Patterns

CellPress
OPEN ACCESS

People of Data

Data science through the lens of systems immunology

David R. Glass[1,2,3,*] and Meelad Amouzgar[1,2]
[1]Department of Pathology, Stanford University, Stanford, CA, USA
[2]Immunology Graduate Program, Stanford University, Stanford, CA, USA
[3]Present address: Vaccine and Infectious Diseases Division, Fred Hutchinson Cancer Center, Seattle, WA, USA
*Correspondence: dglass@fredhutch.org
https://doi.org/10.1016/j.patter.2022.100574

Glass, a post-doctoral researcher, and Amouzgar, a PhD student, in Bendall lab proposed a supervised dimensionality reduction method to explore and analyze single-cell data. Their *Patterns* paper highlights the advantages of supervised learning in single-cell datasets with class labels. They talk about the essential role of data science in this project and in their lives.

What would you like to share about your background (personal and/or professional)?

David R. Glass: My first degree was a bachelor of music in sound recording technology. I spent several years playing in bands, producing albums, and building audio electronics gear for recording studios before returning to school for a second bachelor's in biology. I then joined Stanford's Computational and Systems Immunology PhD program, where I was trained in immunological sciences as well as bioinformatics and data science.

Meelad Amouzgar: I didn't know I wanted to be a scientist growing up, but now I can't imagine myself doing anything else. While my education started in the biosciences, I knew I'd found my niche when I discovered the world of bioinformatics and systems biology through an MS in informatics. As I realized my love for research, I also discovered an equal love for exercise, especially calisthenics and gymnastics. Whether I'm in the lab or walking on my hands, I often daydream about finding new and thrilling ways to integrate my love for data science, human immunology, and athletics— perhaps even turning it into a career!

How did this project you wrote about come to be?

DRG: I came up with the idea to use HSS-LDA for single-cell projection while collaborating with a pathologist to come up with a new diagnostic approach for blood cancers. In that paper, we wanted to empower cytometers to "see" cells, the way that a pathologist does.[1] We therefore quantified cell morphological features by their underlying, antibody-measurable molecular components. I was challenged with finding a reproducible, non-stochastic, 2D visualization scheme that would incorporate all of this information into "morphometric maps" that pathologists could understand and interpret. After completing the project, we realized that HSS-LDA had considerable utility for a variety of biological applications and technologies, so we fully explored the topic in this new manuscript.[2]

Who were the driving forces behind the project?

DRG: While I launched the idea, my co-first author Meelad Amouzgar was the real driving force for the current paper.[2] Meelad rotated in Sean Bendall's group as I was wrapping up my PhD. He came into the group as an already accomplished computational biologist and was very excited and motivated by the project. I mentored Meelad through that rotation, and we continued working on the project together after he officially joined the lab. Initially, I was pretty hands-on about the datasets, visualizations, and benchmarking approaches we should take, but Meelad was so talented that my instruction became less and less necessary throughout the project. In the end, he modified my original code for HSS-LDA, and then generated all figures and performed all analyses himself.

What drew you to your current team and topic?

MA: The manuscript began as a rotation project in Sean Bendall's lab, with David as my rotation mentor. I was drawn to the Bendall lab for my PhD because I could envision myself engaging in similar science that integrates experimental and computational techniques, but my instincts also told me that Sean would cheer me on and have my best interests in mind as my advisor. David cemented my initial impression of the lab during the rotation by being a fair, thoughtful, rigorous, and easygoing rotation mentor. Developing this manuscript with a talented scientist like David was an excellent experience that I'm glad I had at the start of grad school as it taught me many soft skills for future collaborations. This particular rotation project appealed to me for three reasons: (1) I love dimensionality reduction, (2) supervised dimensionality reduction has received little attention with single-cell data, and (3) there is an amazing wave of computationally intensive algorithms and deep learning applications being published with single-cell data, but simple and fast algorithms still offer significant utility to single-cell data analysis that shouldn't be forgotten. I'm curious about the future of supervised dimensionality reduction in single-cell data beyond our manuscript.[2]

Was there a particular result that surprised you, or did you have a eureka moment? How did you react?

DRG: Generally, the earlier parts of the manuscript lean on my ideas, while the latter parts are more composed of Meelad's creative efforts. I was really floored when Meelad showed me the manifold of cell-cycle scores from scRNA-seq found in Figure 6 of our paper.[2] It was such a perfect circular visualization of the continuity of cell cycle, all derived

from discrete labels! Then, when this concept was paired with cell division ID, we were able to explore gene expression changes over the course of multiple cycles of cell division, something I don't believe anyone has done before. I thought it was such an incredible application of LDA, and I was astounded by such a brilliant idea.

Why did you decide to publish in *Patterns*?

MA: *Patterns* reaches a broad audience across multiple domains that apply data science. While our manuscript was biology focused, much of our work extends beyond the biology and into data science.[2] This makes *Patterns* a great fit for the scope of our manuscript.

DRG: All of my research has been collaborative and interdisciplinary. I've had the privilege of publishing in journals specializing in biomedical research, immunology, and biotechnology advances. I'm excited to add a data science journal to that list; I think it demonstrates that by working with a diversity of individuals with different backgrounds and expertise, you can have an impact on many scientific fields.

What is the definition of data science in your opinion? What is a data scientist? Do you self-identify as one?

DRG: I'll defer to The Oracle from the *Matrix*: "*Temet Nosce*. You know what that means? It's Latin. Means, 'Know Thyself.' I'm gonna let you in on a little secret. Being [a data scientist] is just like being in love. Nobody can tell you you're in love. You just know it. Through and through. Balls to bones."

MA: There are many formal definitions of data science out there, but data science to me is detective work. Like a detective, a data scientist gathers leads and searches for clues in the data, then extracts and shares meaningful insights that build to a robust conclusion presented to a jury. For a data scientist, the "crime scene" is the data, and the leads are domain knowledge. Using all the tools at their disposal, the data scientist weaves the clues in the data into a story that is presented to a jury—for us, colleagues in our labs, scientists at conferences, or the peer review process. The Eureka! moments that lead to a clue in the data have come to me while I'm cooking, cleaning, driving, hiking, exercising, or even right before I fall asleep. Taking those ideas and seeing the implementation work as you envisioned them feels like striking gold. I think loving the search for those glimmers of inspiration is a core quality of what it means to be a data scientist, and for that reason I self-identify as a data scientist.

What motivated you to become a (data) researcher? Is there anyone/anything that helped guide you on your path?

DRG: The revolution of "big data" technologies in biosciences has made fluency in high-dimensional data analysis essential for many biologists. When I entered my PhD program in 2015, I saw the writing on the wall. Forward progress was being accelerated by researchers utilizing high-dimensional cytometry, imaging, and sequencing to understand biology at its most fundamental organizational unit: the single cell. Wet lab biologists should take coursework and/or get trained in basic computer science, statistical learning, and computational biology. Likewise, there is a tremendous opportunity for data scientists to apply their skills to answer questions in biology and medicine. I'd recommend interested data scientists seek out collaborations and coursework to enrich their understanding of biology.

MA: Data visualization quickly became one of my favorite parts of data science. To find how I could best communicate and share my results with my colleagues, I would scour the internet reading data science articles. My enthusiasm for studying data, identifying patterns, and storytelling through biological data motivated me to become a data researcher. A significant portion of my interest in research stems from my bioinformatics mentors and colleagues who trained me during my master's degree, and those during my first job after graduating. Having those early-career cheerleaders and seeing them create great futures for themselves helped me envision that for myself.

What barriers have you faced in pursuing data science as a career?

DRG: Imposter syndrome is a huge barrier for biologists attempting to learn data science. The jargon, notation, and approaches are very foreign to those of us with biology degrees. I think many biologists see these unfamiliar things and assume that if they were capable of learning that skill base, they already would have done so. Biologists often undervalue the hard work they went through to become masters at their own craft and put quantitative skills on a pedestal. As someone who began his PhD at age 30, I had moments where I felt like I was too far behind to catch up and too old to start over. Yet, data science is no different from any other skill. If you see an expert in the field, it's probably not because they are a genius. They're probably just another person who has been practicing their craft for many more years than you. It's not beyond your reach if you can dedicate the time.

Which of the current trends in data science seem most interesting to you? In your opinion, what are the most pressing questions for the data science community?

MA: I'm currently very interested in single-cell trajectory inference. Many elegant algorithms have been developed over the last few years, so it's not the newest trend, but I still think it's wild that unsupervised or semi-supervised methods can reconstruct dynamic cellular processes in high-dimensional data. In our manuscript, we highlight the value of using algorithms that leverage prior knowledge to understand single-cell data.[2] High-throughput experiments are becoming more complex, so accurate prior information may help us better organize and understand the massive amounts of data being generated. As the experimental and computational worlds of system biology become more interwoven, generating computational tools that can better leverage prior information may become more valuable. The current paradigm focuses on developing computational tools that serve experiments, but I'd love to see experiments designed to explicitly leverage supervised algorithms for biological discovery.

How do you keep up to date with advances in both data science techniques and in your field/domain?

DRG: While I get tables of contents sent to my inbox from many journals, e-mail is such a slog that I'm usually not in a good mindset to engage with new scientific ideas. Instead, I rely heavily on my

CellPress
OPEN ACCESS

David R. Glass (left) and Meelad Amouzgar (right) of the Bendall lab (bottom)

network of colleagues. Following scientists on Twitter is a great way to keep up with the latest publications and preprints. My lab also posts relevant papers on Slack. Additionally, scientific conferences are a fantastic resource to see a snapshot of a field and to connect with new people and potential collaborators.

MA: PubMed, preprint journals like bioRxiv and arXiv, conferences, my lab posting papers on Slack, and Twitter are my usual resources for updates in biological data science. But I also just love the pursuit of knowledge in a topic I am highly interested in, so I let my random Google searches lead the way. Between all of that, and the fact that the internet on my phone has a way of knowing what I need to know before I know it myself, I find I'm usually kept up to date.

Have you ever used your data science skills in your personal life? If yes, how?

MA: During the COVID-19 pandemic, when exercise facilities were closed, I wanted some home workouts with minimal equipment and found some helpful free, online resources with daily workouts. I wrote a program to recursively download these web pages, remove the excess information, and convert it into a pdf. The final product was a directory with 1,000+ daily workouts organized by date. It was simple but worked very well, and it helped me manage my physical and mental well-being during the lockdown. Another time, for bonding and good fun during a lab retreat, I continued our lab's tradition of

creating an "Intro to Machine Learning" presentation. I mandated our lab members to take a personality test with multiple continuous features and ran a machine learning analysis on the similar/dissimilar personalities using dimensionality reduction, clustering, and other data science tools. It wasn't very educational since we couldn't get through it without laughing at the results. I'd recommend it for every lab.

What is the fun part of being a data scientist?

DRG: I really like the exploratory nature of the work. You can divide the data a thousand different ways and visualize it from all sorts of angles. You can find correlations, build models, or even create pseudotime trajectories. There is an old adage that a work of art is never completed, only abandoned. I think the same applies for data science. The goals of any given project are often nebulous, so sometimes you're seeking both the questions and the answers during the analysis. It's fertile ground for a creative spirit.

REFERENCES

1. Tsai, A.G., Glass, D.R., Juntilla, M., Hartmann, F.J., Oak, J.S., Fernandez-Pol, S., Ohgami, R.S., and Bendall, S.C. (2020). Multiplexed single-cell morphometry for hematopathology diagnostics. Nat. Med. *26*, 408–417. https://doi.org/10.1038/s41591-020-0783-x.

2. Amouzgar, M., Glass, D.R., Baskar, R., Averbukh, I., Kimmey, S.C., Tsai, A.G., Hartmann, F.J., and Bendall, S.C. (2022). Supervised dimensionality reduction for exploration of single-cell data by HSS-LDA. Patterns *3*, 100536. https://doi.org/10.1016/j.patter.2022.100536.

About the authors

David R. Glass is a Cancer Research Institute Irvington postdoctoral fellow at the Fred Hutchinson Cancer Center. In the laboratory of Evan Newell, he applies high-dimensional, multi-omic, single-cell technologies to understand the contribution of T cells and B cells to the immune response in the context of cancer immunotherapy and vaccination. He received a PhD in computational and systems immunology from Stanford University, under the supervision of Sean Bendall. Prior to that, he received a BS in cell and molecular biology from the University of Texas at Austin and a BMus in sound recording technology from Texas State University.

Meelad Amouzgar is a PhD student in the Computational and Systems Immunology program at Stanford. He uses high-dimensional single-cell data to study the extrinsic and intrinsic cell properties that control immune cell fate determination. Prior to Stanford, he attended community college, transferred to UC Davis for his BS in biochemistry, and received an MS in bioinformatics from USF. He worked in cancer immunotherapy clinical trial research in the Bay Area and explored careers outside of science. While in the lab, he loves practicing hand balances, consuming coffee and cookies, and spending time with people he cares about.