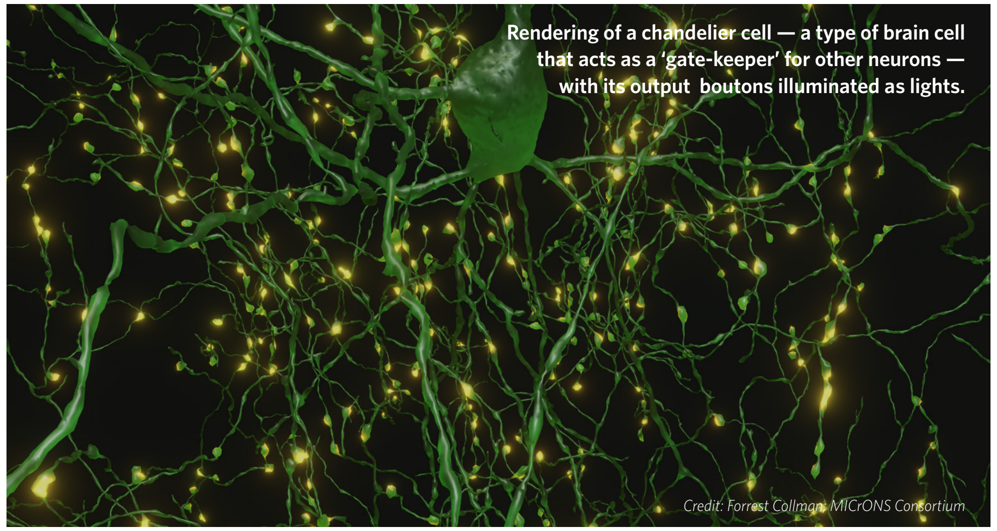


Rendering of a chandelier cell — a type of brain cell that acts as a 'gate-keeper' for other neurons — with its output boutons illuminated as lights.

Credit: Forrest Collman, MICrONS Consortium

# A BETTER DATA ENGINE FOR BRAIN SCIENCE

THIS COMPUTATIONAL PLATFORM is helping labs manage complex data pipelines and laying the foundation for AI-driven discoveries in neuroscience and beyond.

**Scientific research is often celebrated** for its creative chaos — improvisation in the lab, trial-and-error in the field, the occasional serendipitous breakthrough. But as experiments become increasingly data-heavy, with reams of complex inputs and computational analyses, researchers are struggling to scale their operations while maintaining rigour and reproducibility.

A new proposal, outlined in a recent preprint, aims to change that. By adapting best practices from software engineering to scientific research, a coalition of academic and industry partners hopes to streamline data collection, standardize analyses and accelerate discovery — with structured workflows seen as key to this effort.

"The goal, fundamentally, is to produce the same productivity increase in science that we have seen in other disciplines," says Dimitri Yatsenko, a lead architect of the new 'SciOps' framework.

The concept takes inspiration from DevOps, which transformed software development in the 2000s by empowering teams with automated workflows, continuous testing, and seamless code integration. DevOps made it possible to create entirely new industries like cloud computing, e-commerce and streaming.

Academic research needs a similar shift, says Yatsenko, and he has created a data operations platform for scientific laboratories to help make that happen. Known as DataJoint, the platform replaces fragmented data handling processes with an end-to-end computational workflow for data entry, acquisition, processing, analysis and visualization — all unified in a coherent pipeline that unlocks collaboration and integration with artificial intelligence (AI) capabilities.

"A key challenge in data science is to combine computations with data management," notes Yatsenko, founder and chief scientist of a company that shares the platform's name. "DataJoint handles that as a single problem," he says.

Available as an open-source general framework, DataJoint is adaptable for any research discipline, from genomics to climate science. But to date, it has had the most impact in systems-level neuroscience, where data integrity and reproducibility are persistent challenges.

**Supporting MICrONS**

In 2016, the platform was selected as the data backbone for the neurophysiology component of the Machine Intelligence from Cortical Networks (MICrONS) project — a five-year, US$100-million initiative funded by the US Intelligence Advanced Research Projects Activity to map the detailed structure and activity of neural circuits in the mouse brain. DataJoint's adoption helped establish its value in managing large-scale, multimodal neuroscience data.

With an eye toward informing next-generation machine learning models that 'think' like brains, the MICrONS team collected petabytes of data from electron microscopy (to map synaptic connections), calcium imaging (to track neural activity) and behavioural studies (to understand functional responses) in mice.

Another DataJoint pipeline integrated the functional data with structural data from systems like CAVE — which, as reported in Nature, enabled scientists to start teasing apart the computational principles that underlie the function and connectivity of the mammalian brain.

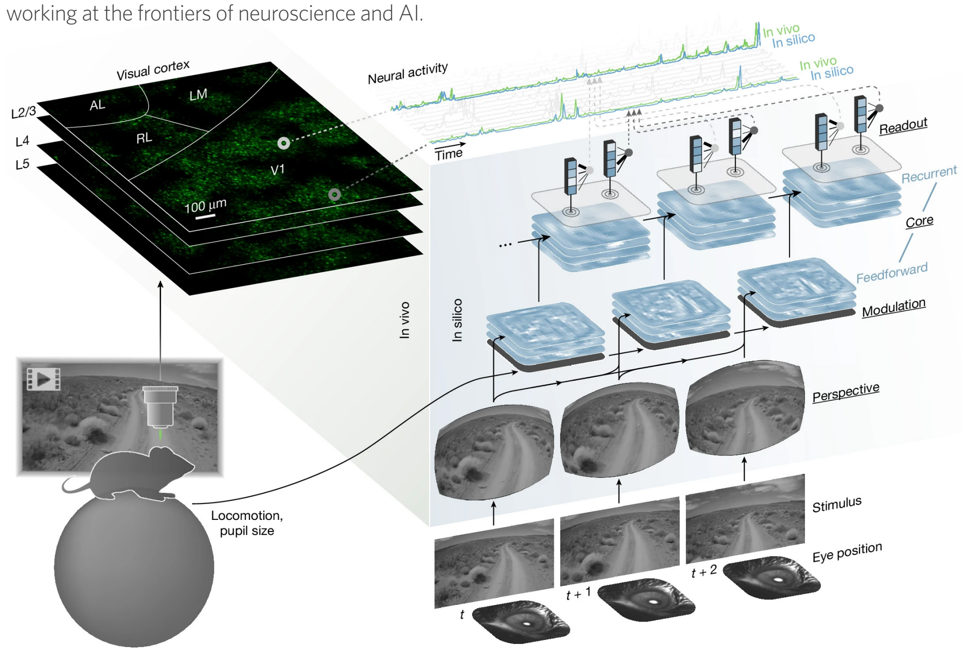
Finishing the job won't be easy. But with much of the data available via DataJoint and the MICrONS Explorer, researchers can continue to explore new questions about neural computation and circuit dynamics. "I think the MICrONS data will be studied for the next decade or more," Yatsenko says.

**'Just essential'**

Beyond MICrONS, over 100 neuroscience labs worldwide use the platform for data pipelines in studies on neurodegenerative disease, neurotransmitter signalling, visual processing, and stroke recovery. Several large, multi-institutional projects, focused on understanding cognition, olfaction, and behaviour, also rely on DataJoint.

"LABS MUST INVEST IN THEIR DATA OPERATIONS TO STAY COMPETITIVE, GIVEN THE GROWING IMPORTANCE OF AI."

**DATAJOINT ENABLED THE MICrONS TEAM** to train a new foundation model by capturing the brain in unprecedented detail. With NIH support, DataJoint has translated the system that powered MICrONS into a scalable platform — bringing the same robust data infrastructure to any lab working at the frontiers of neuroscience and AI.



*In vivo* recorded data on inputs (visual stimulus, eye position, locomotion, and pupil size) and outputs (neural activity) trains an artificial neural network model to generate *in silico* responses. See *Foundation model of neural activity predicts response to new stimulus types*, Figure 1.

Before adopting the platform, researchers often relied on convoluted file naming conventions, custom-written code and incompatible formats that complicated data handling. DataJoint eliminates that chaos by centralizing data management, automating analyses and ensuring consistency — and traceability — across experiments.

"I'm so enthusiastic about it," says Jacob Reimer, assistant professor of neuroscience at Baylor College of Medicine (BCM) in Houston, who was an early adopter of the platform there. "Everything we do is now immediately accessible and easily located," he adds. "There's no other place the data can be."

Reimer had a front-row seat to DataJoint's early development. As a postdoc working in the same BCM lab, Reimer watched Yatsenko, then a graduate student, spend his free time, between imaging experiments of calcium signals in the mouse brain, working on building a better way to manage the growing data deluge.

At first, most colleagues ignored or outright dismissed Yatsenko's effort. "Nobody was convinced it was worthwhile," says Reimer. But then Yatsenko unveiled an early prototype, and, quickly, recalls Reimer, "it became, like, just essential."

The entire lab soon implemented the platform. So too did collaborators from California and Germany. And by 2016, as interest spread, Yatsenko, together with Reimer and two other members of the same BCM lab group, decided to the DataJoint company — of which Reimer remains a shareholder — to scale the platform for broader use.

**Expanding reach**

More labs continue to adopt DataJoint to manage and refine their own data operations, with the company's services and capabilities growing in response to demand.

"Labs must invest in their data operations to stay competitive, given the growing importance of AI and advanced data science techniques," says Marshall Hussain Shuler, a neuroscientist at Johns Hopkins University School of Medicine in Baltimore, Maryland. His lab gathers mounds of microscopy and video imaging to study the electrophysiology and behaviour of neural circuits in mice, aiming to understand how experiences shape sensory processing and decision-making. "Making a platform that does all that in one place is a tall order," he says. DataJoint helps keep it all organized and interpretable, creating what he calls a "lab memory".

"We have already saved months and ran experiments that we never could before," Hussain Shuler says. "DataJoint allows us to create a formal structure for our work that can be understood, extended, and reused."

That structure is helping the team behind Project Aeon, led by the Sainsbury Wellcome Centre at University College London, to better understand the neural basis of natural mouse behaviours, such as foraging, escaping, nesting, and social interactions over naturalistic timescales. The project involves continuous monitoring of mice in large habitats, via dozens of cameras and sensors, for weeks to months, producing high-dimensional quantifications of their behavioural repertoire including pose, position, and identity. "Such in-depth description of mouse behavior, combined with prolonged ephys recordings from implanted arrays, generates very large and complex datasets that are hard to handle," says Dario Campagner, Project Lead Scientist. "DataJoint architecture enables fast and intuitive data querying and provides an easy way to standardize the data format for sharing with researchers all over the world."

Scrutinizing all that data is a task well suited for AI — and these tools require that every detail of an experiment be structured and recorded in order to generate meaningful insights. Yatsenko says the latest advances in the DataJoint platform unify an experiment's design, code, and data — creating a structure that both humans and AI systems can interpret. "For labs seeking rigour, reproducibility, and AI readiness, DataJoint offers a path to organize complexity and unlock the full value of their data." ■