Sanjay Soundarajan

Email: contact@sanjaysoundarajan.dev • sanjaysoundarajan.dev
github.com/megasanjay • **linkedin**.com/in/sanjay-soundarajan

EXPERIENCE

FAIR DATA INNOVATIONS HUB - CALIFORNIA MEDICAL INNOVATIONS INSTITUTE San Diego, CA
Research Software Engineer Nov 2020 - Present

- Envision Portal
  - Leading development of the Envision Portal, a secure web platform enabling researchers to submit retinal imaging datasets. Developed as part of the milestones of the Eye ACT study where investigations on how ophthalmic conditions (macular degeneration, diabetic retinopathy) may offer early signals of Alzheimer's disease.
  - Facilitated AI model development by implementing structured imaging data ingestion pipelines and standardizing metadata for enhanced interoperability.
  - Collaborated closely with neuroscientists and data scientists to architect data schemas and imaging workflows optimized for real-time, AI-ready formatting.
  - Improved integration with AI model testing environments and developed advanced visualization dashboards to support future clinical applications.
- FAIRhub (AI-READI, NIH Bridge2AI)
  - Architect and full-stack developer of FAIRhub, an Azure-hosted platform for curating, managing, and sharing multimodal clinical datasets aligned with FAIR principles.
  - FAIRhub supports datasets for type 2 diabetes research and includes tools for dataset validation, metadata standardization, and user permissions.
  - Integrated GitHub-based workflow tools, version control of metadata, and automated FAIR compliance checking.
- Codefair
  - Contributor to Codefair, an open-source GitHub app that assists researchers in making software compliant with FAIR4RS principles.
  - Supports automated FAIR compliance checks, metadata generation, and integration with repositories like Zenodo.
  - Provides dashboard and interface for software authors to fix code documentation, licenses, persistent identifiers, and versioning.
  - Designed workflows that support Python, Jupyter, and R-based tools.
- FAIRshare
  - Designed and led development of FAIRshare, a cross-platform Electron-based desktop application that helps biomedical researchers prepare software and data for FAIR publication.
  - Emphasized compliance with FAIR4RS principles, automated license generation, metadata creation, and integration with Zenodo, GEO, and ImmPort.
  - Co-authored the FAIR-BioRS guidelines: a community-driven, step-by-step protocol for making biomedical research software FAIR.
  - Worked closely with NIH-funded research teams in infectious disease, immunology, and genomics.
- SODA for SPARC (NIH SPARC Program)
  - Developed intuitive UI and robust backend for SODA, a cross-platform desktop tool supporting the SPARC community in preparing autonomic nervous system datasets.
  - Designed integrations with the Pennsieve data repository for secure, reliable data submissions.
  - Used Electron, Vue.js, and Python Flask to create a scalable and user-friendly tool now used nationally by NIH researchers.
- SPARClink (NIH Hackathon Project)
  - Led team during the 2021 NIH SPARC hackathon to create SPARClink, a tool analyzing citation and reuse data from SPARC publications.
  - Aggregated PubMed metadata and visualized dataset usage trends across institutions.
  - Resulted in a peer-reviewed publication and third-place award at the hackathon.
- Create and maintain all the organizations' websites and all product documentation.
- Published novel development and findings in suitable scientific journals to share our findings with researchers and developers of FAIR data software tools.

INTERNATIONAL OFFICE - CALIFORNIA STATE UNIVERSITY, FRESNO Fresno, CA
Marketing Intern Jan 2020 - May 2020

- Worked with multiple departments by strategizing new methods to connect with students through social media.
- Organized events that allow for interaction between students of multiple diverse backgrounds.

COLLEGE OF SOCIAL SCIENCES - CALIFORNIA STATE UNIVERSITY, FRESNO Fresno, CA
Research Assistant Jun 2018 - Dec 2019

- Assist in the task of information gathering, editing, and verifying for 'American Chinese Restaurants - Society, Culture, and Consumption published by Taylor Francis Group. [**Book**]
- Gather and categorize sources for multiple research articles in the fields of Southeast Asian American anthropology.

PUBLICATIONS

- **Publicly Available Imaging Datasets for Age-related Macular Degeneration: Evaluation according to the Findable, Accessible, Interoperable, Reusable (FAIR) Principles** [**Journal**]
  - Evaluated 16 open-access OCT datasets related to age-related macular degeneration (AMD) for compliance with the FAIR (Findable, Accessible, Interoperable, Reusable) data principles.
  - Developed and applied assessment criteria, revealing critical gaps—especially in data re-usability and proposed guidelines to improve dataset standardization and accessibility for AI/ML research in ophthalmology.
- **AI-READI: rethinking AI data collection, preparation and sharing in diabetes research and beyond** [**Paper**]
  - Supported the development of AI-READI, a cross-disciplinary initiative creating a large-scale, multimodal dataset optimized for AI applications in type 2 diabetes research.
  - Contributed to efforts in data structuring, preparation, and sharing protocols, ensuring the dataset aligns with FAIR principles and supports equitable, high-quality machine learning research.
- **SODA: Software to Support the Curation and Sharing of FAIR Autonomic Nervous System Data** [**Journal**]
  - Co-developed SODA, a cross-platform desktop application that helps researchers prepare and publish autonomic nervous system (ANS) datasets in compliance with SPARC and FAIR standards.
  - Contributed to user interface development, metadata automation, and integration with tools like the SDS Validator and Pennsieve API, streamlining the data curation process for non-coders and advancing biomedical data reusability.
- **Clinical Dataset Structure: A Universal Standard for Structuring Clinical Research Data and Metadata** [**Abstract**]
  - Helped design the Clinical Dataset Structure (CDS), a root-level standard for organizing multi-modal clinical research data (e.g., surveys, vitals, imaging) to ensure FAIR compliance and AI/ML readiness.
  - Contributed to metadata schema integration and pilot dataset evaluation within the AI-READI project, enabling seamless structuring and sharing of complex datasets across clinical and computational teams.
- **Making Biomedical Research Software FAIR: Actionable Step-by-step Guidelines with a User-support Tool** [**Journal**]
  - Co-developed the FAIR-BioRS guidelines, the first practical, step-by-step framework for aligning biomedical research software with the FAIR4RS principles.
  - Contributed to the design and implementation of FAIRshare, an open-source tool that automates FAIR compliance workflows, enabling researchers to curate and share software more effectively
- **SPARClink: an interactive tool to visualize the impact of the SPARC program** [**Journal**]
  - Built SPARClink, a web-based tool that uses APIs, citation tracking, and knowledge graphs to visualize the research impact of SPARC-funded datasets and protocols dynamically.
  - Helped automate FAIR data utilization tracking by integrating APIs from Pennsieve, Protocols.io, NIH Reporter, and NCBI Entrez, contributing to the tool's recognition as Second Prize Winner at the 2021 SPARC FAIR Codeathon.
- **A comprehensive and high-performance motif finding approach on heterogeneous systems** [**Thesis**]
  - Developed and optimized DMF/PDMF, a hash-based DNA motif finding algorithm accelerated via multicore CPU, GPU, SIMD, and heterogeneous computing.
  - Achieved up to 41.48× speedup over serial baseline and demonstrated superior accuracy and runtime performance compared to existing motif-finding methods on real genomic datasets.
- **CPU-GPU Collaborated Computation Models for Biological Sequence Alignment with Mirror-Based Work Load Balancing** [**Paper**]
  - Designed and implemented high-performance CPU-GPU collaborative models for protein sequence alignment (Smith-Waterman algorithm) optimized through advanced load-balancing strategies.
  - Achieved up to 30.5× speedup compared to serial baseline and 2.78× improvement over basic GPU implementations, significantly enhancing computational efficiency on heterogeneous HPC systems.
- **Demystifying Transportation Using Big Data Analytics** [**Paper**]
  - Conducted big data analysis on Chicago's 2016 taxi dataset to identify city hotspots, customer satisfaction trends, and optimal driver earning strategies.
  - Applied linear regression and haversine distance calculations to model tip behaviors and commuting patterns, generating insights useful for transportation planning and service improvement.
- **PDMF: Parallel Dictionary Motif Finder on Multicore and GPU** [**Paper**]
  - Developed high-performance parallel computing models (PDMF) for efficient DNA motif discovery using combinatorial algorithms optimized with tree-based pruning and hash-based heuristics.
  - Implemented OpenMP and CUDA to achieve significant speedups (up to 41.48× over serial versions), demonstrating effective collaboration between multicore CPUs and GPUs on HPC systems.
- **Efficient Branch and Bound Motif Finding with Maximum Accuracy based on Hashing** [**Paper**]
  - Designed an efficient combinatorial motif-finding algorithm enhanced with hash-based heuristics to significantly reduce computational overhead compared to traditional branch-and-bound methods.
  - Demonstrated superior accuracy and practical runtime performance compared to widely-used approximate and suffix-tree approaches across diverse real-world DNA sequence datasets.
- **A Gaze-Based Virtual Keyboard Using a Mouth Switch for Command Selection** [**Paper**]
  - Developed a gaze-based virtual keyboard integrated with a USB-connected mouth switch, effectively addressing the "Midas touch" issue for severely disabled users.
  - Validated the system through user testing (NASA-TLX), demonstrating improved typing speeds (36.6 letters/minute) and higher usability compared to traditional dwell-time methods.

CONFERENCE PRESENTATIONS

- **Codefair: Your Personal Assistant for Developing FAIR Software [Presentation]**
  – Presented at US RSE 2024 Conference - 2nd Annual Conference of the US Research Software Engineer Association.
- **Making FAIR Fair to the Researchers [Presentation]**
  – Presented at FORCE2024 - FORCE11 Annual Conference.
- **Codefair: Make biomedical research software FAIR without breaking a sweat [Presentation]**
  – Presented at BOSC 2024 - 25th annual Bioinformatics Open Source Conference.
- **FAIR-BioRS: Actionable guidelines for making biomedical research software FAIR [Presentation]**
  – Presented at BOSC 2023 - 24th annual Bioinformatics Open Source Conference.
- **Making biomedical research software findable, accessible, interoperable, reusable (FAIR) with FAIRshare [Presentation]**
  – Presented at BOSC 2022 - 23th annual Bioinformatics Open Source Conference.

POSTERS

- **Clinical Dataset Structure: A Universal Standard for Structuring Clinical Research Data and Metadata [Poster]**
  – Presented at ARVO 2024 Conference - Association for Research in Vision and Ophthalmology.
- **Codefair - Your Personal Assistant for Developing FAIR Software [Poster]**
  – Presented at US RSE 2024 Conference - 2nd Annual Conference of the US Research Software Engineer Association.
- **Codefair - Make biomedical research software FAIR without breaking a sweat [Poster]**
  – Presented at BOSC 2024 - 25th annual Bioinformatics Open Source Conference.
- **FAIR-BioRS: Actionable guidelines for making biomedical research software FAIR [Poster]**
  – Presented at BOSC 2023 - 24th annual Bioinformatics Open Source Conference.
- **Making biomedical research software FAIR with FAIRshare [Poster]**
  – Presented at BOSC 2022 - 23th annual Bioinformatics Open Source Conference.

EDUCATION

CALIFORNIA STATE UNIVERSITY, FRESNO Fresno, CA

Master of Computer Science (GPA: 3.75, Magna Cum Laude) Aug 2018 - May 2020

- Specialized in high-performance computing and bioinformatics.
- Master's thesis: Developed and optimized motif-finding algorithms using SIMD and GPU acceleration.
- Published three peer-reviewed IEEE conference papers from thesis research on parallel and GPU-accelerated biological sequence analysis.
- Courses included Distributed Systems, Advanced Algorithms, Bioinformatics, and Software Engineering.

CALIFORNIA STATE UNIVERSITY, FRESNO Fresno, CA

Bachelor of Science in Computer Science (GPA: 3.55, Cum Laude) Jan 2015 - May 2018

- Strong foundation in data structures, databases, and systems programming.
- Completed a capstone project involving database-backed web application development.

ORGANIZATIONS AND INVOLVEMENT

- President of International Student Association at California State University Fresno. Jan 2020 - May 2020
- Campus Involvement Ambassador for Student Involvement at California State University Fresno. Aug 2019 - May 2020
- Board member for Off Campus Student Living at California State University Fresno. Aug 2019 - May 2020
- International Ambassador for the International Office at California State University Fresno. Jan 2019 - Dec 2019

PERSONAL CONTRIBUTION

- Developed and maintained platforms adopted by NIH-funded programs (Bridge2AI, SPARC) used by researchers across the U.S.
- Co-developed FAIR-BioRS guidelines, the first actionable software standard aligned with FAIR4RS principles.
- Led initiatives to ensure AI/ML models used in biomedical contexts are reproducible, traceable, and securely shareable.
- Designed systems that advance early detection of diseases like Alzheimer's using ophthalmic data and retinal imaging.
- Mentored young researchers in best practices of software development, reproducibility, and collaborative coding.
- Contributed to making scientific software more accessible to under-resourced institutions by leading open-source, automated tooling projects.