Diskover

SOLUTION BRIEF | AUTONOMOUS VEHICLE AND SENSOR DATA

Move only the recordings your models still need.

Read each recording where it lands, add the context it lacks, and copy only the data that fills a gap.

500 TB a day expected from a planned 100-vehicle fleet

<9% of 2,067 recordings named for one training need: rain

1,000 recordings with a suspect clock, flagged before any training

At a glance

KEY CHALLENGES

Test fleets write terabytes of camera, LiDAR, radar, and GPS data at every site, every day. The default plan copies all of it to headquarters, at a cost in bandwidth, transfer windows, and duplicate storage. AI teams wait weeks for data the model might not need.

THE SOLUTION

Diskover indexes the storage at each site and reads every MCAP recording during the scan. A workflow step adds outside context, such as the weather at capture. AI teams search the enriched metadata catalog against their own criteria, and a policy copies only the recordings they tag.

WHY IT MATTERS

Gartner predicts that through 2026, organizations will abandon 60% of AI projects unsupported by AI-ready data. Choosing recordings from metadata keeps transfers small, catches bad clocks and missing GPS positions before training, and shows capture teams which conditions the fleet still needs.

Why Diskover

READS THE RECORDING, NOT ONLY THE FILE

A plugin opens each MCAP file during the scan and writes fields such as sensors, capture time, duration, and GPS position to the catalog. The plugin is extensible, so fields are added as a team's questions change.

CATCHES PROBLEMS BEFORE TRAINING

The same scan flags recordings with suspect clocks or missing GPS positions, so teams fix them or set them aside before a model learns from them.

ADDS WHAT THE SENSORS NEVER RECORDED

Place and time work as a key into outside data. A workflow step adds the weather for each recording's capture hour and location, and the same join works with test schedules, road closures, or maintenance logs.

OPEN TO YOUR DATA STACK

A documented REST API exposes the index, tags, and jobs. The catalog runs on Elasticsearch for direct queries, accepts fields from your own extraction models, and exports to Parquet files that Snowflake and Databricks ingest.

ONE CATALOG ACROSS EVERY SITE

One central Diskover instance holds the catalog and dashboards, while Diskover indexes the storage at each remote site. Metadata travels to the center on every scan. Recordings stay at the edge until a policy calls for them.

ASK IN PLAIN LANGUAGE

Ask the AI Connector, "Which recordings were captured in rain with LiDAR and radar?" and get the matching recordings back with a recommended tag, applied only after a person says yes. AI-assisted, human-approved.

About Diskover

Diskover helps enterprises structure unstructured data across cloud, on-prem, and edge environments, making it easier to search, govern, optimize, and deliver file data for compliance, analytics, and AI-ready workflows.

Contact

diskoverdata.com
sales@diskoverdata.com
(800) 560-5853
San Jose, CA

Diskover

SOLUTION BRIEF

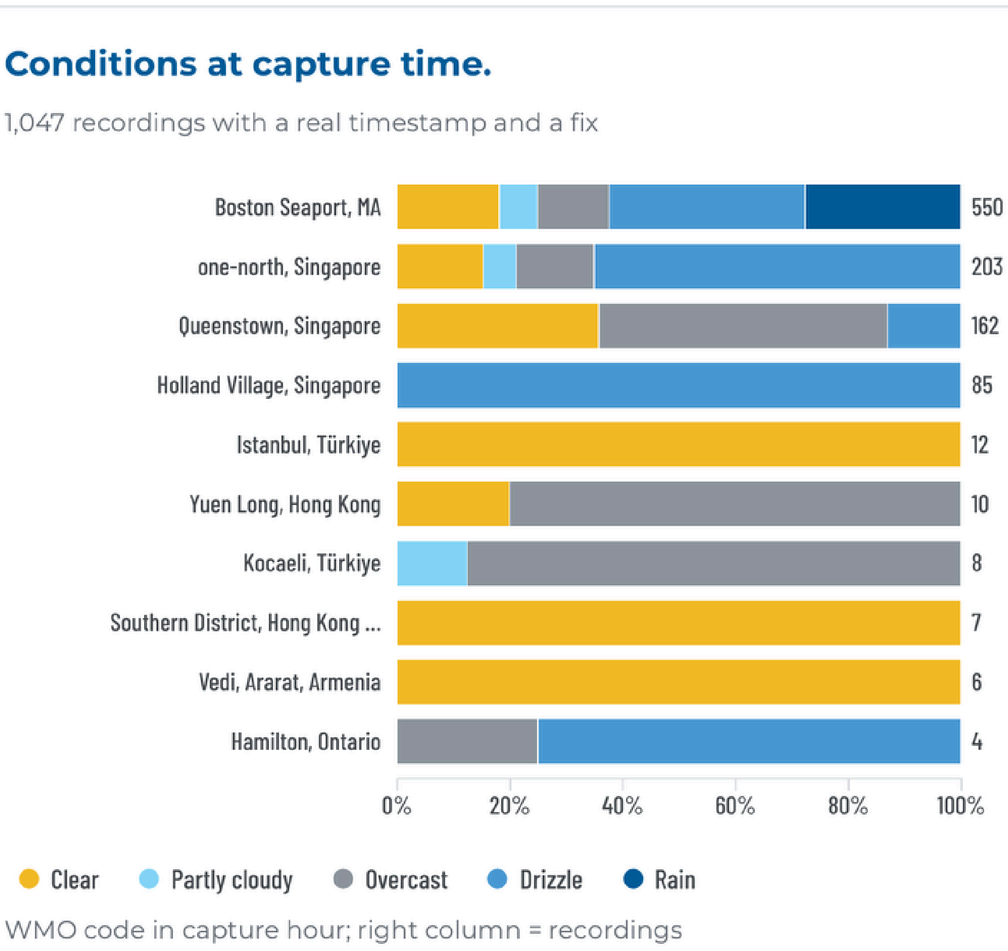
One scan. 2,067 recordings. Every gap visible.

A global automaker asked a simple question: which recordings do our models still need? Its own recordings stayed private, so Diskover's Director of Digital Archaeology gathered nine open robotics and self-driving datasets stored in or converted to MCAP format, from test vehicles to drones and capture rigs, and ran one scan with the MCAP plugin.

2,067	2.07 TB	18.7 h	336 km	99%	1,047
recordings	stored	of recorded time	driven and flown	with a GPS position	matched to weather

Conditions at capture time.

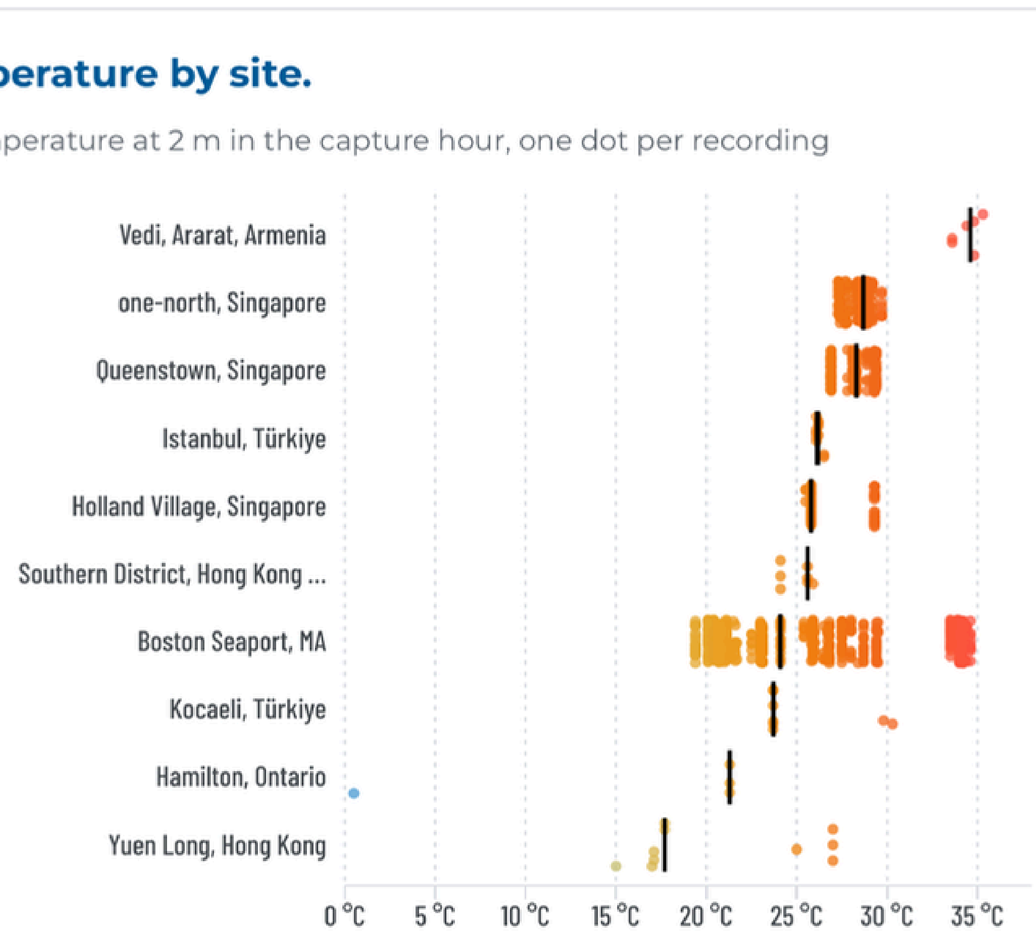
1,047 recordings with a real timestamp and a fix



WMO code in capture hour; right column = recordings

Temperature by site.

Air temperature at 2 m in the capture hour, one dot per recording



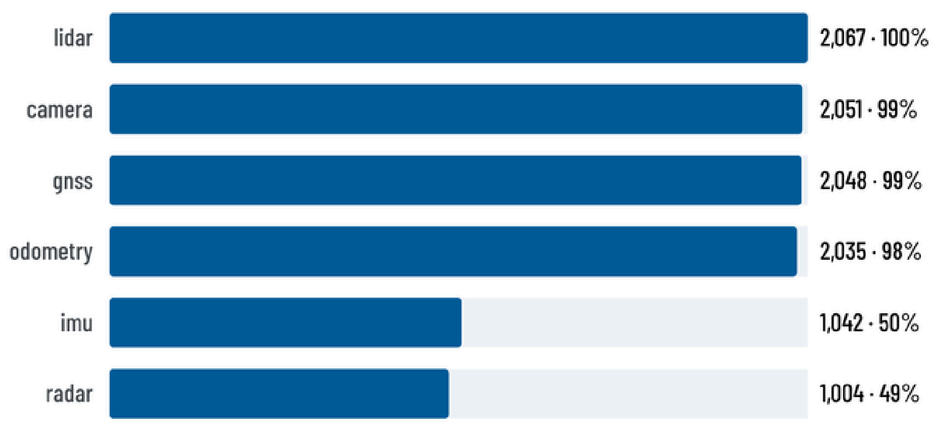
Weather data: Open-Meteo.com, CC BY 4.0

Find the condition your model lacks. Boston shows the full range, from clear to rain. Istanbul, the Southern District of Hong Kong, and Vedi in Armenia were clear in every matched recording, so a rain-focused rule leaves those recordings where they are.

See a gap the team had not asked about. Boston drives span 19 to 35 °C and Singapore sits between 26 and 30 °C. One drive is near freezing. Among the recordings with weather attached, there is no snow and no sub-zero driving. If winter matters, winter recordings become the next candidates.

Sensor coverage.

Recordings carrying each sensor stream



Know the sensor mix. Every recording carries LiDAR and nearly every one carries camera and GPS, but only half carry radar or an inertial measurement unit (IMU).

CAUGHT BEFORE TRAINING

1,000 suspect clocks
One dataset publishes time relative to the GPS epoch, so capture dates and the weather on those dates cannot be recovered.

20 with no GPS position
Held back from the weather lookup, which needs both a place and a time.

A label, not a measurement. The weather comes from an archive for each recording's start hour, so it is a way to find candidates.

About Diskover

Diskover helps enterprises structure unstructured data across cloud, on-prem, and edge environments, making it easier to search, govern, optimize, and deliver file data for compliance, analytics, and AI-ready workflows.

Contact

diskoverdata.com
sales@diskoverdata.com
(800) 560-5853
San Jose, CA

Diskover

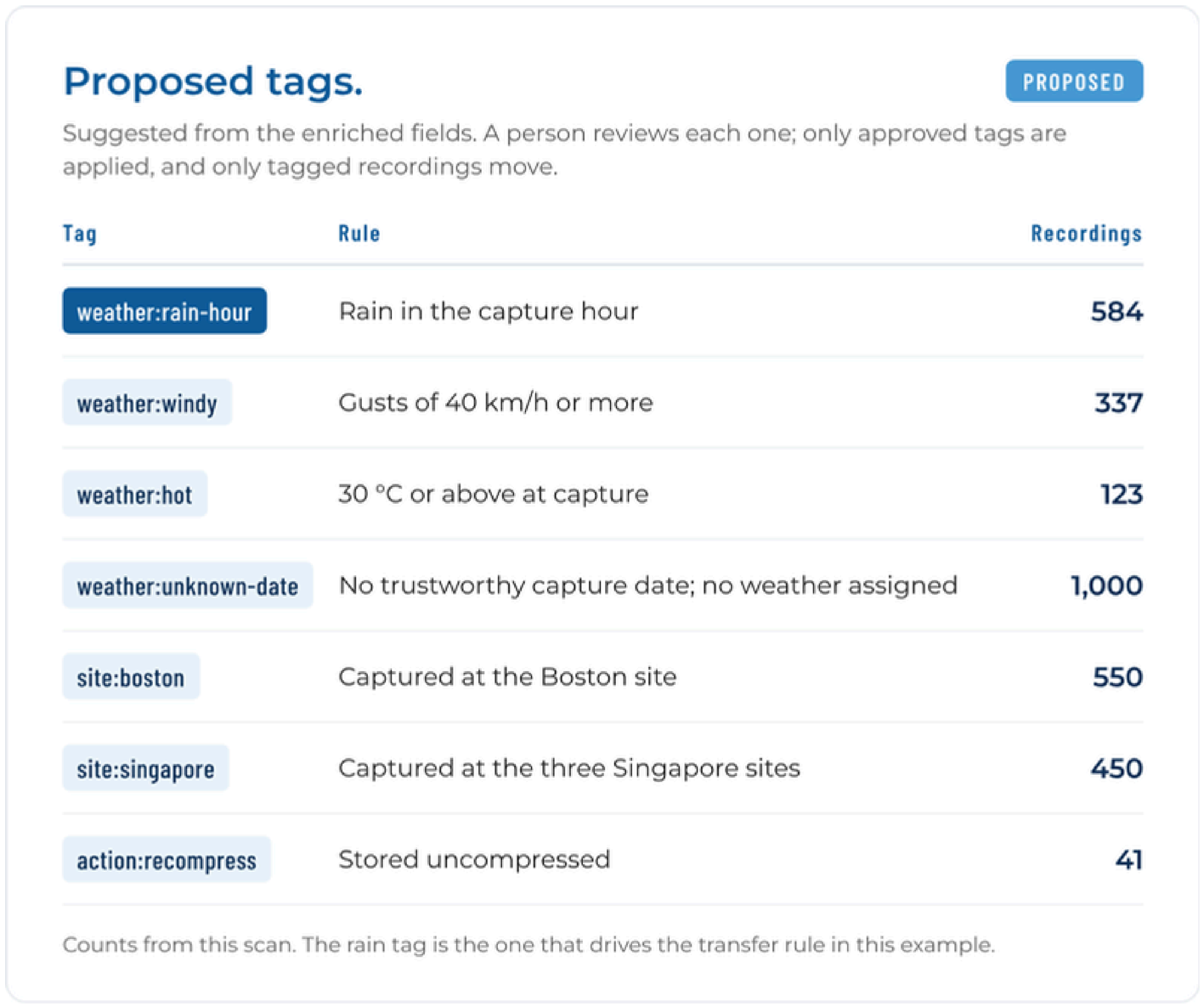
SOLUTION BRIEF

From proposed tag to delivered file.



One afternoon's driving, named file by file. 179 Boston recordings fell in a wet hour, while the weather archive showed rain over the site.

A chart that names files is half the workflow. The other half delivers them, without anyone pasting file paths into a script. In the demo, Diskover proposed tags from the enriched fields, and each one stayed a proposal until a person approved it. A Diskover policy then delivers a copy of each tagged recording to the training store.



Tags waiting for approval: 584 recordings in a wet hour, 337 in gusts of 40 km/h or more, and 123 above 30 °C.

Questions AI architects ask.

Can it read our recordings, or do we send our pipeline's output?

Either path works. A plugin reads the format during the scan, or your team brings the fields its extraction pipeline already produces. Both land in the same catalog.

Can we take the catalog out?

Yes. A documented REST API exposes the index, tags, and jobs, the catalog runs on Elasticsearch, and metadata exports to Parquet for Snowflake and Databricks.

Our value sits deep inside the file. Can we bring our own models?

Yes. Fields your models produce can be added to the catalog alongside the plugin's, then drive search, dashboards, tags, and moves like any other field.

What other dense formats can it read?

Diskover offers plugins for many industries, including genomics files, media containers, images, and PDFs, and teams build their own for formats unique to their work.

> *If I can transfer only the data containing rain back home, rather than all of this data, it makes a world of difference.*
>
> **AI Data Architect, Global autonomous automaker customer**

About Diskover

Diskover helps enterprises structure unstructured data across cloud, on-prem, and edge environments, making it easier to search, govern, optimize, and deliver file data for compliance, analytics, and AI-ready workflows.

Contact

diskoverdata.com
sales@diskoverdata.com
(800) 560-5853
San Jose, CA

Diskover

SOLUTION BRIEF

How it works

1 INDEX
STORAGE OR IT TEAM

Point Diskover at the storage at each test site, such as NAS, object, cloud, or edge servers. Indexing runs where the recordings land, and the data stays in place.

Example: a test site's file share is cataloged where it sits.

2 ENRICH
AUTOMATION

The MCAP plugin writes fields such as sensors, capture time, and GPS position to each record. A workflow step then adds the weather for that place and hour.

Example: a recording now shows LiDAR, radar, 18:30 UTC, and heavy rain.

3 SELECT
ANY AUTHORIZED TEAM

Search and chart the catalog against the AI team's own criteria. Tag recordings directly, or ask the AI Connector, which proposes tags and applies them once a person approves.

Example: an AI lead tags 584 recordings captured in a wet hour.

4 DELIVER
BY POLICY

A Diskover policy delivers a copy of each tagged recording to the training store. Every other recording stays at the edge until its retention period ends.

Result: only the rain recordings reach the AI team, ready for training.

WORKS ACROSS Cloud • On-prem • Hybrid • NAS • SAN • Object • Tape • Archive • Data lakes • Lakehouses • AI/BI/ML

Decide centrally. Keep data at the edge.

Test site A
Local file share
NAS

Test site B
Daily fleet uploads
OBJECT

Test site C
Trackside servers
EDGE

Training store
Tagged recordings only
CLOUD

Diskover

ENRICHED METADATA CATALOG
Examples: sensors • capture time • GPS position • platform • weather at capture
Metadata travels on every scan. Recordings stay at the edge.

Find the gap
Coverage by site, sensor, and weather.

Approve the tags
Applied only after a person says yes.

Deliver by policy
Only tagged recordings are delivered.

Key outcomes

SMALLER TRANSFERS

Only the recordings that fill a gap leave the site. Bandwidth, transfer windows, and storage go to the data the models need, which also keeps costs down.

CLEANER TRAINING DATA

Suspect clocks and missing GPS positions surface before training, so teams fix them first, not weeks into a training run.

FASTER AI CYCLES

AI teams search one catalog across every site and get named candidates from metadata, so the next run starts with the data the model lacks.

GET STARTED

Ready to see it on your recordings?

Start with a 30-day proof of concept on your own data, and recover your investment in 3 to 6 months.

Scan to **book a demo**

Scan to **watch the video**

About Diskover

Diskover helps enterprises structure unstructured data across cloud, on-prem, and edge environments, making it easier to search, govern, optimize, and deliver file data for compliance, analytics, and AI-ready workflows.

Contact

diskoverdata.com
sales@diskoverdata.com
(800) 560-5853
San Jose, CA

Revised November 2026