Candidate Project Overviews

*Please complete the following list for at least three projects to which you have made significant contributions. Please provide as much detail as you can in 2-3 sentences.*

Project 1

Name: Sentiment Analysis of prescription Drug Reviews

Your role: Leader

Duration: two weeks, full time

Problem: Train a model to predict ratings from patients labeled as "Positive", "Negative", and "Neutral" as well as an additional model to predict whether there is complains about side effect in comments (Yes/No).

Action: For prediction ratings, the words in comments were cleaned, stemmed and vectorized using tf-idf, followed by different machine learning algorithms for modeling. For prediction of side effects, word embedding was used for vectorization instead, followed by modeling using long short-term memory (LSTM) for modeling.

Result/Deliverable: For prediction ratings, the model built using Random forest gave the best performance with an accuracy of 0.88. For prediction of side effects, LSTM gave an accuracy of 0.86 and ROC-AUC of 0.84.

Tools Used/Methodologies: Scikit-learn, TensorFlow, Logistic regression, SVM, Random forest, Gradient boosting, LSTM, AWS, Matplotlib and Seaborn.

Your contributions: This is the capstone project of immersive Galvanize Data Science Bootcamp. I led the project and finished it by myself. Knowledge learned from the classes was practiced through the project.

Project 2

Name: Molecular underpinnings of high methane emissions from unrestored industrial salt ponds

Your role: Leader

Duration: 2 years

Problem: Metagenomic analysis of 39 wetland samples together with environmental data in order to study extraordinary high emission of CH4 in a former industrial salt pond.

Action: Analyzed metagenomic DNA from South Bay salt ponds using assembly, lasso and network graphs to figure genetic pathways with the potential to high methane emission. Analysis of 16S rRNA amplicon and genomic bins were then applied to explore what microbes contribute to methane generation.

Result/Deliverable: In addition to methanogenesis genes, the phosphonate pathway expressed for demethylation of methylphosphonates (MPn) due to phosphate starvation was found to more greatly contribute to methane generation as side product.

Tools Used/Methodologies: DEseq2, edgeR, Spades, Metabat, FastTree, MG-RAST, ggKbase, Kbase, JGI Genome Portal

Your contributions: Led the project and responsible for major tasks, including sampling, DNA extraction, Sequencing library preparation, sequence analysis and writing for publication.

Project 3

Name: Diversity, evolution, and classification of virophages uncovered through global metagenomics

Your role: Leader

Duration: 2 years

Problem: Screen a public collection of over 14,000 metagenomes in order to have a collective understanding of the global virophage (a special type of viruses infecting another type of viruses) diversity, distribution, and evolution.

Action: The hmm models of the marker gene were constructed and used for screening, followed by re-assembly, annotation, SNP calling, analyzing their phylogenies and discussing their evolutionary relationships with transposons and other viruses.

Result/Deliverable: 44,229 assembled virophage sequences (334 high-quality genomes) were identified from diverse habitats including human gut, plant rhizosphere, and terrestrial subsurface. Comparative genomic analysis confirmed the presence of four core genes in a conserved block, leading to establishing a revised virophage classification including 27 clades with consistent genome length, gene content, and habitat distribution.

Tools Used/Methodologies: HMMER, FastTree, Spades, Metabat, IMG/VR, pandas, iTOL

Your contributions: Led the project and responsible for major tasks, including sequence analysis and writing for publication.