## Deep Dive Project --- UIUC GPA

#### Project Outline

Professor Ulmschneider’s historical grade distribution visualization tool has been a valuable asset for students at the University of Illinois Urbana-Champaign (UIUC). It allows them to review past grade distributions of courses, aiding in making informed decisions about their class selections. However, while historical data provides a retrospective view, there is a growing interest in leveraging this data to predict future grade distributions for UIUC courses.

The primary problem is to determine whether it is possible to forecast the GPA or grade distribution of UIUC courses using historical data. Such predictions could be beneficial for various tasks:

* Faculty Recruitment: Departments aiming to hire new faculty members can use predicted grade distributions to assess potential impacts on course outcomes.
* Course Development: Instructors planning to modify course structures can anticipate how changes might affect student performance.
* Student Course Selection: Students can utilize predicted grade distributions to choose courses that align with their academic strengths and workload preferences.

#### Data Preprocess Notebook

* data\_extract.ipynb

#### Data Exploration Notebook

* data\_explore.ipynb

#### Baseline Learning Notebook

* baseline.ipynb

#### Deep Learning Notebook

* rnn.ipyn

#### Feature Importance Notebook

* feature\_importance.ipyn

#### Datasets

* Working Dataset: uiuc-gpa.pk
* Debugging Dataset: uiuc-gpa-debug.pk

#### Data Problems:

The University of Illinois Urbana-Champaign (UIUC) GPA dataset provides detailed information on course numbers, the number of students who earned each letter grade, and primary instructors from Fall 2010 to Winter 2023. Despite the richness of the data, several issues may impact the GPA prediction task:

* Discontinued Courses: Some courses are no longer offered after certain semesters, resulting in incomplete time series data for those courses.
* Newly Introduced Courses: There are courses that were introduced during the dataset's time frame, leading to a lack of historical data prior to their inception.
* Infrequently Offered Courses: Certain courses have been offered only once or a limited number of times between Fall 2010 and Winter 2023. Additionally, some courses are scheduled intermittently—such as once every year or every two years—resulting in sparse data points.
* Uneven Course Distribution Across Majors: Different majors offer varying numbers of courses, which could introduce bias or imbalance when comparing data across different departments.

Given the identified data issues, it is better to constrain the prediction of course GPAs to the individual course level.

* Enhance Model Accuracy: Individual courses have specific trends and patterns that are more consistently captured when analyzed separately.
* Mitigate Data Sparsity: Address issues with courses that are newly introduced, discontinued, or infrequently offered by tailoring the model to the available data for each course.
* Reduce Bias: Avoid biases that may arise from uneven course distributions across majors, leading to fairer and more reliable predictive outcomes.

Predicting future GPAs based on majors or aggregated averages may introduce significant inaccuracies due to the variability and inconsistencies in course offerings across different departments.

#### Group Members

* Hongbo Zheng (NetID: hongboz2)
* Zhijie Chen (NetID: lucmon)
* Ning Wang (NetID: ningw5)