Sambit Panda

Cary, NC 27513 | **US Citizen**
919-637-6272 | sampanda501@gmail.com | linkedin.com/in/sampan501 | github.com/sampan501 | sampan.me

SUMMARY

- Highly motivated professional with 10+ years of research experience; interests include machine learning, data science, statistics, cancer genomics, and neuroscience
- Author of 14 publications (h-index: 7, ~250 citations); see all at https://sampan.me/pdf/Sambit-Panda-CV.pdf
- 7+ years of experience using Python and R to develop data science solutions in academic and industry settings

SKILLS

Python (LangChain, FastAPI, PyTorch, Dash, scikit-learn, pandas, TensorFlow), LLM APIs (OpenAI, Gemini, Vercel AI SDK), SQL (Google BigQuery), Cloud Services (Google, AWS, Azure), React (Next.js), R, Cython, Tailwind CSS, Developer Tools (Git, Docker), Continuous Integration (CircleCI, Travis CI) HTML, MATLAB, Unix Shell Scripts, Familiarity with C/C++, Java

RELEVENT EXPERIENCE

MATRIX AI Consortium **Dec 2024 – Present**
AI Research Scientist *Remote*

- Built an AI agent to aid emergency physicians with decision-making using a **Python** (via **FastAPI**, **LangChain)**, and **OpenAI/Gemini API** backend and a **React** frontend (via **Vercel AI SDK**, **Next.js**, and **Tailwind CSS**)
- Collaborated with leading trauma care physicians in Texas and Eric Horvitz (CSO at Microsoft) to build a base of knowledge of the assistant and quantify decision uncertainty of provided recommendations
- Devised a framework for evaluating LLMs (crafting and distributing a survey to ~50 clinicians in the process) and collaborated with top AI researchers to place performance guarantees on LLMs
- Leveraged ~400000 patients' historic data and new data collected by 7 trauma centers across Texas (using **SQL** with data stored in **Google BigQuery**) to develop an AI-driven geospatial tool to inform trauma care policy using **Python** (via **Dash** and **PyTorch**)
- Held a workshop attended by 50+ participants on how to use Generative AI for Biomedical decision

NeuroData Lab, Johns Hopkins **Jan 2019 – Dec 2024**
Researcher *Baltimore, MD*

- Developed multiple algorithms, notably KMERF (random forest-based hypothesis test), Nonparametric MANOVA (a nonparametric multivariate k-sample test), Fast Dcorr (fast approximation to the distance correlation test), and Causal Dcorr (distance correlation for causal inference)
- Authored 11 publications (5 first author, ~150 citations) related to early cancer detection, random forest, neural networks, causal inference, and hypothesis testing using **Python** packages like **TensorFlow** and **PyTorch**
- Created and maintained open-source **Python** packages like hyppo (~150 users, 200+ stars, ~100 forks) and treeple (50+ stars, ~20 forks); ported algorithms from these packages into SciPy.
- Developed and tested code using **Git**, **Docker**, Cloud Services (**AWS EC2/S3**, **Azure VM**), CI (**CircleCI**, **Travis CI**), and **Python** packages (**pandas**, **scikit-learn**)
- Collaborated with Bert Vogelstein, a renowned scientist in cancer genomics, on the MIGHT algorithm that quantifies predictive information in liquid biopsy feature sets; used **Python** packages (**treeple**, **scikit-learn**, **pandas**); wrote manuscript in preparation for PNAS
- Served as SciPy symposium conference chair and reviewer; journal reviewer for SoftwareX; presented work at top conferences like the BRAIN PI meeting and GYSS
- Worked on a project annotating whole body CT scans using **Python**, **Unix shell scripts**

National Institutes of Environmental Health Sciences **May 2023 – Jul 2023**
Data Scientist *RTP, NC*

- Applied the KMERF algorithm (which I created) to discover relationships in neurological data using **Python** packages (**pandas**, **scikit-learn**) and **R**; won 1st place in poster competition
- Collaborated with researchers to publish two manuscripts: (1) neurotransmitter signaling from fear response in mice and (2) the development of a fiber photometry **R** package; developed tutorials interfacing **Python** and **MySQL**

PROJECTS (Highlighting 4 of 6)

iRemedyACT | *Python (LangChain, OpenAI & Gemini API, PyTorch), SQL, Next.js, Google Cloud* **2024 – Present**

- A LLM agent to aid emergency physicians decision making at the point of care.
- A real-time geospatial model leveraging AI to give provide data-driven decisions for policy makers.
- Role: Creator and maintainer of both applications.

scipy.stats.multiscale_graphcorr | *Python, Cython* **2019 – Present**
- Multiscale Graph Correlation is a powerful multivariate test (the 1st and only multivariate test in SciPy).
- Role: Ported this algorithm from hyppo and maintainer.

hyppo (originally mgcpy) | *Python (scikit-learn, pandas), CircleCI, Cloud (AWS, Azure)* **2018 – Present**
- The first Python package for multivariate hypothesis testing, closing the gap with R (~150 users, 200+ stars, ~100 forks).
- Role: Creator and maintainer of this package.

EDUCATION

Johns Hopkins Medical Institute **Baltimore, MD**
PhD, Biomedical Engineering *Jul 2020 – Dec 2024*
- Awards: Computational Biology Fellowship (2020)
- Service: A-Level Capital (VC Firm) Life Sciences Advisor, TA (Neurodata Design I & II)

Johns Hopkins University **Baltimore, MD**
MSE, Biomedical Engineering *Aug 2018 – May 2020*
- Awards: AWS IMAGINE Grant (2018)

NC State University & UNC Chapel Hill **Raleigh & Chapel Hill, NC**
BS, Biomedical Engineering & Biology *Aug 2014 – May 2018*
- Awards: Magna Cum Laude (2018), Honors Program (2018), Dean's List (2014 – 2018), Goodnight Scholarship (Full Ride, 2014), National Merit Scholarship (2014)

PUBLICATIONS (Highlighting 5 of 14)

1. **Panda, S.*,** Shen, C.*, …, & Vogelstein, J. T. (2025). Universally Consistent K-Sample Tests via Dependence Measures. *Statistics and Probability Letters*, *216*(1), 110278. https://doi.org/10.1016/j.spl.2024.110278
2. **Panda, S.**, …, & Vogelstein, J. T. (2024*). hyppo: A Multivariate Hypothesis Testing Python Package*. Manuscript under review in JMLR.
3. **Panda, S.***, Shen, C.*, & Vogelstein, J. T. (2024). *Learning Interpretable Characteristic Kernels via Decision Forests*. Manuscript in preparation for ICML 2025.
4. Curtis, S.*, **Panda, S.***, Li, A.*, …, Vogelstein, B., Vogelstein, J. T.^, & Douville, C.^ (2024). *Detecting and Combining Useful Sets of Predictive Variables*. Manuscript in preparation for PNAS.
5. Shen, C., **Panda, S.**, & Vogelstein, J. T. (2022). The Chi-Square Test of Distance Correlation. *Journal of Computational and Graphical Statistics*, *31*(1), 254–262. https://doi.org/10.1080/10618600.2021.1938585

PRESENTATIONS (Highlighting 3 of 22)

1. **Panda, S.**, & Cruz, C. (2025, May). *Generative AI for Biomedical Decisions* [Oral Presentation]. MATCH DICB AIM-AHEAD program, Virtual.
2. **Panda, S.**, …, & Cushman, J. D. (2023, July). *Elucidating Relationships within Neurological Screening Batteries via Random Forest-Based Hypothesis Testing* [Poster Presentation] RTP, NC, USA.
3. **Panda, S.**, …, & Vogelstein, J. T. (2022, January). *Nonparametric MANOVA via Independence Testing* [Oral Presentation]. Global Young Scientists Summit, Virtual. https://www.youtube.com/watch?v=rJyuTwkgfjQ