DataJoint

Workflow grounding starts at the bench

Integrity established before anything runs.

Dimitri Yatsenko

PhD · Founder and **Chief Science Officer**, DataJoint

AI DRUG DISCOVERY & DEVELOPMENT SUMMIT

© 2026 DataJoint Inc.

DataJoint

Agents work faster than anyone can verify.

- plausible is **perceived as correct**
- the link from data to result is **tenuous**
- human review **cannot keep up**
- **accountability and credit** become ambiguous

Are the conclusions defensible, now and in six months?

© 2026 DataJoint Inc.

DataJoint

The standards require verifiability and oversight.

ISO/IEC TS 5723	"the ability to meet stakeholders' expectations in a verifiable way"
EU AI Act, Art. 12–14	automatic logging, information to interpret the output, and human oversight for high-risk systems
FDA draft guidance, 2025	a model's credibility established for its stated context of use before anyone relies on a result

Each asks for **a record that lets someone else check the work**.

© 2026 DataJoint Inc.

DataJoint

Pharma R&D will run agents in accountable workflows.

	TODAY	TOMORROW
optimized for	velocity	**understanding, and velocity that compounds**
division of work between humans and AI	by capability	**by accountability**
AI grounding	around the model	**around the workflow**
instructions	human prose	**+ formal notation**
integrity	checked after the fact	**established before the run, from the bench**
knowledge	in institutional memory, conversations, and notes	**kept with the work, in curated context**

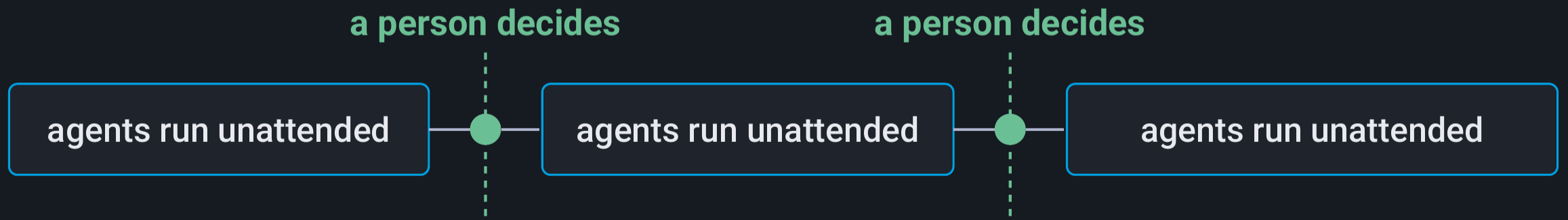
© 2026 DataJoint Inc.

DataJoint

As AI does more, scientists answer for more.

DIVIDED BY	WHO DOES WHAT	AS MODELS IMPROVE
capability	humans do what AI cannot	the line moves
accountability	only a person can be held accountable for consequential decisions	the line holds

Challenge: how can AI agents execute long stretches of work while people direct, decide, and answer for the results?



© 2026 DataJoint Inc.

DataJoint

Grounding moves from the model to the workflow.

GROUNDING		HOW AN ANSWER IS CHECKED
2021–23	fine-tuning	not checkable
2022–24	prompting	by reading the answer
2023–24	retrieval	by following the citation
2025	context engineering	by the context engineer
2025	MCP	by asking the system again
2026	science harnesses	by the code and reasoning the harness records
now	**workflow grounding**	traced to inputs, code, and decisions

An agent is grounded in a workflow when it reads and extends the declarations that run the work.

© 2026 DataJoint Inc.

DataJoint

Prose does not compound. Notation does.

RESPONSIBILITY	LANGUAGE	WHAT THE LANGUAGE LEAVES OUT
pipelines	CWL, Nextflow, Snakemake	the data structure
infrastructure	Terraform, Helm	the derivation
access	Cedar, OPA	what the work is
warehouse transform	dbt	the entity grain
stored structure	SQL DDL, Parquet	computation and state
research workflow	Relational Workflow Model	infrastructure, access, and storage

Open standard · arXiv:2602.16585

© 2026 DataJoint Inc.

DataJoint

Conditions for human accountability with expanded AI autonomy.

THE DECLARATION IS	SO A PERSON CAN VERIFY
declarative and succinct	the design itself, not pages of code
self-describing	what it means, without documentation that can drift
formal	after a machine checks it first
executable	results on real data, not imagined
integrity-bearing	that nothing downstream is invented
state-aware	what is stale

© 2026 DataJoint Inc.

DataJoint

Design review with agents.

Caught in the declaration, before anything runs

DIMENSION	DEFECT	FIX
identity	**missing entity**	declare the entity and key on it
	conflated identity	a table each for specimen, operator, and run
	coordinate as entity	well and spot in the reading's key
grain	**grain flattened**	a part table, one row per trial
dependencies	**linkage by convention**	a foreign key to the upstream table
tiers and modules	**orchestration outside the graph**	populate and its built-in jobs table

© 2026 DataJoint Inc.

DataJoint

Integrity before the run.

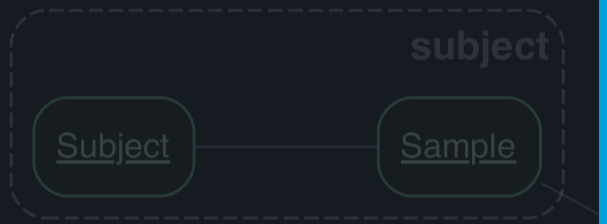
	IN THE DATA MODEL · INTRINSIC	AT THE BENCH · EXTRINSIC
Enforced	at definition, before anything runs	at run time, as the data lands
Held by	the structure of the data model	the record on the row it describes
A failure	a step that reads outside its inputs never deploys	an ingestion without its source record is refused
Answers	“this result must derive from these inputs”	“this file was reported to come from that instrument”

Neither guarantee can substitute for the other.

© 2026 DataJoint Inc.

DataJoint

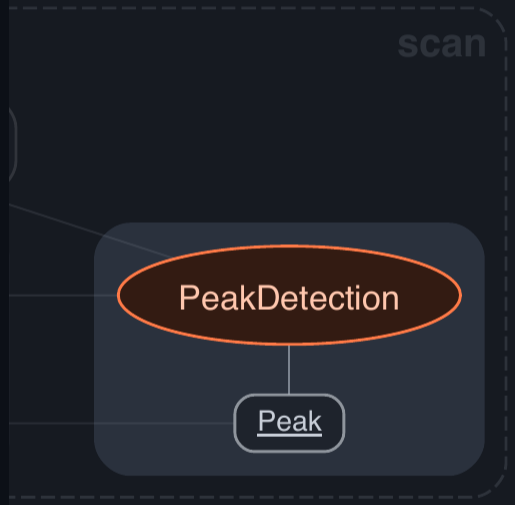
Each table is a step.




```python
class PeakDetection(dj.Computed):
    definition = """
    -> MassAnalysis
    -> PeakDetectionParams
    ---
    total_peaks : int32
    """

    class Peak(dj.Part):
        definition = """
        -> master
        -> MassAnalysis.Spectrum
        peak_idx  : int32
        ---
        mz        : float64
        intensity : float64
        snr       : float64
        """

    def make(self, key):
        params = (PeakDetectionParams & key).fetch1()
        spectra = (MassAnalysis.Spectrum & key).fetch()
        peaks = find_peaks(spectra, **params)
        self.insert1({**key, "total_peaks": len(peaks)})
        self.Peak.insert(peaks)
```



An open standard in development · docs.datajoint.com · datajoint/lcms-demo · arXiv:2602.16585

© 2026 DataJoint Inc.

DataJoint

The schema is the workflow.

Everywhere else

- a repository holds the data
- an orchestrator runs the steps
- a warehouse holds the results

Keeping the repository, the orchestrator, and the warehouse consistent **is** the work.

Relational Workflow Model

- a table is **a workflow step**
- a foreign key is **dependency, order, and provenance**
- the computation lives **on the edge**

One schema.

A schema always matches the data it describes. A catalog can drift.

© 2026 DataJoint Inc.

DataJoint

From surgery to the rig.

UCSF

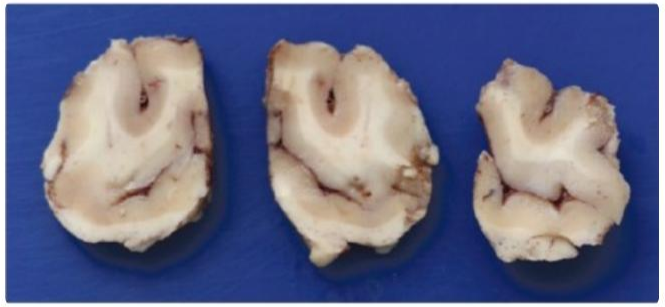
Cadwell Lab · Dr. Cathryn Cadwell

drug-resistant epilepsy, toward regenerative cell therapy



donor: epilepsy surgery

→



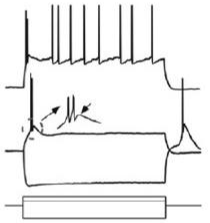
specimen and blocks

→



slice, on the rig

→



patched neuron

Patch-seq measures each patched neuron three ways:
electrical activity, transcriptome, and morphology.

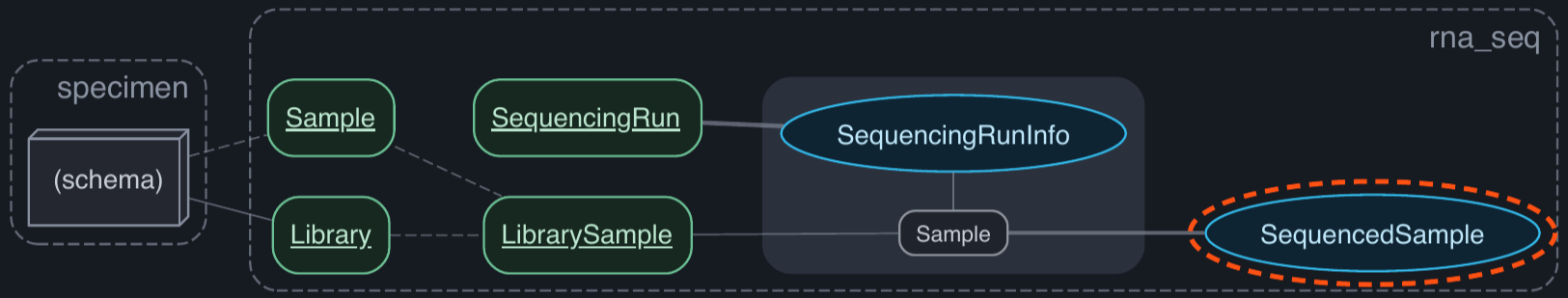
Patch-seq: Cadwell et al., Nature Biotechnology 2016

© 2026 DataJoint Inc.

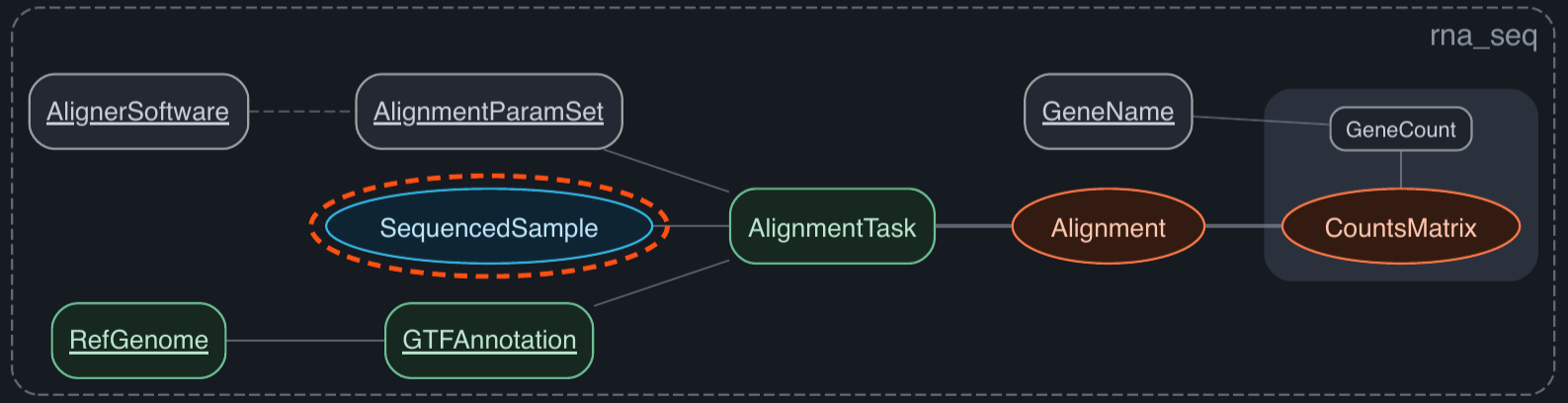
DataJoint

Every result traces to its neuron.

At the bench and the sequencer



↳ **Alignment and counts, continuing from the circled SequencedSample**

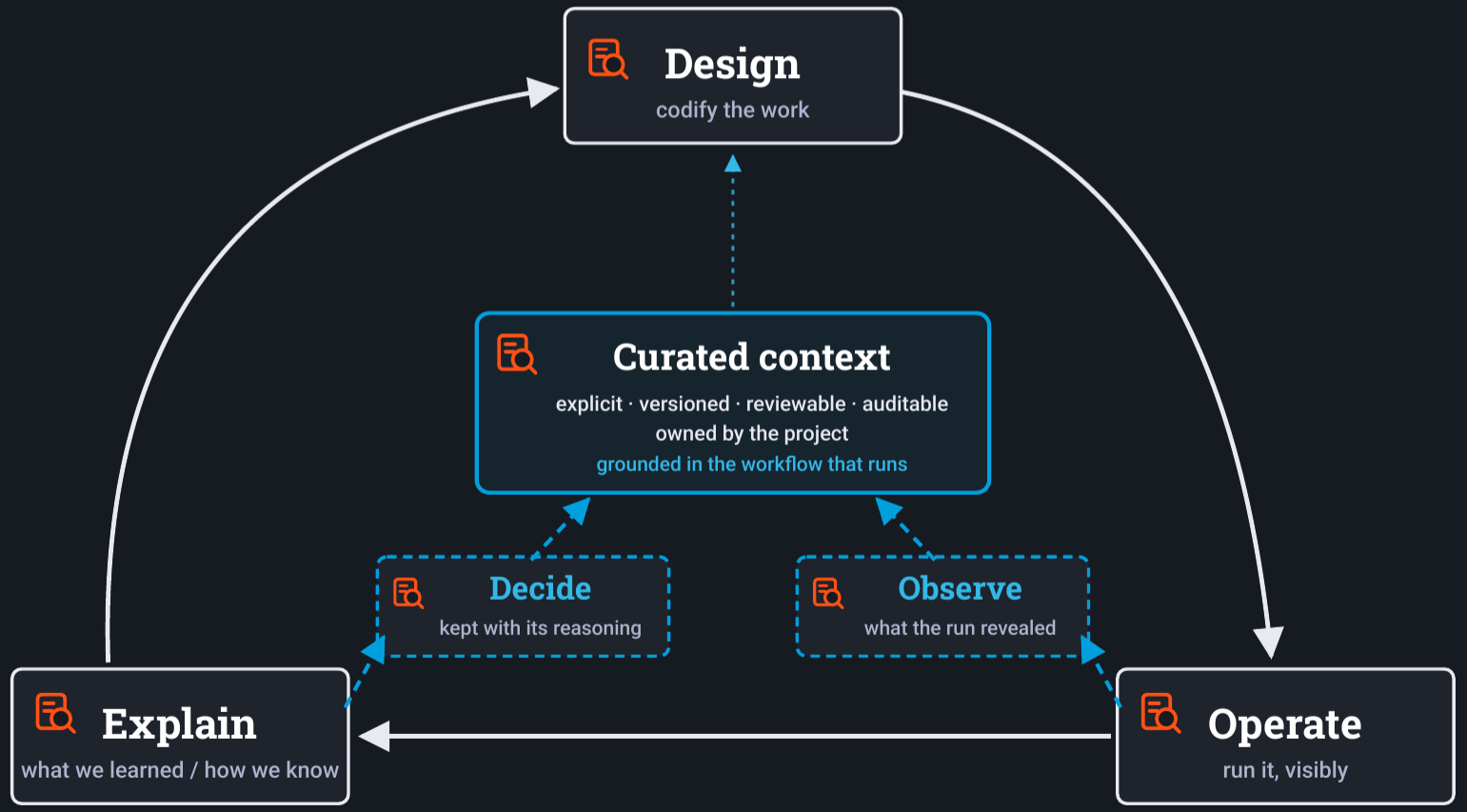


Every input behind one gene count, and nothing else, from the pipeline's own graph · **see it run at booth 5**

© 2026 DataJoint Inc.

DataJoint

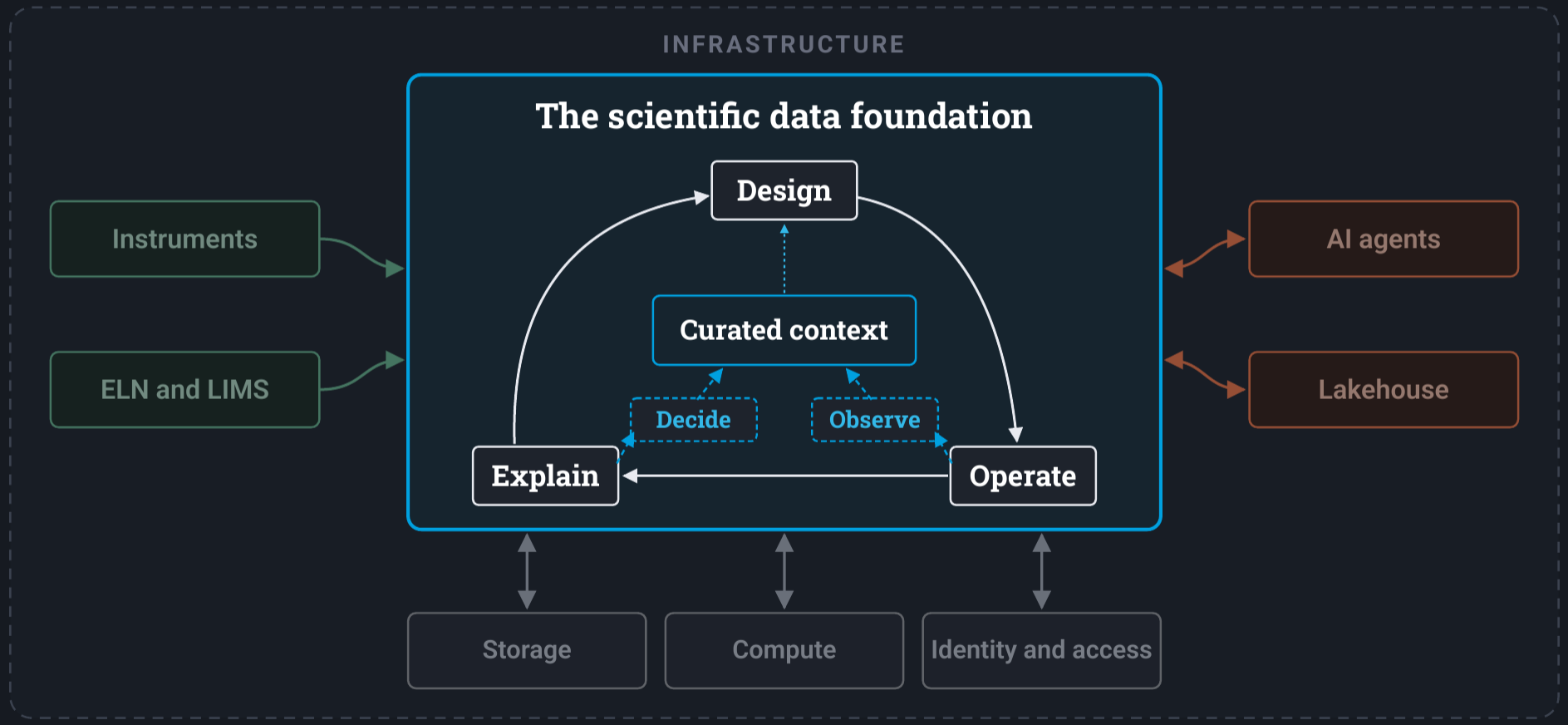
We are building curated context.



© 2026 DataJoint Inc.

DataJoint

The cycle runs in your infrastructure.



DataJoint · The Scientific Data Foundation for Accelerated Life Sciences R&D.

© 2026 DataJoint Inc.

DataJoint

Codified workflows reproduce the scientists' work.

A top-five global pharmaceutical company · inside the company's own cloud, with no data egress by design

Flow cytometry

Exact match

once an unwritten rounding convention was made explicit

Multiplex immunoassay

Exact on the limits

concentrations within 0.005%, below the noise floor

qPCR

Fully explained

every remaining difference traced to its source

© 2026 DataJoint Inc.

DataJoint

Agents are ephemeral. The workflow holds the record.

> People answer for every consequential decision; agents run between checkpoints.

> Agents ground in formal workflows, with integrity **before the run**.

> Trustworthiness is then **built in**, not simply asserted.

What grounds your agents, and where does their work go?

© 2026 DataJoint Inc.

Better science in.
Better intelligence out.



See it run at **booth 5**

© 2026 DataJoint Inc.