Caltech
Student-Faculty Programs

Session H Abstracts

x86-64 to ARM assembly language
Robert R. Walker
Mentors: Adam Blank and Ethan Ordentlich

As machine learning has become the skill of any industry applicant, programming languages such as Python, Java, and C++ dominate every computer scientist's resume. While these languages may remain industry standard and more user friendly languages, they can't interface directly with the hardware of the computer, which is a necessity for understanding the inner-workings of the machine. Because of the different way of thinking required to write efficient assembly code and its ability to interact directly with machine hardware, learning assembly can help students understand how their computers work at a much deeper level. While x86-64 has been perfectly adequate to date, its initial release date in the late 1990's brings into question if it's still the right choice. With the vast majority of new processors (including everything Apple produces) using chips that only support a newer assembly language called ARM and the decrease in learning difficulty that would result, the change to teaching ARM assembly must be made, and must be made now. By updating Caltech's Computer Systems course to the new ARM assembly language, we are better preparing our students for the future soon to come.

Redesigning Caltech CS 2 to support all students' experience
Ellie Chen
Mentor: Adam Blank

Data Structures college courses fall into two types: "implementation" and "client." Implementation-type courses focus on teaching implementing data structures, while client-type courses focus on teaching applying data structures. As of last year, Caltech's CS 2 course, "Introduction to Programming Methods," was heavily weighted toward the "implementation" side. In recent years, the percentage of CS students in CS 2 has decreased significantly. Hence, our project redesigns the course to be more directly applicable to non-CS major students as well. We created many new projects and labs, as well as revised previous assignments to transition CS 2 to weigh heavily toward the "client" side. Moreover, because interacting with teaching assistants (TAs) is a major aspect of the student experience, we also wrote documentation for teaching assistants on various domains and duties they must perform. This research will allow teaching assistants and the course itself to support and benefit CS 2 students in the near future, as well as future students of other core computer science courses.

Restructuring Caltech's CS 2 to support students of all majors
Tobjorn L. Nelson
Mentor: Adam Blank

Caltech's introductory computer science curriculum (consisting of CS 1, 2, 3 and 24) is currently designed with CS 1 to be the only course meant for *all* students at Caltech. The remaining classes are meant primarily for CS majors or students with a generally high interest in the subject. With over half of CS 2's enrollment in the winter term of 2025 consisting of non-CS majors, the overall goal of this research was to make CS 2 a more widely applicable course for students of all majors. As of winter term, 2025, CS 2's curriculum revolved around teaching students how to implement data structures. Projects now instead teach students how to use data structures and various algorithmic techniques through programming real-world applications including, but not limited to, JSON parsers, web crawlers, search engines, navigation apps. By shifting the course's focus away from implementing data structures and towards applying them, we hope that students will learn effective ways to use the tools that computer science offers them, so they can start applying it to their own respective research projects, regardless of the focus.

Improving EV aggregate flexibility with end-to-end learning
Apoorva V. Thanvantri
Mentors: Adam C. Wierman, Christopher T. Yeh, and Nicolas Christianson

As the population of electric vehicles (EVs) rises, meeting their charging demand efficiently while continuing to ensure reliable power grid operation has become increasingly challenging. To facilitate this, aggregators – entities that pool energy resources into one market participant – are tasked with combining the constraints encoding the charging flexibility of each EV into an aggregate flexibility set. Computing this set exactly is intractable, motivating the use of approximation methods. It is vital that the approximation is reliable, meaning infeasible power schedules must not be included in the approximate set, since this can lead to grid instability. Current approximation methods either do not provide this guarantee, or they come at the cost of overly conservative representations that may neglect regions of the true aggregate set important to downstream performance. To push the reliability vs. performance pareto frontier, we develop a novel approach by learning an inner approximation of the aggregate flexibility set via Input Convex Neural Networks (ICNNs). We apply this model to a variety of objectives, including electricity cost minimization. We evaluate our method against several other approximations on real-world load data and compare performance on downstream tasks while guaranteeing reliability.

Prediction selection in two-player games
Yuehan Diao
Mentors: Adam C. Wierman and Tinashe Handina

In a two-player game setting, it is often hard for players to obtain information about the reality, so they have to take actions based on game state predictions that can contain inaccurate information. Assuming players treat the given prediction as the reality and act accordingly to maximize their utilities, we observe that sometimes more accurate predictions lead to worse equilibrium payoffs. This motivates us to formulate a new optimization problem for state prediction selection. We frame the problem into a multi-armed bandit optimization question, where each prediction is treated as an arm with an unknown equilibrium payoff. We then construct confidence bounds over the expected loss incurred by each prediction and use successive elimination to discard predictions whose estimated losses exceed the error bound. The resulting no-regret algorithm, after running for T time steps, can distinguish near-optimal predictions from the set of all predictions. We also conduct numerical experiments to prove our algorithm is applicable in real game settings. Our discovery suggests the need to balance accuracy and equilibrium payoff and provides an effective way to determine the most desirable game state predictions.

Fourier neural operators for time dynamics of antiferromagnetic Mott insulators
Miles M. Waugh
Mentors: Anima Anandkumar and Chuwei Wang

Mott insulators exhibit complex nonlinear photoexcitation dynamics under intense optical driving, which could enable carrier multiplication beyond the Shockley–Queisser limit, making them promising candidates for next-generation solar cells. However, simulating these strongly correlated quantum systems is computationally expensive because classical integrators require fine temporal discretization to fully resolve the nonlinear dynamics. We address this challenge by employing Fourier Neural Operators (FNOs) as surrogate models to predict the time evolution of momentum distributions in optically driven Mott insulators. Our FNO model, trained using data from a fourth-order Runge-Kutta solver, exhibits an approximately 500-fold speedup while maintaining an average relative L2 test error of 0.0153 in predicting post-pulse momentum distributions over a range of system and driving parameters. These results demonstrate the ability of neural operators to serve as efficient surrogates for modeling nonequilibrium dynamics in strongly correlated systems, which could help advance our understanding and discovery of quantum materials.

Neural operators for wildfire spread in California
Akshay Ghandikota
Mentors: Anima Anandkumar and Jiachen Yao

Wildfire spread in California is a pressing forecasting problem driven by multi-scale interactions among fuels, weather, and terrain. We identify three scales in which wildfire can be modelled through neural operator methods. First, a coarse-grained, pointwise time-series Fourier Neural Operator (FNO) predicts next time-step cell-wise marginal burn probabilities, without explicit spatial dependence, operating at ~9km daily-weekly scales. Second, a more fine-grained spatiotemporal Local Neural Operator will predict the next-day burn area fraction profile and associated risk from terrain, fuels, and meteorology input on a 375m-500m grid. Third, for our highest resolution model at <100m spatial resolution hourly, the Local Neural Operator will be used to forecast fire perimeter evolution via a level-set signed-distance-function (SDF) formulation. At such resolutions, the fire plume is known to create its own local 'weather', in which the perimeter spread is modelled by a two-way weather-fire coupling. Thus, we also plan to pair our fire perimeter model with an FNO-based fine-tuned weather predictor that co-evolves near-surface winds and plume-driven buoyant updrafts conditioned on fire heat-flux, approximating fire–atmosphere feedbacks. We aim for our neural operator models at different scales to provide fast, data-driven forecasts that preserve underlying physics dynamics while offering practical lead time for planning and response.

Learning physics causality for inverse PDEs in generative neural operators
Thomas Y. Lin
Mentors: Anima Anandkumar and Jiachen Yao

We investigate methods to let generative neural operators learn physics causality for inverse partial differential equation (PDE) problems with sparse observation. Prior work on generative PDE solvers relies on joint embeddings of parameter and solution spaces, learning their correlation rather than their causal relation. Our key idea is to introduce a surrogate operator into both training and inference, thereby formulating a "vertical" physics loss in addition to the conventional "horizontal" data loss. We also provide a training curriculum across loss terms to improve practical efficiency. To keep the pipeline fully data-driven, we approximate the surrogate operator with a pretrained neural operator, enabling end-to-end optimization without hand-crafted solvers.

Forecasting Alzheimer's disease progression using brain-age slopes and longitudinal MRI
Dean Yao
Mentors: Pratik Chaudhari and Elena Mantovan

Early detection of accelerated neurodegeneration is essential for timely intervention in Alzheimer's disease. Leveraging longitudinal MRI data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) – 10,755 scans from 2391 participants – I trained an AutoGluon ensemble model to estimate "brain age", achieving a root mean squared error of ~30 months on a held out test set. For each longitudinal subject, the slope of brain age and biological age was calculated and slopes greater than 1 indicated accelerated aging. A polynomial regression that combines the slope of a patient's progression with their respective MRI scan and demographic factors was used to predict the slope of their next progression. These findings show that a patient's historical progression and MRI data can forecast disease progression and provide an interpretable metric for monitoring neurodegenerative risk.

Faster attention for large language models
Kailen A. Hargenrader
Mentors: Yisong Yue and Ivan Jimenez Rodriguez

The attention mechanism is the heart of the transformer architecture underpinning Large Language Models (LLMs). Computing attention scales asymptotically as O(N2), where N is the model's context length, which is notoriously slow. Our paper provid es an alternative formulation of attention which improves asymptotics when combined with sampling without making sparse or linear assumptions, as done in previous works such as Performer or Reformer. We present TreeAttention, a version of hierarchical attention using a binary tree. We find that TreeAttention is similarly expressive to

standard attention on tasks such as image classification using vision transformers and autoregressive sampling with GPT-2. Additionally, sampling methods provide a trade off between accurate attention approximation and computational efficiency.