|  |  |
| --- | --- |
| **UW DIRECT Capstone Project Proposal**  The project should allow trainees to cement the acquisition of data science skills and develop proficiency in the conduct of team-based interdisciplinary research | |
| **PROJECT NAME:** High Resolution Mass Spectrometry Data Analysis for Water Quality Assessment | |
| **SPONSOR NAME:**  **May we list you on our website as a partner DIRECT project partner? YES**  **Will graduate students be asked to sign a non-disclosure agreement? NO**  <<This information is for planning purposes only. However, there must be a final ‘public-facing’ version of the project for student portfolio and presentations. Please ask clarifying questions as needed>> | |
| **PROJECT DESCRIPTION:** This project will better utilize High Resolution Mass Spectrometry (HRMS) data for water quality assessment. The analysis of HRMS data for water quality assessment is still in its infancy, with many basic aspects of data reduction, analysis, and interpretation still lightly developed. Here, students will work on developing improved data alignment and feature clustering methods to facilitate the interpretation of the data (e.g. relate detections to sample or chemical characteristics).  **DESCRIPTION OF DATA TO BE USED: 1)** HR-LC/MS (High Resolution Liquid Chromatograph Mass Spectrometry) data from a regional survey of Puget Sound receiving waters representing an “urbanization gradient”. 2) HRMS data representing a dilution series of a chemical source represented by a complex chemical mixture (e.g. tire leachate)  The main variables in the dataset: mass-to-charge ratio (m/z), retention time (RT), detector intensity and for stretch works: MS/MS fragment relationships | |
| **PROJECT START DATE:** 3/26/18 | **PROJECT END DATE:** no later than 6/22/18 |
| **PROBLEM TO SOLVE/OBJECTIVE:**  1. MS feature detection: based on the detected intensity trend over time, resolve chemical features (representing detected compounds) out from the complex background noises and data interferences.  2. Feature alignment: Use data sorting methods to correct the distortion arising from detector variation and replicate variability to match the features from different samples; filling the gap of peaks and convert them into centroid output.  3. Feature clustering: By comparing detected features within samples and across sample types, group similar features together as clusters based on developed algorithms, which either reflect a chemical commonality or shared statistical relationship of the detected compounds.  4. Stretch goal: a) Calculate sample similarity based on various cluster of features; b) Identify a shared source or chemical property based on feature clustering. c) Prediction and identification of chemical identity based on the cluster information; d) Create a handy interface for the tool | T**IMELINES AND DELIVERABLES: (13 weeks total)**  Programming language & environment: Python 3.0+, SKLearn package and will be preferred, CNN tools is potentially required  Timelines:  Week 1 - 2: Background knowledge preparation: literature review, MS data import and get familiar with the data.  Week 3 - 4: MS feature detection method development.  Week 5 - 7: Feature alignment method development. Verification along with results from existing workflows, testing and debug.  Week 8 - 12: Feature clustering method development and validation, package test and debug.  Week 13 ~: summarize work and finish the final reports.  Deliverables:  A python package that functioned with:  1. MS data feature detection and alignment.  2. MS1 feature clustering according to different type of samples (dilution series or different backgrounds).  3. Prediction of dilution rate based on the cluster behavior or assignment/identification of known clusters in the new samples have different background noises. |
| **PROJECT MENTOR(S):** Edward Kolodziej  **UW FACULTY CO-ADVISOR:** David Beck  **PROJECT TEAM MEMBERS:**  Ximin Hu | |