DataJoint

June 11, 2025

neuromatch

Building Data Workflows for Neuroscience and AI

Infrastructure × Operations × Collaboration

Dimitri Yatsenko, PhD
Thinh Nguyen, PhD

Copyright 2024 DataJoint Inc. All Rights Reserved.

The Future of Research Operations

operational excellence - data integrity - governance - quality control - AI readiness

DataJoint Platform - computational workflows, future developments

DataJoint Elements - standardized workflows for neurophysiology

Data Publishing - DANDI & EMBER

MICrONS Project

Data access and analysis - DANDI, DataJoint, VORTEX

Analysis collaboration with Neuromatch, Stanford, led by Nima Dehghani MIT

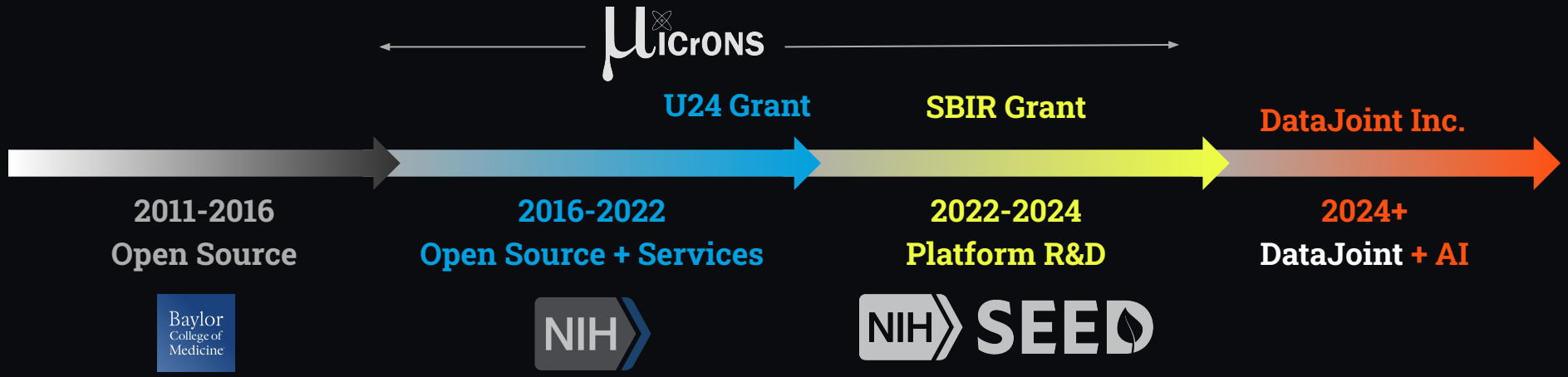
neuromatch

AI in Research Operations

Co-Pilot → Co-Scientist → Co-Strategist

Copyright 2024 DataJoint Inc. All Rights Reserved.

DataJoint: A Journey of Collaboration



*See *The U.S. Government Launches a $100-Million "Apollo Project of the Brain"*, Scientific American (Mar. 8, 2016); *A milestone map of mouse-brain connectivity reveals challenging new terrain for scientists*, Nature (Apr. 15, 2024).

OUR MISSION

To create a healthy operating environment for science that yields rapid, reproducible findings and aggregates data into a foundation for breakthrough discoveries.

Proprietary & Confidential. Copyright DataJoint 2024. All Rights Reserved.

Based on first principles of science & engineering.

SciOps Requirements

"SciOps" ≅ DevOps for Science

Reproducibility

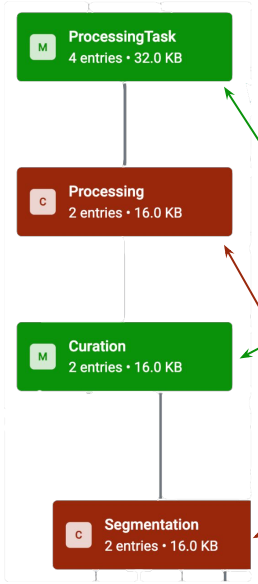
✓ Requires **relational integrity** among data, code, and process.

Flexibility + Resilience to Change

✓ Requires **unified management** of data, code, and process.

The Computational Database

A new database model



It's like a spreadsheet, with entire data tables instead of cells.

Manual tables contain user-entered data or parameters.

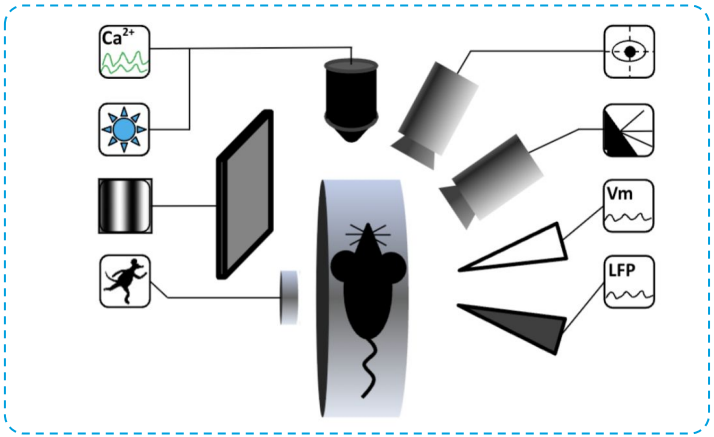
Computed tables run code that inserts new rows.

a16z

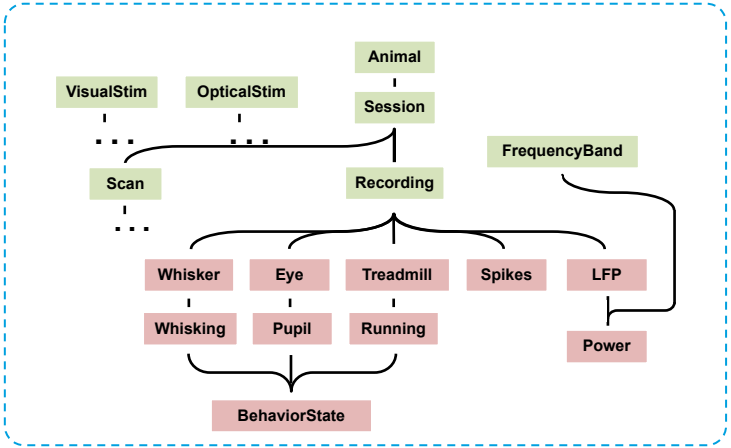
Proprietary & Confidential. Copyright DataJoint 2024. All Rights Reserved.

Express a study as a *pipeline* of data transformations.

Experimental Setup



Computational Schema



Yatsenko et al (2015)
https://arxiv.org/abs/1807.11104

Copyright 2024 DataJoint Inc. All Rights Reserved.

Demo

Open Source + Infrastructure + Operations

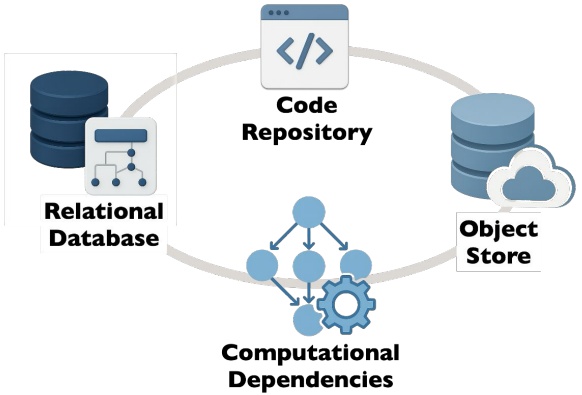
Open-source core

A flexible open standard for scientists to define all aspects of a study — so it can be ***understood***, ***validated***, ***shared***, and ***automed***.

Captures: code + data + dependencies

Operating in 100+ labs

https://github.com/datajoint/datajoint-specs
https://github.com/datajoint/datajoint-python



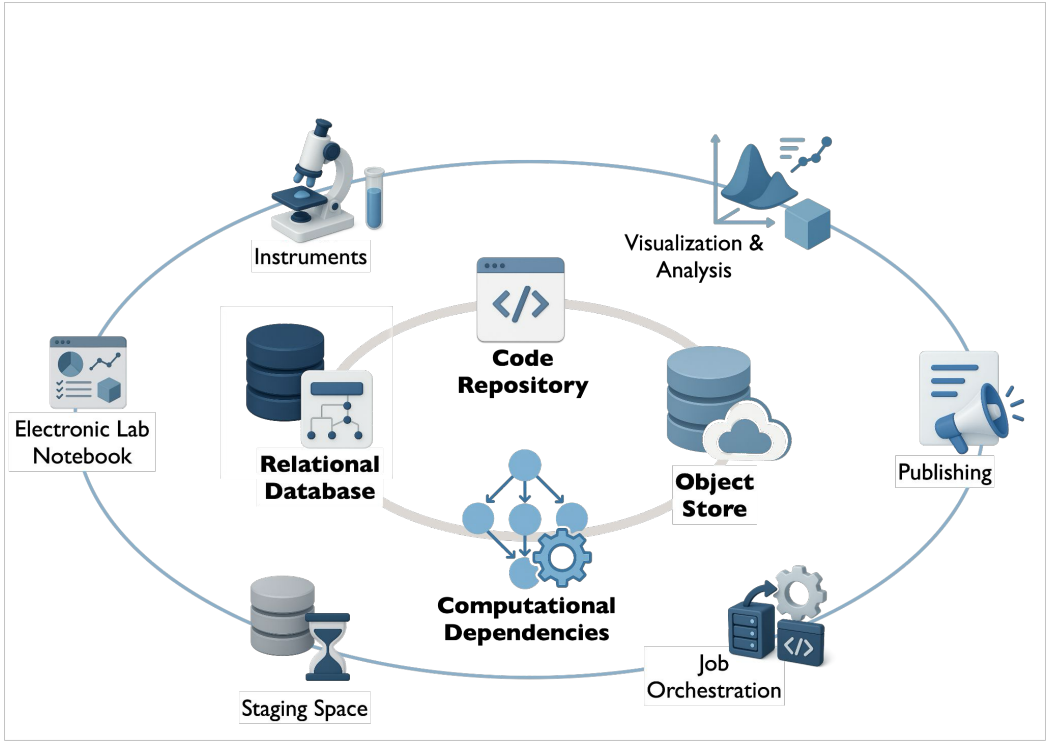
Open Source + Infrastructure + Operations

Open-source core

A flexible open standard for scientists to define all aspects of a study — so it can be ***understood***, ***validated***, ***shared***, and ***automed***.

Functional extensions for efficiency

Integration, automation, interfaces, security & compliance, sharing and publishing.



Copyright 2024 DataJoint Inc. All Rights Reserved.

Open Source + Infrastructure + Operations

Open-source core

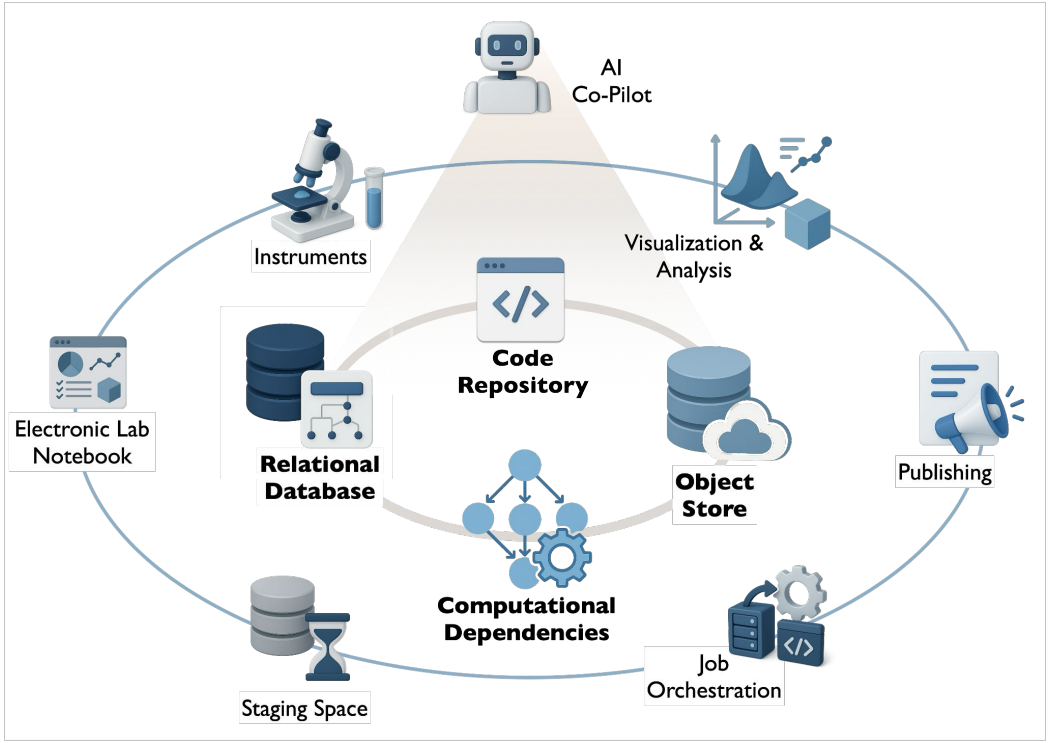
A flexible open standard for scientists to define all aspects of a study — so it can be ***understood***, ***validated***, ***shared***, and ***automed***.

Functional extensions for efficiency

Integration, automation, interfaces, security & compliance, sharing and publishing.

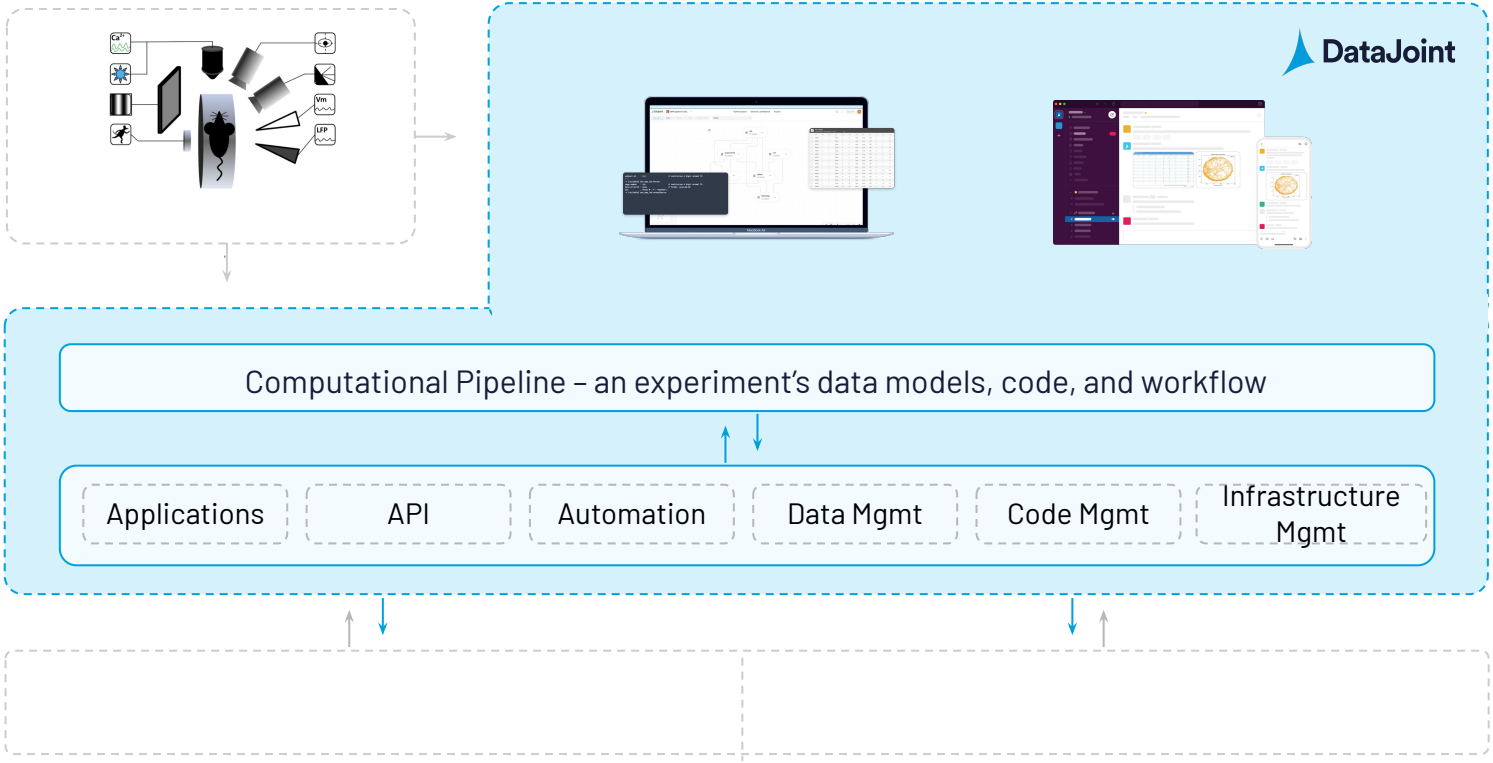
AI integration to expand capability

Data intelligence, coding assistance, knowledge tracking, writing assistance, strategy



Copyright 2024 DataJoint Inc. All Rights Reserved.

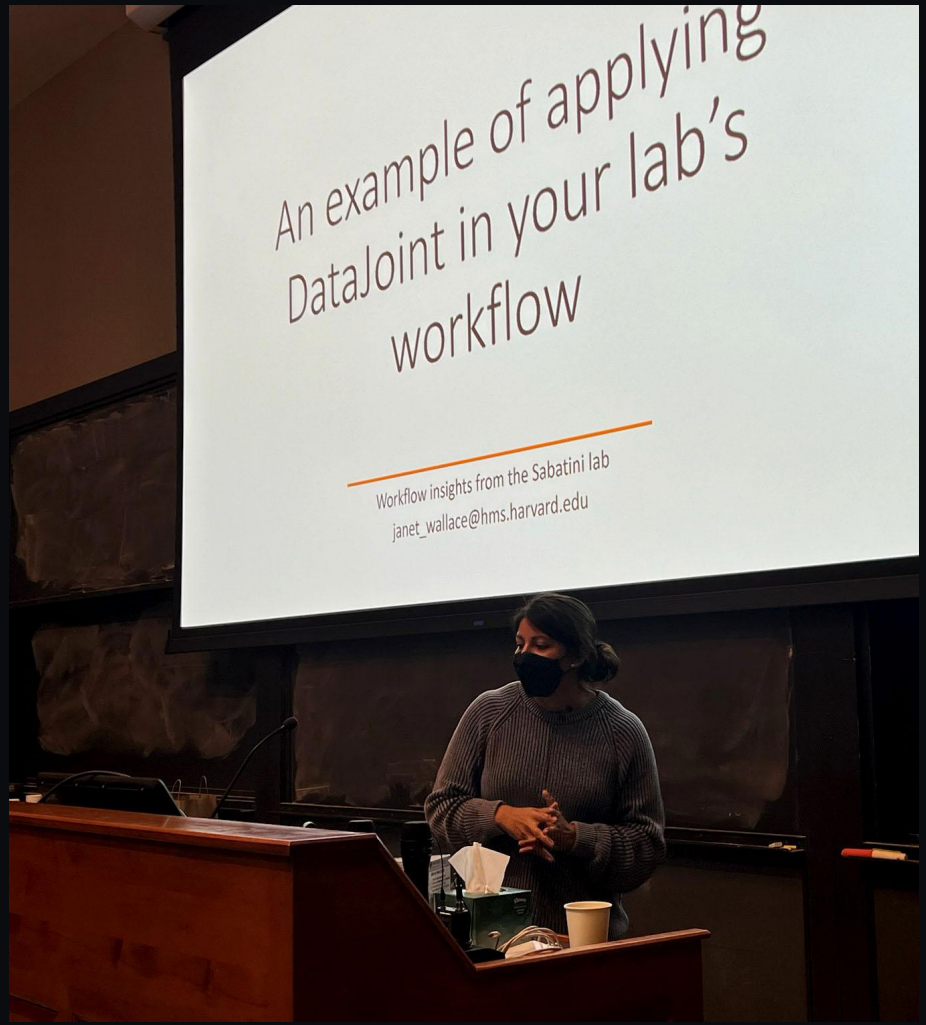
A full-stack operating platform for science.



Proprietary & Confidential. Copyright DataJoint 2024. All Rights Reserved.

Research teams

define their pipelines and workflows



DataJoint **Elements**

https://docs.datajoint.com/elements/

Modular pipelines

https://docs.datajoint.com/elements/

Common language

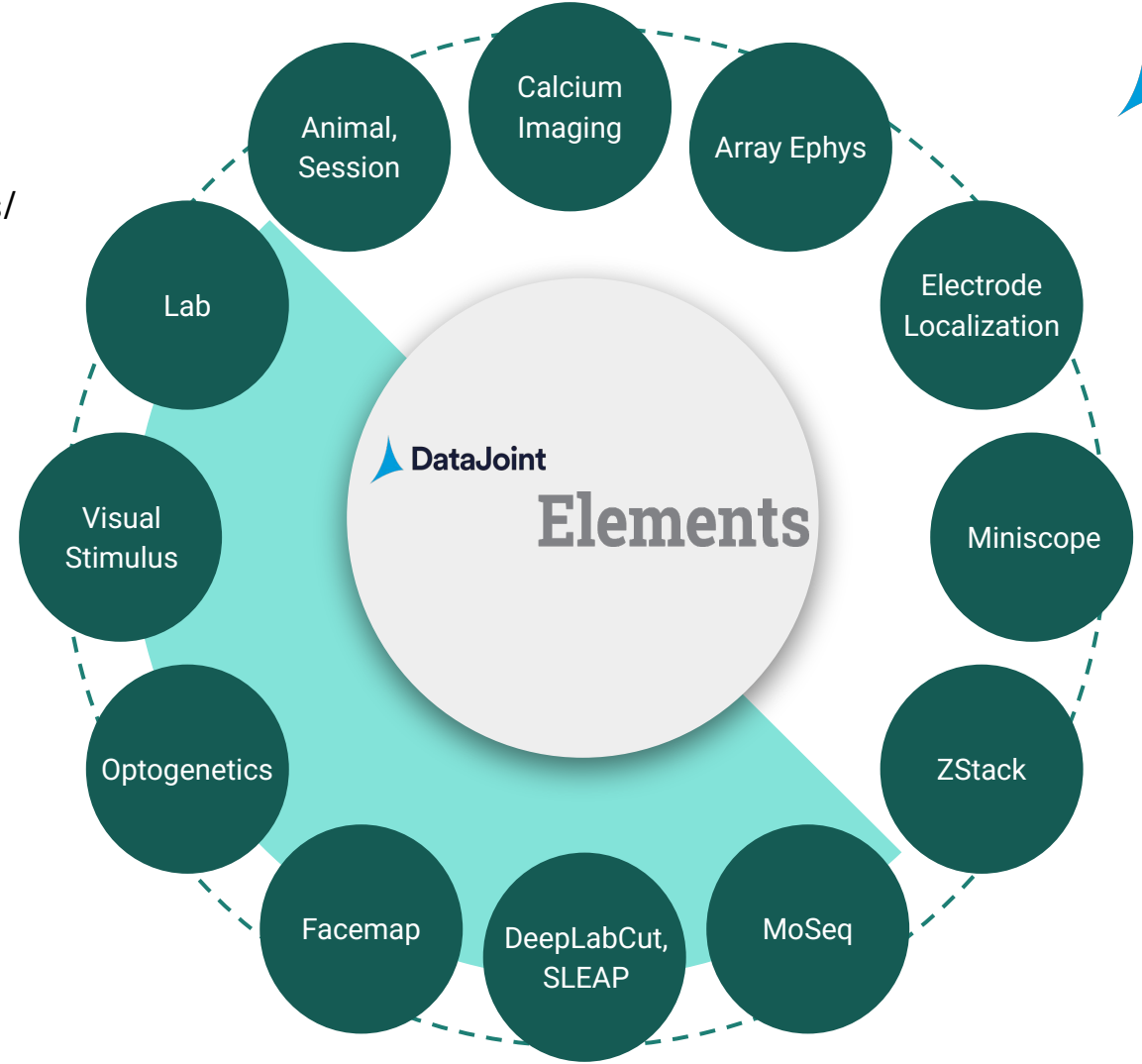
A flexible open standard for scientists to define all aspects of a study — so it can be ***understood***, ***validated***, ***shared***, and ***automed***.

Standardized modules

Validated open-source modules.
Integration, interfaces, customization

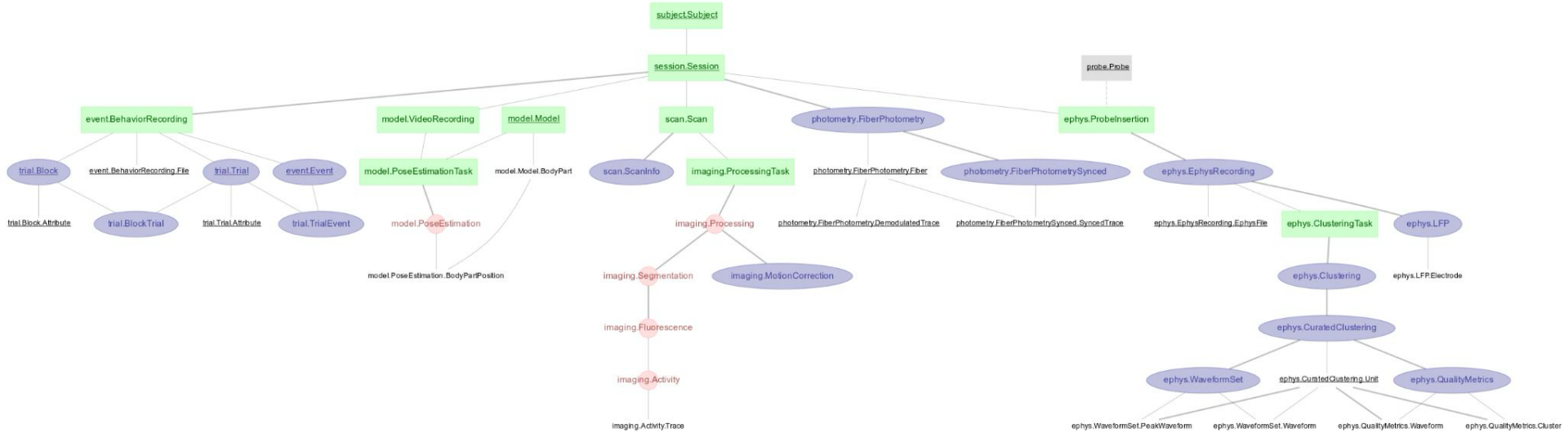
Uniform processes

Navigation, automation, queries, visualization, sharing, publishing



Copyright 2024 DataJoint Inc. All Rights Reserved.

Multimodal data pipeline



Operant Behavior (trials, events)	Pose Estimation (DLC)	Calcium Imaging (Suite2p)	Fiber Photometry	Ephys (SpikeGLX, Kilosort)

DataJoint

Open-source tools and informatics resources

Community, licensing, governance.

open ephys

NEURODATA WITHOUT BORDERS

DANDI

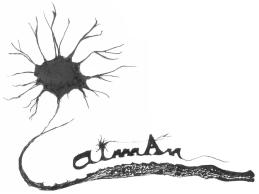
Suite2p

Facemap

DeepLabCut:
a software package for animal pose estimation

ALLEN INSTITUTE

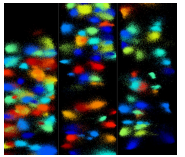
Allen Brain Observatory



Miniscope

UCLA

neuro PIXELS



Kilosort

MoSeq

bonsai
VISUAL REACTIVE PROGRAMMING

ScanImage®

SpikeInterface

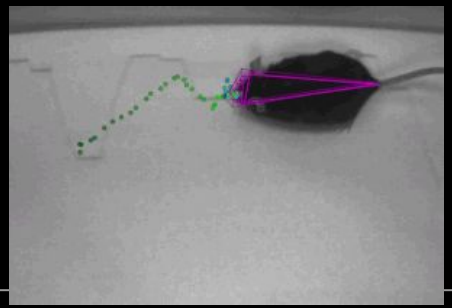
incf International Neuroinformatics Coordinating Facility

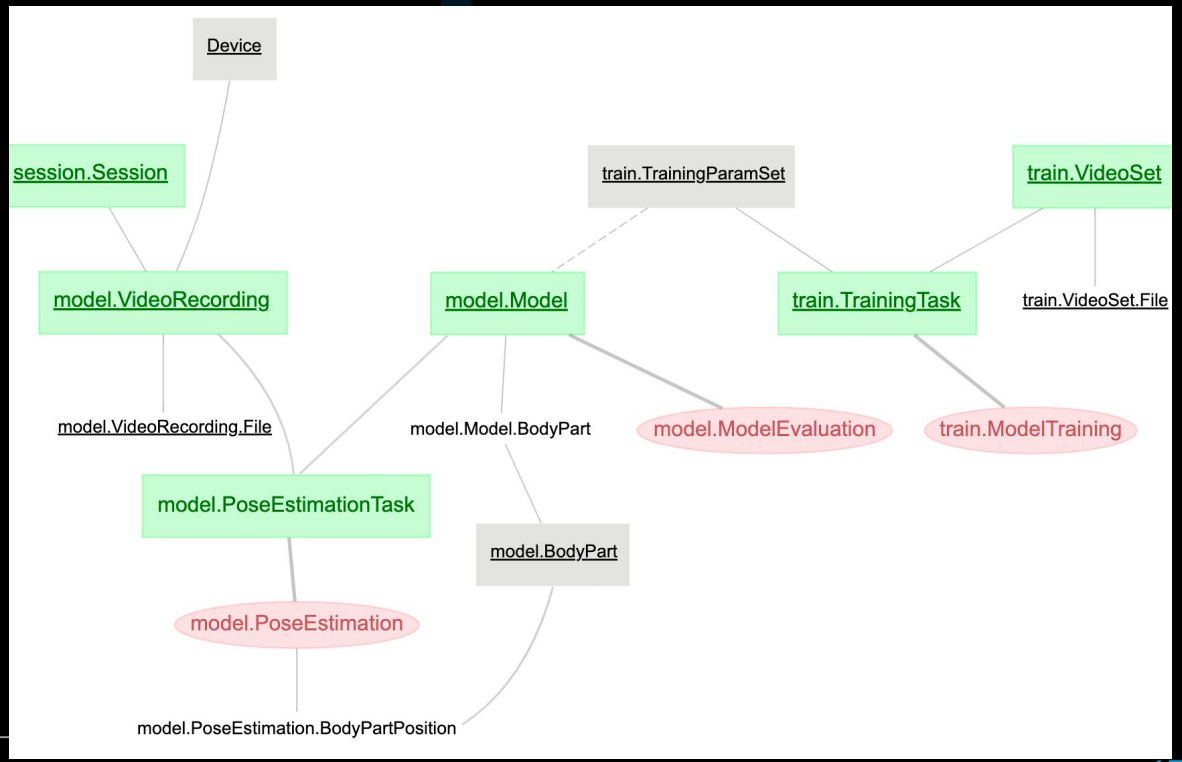
Element DeepLabCut

https://github.com/DeepLabCut/DeepLabCut

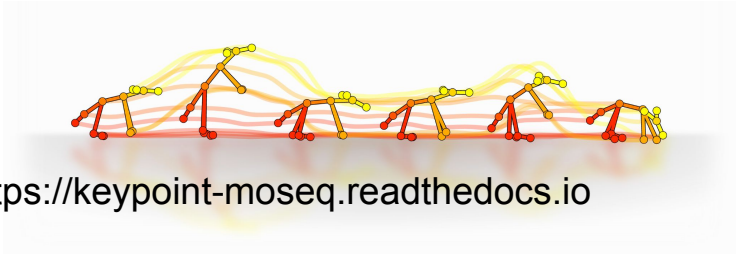
Analysis
- Video Management
- Model Training
- Pose Estimation

Projects
- Mesoscale Activity Project
- Mathis Lab @ EPFL
- Lu Lab @ Indiana U
- Rose Lab @ Bonn U
- Moser Group



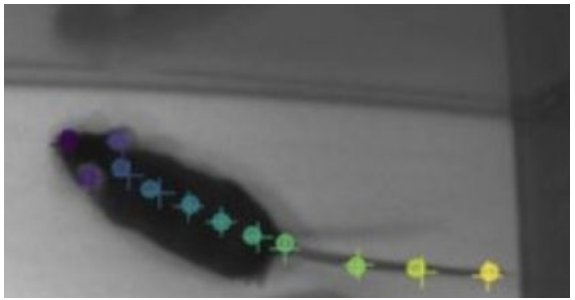


Keypoint-MoSeq

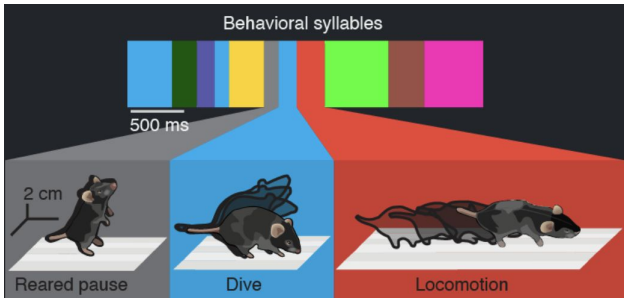


https://keypoint-moseq.readthedocs.io

Unsupervised machine learning algorithm segmenting continuous behavior into "syllables": e.g. rear, turn and pause



1. Model Training
2. Model Inference



Article | Open access | Published: 12 July 2024

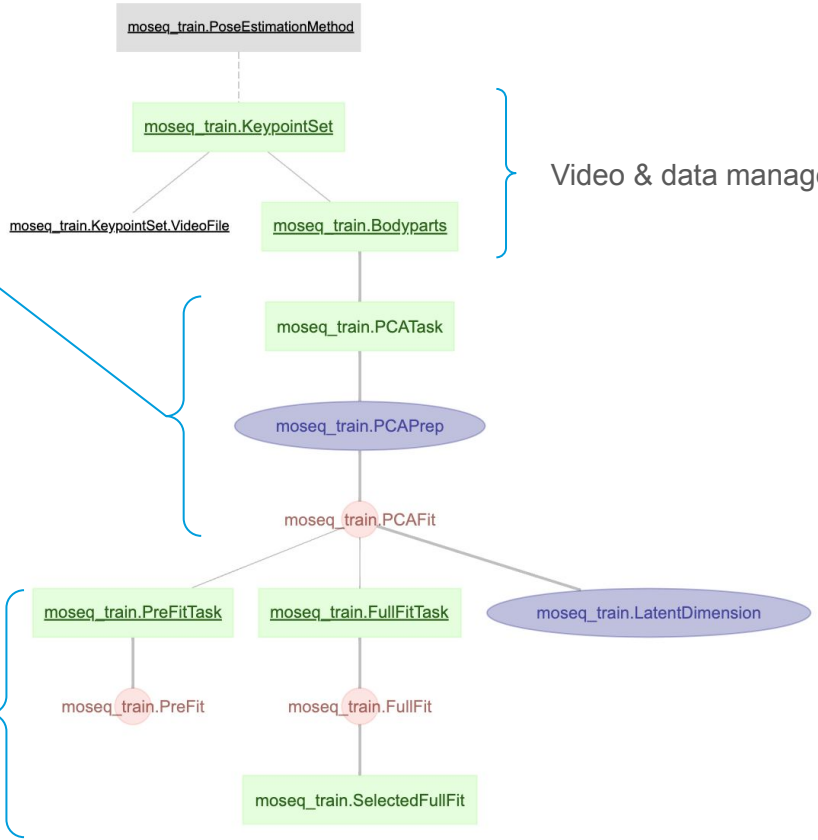
Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics

Proprietary & Confidential. Copyright DataJoint 2024. All Rights Reserved.

https://github.com/datajoint/element-moseq

MoSeq Model Training pipeline

- Project setup
  - Edit the config file
  - Load data
  - Fit PCA
- Model fitting
  - Setting kappa
  - Initialization
  - Fitting an AR-HMM
  - Fitting the full model
  - Sort syllables by frequency
  - Extract model results
    - [Optional] Save results to csv
  - Apply to new data
- Visualization
  - Trajectory plots
  - Grid movies
  - Syllable Dendrogram

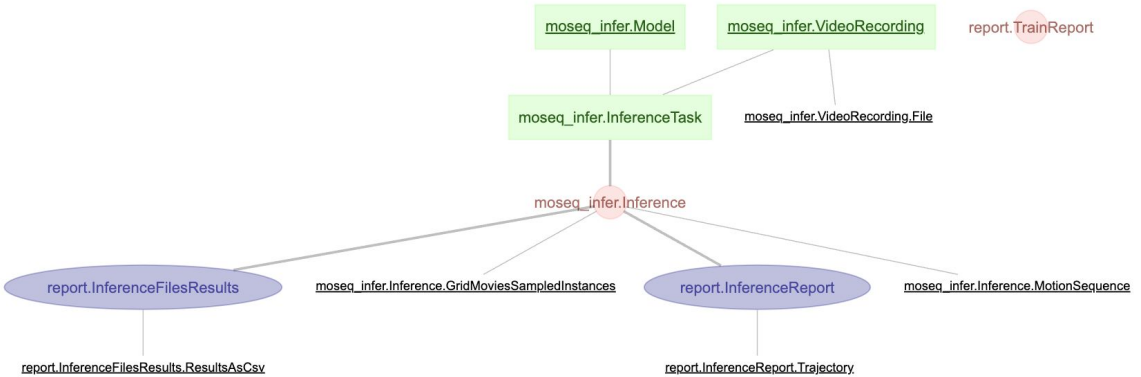


Proprietary & Confidential. Copyright DataJoint 2024. All Rights Reserved.

https://github.com/datajoint/element-moseq

MoSeq Model Inference pipeline

- Project setup
  - Edit the config file
  - Load data
  - Fit PCA
- Model fitting
  - Setting kappa
  - Initialization
  - Fitting an AR-HMM
  - Fitting the full model
  - Sort syllables by frequency
  - Extract model results
    - [Optional] Save results to csv
  - Apply to new data
- Visualization
  - Trajectory plots
  - Grid movies
  - Syllable Dendrogram



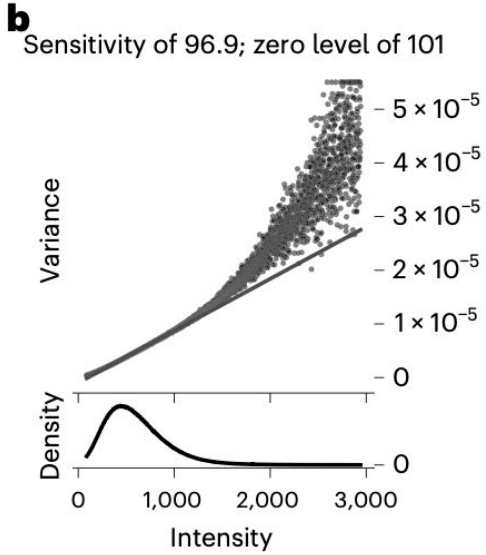
Proprietary & Confidential. Copyright DataJoint 2024. All Rights Reserved.

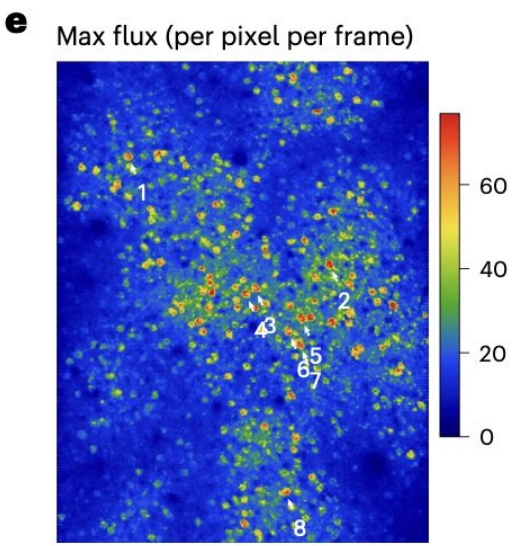
Standard quality control for every modality — and compression!

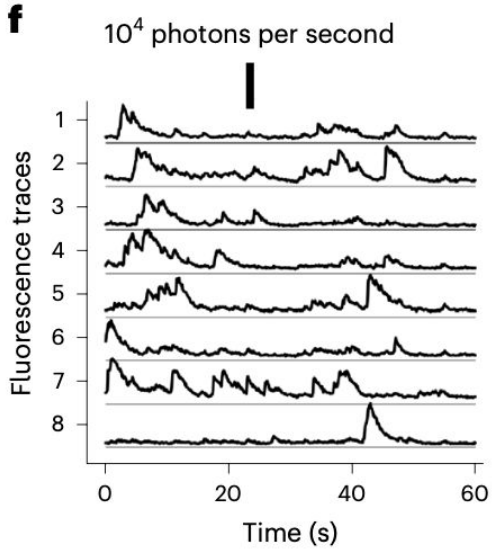
Uniform QC metrics for data quality across all data modalities

Protocol | Published: 17 March 2025

Standardized measurements for monitoring and comparing multiphoton microscope systems







Open Data In Neurophysiology:
Advancements, Solutions & Challenges

Colleen J. Gillon[†,1] Cody Baker[†,2] Ryan Ly[†,3] Edoardo Balzani,[4] Bingni W. Brunton,[5] Manuel Schottdorf,[6] Satrajit Ghosh,[7] and Nima Dehghani[7,8]

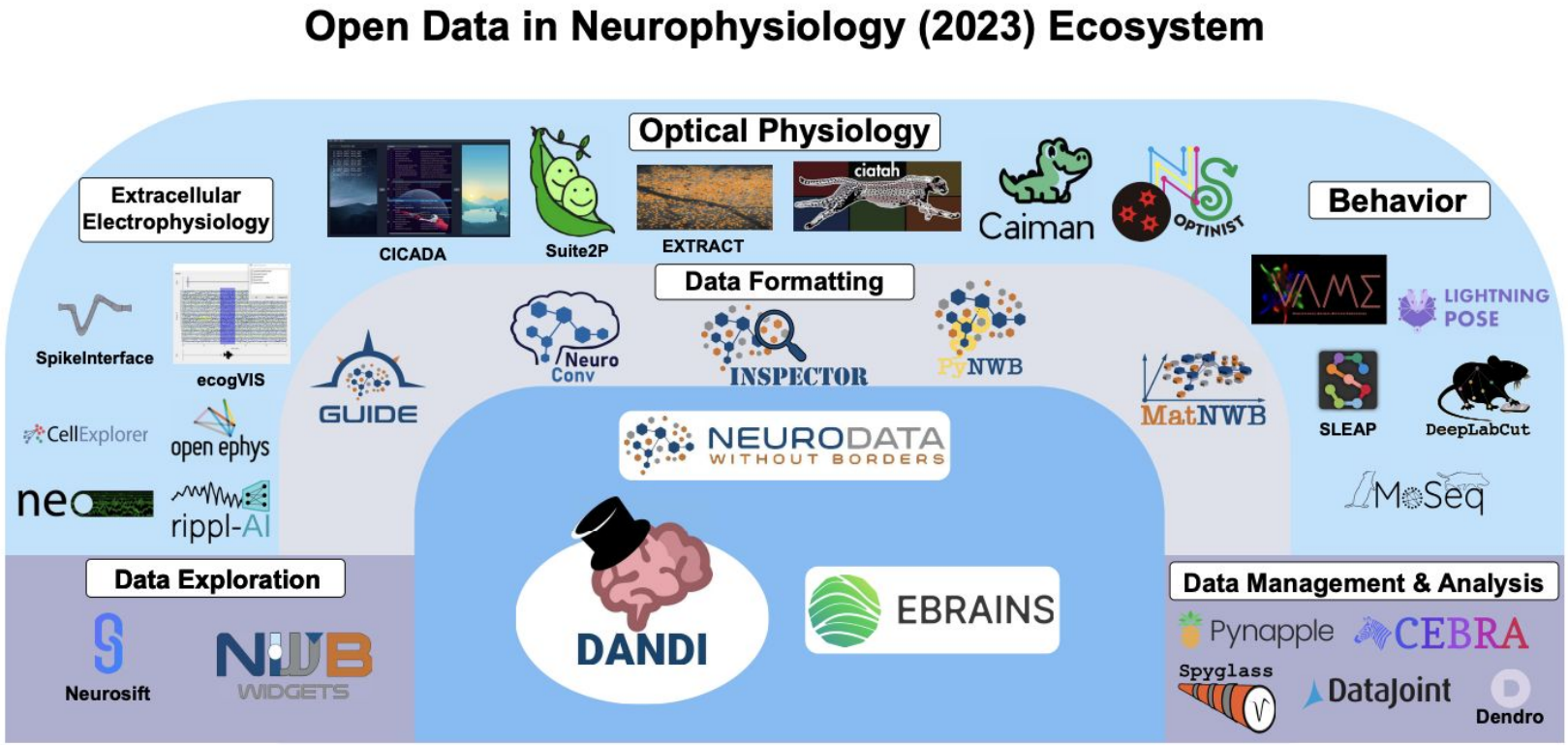
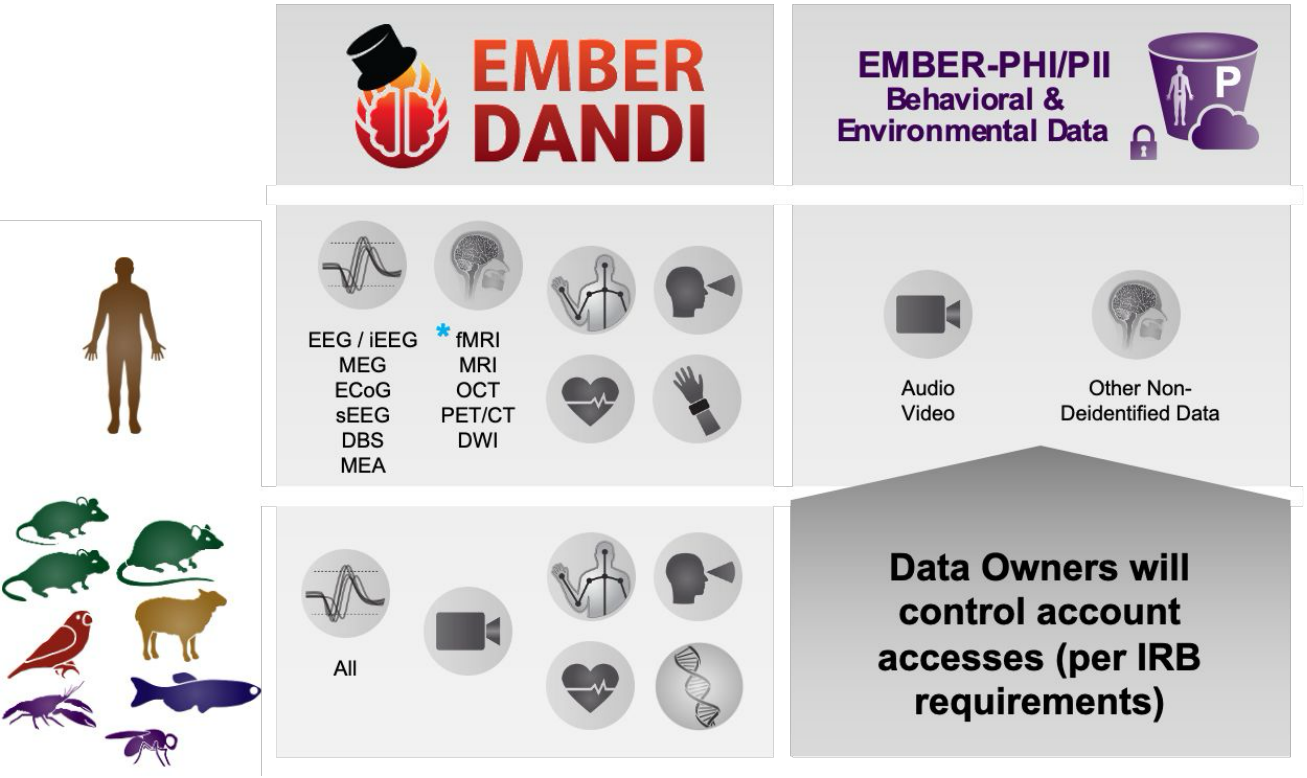


FIG. 1: The ecosystem of open source neurophysiology toolkits presented or discussed during ODIN 2023. See Table I for more information about each toolkit.

EMBER Multimodal Data Storage Strategy

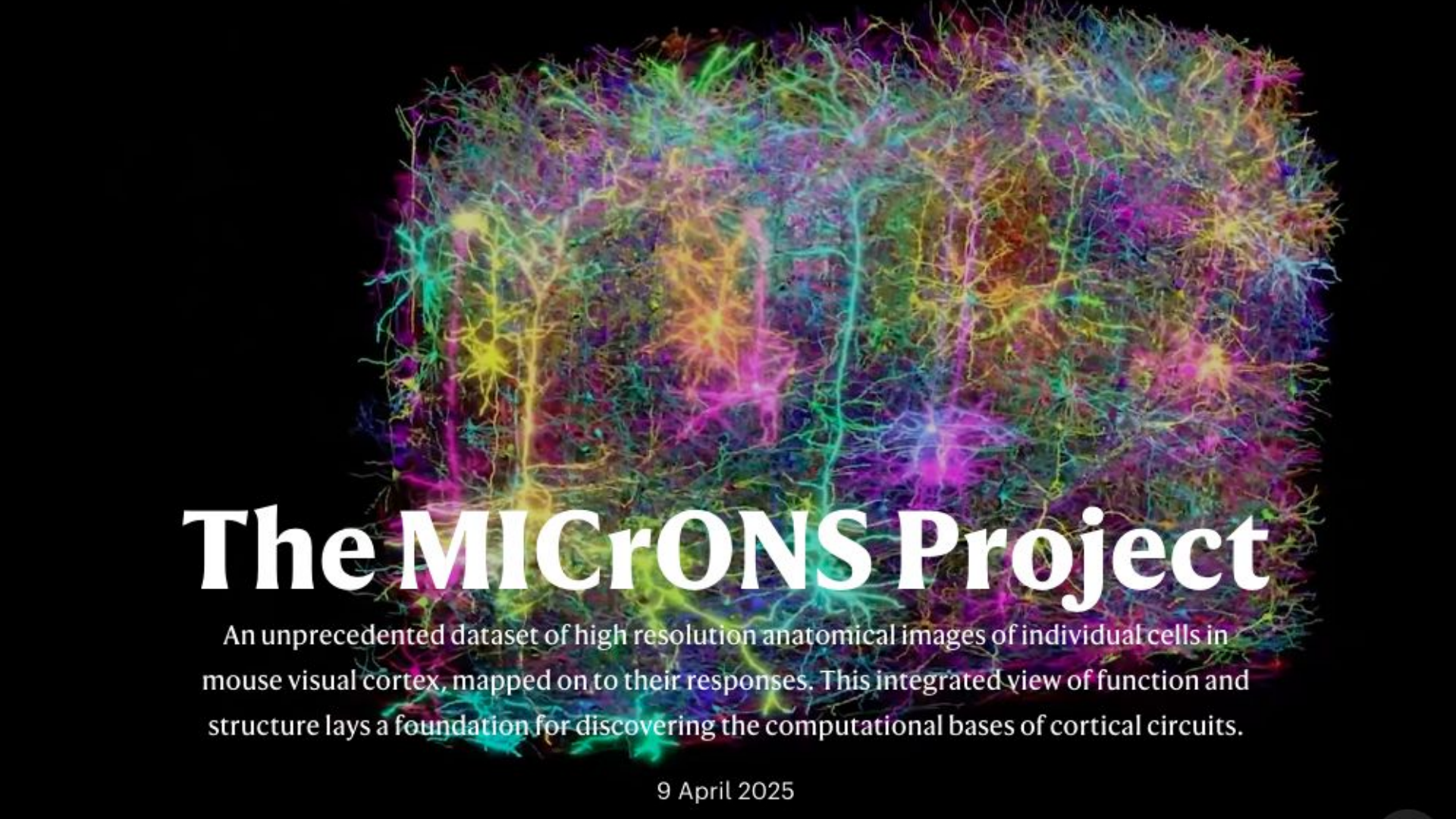
Updated Plan

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY



***Deidentified**

We will be working directly with teams to assess and guide what data can be placed in DANDI



The MICrONS Project

An unprecedented dataset of high resolution anatomical images of individual cells in mouse visual cortex, mapped on to their responses. This integrated view of function and structure lays a foundation for discovering the computational bases of cortical circuits.

9 April 2025

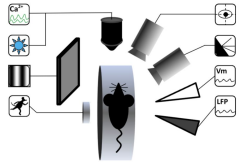
MICrONS: $100M "Apollo Project of the Brain."

2016 — IARPA — *The "Machine Intelligence from Cortical Networks" program reverse-engineers the algorithms of the brain to develop less artificial AI.* — MICrONS — 2025

✓ Collaboration ✓ Data Collection ✓ Analysis Publication

DataJoint: the system of record

- *In vivo* SciOps
- Structure / function link
- Neural Data Access hosting

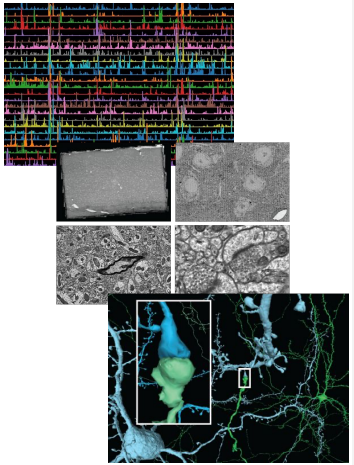


Experimental Methods

BCM Baylor College of Medicine — **Function.** Two-photon imaging records neuronal responses to visual stimuli in a cubic millimeter of mouse visual cortex.

ALLEN INSTITUTE — **Neuroanatomy.** Serial-section electron microscopy images the same cubic millimeter of visual cortex.

PRINCETON UNIVERSITY — **Connectome.** Convolutional nets align the image slices to reconstruct neurons and synapses in 3D and segment cells by class.



"We now have the technology to look into the detailed organization of the brain."

Andreas Tolias

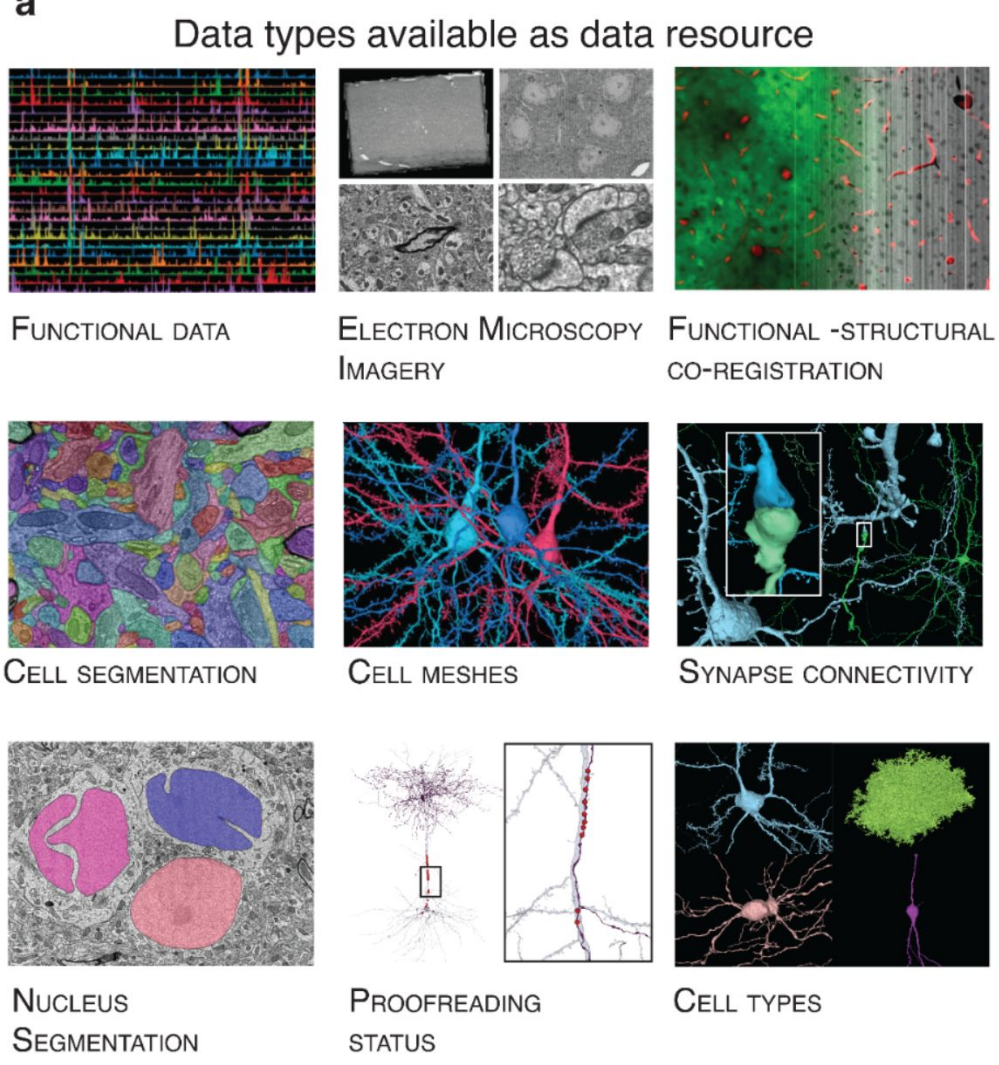
75,000 functional neurons

500,000 synapse junctions

2 petabytes of data

Proprietary & Confidential. Copyright DataJoint 2024. All Rights Reserved.

MICrONS Data



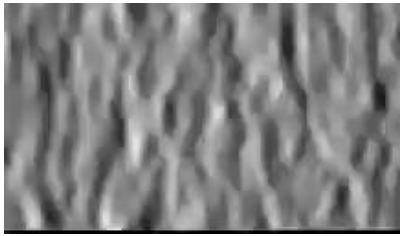
https://www.microns-explorer.org

https://datajoint.com/microns

Visual Observatory of Cortex (VORTEX)
https://www.microns-explorer.org/vortex

MICrONS Visual Stimuli

"Monet"



"Trippy"

Videos

Receptive fields
Orientation tuning
Direction tuning

Responsiveness
Signal correlations

ML models
“Digital twin”

Modeling
Validation
Ethological relevance

Accessing Physiology Data

1. (Difficult) Download the SQL data
   - Set up your own SQL server, load the data
   - Access with the DataJoint API

2. (Easier) Connect to the database instance hosted by DataJoint.com
   - Access with public credentials using the DataJoint API
   - Complete tutorial in DevContainer - Codespaces:
     https://github.com/datajoint/microns_phase3_nda/

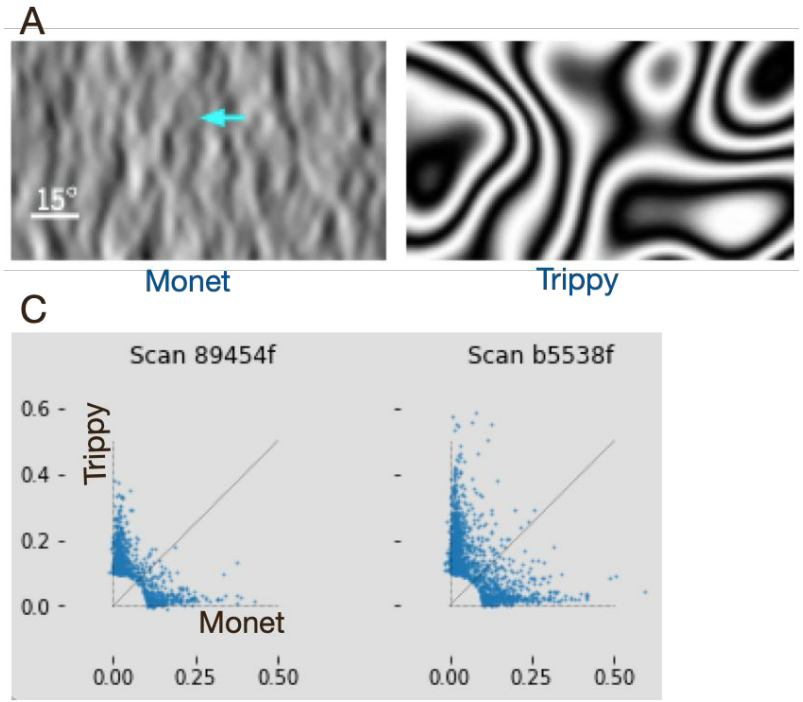
3. (Easiest) Download NWB files from DANDI Archive
   https://doi.org/10.48324/dandi.000402/0.230307.2132

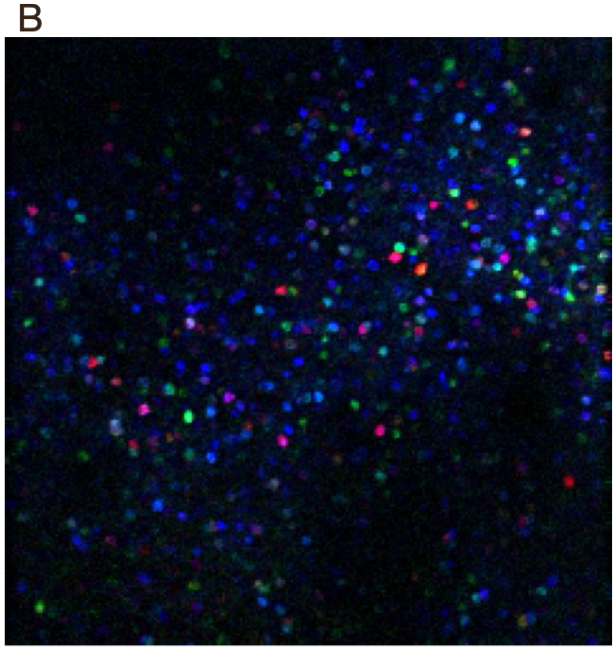
DANDI

4. (Pro) Work on a complete data pipeline from raw data
   Analysis collaboration: DataJoint + Neuromatch + Stanford

Analysis Collaboration

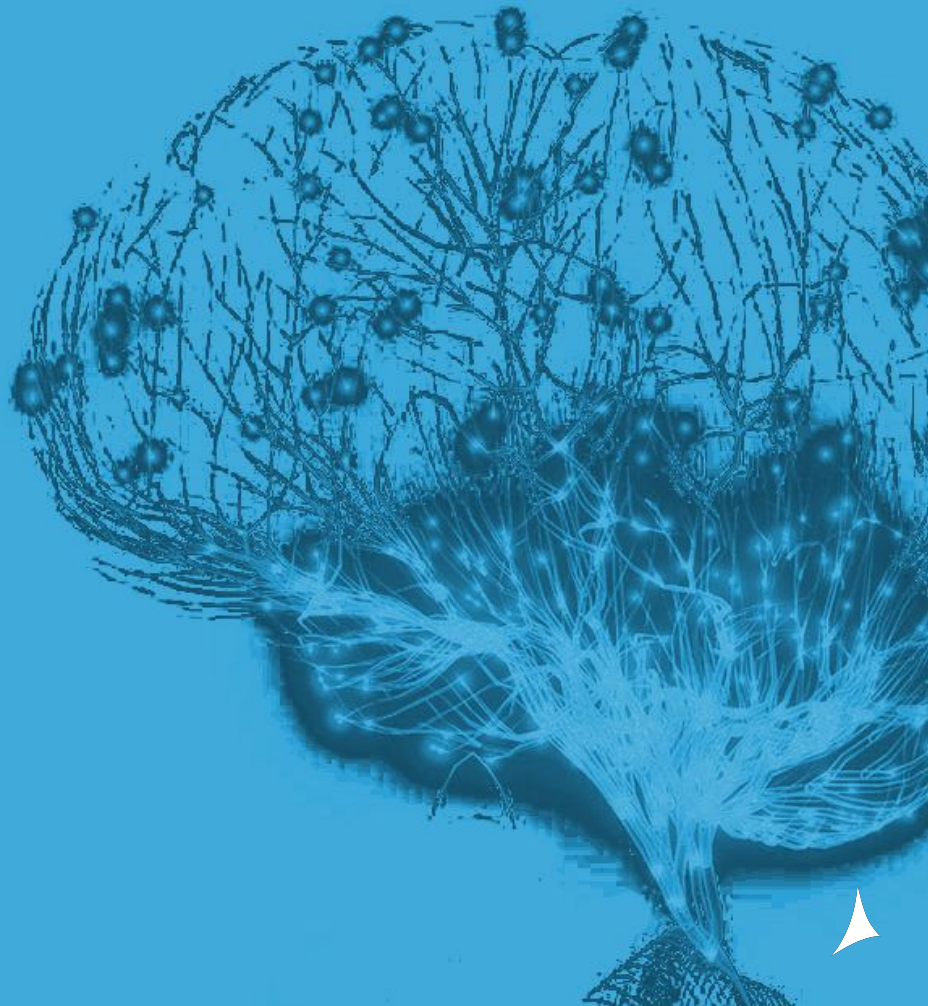
*D. YATSENKO[1], P. FAHEY[2], T. MUHAMMAD[3], M. SHAKIBA[4], R. ROKNI[4], M. MOHAMMADI[4], N. DEHGHANI[5], A. S. TOLIAS[2]





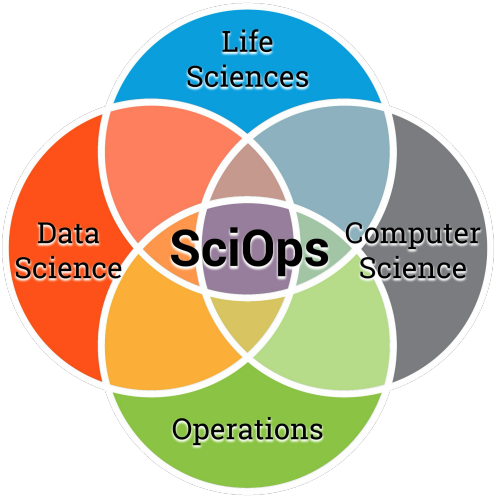
Synthetic stimuli divide neuronal populations into two functional circuits

The NeuroAI Future



Copyright 2024 DataJoint Inc. All Rights Reserved.

New Discipline: SciOps



Skills + Standards + Systems

DataJoint

APL Johns Hopkins Applied Physics Laboratory

CATALYST NEURO

Pfizer

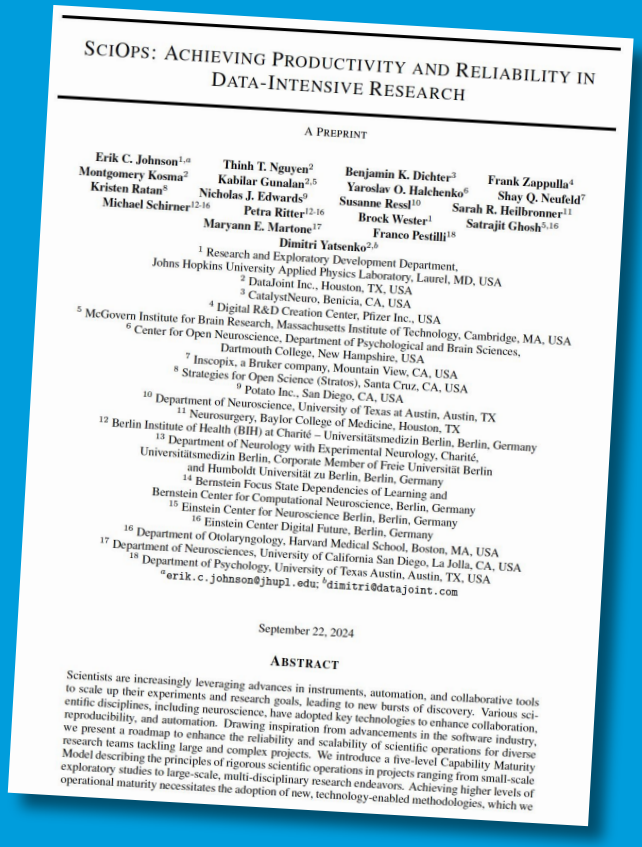
DARTMOUTH

INSCOPIX A BRUKER COMPANY

CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN

UC San Diego School of Medicine

TEXAS The University of Texas at Austin



SciOps: Achieving Productivity and Reliability in Data-Intensive Research

A Preprint

Erik C. Johnson[1,a] Thinh T. Nguyen[2] Benjamin K. Dichter[3] Frank Zappulla[4]
Montgomery Kosma[2] Kabilar Gunalan[2,5] Yaroslav O. Halchenko[6] Shay Q. Neufeld[7]
Kristen Ratan[8] Nicholas J. Edwards[9] Susanne Ressl[10] Sarah R. Heilbronner[11]
Michael Schirner[12-16] Petra Ritter[12-16] Brock Wester[1] Satrajit Ghosh[5,16]
Maryann E. Martone[17] Franco Pestilli[18]
Dimitri Yatsenko[2,b]
[1] Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA
[2] DataJoint Inc., Houston, TX, USA
[3] CatalystNeuro, Benicia, CA, USA
[4] Digital R&D Creation Center, Pfizer Inc., USA
[5] McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA
[6] Center for Open Neuroscience, Department of Psychological and Brain Sciences, Dartmouth College, New Hampshire, USA
[7] Inscopix, a Bruker company, Mountain View, CA, USA
[8] Strategies for Open Science (Stratos), Santa Cruz, CA, USA
[9] Potato Inc., San Diego, CA, USA
[10] Department of Neuroscience, University of Texas at Austin, Austin, TX
[11] Neurosurgery, Baylor College of Medicine, Houston, TX
[12] Berlin Institute of Health (BIH) at Charité – Universitätsmedizin Berlin, Berlin, Germany
[13] Department of Neurology with Experimental Neurology, Charité, Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany
[14] Bernstein Focus State Dependencies of Learning and Bernstein Center for Computational Neuroscience, Berlin, Germany
[15] Einstein Center for Neuroscience Berlin, Berlin, Germany
[16] Einstein Center Digital Future, Berlin, Germany
[17] Department of Otolaryngology, Harvard Medical School, Boston, MA, USA
[18] Department of Neurosciences, University of California San Diego, La Jolla, CA, USA
Department of Psychology, University of Texas Austin, Austin, TX, USA
[a] erik.c.johnson@jhupl.edu; [b] dimitri@datajoint.com

September 22, 2024

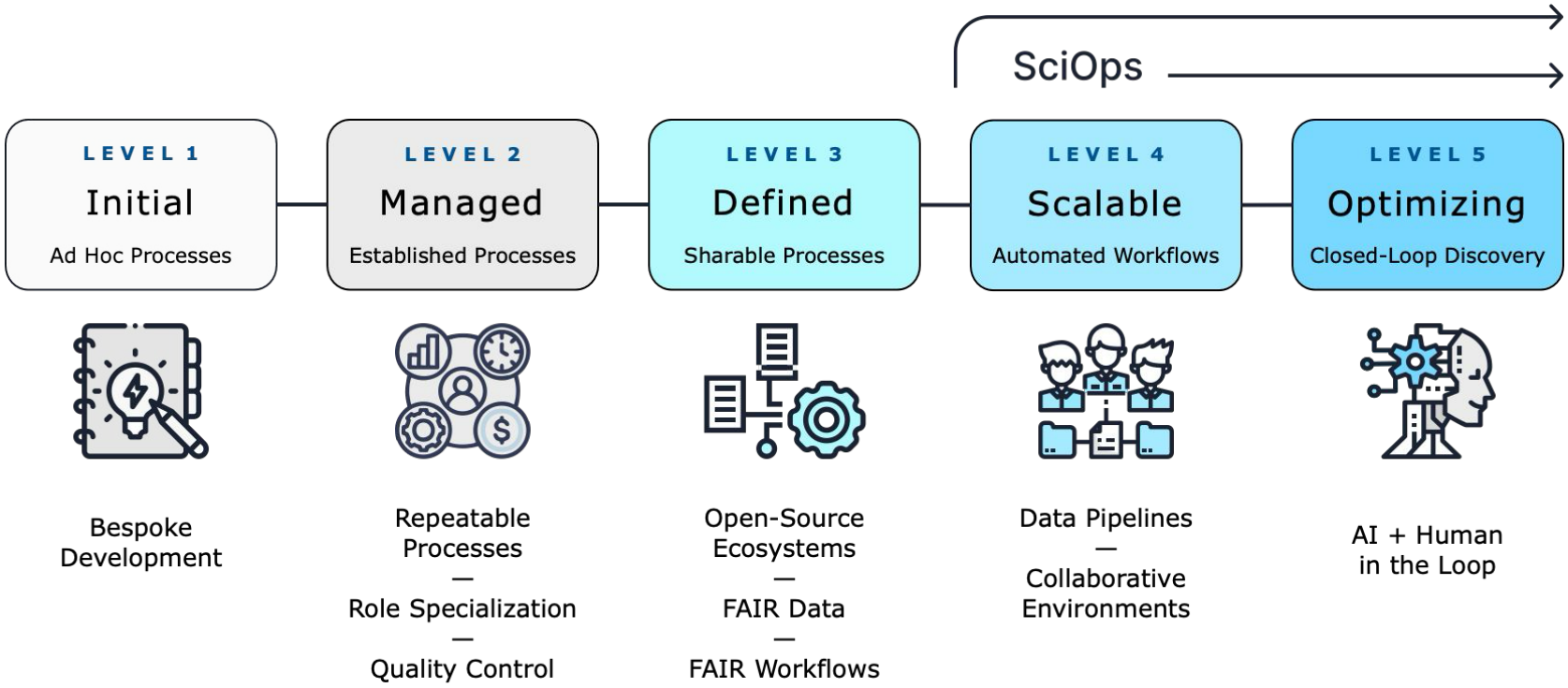
Abstract

Scientists are increasingly leveraging advances in instruments, automation, and collaborative tools to scale up their experiments and research goals, leading to new bursts of discovery. Various scientific disciplines, including neuroscience, have adopted key technologies to enhance collaboration, reproducibility, and automation. Drawing inspiration from advancements in the software industry, we present a roadmap to enhance the reliability and scalability of scientific operations for diverse research teams tackling large and complex projects. We introduce a five-level Capability Maturity Model describing the principles of rigorous scientific operations in projects ranging from small-scale exploratory studies to large-scale, multi-disciplinary research endeavors. Achieving higher levels of operational maturity necessitates the adoption of new, technology-enabled methodologies, which we

Our SciOps paper is under review by *Nature Methods*.

Copyright 2024 DataJoint Inc. All Rights Reserved.

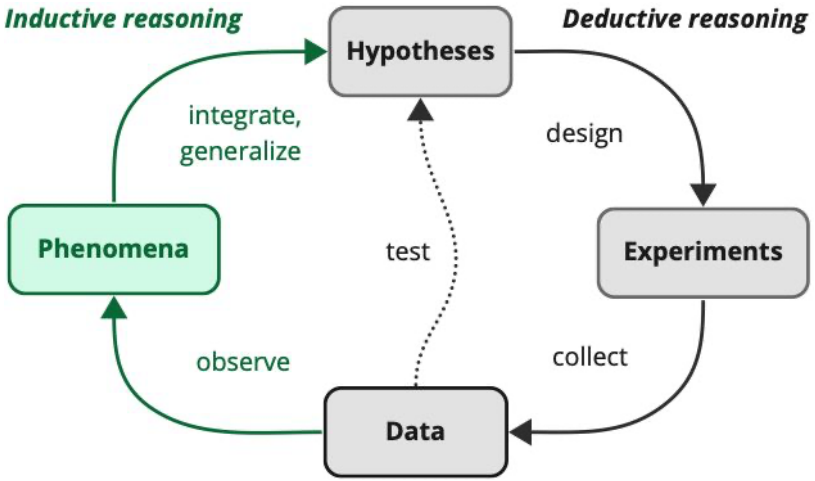
NeuroAI requires SciOps maturity.



https://doi.org/10.48550/arXiv.2401.00077

Copyright 2024 DataJoint Inc. All Rights Reserved.

Closed-loop system from hypothesis to verification



AI requires mature scientific operations

SciOps: **A Comprehensive Platform**

Productivity through automation

- Ingest and store
- Prepare and standardize
- FAIR
- Flexible queries
- Collaborate, controlled sharing, and publish

AI Enablement

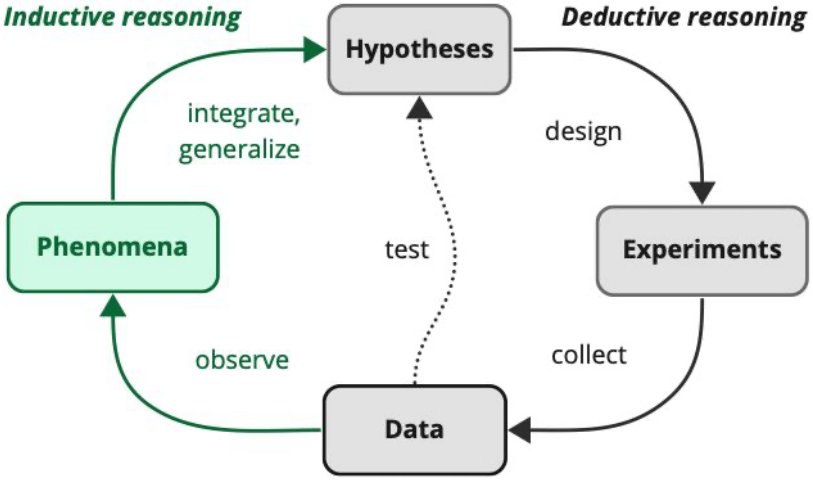
- Intuitive UI and common language
- Analyze and build models
- Integrate pipelines, tools (existing and 3rd party)
- Dynamic and flexible
- Compute orchestration

Strong Governance

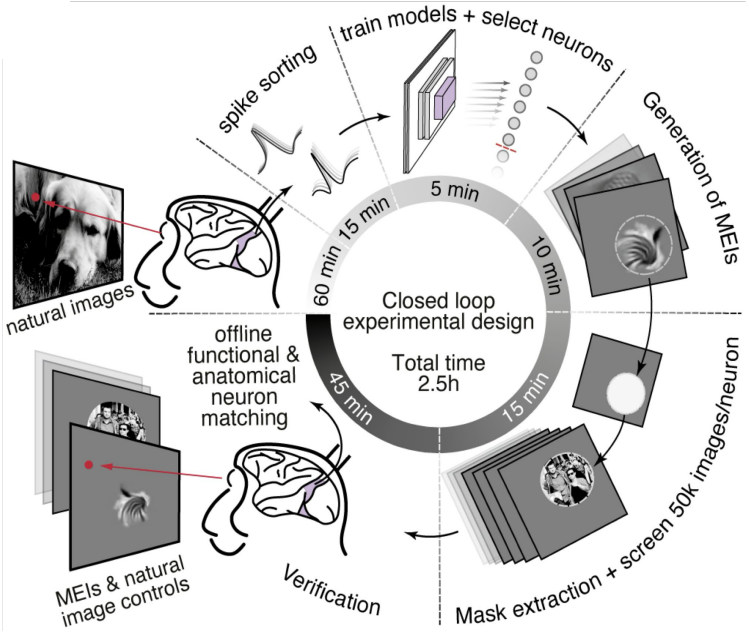
- System of record
- Code + data + dependencies + environment
- Data Integrity enforced and transparent
- Reproducibility
- Protect IP
- Well documented processes and transformations

Copyright 2024 DataJoint Inc. All Rights Reserved.

Closed-loop neuroscience with live ML modeling.



AI embedding requires mature operations



courtesy Prof. Andreas Tolias

Stanford University

Copyright 2024 DataJoint Inc. All Rights Reserved.

The Three AI Partners

A synergistic framework of AI collaborators. Click on any layer to learn more.

Level 2: AI Co-Visionary

Strategy & Vision

Guides long-term research trajectory, identifies funding and partnership opportunities, and helps map the strategic impact of scientific discoveries.

Level 1: AI Co-Scientist

Creativity & Exploration

Acts as an intellectual partner, generating novel hypotheses, synthesizing literature, and assisting in manuscript preparation through a closed-loop discovery process.

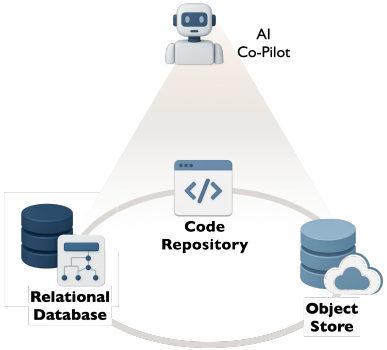
Level 0: AI Co-Pilot

Rigor & Precision

Ensures absolute data integrity by translating natural language into precise, validated queries and orchestrating complex analysis workflows with perfect reproducibility.

Deepinvent

- FutureHouse.org
- NovelSeek
- ReadySetPotato.com
- Google Co-Scientist
- Microsoft Discovery Platform
- HypSynth



Areas for collaboration

NIH AI: NOT-OD-25-117

- Operational excellence:

  governance + infrastructure + operations + dev support + training
- AI-powered closed-loop science: Grant Opportunities
- Industry sponsored research
- Research communication and publishing: data + computation
- Technology dissemination & commercialization. Software licensing.

Copyright 2024 DataJoint Inc. All Rights Reserved.

THANK YOU