tool for the study of uncharacterized proteins based on gene expression correlation and the PageRank algorithm. J. Proteome Research, 2020 19 (12), 4795-4807. DOI: 10.1021/acs.jproteome.0c00364

Advisor: Prof. Adjeroh

Project 4: Cardiovascular Image Analysis
Specific Problems: (1) Analyzing echocardiogram– describing key attributes based on observable features of the heart; (2) Large variability in cardiovascular images, image quality, image resolution, etc.; (3) Connecting observed image features to cardiovascular diseases, disease stage/severity, etc.;

Relevant material:
Madani A, Arnaout R, Mofrad M, and Arnaout R, "Fast and accurate view classification of echocardiograms using deep learning", npj Digital Medicine (2018) 1:6 ; doi:10.1038/s41746-017-0013-1.

Advisors: Profs. Adjeroh & Doretto

Project 5: Person Re-Identification
It is not unusual to walk around public places and to be recorded by a network of cameras. Video analytics algorithms are now being developed that attempt to understand our behavior by leveraging video data being recorded. In order to do so, one of the problems that needs to be solved is to "link" the presence of a person under a camera view with the presence of the same person under a different camera view and usually at a different time. This is the person re-identification problem. The basic task to be addressed has to answer the question of whether the whole-body appearance of two people is close enough to believe that they are the same person.

Person re-identification is the subject of intense research, and now there are works that claim better that human performance, like this one:

"**AlignedReID: Surpassing Human-Level Performance in Person Re-Identification**"
https://arxiv.org/abs/1711.08184

This is a simple description of how it can be used:
https://medium.com/@niruhan/a-practical-guide-to-person-re-identification-using-alignedreid-7683222da644

and this is a third-party implementation:
https://github.com/huanghoujing/AlignedReID-Re-Production-Pytorch

The following is a common benchmark dataset, Market-1501, which you can download
https://drive.google.com/drive/folders/1CaWH7_csm9aDyTVgjs7_3dlZIWqoBlv4

Goal 1: Your first goal is to adapt and run the implementation of the AlignedReID algorithm, and you need to reproduce the results on Market-1501 reported in the original paper.

Implementation: Your implementation should be based on the following library: PyTorch https://pytorch.org. In addition, your implementation has to leverage and run on the following online computing environment provided by Google, called Google Colab (which also integrates with your MIX Google Suite): https://colab.research.google.com/notebooks/welcome.ipynb

You will need to produce and deliver a Colab notebook similar to a Jupyter notebook which runs and produces results about AlignedReID.

Goal 2: If your project group is made by more than one person, you should compare the results of your Collaboratory-PyTorch implementation of AlignedReID with the Collaboratory-PyTorch implementation, that you will create, of the following approach

"Deep Cosine Metric Learning for Person Re-Identification"
https://elib.dlr.de/116408/1/WACV2018.pdf
which also has already a public implementation provided here
https://github.com/nwojke/cosine_metric_learning

Note that in order to perform the comparison you will need to follow either the protocols used in the first paper, or the protocols used in the second paper.

Again, you will need to produce and deliver a Colab notebook which runs the algorithms and generates the results.

Bonus points: Extend the evaluation of the algorithms above to other public benchmark datasets found in any of the two papers above.

Advisor: Prof. Gianfranco Doretto

Project 6: Gaze Estimation

Gaze estimation is the problem of estimating the viewing direction of a person. We will focus of the version of the problem that entails the use of eye images to perform such estimation. This basic technology is the building block for eye trackers of various types that usually serve as sensing devices to enable the inference of more complex behavior of the person being observed. For instance, the ability to understand where people look when they go shopping allows to better present or place products in points of sale. It has been shown that accurate "product placement" has the potential to significantly increase product sales. Several other applications of this technology often have to do with the analysis of certain specific behavior of people. The basic task to be addressed is: given an image of an eye, provide the direction of sight of that eye with respect to the camera reference frame that has taken the image.

Despite its usefulness, and the presence of companies selling eye tracking products, eye tracking is still an active area of research. This work provides a relevant example:

"Eye Tracking for Everyone'
https://arxiv.org/abs/1606.05814

From this website you will be able to download a training dataset and there is also a link to a public implementation: https://gazecapture.csail.mit.edu/ and https://github.com/CSAILVision/GazeCapture

Goal 1: Your first goal is to reimplement the iTracker algorithm and to reproduce the cross-dataset generalization results between the GazeCapture and the TabletGaze datasets reported in the original paper.

Implementation: Your implementation should be based on the following library: PyTorch https://pytorch.org. In addition, your implementation has to leverage and run on the following online computing environment called Google Colab (which also integrates with your MIX Google Suite): https://colab.research.google.com/notebooks/welcome.ipynb

You will need to turn in a Colab notebook that executes the code and produces the results.

Goal 2: If your project group is made by more than one person, you should compare the results of your Collaboratory-PyTorch implementation of iTracker with the Collaboratory-PyTorch implementation, that you will create, of the MPIIGaze tracker, described here

"Appearance-Based Gaze Estimation in the Wild" https://arxiv.org/pdf/1504.02863.pdf , also

https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/gaze-based-human-computer-interaction/appearance-based-gaze-estimation-in-the-wild/

Note that in order to perform the comparison you will need to follow the protocol used in the first paper (results in Table 4). You will need to turn in a Colab notebook that executes the code and produces the results.

Bonus points: Extend the cross-domain evaluation of the algorithms above by replacing the TabletGaze benchmark datasets with the MPIIGaze dataset.

Advisor: Prof. Gianfranco Doretto.

Project 7: Big Data Analysis of Weather and Climate:
Consider the issue of weather and environmental conditions around the world especially over time. Things like temperatures, sea, level rise, etc. Develop a model for predicting storm intensity – hurricane strength, average wind speed, rainfall amount, sea levels in coastal areas, etc. Maybe extend this to predicting magnitude of damage. This project could focus the on a specific region. Data for this project are generally available from various publicly available governmental and non-governmental datasets. This is a timely project since, for example, the number of named (major) storms to hit the United States this year has been so large that weather forecasters have exhausted their list of names for these storms and are now well into the Greek alphabet.

On a related note, iIt would be interesting to explore the apparent increase in the number and intensity of severe weather events particularly in light of predicted trends in global warming. Maybe playing out the consequences of certain interventions like (I believe) California says no more gas powered cars to be sold after 2030. What will that mean?

Advisor: Prof. McLaughlin

Projects 8: Social Computing -- Social Media Analytics of Current topics
Imagine using social media to analyze some current topic or trend. Perform analyses to show how the topic or issue has developed over some period of time. Perform analyses such as sentiment analyses to learn how the social media sphere "feels" about the topic. Also perform analyses to break down social media data geographically, demographically, by source and other segmentations. Construct a presentation (dashboard) of the results that shows what is learned from the analyses.

For this project, students will design a system to measure public sentiment on a given topic of widespread discussion (e.g., 2020 presidential election, Black Lives Matter, COVID-19). Students will analyze data from a social media service such as Facebook or Twitter. The data can be sourced from the social media service's API, publicly available datasets, or a combination of both.

The project must be able to broadly categorize social media posts, and the users posting them, into groups for, against, and neutral on the topic. By analyzing the aggregate postings on the topics, new versus shared posts, the volume of postings by users, and any other measurements the students believe appropriate, the project should quantify overall public sentiment on the topic and identify keywords, post topics, shared articles, or other identifiers indicative of each class of sentiment.

Advisors: Profs. McLaughlin & Powell

Project 9: Data Visualization Project
For this project, students will design a series of interactive data visualizations to allow users to explore socio-economic trends over a geographic area. This project will make use of multiple types of visualizations (maps, charts) to illustrate at least 5 different measures. Users must be able to seamlessly explore and switch between different data measures. The visualizations can be implemented in a platform of the students' choice.

One possible implementation of this project would be to explore the root causes, growth, and impact of the drug epidemic in West Virginia. Using data covering many years, users would be able to explore county-level data on population, economic conditions, life expectancy rates, crime rates, drug use and drug overdose rates for the entire state. We have data available for this project. Students are welcome to propose other subjects to explore through the use of data visualizations.

Advisor: Prof. Powell

Project 10: Modeling the geographic and demographic spread of COVID-19 in the United States
The tragic COVID-19 pandemic continues to unfold around us in the United States. A casual observation of the infection maps (like https://coronavirus.jhu.edu/map.html) suggests patterns of infection spread. For example, at its onset in the US, the concentration of infection was centered on urban and coastal areas. In fact, the spreads in multiple locations appears to look like waves in a pond. This project would have two broad goals – to map and display the spread of the virus at the county level from its onset to the current time, and to model the continuing spread of the disease based on what is learned in the historical data.

There are many variations of this issue that could, and perhaps should, be explored. For example, what have the geographic spread patterns looked like for different age groups, ethnic groups, and economic status. Where have the "pebbles in the pond" been with respect to the spread of the virus – major urban areas, major airports, transportation corridors, etc.? And how has population and population density impacted the rate and extent of the spread.

There are numerous data resources available for the study of the COVID-19 virus. The first task of the project would be to define specific goals and methods, Then, project investigators will need to explore and select appropriate datasets to carry out this research effort.

Other potential COVID-19 related topics
(Students will work with the instructors to identify available datasets for these, as needed).

Predicting hospitalizations and deaths due to COVID based on local vaccine hesitancy rates. This might include using demographics, locality and other variables as predictors. It would be interesting to see where hospitals are full and other medical needs are not being met or curtailed due to the lack of hospital beds, ICUs, etc.

Predicting the trajectory of COVID infections, related deaths and hospitalizations based on regions, vaccinations, infection rates, etc. In other words, when might the pandemic start to trail off and will this be different based on different local characteristics.

Explore possible connections between different types of social media and COVID recovery, vaccinations, etc.

The COVID pandemic has taken the attention away from the opioid crisis. I don't think it is any better. I am not sure though. It might be interesting to explore the prevalence of opioid additions, overdoses in relation to things like COVID infections, vaccination rates, unemployment rates, local economic indicators, etc.

Advisor: Prof. McLaughlin & Prof Powell

Project 11: Domain Adaptation for Plants in Herbarium Collections.

From the PlantCLEF 2020 challenge: https://www.imageclef.org/PlantCLEF2020 :

"*For several centuries, botanists have collected, catalogued and systematically stored plant specimens in herbaria. These physical specimens are used to study the variability of species, their phylogenetic relationship, their evolution, or phenological trends. One of the key step in the workflow of botanists and taxonomists is to find the herbarium sheets that correspond to a new specimen observed in the field. This task requires a high level of expertise and can be very tedious. Developing automated tools to facilitate this work is thus of crucial importance. More generally, this will help to convert these invaluable centuries-old materials into FAIR data.*"

Since the image domain of herbaria images used to train species classifiers is different from the image domain of new specimen observed in the field, the classification accuracy drops significantly. The problem can be addressed by leveraging so called ***domain adaptation*** techniques. Based on the PlantCLEFF 2020 challenge report http://ceur-ws.org/Vol-2696/paper_140.pdf among the methods that stood out, there was a domain adaptation approach developed here at WVU [Motiian et al. NIPS 2017]. It was used as described by the France and Costa Rica team in this report http://ceur-ws.org/Vol-2696/paper_141.pdf

Your first goal is to deliver a PyTorch implementation that reproduces the methodology and main results in the France and Costa Rica report.

Your second goal is to experiment with techniques that were not included in the initial report. For instance, what happens if you try to use a different pretext task for self-supervised learning, or what happens if you use a contrastive loss to do a pre-training of the network, or perform a comparison between different network architectures for the backbone network. Here you have the freedom to investigate and try to experiment or propose any new variation of the original methodology. You will then explain what you were able to learn from this experience.

[Motiian et al. NIPS 2017] Motiian, S., Jones, Q., Iranmanesh, S. M., and Doretto, G., Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. https://dl.acm.org/doi/10.5555/3295222.3295412

Advisor: Prof. Gianfranco Doretto.