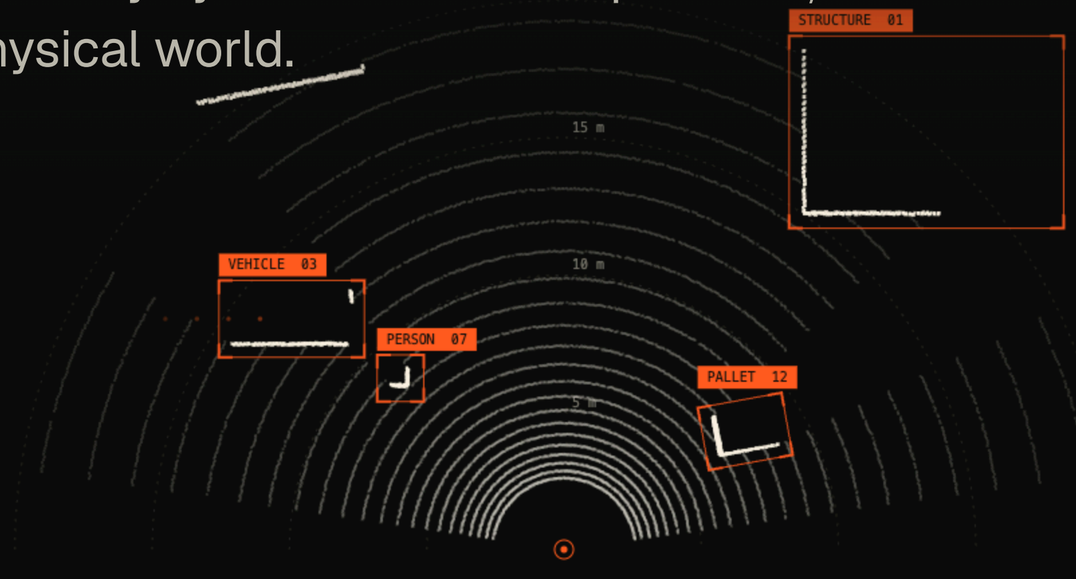
INVESTOR OVERVIEW

Training data for the real world.

Real-world sensor data for any system that has to perceive, move or decide in the physical world.



AI is moving into the physical world. Its data is not ready.

01

Scarce

Unlike text and images, real-world sensor and action data does not exist at internet scale. It has to be collected on real hardware, in real places.

02

Hard to verify

A 10 millisecond timing offset or a noisy label is invisible in a demo and expensive in deployment. Most buyers cannot audit how their data was made.

03

Hard to handle

Defense, healthcare and other controlled programs cannot route sensitive collections through generic crowd pipelines. Handling has to be designed in.

THE GAP

There is no internet for the physical world.

~300,000

hours of robot manipulation data in the world, estimated

VS

~1 billion

hours of internet video available to train on

Robot data is about 0.03% of the internet video corpus

Source: Bessemer Venture Partners, Robotics and Physical AI (estimate).

Training data became a multi-billion dollar business.

Bases differ: recognized vs annualized, gross vs net. Estimates, not audited, and not directly comparable. USD billions.

The market is already bigger than the reports say.

$3.4B to $4.6B

Analyst estimate, 2026

Six research firms size AI training datasets in this range, growing roughly 16% to 23% a year.

~$5B+

Four vendors' annualized revenue

Scale AI, Surge AI, Mercor and Handshake alone, on gross, estimated figures. Net revenue is lower.

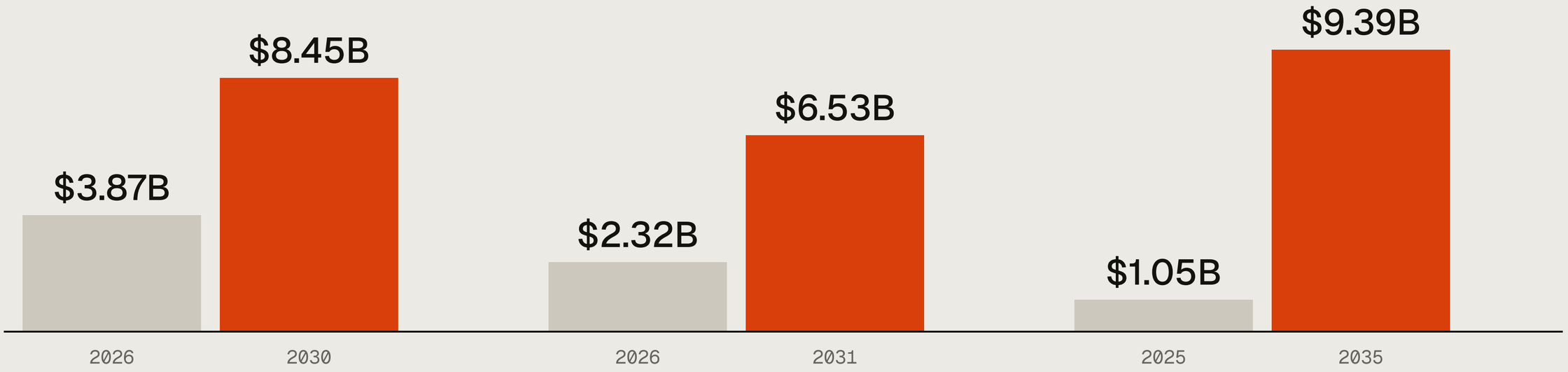
$1B+

Reported per frontier lab, per year

Widely repeated figure for human data spend at leading labs. It traces to a few sources, so treat as indicative.

Sources: Research and Markets and five other firms; Sacra; Forbes; TIME via micro1; Labelbox. See Sources.

Analysts expect 20%+ annual growth.



AI training datasets
21.6% CAGR. Research and Markets

Data labeling services
22.95% CAGR. Mordor Intelligence

Robot training data
24.5% CAGR. Emergen Research

Scopes differ and should not be added together. Labeling services exclude raw data collection and synthetic data. USD billions.

Demand is moving to real-world data.

Rising

- Egocentric and robotics demonstration data
- Expert judgment in professional domains
- Wet-lab and experimental science data
- RL environments

Cooling

- Generalist annotation and microtasks
- Static math and verifiable coding tasks
- Commodity labeling priced on volume

Oct 2026

Classification of lab data listings and reported deals, covering Jun 2023 to Oct 2026.

Source: aitraining.jobs, What AI Labs Are Buying.

Investors are paying up for data.

COMPANY	WHAT IT DOES	REPORTED VALUATION	BASIS
Scale AI	Data labeling and evaluation	~$29B	Implied by Meta's $14.3B purchase of 49%, Jun 2025
Surge AI	Expert human data	~$24B	Forbes estimate, Sep 2025. No outside capital raised
Mercor	Expert data marketplace	$10B	$350M Series C, Oct 2025. Reported talks at $20B
Physical Intelligence	Robot foundation models	~$20B	Reported talks, Oct 2026, to raise $1B or more
Mecka AI	Human motion data for robots	~$500M	Reported Sequoia-led round, Sep 2026, three months after a $60M round

Valuations are reported or estimated, not audited, and several are talks rather than closed rounds. Shown for context, not as comparables for Cato Labs.

Every robot that ships needs data to learn from.

6.5M

Humanoid shipments forecast for 2035

Goldman Sachs, Sep 2026, up from a previous forecast of about 1.4 million.

890,000

Forecast for 2030

Goldman Sachs. Its earlier 2030 forecast was 256,000, so the estimate has more than tripled.

50,000

China humanoid shipments, 2026

Morgan Stanley, Jun 2026, raised from 14,000 at the start of the year.

Sources: Goldman Sachs and Morgan Stanley forecasts, as reported by 24/7 Wall St., Humanoid Guide and SCMP. See Sources.

SIZING

A small share of this market is a large business.

$85M

1% of the 2030 market

1% of the $8.45B AI training dataset market that Research and Markets forecasts for 2030.

$254M

3% of the 2030 market

Arithmetic on the same forecast. An illustration of scale, not a revenue forecast.

$423M

5% of the 2030 market

The physical-world segment is forecast to grow faster than the whole, at about 24.5% a year.

Sources: Research and Markets (21.6% CAGR, $8.45B in 2030); Emergen Research (24.5% CAGR, robot training data).

Incumbents were built for screens, not sensors.

01

Built for text

Leading vendors scaled on text, code and expert judgment: people at laptops. Physical data needs hardware, sites and safety.

02

Cannot be bought or scraped

Robot and sensor data has to be created for each task and environment. There is no corpus to license.

03

Capture competes on volume

Many capture vendors sell hours. Rates are falling, and volume alone is a weak moat.

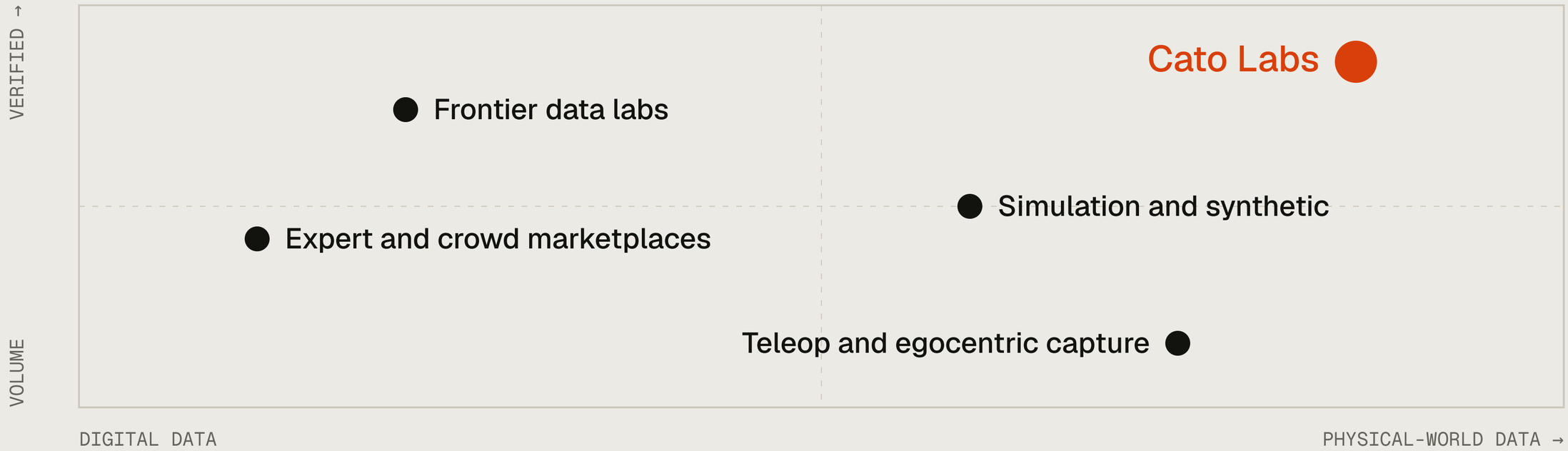
04

Controlled domains are underserved

Defense, healthcare and critical infrastructure cannot use open crowd pipelines. Synthetic data does not fully replace real data.

HOW WE ARE DIFFERENT

The verified real-world data layer.



Our view of the landscape, by category. Illustrative positioning, not a measurement.

One accountable process, from capture to benchmark.

01

Capture

Synchronized, calibrated multi-sensor recording on your hardware or ours.

02

Annotate

Labels against a versioned schema, reviewed and adjudicated.

03

Evaluate

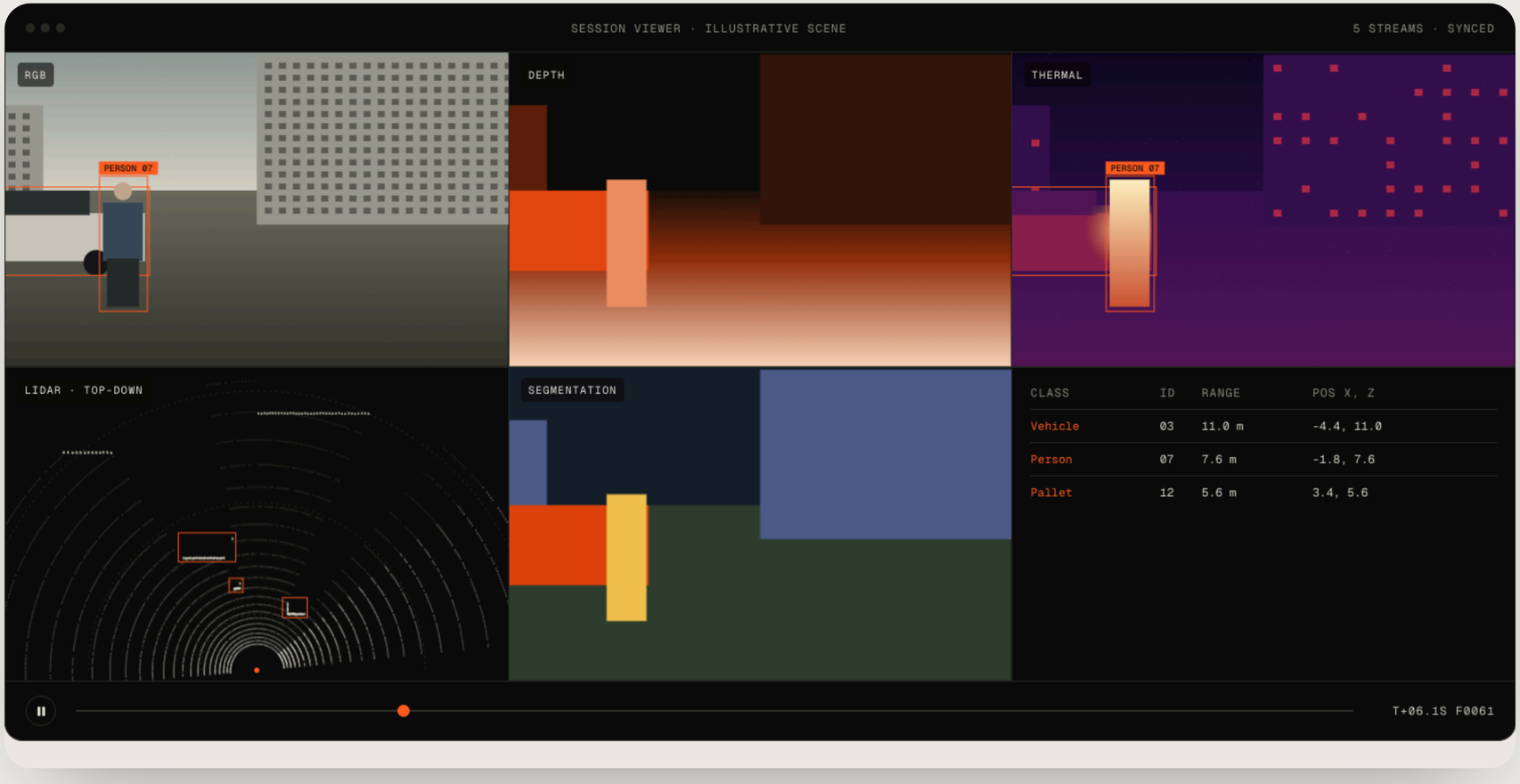
Held-out benchmarks built to your deployment conditions.

04

Deliver

Versioned releases in the formats your pipeline already reads.

What the data looks like.



Illustrative scene. Five synchronized streams on one timeline, with labels that line up across sensors.

HOW IT WORKS

From a spec to a dataset you can train on.

01

Scope

Task, sensors, environments, schema and acceptance criteria agreed first.

02

Collect

Field capture against a written plan, with every session logged.

03

Annotate

Labeling against guidelines that are versioned and shared.

04

Verify

Multi-pass review and expert adjudication. QA ships with the data.

05

Deliver

Versioned releases with a changelog for every update.

WHERE IT IS USED

Six domains, one standard.

Robotics

Manipulation, navigation and humanoid demonstrations.

Autonomy and mobility

LiDAR and camera data for ground, aerial and maritime vehicles.

Defense and security

Perception in degraded, contested conditions.

Industrial and infrastructure

Inspection of plants, pipelines, grids and sites.

Healthcare and life sciences

Device and procedure capture where consent and compliance allow.

Geospatial and environment

Aerial and satellite imagery, mapping and change detection.

23% of data labeling revenue comes from automotive and mobility, the largest vertical. Healthcare is the fastest growing, at 24.6% a year. Video, including LiDAR work, grows 31% a year.

Source: Mordor Intelligence, AI Data Labeling Market.

BUSINESS MODEL

Paid pilots that grow into recurring programs.

01

Pilot

A fixed-fee, scoped dataset with written acceptance criteria. Low risk for the buyer, and our proof of quality.

02

Program

Recurring collection and annotation against an evolving spec, on annual or multi-quarter contracts.

03

Evaluation

Versioned, held-out benchmarks sold as a subscription, so customers can measure progress release to release.

04

Controlled delivery

Premium scoping for defense, healthcare and infrastructure work that needs special handling.

Land with a pilot, win on evidence, expand by domain.

01

Target

Robotics, autonomy and defense teams already buying data and short of verified physical-world data.

02

Land

A paid pilot against a written spec, sold founder-to-buyer.

03

Prove

A held-out evaluation that shows the data improves the customer's model.

04

Expand

Move from a pilot to a recurring program, then add sensors and sites.

05

Cross-sell

Carry the same process into adjacent domains, from industrial to healthcare.

TEAM

Founders who sell, source and hold the line on evidence.

Garrett Martel

Investment work in venture capital and private equity at Fidelity

Joined Corgi on the go-to-market team

Oz Harkavi

Served in the Israel Defense Forces

Worked at Corgi

$1M+

in revenue sourced together in six months at Corgi.

Why it matters: this business is won by sourcing buyers and running a tight pilot, and defense buyers want to see handling taken seriously.

WHAT SETS US APART

Built on standards, not promises.

01

Synchronized by design

Hardware triggering or a shared clock, with both sensor and host timestamps logged.

02

Failures are kept

Failed attempts are labeled and retained, so policies learn recovery.

03

Every label has an owner

Versioned guidelines, independent review and documented adjudication.

04

Handling comes first

Controlled and restricted requirements are scoped before collection begins.

Funded players cluster on raw capture.

CATEGORY	EXAMPLES AND DISCLOSED FUNDING	STRENGTH	WHERE CATO FITS
Annotation platforms	Encord ($110M), Scale AI	Tooling and scale for labeling	Real-world capture and evaluation in one process
Teleoperation and capture vendors	Tacta ($75M), XDOF ($70M), Mecka (~$68M), Config ($35M), Micro1 ($35M)	Demonstrations collected at volume	Provenance, review and controlled handling as standard
Simulation and synthetic data	Lightwheel, Applied Intuition, DiffuseDrive	Generating scenarios cheaply	Real captures to validate how well simulation transfers

Disclosed funding as of Aug 2026, per Teahose and company announcements. Named for context only, not a claim about any company's full offering.

RISKS

What could go wrong, and what we do about it.

RISK	WHY IT MATTERS	HOW WE RESPOND
Labs build it in-house	The largest buyers can hire their own collectors.	Be the verified outside layer: provenance, review and evaluation that internal teams rarely staff.
Price compression	Capture vendors compete on hours, and rates are falling.	Sell accepted, evaluated datasets and subscriptions, not raw hours.
Simulation replaces real data	Synthetic data keeps improving.	Sell the real captures that validate simulation. The two are complements.
Operations at scale	Field collection is labor-heavy and hard to standardize.	Written protocols, a logged chain of custody and staged growth by domain.

Our assessment. Market risks are drawn from the competitive landscape on the previous slides.

USE OF FUNDS

What this round pays for.

01

Collection and annotation

Capture capacity, operator training and review teams, so pilots turn around quickly and to a consistent standard.

02

Evaluation and tooling

The pipeline that versions datasets, scores them against held-out sets and ships quality results with every release.

03

Security and go-to-market

Controlled-handling practices that unlock defense and healthcare buyers, and a sales motion built on paid pilots.

Where the numbers come from.

Vendor revenue: Scale AI, Mercor	Sacra	sacra.com/c/scale-ai, sacra.com/c/mercor
Vendor revenue and valuation: Surge AI	Forbes, via Contrary Research	research.contrary.com/company/surge-ai
Scale AI valuation and Meta transaction	Teahose	teahose.com/guides/scale-ai-valuation
Mercor round and reported talks	ValueAdd VC	valueaddvc.com/blog (Mercor valuation)
Demand by data category; Mecka and Snorkel run rates	aitraining.jobs, What AI Labs Are Buying	aitraining.jobs/research/what-ai-labs-are-buying
Robot data scarcity (300,000 hours vs ~1B hours)	Bessemer Venture Partners	bvp.com/atlas/bessemer-predicts-robotics-and-physical-ai
AI training dataset market, $3.87B (2026) to $8.45B (2030); range $3.4B to $4.6B	Research and Markets; Fortune Business Insights, Grand View, Precedence, Straits, MarketIntelo	researchandmarkets.com/reports/5948794
AI data labeling services; vertical shares and growth	Mordor Intelligence	mordorintelligence.com/industry-reports/ai-data-labeling-market

SOURCES, 2 OF 2

Where the numbers come from.

Robot training data, $1.05B (2025) to $9.39B (2035)	Emergen Research, via OpenPR	openpr.com/news/4667974
Vendor funding; real-world data spend	Teahose, Robotics Training Data guide (Aug 2026)	teahose.com/guides/robotics-training-data
Mecka AI round near $500M	TechCrunch, 11 Sep 2026	techcrunch.com/2026/09/11/mecka-ai-nears-500m-valuation
Physical Intelligence talks near $20B	Bloomberg, 27 Oct 2026	bloomberg.com/news/articles/2026-10-27
Humanoid shipment forecast, 6.5M by 2035	Goldman Sachs, via 24/7 Wall St. and Humanoid Guide	humanoid.guide/goldman-raises-2035-humanoid-robot-forecast-to-6-5-million
China humanoid shipments, 50,000 in 2026	Morgan Stanley, via SCMP	scmp.com/tech/article/3358210
Teleoperator pay, $25 to $35 an hour (context for cost drivers)	Developer job postings, via vendor guides	roboticscenter.ai/blog/robot-data-collection-cost
Team background and the $1M+ revenue figure	Company information	Provided by the founders

Tell us what your system needs to see.

operations@catolabs.co catolabs.co