

AIKO TAU

Digital infrastructure for developing and evolving AI/ML models

Navindu Amerasinghe · Antti Kolehmainen · Sergio Moreschini · David Hästbacka

TASE Tampere Software Engineering Research Centre

AI-native software systems

AI is becoming part of almost every software system



Data is essential



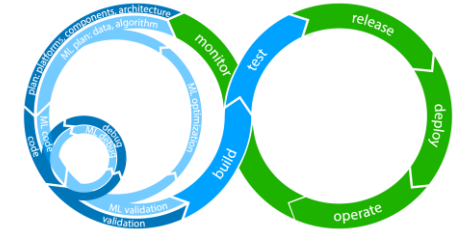
A system's behaviour is only as good as the data behind it.

Models keep evolving



Performance and capabilities improve with new generations of a model.

Evolution runs on data



Better models demand more data — accurate, contextualized, ...

TAU objectives in AIKO

1

Data for continuous AI/ML model development

Research how data is collected, how it is versioned, and how it is taken into continuous practices for evolving and deploying AI/ML models.

2

Practices companies can adopt

Develop working practices that industry can take up.

As vendor-neutral and platform-independent as possible.

3

Information management

Study how data access and data sharing could be managed.

Applicable guidance for companies what to consider.



Euroopan unionin
osarahoittama



AIKO TAU DEMO progress

Applying Artificial Intelligence in Business Automation Solutions

Traffic Speed Forecasting — an end-to-end MLOps pipeline

From raw motorway sensor data to live predictions, orchestration, and governance

TUNI / TAMK / Fintraffic



AIKO • RESEARCH GOALS

What the dataset has to support

Data collection and integration

Digitraffic • LAM • TomTom • Waze (later)

All sources in one linked dataset: every sign, station and incident placed on the road network, all times in UTC.

Statistical analysis and visualization

Time series • before-after • regression • GIS

Exact sign-change times, 5-minute and per-vehicle traffic, and coordinates plus road addresses for maps.

Impact assessment

Flow • safety • optimization • driver experience

Speeds and volumes up- and downstream of each change, with incidents and road weather as safety context.

Reliability assurance

Sensitivity • validation • bias assessment

Independent sources cross-checked (LAM vs TMS); gaps and faulty records flagged, never silently dropped.

AIKO • DATASETS

The datasets we use

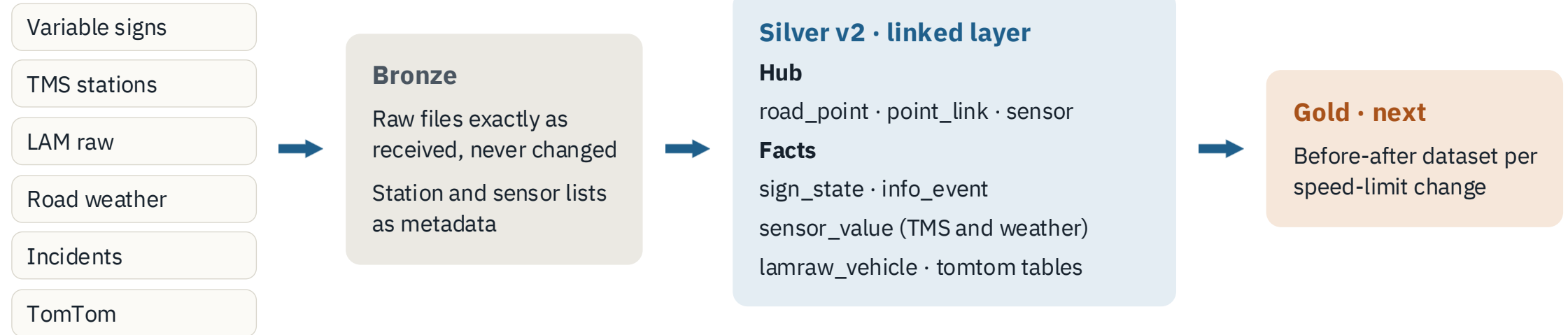
Digitraffic (Fintraffic) and TomTom; signs, TMS and road weather are snapshots about every 5 minutes.

Source	What it gives	Analysis week (24 Sep – 1 Oct)
Variable signs	Speed-limit and message changes	567 signs, 3,795 limit changes
TMS stations	Speed and volume per direction, 5-minute	519 stations
LAM raw	Every vehicle: speed, length, lane, direction	4,477 station-days
Road weather	Air and road temperature, friction, rain	530 stations
Incidents	Traffic announcements and the road stretch affected	113 incidents
TomTom	Segment speeds around query points	23–26 Sep, more uploading
Waze	Crowdsourced incident reports	Not included yet

AIKO · CURRENT INTEGRATION

From raw files to one linked dataset

Sources



Joined by **location** (one point id, road addresses, nearest-station links) and by **time** (UTC everywhere).

AIKO · DEMO PREPARATION · STATUS

A linked dataset for the sign-impact demo

Signs, traffic, weather and incidents joined by location and UTC time for 24 Sep – 1 Oct 2026, the week where every source overlaps.

3,795

speed-limit changes
captured on 303 signs

52.8 M

TMS and road-weather
readings cleaned and linked

4.55 GB

vehicle-level LAM data,
4,477 station-days

0.07 %

LAM vs TMS count
difference in cross-check

Done

Location hub: 1,616 points, signs → nearest stations
Sign changes, incidents, TMS and weather values

In progress

Vehicle-level LAM data for the full
week

Next

TomTom speeds, full-week
validation

PROGRESS OVERVIEW

linked data to the demo

An end-to-end, open-source MLOps pipeline forecasting motorway speed 15 minutes ahead - from raw sensor data to a live dashboard, fully orchestrated and governed.

✓ Data pipeline

Bronze → Silver → Gold on open-source tools (Spark, Delta, SeaweedFS)

✓ Forecasting model

LightGBM trained, evaluated, and registered (MAE 5.43 km/h)

✓ Live dashboard

Predicted vs. actual speed in Grafana, per station & direction

✓ Orchestration

Full pipeline runs as one Dagster graph — one click or scheduled

✓ Governance

All tables catalogued in Unity Catalog (silver / gold schemas)

✓ Open-source

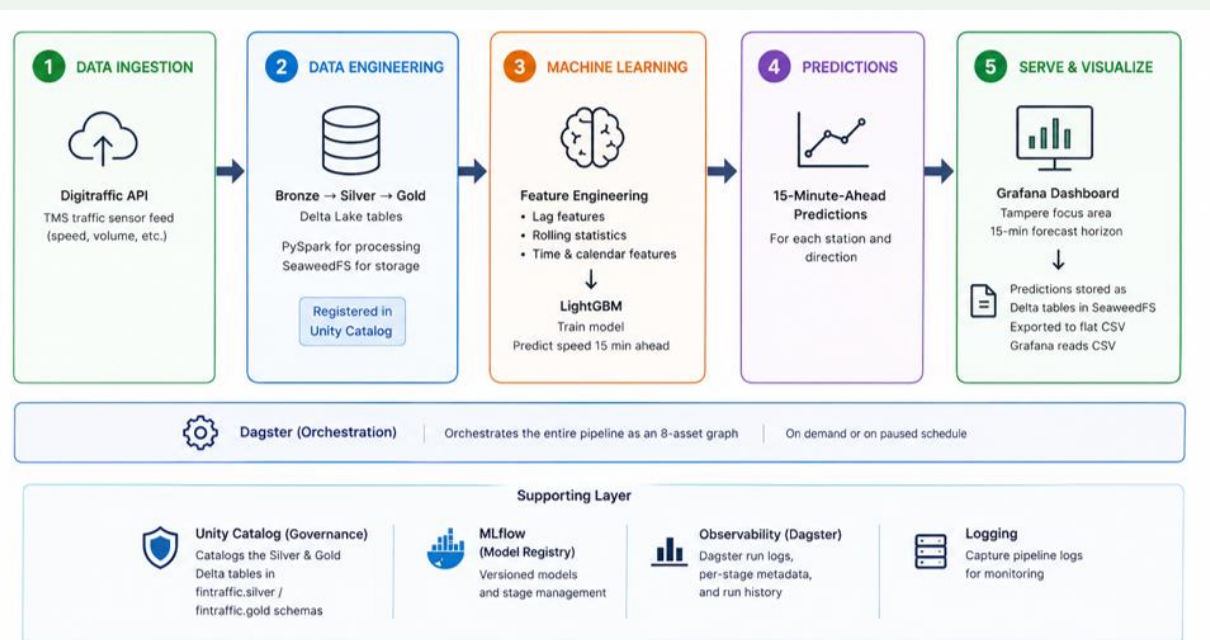
Entirely non-proprietary, self-hosted - no vendor lock-in

STATUS & ROADMAP

Where we are — and what's next

✓ Done

- ✓ End-to-end data pipeline (Bronze→Silver→Gold)
- ✓ LightGBM model trained, evaluated, registered
- ✓ Live Grafana forecasting dashboard - locally implemented
- ✓ Full Dagster orchestration (8-asset graph) - locally implemented
- ✓ Unity Catalog governance (5 tables)



→ Next steps

DATAOPS

1. Integrate TomTom data alongside LAM/TMS stations (LAM primary, TomTom secondary)
2. Version 2 - scale to all stations & a larger dataset (finer 5-minute data)
3. Automate catalog sync inside the pipeline

MODELOPS

1. Analyze how variable speed limits affect traffic speed and with what delay
2. Model monitoring & drift detection

DEVOPS

1. Scale Spark from local to the cluster
2. Deploy Grafana and Dagster on the server

