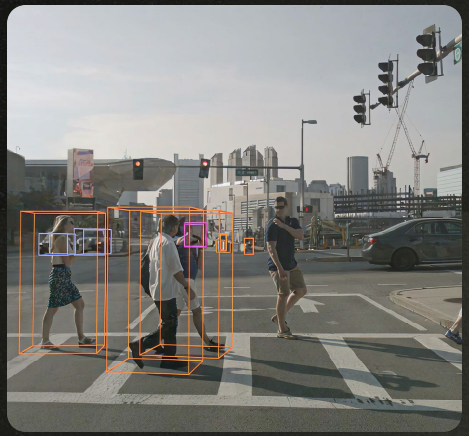
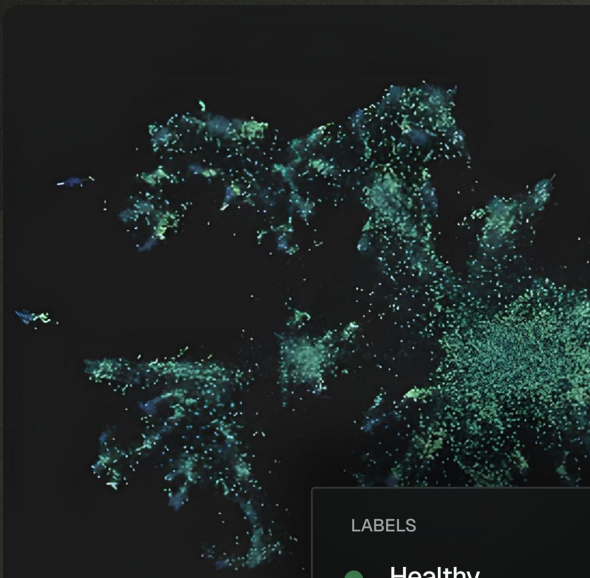
VOXEL51

Scaling Computer Vision in the Enterprise





```python
import fiftyone.brain as fob
from fiftyone import ViewField as F

fob.compute_mistakenness(
    dataset,
    "predictions",
    label_field="ground_truth"
)

mistake_view = dataset.sort_by("mistakenness", reverse=True)
session.view = mistake_view
```

LABELS

- Healthy
- Leaf Spot
- Blight
- Pest Infestation

VOXEL51

Why Data-Centric Computer Vision?

Agenda

- Why data-centric computer vision?
- Dataset management at scale
- Data annotation and auto-labeling
- Out-of-the-box workflows
- Collaborating on datasets
- Safety and security
- Next steps

DATA EATS MODELS FOR LUNCH

Unlocking the potential of your data is key to competitive advantage

“I do believe the models are getting commoditized…models by themselves are not sufficient, but having a full system stack and great successful products, those are the two places” where companies need to focus now.”

Satya Nadella, Microsoft CEO

AI projects fail 85% of the time

Amazon's "Just Walk Out" store vision system could not handle the full diversity of real shopper behavior and product interactions. Human reviewers had to verify about 70% of transactions, because the AI wasn't accurate enough on its own.

GUARDIAN

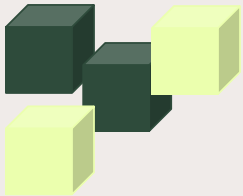
NHTSA has reports of 16 crashes, including seven injury incidents and one death, involving Tesla vehicles in Autopilot that had struck stationary first-responder and road maintenance vehicles.

REUTERS

GE's Caption AI miscalculated key heart metrics during ultrasound exams due to a data-handling bug that included incorrect video frames. The flaw triggered a Class II FDA recall.

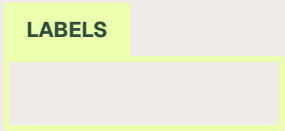
FDA

30% of model errors are due to bad data



Unstructured data is difficult to search, slice, and debug

Complex integration across modalities, labels, and metadata



Hard to annotate and validate at scale

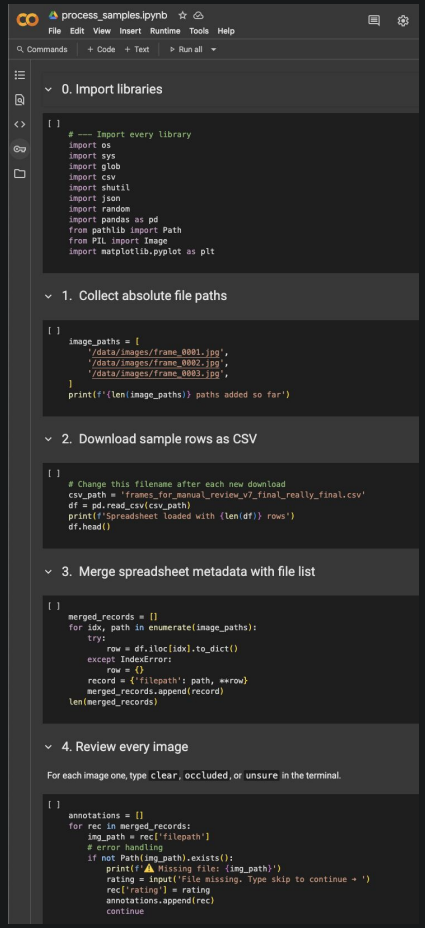
ML teams spend 39% of their time wrangling data

It is simply *impossible* to touch each data sample manually to ensure quality. Yet, even top teams have no good alternatives

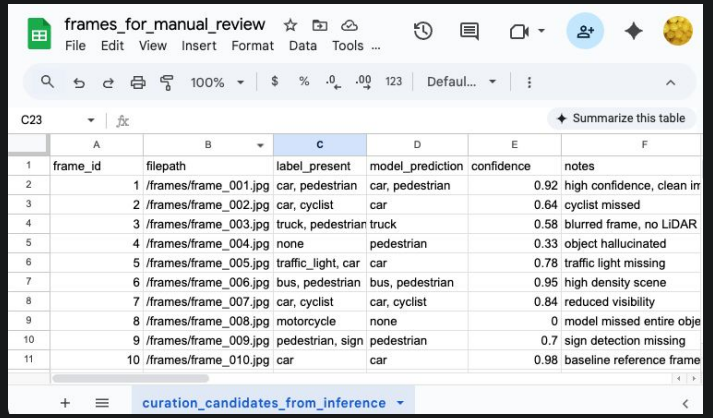
This is what it looks like when your data stack wasn't built for visual data.

It's time-consuming, painful, and impossible to scale.

Python scripts and Jupyter notebooks

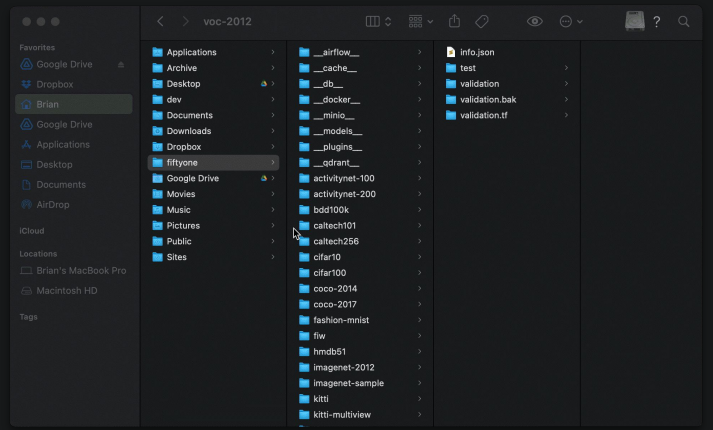


Spreadsheets



	A	B	C	D	E	F
1	frame_id	filepath	label_present	model_prediction	confidence	notes
2	1	/frames/frame_001.jpg	car, pedestrian	car, pedestrian	0.92	high confidence, clean im
3	2	/frames/frame_002.jpg	car, cyclist	car	0.64	cyclist missed
4	3	/frames/frame_003.jpg	truck, pedestrian	truck	0.58	blurred frame, no LiDAR
5	4	/frames/frame_004.jpg	none	pedestrian	0.33	object hallucinated
6	5	/frames/frame_005.jpg	traffic_light, car	car	0.78	traffic light missing
7	6	/frames/frame_006.jpg	bus, pedestrian	bus, pedestrian	0.95	high density scene
8	7	/frames/frame_007.jpg	car, cyclist	car, cyclist	0.84	reduced visibility
9	8	/frames/frame_008.jpg	motorcycle	none	0	model missed entire obje
10	9	/frames/frame_009.jpg	pedestrian, sign	pedestrian	0.7	sign detection missing
11	10	/frames/frame_010.jpg	car	car	0.98	baseline reference frame

Manual debugging loops



DATA CURATION

Garbage in, garbage out

Poor dataset curation leads to:

- Misleading metrics and **poor generalization** in production
- Missed edge cases that **compromise safety and robustness**
- Incorrect labels that **silently degrade model performance**

Your data

- Data collection
- Cloud storage
- Data lakes
- Vector DBs

FiftyOne

Curation

Inspect, manage, and visualize your datasets with full control at every step

Annotation

Streamline your **annotation workflow** with Verified Auto Labeling with built-in QA

Evaluation

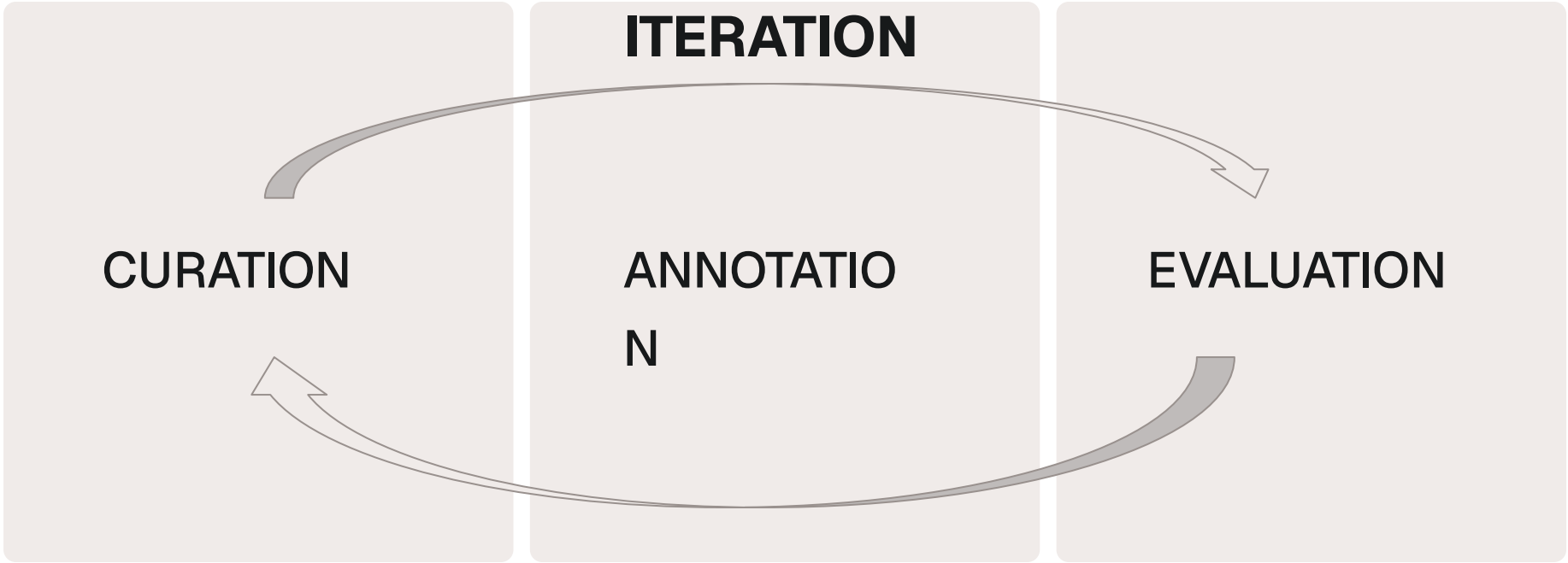
Confidently measure, compare, and debug **model performance**

Your models

- Training frameworks
- Model registry
- Model serving
- Model monitoring

Your compute

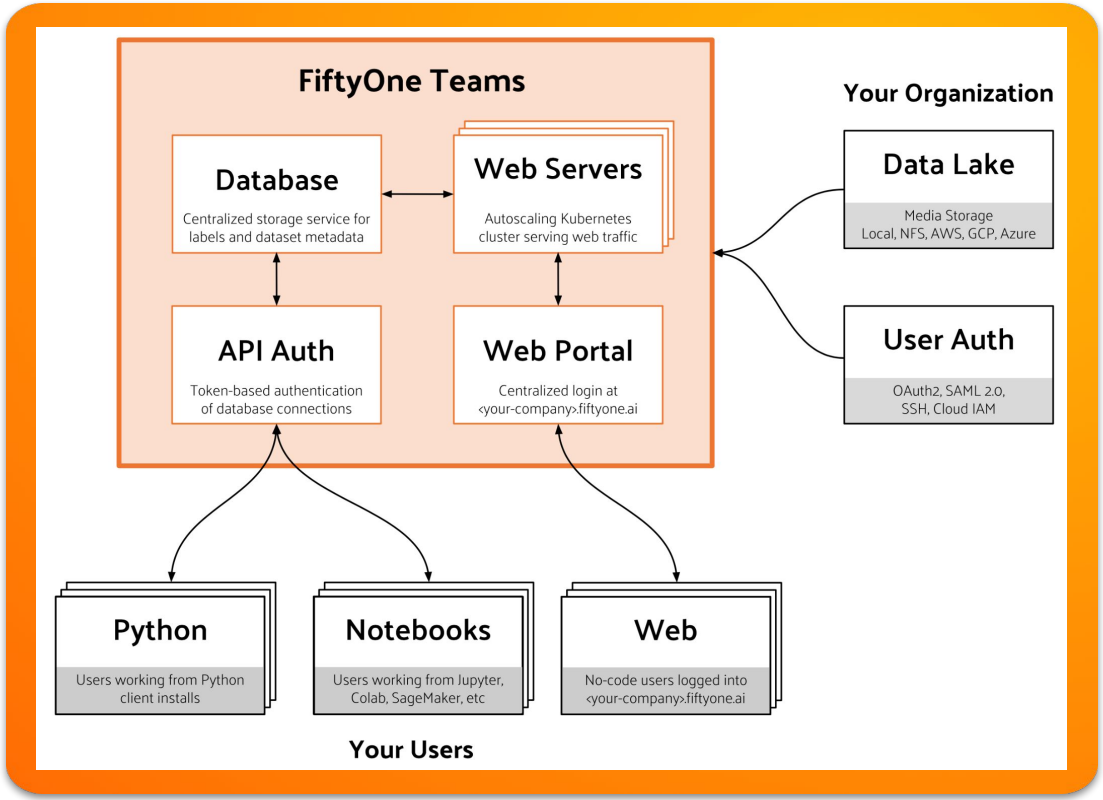
Building a Successful Pipeline



VOXEL51

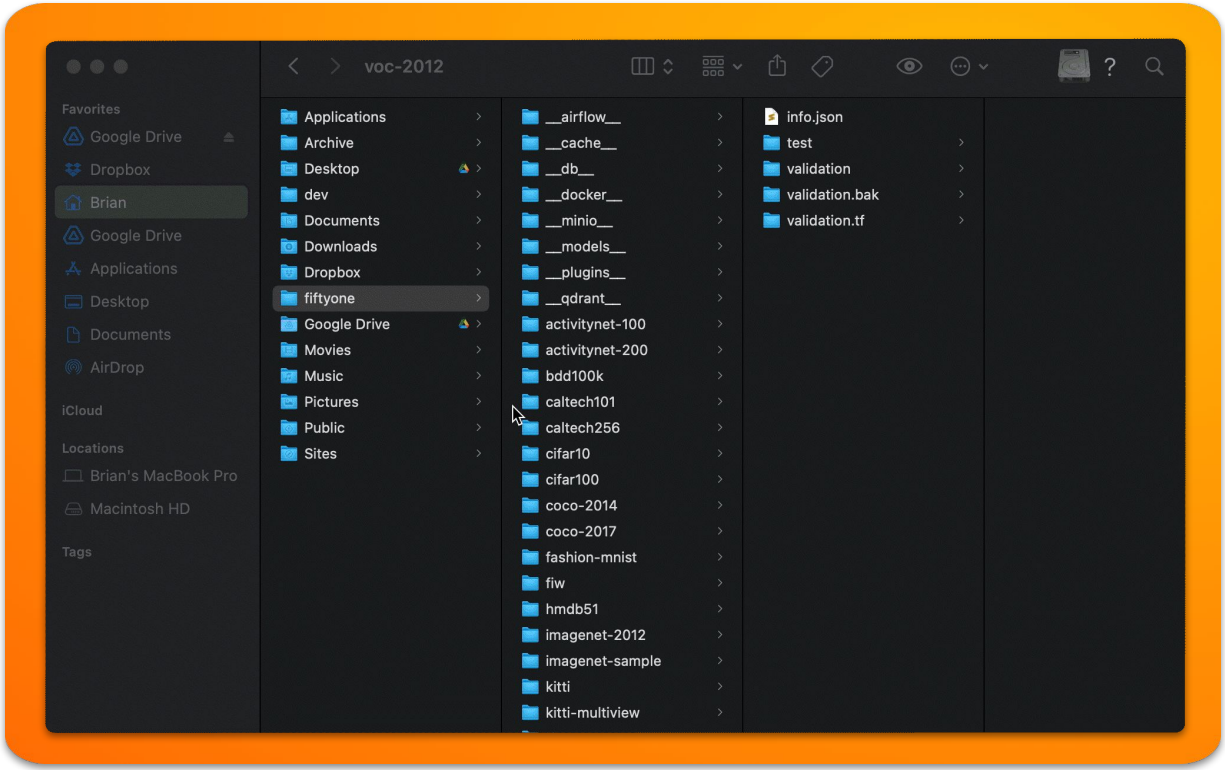
Dataset Management at Scale

Platform Architecture



Where Do Datasets Live?

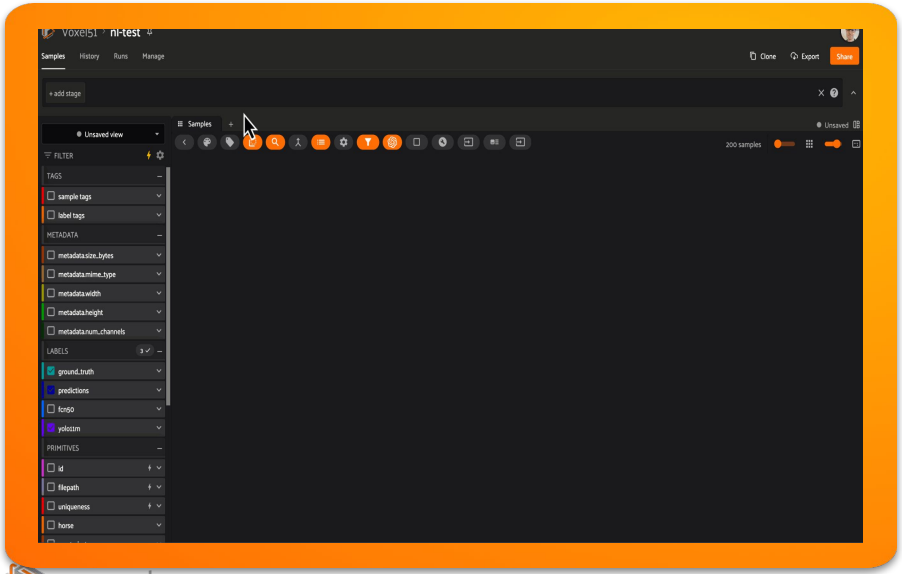
For DIY and some open source tools, it's (generally) on your computer's local filesystem.



Where Do Datasets Live?

For enterprise vendors, cloud-backed media or data lake

Cloud-Backed Media



Data Lake

Dataset Storage

Cloud-Backed Media

- ✓ Any S3-compatible object store
- ✓ Cached locally when running workflows requiring pixel-level access
- ✓ The sample's file path will be the object storage path
  - ❑ E.g., gs://path/to/sample
- ✓ Admins can configure cloud credentials in the App

Data Lake Integration

- ✓ Connects to and imports from your data lake (e.g., Databricks, BigQuery)
- ✓ Makes use of operators for defining connection to data lake, query parameters, preview behavior
- ✓ Import selected number of samples into new or existing dataset

Data Lens: Powered by FiftyOne Operators

```python
import json
import time
from typing import Generator

import fiftyone as fo
from databricks.sdk import WorkspaceClient
from databricks.sdk.service.sql import (
    StatementResponse, StatementState, StatementParameterListItem
)
from fiftyone import operators as foo
from fiftyone.operators import types
from fiftyone.operators.data_lens import (
    DataLensOperator, DataLensSearchRequest, DataLensSearchResponse
)
```

```python
class DatabricksConnector(DataLensOperator):
    """Data Lens operator which retrieves samples from Databricks."""

    @property
    def config(self) -> foo.OperatorConfig: ...

    def resolve_input(self, ctx: foo.ExecutionContext): ...

    def handle_lens_search_request( ...


class DatabricksHandler:
    """Handler for interacting with Databricks tables."""

    def __init__(self): ...

    def handle_request( ...

    def _init_client(self, ctx: foo.ExecutionContext): ...

    def _start_warehouse(self) -> str: ...

    def _iter_data(self, request: DataLensSearchRequest) -> Generator[dict, None, None]: ...

    def _transform_sample(self, sample: dict) -> dict: ...

    def _build_detections(self, sample: dict) -> fo.Detections: ...

    def _response_to_dicts(self, response: StatementResponse) -> list[dict]: ...

    def _check_for_error(self, response: StatementResponse): ...
```

VOXEL51

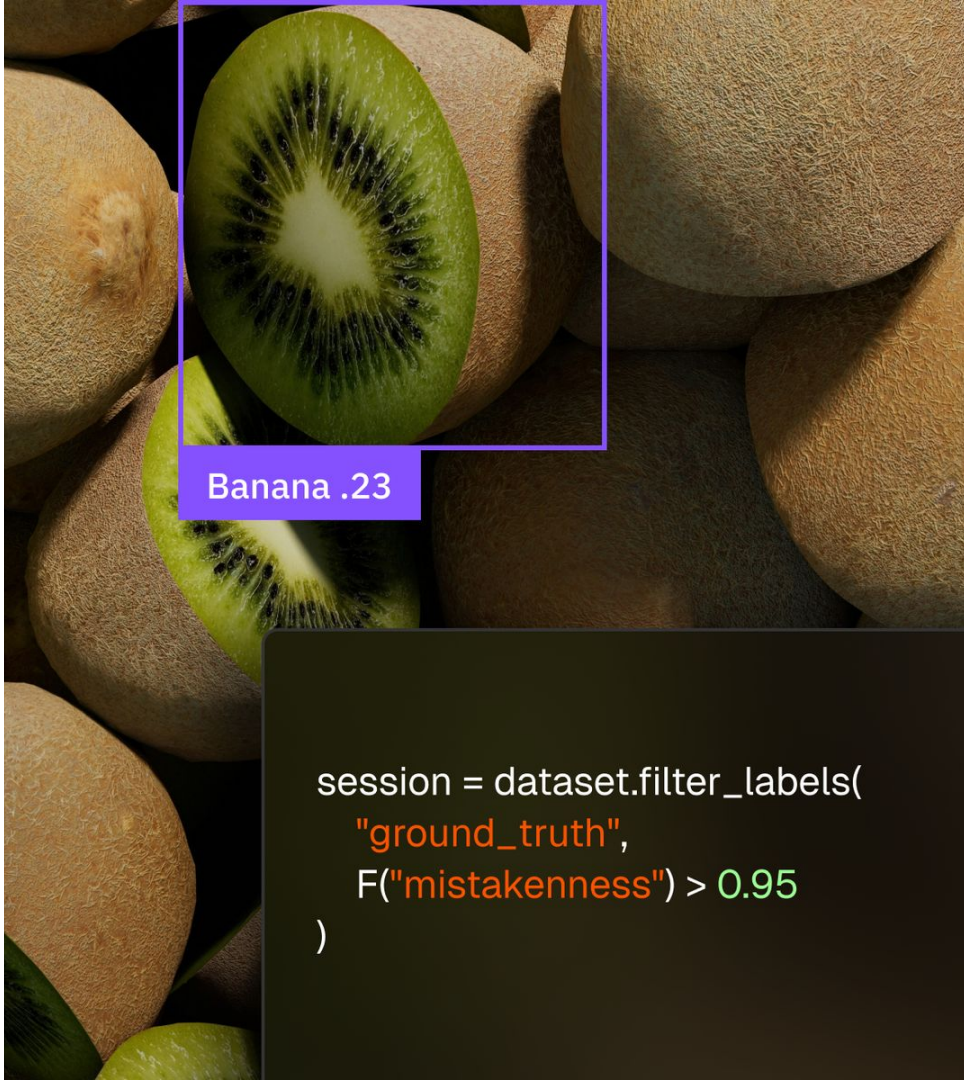
Auto Labeling

ANNOTATION

Annotation bottlenecks stall AI development

Manual annotation is slow, inconsistent, and expensive.

- High cost with little scalability
- Painfully slow for large datasets
- Quality varies across annotators and vendors
- Delays model iteration and validation

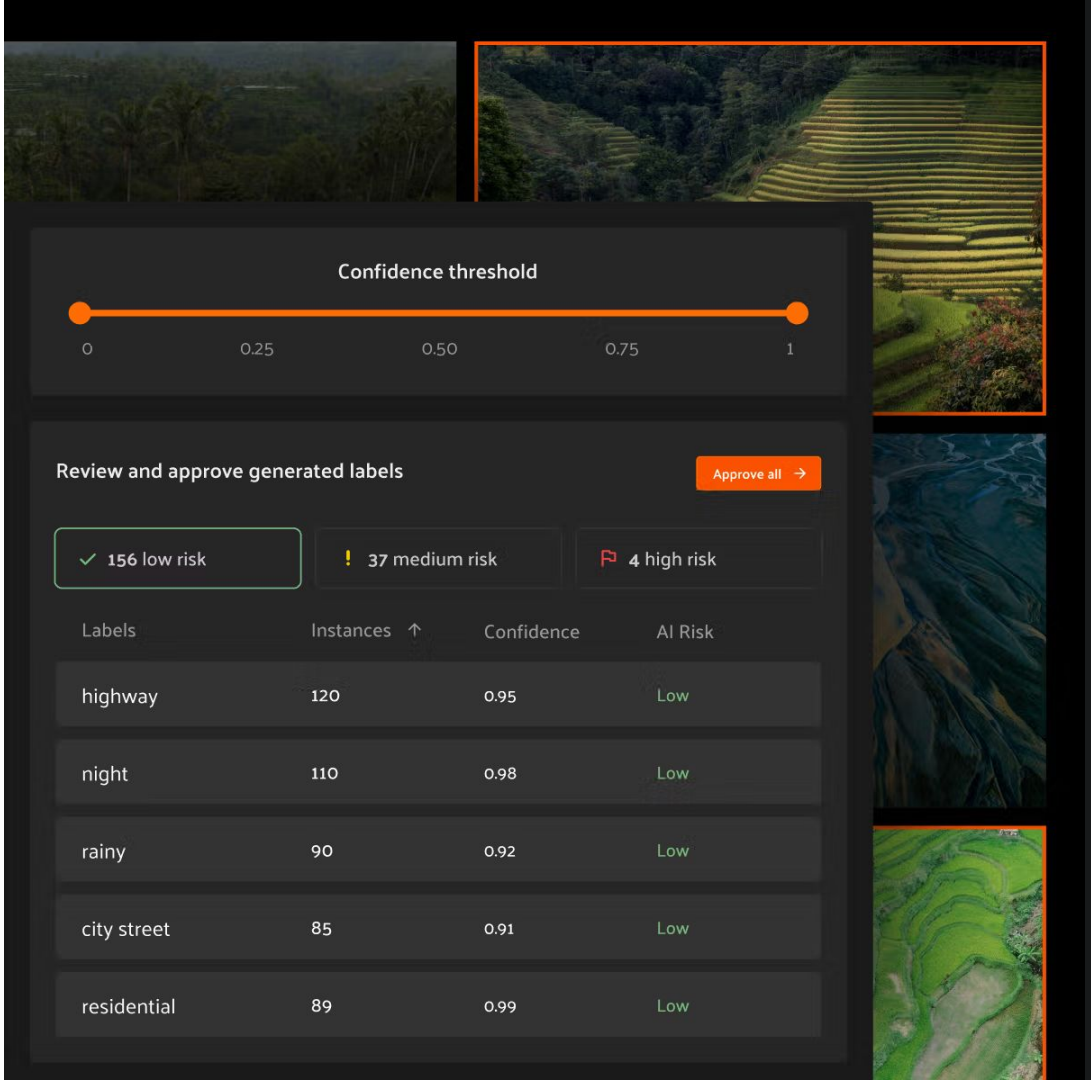


ANNOTATION

Verified Auto Labeling

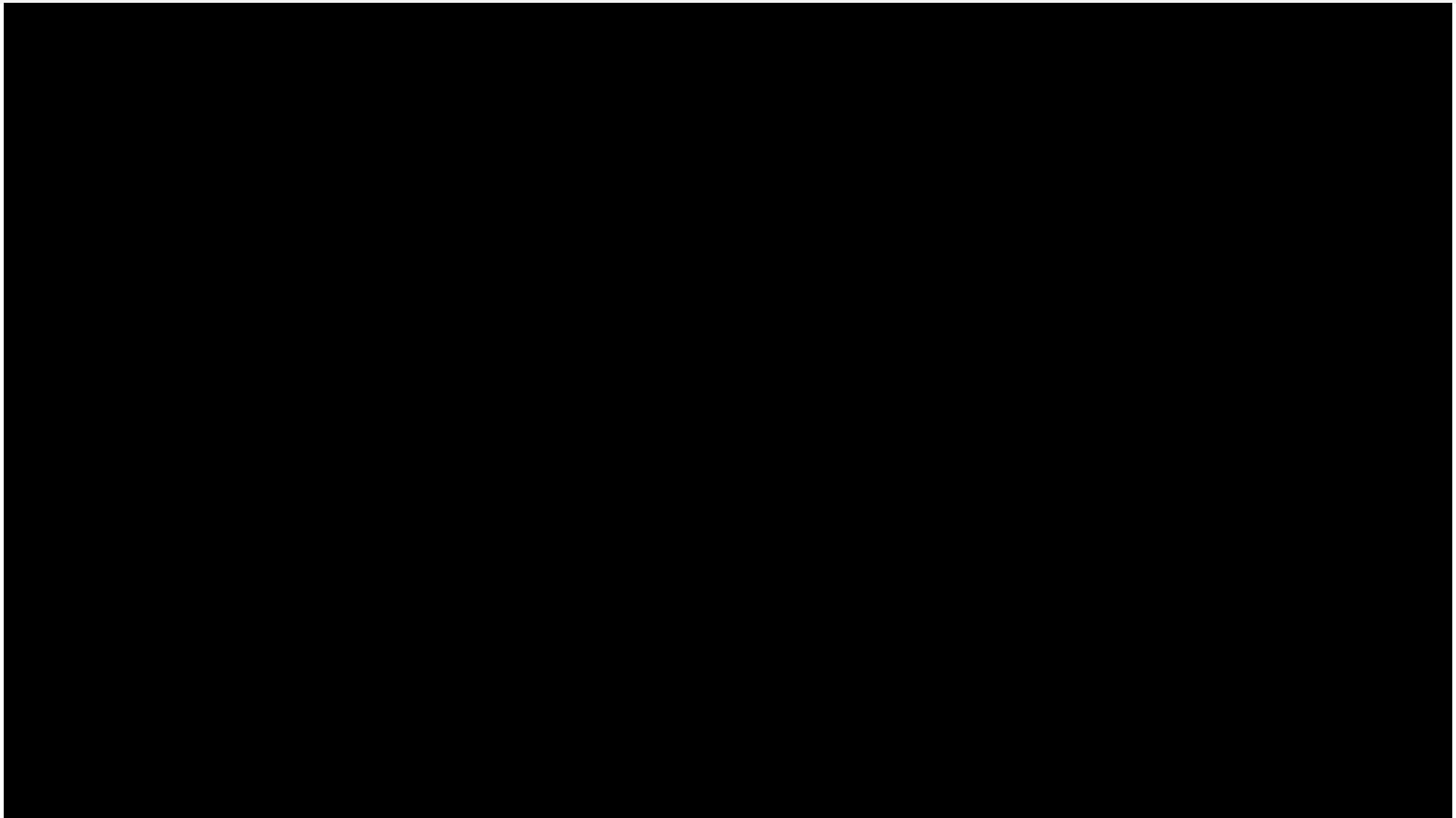
VAL uses foundation models to automatically generate labels — and adds confidence scoring to prioritize the ones that require human review.

- **Reduce QA and annotation costs** while maintaining near-human accuracy
- **Bring your own** foundational models
- Ensure **data sovereignty** by keeping more of your proprietary training data in-house



Demo →

Verified Auto Labeling with FiftyOne



VOXEL51

Model Evaluation

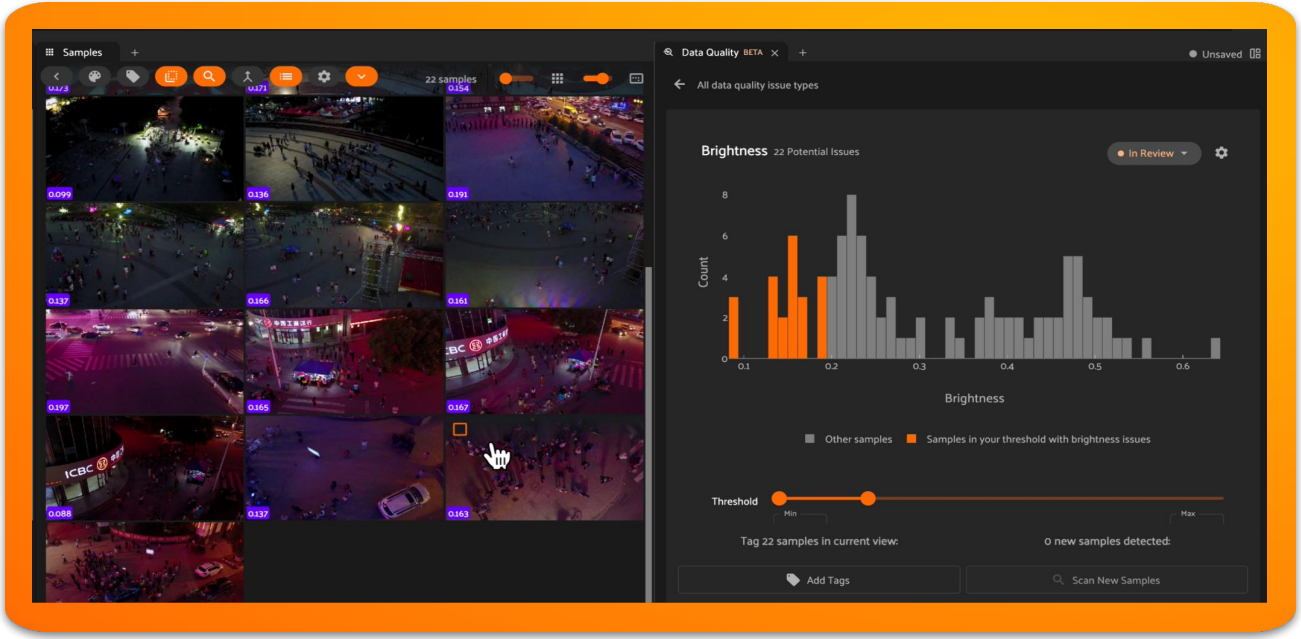
My model is broken! A mystery…

My model is broken! A mystery…

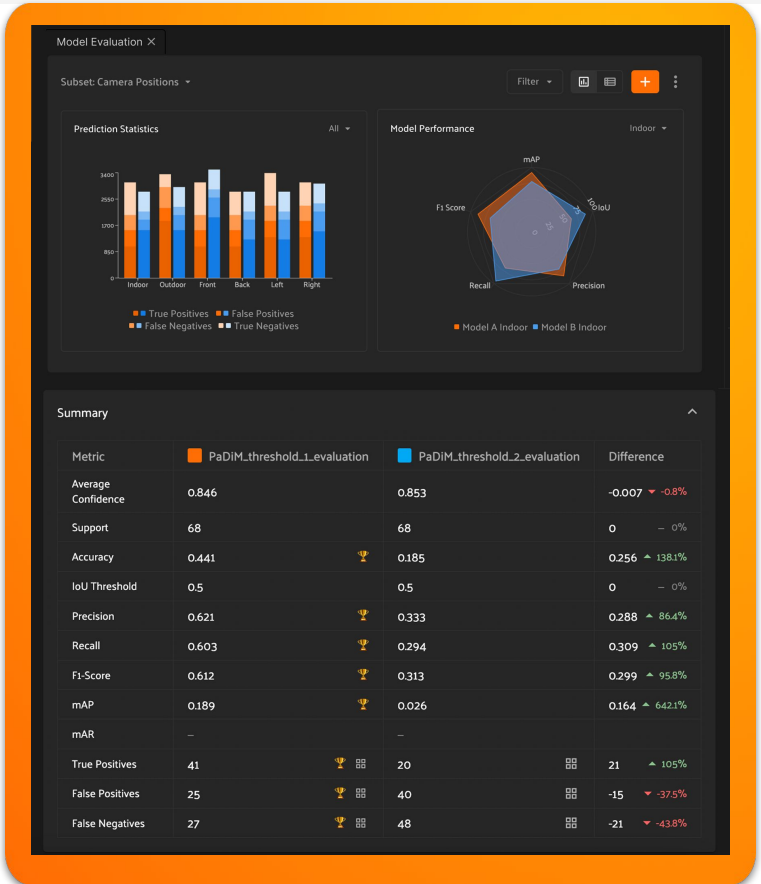
Data Quality

Scan for and analyze several different types of data quality issues

- ✓ Brightness
- ✓ Blurriness
- ✓ Aspect Ratio
- ✓ Entropy
- ✓ Near Duplicates
- ✓ Exact Duplicates



Model Evaluation and Scenario Analysis



- Evaluate predictions visually and numerically
- Slice performance by conditions
- Compare multiple models or thresholds side-by-side
- Detect hidden failure modes and edge cases at a glance

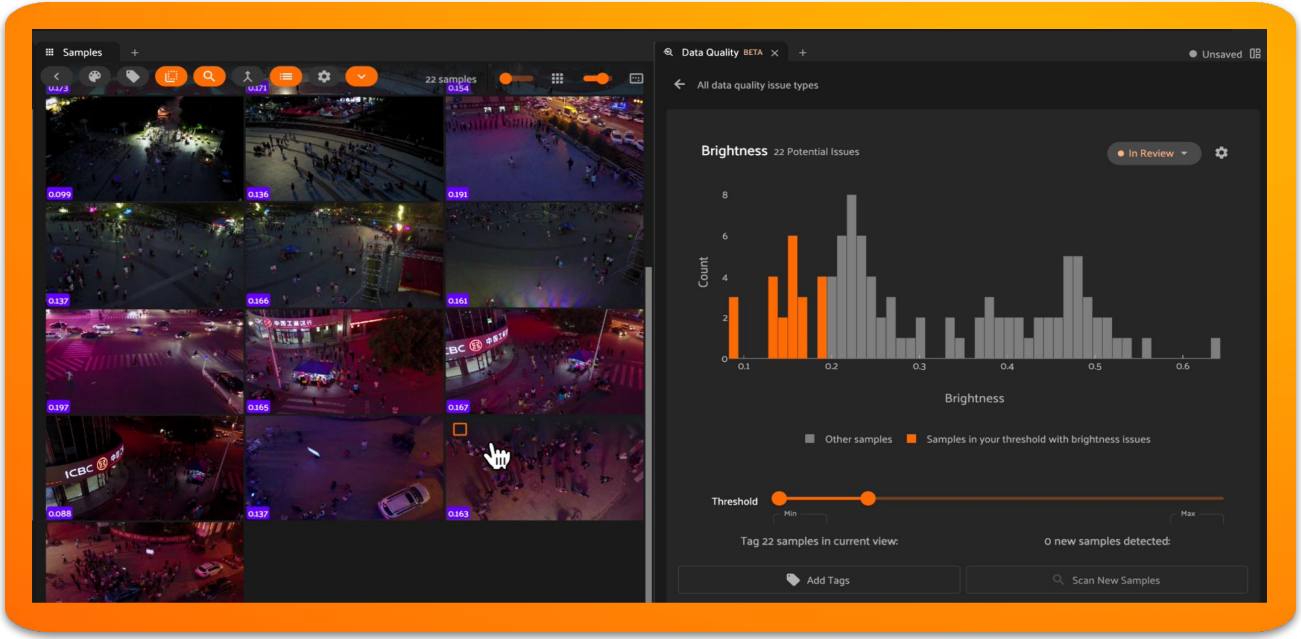
VOXEL51

Iteration

Data Quality

Scan for and analyze several different types of data quality issues

- ✓ Brightness
- ✓ Blurriness
- ✓ Aspect Ratio
- ✓ Entropy
- ✓ Near Duplicates
- ✓ Exact Duplicates

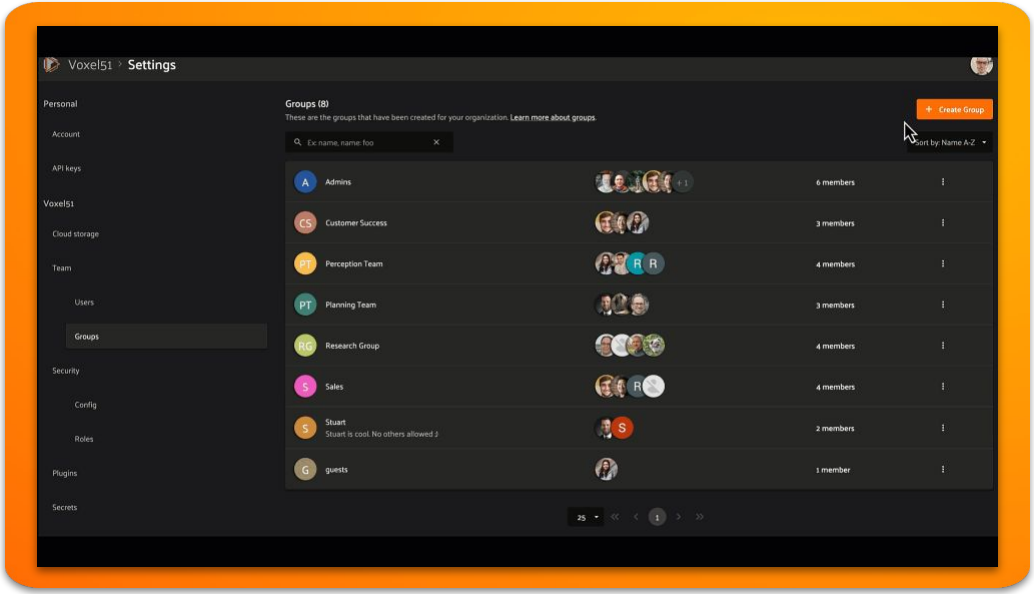


VOXEL51

Collaborate
Securely on Datasets

Collaborate With Your Team

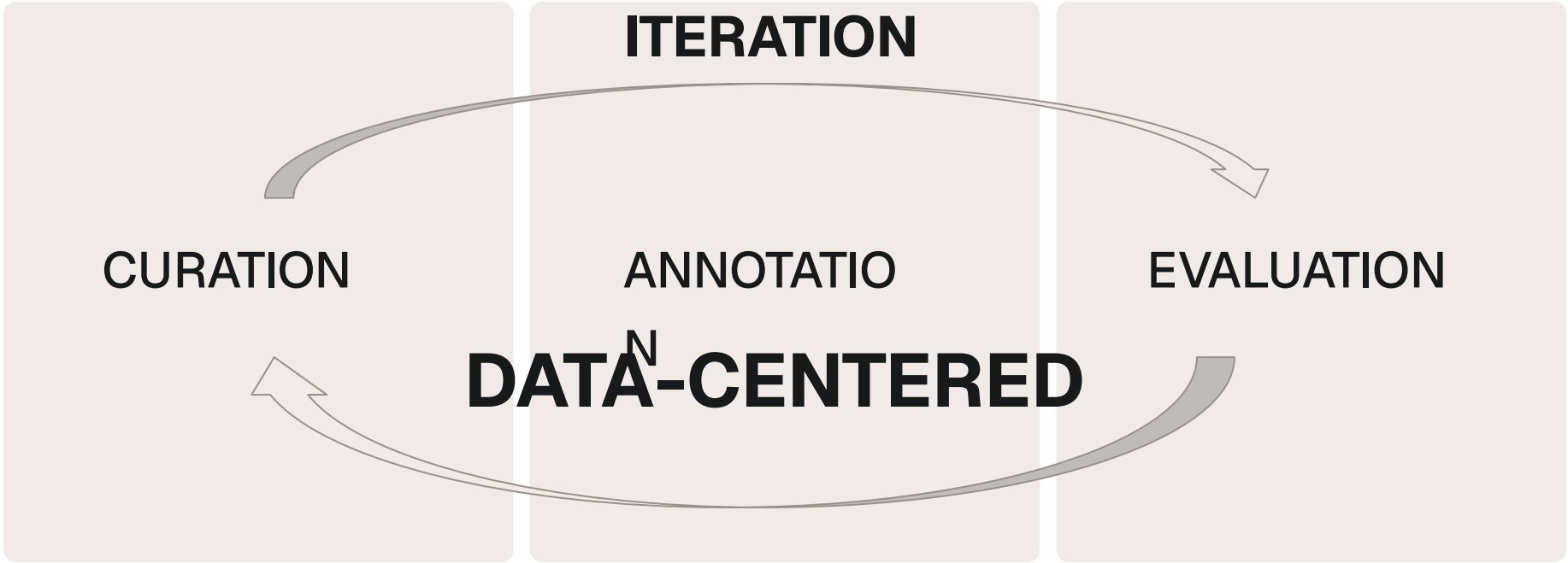
- Many free and open source tools are (generally) single-user
- Enterprise platforms generally multi-user by design
  - Hooks into your org's auth provider (AD / LDAP / SAML)
- Role-based access control for users and datasets
  - Limit datasets to particular users/groups
  - Users/group actions limited by their role



Dataset Versioning

- Clone and export datasets
  - DVC
- Create, share, and roll back dataset snapshots
- Compare diffs between snapshots

Building a Successful Pipeline



VOXEL51

Thank you!
Questions?

— @voxel51

voxel51.com