Haoyun Lei

E-mail: haoyunl@andrew.cmu.edu | Phone: +1(412)969-3798
LinkedIn: linkedin.com/in/haoyunlei/ | Website: https://leovam.github.io/

EDUCATION

Carnegie Mellon University Aug 2016 – May 2021
Ph.D. in Computational Biology (expected)
Joint Carnegie Mellon-University of Pittsburgh Ph.D. Program in Computational Biology
Computational Biology Department, School of Computer Science

Huazhong University of Science and Technology Sep 2008 – Jun 2012
B.S. in Biological Science
College of Life Science and Technology

SKILLS

Programming Languages: Python (proficient), R (fluent), MATLAB (fluent), Shell (fluent), Java (familiar)

Technical Skills and Tools: Machine Learning (scikit-learn), Deep Learning (PyTorch, TensorFlow), Bioinformatics (GATK, SAMtools, bedtools CNVkit etc.), Data Analysis (Numpy, Scipy, Pandas), Data Visualization (Matplotlib, Seaborn), Combinatorial Optimization (Gurobi, SCIP), Cloud Computing (AWS), Web Development (HTML/CSS/JS)

WORK EXPERIENCE

Laboratory Corporation of America Holdings (LabCorp) May 2020 – Jul 2020
Bioinformatics Summer Intern Westborough, MA

Converting Free-text Patient Data to ICD Codes using Natural Language Processing (PyTorch, TensorFlow)

- Explored language tools (**BioBERT**, **medaCy**) to annotate and chunk the important information in medical text
- Fine-tuned **BERT** model on ICD-10 code classification at chapter and block (first three characters) level
- Designed a **two-step BERT** model to predict multiple ICD-10 codes in LabCorp's patient medical text
- Managed to work on a small dataset and reached **84%** on multi-label clarification at chapter level

Benchmarking CNV Detection Tools (Python, R, Perl)

- Tested and compared public CNV detection tools for calling CNVs in targeted NGS data with a very small panel
- Explored combinations of parameters of tools to increase true positive detection in **CNVkit**, **DECoN** & **CoNVaDING**
- Designed algorithms to rescue and recover CNVs with a weaker signal in a very small panel of targets
- Reached over **94%** in sensitivity while kept specificity around **90%**

RESEARCH EXPERIENCE

Ph.D. Thesis: Integrating Multiple Data Types to Infer Tumor Evolution (Python, R, MATLAB) May 2017 - Present

- Created a mixed membership model for the **Non-negative Matrix Factorization (NMF)** problem
- Developed an efficient **coordinate descent algorithm** to solve the NMF problem in **Python**
- Designed a **Mixed Integer Linear Programming Model** with the popular optimization solvers of **Gurobi** and **SCIP**
- Reached **~95% accuracy** with only small set of data and no other existing methods could do this

Detection of Cancer Types and Relevant Features using Deep Learning with RNA-seq Data (PyTorch) Spring 2020

- Designed and fine-tuned **1D CNN**, **2D CNN** and a **hybrid CNN** models to detect cancer types
- Designed a **Stacked Denoising Autoencoder Classifier** to improve the detections (**~96% accuracy**)
- Applied **embedding** method to find implicit relationships between cancer samples and genes

Footprint Match and Pattern Detection using Machine Learning (scikit-learn) Spring 2017

- Classified ~ 10,000 footprint images with **Neural Network** and **SVM** using **scikit-learn (~95% accuracy)**
- Applied the **Scale-invariant feature transform (SIFT)** algorithm to the match of saved and new images
- Extracted the image patterns with **K-Means** and **Gaussian Mixture Model**

Predict Proto Genes using **Logistic Regression, Naïve Bayes Classifier** and **Decision Tree** Spring 2017
Copy Number Extraction from DNA Sequencing Data with **Numpy**, **Scipy** and **Regular Expression** Fall 2016
Model Gene Regulatory Network by combining **Boolean network** and **Ordinary Differential Equation** models Fall 2016