# Joonyoung (Aaron) Bae

**Email:** joonyoungbae.aaron@gmail.com | **Personal Website:** n99joon.github.io

## Research Interests

Manifold Learning, Algorithmic Game Theory, Computational Biology, Statistical Machine Learning

## Education

**University of Southern California (USC)** Los Angeles, CA
*Master of Science in Computer Science (GPA: 4.0 / 4.0)* *Jan 2024 – Dec 2025*
- Thesis : Clustering and Structural Analysis of High-Dimensional Data on Manifold | Advisor: Jiapeng Zhang

**University of Hong Kong (HKU)** Hong Kong
*Bachelor of Engineering (Computer Engineering & Minor in Mathematics) (GPA: 3.44 / 4.3)* *Sep 2017 – Jun 2023*
- Leave for South Korean Mandatory Military Service (Sep. 2020 - Jun. 2022)

## Research Experience

**Graduate Research Assistant** USC
*Advisor: Jiapeng Zhang* *Aug 2024 – Present*

- Worked on improving clustering and visualization methods mainly for high-dimensional single-cell and bulk RNA sequencing biology datasets and image datasets. Main direction is to uncover and formulate latent structures of the data and devise algorithms leveraging the intrinsic information to yield meaningful improvements

**Main deliverables:**

- **CoreSPECT:** Designed a plug-and-play clustering enhancement framework, based on density-geomtery correlation observations, boosting existing algorithms (e.g. Kmeans & HDBSCAN) to near-SOTA performance with $\sim 50$x faster runtime. Developed the initial model assumptions and established theoretical guarantees for our framework
  - Manuscript: **CoreSPECT: Enhancing Clustering Algorithms via an Interplay of Density and Geometry**
    *Chandra Sekhar Mukherjee*,**Joonyoung Bae***,Jiapeng Zhang (*Equal Contribution) arXiv (2025)*
    (Under Submission for a Conference on AI&Statistics with High Initial Peer-Review Scores)
- **CoreMAP:** Developing a visualization algorithm that enables hierarchical display of most-to-least separable points by incorporating a novel anchoring idea and clustering on core nodes to UMAP's attraction-repulsion dynamics
  - The manuscript is currently under development
- **cplearn:** A clustering and visualization python package of **CoreSPECT** and **CoreMAP** published on PyPI

## honors and awards

**CSCI567: Machine Learning** - *Instructor: Vatsal Sharan* **(Ranked 1st among 250 students)** 2024
**Credit Suisse Global Coding Challenge 2021 (7th in World, 2nd in Asia)** 2021
- Algorithm problems scored on accuracy and runtime among $+20,000$ participants (Coded in C & C++)

**Republic of Korea Army Torchlight Award** 2020
- Top distinction in performance in my cohort of Cybersecurity Division (among $\sim$20 selected individuals)

**HKU Worldwide Undergraduate Student Exchange Scholarship** 2019-2020
**HKU Dean's Honours List** 2017-2018, 2019-2020

## Teaching Experience

**Course Grader of CSCI567: Machine Learning** USC
*Instructor: Haipeng Luo* *Jan 2025 – Jun 2025*

- Graded on the topics of Supervised/Unsupervised Learning, SVM, Kernel Methods, Neural Networks, Reinforcement Learning, Multi-armed Bandits, Learning in Games, etc

## Experience

**USC CS Theory Group** USC

*Member* *Jan 2024 – Present*

- Presented "Clustering and Structural Analysis of High Dimensional Data on Manifold" at the weekly gathering
- Attended weekly talks and gained insights on wide variety of theoretical computer science topics

**Full-Time Research Assistant** HKU

*Advisor: Eric Schuldenfrei, Faculty of Architecture* *Jun 2023 – Sep 2023*

- Developed 3D-Multilateration program using gradient descent to locate a device only using noisy signal strength
- Implemented object detection on a microprocessor, ESP32-CAM, with limited computation power using decision tree based algorithms (XGBoost and RandomForest) in C++, for smart building applications

**Cybersecurity Division** Republic of Korea Army

*Mandatory Military Service* *Oct 2020 – Apr 2022*

- Selected after interviews & exams and executed cyber-emergency response, real-time monitoring, packet analysis, and general maintenance of the Korean Army network systems, using tools such as SIEM, Firewall, and IPS

**Undergraduate Research Assistant** KAIST

*Advisor: Doo-Hwan Bae* *Jul 2020 – Sep 2020*

- Mainly conducted and published a Systematic Literature Review on Self-Adaptive System and its environment

- **Concepts and Models of Environment of Self-Adaptive Systems: A Systematic Literature Review**
  *Yong-Jun Shin,* ***Joon-Young Bae****, Doo-Hwan Bae* *APSEC (2021)*

## Projects

**Co-occurrence Embedding based Knowledge Base Compression in RAG** Sep 2025 – Present

- Developing Retrieval Augmented Generation(RAG) with memory efficient document retrieval by Word2Vec-like embedding of documents and human-feedback guided multi-document summarization, without loss of accuracy

**Justified Representation based Graph Centrality and Clustering** Jan 2025 – Jun 2025

- By viewing nodes of a graph as voters/candidates and edges as approval ballots, devised a graph centrality measure that gives balanced score on various-sized clusters by committee selection with justified representation guarantees

**Regularization on Various Stages of Convolutional Neural Network(CNN)** Jan 2024 – Jun 2024

- Demonstrated that adding various regularization methods on input, internal layers, and labels on CNN can boost performance of 450k parameters baseline model to achieve similar accuracy as AlexNet with 15x more parameters

## Audited Coursework

Below is the list of audited courses that I fully attended at USC, but not listed in the transcript

**MATH435: Introduction to Differential Geometry** - *Instructor: Soham Chanda*

**MATH505B: Applied Probability** - *Instructor: Jianfeng Zhang*

**EE546: Mathematics of High-dimensional Data** - *Instructor: Mahdi Soltanolkotabi*

**EE588: Optimization for the Information and Data Sciences** - *Instructor: Mahdi Soltanolkotabi*

## Skills

**Programming:** C/C++, Python & its ML libraries (Tensorflow, Torch, etc.), MATLAB, Java

**Academics:** Algorithms, Statistical ML, Probabilities, Differential Geometry, Spectral Graph Theory, Algorithmic Game Theory, Optimization, Deep Learning, High-Dimensional Statistics

**Language:** **(Native)** Korean **(Fluent)** English **(Intermediate)** Mandarin **(Beginner)** Cantonese