Patterns

CellPress
OPEN ACCESS

People of Data

Data science, human intelligence, and therapeutics discovery: An interview with Sean Escola, Saul Kato, and Pavan Ramkumar

Pavan Ramkumar,[1] Saul Kato,[1,2,*] and G. Sean Escola[1,3,*]
[1]Herophilus, Inc, San Francisco, CA 94107, USA
[2]Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA 94143, USA
[3]Zuckerman Institute, Department of Psychiatry, Columbia University, New York City, NY 10032, USA
*Correspondence: saul@herophilus.com (S.K.), gse3@columbia.edu (G.S.E.)
https://doi.org/10.1016/j.patter.2022.100490

Sean Escola, Saul Kato, and Pavan Ramkumar explain the importance of data science in their research. They have developed a simple non-parametric statistical method called the Rank-to-Group (RTG) score that identifies hierarchical confounder effects in raw data and machine learning-derived data embeddings. This approach should be generally useful in experiment-analysis cycles and to ensure confounder robustness in machine learning models.

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

Sean Escola: I'm driven to understand intelligence: how the brain computes the functions that permit the range of human experience and behavior. Our very humanity and culture—from the emotions felt at hearing a child laugh to great works of art and literature—are computable functions of the human mind. Understanding this function space will necessitate (1) collecting large datasets from multiple lines of inquiry across neuroscience, psychology, and other fields of research; (2) using the tools of machine learning to integrate and characterize those data and their interactions; and (3) developing computational and mathematical frameworks within which quantitative theories that describe neural computation can be formulated.

It is also my strong belief that therapeutics discovery for neurological and psychiatric disorders will need an understanding of the computational nature of neural processing that is similar in richness to that needed to understand neural computation itself. That is because symptoms—hallucinations in schizophrenia, memory deficits in Alzheimer's, social deficits in autism—are fundamentally expressions of faulty neural computation.

These two professional drives—the understanding of neural computation for the sake of understanding itself and for the sake of the development of novel, more efficacious therapeutics—have manifested in my career having two different prongs that nonetheless share the foundational core described above. Along the first prong, I have an academic lab at Columbia University where my group builds computational models and collaborates with experimentalists to analyze data in the service of investigating the brain circuits involved in motor function and learning. Along the second, I have co-founded a company, Herophilus, which uses human stem cell-derived neural organoids as a platform for phenotyping disease states and testing pharmaceuticals.

Organoids are incredible in terms of the richness of the biology that they exhibit: they grow to about 1–2 mm in diameter, contain a large complement of neural cell types including excitatory and inhibitory neurons as well as astrocytes, assemble in layers that are molecularly identifiable as the layers of cerebral cortex, and of greatest interest to me, their neurons wire up into networks and are active. This final point means that it may be possible to identify computational disease phenotypes by comparing the neural activity in organoids that are cultured from stem cells derived from patients with particular diseases to healthy controls. However, this promise can only be realized through the collection of large datasets and the application of sophisticated computational tools that permit the characterization of these complex data.

Saul Kato: I'm a computer scientist turned physicist turned neuroscientist. I grew up programming computers and wondering why we couldn't emulate human intelligence very well. In third grade, I designed my first software app AutoTessellator, a drawing program to make Escher-like plane tilings, written in LOGO on an Apple II. Then I fell in love with physics as a quest for ground-truth understanding of the universe. But I was distracted by Silicon Valley and founded two tech startups over 10 years, before going back to get a PhD in neurobiology to study the deep mystery of the power of biological brains—another understand-the-universe question. I started a neuroscience lab at the University of California, San Francisco, and shortly thereafter co-founded Herophilus because we saw an opportunity—an obligation, really—to leverage recent technology breakthroughs to do drug discovery "right."

Pavan Ramkumar: I apply computational methods to a wide range of biological areas, develop software to do so scalably, and drive discovery and decision making as a consequence. Before drug discovery, I have worked in cybersecurity risk modeling and systems neuroscience.

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

SE: My father is an electrical engineer who always encouraged me to study computer

science. I also discovered a love for chemistry at the age of about 12. These two paths converged in my study of biochemistry and computer science as an undergraduate, followed by MD-PhD degrees with a thesis in statistical modeling of neural network activity.

SK: In high school I won Super-Quest, a wonderful Westinghouse/Intel/Regeneron-like research contest in the 90s that awarded summer stints at NSF supercomputing centers and free supercomputer time to high school students to use for their projects. My project was to compute, from physics first principles, the fastest strategy and path for a racecar to take given an arbitrary course. That got me hooked on using computers—big, heavy metal machines—for research.

PK: I have had several influences, most notably my PhD and postdoc advisors but also several peers in former and current teams I work in. Data science was a natural transition to the applied work I had been doing throughout my academic career as a computational scientist focused on developing methods and tools to interpret experimental data in neuroscience.

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

SE: All scientists are data scientists. We live in a golden age of experimentation where tera- and peta-scale datasets are the norm, not the exception. For example, recording many thousands of neurons simultaneously across multiple brain regions at sub-millisecond resolution in behaving animals is becoming routine. Historically in neuroscience, a relatively rarified group of individuals had high-quality computational skills—i.e., ability to code complex algorithms and the good practices needed to generate shareable software. These individuals would then collaborate with experimentalists to assist in the interpretation of data. This pattern was very much in place when I began my career in neuroscience during my PhD. However, in the past 20 years, this picture has changed substantially. Almost all of our incoming students have formal training in programming. This obviates the split between experimentalists and data scientists and will accelerate understanding especially in the setting of massive modern datasets. This is true across science, not just for neuroscience.

SK: Data science is the methodological field of using complex and large datasets to derive human insight and knowledge about the world. Across virtually all fields of science, it is becoming synonymous with simply "scientist." I am perfectly happy to call myself a data scientist.

PK: Data science is an approximately 15-year-old discipline of practice at the confluence of statistics, computer science, software engineering, and machine learning. Data science needs to exist as a distinct discipline from its antecedents because of the specific ways in which it combines them. Yes, I do identify as a data scientist.

Why did you decide to publish in *Patterns*?

SE: We wanted our story to get the treatment that journal publication permits, rather than limiting it to the framework of a conference proceeding. While we considered publishing in a biological journal, our paper is fundamentally about a statistical technique that can be applied in any domain, not just in biology. We ultimately chose *Patterns* because it provides the best of both worlds. As a member of the Cell Press family, *Patterns* allowed us to remain close to the biology community that we are part of, and as a data science journal, *Patterns* afforded us the opportunity to discuss a computational approach in detail.

SK: Cell Press journals have a reputation for rigorous, cutting-edge, high-impact life science. We wanted to support the recognition of data science for biology as a major field in its own right.

PK: *Patterns* is a new journal with the mission to bring computational methods and tools to the Cell Press readership community, an ideal venue for this work.

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

SE: At Columbia we have a computational neuroscience weekly seminar. Many of the talks we host relate to advances in machine learning. This is one of the advantages of being at a top institution that is not shared by many. In the wake of COVID-19, it has been very exciting to see the development of high-quality virtual seminar series such as WorldWide-Neuro that provide a similar opportunity on a global scale.

PK: I take a top-down approach. I go looking for ideas to solve problems I face in the domains I work in rather than trying to stay up to date with methods and tools, which is increasingly impossible. I learn from peers who take a bottom-up approach. I also like the Feynman method: the intersection of a bag of 10 open problems and 10 tools will eventually result in a novel application.

In your opinion, what are the most pressing questions for the data science community?

SE: The most pressing question, in my opinion, is understanding why many tasks that are so easy for humans and other animals to do remain out of reach for machine learning models and robots. The human is an existence proof for artificial general intelligence (AI). I believe that the coming years will see artificial intelligence researchers turn back to neuroscience—as was the case when AI research was in its infancy—to gain insights on how to improve models.

What is the role of data science in your domain/field? What advancements do you expect in data science in this field over the next 2–3years?

SE: In my field, data science is necessary for everything that we do. Datasets in biology are too big to be interpreted without complex tools. Furthermore, simple analyses by necessity yield highly reductionistic views of whatever process is being studied. For the kinds of subtle characterizations of data that will enable us to gain deep insight into the nature of neural computation and disease processes, we need to integrate, not reduce. This can only be done with good data science.

PK: Biology, particularly cell and molecular biology for disease modeling and therapeutics discovery is experiencing a rapid industrialization. Therefore, the rich datasets in the domain are scaling rapidly and require a concomitant scaling of data science (algorithms, software architecture, computer and data infrastructure). The ongoing transition has opened up previously unsolved problems for data

scientists. For instance, multi-channel ultra-high-resolution imaging is the norm for several kinds of immunofluorescence, but the state of art in machine vision has primarily been developed for natural images from commodity cameras. The field is therefore inventing new standards for microscopy, adapting or reinventing machine vision for ultra-high-resolution imaging, both for data management and computation. I believe these advances will converge toward industry standards in the next 2–3 years.

What is the fun part of being a data scientist?

SE: Implementing algorithms in code has always been a thrill for me going back to my undergraduate days. In machine learning, it's often useful to test code on artificial data, and the first time I see that an algorithm I implemented works as expected is a big dopamine releaser for me. Of course, then you try it on real data and nothing works.

PK: Data science is the central function of organizations interested in seeking truth and making informed decisions. Iteratively going from ambiguity to clarity by bringing together dozens of techniques and disciplines is the most exhilarating aspect of being a data scientist.

What is your advice for future data scientists?

SK: Always scrutinize the quality and relevance of your data. "Garbage in, garbage out" is an axiom in the age of big data.

How did this project come to be?

SE: This project arose out of necessity. At Herophilus we culture hundreds of thousands of organoids from ~1,000 human donors, with ~10 clones per donor, across many dozens of experimental batches. As we scaled up, we immediately realized that we needed to answer the question: how can we meaningfully compare all these data in light of the variability introduced by factors such as donor, clone, and batch swamping the signal? The challenge is that certain factors are hierarchically organized—e.g., many clones can come from the same donor—obviating many existing approaches for the assessment of bias. The technique we present in our article—the rank-to-group (RTG) score—is what we developed to handle this issue.[1]

Who were the driving forces behind it?

SE: This was a real team effort. The first author, Alex Rogozhnikov, had the initial idea; Rishi Bedi and Pavan Ramkumar developed the code and applied it to multiple datasets; I did the analytics; and we all conceptualized and wrote the paper. And our biologists "beta-tested" the heck out of it.

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

SE: A big surprise for me was how noise robust RTG scoring is compared to linear methods for example. This is explicitly shown in Figure 3 and is also clear by looking at the confidence intervals in Figures 5 and 6. However, in retrospect this makes sense because noise is most likely to affect the ordering of data points that are far from the "query" and therefore don't affect the RTG score (see Figure 1 for an illustration of the algorithm). This helped me understand the strength of rank-based statistics as opposed to those defined by continuous measures.

How did you celebrate acceptance?

A ton of Slack emojis!

What drew you to this area of research? How has the research focus of your team evolved over the years?

SE: In my lab at Columbia, we study the brain computations underlying motor function and learning. This research program grew out of my graduate studies where I developed a modeling framework—the generalized linear model/hidden Markov model—that permits the inference of state switches in neural data and thus allows for the study of sequencing. Motor behaviors can be thought of as sequences of primitive motor elements. We primarily focus on the structures known as the basal ganglia, motor thalamus, and motor cortex and construct models of how these regions interact and learn in order to explain features in the experimental data of colleagues who probe these systems in animals. One way in which our work has evolved over time is that we are beginning to look at how our models might dysfunction and thereby provide hints to disease phenomenology.

At Herophilus, the focus of the machine learning group, led by Pavan Ramkumar, is to infer phenotypes of disease from the large datasets we collect from our organoids. These datasets include transcriptomic and proteomic data, cellular- and tissue-level imaging including whole cleared organoids, and neural recordings. While phenotyping remains the focus of the group, we have, out of necessity, also developed many tools for the quantitative assessment of data quality. It was this thrust of our work that led to the research presented in our *Patterns* article.[1]

Where is the team based currently, and how long have you been there?

SE: The Herophilus machine learning team is composed of computational scientists of various backgrounds including computer science, neuroscience, physics, and bioengineering. Importantly, members of this team are embedded in every experimental effort at the company from conception and planning to experimentation and interpretation of results. This allows our science to have a data-first ethos that optimizes for reliability and replicability in our results.

What kind of atmosphere do you look to foster in your team? Is there anything you try to replicate or avoid from your own experiences or that you have learnt over the years?

SE: Collaboration is key and ideally that collaboration begins prior to experimental design. I have seen this to be true in both my research at Columbia and the research at Herophilus. When data scientists are introduced later in a project, it is inevitable that there is an experimental tweak or additional control whose importance becomes apparent.

Which achievement/discovery in your career are you most proud of?

SE: Last year, my postdoc Laureline Logiaco published an article in *Cell Reports*[2] that was a culmination of several years of work. I like this article in particular because we are able to offer a theoretical understanding of a computational problem that must be solved as part of motor sequence generation. Excitingly, several features of our model have been or are being followed up on with experiments.

A lot of data scientists continue their career outside of academia; what is your view on that? Do you encourage your students and postdocs to continue their careers in academia and establish their own teams? Are you supportive of careers outside of academia?

SE: My own career is both within and outside of academia. I believe it is a huge disservice to our incredibly talented students and postdocs to perpetuate the conceit that a job in industry is some kind of failure. Happily, the success of Deepmind, Meta Control Labs, and Herophilus have begun to attenuate that pejorative.

REFERENCES

1. Rogozhnikov, A., Ramkumar, P., Bedi, R., Kato, S., and Escola, G.S. (2022). Hierarchical confounder discovery in the experiment-machine learning cycle. Patterns *3*, 100451.

2. Logiaco, L., Abbott, L.F., and Escola, S. (2021). Thalamic control of cortical dynamics in a model of flexible motor sequencing. Cell Rep. *35*, 109090. https://doi.org/10.1016/j.celrep.2021.109090.

About the authors

Saul Kato is co-founder and CEO of Herophilus. He is an assistant professor at the Weill Institute for Neurosciences at UCSF where he heads the Foundations of Cognition lab. He has a background in computer science, physics, and neurobiology, holding a PhD in neurobiology from Columbia University and an MS in electrical engineering and BS in physics from Stanford. Before moving to the study of the brain, he was founder of Sven Technologies, a 3D graphics software company acquired by Dassault Systemes, and WideRay Corp., a wireless infrastructure company acquired by Dimensional Associates.

G. Sean Escola is an assistant professor in the Center for Theoretical Neuroscience and the Department of Psychiatry at Columbia University. His research, funded by the NIH and NSF, studies circuit models of motor function and learning. He is a co-founder of Herophilus, where he is involved in multiple aspects of the company's scientific program. As a psychiatrist, Dr. Escola sees patients in the outpatient and emergency room settings. Dr. Escola received his undergraduate degrees in biochemistry and computer science, master's in computer science, and MD-PhD degrees with a thesis in computational neuroscience, all at Columbia. He completed a residency in clinical psychiatry also at Columbia.

Pavan Ramkumar is director of machine learning at Herophilus. He has a background in building data science organizations, production machine learning systems, and tech-enabled biological discovery. He holds a PhD in neuroscience, signal processing, and machine learning from Aalto University, an MSc in bioinformatics from TKK in Finland, and a B. Tech from the Indian Institute of Technology, Guwahati.