ACHILLE NAZARET

Website: anazaret.github.io
Github: github.com/anazaret
Email: achille.nazaret@columbia.edu
Phone: (917) 216-1327

EDUCATION

Columbia University — New York, NY
Ph.D. in computer science. (4.0/4.0) – Advisors: Prof. David Blei and Prof. Elham Azizi. — Jan 2021 - May 2025
M.S. in computer science. (4.0/4.0) — Aug 2019 - Dec 2020

École Polytechnique — Palaiseau, France
M.S. in mathematics and computer science. — Aug 2018 - Apr 2019
B.S. in mathematics and computer science. (top 5% students) — Aug 2016 - Jun 2018

École Spéciale Militaire de Saint-Cyr — Coëtquidan, France
Accelerated track to the rank of army officer (second lieutenant). — Aug 2016 - Nov 2016

Lycée Privé Sainte-Geneviève (MP*) — Versailles, France
Competitive undergraduate program in mathematics, physics, and computer science. (ranked 1st) — Aug 2014 - Jul 2016

ACADEMIC RESEARCH EXPERIENCE

Blei Lab, Columbia University — New York, NY
Ph.D. candidate. Advisor: Prof. David Blei — Jan 2021 - Present

- Designing efficient probabilistic prediction models combining decision trees with diffusion models [1].
- Created a Python library and mathematical criteria for mechanistic interpretability in large language models (LLMs) [2].
- Scaled up causal discovery methods: $\times 10$ more variables, and $\times 100$ speed improvement [3] and [4].

Azizi Lab, Columbia University — New York, NY
Ph.D. candidate. Advisor: Prof. Elham Azizi — Jan 2021 - Present

- Developing interpretable multimodal generative models to understand cancer progression, [6] and [8].

Yosef Lab, University of California, Berkeley — Berkeley, CA
Research Assistant. Advisors: Prof. Nir Yosef, Prof. Michael I. Jordan, and Romain Lopez — Apr 2019 - Aug 2019

- Developed an open-source Python package [14] for single-cell data analysis: `scvi-tools` (1.3k+ GitHub stars).
- Designed multimodal generative models to impute unobserved genes in spatial genomics [13]. Still state-of-the-art in 2024.

WORK EXPERIENCE

Apple – Health AI — New York, NY
Machine learning research scientist (part-time, alongside Ph.D.) — Feb 2024 - Dec 2024
Machine learning research scientist (intern) — May 2021 - Aug 2021 and Jan 2022 - Aug 2022

- Estimated the causal impact of the Apple Watch's notifications on user behavior [10].
- Created foundation models of health biomarkers from time series data of wearables, to understand user health and fitness [9].

Palantir Technologies — San Francisco, CA
Forward deployed software engineer (intern) — Jun 2020 - Aug 2020

- Scoped, prototyped, and deployed data-driven algorithms to reduce costs for a US healthcare insurer.

Akwa Group — (remote) Casablanca, Morocco
Machine learning consultant (alongside M.S.) — Sep 2018 - May 2019

- Created datasets and built models to predict the performance of new gas stations – surpassed human experts by 25 %.

IMC Trading — Amsterdam, Netherlands
Software engineer (intern) — Jun 2018 - Sep 2018

- Distributed model training pipelines on a cluster for faster overnight training (HFT, futures).

Bernardaud — Paris, France
Operations research consultant (alongside B.S.) — Feb 2018 - Jun 2018

- Designed algorithms to find optimal production processes under factory constraints; created a full-stack website to use them.

Ministry of Defense — Paris, France
Junior data-scientist (intern) — Nov 2016 - Apr 2017

- Developed graph-mining and NLP models for social network analysis to produce intelligence and detect bot farms.

SELECTED AWARDS

PhD Fellowship from the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard (2022-2024).

Research in computer science.

- Best paper, 3rd prize, ICML 2024, Workshop on Mechanistic Interpretability [2] – Oral presentation.
- First place, ICLR 2023, GSK.ai CausalBench Challenge, Causal discovery contest [16].
- Best poster award, ICML 2022, Workshop on Computational Biology [11] – Spotlight presentation.
- Best poster award, ICML 2019, Workshop on Computational Biology [13] – Spotlight presentation.

Competitive programming.

- Google HashCode 2020, team `Optimistic`. Rank 180th (8th in the US)
- GSA Ultra 2019 Programming Competition. Rank 14th
- International Collegiate Programming Contest (ICPC), SW Europe Region, 2019. Rank 20th
- Ecole Polytechnique ICPC qualifier, 2019. Rank 2nd

Other contests.

- Citadel Securities: East Coast Statistics Datathon 2020. Rank 2nd
- DGSE (French intelligence agency): TRACS 2019: Cybersecurity & data-analysis challenge. Rank 2nd
- IMC Trading: Algorithmic trading competition 2018. Rank 3rd

National decorations.

- National Defense Medal, Bronze Echelon – *For exceptional services rendered for the defense of France.*

PROGRAMMING SKILLS

Expert in Python (for research, open-source development, and industry); experience with C++, Java, OCaml, and Go.

PATENTS

- Physiological predictions using machine learning. Applicant: Apple Inc. Inventors: **A. Nazaret** et.al. 03/2024

FIRST AUTHOR PUBLICATIONS (CO-FIRST INDICATED BY *)

[1] N. Beltran-Velez*, A. A. Grande*, **A. Nazaret***, A. Kucukelbir, D. Blei. Treeffuser: Probabilistic Predictions via Conditional Diffusions with Gradient-Boosted Trees. *NeurIPS 2024*

[2] C. Shi*, N. Beltran-Velez*, **A. Nazaret***, C. Zheng*, A. Garriga-Alonso, A. Jesson, M. Makar, D. Blei. Hypothesis testing the circuit hypothesis in LLMs. (earlier version at ICML 2024 MI Workshop – **3rd best paper**) *NeurIPS 2024*

[3] **A. Nazaret**, D. Blei. Extremely Greedy Equivalence Search. *(Spotlight) UAI 2024*

[4] **A. Nazaret***, J. Hong*, E. Azizi, D. Blei. Stable Differentiable Causal Discovery. *ICML 2024*

[5] J. Fan*, **A. Nazaret***, E. Azizi. A thousand and one tumors: the promise of AI for cancer biology. *Nature Methods, 2024*

[6] **A. Nazaret***, J. Fan*, V. Lavallée, D. Pe'er, E. Azizi. Deep generative modeling for mapping derailed trajectories in Acute Myeloid Leukemia. *under review at Genome Biology, bioRxiv preprint*

[7] **A. Nazaret**, C. Shi, D. Blei. On the Misspecification of Linear Assumptions in Synthetic Control. *(Oral) AISTATS 2024*

[8] S. He*, Y. Jin*, **A. Nazaret***, L. Shi, X. Chen, S. Rampersaud, E. Azizi *et.al.* Starfysh integrates spatial transcriptomic and histologic data to reveal heterogeneous tumor–immune hubs. *Nature Biotechnology, 2024*

[9] **A. Nazaret**, S. Tonekaboni, G. Darnell, S. Ren, G. Sapiro, A. Miller. Modeling Heart Rate Response to Exercise with Wearables Data. *Nature Digital Medicine, 2023*

[10] **A. Nazaret** and G. Sapiro. A large-scale observational study of the causal effects of a behavioral health nudge. *Science Advances, 2023*

[11] **A. Nazaret**, J. Fan, D. Pe'er, E. Azizi. Probabilistic basis decomposition for characterizing temporal dynamics of gene expression. **Best poster award** – *Workshop on Computational Biology, ICML 2022*

[12] **A. Nazaret**, D. Blei. Variational Inference for Infinitely Deep Neural Networks. *(Spotlight) ICML 2022*

[13] R. Lopez*, **A. Nazaret***, M. Langevin, J. Samaran, J. Regier, M. I Jordan, N. Yosef. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. **Best poster award** – *Workshop on Computational Biology, ICML 2019*

OTHER PUBLICATIONS

[14] A. Gayoso, R. Lopez, G. Xing, P. Boyeau, J. Hong, K. Wu, M. Jayasuriya, E. Mehlman, M. Langevin, Y. Liu, J. Samaran, G. Misrachi, **A. Nazaret**, O. Clivio, C. Xu, T. Ashuach, M. Gabitto, M. Lotfollahi, V. Svensson, E. Beltrame, V. Kleshchevnikov, C. Talavera-López, L. Pachter, F. J Theis, A. Streets, M. I Jordan, J. Regier, N. Yosef. A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology, 2022*

[15] K. Choromanski, D. Cheikhi, J. Davis, V. Likhosherstov, **A. Nazaret**, A. Bahamou, X. Song, M. Akarte, J. Parker-Holder, J. Bergquist, Y. Gao, A. Pacchiano, T. Sarlos, A. Weller, V. Sindhwani. Stochastic flows and geometric optimization on the orthogonal group. *ICML 2020*

[16] M. Chevalley, J. Sackett-Sanders, Y. Roohani, P. Notin, A. Bakulin, D. Brzezinski, K. Deng, Y. Guan, J. Hong, M. Ibrahim, W. Kotlowski, M. Kowiel, P. Misiakos, **A. Nazaret**, M. Püschel, C. Wendler, A. Mehrjou, P. Schwab, 2023. The CausalBench challenge: A machine learning contest for gene network inference from single-cell perturbation data. *in preparation for submission, arXiv preprint*

OPEN SOURCE SOFTWARE

- `Treeffuser`: An easy-to-use package for probabilistic prediction on tabular data with tree-based diffusion models.
- `SDCD`: A method for inferring causal graphs from labeled interventional data.
- `scvi-tools`: A library for analyzing single-cell data with deep generative models.