FALL 2020

DEPAUL CENTER FOR DATA SCIENCE

IN THIS ISSUE

DePaul | COLLEGE OF COMPUTING AND DIGITAL MEDIA

Dear Students and Faculty,
Welcome to the new academic year 2020-2021!

Although this was for sure a very unusual summer, I hope you took some time off to relax and recharge your batteries to start with vigor a new academic year!

In this Fall issue of our Center Newsletter, you will learn about our Data Science faculty and students' accomplishments, new projects including COVID-19 related research work, new courses that you can enroll as part of the Data Science Program, a collection of COVID-19 datasets that was compiled by the Center, and an interview with the Associate Director for Center Industry Relations, Dr. Ilyas Ustun, focusing on how you can get involved in research and capstone project activities.

As most of our classes will be online, the DePaul Center for Data Science will continuously serve as a bridge between students, faculty, and the industry. We will keep sharing job opportunities, research opportunities as well as industry news to help you better prepare for careers in the Data Science field. You might also find helpful "The 5 tips for succeeding as an online student", an article that offers useful tips for adjusting to e-learning in the new quarter.

DePaul faculty will always be here for you to guide you along the way. I am confident that together we will find new ways to learn and advance knowledge in our Data Science field.

Best wishes for a successful new Academic Year!
Daniela Raicu, PhD
Director of the DePaul Center for Data Science

Faculty News

- Dr. Bamshad Mobasher received a $24,834 DePaul Academic Growth and Innovation (AGIF) award for his project "Using Machine Learning to Battle the Infodemic of Misinformation in the Age of COVID-19". The AGIF award was established in 2018 with a yearly budget of $2M to find, encourage, and support innovative academic ideas that will positively impact DePaul revenue within the next 3 years.

- Dr. Daniela Stan Raicu served as the keynote speaker at the 2020 IEEE International Symposium on Computer Based Medical Systems (CBMS) in July, 2020, where she delivered a presentation on the topic "Deep Learning for Explainable Computer-Aided Diagnosis".
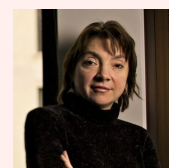
New Research Lab

We are excited to announce the inception of the Computational Biology and Applied Bioinformatics (CoBaAB) laboratory under the leadership of Drs. Ramaraj & Rogers, SoC, CDM. CoBaAB lab will primarily focus on research problems associated with large-scale biological sequence (Nucleic & Amino acid) data. Some of the ongoing research projects include (i) Phylogenomics, (ii) Pangenomics, (iii) Biofilm data analysis and (iv) Machine & Deep Learning approaches towards functional annotation of sequence data. CoBaAB lab will be affiliated with DePaul Center for Data Science (CDS). Besides current ongoing research, the lab will also promote and coordinate interdisciplinary research activities within other research labs at CDS like VIDA and MedIX labs and also relevant units/schools at DePaul University.

Current Lab Members – Drs. Ramaraj & Rogers (PIs); Students: DeAngelo Wilson, Zonglin Yang and Bo Wen Liu (Graduate); Lilian Cornell (Undergraduate); Peter Gandy (SoC Alumnus & Prospective PhD Student).

Students and collaborators interested in being part of CoBaAB lab please contact Dr. Ramaraj at tramaraj@depaul.edu or Dr. Rogers at john.rogers@depaul.edu

New Publications

- Tan T., Montague E., Furst J., Raicu D. "Robust Physician Gaze Prediction Using a Deep Learning Approach". The 20th IEEE International Conference on BioInformatics And BioEngineering (BIBE 2020), Virtual Conference, October 26-28, 2020.

- Shang M., Furst J., Raicu D. "Assessment of medical reports uncertainty through topic modeling and machine learning". 2020 IEEE International Symposium on Computer Based Medical Systems (CBMS), Rochester, Minnesota, July 28-30, 2020.

- Tan T., Montague E., Furst J., Raicu D. "Developing parameters for a technology to predict patient satisfaction in naturalistic clinical encounters". In: Duffy V. (eds) Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Posture, Motion and Health. International Conference on Human Computer Interaction (HCII 2020), Lecture Notes in Computer Science, vol 12198. Springer, Cham. https://doi.org/10.1007/978-3-030-49904-4_35.

- Govindaswamy A., Farooq W., Wang Y., Ustun I., Raicu D., Furst J., Kim H. "Measurement of similarity in C. elegans healthspan using dynamic time warping on movement features". 12th International Conference on Bioinformatics and Computational Biology (BICOB), San Francisco, CA, March 23-25, 2020.

1. DSC 510: Health Data Science

Mondays, 5:45PM-9:00PM & Online
Description: The course will focus on data science methods used in clinical studies and public health applications. Students will be introduced to a variety of health care data from electronic health records to payer data, geospatial and unstructured data, and will learn how to solve data science problems in the health sector. Topics include overview of healthcare analytics and typical research questions, epidemiology, data ethics, governance and security, applications of modeling techniques and machine learning methods to a variety of case studies in health care. Prerequisite DSC 441 Fundamentals of Data Science.
This course is part of the revised concentration in Health care and will count as an elective for all other concentrations.

2. CSC 521: Monte Carlo Algorithms

Wednesday, 5:45PM-9:00PM & Online
Description: A course about the use of random numbers for numerical computation with particular emphasis on implementation issues and applications in science and finance. Covered topics include: pseudo random number generators, the inversion method, the accept-reject method, discrete event simulations, multi-dimensional integration, the Metropolis and the Bootstrap algorithms.
Prerequisite(s): (CSC 402 or CSC 404) and DSC 423 or consent of instructor.
Course requires strong programming skills contact your faculty advisor if you are interested in enrolling.

3. MAT 427: Bayesian Statistics

Thursdays, 6:00PM-9:15PM & Online
Description: Comparison of Bayesian and frequentist methods, conditional probability, Bayes theorem, conjugate distributions, computational methods, hands-on Bayesian data analysis using appropriate software, interpretation and presentation of analysis results. Students will learn to use software packages including OpenBUGS. The free software program R will be utilized for data analysis. MAT 349 or MAT 351 or MAT 451 is a prerequisite for this class.
Contact Prof. Phil Yates (pyates@depaul.edu) for questions. This class can count as an elective.

4. FIN798 Special Topics course on Blockchain (Graduate School of Business)

Wednesday, 6:00PM-9:00PM & Online
Description: This course is for students who want to learn more about the full scope of blockchain applications. In this course, you will learn about the uses of blockchain as well as the technical fundamentals. DePaul University has partnered with Deloitte's Blockchain Lab to help students understand the current frontier of real-world blockchain use cases. Student teams comprised of business and CDM students will work on cases that simulate real-world blockchain engagements as students learn how to use blockchain to help clients meet challenging objectives.
CDM students should contact their faculty advisor to get this course approved as an elective in their program. Contact Prof. Lamont Black (lblack6@depaul.edu) for questions.

CDPH COVID-19 PROJECT

Project Description

Many people from racial and ethnic minority groups are at a higher risk of getting sick from COVID-19. To understand why minority groups are being disproportionately affected by COVID-19 and to find possible solutions of racial health disparities, the Chicago Department of Public Health (CDPH) began building a COVID-19 surveillance dataset in January 2020. Similar with datasets in other medical studies and surveys, CDPH COVID-19 Data has a significant amount of unreported race/ethnicity (RE) information. Among a total of 256,483 records (Jan 5 - Jun 18), there are 121,124 records (47.22%) with unreported RE information, which can impact the research and understanding of COVID-19 impact on different RE groups. CDPH and DePaul have partnered to look into ways for data imputation, modeling, and evaluation.

Project Results

Using a statistical approach, the Bayesian Improved Surname Geocoding (BISG) model, the CDPH and DePaul team successfully imputed 76% of missing RE values in CDPH COVID-19 Data with 81% accuracy. The results helped researchers have a better understanding of racial equity within the COVID-19 epidemic in Chicago.

CDPH COVID-19 Data Racial Proportion Before and After Imputation
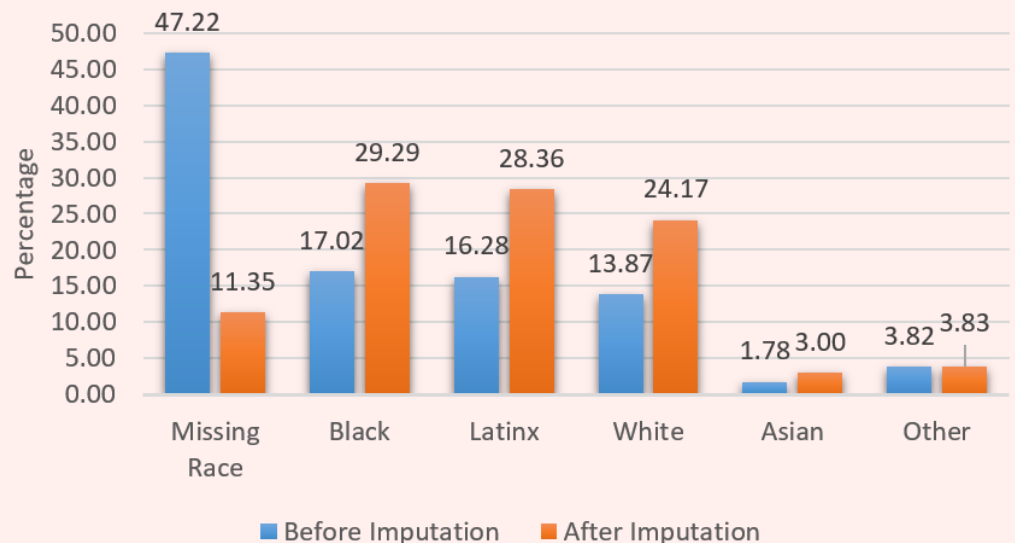
DePaul Team Members
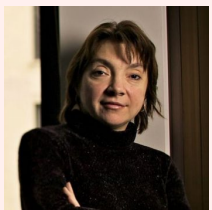
Dr. Daniela Raicu
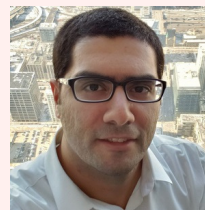College of Computing and Digital Media

Dr. Fernando De Maio
College of Liberal Arts & Social Sciences

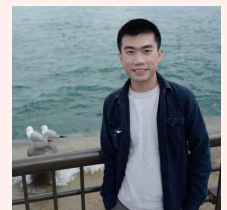**Dr. C. Scott Smith**
College of Liberal Arts & Social Sciences

Dr. Jacob Furst
College of Computing and Digital Media

Dr. Ilyas Ustun
College of Computing and Digital Media

Hao Wu
College of Computing and Digital Media

Yiyang (Ian) Wang
College of Computing and Digital Media

COVID-19 DATASETS

1. <u>COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University</u>

This is the data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Also, Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL).

2. <u>COVID-19 Open Research Dataset Challenge</u>

In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 200,000 scholarly articles, including over 100,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in new coronavirus literature, making it difficult for the medical research community to keep up.

3. <u>COVID-19 Global Forecasting</u>

Kaggle is launching a companion COVID-19 forecasting challenges to help answer a subset of the NASEM/WHO questions. While the challenge involves developing quantile estimates intervals for confirmed cases and fatalities between May 12 and June 7 by region, the primary goal isn't only to produce accurate forecasts. It's also to identify factors that appear to impact the transmission rate of COVID-19.

4. <u>COVID-19 image data collection</u>

A public open dataset of chest X-ray and CT images of patients which are positive or suspected of COVID-19 or other viral and bacterial pneumonias (MERS, SARS, and ARDS.). Data will be collected from public sources as well as through indirect collection from hospitals and physicians. All images and data will be released publicly in the GitHub repo.

ALL DATASETS CAN BE FOUND ON THE CENTER WEBSITE UNDER <u>RESOURCES/DATASETS</u>

Ilyas Ustun, PhD

Associate Director for Industry Relations, DePaul Center for Data Science

Ilyas earned his Bachelor's in Industrial Engineering with summa cum laude. He earned his Ph.D. with a focus on transportation analytics and worked as a data scientist in digital agriculture. He has taught data science courses at the University of California Berkeley and the University of Virginia and mentored students at Thinkful. He published an online video course titled "Learning Python for Data Science". Currently, he is a faculty in Data Science at DePaul University in Chicago. His research interests include data mining in the areas of business analytics, transportation analytics, health analytics, and deep learning.

Q1: What are your responsibilities serving as the Associate Director of the Center?

As the Associate Director of the Center for Data Science my responsibilities include a wide array of topics from initiating data science projects that will lead to Capstone and/or research papers, collaborating with the industry, city or government, and research institutes so that our students can take part in joint projects, to establishing internships with these institutes so that our students can get real world data science experience. I am also responsible for teaching the Data Science Capstone class in which students take part in a project coming up with a data science solution to a challenging problem. I am very passionate about data science and always look out for new projects that can be of high interest for students.

Q2: How will the Center better serve DS students in this special year dealing with a pandemic?

Switching everything to online has been a challenge for both the students and the faculty. Many internships for the summer have been canceled. However, we offered the Capstone Data Science course and engaged students to work on new projects including one of my new research on the effects of Covid19 on transportation networks. In this project, students have had the opportunity to develop and implement visual and statistical analysis of transportation related data. I have several more projects waiting in the pipeline. I am also in talks with Argonne National Lab to start a project and a possible internship in the field of supply chain and business analytics. I have made a lot of new connections with professionals doing data science in the industry and plan to invite them for giving virtual talks.

Q3: How can students get more involved with the Center's research projects in the new school year?

The students should be receiving an email when a new project or internship is established. Regardless, my advice is for them to check the Center for Data Science website (cds.cdm.depaul.edu) and look for projects with different professors. There are many projects and the students should be able to find something of interest to them. We are also very open to new ideas and projects, and as such if a student has an idea or project that is data science or machine learning related, they should not hesitate to contact me at iustun@depaul.edu.

EDUCATIONAL MISSION

To nurture the growth of the next generation of data scientists and computer scientists to better prepare them for data-related computing careers. Students work on state-of-the-art research and practice activities under the supervision of faculty members.

CONTACT US

✉ dsc@depaul.edu

🌐 http://cds.cdm.depaul.edu/

🐦 https://twitter.com/DePaulCDS

🏠 243 S Wabash Ave, Chicago, IL 60604