THE UNIVERSITY OF BRITISH COLUMBIA

Department of Computer Science, Mathematics, Physics and Statistics
Okanagan Campus

CMPS Undergraduate Research Conference

April 2025

General Information

- Date: Friday, April 25, 2025
- Time: 8:45 am – 6:00 pm
- Discipline: COSC, MATH, PHYS, and STAT/DATA
- Location:
  - ⋆ COSC: SCI 333 and SCI 236
  - ⋆ MATH/PHYS/STAT/DATA: SCI 337
- Organizers: Alex Hill, Ifeoma Adaji, Paul Tsopméné, and Yves Lucet
- Contact:
  - ⋆ COSC: Yves Lucet (yves.lucet@ubc.ca)
  - ⋆ MATH/PHYS/STAT/DATA: Paul Tsopméné (paul.tsopmene@ubc.ca)
- Website: https://cmps.ok.ubc.ca/undergraduate/conference/

Contents

1 Computer Science – Parallel Session I

1.1 Names of Speakers

1. Noah Geerts
2. Karel Harjono
3. Soren Stenback
4. MacKenzie Richards
5. Komal Singh
6. Murad Shahmammadli
7. Jerry Fan
8. Samantha Hodge
9. Oluwadabira Omotoso
10. Aidan Morris
11. Bridgette Hunt
12. Noah Stasuik
13. Stefan Spataro
14. Miles DeBoer
15. Karam Hejazin
16. Daniel Penner
17. Cheng Fang

1.2 Location

- SCI 333

1.3 Zoom Link

- https://ubc.zoom.us/j/64827821543?pwd=lR0POOOIMNjlSAJfp2RUk2rS5onpij.1

1.4 Length of Each Talk

- 16 minutes followed by a 4-minute question period.

1.5 Schedule

Welcome (SCI 333): 8:45 am – 9:00 am

Session 1 (SCI 333): 9:00 am – 10:00 am

- 9:00 am – 9:20 am: **Noah Geerts**
  - ⋆ Title: Empirical Analysis of PCA-based Distance Comparison Operations
  - ⋆ Supervisor: Ramon Lawrence
- 9:20 am – 9:40 am: **Karel Harjono**
  - ⋆ Title: From Embeddings to Entities: A Comparative Analysis of RAG Architectures in Academic Domains
  - ⋆ Supervisor: Ramon Lawrence
- 9:40 am – 10:00 am: **Soren Stenback**
  - ⋆ Title: Descript-SQL: Enriched Database Descriptions in Text-to-SQL
  - ⋆ Supervisor: Ramon Lawrence

Break: 10:00 am – 10:15 am

Session 2 (SCI 333): 10:15 am – 11:15 am

- 10:15 am – 10:35 am: **MacKenzie Richards**
  - ⋆ Title: ECA Rules in EmbedDB: Enabling Lightweight, Event-Driven Decision Making for Resource-Constrained Devices
  - ⋆ Supervisor: Ramon Lawrence
- 10:35 am – 10:55 am: **Komal Singh**
  - ⋆ Title:
  - ⋆ Supervisor: Bowen Hui
- 10:55 am – 11:15 am: **Murad Shahmammadli**
  - ⋆ Title:
  - ⋆ Supervisor: Bowen Hui

Break: 11:15 am – 11:30 am

Session 3 (SCI 333): 11:30 am – 12:30 pm

- 11:30 am – 11:50 am: **Jerry Fan**
  - ⋆ Title:
  - ⋆ Supervisor: Bowen Hui
- 11:50 am – 12:10 pm: **Samantha Hodge**
  - ⋆ Title: Discovering Linguistic Correlates in Software Code Reviews" and my abstract is as follows
  - ⋆ Supervisor: Bowen Hui
- 12:10 pm – 12:30 pm: **Oluwadabira Omotoso**
  - ⋆ Title: Trash or Recycle: Addressing the Challenge of Improper Waste Disposal
  - ⋆ Supervisor: Patricia Lasserre

Lunch (SCI 333): 12:30 pm – 1:30 pm, we will offer the lunch (pizzas)

Session 4 (SCI 333): 1:30 pm – 2:30 pm

- 1:30 pm – 1:50 pm: **Aidan Morris**
  - ⋆ Title: Enhancing Sorting Efficiency in Embedded Systems: Implementation of Adaptive Sort within EmbedDB
  - ⋆ Supervisor: Ramon Lawrence
- 1:50 pm – 2:10 pm: **Bridgette Hunt**
  - ⋆ Title: Chatbots in the Classroom: A Survey of Student Sentiment Toward AI in Education

- ⋆ Supervisor: Ramon Lawrence

- 2:10 pm – 2:30 pm: **Noah Stasuik**
  - ⋆ Title: Evaluating LLM Performance in Essay Assessment: A Comparative Analysis of AI Grading and Feedback Systems for University English Courses
  - ⋆ Supervisor: Ramon Lawrence

Break: 2:30 pm – 2:45 pm

Session 5 (SCI 333): 2:45 pm – 3:45 pm

- 2:45 pm – 3:05 pm: **Stefan Spataro**
  - ⋆ Title: Automating Assessment: Canvas Quiz Generation for Supplemental Learning
  - ⋆ Supervisor: Yves Lucet
- 3:05 pm – 3:25 pm: **Miles DeBoer**
  - ⋆ Title: AI-Driven Extraction and Optimization for Intelligent Event Scheduling
  - ⋆ Supervisor: Ifeoma Adaji
- 3:25 pm – 2:45 pm: **Karam Hejazin**
  - ⋆ Title: Evaluating Multimodal Emotion Recognition Models on Real-World Data
  - ⋆ Supervisor: Ifeoma Adaji

Break: 3:45 pm – 4:00 pm

Session 6 (SCI 333): 4:00 pm – 4:40 pm

- 4:00 pm – 4:20 pm: **Daniel Penner**
  - ⋆ Title: CampusLink: A Social Networking Platform for UBC Okanagan International Students
  - ⋆ Supervisor: Ifeoma Adaji
- 4:20 pm – 4:40 pm: **Cheng Fang**
  - ⋆ Title: The Impact of the Emergence of Large Language Models on Stack Overflow
  - ⋆ Supervisor: Ifeoma Adaji

Break: 4:40 pm – 5:00 pm

Prize Draw and Closing/Party (SCI 333):
5:00 pm – 6:00 pm (see Section 4)

1.6 Abstracts

1. **Noah Geerts:** Vector databases are increasingly used in applications such as information retrieval, recommendation systems, and natural language processing due to their ability to perform efficient similarity searches. These searches rely on k-nearest neighbor (KNN) algorithms, but exact KNN methods become computationally expensive as dataset size and dimensionality grow. Approximate nearest neighbor (ANN) algorithms address this issue by trading slight accuracy losses for substantial speed gains. However, even state-of-the-art ANN methods remain bottlenecked by distance comparison operations (DCOs), which dominate computational costs. To accelerate similarity searches further, we investigate replacing exact DCOs with approximate ones. We first examine ADSampling, an existing approximate DCO that reduces computation by randomly subsampling dimensions. We then propose a novel DCO, DDCres, which performs comparisons in Principal Component Analysis (PCA) space, prioritizing dimensions with high variance to maximize information retention. We benchmark DDCres against ADSampling and standard full-distance scanning within the HNSW and IVF algorithms, analyzing its impact on search speed and accuracy. Our results demonstrate that DDCres offers a significant performance boost with minimal accuracy loss, highlighting the potential of dimension-aware approximate DCOs for enhancing large-scale vector search systems.

2. **Karel Harjono:** Retrieval-Augmented Generation (RAG) systems are transforming how AI models access and utilize external knowledge, specifically in domain-specific applications such as education. Traditional RAG methods typically rely on vector store retrieval, which excels in semantic similarity but struggles with transparency and structured reasoning. This thesis explores an alternative approach, GraphRAG, which uses knowledge graphs to encode explicit relationships between entities from a given passage, potentially offering improved context relevance and explainability. Through a controlled evaluation involving curated datasets across seven academic disciplines and six question types, this thesis compares the performance, retrieval accuracy, and transparency of GraphRAG and vector-based RAG systems. Results show comparable performance across most metrics, with GraphRAG offering notable advantages in source traceability and structured retrieval. Additionally, this study introduces a domain-specific benchmark dataset to assess RAG systems in educational contexts. The findings highlight the value of structured retrieval in enhancing trust and interpretability in AI-assisted learning environments and suggest directions for future research on evaluation methodologies and user interface improvements.

3. **Soren Stenback:** Text-to-SQL research has developed over several years to make use of Large Language Models (LLMs). One avenue of research is to create models that prompt an existing LLM such as ChatGPT and provide specific context that will assist the LLM in generating correct SQL queries based on a provided database schema and a text question. However, there are certain functions or techniques that do not assist the LLM and reduce the accuracy of the model. Researchers must determine which techniques are beneficial and which are detrimental to improving an LLM for text-to-SQL. The focus of this thesis is to assess an existing model called E-SQL and attempt to apply improvements through generative database descriptions to increase the accuracy of the model. Additional discoveries through testing are presented.

4. **MacKenzie Richards:** Resource-constrained IoT and edge devices require autonomous, real-time decision-making capabilities without relying on cloud systems or energy-intensive polling. This thesis introduces event-condition-action (ECA) rule support in EmbedDB, transforming it into an active database system that executes reactive queries with minimal latency and memory overhead. We present a lightweight API for defining ECA rules that trigger on data insertion, enabling devices to evaluate conditions (e.g., sliding-window aggregates) and execute actions (e.g., alerts or actuation) entirely on-device. Our implementation maintains EmbedDB's efficiency, operating in under 4 KB of RAM, while adding active capabilities. Benchmarks show that EmbedDB processes ECA rules in 49 μs per insert, outperforming InfluxDB/Kapacitor by over 200 times in end-to-end latency.

5. Komal Singh

6. Murad Shahmammadli

7. Jerry Fan

8. **Samantha Hodge:** Automated Text Classification has the potential to save dozens of human hours spent manually labeling data. Modern techniques typically use deep learning approaches that can be tuned to a high accuracy but require large datasets for their training, which is often not available in practice. In this thesis, we explore multiclass text classification techniques with a small dataset involving code reviews written as part of students' software engineering course projects. Previous research has manually labeled this data using a set of communicative acts that represent 13 task-related and social behaviours. A key challenge faced is that deeply imbalanced data, with labels ranging from 39 to 3,728 data points. Textual and linguistic analysis was applied to uncover patterns potentially embedded in the structure of each class's language. These observations were then used in combination with three machine learning models to investigate if a linguistically targeted approach could provide greater accuracy than the common, sheer brute force approach driven by large amounts of data. Results show that text analysis provides some insights into communication styles, however, none of the examined models produced results with a level of accuracy that would make human replacement viable.

9. **Oluwadabira Omotoso:** This paper addresses the pervasive issue of improper waste disposal, primarily stemming from a lack of awareness regarding the appropriate sorting of trash. To tackle this problem, we propose two significant contributions: a comprehensive labeled dataset for training deep learning image recognition models and an object detection mobile application designed to assist users in identifying items for disposal and locating the nearest appropriate bins. Our dataset includes 11 major categories of waste, featuring specific subcategories that reflect the unique disposal challenges encountered on campus. We explore the limitations of existing datasets, including their lack of multi-classification capabilities and controlled environments that do not accurately represent real-world conditions. Using YOLOv8 for object detection, our application enables users to capture images of waste items, receive immediate identification with confidence scores, and access guidelines for proper disposal. We aim to increase recycling efforts and encourage responsible waste management practices within the community, ultimately contributing to a more sustainable environment.

10. **Aidan Morris:** This thesis presents the implementation of the Adaptive Sort algorithm within EmbedDB, focusing on its integration into embedded systems. The work highlights the challenges encountered, key design decisions, and optimization strategies tailored for resource-constrained devices. Performance evaluation is conducted using real-world datasets of varying sizes, demonstrating the algorithm's efficiency and adaptability. This research contributes to enhancing the sorting capabilities of embedded systems, improving their robustness, versatility, and overall performance in diverse applications.

11. **Bridgette Hunt:** This thesis explores the role of generative artificial intelligence (AI), specifically chatbots, in higher education by examining student sentiment, engagement, and the impact on instructor workload. The study focuses on the HelpMe system, a web application help center incorporating a Retrieval-Augmented Generation (RAG)-based chatbot adapted from the ChatEd architecture. Implemented in an upper-year computer science course at UBC Okanagan, the chatbot was designed to assist students with asynchronous questions using a knowledge base built from course materials and instructor-provided answers.

    To ascertain the chatbot's effectiveness and students' perception of it, survey data was collected from 83 participants, accompanied by backend system usage data and instructor-reported email correspondence. The survey gauged attitudes toward AI in education, willingness to use chatbots, preference for course-provided versus external tools, and comparative usefulness against traditional help-seeking methods like email and office hours.

    Results indicate that a majority of students are open to using chatbots for academic assistance, mainly if provided directly by the institution. While many participants responded neutrally to questions about the chatbot's capabilities, a significant portion acknowledged its utility and expressed support for broader integration of such tools in other courses. However, usage data revealed that most students interacted with the chatbot only once or twice, with a sharp decline in engagement after the initial weeks of the term.

    The findings suggest that while students are receptive to AI-based tools in the classroom, sustained engagement requires further improvements in usability, integration, and trust. The HelpMe system's chatbot may serve best as a supplemental resource rather than a replacement for traditional support channels. Ultimately, this thesis highlights both the promise and limitations of generative AI in education, offering insight into how such systems can be refined to better support students and educators alike.

12. **Noah Stasuik:** With artificial intelligence rapidly transforming industries globally, its integration into higher education appears increasingly inevitable. This thesis explores the potential of using LLMs (Large Language Models) to grade students' essays and provide feedback on their writing in 100-level university English courses. With grading consuming significant portions of professors' and TAs' time, there often remains insufficient opportunity to directly engage with students.

    In this study, various LLM models and assessment strategies were implemented to evaluate the quality of feedback and accuracy of grades delivered by AI systems in comparison to the original human graders. All participating students consented to be evaluated by both locally hosted AI models (Llama 3.1, 3.2) as well as OpenAI's commercial offerings (GPT-4o-mini, o1, and o3-mini). The findings indicate

that while AI currently lacks the consistency necessary to fully replace human assessment, newer and more powerful LLMs demonstrate progressively better performance in both grading accuracy and feedback quality. Furthermore, when these models are combined with specialized assessment methodologies, the results show even greater accuracy in both grading and feedback semantic similarity.

Although the results confirm that AI cannot completely substitute human grading expertise, they strongly suggest that these technologies could serve as valuable assistive tools in the assessment process. The AI-generated feedback showed particular promise for helping students improve their work, with semantic similarity metrics achieving acceptable scores when compared to human-provided guidance.

13. **Stefan Spataro:** This project is a web-based application built to help instructors efficiently create Canvas quizzes for supplementary course material. Instead of consuming excess time to manually create questions and answers, simply enter the concepts and the tool creates the content zipped into a Canvas-ready format. Generating insightful questions and right/wrong answers related to reading content can be time consuming. This tool can increase productivity by enabling instructors to generate quizzes in far less time, making it easier to scale quality teaching materials. The application uses a React (Vite) frontend and a Node.js + Express backend, with a PostgreSQL database for storing user data and quiz history. It utilizes accessible large language model APIs like OpenAI's GPT-4o to produce questions from user input suggestions, formatting outputs in QTI 2.2. By applying AI and modern web development tools, it successfully generates plug-and-play .zip files with QTI 2.2 quizzes while reducing workload. Future additions will include generating quizzes based on content input (PDF), token/credit currency system to manage use, and optimizations to the quiz generation prompt and logic.

14. **Miles DeBoer:** In an increasingly fast-paced digital world, individuals are receiving information from a variety of different sources. Important dates and events are often lost among large amounts of data, leading to missed appointments and forgotten deadlines. Despite the existence of calendar applications, these issues remain present as these applications still require manual input.

    To address this challenge, I designed and implemented a scheduling assistant application leveraging artificial intelligence for the extraction and scheduling of events from various data sources on a mobile device. Using Support Vector Machines (SVM), Naive Bayes (NB), Neural Networks (NN), and a custom AdaBoost Hybrid Model, I developed a model that takes in text values and detects the presence of an event. This model was then expanded upon to allow for image uploads, using EasyOCR, enabling the extraction of text from images, which the model then processes and an event is scheduled if one is identified from the image.

    The model was integrated into an Android mobile application where events were identified from Gmail, Google Docs, and Google Photos Library. The goal is to minimize reliance on manual calendar entry, streamlining the event-tracking process, and saving users time and effort.

15. **Karam Hejazin:** This thesis examines how well state-of-the-art multimodal emotion recognition models perform on real-world data. Most existing models are trained on clean, structured datasets like IEMOCAP, which feature studio-quality recordings and acted dialogue. However, real-world applications—such as analyzing social media content—introduce significant challenges like background noise,

low-resolution visuals, and spontaneous, unscripted speech.

To evaluate generalization, we created a custom test set of manually labeled YouTube clips from the News and Education domains. These clips were annotated for six core emotions and used to test two leading models: CORECT, a transformer-based model, and COGMEN, a graph-based architecture. Both models were tested in a zero-shot setting using combinations of text, audio, and visual input.

Our findings show a clear drop in performance across all modalities, with the highest F1 score reaching only 0.35—substantially lower than training performance. The results emphasize the limitations of current models when exposed to noisy, in-the-wild data, and point to the need for domain adaptation, robust feature extraction, and evaluation benchmarks that better reflect real-world conditions.

16. **Daniel Penner:** In response to the variety of social, cultural, religious, and even linguistic challenges faced by international students studying at smaller campuses, the project aimed to create a prototype system that demonstrates the viability of software in creating spaces for international students to establish strong interpersonal connections and adapt more easily to the local environment.

    The system offers three key features. First, users can add and maintain direct connections with one another and can message or call their connections. Users can also create or join chat rooms called servers, which allow them to meet other students and discuss shared interests or areas of study in a designated space. Finally, users can add locations they have visited to a city map, allowing students to share businesses or public spaces that other students can rate and review.

    The application was developed using a MERN technology stack and a variety of external APIs for asynchronous messaging, audio calls, map functionality, and automated email verification. The system was containerized using Docker and deployed to the domain campuslink.online via DigitalOcean. To test the usability of the site and its adherence to our intended use cases, we conducted an informal poll with a handful of undergraduate and graduate students at UBC Okanagan. This allowed us to analyze the effectiveness of our final product, and find areas of weakness for ongoing development and potential future application of the site.

17. **Cheng Fang:** Stack Overflow, launched in 2008, is a Q&A website mainly for computer programmers. It enables developers around the world to collaboratively solve coding problems, share expertise and learn from peers, making it a cornerstone of informal technical education. The website serves as a platform for users to ask and answer questions, and, through membership and active participation, to vote questions and answers up or down and edit questions and answers in a fashion similar to a wiki. In the past, learning IT skills via Stack Overflow was a common and mainstream approach. However, with the development and popularity of the Large Language Models, peoples' attention seems to turn to these LLMs. The emergence of LLMs like ChatGPT has significantly changed the landscape of online technical Q&A platforms, particularly Stack Overflow. ChatGPT uses online data sources such as Stack Overflow data to train its own model, and as the ChatGPT model is optimized, users seem to prefer ChatGPT as a source of programming knowledge. This thesis investigates the impact of the emergence of LLMs on Stack Overflow by analyzing shifts in user behavior, question quality, engagement metrics, popular topics before and after the introduction of ChatGPT. In order to identify trends in activity, response patterns, and sentiments of users on

Stack Overflow, the relevant datasets were extracted and examined. Additionally, machine learning techniques were deployed to make a forecast of future trends in Stack Overflow. The findings provide insights into the evolving role of community-driven knowledge-sharing platforms and offer predictions on their future sustainability in the era of AI-assisted learning. The study aims to inform both platform developers and users about potential adaptations needed to maintain the integrity and usefulness of collaborative online learning environments.

2 Computer Science – Parallel Session II

2.1 Names of Speakers

1. Beck Corkle
2. Divyajot K. Dadiala
3. Ryhs Smith
4. O. Omotoso
5. Wenqi Marshall Guo
6. Putri Leksono
7. Chinmay Arvind
8. Imoudu Ibrahim
9. Muhammad Ashar

2.2 Location

- SCI 236

2.3 Zoom Link

- https://ubc.zoom.us/j/67543234740?pwd=qhyiXqsSsWaxtjAlne3k7zdjakHJAo.1

2.4 Length of Each Talk

- 16 minutes followed by a 4-minute question period.

2.5 Schedule

Welcome (SCI 236): 8:45 am – 9:00 am

Session 1 (SCI 236): 9:00 am – 10:00 am

- 9:00 am – 9:20 am: **Beck Corkle**
  - ⋆ Title: An LLM Approach to Detecting Cross-Project Correlated Bugs in GitHub Issues
  - ⋆ Supervisor: Gema Rodriguez-Perez
- 9:20 am – 9:40 am: **Divyajot Kaur Dadiala**
  - ⋆ Title: A Systematic Literature Review: Integrating Psychological Well-being Model to Assess Software Developer Well-being
  - ⋆ Supervisor: Gema Rodriguez-Perez
- 9:40 am – 10:00 am: **Ryhs Smith**
  - ⋆ Title: Annotating Cancer Pathology Reports to Facilitate the Training of a Natural Language Processing Model
  - ⋆ Supervisor: Fatemeh Fard and Rasika Rajapakshe

Break: 10:00 am – 10:15 am

Session 2 (SCI 236): 10:15 am – 11:15 am

- 10:15 am – 10:35 am: **Oluwadabira Omotoso**

  - ⋆ Title: Bugspyter: Detecting Code Bugs In Jupyter Notebooks Using LLMs
  - ⋆ Supervisor: Fatemeh Fard and Jie Wu
- 10:35 am – 10:55 am: **Wenqi Marshall Guo**
  - ⋆ Title: LangGas: Introducing Language in Selective Zero-Shot Background Subtraction for Semi-Transparent Gas Leak Detection with a New Dataset
  - ⋆ Supervisor: Shan Du
- 10:55 am – 11:15 am: **Putri Leksono**
  - ⋆ Title: Semantic Segmentation of Sentinel-2 Imagery for Crop Type Mapping in Saskatchewan Using Deep Learning and Targeted Band Selection
  - ⋆ Supervisor: Shan Du

Break: 11:15 am – 11:30 am

Session 3 (SCI 236): 11:30 am – 12:30 pm

- 11:30 am – 11:50 am: **Chinmay Arvind**
  - ⋆ Title: PNRL: A System to Support Reinforcement Learning Development in an Undergraduate AI Course
  - ⋆ Supervisor: Yong Gao
- 11:50 am – 12:10 pm: **Imoudu Ibrahim**
  - ⋆ Title: Enhanced Degree Planning with Course Planner
  - ⋆ Supervisor: Abdallah Mohamed
- 12:10 pm – 12:30 pm: **Muhammad Ashar**
  - ⋆ Title: Study on the Use of Gamified Platform
  - ⋆ Supervisor: Abdallah Mohamed

Lunch (SCI 333): 12:30 pm – 1:30 pm, we will offer the lunch (pizzas)

2.6 Abstracts

1. **Beck Corkle:** Cross-Project Correlated Bugs (CPCB) are code defects reported in Issue Tracking Systems (ITS) like GitHub, where the root cause originates in a different repository. The increasing adoption of open-source software development increases reliance on external dependencies, raising the likelihood that reported bugs originate in an upstream dependency rather than the project's own code. Previous empirical studies show that CPCB are typically more time-consuming, spark more discussion and are more difficult to solve than within-project bugs. Existing automated detection approaches have used traditional NLP on the broader category of Cross-Project Correlated Issues (CPCI) and have shown limitations. Motivated by this, this work explores the use of Large Language Models (LLMs) for

CPCB detection, exploring various prompt engineering techniques and input data combinations using issues from popular GitHub scientific computing repositories (eg., Numpy and Matplotlib). The results demonstrate strong performance, scaling positively with model size, demonstrating the utility of LLMs in detecting these complex, cross-repository dependencies documented in ITS.

2. **Divyajot K. Dadiala:** The well-being of software developers plays a critical role in their productivity and the overall software engineering process. Despite its significance, this aspect has often been overlooked. This paper examines how the PERMA model, a psychological framework, has been applied in various studies to assess and understand the well-being of software developers. Through a systematic review of research conducted in the software industry, we identify gaps in existing studies. For our study, we extracted relevant papers from three widely recognized academic databases: ACM, IEEE Xplore, and Scopus. We then applied a filtering process to identify studies closely aligned with our research objectives and used the final selection to derive our findings. Our analysis categorizes prior research based on different aspects of the PERMA model, revealing that while some factors have been explored, they remain insufficiently studied. Additionally, we investigate the well-being themes and factors used in the studies. Our findings highlight opportunities for further research and methodological advancements in assessing software developers' well-being.

3. **Ryhs Smith:** Natural language processing (NLP) is a form of artificial intelligence that allows computer models to rapidly analyse large volumes of complex text and summarize essential information. In healthcare, NLP can assist medical professionals with extracting key information from medical records to support clinical decision making. To achieve this, an NLP model must be trained to identify the critical medical information. This project explores the potential for training an NLP model to process breast cancer pathology reports.

   The starting point in this process was to manually annotate breast cancer pathology reports from BC Cancer data to standardize and categorize pathology information. The pathology reports were initially sorted by cancer type to ensure that only breast cancer pathology reports were selected for annotation. The three annotation software programs used were Prodigy, Doccano, and Label-Studio to help create the annotated breast cancer dataset. This dataset included categories for patient information, cancer stage, biopsy characteristics and other relevant details. The outputs of each annotation software have been converted to a standardized data type to allow for easier training of an NLP model.

   Annotating and converting breast cancer pathology data for this project is the first step in reaching the goal of developing a medical NLP model capable of summarizing critical pathological information for medical professionals.

4. **Oluwadabira Omotoso:** Computational notebooks are increasingly used in the fields of data science, computer science, classrooms, the software industry, and various fields. However, users often encounter errors, bugs, and vulnerabilities related to modularized code, unexecuted cells, and outdated library versions. This paper presents a tool, Bugspyter, designed to detect and repair code bugs in Jupyter Notebooks using LLMs. We develop an agent-based model to test the performance of the LLM in identifying the bug types and the root causes of these bugs in the notebooks, along with enhancing the model with static analysis results. Our results show that Bugspyter can identify buggy notebooks and has a high accuracy

for identifying implementation bug types in notebooks. Additionally, it can identify coding errors as the root cause of bugs in a notebook but fails to perform well in other root causes of bugs. Furthermore, we see an improvement in the performance of LLMs when identifying bugs in executed notebooks but no change in performance with the inclusion of static analysis results. This study contributes valuable insights into enhancing the reliability of computational notebooks, as it helps to reduce the need for many manual evaluations to fix these issues.

5. **Wenqi Marshall Guo:** Gas leakage poses a significant hazard that requires prevention. Traditionally, human inspection has been used for detection, a slow and labour-intensive process. Recent research has applied machine learning techniques to this problem, yet there remains a shortage of high-quality, publicly available datasets. This paper introduces a synthetic dataset featuring diverse backgrounds, interfering foreground objects, diverse leak locations, and precise segmentation ground truth. We propose a zero-shot method that combines background subtraction, zero-shot object detection, filtering, and segmentation to leverage this dataset. Experimental results indicate that our approach significantly outperforms baseline methods based solely on background subtraction and zero-shot object detection with segmentation, reaching an IoU of 69% overall. We also present an analysis of various prompt configurations and threshold settings to provide deeper insights into the performance of our method.

6. **Putri Leksono:** This thesis explores semantic segmentation of Sentinel-2 satellite imagery for crop type mapping in Saskatchewan, Canada, using deep learning models. Accurate segmentation of agricultural land is essential for precision agriculture, and satellite imagery offers a scalable solution for large-area monitoring. However, existing approaches often rely on all available spectral bands, potentially causing model overfitting and increased computational complexity. This research addresses these challenges by evaluating the impact of targeted band selection and model architecture on segmentation performance.

   The study focuses on three major crops in Saskatchewan: barley, spring wheat, and rapeseed/canola. Two Sentinel-2 band sets were tested. The first, derived from existing literature, includes B7 (NIR), B8 (NIR), B11 (SWIR), and WVP (Water Vapour Product). The second set was selected through a combination of statistical and spectral analysis, including ANOVA, Random Forest feature importance, and Principal Component Analysis, applied specifically to the reflectance profiles of each crop type. This process identified B1, B3, B5, and B12 as the most informative bands, reflecting their distinct spectral responses across the selected crops. Ground truth segmentation masks were generated from Canadian open government crop type data.

   Two segmentation models were implemented: a U-Net convolutional neural network and SegFormer, a transformer-based architecture. Preprocessing included radiometric normalization, resizing to 256×256 pixels, and data augmentation via horizontal flipping and rotation.

   Across both models, the feature-extracted band set consistently improved segmentation performance. For instance, U-Net's pixel accuracy increased from 0.6100 to 0.6441 with the selected bands, along with improvements in IoU and Dice coefficient. SegFormer achieved the best overall performance using the selected bands, reaching a pixel accuracy of 0.6908. These results demonstrate a strong correla-

tion between the chosen bands and the spectral characteristics of barley, spring wheat, and rapeseed/canola.

This study highlights the benefits of combining domain knowledge with crop-specific spectral analysis for band selection and demonstrates a practical approach for high-resolution crop segmentation. The findings support future developments in agricultural monitoring, land-use classification, and environmental analysis.

7. **Chinmay Arvind:** Reinforcement Learning (RL) is a key learning model in Artificial Intelligence (AI) and has been used to create game-based learning environments for training AI agents. However, these environments have been either too basic or too advanced to be adopted as a tool to teach RL concepts within an undergraduate AI course. In this thesis, a system called PNRL (Puddle Navigator Reinforcement Learning) is proposed, that uses a client-server architecture to create an environment for game-based learning to take place and support student learning of RL concepts at a sufficiently complex level for undergraduate students in an AI course. The puddle world problem is a benchmark problem in which an AI agent needs to learn how to navigate a grid world to achieve a goal by exploring the environment. This thesis explains the architecture of PNRL, its communication protocol that enable students to write parts of the client-side agent so moves are made and sent to the server, allowing instructors to evaluate the student's understanding of RL concepts. The instructor's evaluation can be guided by assessing the client-side (student's) agent's learning stability in its training process. The system as a whole is tested for its effectiveness in handling complex communication protocols and the client-side agent is tested for its effectiveness in achieving learning stability. A comparison of agent training termination conditions in terms of time-feasibility and learning stability is also conducted to find the most feasible condition in supporting student learning of RL concepts. Our evaluation indicates that the agent implemented in PNRL achieves learning stability, the system is reliable in handling communication and the maximum steps training termination condition is feasible to support student learning. This thesis also briefly discusses our investigation into the possibility of Python-Java integration to support the enhancement of the gaming server system in UBC's COSC 322 AI course.

8. **Imoudu Ibrahim:** Course Planner is a web application designed to simplify degree planning for university students. Built using the Remix framework on React, it provides an intuitive interface for organizing courses across terms. My contribution focused on integrating Workday, allowing the platform to retrieve real-time course availability, schedules, and enable course selections to be saved seamlessly. The result is a more complete and responsive tool that supports students and academic advisors in planning efficient, personalized academic paths.

9. **Muhammad Ashar:** This study evaluates the impact of gamification on student engagement and academic performance by integrating the NuTeach platform with UBC Canvas. NuTeach incorporates game-like elements—such as coins, flashcards, and deadline extensions—to motivate students in COSC 121, a first-year computer science course. After implementation, an anonymous survey assesses whether these gamified features improve learning outcomes. The research aims to provide insights into effective educational gamification strategies while maintaining strict ethical guidelines, including voluntary participation and data anonymity. Results may inform future teaching tools and course designs.

3 Mathematics, Physics, and Statistics/Data

3.1 Names of Speakers

1. Divyesh Dagia
2. Rachel Brookes
3. Rory Campbell
4. Junhao Liu
5. Isaac Jasper
6. Aiden Stock
7. Syed Aamir Ahmed
8. Danny Santano
9. Ruoyan Hu
10. Samuel Street
11. Elise Arbo
12. Zahra Kagda
13. Wasi Naqvi
14. James Dufresne
15. Jayden Ryga
16. Marcus Cosentino
17. Betty Zhang

3.2 Location

- SCI 337

3.3 Zoom Link

- https://ubc.zoom.us/j/63259577548?pwd=CCSWXn46YN8yXoX5BEbsi4kGvyzkxU.1

3.4 Length of Each Talk

- 16 minutes followed by a 4-minute question period.

3.5 Schedule

Welcome: 8:45 am – 9:00 am

Session 1 (SCI 337): 9:00 am – 10:00 am, PHYS

- 9:00 am – 9:20 am: **Divyesh Dagia**
  - ⋆ Title: Modifications of the Einstein Lagrangian
  - ⋆ Supervisor: Dan Vollick
- 9:20 am – 9:40 am: **Rachel Brookes**
  - ⋆ Title: Development of a Home-Built Radio Telescope Capable of Monitoring Solar Transits and Detecting the 21 cm Line Due to Galactic Hydrogen
  - ⋆ Supervisor: Jake Bobowski
- 9:40 am – 10:00 am: **Rory Campbell**
  - ⋆ Title: The Effect of PrairieLearn on Students' Conceptual Understanding in Introductory Physics Courses
  - ⋆ Supervisor: John Hopkinson

Break: 10:00 am – 10:15 am

Session 2 (SCI 337): 10:15 am – 11:15 am, MATH

- 10:15 am – 10:35 am: **Junhao Liu**
  - ⋆ Title: Algebraic Topology: An Introduction to the Fundamental Group
  - ⋆ Supervisor: Paul Tsopméné
- 10:35 am – 10:55 am: **Isaac Jasper**
  - ⋆ Title: Algebraic Topology: An Introduction to Homology
  - ⋆ Supervisor: Paul Tsopméné
- 10:55 am – 11:15 am: **Aiden Stock**
  - ⋆ Title: Introductory Category Theory
  - ⋆ Supervisor: Paul Tsopméné

Break: 11:15 am – 11:30 am

Session 3 (SCI 337): 11:30 am – 12:30 pm, STAT/DATA/MATH

- 11:30 am – 11:50 am: **Syed Aamir Ahmed**
  - ⋆ Title: Analyzing the Competencies of a Modern-Day Data Scientist Using Topic Modeling
  - ⋆ Supervisor: Irene Vrbik
- 11:50 am – 12:10 pm: **Danny Santano**
  - ⋆ Title: Change-point Detection of Financial Market Crashes via Density Ratio Estimation between Sliding Windows
  - ⋆ Supervisor: John Thompson
- 12:10 pm – 12:30 pm: **Ruoyan Hu**
  - ⋆ Title: Epidemic shutdown strategies: Finding the best in a population with opinion dynamics
  - ⋆ Supervisor: Rebecca Tyson

Lunch (SCI 333): 12:30 pm – 1:30 pm, we will offer the lunch (pizzas)

Session 4 (SCI 337): 1:30 pm – 2:30 pm, MATH

- 1:30 pm – 1:50 pm: **Samuel Street**
  - ⋆ Title: Motivating First-Year Students in Calculus
  - ⋆ Supervisor: Paul Tsopméné

- 1:50 pm – 2:10 pm: **Elise Arbo**
  - ⋆ Title: Authentic Applications of Integral Calculus
  - ⋆ Supervisor: Paul Tsopméné
- 2:10 pm – 2:30 pm: **Zahra Kagda**
  - ⋆ Title: Approximating Non-Convex Curves via Mixed Integer Linear Programming
  - ⋆ Supervisor: Warren Hare

Break: 2:30 pm – 2:45 pm

Session 5 (SCI 337): 2:45 pm – 3:45 pm, PHYS

- 2:45 pm – 3:05 pm: **Wasi Naqvi**
  - ⋆ Title: Simulating In-orbit performance for the Cosmological Advanced Survey Telescope for Optical and uv Research (CASTOR)
  - ⋆ Supervisor: Alex Hill
- 3:05 pm – 3:25 pm: **James Dufresne**
  - ⋆ Title: Peering Behind the Interstellar Curtain: Investigating the Link Between Faraday Depth and Physical Structures in the Interstellar Medium
  - ⋆ Supervisor: Alex Hill
- 3:25 pm – 3:45 pm: **Jayden Ryga**
  - ⋆ Title:
  - ⋆ Supervisor: Andrew Jirasek

Break: 3:45 pm – 4:00 pm

Session 6 (SCI 337): 4:00 pm – 4:40 pm, MATH

- 4:00 pm – 4:20 pm: **Marcus Cosentino**
  - ⋆ Title: An Ordinary Differential Equation Model for Replicating Species Response to Fire in a Two-Patch System
  - ⋆ Supervisors: Rebecca Tyson and Michael Noonan
- 4:20 pm – 4:40 pm: **Betty Zhang**
  - ⋆ Title: From Disease Dynamics to Economic Shifts: How School Closures Affect Female Workforce Participation and Economic Health Amidst a Pandemic
  - ⋆ Supervisors: Rebecca Tyson

Break: 4:40 pm – 5:00 pm

Prize Draw and Closing/Party (SCI 333):
5:00 pm – 6:00 pm (see Section 4)

3.6 Abstracts

1. **Divyesh Dagia:** Einstein's theory of general relativity was a tremendously successful theory of gravity. This presentation examines modifications of the Einstein field equations. To do this we modify the Einstein-Hilbert Lagrangian and subject it to variations of the metric. This is known as metric formalism. We mainly look at the Palatini formalism where we subject it to independent variations of the metric and the affine connection, which is a more general variation and gives the same field equations for the Einstein-Hilbert Lagrangian (but not generally the same). We look at $f(R)$ theories in the metric and Palatini cases, as well as a more general $f(R)$ theory where the connection is allowed to be non-symmetric. We found different field equations for the metric and Palatini $f(R)$ theories. We also looked at Born-Infeld-Einstein (BIE) thoeries, as well as theories where non-metricity is part of the Lagrangian. We find that these theories provide a possible description of Einstein-Maxwell theories purely based on the geometric Lagrangian. Lastly we look at a cosmological application of BIE theory with a scalar field acting as the matter part, by subjecting it under the spatially flat FLRW metric. We were able to see an accelerated expansion solution when considering the potential $V(\phi) = e^{-\phi}$, and a decelerated expansion solution when considering $V(\phi) = \phi^2$.

2. **Rachel Brookes:** The UBC Okanagan Physics pyramidal horn radio telescope originally constructed by Dr. Jake Bobowski and students from the physics department in 2010. As part of my directed studies project, I have substantially upgraded the radio telescope to enhance its sensitivity and stability. Major improvements include the addition of a horn extension to increase the area of the aperture by a factor of four, the use of temperature-controlled Peltier coolers to minimize thermal noise in the low-noise amplifiers, refinement of the waveguide-to-coax adapter's impedance match to 50 Ω, and the integration of RF isolators along the signal chain to minimize reflections. Although not yet implemented in the setup, a novel low-loss (< -1 dB) bandpass filter based on coupled loop-gap resonators was developed. With these upgrades, the telescope has been optimized to achieve a signal-to-noise ratio > 3 when monitoring solar transits. This improved sensitivity also allowed for the successful detection of the 21 cm line emitted by neutral hydrogen in the plane of the Milky Way galaxy. Near 1420 MHz, a stable and repeatable signal was observed when the telescope was directed toward the galactic plane. The signal strength was strongest near the centre of the galaxy where a supermassive black hole called Sagittarius A^* is located. This black hole is about 4 million times the mass of the sun and is surrounded by a dense cluster of stars.

3. **Rory Campbell:** In 2021, the open-source homework system PrairieLearn was introduce into the PHYS 111 course at the University of British Columbia's Okanagan campus, followed by its implementation in PHYS 112 the following year. The Force Concept Inventory (FCI), a third-question standardized test used to access conceptual understanding in physics, was administered prior to the first lab (pre-test) and again in the 11th or 12th week of classes (post test). In this study, pre- and post-test scores, as well as normalized and, where possible, compared to data from the previous years. Common misconceptions, the FCI gender gap, and the effect of taking calculus prior to instruction were also examined. This research aims to assess whether the introduction of PrairieLearn had a positive, negative, or negligible impact on students' conceptual understanding in first-year physics

courses.

4. **Junhao Liu:** Topology is the field of math that looks at shapes and spaces, focusing on features that do not change when you stretch or bend them – just no tearing or gluing. Within topology, there is a branch called algebraic topology, which uses tools from algebra to study spaces. The central idea is to systematically associate algebraic objects to spaces – one important example being the fundamental group, a way of capturing information about loops in a space. This presentation will introduce the concept of the fundamental group, explore a cool application via Brouwer's Fixed Point Theorem and its real-world implications, and conclude with a discussion of the Van Kampen Theorem, a powerful method for computing the fundamental group of complex spaces by breaking them into simpler pieces.

5. **Isaac Jasper:** Topology is the branch of math that studies properties of spaces that are invariant under any continuous deformation. A key area within topology is algebraic topology, which uses tools from algebra to study spaces. A typical question in algebraic topology involves identifying and classifying the holes in a space. One of the main tools for addressing this question is homology. The goal of this presentation is to introduce the concept of homology, with an emphasis on developing intuition over technical detail. Through the use of visual examples and elementary computations, we will determine the homology of some familiar topological spaces, such as the circle and the torus. We will also present a cool application of homology in geometry.

6. **Aiden Stock:** This presentation explores the fundamental concepts of category theory. This presentation has very limited formal prerequisite requirements in terms of mathematical knowledge, however, "mathematical maturity" and exposure to abstract mathematics will help greatly. The goal is to get students interested in category theory, as well as try to make it more approachable, as historically it has been somewhat difficult for students. In particular, this presentation aims to cover first definitions of a category, functors, natural transformations, limits and colimits, and adjunction of functors.

7. **Syed Aamir Ahmed:** The surge in demand for data scientists prompts the question: "what exactly does it mean to be a data scientist?" We investigate this by quantitatively comparing academic curricula and industry requirements using Latent Dirichlet Allocation (LDA) topic modeling. To understand the differing perspectives of academia and industry in answering the research question, we analyzed graduate lecture slides from UBCO's Master of Data Science program, and over 250 job descriptions for data science roles at various levels, aiming to identify and compare the central themes and their significance within each context. Our findings reveal discernible gaps between the two domains. Industry roles emphasize skillsets surrounding Big Data ETL pipelines, modern deep learning techniques, including recent demand for Generative AI, as well as an emphasis on cross-functional collaboration, which were all not as prominent in the curriculum. As the field continues to rapidly evolve, these findings highlight the need for academic curricula to adapt by integrating these applied skills alongside core theory, ensuring graduates possess the comprehensive technical and soft skills required of an adaptable modern data science professional.

8. **Danny Santano:** Market crashes are unstable financial periods characterized by substantial shifts in the distribution of asset returns. Detecting these shifts

can provide economic insights and help identify market disruptions. This research uses density ratio estimation (DRE) to identify distributional changes surrounding market crash events. Specifically, we use a relative unconstrained least-squares importance fitting (RuLSIF) DRE method in a double forward sliding window (DFSW) approach that compares return distributions of pre- and post-crash windows with customizable timeframes. We apply this method to stock returns data surrounding the 2008 financial crash and find differences in density ratios compared to stable periods. We show that our decoupled DFSW approach effectively captures full-distribution shifts, demonstrating a contrast between stable and crash periods with elevated density ratios during market instability. This research demonstrates that RuLSIF-based DRE can successfully detect distributional changes during financial crashes by capturing shifts in the overall return distribution.

9. **Ruoyan Hu:** My research aims to find the optimal level of shutdowns during a pandemic by analyzing people's opinions towards the disease and their subsequent behaviours. By examining the balance between economic costs and health impacts, I want to create a valuable tool for policymakers in the future. Understanding this balance is crucial for effectively managing pandemics by minimizing both the health and economic impacts.

10. **Samuel Street:** Sometimes it can be difficult to find motivation for students to learn mathematics apart from the beauty of the logic and symbols themselves. This presentation is about bringing some "real world" problems to the level of first-year differential and integral calculus allowing students early in their journey of mathematics to see its practicality and allow professors to have easily found examples when motivating students to learn calculus. The presentation will go through an application of calculus being used to optimize production using a rough Cobb-Douglas type model of Apple's iphone production, followed by the use of a linear approximation technique used to determine whether the cell density of cancer cells is stable at different equilibrium points of a cancer model, and finished by a cool application of improper integrals in cancer treatment.

11. **Elise Arbo:** This presentation explores authentic, real-world applications of integral calculus. The information presented has been gathered over this past semester and is in the process of being combined-along with the contents of the previous presentation-into a comprehensive report on authentic applications of differential and integral calculus. Today we will focus on several key topics: indefinite integrals, definite integrals, and improper integrals. We will also dive into some practical applications of these topics including solving differential equations, the present value of continuous money flow, and pharmacokinetics-fields in which integrals play a crucial role. One of the key goals is to motivate students highlighting the relevance of the mathematics they are learning and demonstrate its importance in real-world applications.

12. **Zahra Kagda:** In analyzing several master's theses about Road Optimization, I discovered an error in a numerical method used to approximate non-convex curves. In my talk, I will discuss this numerical method, the error that was present and propose a corrected version.

13. **Wasi Naqvi:** The Cosmological Advanced Survey Telescope for Optical and uv Research(CASTOR) has been undergoing detector testing for the new CIS-303 de-

tectors. Understanding the detector effects and optimizing CASTOR's capabilities is a critical element needed to prepare the mission for launch by the end of this decade. To this end, a simulation pipeline was developed to generate in-flight simulation and characterization of these CMOS-based detectors. The detector simulation pipeline combines ESA's Pyxel framework with various photometric tools to generate dark current profiles, cosmic ray effects, readout electronics, and optical point spread functions for each of CASTOR's three passbands. The dark current is characterized by a widget that allows visualization of the expected number of "hot" pixels for varying current per unit area. A Cosmic ray generation model is fine-tuned for the circumpolar orbit of CASTOR. A machine learning-based rejection model is trained specific to CASTOR parameters. These tools are then applied to a realistic stellar catalogue from CASTOR's phase 0 survey, yielding key insights for in-orbit performance. There is a specific focus towards recovering the point spread function of faint stars in the Milky Way. Detector effects on sersic profiles of galaxies are also investigated.

14. **James Dufresne:** Faraday rotation of polarized radio emission is a powerful tool for probing the magneto-ionic structure of the interstellar medium (ISM). In this thesis, we present a simulation-based framework for constructing synthetic Faraday depth spectra from magnetohydrodynamic (MHD) models of a stratified, supernova-driven ISM. Using data from existing Flash simulations of (Hill et al., 2012), we apply post-processed photoionization via CMacIonize to self-consistently determine the spatial distribution of free electrons. We develop a preliminary catalog linking identifiable features in the Faraday spectra to underlying physical structures in the ISM. The results demonstrate the utility of synthetic observations for interpreting radio polarization surveys and this work establishes a testbed for enabling a clearer mapping between Faraday complexity and the interstellar environment.

15. Jayden Ryga

16. **Marcus Cosentino:** Species response to wildfire is highly context-dependent, and the habitat conditions caused by fire are difficult and costly to replicate experimentally. I aimed to determine the conditions (initial population values, patch regeneration stage, distance between patches, wildfire frequency, and fire offset between patches) that maximized species success and population stability in a fire-prone ecosystem. To better understand the biological impacts of a changing fire regime, I developed a system of ordinary differential equations (ODEs) replicating species' response to fire in a two-patch system. Previous ODE models have done little to capture patch regeneration and the variation in successional communities after disturbance. Adapting the model for North American passerines and solving numerically, I determined that fires every 2/3rds to 5/6ths of the fire period in initially moderately burned patches created the most biodiverse landscape with the greatest population size. This modelling approach could generate insights that improve conservation and management decisions.

17. **Betty Zhang:** The interaction between disease dynamics and GDP, often used as an indicator of economic health, has been poorly understood. We present a simple model that combines the standard SLIQR (Susceptible-Latent-Infectious-Quarantined-Recovered) disease model framework with economic logistic equations to explore the mutual influence between disease spread and economic performance. Our model specifically examines the impact of school closures on GDP,

particularly focusing on how the reduced contribution of the female workforce to GDP is affected when females are sent home due to school closures. Our findings reveal that as more schools close, GDP declines gradually and recovers slowly. Conversely, with fewer school closures, GDP tends to decline rapidly but also recovers more quickly.

4 Prize Draw and Closing/Party

Location: SCI 333

- 5:00 pm – 5:30 pm: Prize Draw
- 5:30 pm – 6:00 pm: Closing/Party