

Hi!



Francesco Di Corato, PhD
Co-founder & Scientific Director



The Data Layer for the Physical World



License: **Apache-2.0**

Supporting





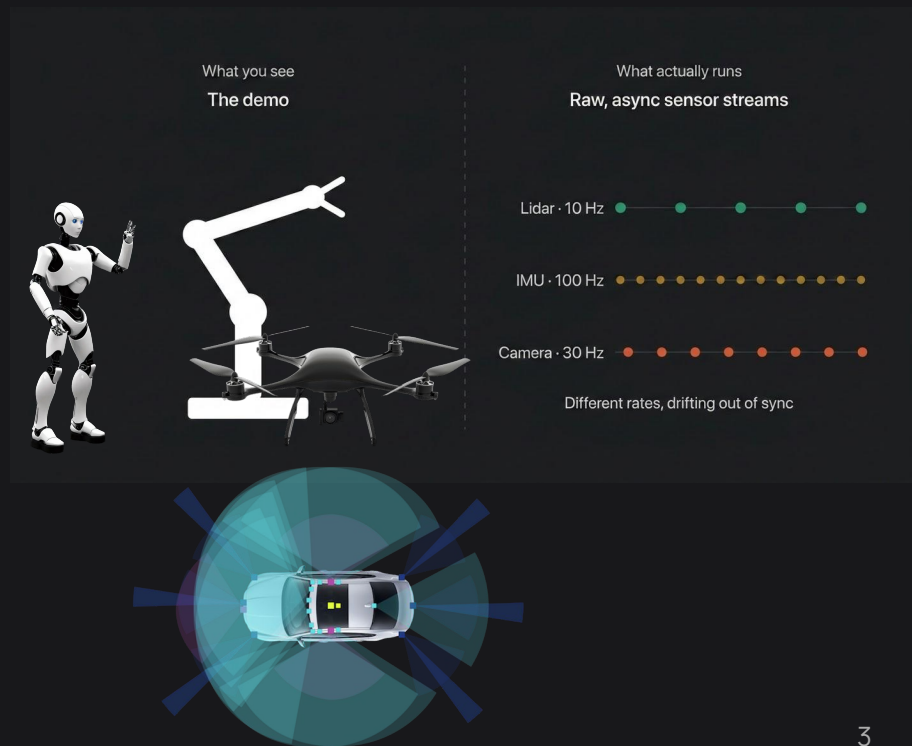
The Data Shift

From Classical Robotics to Physical-AI

The bottleneck moved

For years, robotics was gated by control and algorithms.

Today it's **gated by data**

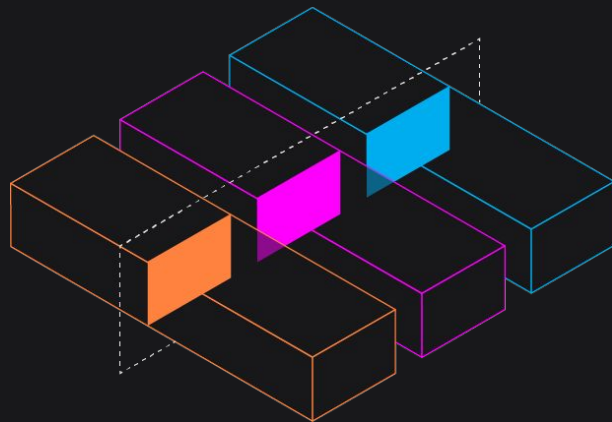


The multi-modal timeseries problem

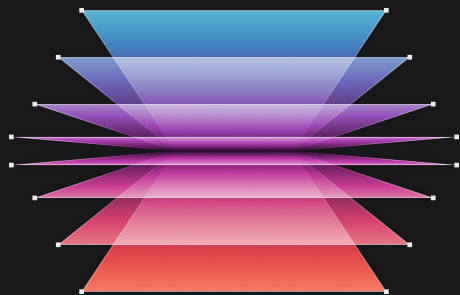
Robotics data is acquisition-bound.

A GPS reading, an IMU sample, a camera frame are **meaningful together** because they describe the same physical moment.

Async, multi-rate, physically bound: the **timestamp** isn't metadata. **It's the meaning**.

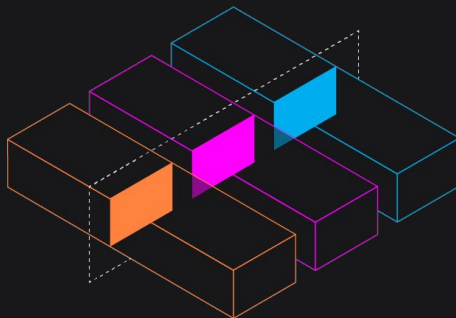


The Robotics Data pillars



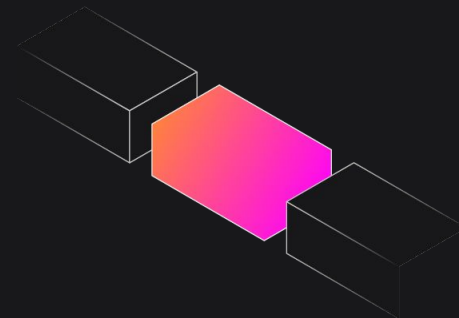
Schema

The Semantic rules



Channel

The data stream



Recording

The Ensemble



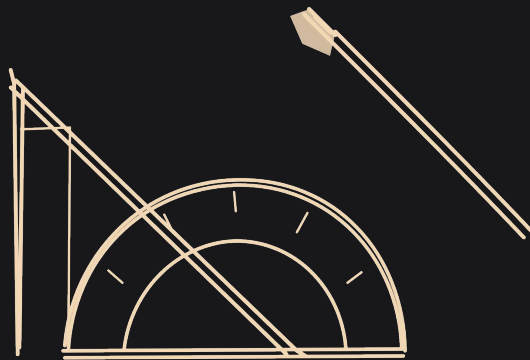
Data as *by-product*

Recordings for tuning, debug and replay

Linear stream of messages as files.

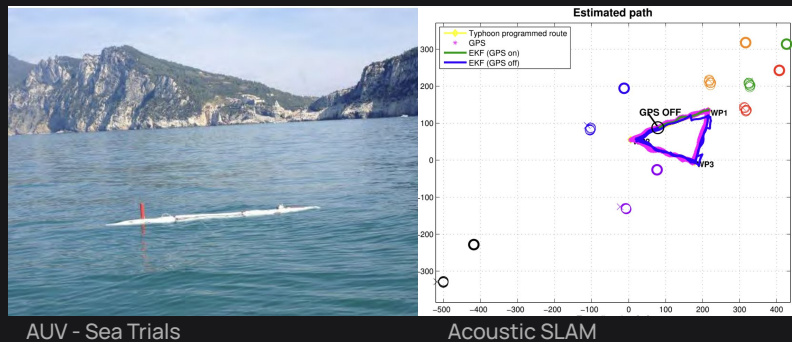
Easy to write, **hard to index, query or stream at scale**.

No need for complex data architectures

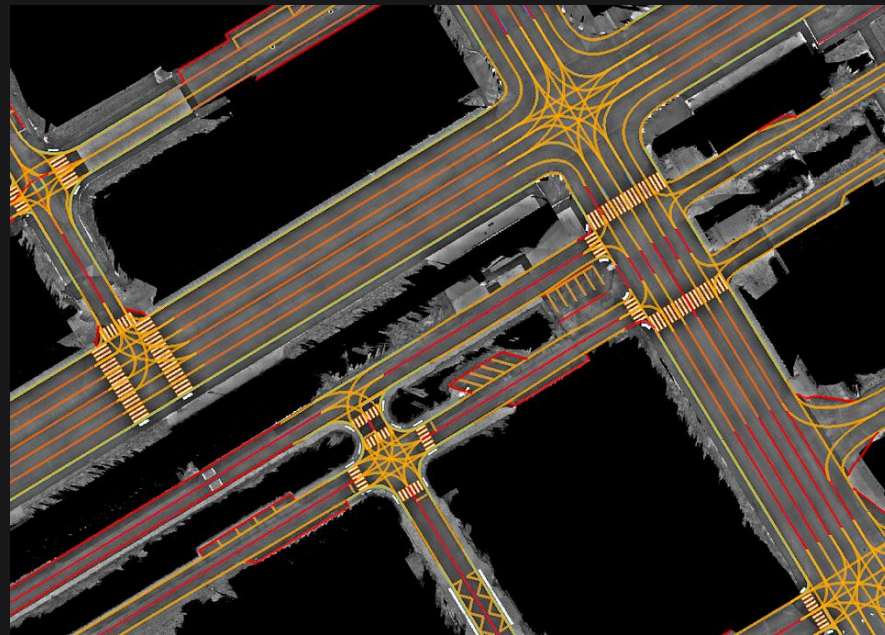


Some context from my past life...

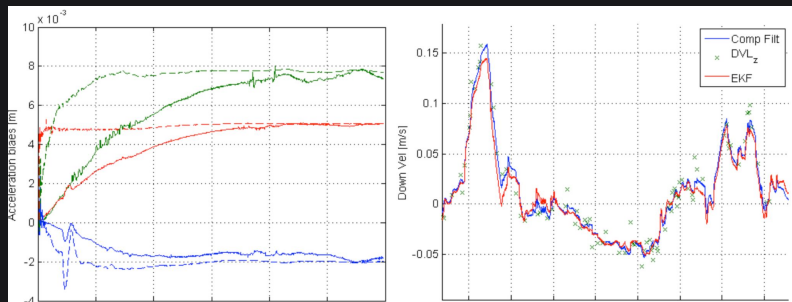
At sea: **online**, real-time, GPS-denied



On the road: **offline**, batch, at scale



HD map – aligned imagery, lane geometry (*image from the web*)



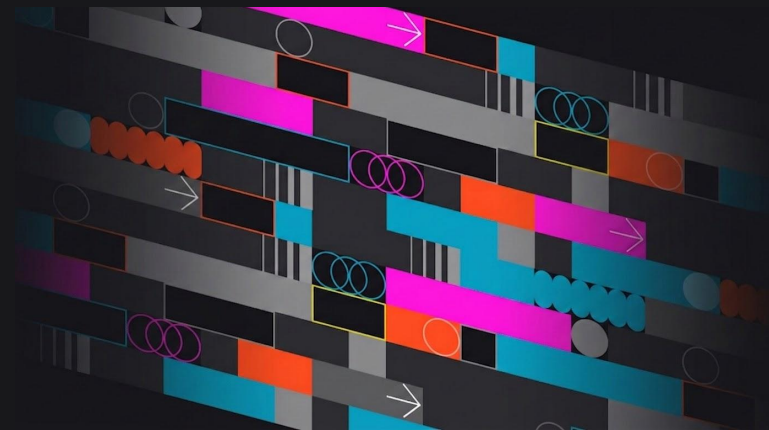
EKF tuning - bias & DVL velocity



Data became the Fuel

The model learns from experience.

The data becomes the **asset**.



01

Aligned

so a model can train

02

Organized

there's a lot of it

03

Queryable

find, fix, rebalance

04

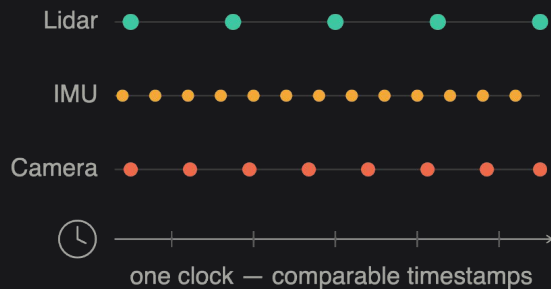
Trustworthy

certify what you built

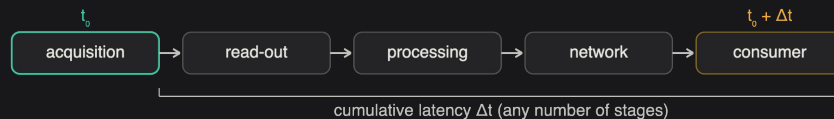


Aligned, i.e. “Synchronized”

Same time origin : One clock across all measurements



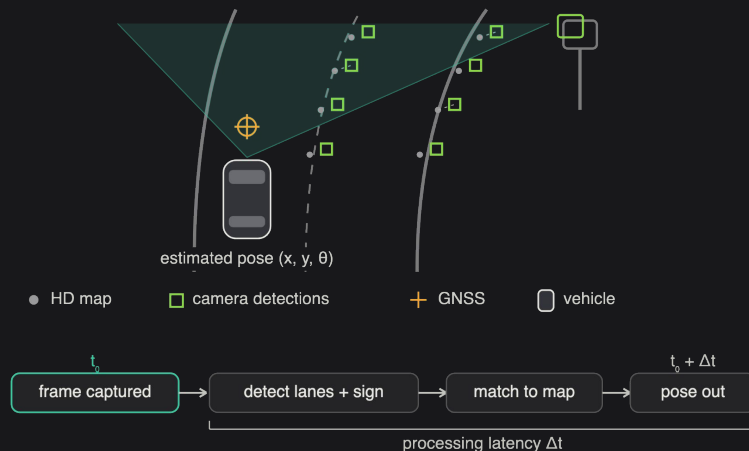
Timestamp closest to **sensor acquisition**



The classical “Synchronization” concept

Practical Example

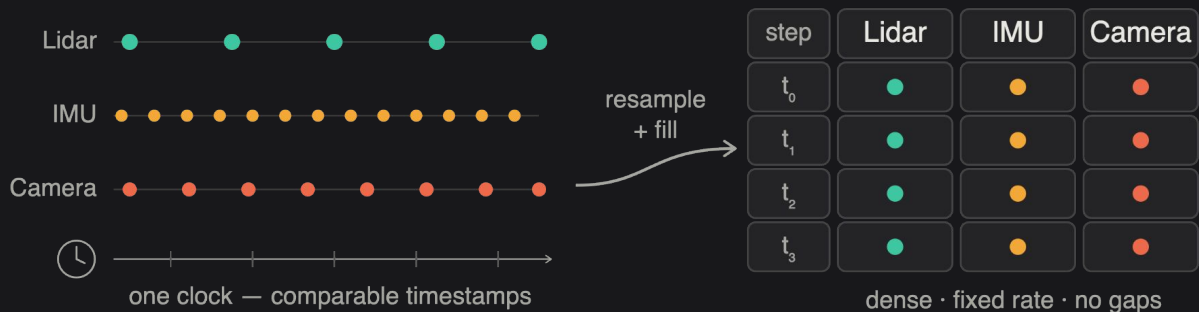
Map-based **on-line** localization



“Synchronized” now means 2 things

Same time origin: One clock across all measurements

Dense dataframe: One row per tick, every sensor filled.



The second doesn't replace the first. It's **built on top of it!**



Organized: too much to keep in *bags*

≈ 1–5 TB

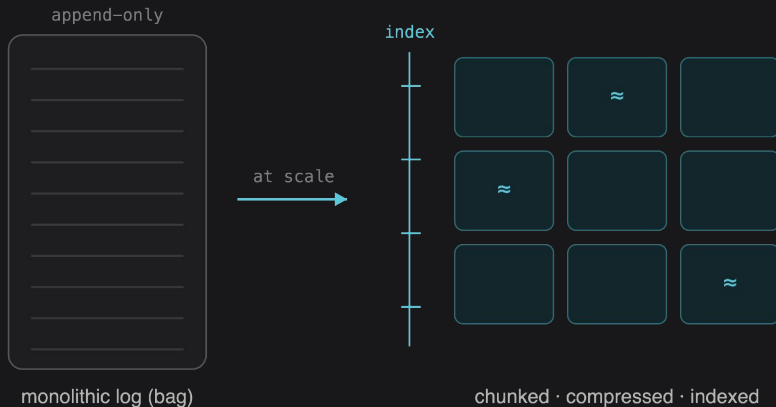
/ hour · per vehicle

100+ TB/day at fleet scale – you can't just append anymore

Storage

Compression

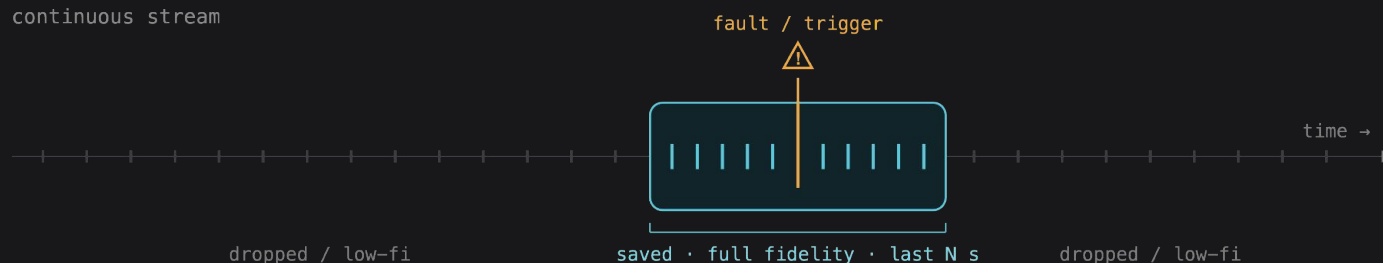
Indexing



Saving is easy. Knowing when is not

Keep what matters.

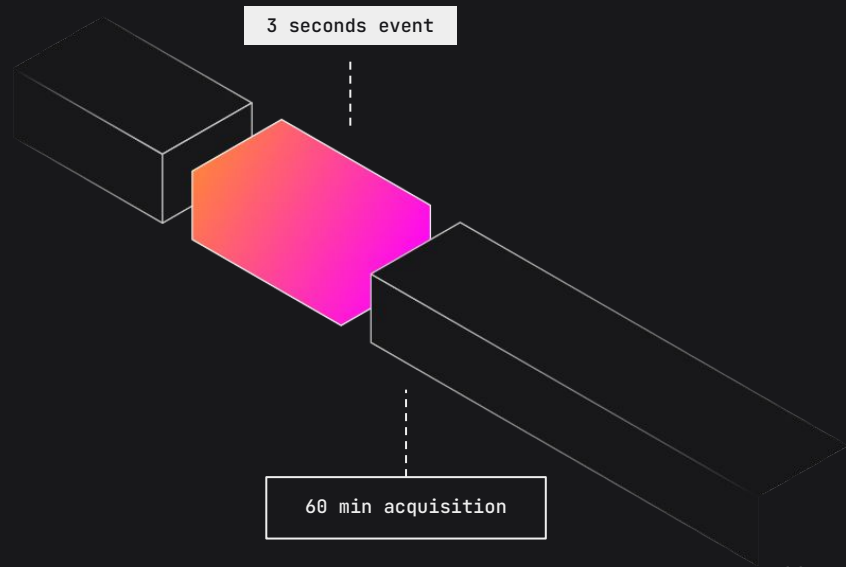
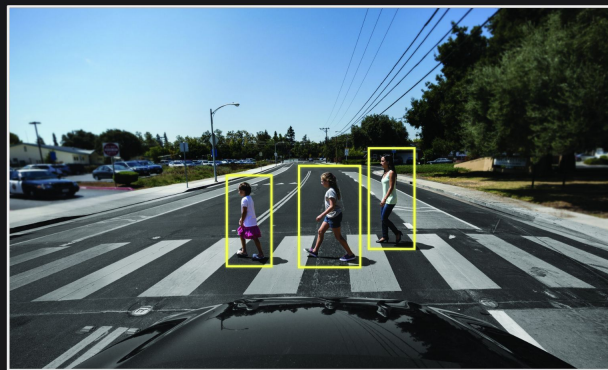
Faults, disengagements, edge cases. What you keep is a **SOTIF-like decision**.



Searching needs indexing

Typical autonomous driving
acquisition sessions last ~60'.

How do we find the **exact 3-second
emergency braking event** of a pedestrian
crossing the road?



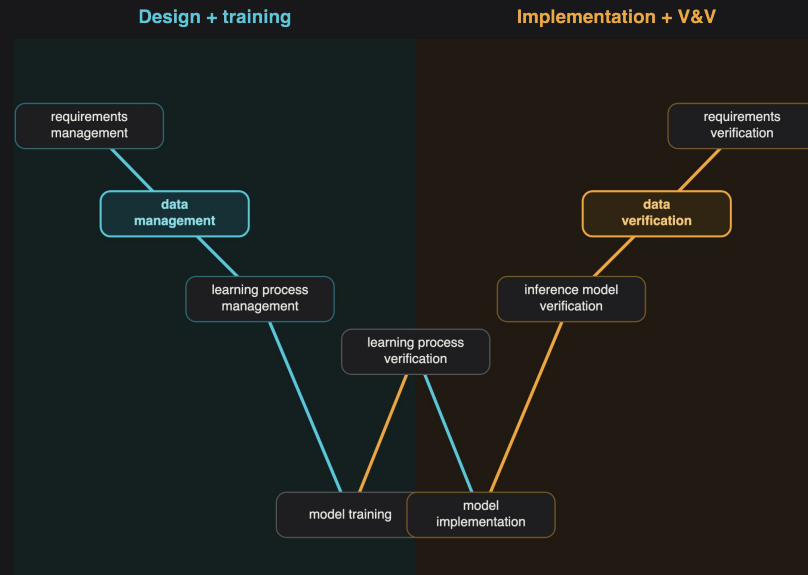
The AI-Certification Plan

Regulators already treat data as the safety case

EASA's AI Roadmap 2.0 puts the data lifecycle at the **center of assurance**.



<https://www.easa.europa.eu/en/downloads/137919/en>



Certification and Sovereignty

Trustworthy = measurable, traceable, monitored.

CERTIFIABLE

01

Measured data quality

Accuracy, completeness,
representativeness, balance

02

Lineage & Immutability

Every dataset, transform and
split traceable to the raw source

03

Traceability by design

requirements ↔ data ↔
scenarios ↔ model ↔ tests

04

Continuous monitoring & change control

Certify a process for safe
evolution, not a frozen model.

SOVEREIGN

05

Sovereignty by design

Businesses and individuals
control the data they generate

06

Privacy

Sensitive, proprietary. Need infrastructure
for sharing and reuse, keeping ownership.



Many other needs...

Data-as-a-Service

Moving beyond rigid, unsearchable files into a unified, instant-query data platform.

Serverless Processing

On-demand, event-driven execution to enable data-driven processing pipelines.

Scale-to-Zero

Heavy GPU/CPU workloads scale up to n instantly, and drop to zero when empty. No infrastructure waste.



Data is the Fuel → Needs an Engine

01

Aligned

so a model can train

02

Organized

there's a lot of it

03

Queryable

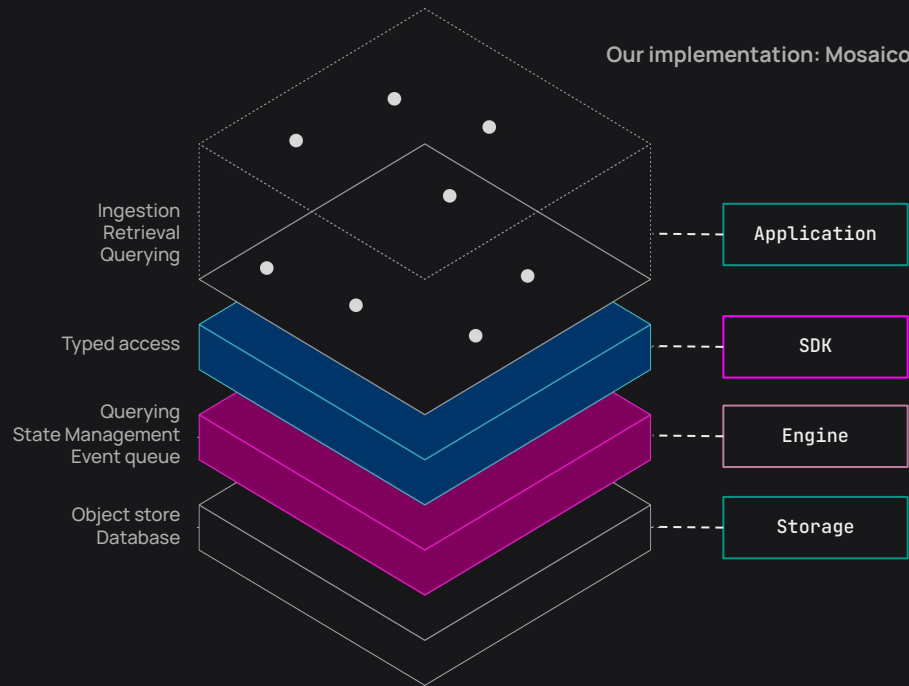
find, fix, rebalance

04

Trustworthy

certify what you built

None of that comes from a **file**



It comes from an **architecture**



Want to know more?



`francesco.dicorato@mosaico.dev`



`github.com/mosaico-labs/mosaico`



`docs.mosaico.dev`

