tvmcon

Monitoring ML on the Edge

Alessya Visnjic, CEO @ WhyLabs

tvmcon

Monitoring ML ~~on~~ at the Edge

Alessya Visnjic, CEO @ WhyLabs

Your product is out in the wild!

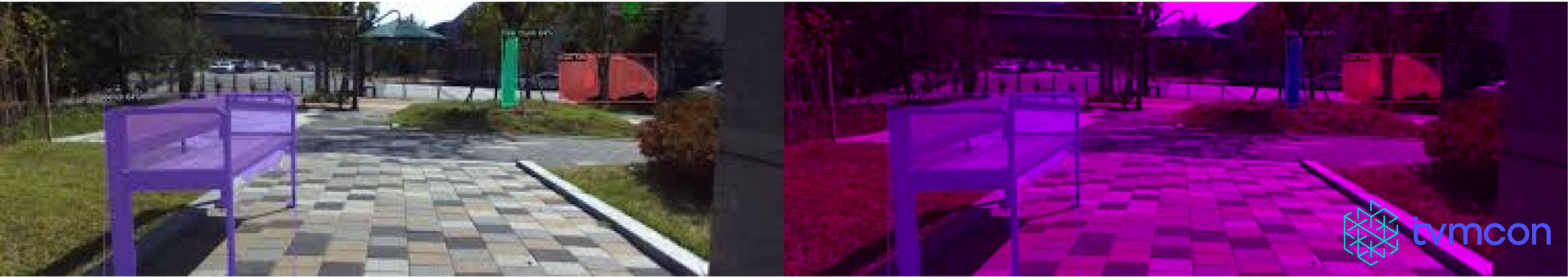
Now, the fun part:

- Ensure that it works the way you intended
- Prevent it from RapidUncontrolledThermalEvents
- Figure out how to improve it

Suddenly: performance degraded by 10%

Time to notice the problem: 3 weeks
Time to root-cause the problem: 1 week
Time to rollback firmware across the fleet: 1 day

~1 month to fix a buggy code change:
swapped the colors between channels

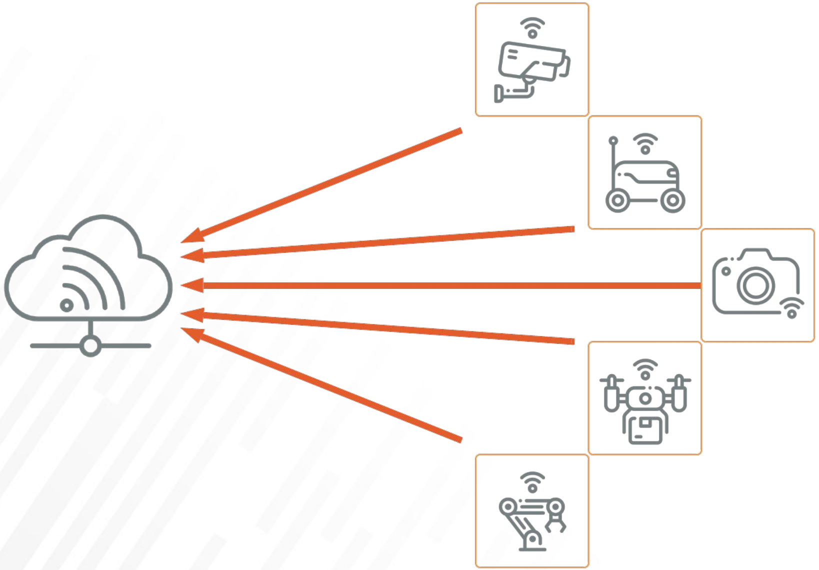


Suddenly: # of objects detected in a scene down by 25%

Next release is blocked for an unknown reason
Time to root-cause the problem: 1 week
Time to update configuration on the fleet: 1 day

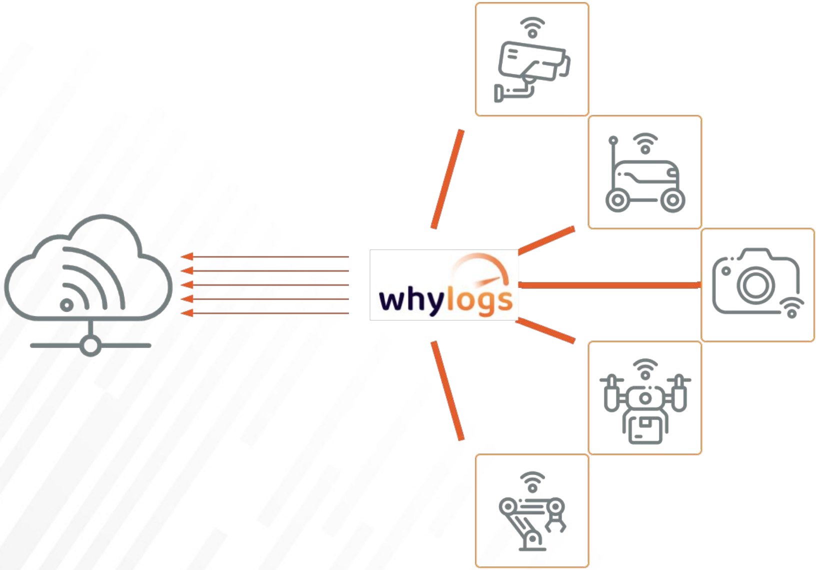
Dozens of teams blocked for weeks by a bug:
default depth encoding changed from 1 meter to 10 meters

Monitoring ML at the edge is very hard



- ❖ A lot of data is just ephemeral
- ❖ Massive volumes of sensor data (>10 GB/min)
- ❖ Limited and intermittent connectivity
- ❖ Expensive data transfer / expensive post-processing
- ❖ Sensor problems vs. model problems vs. system problems
- ❖ Diversity in the fleet
- ❖ Time-to-resolution can be weeks or months

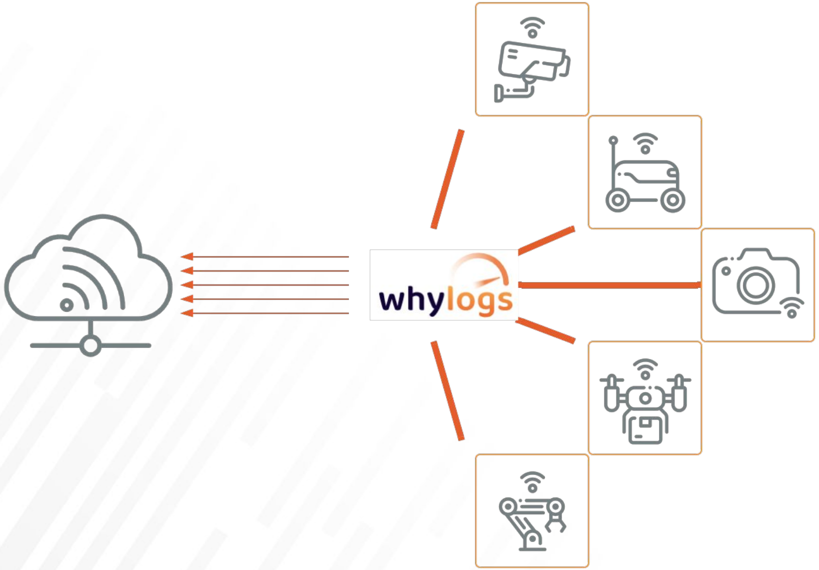
Monitoring ML at the edge, with whylogs



- Sketch key metrics locally, on the device
- Introduce meaningful batches (1min, 5min, 1hr)
- Sync tiny sketches when connectivity is available
- Merge statistics over time and over devices
- Collect sketches on signal data, metadata, model outputs, custom metrics
- Monitor metrics as soon as it's synced
- Pinpoint the problem to a batch to debug faster

github.com/whylabs/whylogs

whylogs: open standard for data logging



- ❖ Lightweight
- ❖ Configurable
- ❖ Portable
- ❖ Mergeable

Monitoring Data Health:

- Metadata from messages
- Model input statistics:
  - Volume
  - Structured data statistics & distributions
  - Missing data / incomplete data
  - Statistical features from non-structured data

Monitoring Model health:

- Predictions
  - Labels
  - Tracks
  - Bounding boxes / polygons

Monitoring System health:

- Device / sensor metadata
  - Volume
  - Settings / configurations
- Runtime metadata
  - Latency
  - Throughput
  - Temperature
  - Voltage

Thank you!

Help build the open standard

for data logging:

github.com/whylabs/whylogs

alessya@whylabs.ai

@zalessya

Monitor with WhyLabs

whylabs.ai/free

Join.slack.whylabs.ai

Thank you!

Help build the open standard for data logging:

github.com/whylabs/whylogs

alessya@whylabs.ai

@zalessya

Monitor with WhyLabs

whylabs.ai/free

Join.slack.whylabs.ai

bit.ly/whylogs:

Logging for the ML stack

Help build the open standard for data logging:

github.com/whylabs/whylogs

join.slack.whylabs.ai

Thank you!

alessya@whylabs.ai
@zalessya

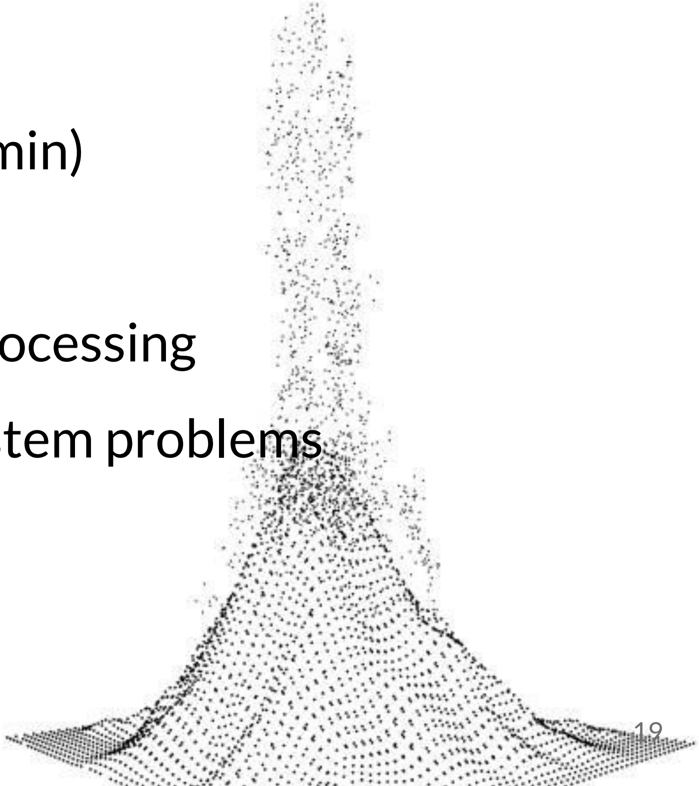
Monitoring an ML applications means knowing answers to there questions w/ doing manual work:

- ❖ Is the system running as expected?
- ❖ Is the model receiving healthy data?
- ❖ Do my predictions look healthy?
- ❖ Is my model degrading over time?
- ❖ Why is performance dropping suddenly?
- ❖ What can I change to improve the model performance?

What if you could monitor for changes?

- Is the model seeing healthy data?
- What do the predictions look like?
- How is the model performing?
- How is the current data different from training data?
- How was yesterday’s data different from today?
- How was yesterday’s data different from last week’s data?

- Massive volumes of telemetry data (10GB/min)
- Limited and intermittent connectivity
- Expensive data transfer / expensive post-processing
- Sensor problems vs. device problems vs. system problems
- Time-to-resolution is weeks or months

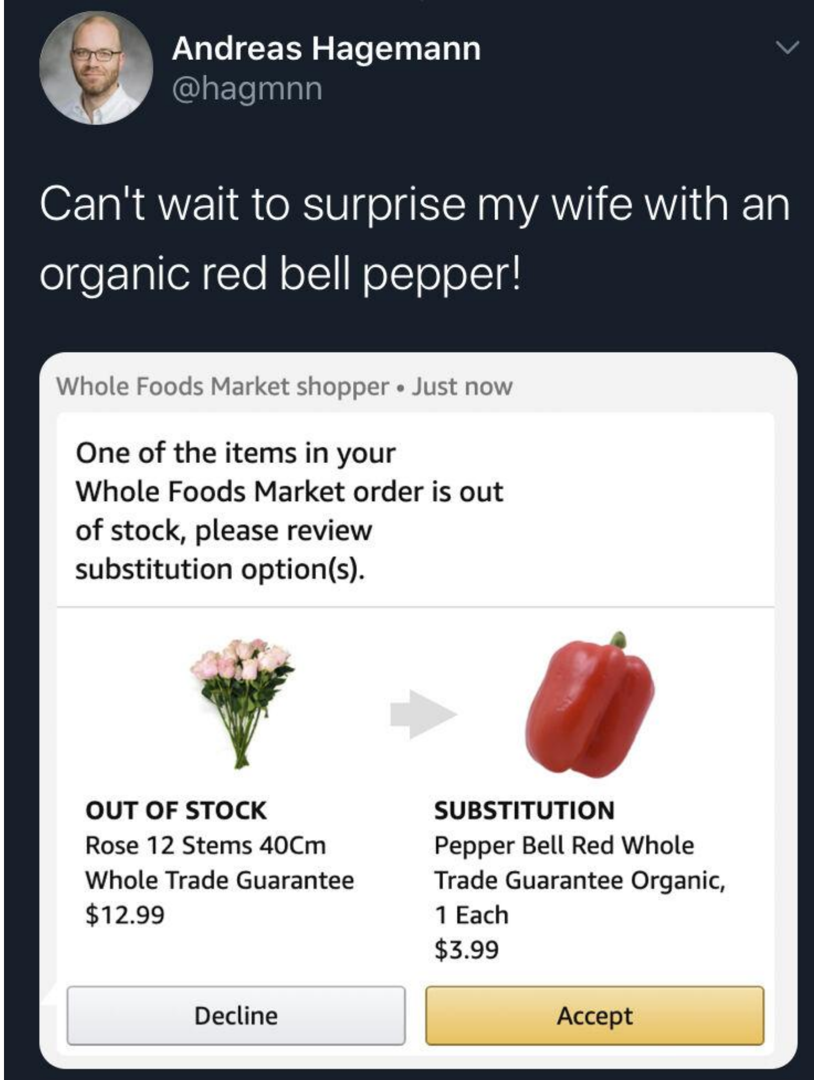


Agenda

- ML Stack: what is missing?
- How to design data logging
- whylogs: open standard for data logging
- Use cases
- Q&A

If you are not keeping an eye on the model in production…

… your customers are!



Andreas Hagemann
@hagmnn

Can't wait to surprise my wife with an organic red bell pepper!

Whole Foods Market shopper • Just now

One of the items in your Whole Foods Market order is out of stock, please review substitution option(s).

OUT OF STOCK
Rose 12 Stems 40Cm Whole Trade Guarantee
$12.99

SUBSTITUTION
Pepper Bell Red Whole Trade Guarantee Organic, 1 Each
$3.99

Issues **encountered in production** (small sample)…

- Experiment/production environment mismatch
- Wrong model version deployed
- Underprovisioned hardware
- Inappropriate hardware
- Latency/SLA issues
- Data permissions misconfigured
- Untracked changes broke prod
- Traffic sent to the wrong model
- Computational instability
- Customers gaming the model (adversarial attacks)
- PII data exposed
- Expected accuracy doesn’t materialize

- Pre-processing mismatch in experiments vs. production
- Retrained on faulty data
- Accuracy improves on one segment, regresses in others
- Outliers predicted incorrectly
- Bias identified
- Correlation with protected features
- Overfitting on training/test
- Surge in missing values
- Surge in duplicates

- Poor performance on outliers
- Data quality issues affect accuracy
- Production data doesn’t match test/training
- Accuracy is decaying over time
- Data drift in inputs
- Concept drift in outputs
- Extreme predictions for out of distribution data
- Model not generalizing on new data / new segments
- Major consumer behavior shift

…or it simply doesn’t work, and nobody know why…

Issues encountered in production (small sample)...

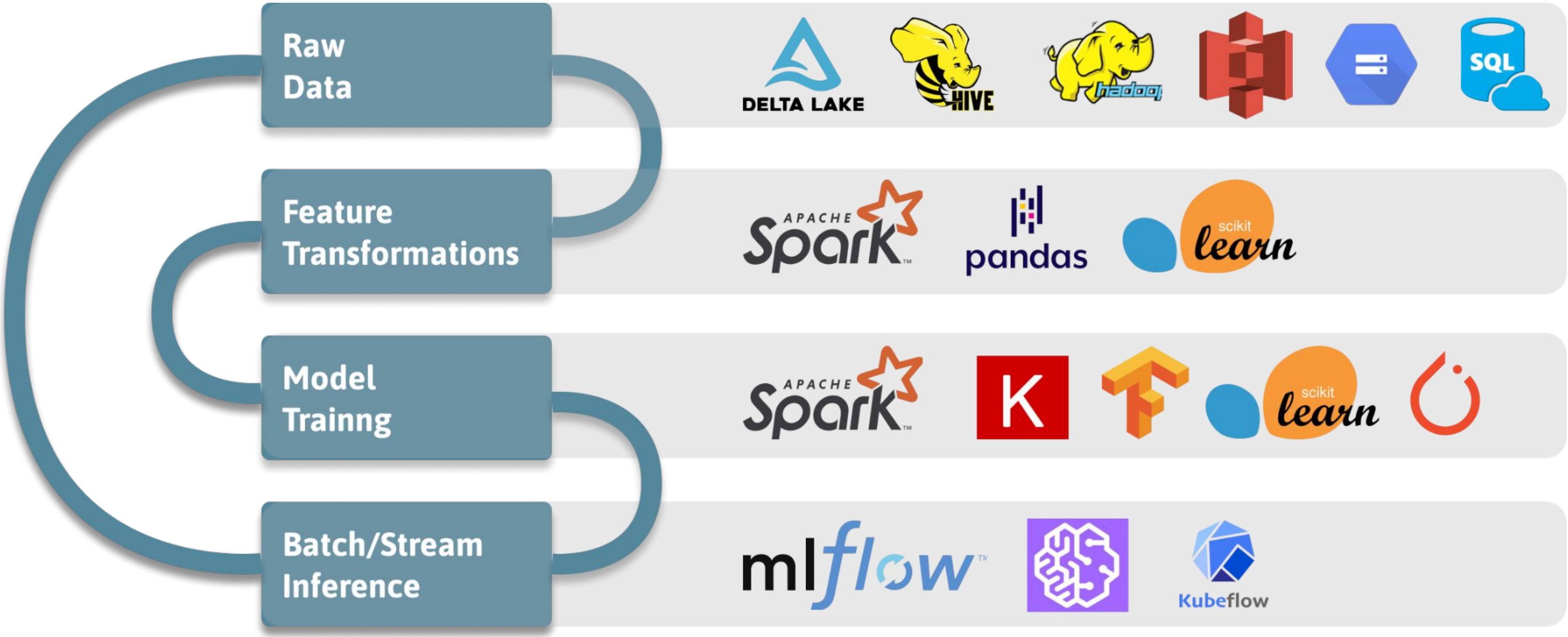
- Experiment/production environment mismatch
- Wrong model version deployed
- Underprovisioned hardware
- Inappropriate hardware
- Latency/SLA issues
- Data permissions misconfigured
- Untracked changes broke prod
- Traffic sent to the wrong model
- Computational instability
- Customers gaming the model (adversarial attacks)
- PII data exposed
- Expected accuracy doesn’t materialize

- Pre-processing mismatch in experiments vs. production
- Retrained on faulty data
- Accuracy improves on one segment, regresses in others
- Outliers predicted incorrectly
- Bias identified
- Correlation with protected features
- Overfitting on training/test
- Surge in missing values
- Surge in duplicates

- Poor performance on outliers
- Data quality issues affect accuracy
- Production data doesn’t match test/training
- Accuracy is decaying over time
- Data drift in inputs
- Concept drift in outputs
- Extreme predictions for out of distribution data
- Model not generalizing on new data / new segments
- Major consumer behavior shift

Data problems!

ML stack: moving massive volumes of data



Raw Data

Feature Transformations

Model Trainng

Batch/Stream Inference

How do you

Test

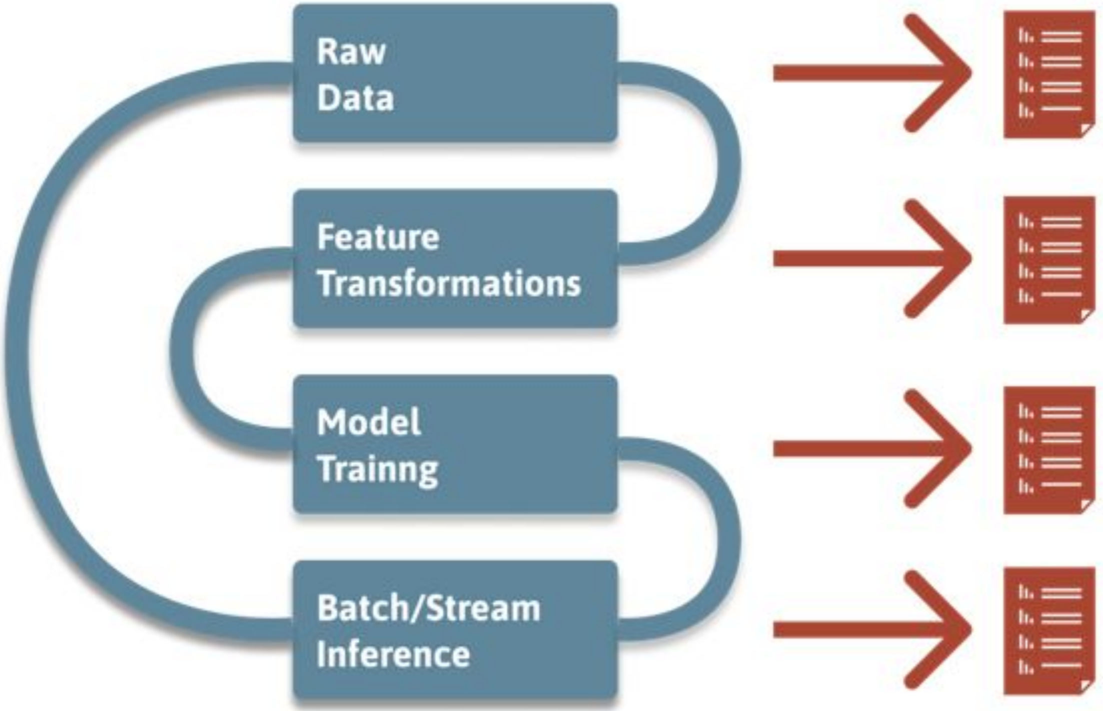
Monitor

Debug

Document

data?

ML Stack is missing **data & model logging**



Log metadata & statistical properties of data

A good data log should capture:

- Metadata
- Counts
- Statistics
- Distributions

Stratified sam

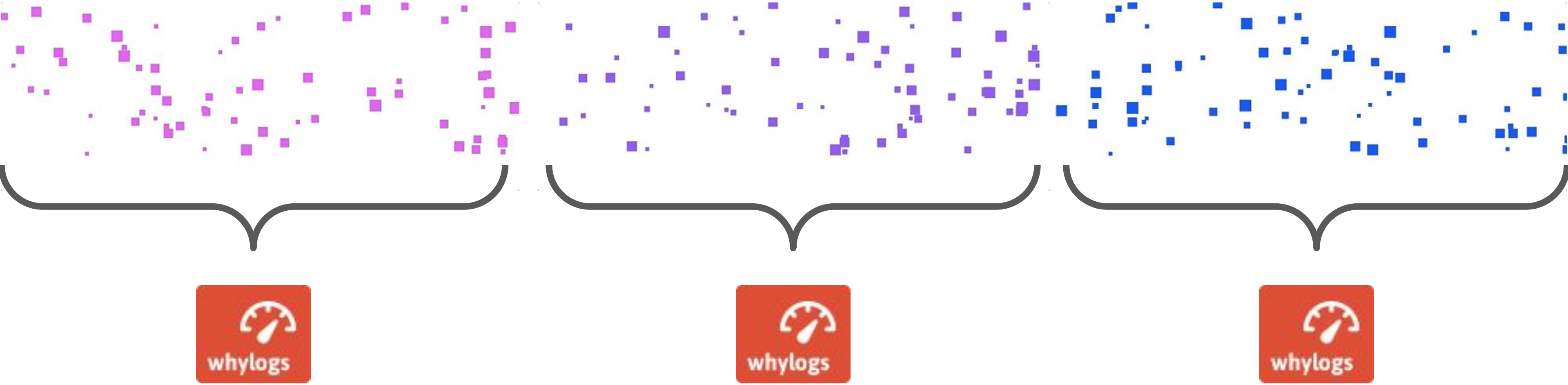
Key properties of a data log:

- Lightweight
- Portable
- Mergeable
- Configurable
- Close to code

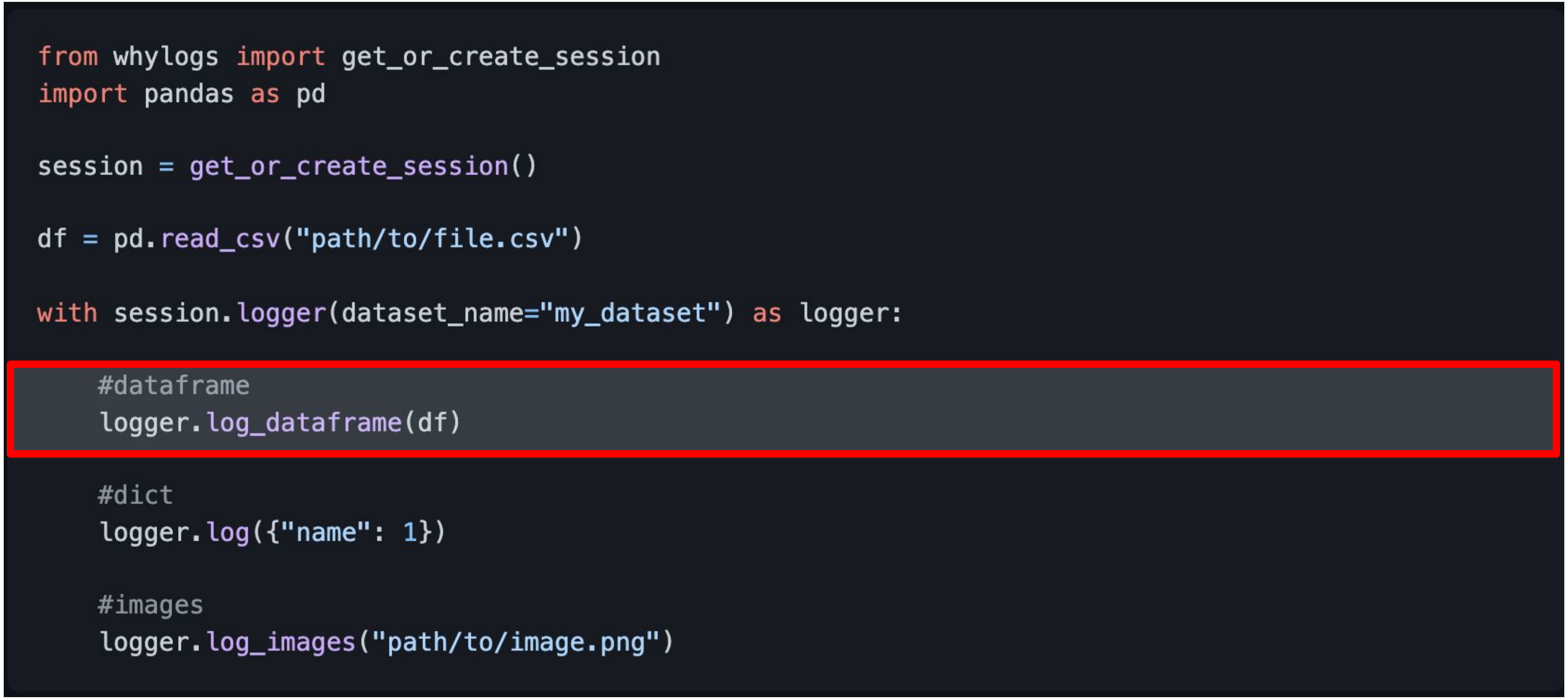
Before whylogs: **data is ephemeral**

whylogs:

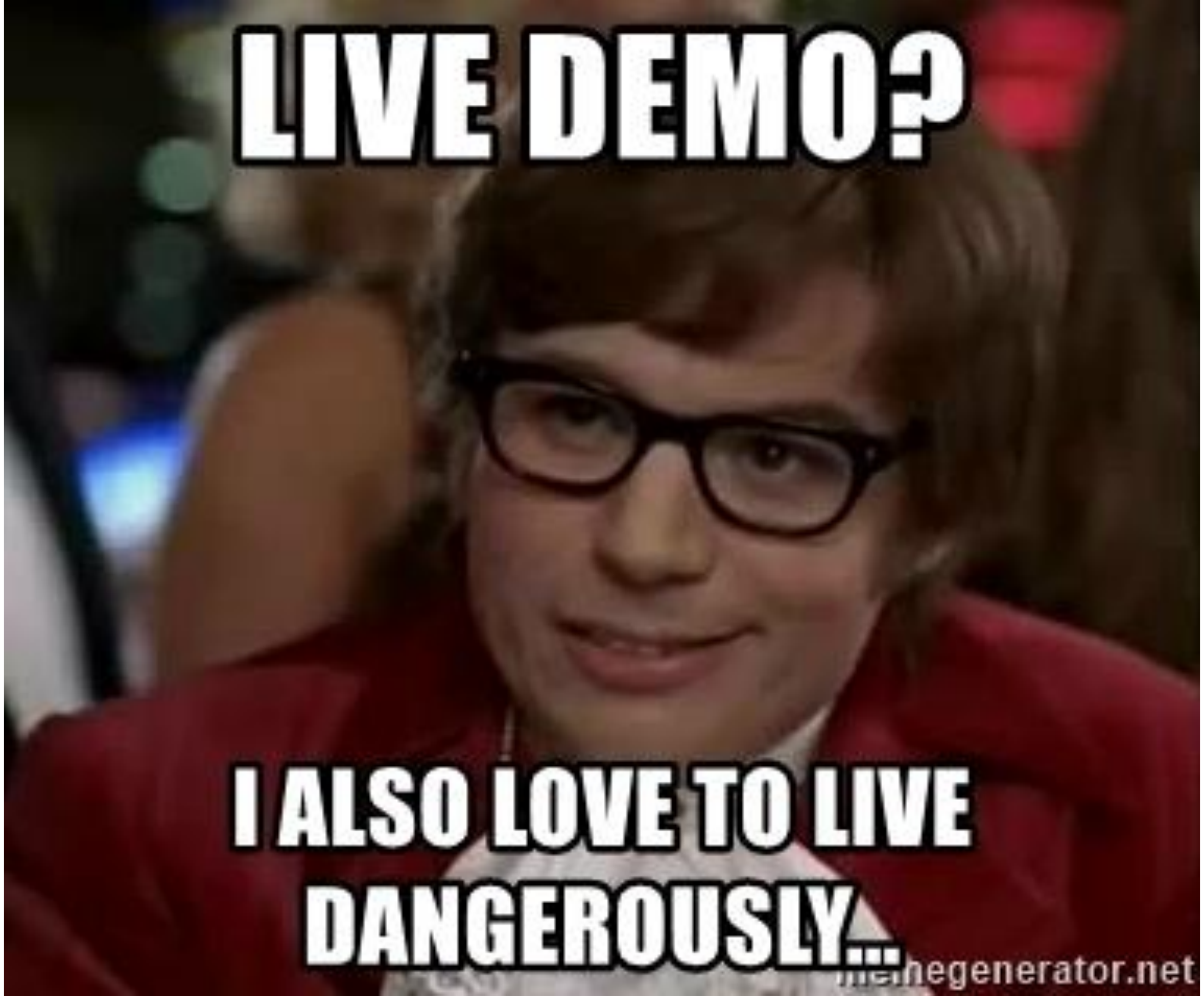
a standard format for representing a snapshot of data



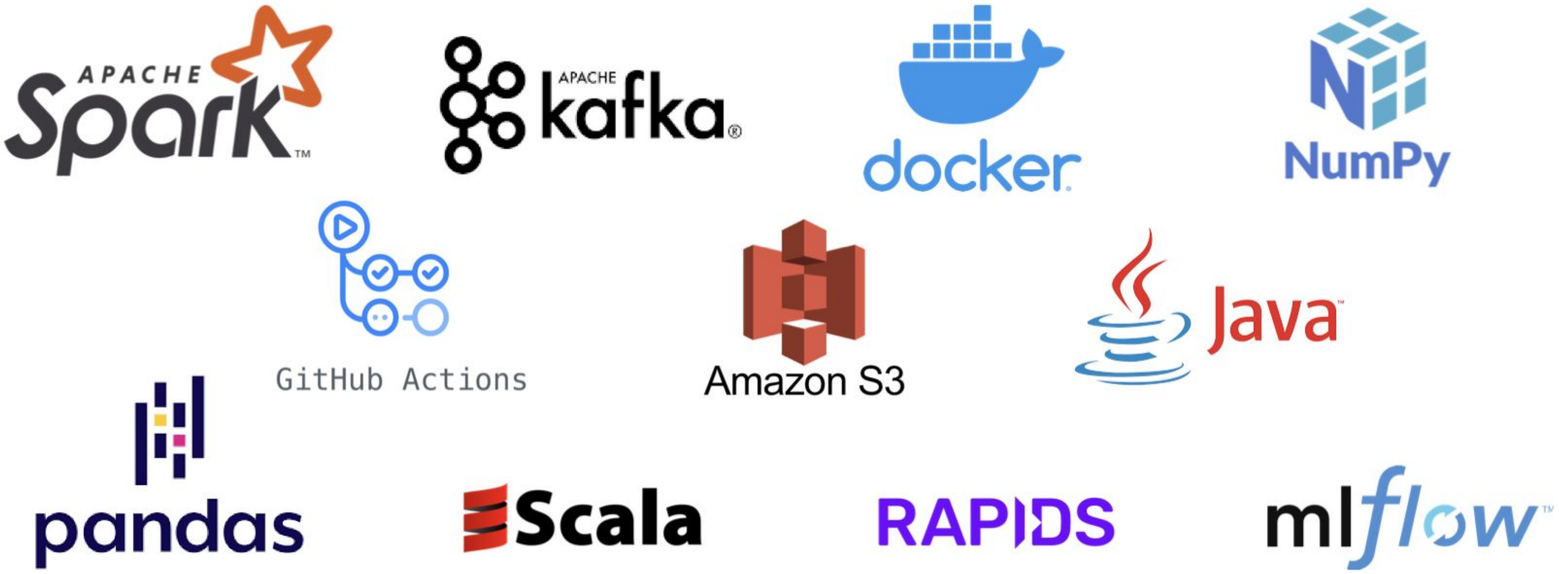
Profile and log with **1 line of code**



```python
from whylogs import get_or_create_session
import pandas as pd

session = get_or_create_session()

df = pd.read_csv("path/to/file.csv")

with session.logger(dataset_name="my_dataset") as logger:

    #dataframe
    logger.log_dataframe(df)

    #dict
    logger.log({"name": 1})

    #images
    logger.log_images("path/to/image.png")
```


LIVE DEMO?
I ALSO LOVE TO LIVE DANGEROUSLY...
memegenerator.net

Works with you favorite tools out of the box



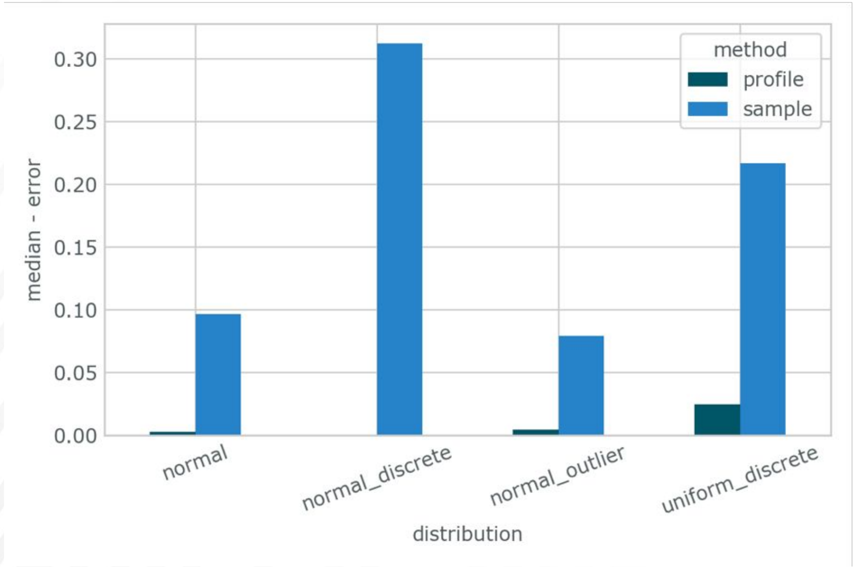
Run **data logging** without overhead

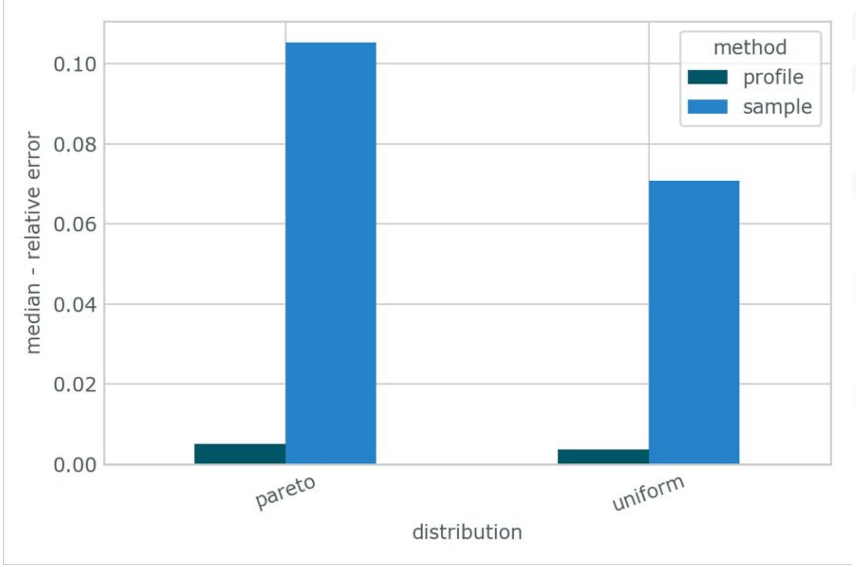
Using streaming algorithms to capture data statistics, whylogs ensures a constant memory footprint, scales with the number of features in the dataframe, and outputs lightweight log files (json, protobuf, etc).

Dataset	Size	No. of Entries	No. of Features	Est. Memory Consumption	Output Size (uncompressed)
Lending Club	1.6GB	2.2M	151	**14MB**	7.4MB
NYC Tickets	1.9GB	10.8M	43	**14MB**	2.3MB
Pain pills in the USA	75GB	178M	42	**15MB**	2MB

Capture **accurate data distributions**

Whylogs profiles 100% of the data to accurately capture distributions. Capturing distributions from sampled data is significantly less accurate. This chart presents median errors for distributions estimated with whylogs vs. from sampled data.





You can

Test

Monitor

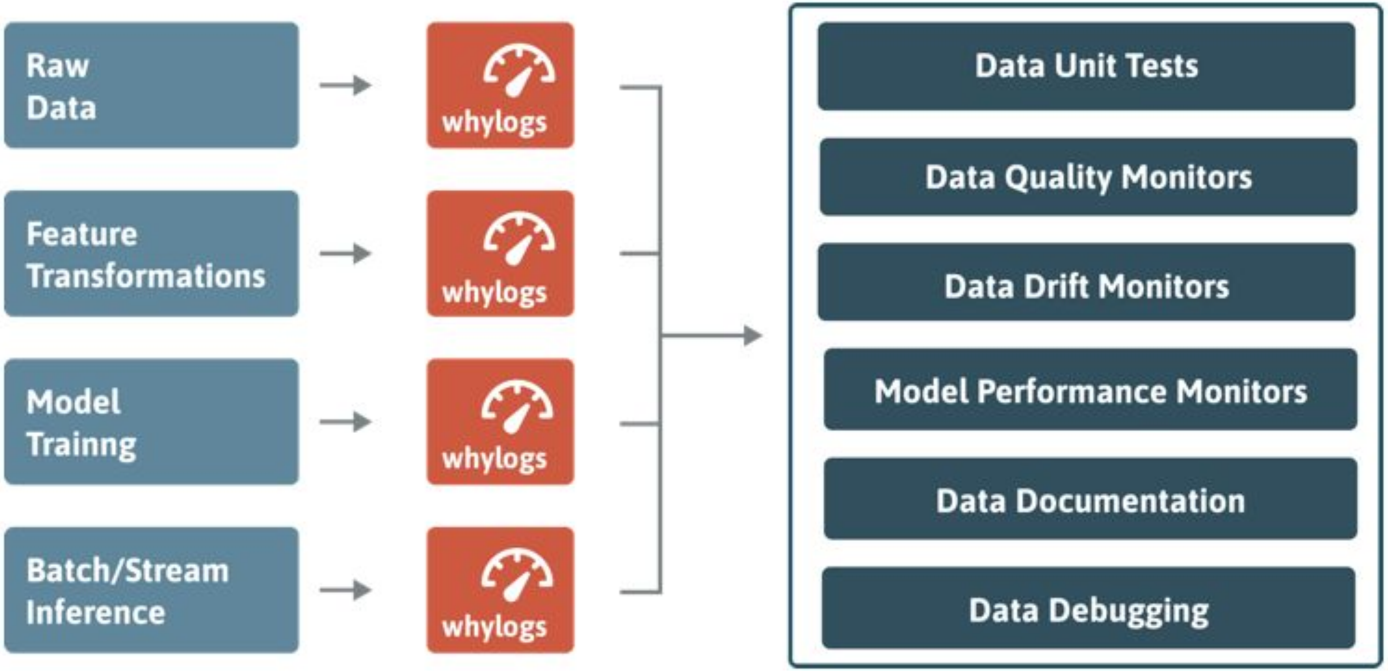
Debug

Document

data with whylogs!!

Logging enables all key **MLOps** activities

Once data is logged systematically, whylogs outputs can be used to test, monitor, and debug data. Use whylogs at any point of the ML stack and through the lifecycle of the ML application.



Raw Data

Feature Transformations

Model Trainng

Batch/Stream Inference

whylogs

whylogs

whylogs

whylogs

Data Unit Tests

Data Quality Monitors

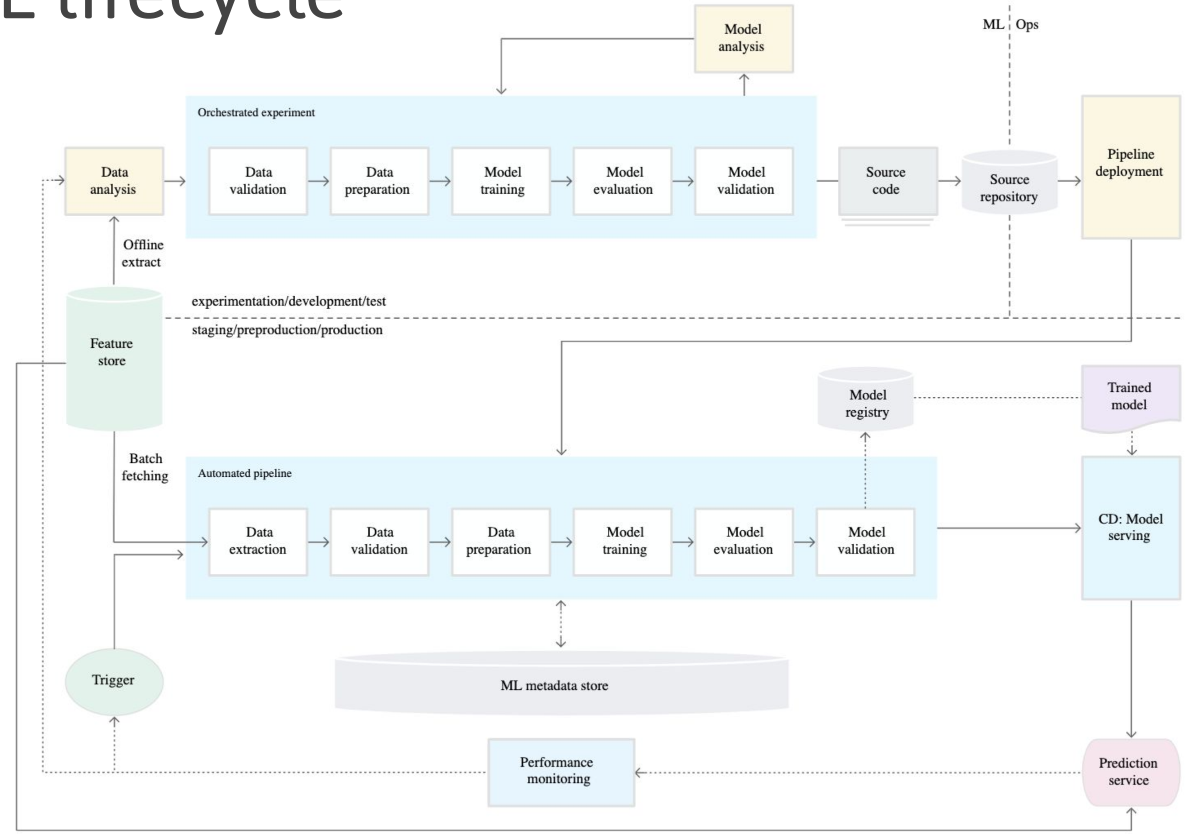
Data Drift Monitors

Model Performance Monitors

Data Documentation

Data Debugging

2021 ML lifecycle



Source: Google Cloud AI

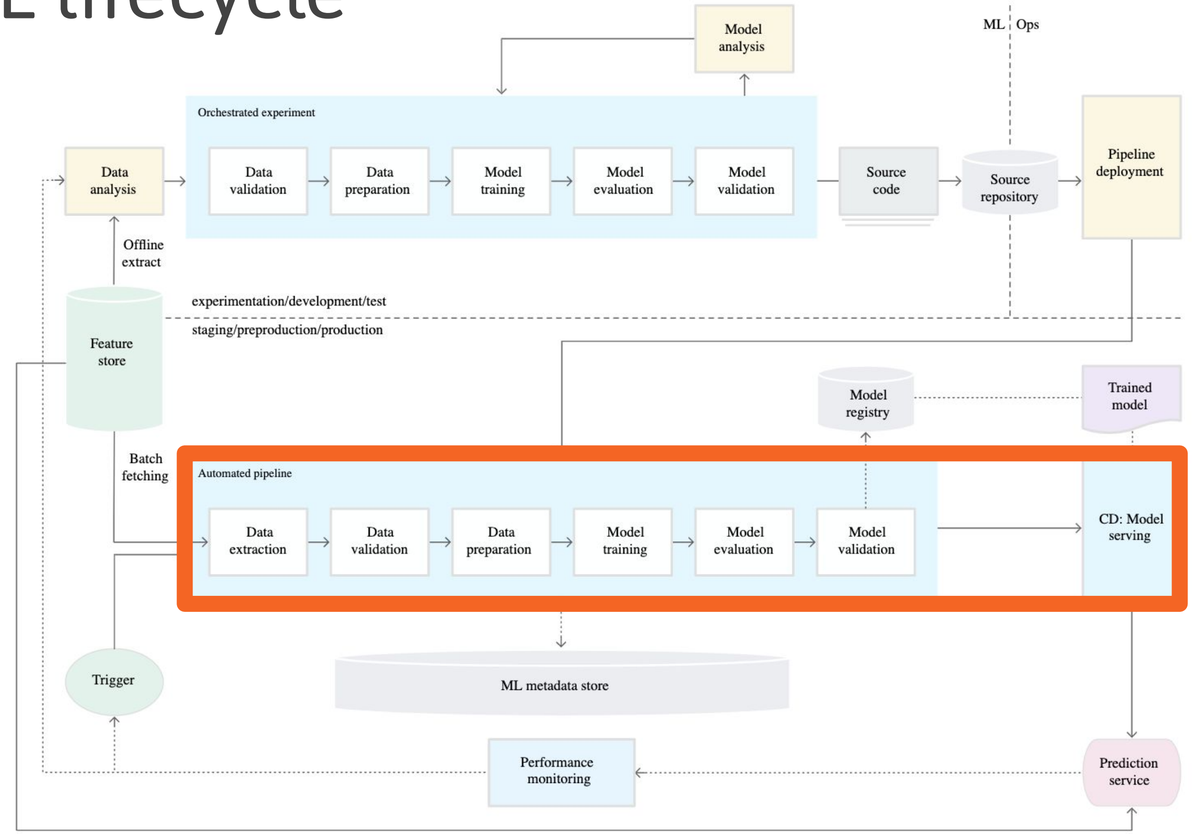
whylogs @ Data + AI Summit

May 27, 2021 05:00 PM: Re-imagine Data Monitoring with whylogs and Spark

May 26, 2021 03:50 PM PT: The Critical Missing Component in the Production ML Stack

May 27, 2021 11:00 AM PT: Semantic Image Logging Using Approximate Statistics & MLflow

2021 ML lifecycle



Source: Google Cloud AI

Logging in **DevOps** vs **MLOps**

DevOps:

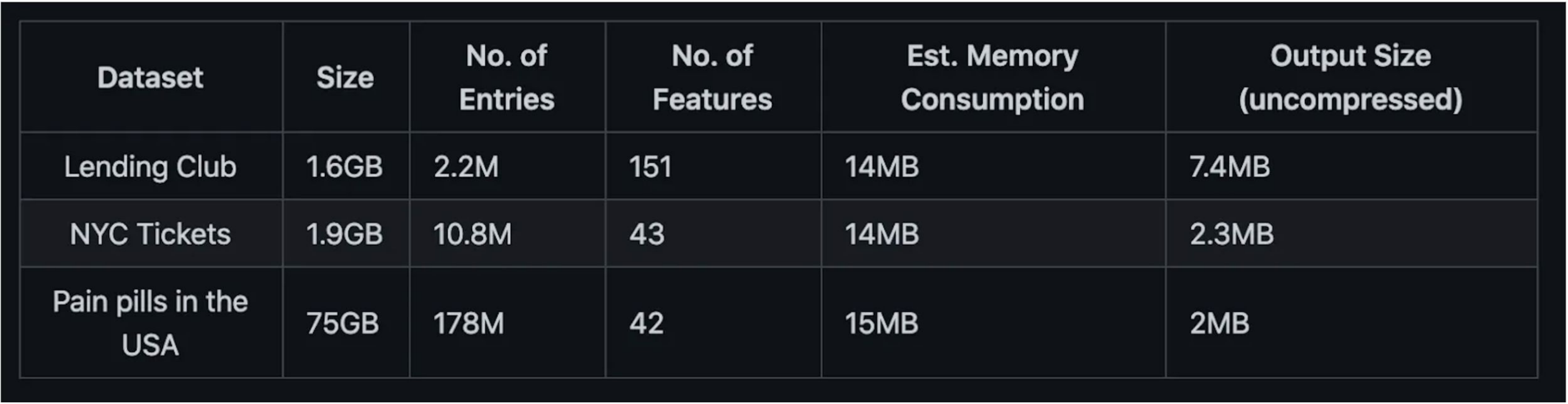
- System logs: Hostname, process type, log type, application, action, and TCP socket status
- Monitor:
  - Load
  - Response time
  - CPU
  - Memory
  - Storage
  - I/O

ML/AI:

- Data logs (sampling): an entire feature vector
- Data logs (profiling): statistical properties of each feature
- Monitor:
  - Missing values
  - Cardinality
  - Data Type
  - Summary statistics
  - Data distribution
  - Top K items distribution
  - Unstructured data stats

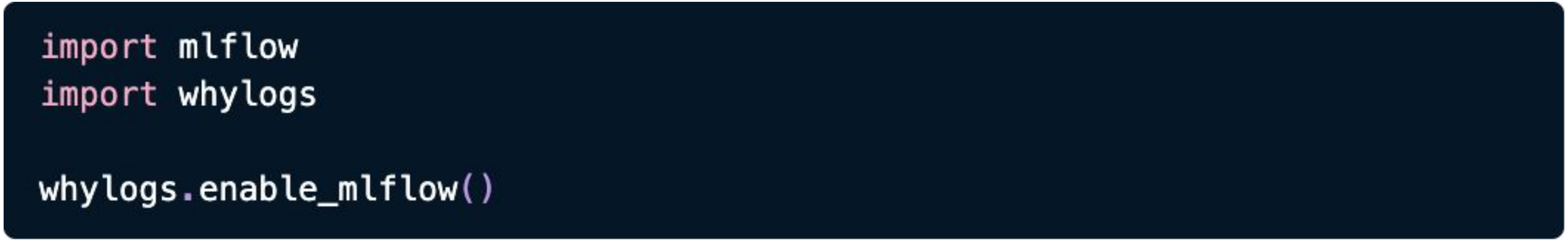
ML metrics requires massive data cardinality support. A model tracking 105 features, 4 metrics per feature, 420 metrics. Adding tracking for 3 segments, that’s 1680 metrics!

Data logging **must be scalable**



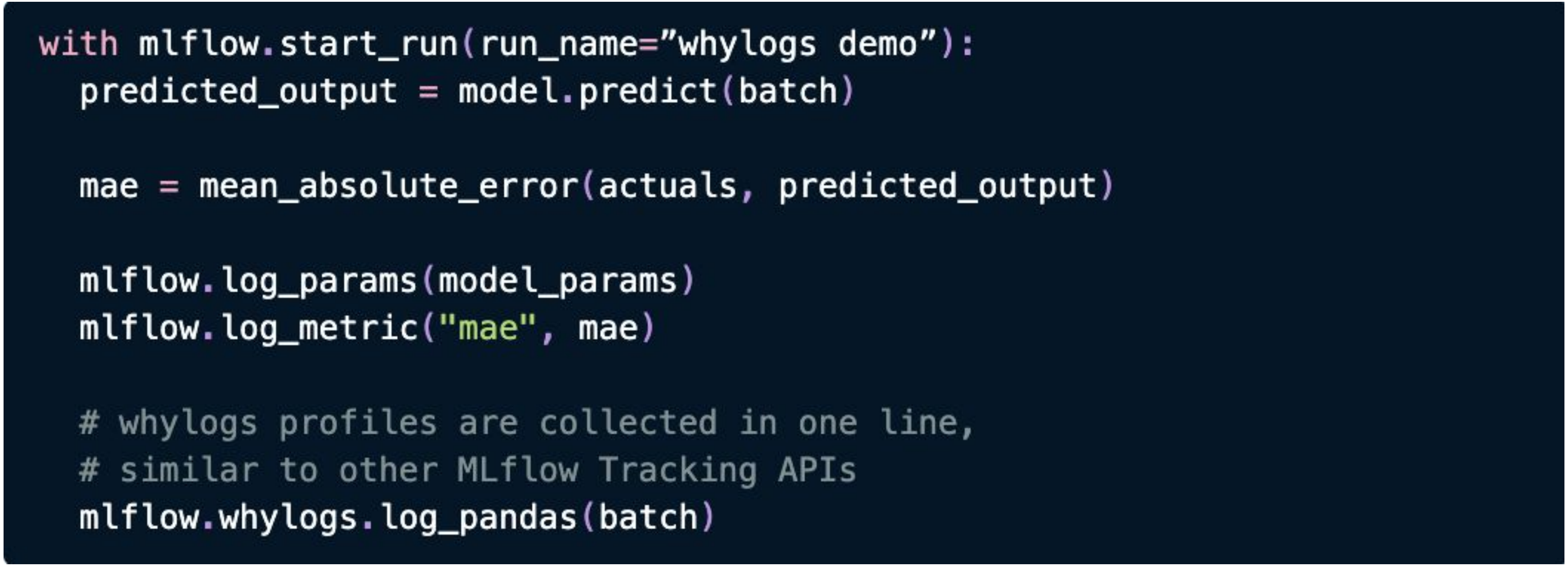
Dataset	Size	No. of Entries	No. of Features	Est. Memory Consumption	Output Size (uncompressed)
Lending Club	1.6GB	2.2M	151	14MB	7.4MB
NYC Tickets	1.9GB	10.8M	43	14MB	2.3MB
Pain pills in the USA	75GB	178M	42	15MB	2MB

Data logging with mlflow™



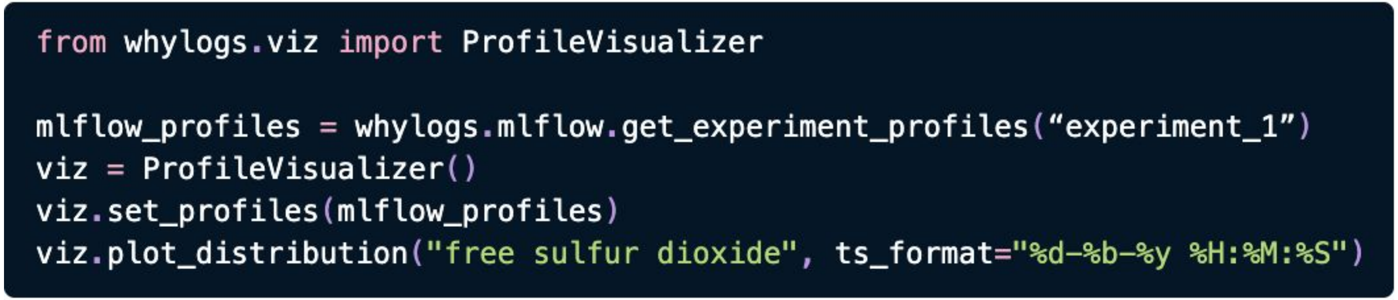
```python
import mlflow
import whylogs

whylogs.enable_mlflow()
```

Data logging with mlflow™

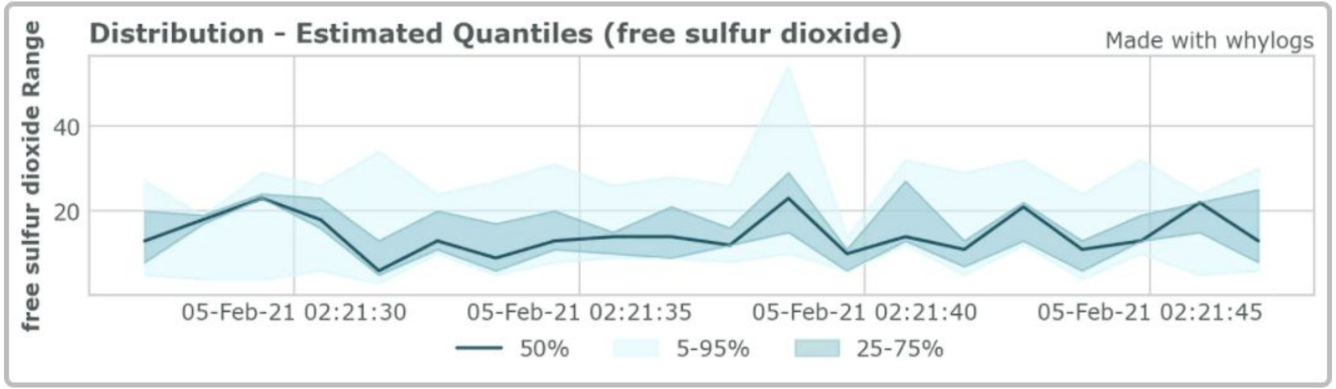


```
with mlflow.start_run(run_name=”whylogs demo”):
  predicted_output = model.predict(batch)

  mae = mean_absolute_error(actuals, predicted_output)

  mlflow.log_params(model_params)
  mlflow.log_metric("mae", mae)

  # whylogs profiles are collected in one line,
  # similar to other MLflow Tracking APIs
  mlflow.whylogs.log_pandas(batch)
```

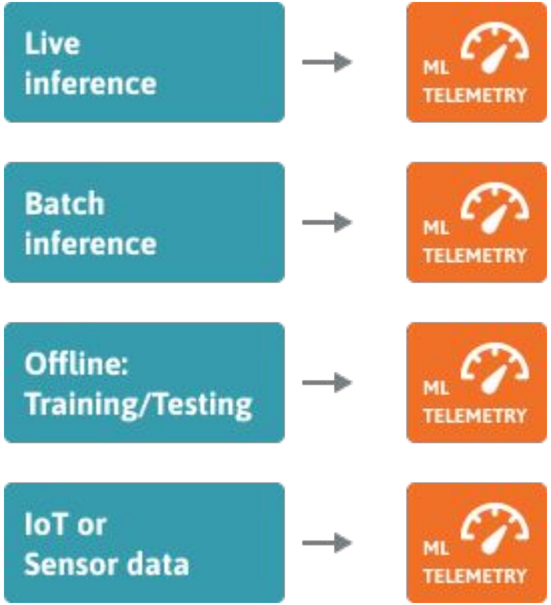
From logging to visualizing **data drift**



```
from whylogs.viz import ProfileVisualizer

mlflow_profiles = whylogs.mlflow.get_experiment_profiles(“experiment_1”)
viz = ProfileVisualizer()
viz.set_profiles(mlflow_profiles)
viz.plot_distribution("free sulfur dioxide", ts_format="%d-%b-%y %H:%M:%S")
```



Enabling ML monitoring: **Instrument data**



Live inference → ML TELEMETRY

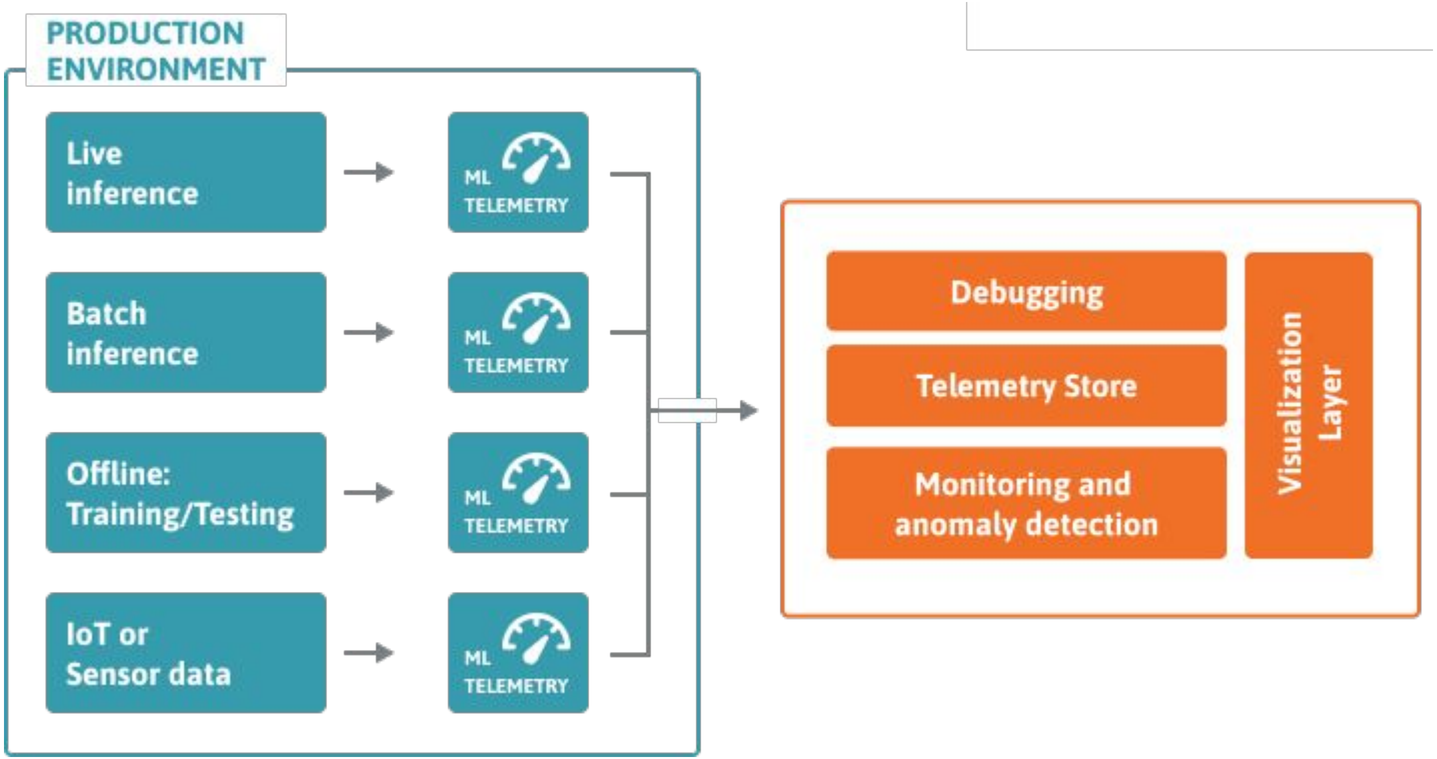
Batch inference → ML TELEMETRY

Offline: Training/Testing → ML TELEMETRY

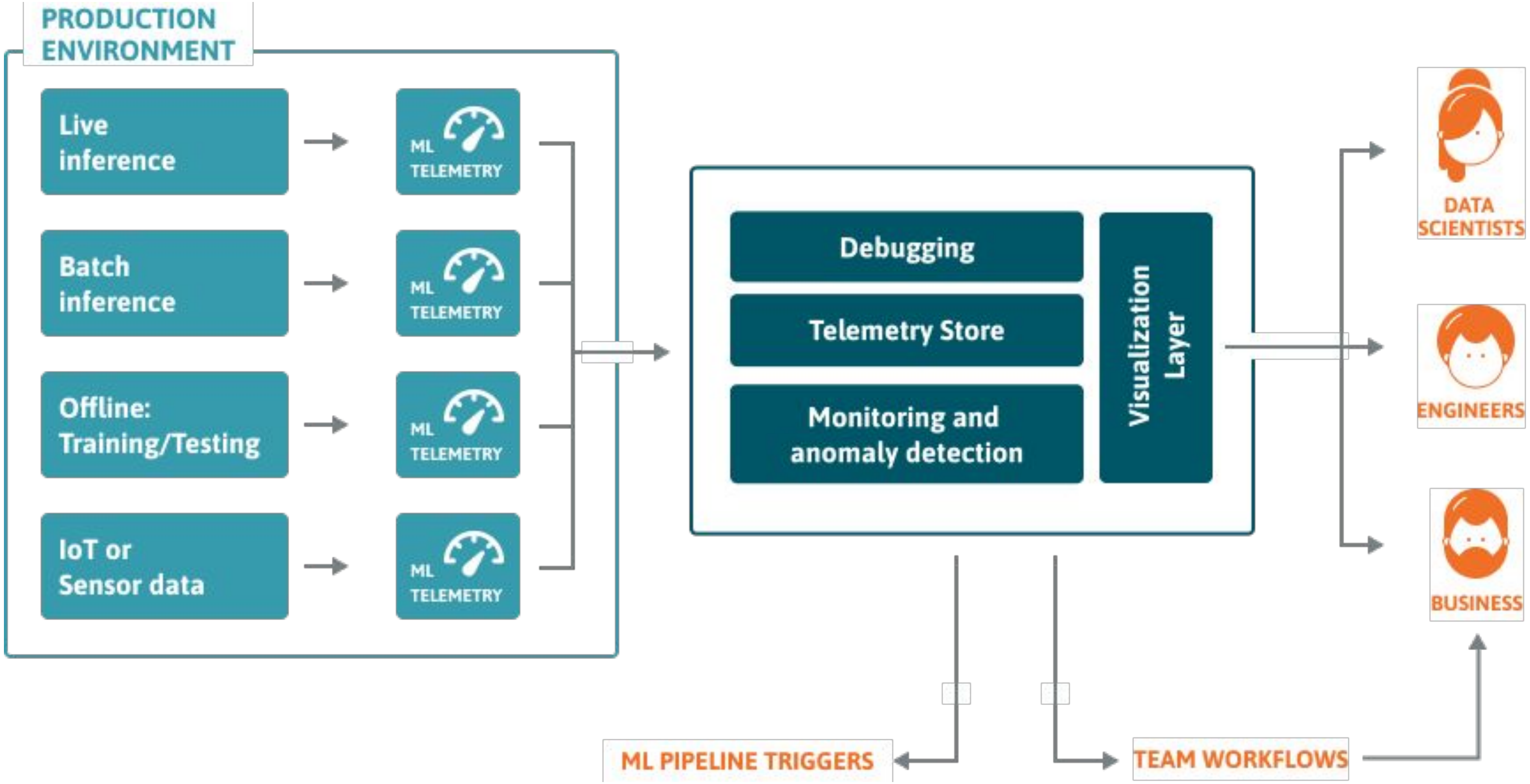
IoT or Sensor data → ML TELEMETRY

bit.ly/whylogs

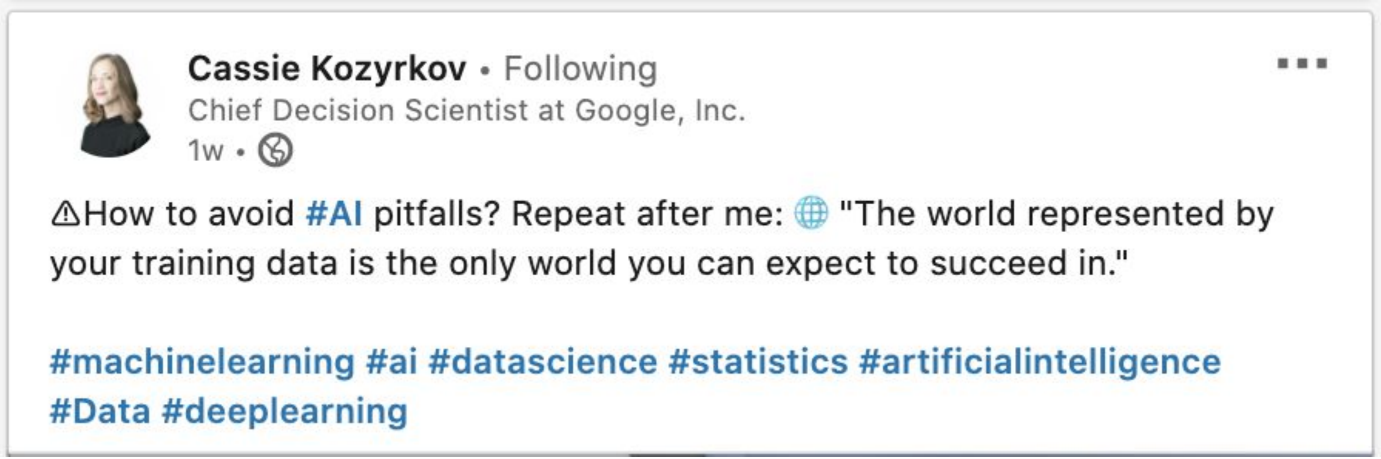
Enabling ML monitoring: **Identify anomalies**



Enabling ML monitoring: **notify stakeholders**



ML models struggle in the “wild”



Cassie Kozyrkov • Following

Chief Decision Scientist at Google, Inc.

1w • 🌐

⚠How to avoid #AI pitfalls? Repeat after me: 🌐 "The world represented by your training data is the only world you can expect to succeed in."

#machinelearning #ai #datascience #statistics #artificialintelligence #Data #deeplearning