Sacred Heart University
DigitalCommons@SHU

WCBT Working Papers | Jack Welch College of Business & Technology

2023

Thinking Local with Original Data In AI and Machine Learning Research

David G. Taylor
Sacred Heart University

Robert McCloud
Sacred Heart University

Follow this and additional works at: https://digitalcommons.sacredheart.edu/wcob_wp

 Part of the Artificial Intelligence and Robotics Commons, Business Analytics Commons, and the Data Science Commons

Recommended Citation

Taylor, D. G., & McCloud, R. (2023). *Thinking local with original data In AI and machine learning research* [White paper]. https://digitalcommons.sacredheart.edu/wcob_wp/31/

This White Paper is brought to you for free and open access by the Jack Welch College of Business & Technology at DigitalCommons@SHU. It has been accepted for inclusion in WCBT Working Papers by an authorized administrator of DigitalCommons@SHU. For more information, please contact lysobeyb@sacredheart.edu.

JACK WELCH COLLEGE
OF BUSINESS & TECHNOLOGY
Sacred Heart University

Thinking Local with Original Data In AI and Machine Learning Research

By David G. Taylor, Dean and Robert McCloud, Associate Dean
Welch College of Business & Technology, Sacred Heart University

Executive Summary

Sacred Heart University spent significant funds to establish an AI lab. Initially there is no ongoing research and no real plan for a research agenda. This paper details how the Jack Welch College of Business and Technology created and implemented an active meaningful research plan. It involves two key elements: thinking local and using business connections to foster active, impactful research. Surrounding communities, business connections, area environment, and other Sacred Heart University departments all played a part. The research plan also identifies a specific issue in working with local and business contact sources: the AI researcher almost never gets data that is ready to use. Typically, there are missing or mistaken data points. While one enters a research project thinking of structure, algorithms and potential results, the reality is that a substantial amount of time will be devoted to cleaning data. For business students is an important lesson: structure your AI input so that the results have meaning.

Background

In 2017 the University merged its School of Computing Science and Engineering with the Jack Welch School of Business. The resulting entity was rechristened the Jack Welch College of Business and Technology (WCBT). A critical part of WCBT's mission is to prepare business students to be technologically literate and experienced as they begin their careers. Shortly after the merger decision, Sacred Heart purchased the former General Electric headquarters complex in Fairfield, Connecticut. Moving to this new campus presented an opportunity:

build a group of laboratories that would offer our students significant educational and research possibilities. As the GE headquarters transitioned into an academic campus, several new laboratories emerged: Artificial Intelligence; Engineering; Virtual Reality/Augmented Reality; Motion Capture; and Cybersecurity. All WCBT students and faculty can work together in these labs. For this paper we focus on the AI lab and its active projects.

Problem

Sacred Heart University is a strong teaching institution with a typical course load of 3-3. A successful AI research program understands that teaching and research are closely aligned. Recognizing that we do not have R1 resources, but are working toward R2 status, WCBT needed to create a meaningful AI research program. Could we do research that matters? How do we best incorporate students as partners in that research?

Solutions

Our solution was to look local and inward. We want to do our own research, create curated databases, and offer students and faculty programming choices.

We tap into our student expertise. Many of our undergraduate business students have become adept at programming. They recognize that programming, data analysis and solving business problems go hand in hand. Nurturing that attitude brings a student population that is happy to do AI research with their professors.

WCBT also has a substantial international graduate student(Masters level) population. Many of these students come to us with solid work experience. They are also up-to-date on industrial database trends. For example, they use mongoDB, a source-available database. Adapting MongoDB for some solutions helped us deal with the fact that students come equipped with different hardware across operating systems. mongoDB is compatible with Windows Vista and later, Linux, and OSX10.7 and later. Thus, students could work either on their own equipment or using command-line programming in our lab. This second option was attractive because it helped promote our Red Hat certification program.

Our solution orientation led us to implement several successful AI and machine learning initiatives:

Student-Professional Mentoring: A member

of WCBT's business team, Richard Robustelli, participates in an professional organization known as TechPACT. TechPACT consists of successful computer professionals. Among the organizations who participate are: Google, AWS, Cummins, L'Oréal, J.P. Morgan, MetLife, Estee Lauder, Deloitte, Ralph Lauren, Novant Health, and Becton, Dickinson and Company (BD): all substantial corporations dedicated to improving equity in computer fields. TechPACT's goal is "…to reduce the digital divide and pursue representative diversity in technology across all levels." The organization selected WCBT to develop an AI-driven process to match mentors with young minority and female university students who seek guidance in entering computer fields. The AI-driven software development is a joint effort by WCBT faculty and students which involves taking input from the Internet, structuring the data, and developing algorithms for obtaining the best matches. The second project phase will use machine learning to study qualitative input and resumes. Sacred Heart is leading the development effort in collaboration with Boston and Fordham Universities. Over time we will implement our continuous improvement focus to study what makes a good mentor-mentee match. Python and mongoDB are the primary development softwares.

Tracking Horseshoe Crabs: This local initiative, project Limulus takes advantage of Sacred Heart's location close to Long Island Sound. It came to WCBT through the University's Biology Department. The goal is to use data collected by students, create a curated database and to analyze the date using code developed in the R language. Using AI, students and professors create statistical data charts and maps. This information helps track horseshoe crab movement throughout Long Island Sound. An interesting aspect of the program is the inclusion of area high school students. They tag the horseshoe crabs and help locate them as they reappear on different beaches. Data analysis results are uploaded to a server maintained by the International Union for Conservation of Nature, Horseshoe Crab Species Specialist Group. Through data gathering, analysis and sharing, WCBT professors and students collaborate with biology professors and students and area high school students. Project Limulus merges local experience with an international AI study. As their first

exposure to AI driven studies, students learned the difficulty in collecting accurate information for their datasets. According to WCBT professor Samah Senbel, data collection problems included: horseshoe crabs with no recorded gender, found at the wrong location, or two different sizes recorded for the same crab. She commented, “That is my biggest issue in my research. I never get data ready to go, they all come as a mess and need weeks or months of cleaning!” Process turns into benefit: students learn the difficulty and time intense nature of solid research. This provides them with valuable experience for analyzing real world problems and studies.

Improving Athletic Performance: Sports data analytics is an emerging field for AI studies. With the participation of our Division 1 women’s basketball team a WCBT faculty and students partnered with colleagues from the Sacred Heart Exercise Science Department. Their goal was to assess the impact of sleep and learning on athletic performance. Data was collected by having players wear heart rate measuring devices.

As described by WCBT professor Tolga Kaya, “This holistic approach includes multi-modal data from athletes including training, performance testing, questionnaires, game performance, and sleep metrics. We use two wearable devices to collect these data and utilize machine learning techniques to predict athlete’s readiness, performance, and injury risks.” Undergraduate and graduate students participated. The WCBT study team worked closely with Sacred Heart Exercise Science faculty and students, as well as researchers from New York Institute of Technology and Ahmedabad University, India.

Using machine algorithms, the investigators showed they could predict game performance and injuries with over 90% accuracy. Data imputation, an interpretable feature set, data balancing and classifiers were all used to build a reliable database. Human subjects are not always invested in research. For example, a basketball player might forget to wear her heart rate measuring device. With missing data, the researchers are forced to use imputation methods. This is a common problem in creating your own data. However, having your own original, real data makes for more publishable research. For leading journals this is preferable to downloading ready-made, clean datasets.

Planting Mature Trees: The Town of Fairfield, Connecticut has the following announcement on its web site. “For $200, a member of the Fairfield Forestry Committee will walk you through the process of selecting a tree appropriate for your site and the Town will plant it for you in the public right-of-way along your property line…To achieve the full benefits of a tree, it must be selected and planted according to a well thought-out plan suitable for the planting site and surroundings…”

While the program intentions were laudatory, implementation caused a problem. There was no way to track individual trees and to use that data to predict what type of tree would thrive in a given environment. The Fairfield Forestry Committee turned to WCBT for help. The solution was to develop a dataset and to use AI to determine what tree would be recommended for the varied locations throughout the Town.

Alyssa Dunn began work on the trees project as an undergraduate. She described the AI dataset situation, “In this project, it was essential to ensure the data was recorded correctly and properly because we were looking for particular details about the trees to help solve why some trees were dying and others were not. When I first received the data, there was no clear start or end consistent with the inputted data. I was able to use MySQL to organize the data into clear and concise columns and rows. Once the data was readable, I was able to clean up, in other words, split specific columns and delete unnecessary columns, the data, and make it usable. After it was clean and readable I was able to transport it to other platforms and languages to visualize the data to help solve the problem.”

Through hands-on experience Alyssa learned that AI can only be successful when you pay careful attention to data cleansing. Because the Fairfield tree dataset was small (about 500 trees), it was not desirable to throw away individual entries. But some had to be discarded because the researchers discovered double entries, such as two trees being given the same location. One solution was to return to the data collectors to resolve ambiguity. As a result of working with the Fairfield citizens, WCBT researchers gradually improved the data quality. Continuous improvement is the watchword as the tree project enters its second year. It is a solid example of finding local sources for

meaningful AI research.

Breast Cancer Detection: Sometimes we tap into existing global databases. A WCBT graduate student, Genevieve Gish Alouche, proposed a research project to study ways of detecting breast cancer through analyzing mammograms. In this case a database was available for download from the Radiological Society of North America. Our lab computers have 68 terabytes of storage, so importing even a sizeable database is not a problem.

According to Genevieve's proposal, "Breast cancer is the most diagnosed cancer and is the leading cause of cancer death among women worldwide. Preventative measures are increasingly used to detect cancer at an early stage. A primary detection tool is annual mammograms starting at the age of 40." This student led research aims to accurately identify breast cancer in a dataset provided by the Radiological Society of North America. This dataset contains thousands of mammograms along with other attributes, such as whether the individual has implants or had a biopsy. The research project involves cleaning and normalizing the data, processing images, creating machine learning models, and assessing these models with their KPIs (accuracy, precision, recall, ROC, AUC, and f1-score). Python and its associated packages are the primary tools.

Conclusion

Building a successful AI student-faculty business school research program benefits from careful project definition. Know your capabilities. Don't promise to compete outside your capabilities. There are numerous possibilities for AI research. Bringing the business faculty and students into AI research can be difficult. But the potential learning results, for both students and faculty, can be enormous. Here are our own lessons learned:

1. Find areas where your institution can bring local and area expertise. A business school benefits in many ways from connections in the local community. AI research offers one more way to connect.
2. Define each project's scope. In this time of available dataset downloads it is tempting to take on large-scale, dramatic projects. Using pre-curated, public databases means eliminating an important part of the AI research process.
3. An important process is dataset structure and cleansing. By creating your own datasets, you have the opportunity to teach

your students valuable research lessons. This experience also helps in publishing your research in quality journals.

4. There are curriculum opportunities for our business undergraduates:

-- Teach database design, cleansing and normalization. Let our students know about the shift from relational to object data. But also teach them that relational databases are sometimes preferable in the case of legacy datasets.

-- Our students should be proficient in programming basics. The most obvious teaching opportunity is Python. At WCBT we prefer the combination of Python and mongoDB. We recognize that there are other choices just as valid. Whatever choice you make, sending students in the world understanding the building blocks of AI gives them a competitive career advantage.

5. It helps to have a strong IT person on the lab development and operational team. Understanding your university's IT structure and utilizing its support keeps AI efforts running smoothly. At WCBT we hired an IT staff member as Visiting Instructor.

When AI learning requirements are implemented, you might experience student pushback. The thought of doing *any* programming is unappealing to some business students. Yet, if we send our students into the world without this basic knowledge, we fail in our educational responsibility. The WCBT examples, we believe, show how an AI curriculum can tie in with local involvement, service, mission, collaboration, and fun.