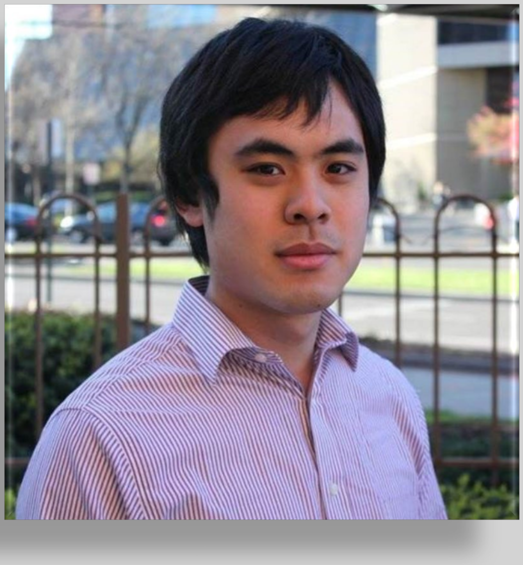
SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING

COMPUTER SCIENCE COLLOQUIUM SERIES



Yun William Yu

Harvard Medical School

Friday, February 15, 2019
3:00 pm
Luddy Hall, Rm. 1106

Compression for biological data analysis

Abstract: Compression has for decades served primarily the utilitarian purpose of enabling easier storage and transmission of data. Here however, I show how compression can be used to better understand biological processes and assist in data analysis.

First, I will demonstrate the relationship between lossy compression and understanding the perceptual characteristics of downstream agents. Quartz, my lossy compression program for next-generation sequencing quality scores counterintuitively improves SNP calling, despite discarding 95% of quality scores, showing the oversensitivity of variant callers to sequencer noise. More recently, I developed HyperMinHash, a lossy floating-point compression of the popular MinHash Jaccard index sketch that reduces the space-complexity from log(n) to loglog(n) by using the understanding that MinHash cares less about large hash values than smaller ones.

In the second part of this talk, I show how we exploit the compressive structure of biological data to speed up similarity search. I prove that by organizing the database to facilitate clustered search, our time-complexity scales with metric entropy (number of covering hyperspheres) if the fractal dimension of a dataset is low. This is the key insight behind our compressively accelerated versions of standard tools in genomics (CORA, 10-100x speedup for all-mapping of NGS reads), metagenomics (MICA, 3.5x speedup Diamond), and chemical informatics (Ammolite, 150x speedup SMSD).

Biography: Yun William Yu is a Research Fellow in the Department of Biomedical Informatics at Harvard Medical School, where he works on sketching and streaming algorithms for aggregate patient medical records. He received a BS in mathematics and a BA in chemistry from Indiana University, and completed an MRes in biomedical physical chemistry and an MPhil in mathematics at Imperial College London on a Marshall Scholarship. Supported by a Hertz Fellowship, he did his PhD in applied mathematics under Professor Bonnie Berger at the Massachusetts Institute of Technology.

INDIANA UNIVERSITY
**SCHOOL OF INFORMATICS,
COMPUTING, AND ENGINEERING**