Sponsor // Title

1. Pacific Northwest National Lab (Ryan Renslow): Generating novel, chemically sound molecules using latent dimension reaction manifolds from deep learning

2. Alaska Center for Energy and Power (ACEP, Heike Merkel): Data Driven Solar Performance Analysis for Alaska

3. Alaska Center for Energy and Power (ACEP, Chris Pike): Demand Charge Reduction at Poker Flats Research Range (PFRR)

4. Argonne National Lab (Maria Chan): Machine learning impurity energy levels in semiconductors

5. Optimum Energy: Generalized Chiller Efficiency Predictor

6. King County Metro: Predictive Management of Battery Degradation for King County Metro Buses

7. Novo Nordisk: Sequence features influencing yeast promoter strength – application of data science in biopharmaceutical development

8. KPMG: Signals of Energy Demand

9. UW, Mechanical Engineering (Igor Novsselov): Raman Spectra Decomposition Rate Analysis and Machine Learning based Prediction

10. UW (Lilo Pozzo): High-throughput measurement of deep eutectic solvent melting points using IR bolometry

11. UW (Dave Beck): MetaMolES: metabolite retrosynthetic discovery, prediction and analyses

12. Pacific Northwest National Lab (Robert Rallo): Applications of machine learning for chemical systems

13. Pacific Northwest National Lab (Schram/Oblath): Applications of machine learning in a track finder algorithm for Project8 (a fundamental nuclear physics experiment).

14. Pacific Northwest National Lab (Robert Rallo): Domain-aware embeddings for high-dimensional data clustering

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: Generating novel, chemically sound molecules using latent dimension reaction manifolds from deep learning

SPONSOR NAME: Pacific Northwest National Lab
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, UW-PNNL partnership

PROJECT DESCRIPTION: The vast majority of small organic molecules have yet to be identified, many of which are constituents of biofuels and biofuel crops. Deep learning can be used to both predict molecular properties and help identify novel molecular structures, which can subsequently be related to fuel properties and metabolic pathways. Moreover, a large portion of DOE (Department of Energy) research programs seeks to understand the biological, biogeochemical, and physical processes at the molecular scale, requiring knowledge of the molecules present in complex samples. Specifically, DOE is interested in metabolic pathways, biological systems, active phenotypes/functions, industrial reactions, and these topics as they relate to biofuels, bioproducts, earth systems, and climate.

One bottleneck in this field is the generation of novel candidate molecules that may be present in real biofuel and related systems. The goal of this project is to determine if deep learning networks, specifically variational autoencoders (VAEs), can encode known chemical reactions. If possible, this could be used to generate lists of candidate molecules that may be identified in real samples. A software tool will be developed that will search the latent space created by an already-trained VAE for candidate molecular structures. Ultimately, the output from this tool will be a set of molecules, ranked for their proclivity to produce identical experimental signatures to those observed in the input data. The key of this software will be intelligent algorithms to search for candidate molecules. Approaches will focus on “latent space math”, wherein the latent signatures of structural motifs are isolated and subsequently added/subtracted to/from the latent representation of known structures. It is to be determined how robust latent space math is for chemical reactions. In practice, every possible molecular structure in this universe (within the confines of chemistry rules encoded in molecular identifiers, e.g., SMILES) will be representable in the software’s latent space. Thus, by traversing this space with intelligent methods, one would be able to generate the vast majority of molecules yet undiscovered. Traversing latent space using latent space math is an example of an advanced search method of latent space. Briefly, known chemical reactions (e.g., known biotic and abiotic reactions; metabolic pathway ontologies from MetaCyc) would be mapped onto latent space, and new molecules would be generated following these patterns into new latent space locations. The appeal to this approach would be that chemical reaction “math” within latent space could be possible. As an illustration, theobromine minus caffeine (representing the demethylation reaction) plus ethyltheophylline (T-C+E) would result in the latent space location of demethylated ethyltheophylline (i.e., 7-ethyl-3-methyl-3,7-dihydro-1H-purine-2,6-dione). In this manner, new molecules that are chemically possible through known reactions could reduce the massive search space and result in high probability structures that result from existing biological, biogeochemical, and industrial processes.

DESCRIPTION OF DATA TO BE USED: Data will consist of a trained VAE, thousands SMILES-based molecule structures, and chemical reactions (e.g. from BioCyc). The latent space math and searching software will be written in Python, and the output will be SMILES.

PROJECT START DATE: 3/29/19	**PROJECT END DATE:** no later than 6/14/19

PROBLEM TO SOLVE/OBJECTIVE: The objective of this project is to determine the robustness of chemical reaction encoding in VAE latent space and automate molecular structure generation from the latent space dimensions. Input into the Python software, to be produced by the DIRECT project, will be both known chemical reactions and reactants, as supplied as SMILES and/or SMIRKS. Initially, the team is to determine how well known common reactions are represented in latent space. This will require becoming familiar with utilizing a trained VAE written in Keras, the Python Deep Learning Library. If the latent space encoding of known reactions is found to be robust, the team will write automation software to explore this space for input reactants. If not, alternative latent space searching algorithms, or alternative latent space encoding/shaping methods, will be explored, with the goal of increasing the number of novel, yet chemically-sound, molecular structures.

T**IMELINES AND DELIVERABLES:**

- By April 2nd: Site visit at PNNL, tour, and project discussion
- By April 26th: Initial assessment of 10 known chemical reactions in latent space (Statistical analysis of these reactions)
- By May 10th: Presentation on results. Go/no-go decision point: continue latent space math for reactions or begin investigation of other latent space investigation methods.
- By May 31st: Python package/module for latent space exploration automation (with error checking, documentation, and PEP-8 formatting.)
- Stretch goal: Mapping of all known metabolic pathways from MetaCyc into latent space. Generation of 10k+ novel biologically-relevant molecules using MetaCyc mapping.
- By June 11th: Final presentation at PNNL

PROJECT MENTOR(S): Dr. Ryan Renslow, Pacific Northwest National Laboratory, ryan.renslow@pnnl.gov
Sean Colby, Pacific Northwest National Laboratory, sean.colby@pnnl.gov

UW FACULTY CO-ADVISOR: Dr. Jim Pfaendtner, University of Washington

PROJECT TEAM MEMBERS:

UW DIRECT Capstone Project Proposal

PROJECT NAME: Data Driven Solar Performance Analysis for Alaska

SPONSOR NAME: Alaska Center for Energy and Power (ACEP), University of Alaska Fairbanks (UAF)
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? We will ask students to anonymize any project presentations so that identifying information such as addresses are not shared with outside audiences.

PROJECT DESCRIPTION:
The large scale adoption of residential solar photovoltaic (PV) technology in Alaska is relatively new and a large scale performance data set is needed. This data set will assess whether solar PV systems are performing as modeled and more clearly demonstrate what solar PV performance has been realized by existing systems. This study is motivated by the growing economic viability of solar PV in Alaska due to lower installation and module costs.

The Alaska Center for Energy and Power has established relationships with the major solar installers in the state. Nearly 100 systems have been installed for over one year that have performance data available to ACEP. This data is in a variety of formats and will need to be parsed.

The most useful overall metric will be kWh produced (AC) per kW (DC) installed on a monthly and annual basis. Once this analysis is complete a deeper dive can occur using additional sets of parameters to try to investigate why certain systems seem to be outperforming others. .

DESCRIPTION OF DATA TO BE USED: Data downloaded from inverter websites is being acquired from homeowners and installers. Typically the data is shown as kWh produced each day. Data sets are in excel files, with a separate file for each installation. Metadata will be listed in a separate tab on the excel workbook. Geospatial data is in a separate format.

PROJECT START DATE: 3/26/18	**PROJECT END DATE:** 6/1/2019

PROBLEM TO SOLVE/OBJECTIVE: << describe the problem in a bit more detail and what are the specific objectives>>

Residential solar PV in Alaska is in the early stages and currently in a situation where the conversation typically revolves around the exceptional systems that are producing more than their owners expected, or alternatively, the systems that are underperforming. A broad data collection effort and analysis is needed. This project will accurately portray the performance of a wide swath of rooftop solar PV systems. The angle and azimuth of the PV modules have a major effect on their performance, but the meta data recorded by installers is often incomplete. Hi resolution LiDar data can be used to measure the angle and azimuths of the roofs in question. This data is available through state websites. In addition, google sunroof could be a source of useful data.

This project will use the actual performance data of systems cross referenced with geospatial data to demonstrate performance based on standard industry metrics such as kWh (AC) per kW (DC) installed on both a monthly and annual basis.

T**IMELINES AND DELIVERABLES:** << outline the expected work plan as well as what is expected to be delivered at the end of the project. The work plan should include the use or development of Data Science software/tools >>

- Parse data set into a usable format
- Use geospatial data to add metadata (azimuth and tilt angle) to the production data where it is missing.

Deliverables will include:

- Tables of actual produced kWh/installed kW monthly and annually for each system
- A contour map for the best, middle and worst producing systems.
- An analysis of the location of the best and worst performing systems as well as the type of systems that perform best and worst based on the metadata such as DC to AC ratios, tilt angles locations etc. If possible, commonalities between the similarly producing systems will be identified.
- ACEP envisions that the results of this project will be crafted into a short presentation for solar installers as well as interested homeowners to better understand the solar PV production in the region.

<< DIRECT students are versed in Python and SKLearn environment but can quickly pick up other languages/environments>>

PROJECT MENTOR(S): Chris Pike, ACEP Research Engineer, cpike6@alaska.edu, 907-799-6731.

UW FACULTY CO-ADVISOR:

PROJECT TEAM MEMBERS:

UW DIRECT Capstone Project Proposal

PROJECT NAME: Demand Charge Reduction at Poker Flats Research Range (PFRR)

SPONSOR NAME: Alaska Center for Energy and Power (ACEP), University of Alaska Fairbanks (UAF)
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? Unlikely

PR OJECT DESCRIPTION: The Poker Flat Research Range is a university-owned rocket launch facility near Fairbanks, Alaska. Utility demand charges on 4 installed meters at the range are substantial, and the prospect of combining them into one virtual meter with a single demand charge plus load leveling is assumed to decrease energy expenditures significantly. To understand actual savings that could be realized, 1-minute time resolution data were collected with 4 meters for a year. The project aims to use those data to show cost savings with one hypothetical virtual meter as well as demonstrate increased reliability with the possibility of uninterrupted transition to backup power sources during a utility outage by integrating energy storage and renewable/non-renewable generation resources.

This work will be applicable to similarly designed microgrids (e.g. U.S. military) where large temporary peak loads can introduce challenges to the operation of a microgrid and a solution that involves high cost expenditures is undesirable.

The student(s) analyzing the data for this project should have a background in electrical engineering and base-level training in microgrids and renewable energy sources.

DESCRIPTION OF DATA TO BE USED: We collected data at the range over a 12-month period from 4 power meters: 1 PQube meter and 3 Watts-On meters. The meters were installed at 4 locations across the range. Data sets contain samples at 1-2 Hz for each of the 48 channels. Channels contain electrical properties like voltage, frequency, current, active power, power factor, etc. The datasets are in netCDF format and follow the ASIM naming convention.

PROJECT START DATE: 3/26/19	**PROJECT END DATE:** no later than 6/22/18
PROBLEM TO SOLVE/OBJECTIVE: The project attempts to understand all necessary factors to develop an economically viable microgrid design and develop cost saving renewable and microgrid measures. These factors will be analyzed in the context of (1) reducing demand charges to the utility by deploying one virtual meter with a single demand charge, (2) load leveling using data from the 4 meters, and (3) offsetting imported power with locally produced, renewable power (solar or wind).	T**IMELINES AND DELIVERABLES:** Use MATLAB or Python to aggregate the datasets and perform quality control on the results. Identify missing data, peak loads, average daily loads. Obtain meteorological data and estimate solar and wind potential. Calculate load leveling potential. Use HOMER software for microgrid and distributed generation power system design and optimization to run simulations on the load profiles, estimate cost savings from using a virtual meter in place of the existing utility meters. Include renewable power generation in analysis and cost savings estimates. Compile report with potential renewables generation, load leveling, and cost savings measures.

PROJECT MENTOR(S): Heike Merkel hmerkel@alaska.edu

UW FACULTY CO-ADVISOR: << project sponsors w/external partners should have a UW faculty member as a co-advisor if possible >>

PROJECT TEAM MEMBERS:

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: Machine learning impurity energy levels in semiconductors

SPONSOR NAME: Maria K.Y. Chan / Arun Mannodi-Kanakkithodi
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

PROJECT DESCRIPTION: This project deals with the automated design of novel semiconductor materials with tailored impurity properties, enabled by a combination of atomistic simulations and machine learning.

DESCRIPTION OF DATA TO BE USED: The nature of the dataset will be a set of computed property values for several candidate materials which are represented by input vectors defined using standard material fingerprinting/descriptor practices.

PROJECT START DATE: TBD	**PROJECT END DATE:** TBD

PROBLEM TO SOLVE/OBJECTIVE:

This objective of this project is to create an artificial intelligence-based materials design framework for the targeted discovery of semiconductors containing certain impurities that lead to desired electronic and optical properties. The project will involve the generation of data on electronic, optical and defect properties of candidate semiconductors using first principles-based density functional theory calculations. Following this, modern machine learning methods such as neural networks and Bayesian optimization will be applied to develop forward property prediction models and inverse design loops that optimize material compositions for desired electronic and impurity behavior.

T**IMELINES AND DELIVERABLES:**
Months 1 and 2: Use of ScikitLearn and TensorFlow to perform dimensionality reduction of from complex, mixed types of input representation of materials.
Months 2 and 3: Collect atomistic simulation data (already available), train predictive neural networks using TensorFlow.

<< DIRECT students are versed in Python and SKLearn environment but can quickly pick up other languages/environments>>

PROJECT MENTOR(S): Maria Chan (mchan@anl.gov)

UW FACULTY CO-ADVISOR: David Ginger

PROJECT TEAM MEMBERS: Arun Mannodi Kanakkithodi (mannodiarun@anl.gov)

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: Generalized Chiller Efficiency Predictor

SPONSOR NAME: Optimum Energy, LLC
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? Yes

PROJECT DESCRIPTION: Predict the efficiency of chillers based on design specifications and model numbers, using data from existing chillers.

DESCRIPTION OF DATA TO BE USED: Chiller temperature, pressure, and energy consumption data for a variety of chillers and sites over several years.

PROJECT START DATE:	**PROJECT END DATE:**

PROBLEM TO SOLVE/OBJECTIVE:
Centrifugal chillers are large, centralized refrigeration machines often used to generate cooling for large HVAC systems. Chillers are generally custom-built based on customer requirements with a variety of bundle sizes and motor sizes, which makes a general prediction of chiller electrical efficiency difficult. Optimum can predict the efficiency of one chiller based on operational data from that chiller, but we can only approximately predict the performance of another chiller based on chillers of similar makes and models. Given a dataset of chiller performance and design specifications, predict the efficiency of another chiller based on design data and specifications.

T**IMELINES AND DELIVERABLES:** Develop a method of predicting chiller efficiency based on model numbers and design specifications. Optimum will recommend the features to be used in the prediction; don't worry about feature selection. Students may need to do feature engineering or elimination. We recommend a method using linear or non-linear regression, or a similarly straightforward prediction method; neural networks, deep learning, and decision trees will not create an auditable equation that we can use. Pay attention to data cleaning and validation; the data may contain values that are invalid due to communication errors or instrument error.

PROJECT MENTOR(S): Fred Woo, Michael Huguenard

UW FACULTY CO-ADVISOR:

PROJECT TEAM MEMBERS:

UW DIRECT Capstone Project Proposal

PROJECT NAME: Predictive Management of Battery Degradation for King County Metro Buses

SPONSOR NAME: King County Metro

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

PROJECT DESCRIPTION: This project will set the framework for a predictive battery replacement system for the King County Metro hybrid-electric bus fleet. Using known vehicle dynamics models and geospatial route data, this project aims to establish a model for estimating load required by the battery modules, generate a database of load history for each bus, and find relationships between specific routes and battery fatigue.

DESCRIPTION OF DATA TO BE USED: Data is available on the King County GIS portal, including the shapefiles (.shp) for the metro routes. LiDAR elevation data is available from the Washington Department of Natural Resources LiDAR portal. The OneBusAway API shows the route for a given bus ID number. Battery maintenance data is also available from the hybrid-electric bus battery management system.

PROJECT START DATE: 4/1/2019 (start of spring quarter)

PROJECT END DATE: no later than 6/28/2019 (start of summer quarter)

PROBLEM TO SOLVE/OBJECTIVE:

For the BAE Systems HybriDrive bus fleet used by King County Metro, each bus uses 16 lithium ion battery modules, providing ~12 kWh of energy storage. Around 100 modules per month are replaced after they have reached the end of life specifications determined by the battery management system (BMS). The cost of replacing the modules is high, due to the price of the modules and unpredictable nature of the replacement criteria. A predictive protocol, based on vehicle dynamics models, geospatial route parameters (from GIS data), and known battery degradation factors, will lead to a better understanding of the modules' remaining usable life and lead to a more economical replacement protocol.

T**IMELINES AND DELIVERABLES:**

Example workflow:

1. Define parameters needed for the vehicle dynamics model (e.g. route grade, velocity, acceleration, mass, weather conditions) and create procedures for estimating them
2. Create battery load profiles for specific buses
3. Create a database of load estimations and maintenance data
4. Use visualization tools and route metrics to compare battery fatigue between buses

Deliverables will include:

1. Python package that estimates the load required by the battery modules when given the vehicle dynamics parameters as inputs.
2. Database that includes BMS data, load estimation results from python package, and routes driven for each hybrid-electric bus.

PROJECT MENTOR(S): Mark Parsons, Mark.Parsons@kingcounty.gov
UW FACULTY CO-ADVISOR: Daniel Schwartz
PROJECT TEAM MEMBERS:

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: Sequence features influencing yeast promoter strength – application of data science in biopharmaceutical development

SPONSOR NAME: Novo Nordisk
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

PROJECT DESCRIPTION:

Heterologous protein expression in eukaryotes such as yeast can be challenging, but with a greater understanding of transcription factors we should be able to fine tune our levels of expression, thus improve the efficiency and ultimately reduce the cost and environmental burden of producing important biopharmaceuticals such as insulin.

In the past few years there has been a concerted effort to systematically explore these transcription factor impacts using natural and synthetic biology which has generated millions of data points. We propose integrating these different datasets together and applying machine learning techniques to help identify transcription factor motifs, sequences, and positions that have the greatest impact on protein expression and ultimately build a predictive model for protein expression based on this information.

DESCRIPTION OF DATA TO BE USED:

Databases

Element	Database
Transcription factor	http://www.yeastract.com/ http://yetfasco.ccbr.utoronto.ca/
Transcription Start Site	http://www.yeastss.org/

Publications

Element	Paper
Upstream untranslated region	Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences https://www.ncbi.nlm.nih.gov/pubmed/29097404
Upstream/downstream region	Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. https://www.ncbi.nlm.nih.gov/pubmed/23599004
Synthetic Promoters	The development and characterization of synthetic minimal yeast promoters DOI: 10.1038/ncomms8810
Synthetic Promoters	Design of synthetic yeast promoters via tuning of nucleosome architecture DOI: 10.1038/ncomms5002
Synthetic Promoters	Deciphering cis-regulatory logic with 100 million synthetic promoters http://dx.doi.org/10.1101/224907

Promoter regions	Core promoter sequence in yeast is a major determinant of expression level DOI:10.1101/gr.188193.114
Promoter regions	Asymmetric nucleosomes flank promoters in the budding yeast genome DOI:10.1101/gr.182618.114

PROJECT START DATE:	**PROJECT END DATE:**
PROBLEM TO SOLVE/OBJECTIVE: Combine existing datasets and cross validate previous modeling approaches exploring transcription factor effects that strongly influence protein expression in yeast. Build a comprehensive predictive/optimization model of protein expression in yeast based on the configuration (sequence, position, etc.) of transcription factors.	**TIMELINES AND DELIVERABLES:** Week 1 – Introduction Project kickoff & introduction meeting. Knowledge transfer & familiarization with datasets. Weeks 2-3 – Dataset integration, feature exploration, harmonization, and engineering Weeks 4-6 – Understanding different modeling approaches taken to date. Cross data validation of different approaches. Develop approaches to define comprehensive model. Weeks 7-10 – Build, test, and refine comprehensive model of host expression Weeks 11-12 – Conclusions and Presentation: Finalize results and conclusions and prepare final project reports and presentation

PROJECT MENTOR(S): Sally Lyons-Abbott(main), Jon Rue (co), Chung-leung Chan(co)

UW FACULTY CO-ADVISOR:

PROJECT TEAM MEMBERS:

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: Signals of Energy Demand

SPONSOR NAME: KPMG
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? Yes

Yes, students will be required to sign a non-disclosure agreement (NDA) with regards to all KPMG technologies and intellectual property. However, a public-facing presentation and poster highlighting the high-level results will be possible. KPMG requests ownership of all data, code, and model-related intellectual property developed during the project.

PROJECT DESCRIPTION:

Understanding and forecasting domestic and commercial energy demand is a great concern to utility companies, facilities managers, and building commissioning projects for energy-saving initiatives. Utility companies use demand estimates to reduce operating costs by ensuring the right amount of energy is produced, avoiding wasteful extra energy production or costly outages. Locally, facilities and operations managers make plans to optimize the operations of chillers, boilers, and energy storage systems. While energy demand is largely cyclical, is it also highly dependent on a number of climatic, economic, geospatial, and demographic variables that are publicly available. Understanding the role of these local variables in driving energy demand can guide future infrastructure investment and also feed-back into preditvie mores for improved forcasting accuracy and can serve as a method to predict local energy demand. In this project, students would utilize existing publicly-available and KPMG Signals Repository datasets of energy consumption and local economic, geospatial, and demographic variables to develop statistical insights and Machine Learning models around local consumer or business energy consumption. Insights from this model would be better used to understand and predict local energy demand, and would potentially be used to as a general method to forecast local energy consumption for future KPMG Advisory Engagements.

DESCRIPTION OF DATA TO BE USED:

Students will utilize KPMG's Signal Repository platform (S-R) to access publicly-available datasets and develop statistical insights and Machine Learning models around local consumer and business energy consumption. Students will explore multiple publicly-available energy consumption datasets. The following energy-demand datasets will be used as a starting point, but students will be expected to explore the publicly-available data for more signals:

https://www.kaggle.com/lucabasa/dutch-energy
https://www.kaggle.com/new-york-city/ny-dcas-managed-building-energy-usage
https://www.kaggle.com/city-of-seattle/sea-building-energy-benchmarking
https://www.energy.gov/data/downloads/open-data-catalogue
https://www.kaggle.com/claytonmiller/building-data-genome-project-v1
https://www.kaggle.com/chicago/chicago-energy-usage-2010

The above energy utilization signal will be combined with local economic, geospatial, and demographic variables contained within the KPMG Signals Reposity. Here there are many possible datasets to use here but these include:

- US Census Data
- EIA Fuel Price Data
- IRS Tax Return Data
- Economic Indicators (e.g., S&P 500, Small Business Loan Data, US Bankruptcy Filings)
- Local Business and Commerce Data
- Local Medicare Provider Locations

Students will work with KPMG professionals to ideate around which datasets will be most appropriate. All datasets will be available by the project start date, subject to each participant signing an NDA and acknowledging agreement to all Signals Repository user policies. Access to the Signals Repository will be terminated after the project end date.

PROJECT START DATE: 3/26/18	**PROJECT END DATE:** no later than 6/22/18
PROBLEM TO SOLVE/OBJECTIVE: Students will be expected to analize and develop a number of models around energy use: • Exploratory Data EDA: Students will be expected to do a comprehensive exploratory data analysis around the publicly available and KPMG Signals Reponsity datasets to ideate around appropriate variables for a predictive and explanatory models • Explainability Model: An explainable, regression-based model that shows the forcing of different local variables on energy consumption • Energy Predictive Model: Time-series forcasting-based energy predictive model • Enhanced Energy Predictive Model: Advanced model that enhances the previous forcasting-based model with local variables to predict energy consumption	T**IMELINES AND DELIVERABLES:** Students are open to use Python, SKLearn, R packages, particularly around time-series, geo-spacial analysis, and visualization packages. KPMG Data Scientists can provide support, but students are expected to be proactive about their required learnings. Work Plan: • 4/10: Initial touch point and brainstorming of model ideas. • 4/15: Introduction to the signals repository and APIs • 5/01: Exploratory Data Analysis Complete • 5/15: Model development around Exploratory, Energy-predictive, and Enhanced Energy Predictive models • 06/01: Model evaluation complete • 06/22: KPMG Signals Reposoitory Function Development, Write-up and Publicly-facing Poster-report complete Note: We will also include weekly- or bi-weekly check-in with students as appropriate. Expected Deliverables: • Comprehensive exploratory analyses documents (Notebooks) describing the exploratory process and datasets used. • Model code delivered in the KPMG internal Gitlab

	• Any signal harvesting code, including API calls to the S-R or othet data sources • Model evaluation Notebook • KPMG Signals repository data dictionaries and documentation around data used • KPMG Signals function code in KPMG internal Gitlab • A Signals 'archive' listing the predictive signals

PROJECT MENTOR(S):

Niels Hanson, Lead Specialist, Data Science, Data & Analytics, KPMG
Shreedhar Sasikumar, Manager, D&A Modeler, Data & Analytics, KPMG
William Nowacki, Managing Director, Data & Analytics, KPMG
Matteo Colombo, Principal/Partner, Data & Analytics, KPMG

Please reach out to Niels Hanson (nhanson@kpmg.com), Shreedhar Sasikumar (shreedharsasikumar@KPMG.com), Matteo Colombo (matteocolombo@kpmg.com), William Nowacki (wnowacki@kpmg.com).

UW FACULTY CO-ADVISOR:

Jim Pfaendtner (jpfaendt@uw.edu)

PROJECT TEAM MEMBERS:

UW DIRECT Capstone Project Proposal

PROJECT NAME: Raman Spectra Decomposition Rate Analysis and Machine Learning based Prediction

SPONSOR NAME: Mechanical Engineering Department, Igor Novosselov
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

PROJECT DESCRIPTION: Team will identify components in a mixture Raman spectrum and then define the decomposition and formation rates of the substances across varying temperature and resonance times. Ideally, the team will also be able to apply machine learning to predict decomposition, and we can confirm by performing additional experiments.

DESCRIPTION OF DATA TO BE USED: We have a supercritical gasification reactor that we have used to collect over 120 data sets of decomposition for 3+ substances (formic acid, ethanol, and methanol, with plans for dimethylformamide (DMF)) over 8 temperatures and 5 resonance times. The data is in '.txt' format.

PROJECT START DATE: 4/01/18

PROJECT END DATE: 6/22/18

PROBLEM TO SOLVE/OBJECTIVE:
Currently the process to analyze the large amount of data collected is to use a mix of Matlab code and another paid software 'Pexact' which is based off of research performed by a German research group around their work dating between 2004 – 2006. The process is tedious, very manual, and does not allow for custom built add-ons limiting the analysis to dated research. Using our reactor's generated data as well as the NIST open Raman spectra database the team will solve this problem by completing the objective of an open source platform for researchers to upload an effluent Raman spectra mixture set(s) and the software will read the spectra, identify components, define decomposition rates across varying parameters such as temperature, resonance time, possibly pressure. Then from the defined decomposition rate the software can predict the decomposition using machine learning beyond the known data set limits. *Note: this is a continued project from the winter quarter class*

TIMELINES AND DELIVERABLES:
Week 1 – 3 (deliverable at end of week 3): Code that can identify components in formic acid, ethanol, and dimethylformamide (DMF).

Week 4- 6 (deliverable at end of week 6): Code that can compute decomposition rate of one of the substances (either formic acid, ethanol, or DMF). Also, a comparison of the rates with published literature values.

Week 7-9 (deliverable at end of week 9): Code that uses machine learning to predict the decomposition rate of substance within data set range. The stretch goal would be to also predict outside of the bounds of the given data sets – i.e. higher temperatures or resonance times.

Week 10: All code committed (and open) to Github, Final report document complete and submitted to sponsor.

Note this should/will be in Python code

PROJECT MENTOR(S): Igor Novosselov, ivn@uw.edu
UW FACULTY CO-ADVISOR: Dave Beck, dacb@uw.edu
PROJECT TEAM MEMBERS: Elizabeth Rasmussen, Parker Steichen, Brandon Kern, Jon Onorato

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-throughput measurement of deep eutectic solvent melting points using IR bolometry

SPONSOR NAME: Prof. Lilo Pozzo
May we list you on our website as a partner DIRECT project partner?
Will graduate students be asked to sign a non-disclosure agreement? <<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: IR image data (bolometry) will be used to determine melting points of variable deep eutectic solvent compositions in high-throughput. A software algorithm and GUI will be developed to convert time-dependent IR bolometry video into accurate measurement of melting points of organic deep eutectic solvents with variable composition.
DESCRIPTION OF DATA TO BE USED: IR bolometry video data will already be available for students and will also be acquired by students at the project start time to ensure they are familiar with the experimental setup.

PROJECT START DATE: 3/26/18

PROJECT END DATE: no later than 6/22/18

PROBLEM TO SOLVE/OBJECTIVE: Deep eutectic solvents (DES) are promising materials due to their 'green' footprint, low cost, solvent quality and molecular versatility. DES are produced from two organic compounds that are mixed at a specific ratio that causes a depression of the melting point when compared to the pure materials. We are interested in the high-throughput evaluation of promising DES systems for use as electrolytes for energy storage. However, the possible design space is enormous due to the 1,000's of pairs of chemicals that could be combined to produce DES electrolytes. In order to be useful, we need to quickly identify DES compositions that will have useful melting points (low) for use in energy applications. IR bolometry (thermal imaging) is proposed as a high-throughput approach to measure melting point of a large array of samples in an effective and accurate way. However, this requires that we develop the software tools to obtain melting points from the IR video, removing aberrations and quantifying T_{melt} from the data along with saving it with the associated metadata.

T**IMELINES AND DELIVERABLES: S**tudents will work with an IR video bolometer (calibrated thermal camera) and a temperature controlled plate to understand how the technique can be used to evaluate melting points of solid samples in a parallel and scalable approach. The goal is to analyze 96 samples in less than 15 minutes.
After understanding the experimental setup and requirements, students will apply established AI and ML methods to automate the analysis of the IR bolometry data. The objectives and deliverables at the end of the project is to develop a software package (Python) that achieves:
- Data pre-treatment / cleanup and co-registration of sample information (i.e. metadata) with the IR video data.
- Analysis of local sample temperature and local plate temperature based on time-dependent IR image data.
- Plot data and automatically / simultaneously calculate melting points for all DES 96 samples using the IR video bolometry
- Optional: Use chemical information data (e.g. chemical structure, DES sample composition) to predict DES melting point through application of a machine learning algorithm.
- Optional: Compare accuracy of high-throughput IR bolometry melting point determination with the state of the art Differential Scanning Calorimetry (DSC) technique. Estimate error and sources of uncertainty.

PROJECT MENTOR(S): Prof. Lilo D. Pozzo
UW FACULTY CO-ADVISOR: Prof. Pozzo
PROJECT TEAM MEMBERS:

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: MetaMolES: metabolite retrosynthetic discovery, prediction and analyses

SPONSOR NAME:
May we list you on our website as a partner DIRECT project partner?
Will graduate students be asked to sign a non-disclosure agreement? <<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 aims to utilize data science and software engineering intuition to find and predict a plausible metabolic pathway for production of a given molecule with a retrosynthetic analysis approach. The current focus is to recommend enzymes that may produce a given molecule based on the enzyme's substrate promiscuity. The recommendation is based on a ranked comparison of the target molecule with the max common substructure of an enzyme's native products. Basically, it will save the world by repurposing biology.

DESCRIPTION OF DATA TO BE USED: Published data from MetaCyc, KEGG, NCBI, etc. All data are public domain and completely available.

PROJECT START DATE: 3/26/18	**PROJECT END DATE:** no later than 6/22/18
PROBLEM TO SOLVE/OBJECTIVE: The team will identify existing metabolic pathways that can be repurposed to synthesize industrially relevant molecules via synthetic biology.	T**IMELINES AND DELIVERABLES:** At the end of the project, it is expected that the team will have constructed an open source software package in Python that when provided with a target molecule to be synthesized biologically, will identify existing enzymes that can be used to construct a new metabolic pathway in a microbial host that will synthesize that molecule or a closely related molecule that can be catalytically upgraded by a successive step.

PROJECT MENTOR(S): Dave Beck, dacb@uw.edu

UW FACULTY CO-ADVISOR:

PROJECT TEAM MEMBERS: Dave Beck

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: Applications of machine learning for chemical systems

SPONSOR NAME: N/A
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? Yes

PROJECT DESCRIPTION:
The overall project aims to develop domain aware machine learning models to improve predictions of the temporal evolution of chemical system. The effort is conducted in collaboration with data scientists, applied mathematicians, computational chemists, and experimental chemists.
DESCRIPTION OF DATA TO BE USED:
Existing experimental data will be provided in the form of 3D volume XCT scans and will be used to develop the data pipeline and machine learning models.

PROJECT START DATE: 3/26/18

PROJECT END DATE: no later than 6/22/18

PROBLEM TO SOLVE/OBJECTIVE:

The objective for this project is to develop a data driven machine learning models to provide:

1) a predictive forward model for chemical systems with temporal 2D image data
2) build on the aforementioned model to create generative model (stretch goal)

T**IMELINES AND DELIVERABLES:**

Workplan includes (but not limited to):

1) developing an understanding of container technologies, use of various computing resources (CPU/GPU/HPC), and learn common machine learning tools (keras, mlflow, etc.)
2) developing data processing pipeline for temporal 2D images
3) developing/extending machine learning models to study various chemical systems
4) developing/extending the existing machine learning tools to conduct large scale hyperparameters search

Deliverables:

1) Provide code that processes data, creates/tests/validate ML models
2) Provide a written report summarizing the work
3) Provide/deliver a talk on the research conducted at PNNL

PROJECT MENTOR(S):

- Dr. Malachi Schram
- Dr. Robert Rallo

UW FACULTY CO-ADVISOR:

- ??

PROJECT TEAM MEMBERS:

- Dr. Jan Strube
- Dr. Carlos Ortiz
- Dr. Jenna Pope

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: Applications of machine learning in a track finder algorithm for Project8 (a fundamental nuclear physics experiment).

SPONSOR NAME: N/A
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? Yes

PROJECT DESCRIPTION:
The overall goal of the Project 8 experiment is to measure the absolute neutrino mass using tritium beta decays. For this proposal, the goal is to develop machine learning model for track reconstruction. The effort is conducted in collaboration with data scientists, physicists, and engineers.

DESCRIPTION OF DATA TO BE USED:
Existing simulated data will be provided in the form of 2D images and will be used to develop the data pipeline and machine learning models.

PROJECT START DATE: 3/26/18	**PROJECT END DATE:** no later than 6/22/18

PROBLEM TO SOLVE/OBJECTIVE:

The objective for this project is to develop machine learning models to improve the efficiency and accuracy of track reconstruction for Project 8.

T**IMELINES AND DELIVERABLES:**

Workplan includes (but not limited to):

1) developing an understanding of container technologies, use of various computing resources (CPU/GPU/HPC), and learn common machine learning tools (keras, mlflow, etc.)
2) developing data processing pipeline for 2D track images
3) developing/extending the existing machine learning tools to conduct large scale hyperparameter searches

Deliverables:

1) Provide code that processes data
2) Create/test/validate ML models
3) Provide a written report summarizing the work
4) Provide/deliver a talk on the research conducted at PNNL

PROJECT MENTOR(S):

- Dr. Malachi Schram
- Dr. Noah Oblath

UW FACULTY CO-ADVISOR:

- ??

PROJECT TEAM MEMBERS:

Dr. Brent VanDevender

- Dr. Ben LaRoque

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: Domain-aware embeddings for high-dimensional data clustering

SPONSOR NAME: N/A
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

PROJECT DESCRIPTION:
The project will explore different approaches to include domain knowledge into machine learning algorithms with focus on unsupervised learning. In particular, we will adapt the loss function of deep autoencoders to develop embeddings which are optimized for clustering tasks.

DESCRIPTION OF DATA TO BE USED:
The project will combine large unlabeled datasets (e.g., ChEMBL with 1.8M chemicals) suitable to develop structure-property relationships for chemicals with smaller manually curated datasets. Specifically, data will include molecular descriptors, physicochemical properties and biological activity.

PROJECT START DATE: 3/26/18

PROJECT END DATE: no later than 6/22/18

PROBLEM TO SOLVE/OBJECTIVE:

The objective of this project is to explore methods for developing domain-informed embeddings optimized for clustering tasks. To this end we will couple the autoencoders with clustering algorithms.

T**IMELINES AND DELIVERABLES:**

Workplan includes (but not limited to):

1) Machine learning architectures for coupling autoencoders with clustering algorithms such as K-means and self-organizing maps.
2) Modifying the loss function of autoencoders to include terms related to cluster quality or to the preservation of chemical properties.

Deliverables:

1) Documented code available in a github repository.
2) Written report summarizing the work and the main results.
3) Talk on the research conducted at PNNL.

PROJECT MENTOR(S):

- Dr. Robert Rallo
- Dr. Malachi Schram

UW FACULTY CO-ADVISOR:

PROJECT TEAM MEMBERS: