



FUSE

Record-agnostic multimodal semantic
embeddings for industrial data

Like finding a needle in a haystack,
except there are many haystacks...
...and the needle is fragmented across them.



Agenda

 Problem & core idea

 System design

 Tasks & governance

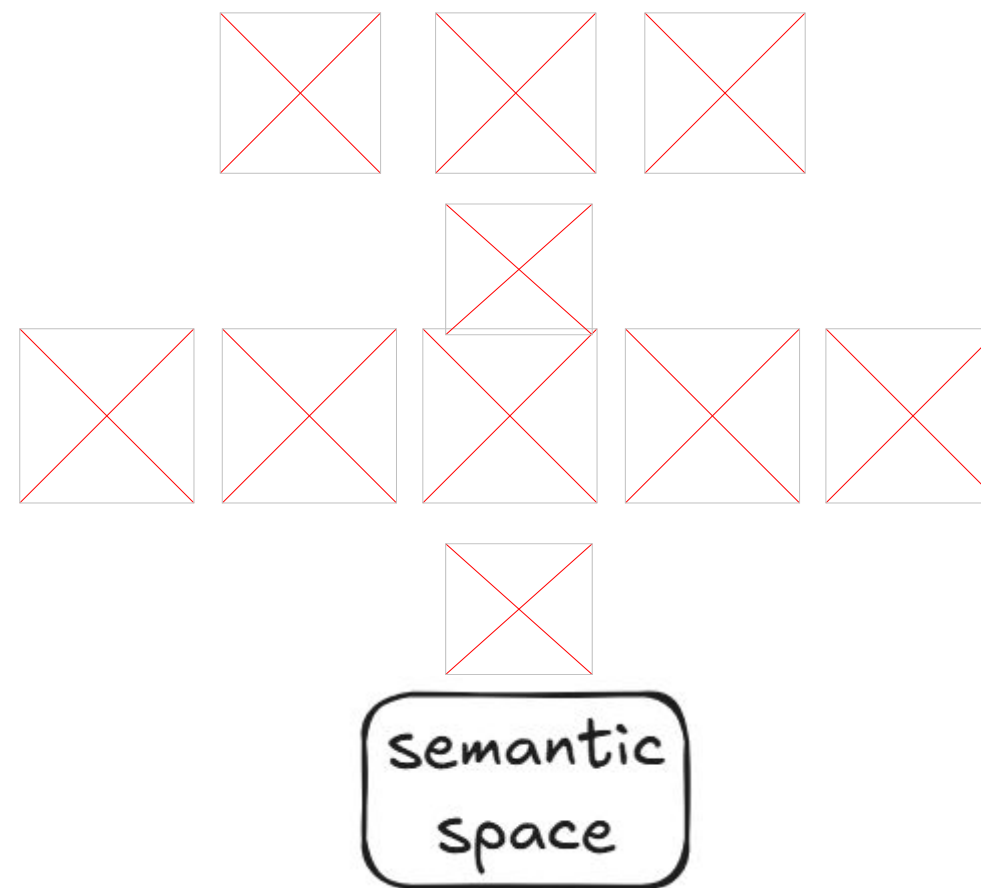
 Evaluation

 Implications

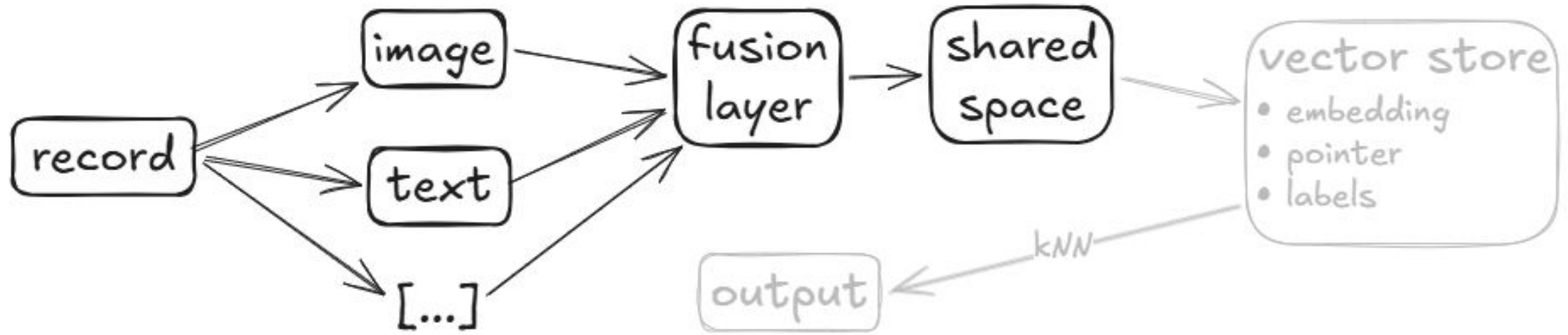
Problem & core idea

A "semantic join" for siloed data

- Operational evidence distributed across stores, records, and modalities w/out joins
- Modality encoders and fused embedding space for case-based reasoning tasks

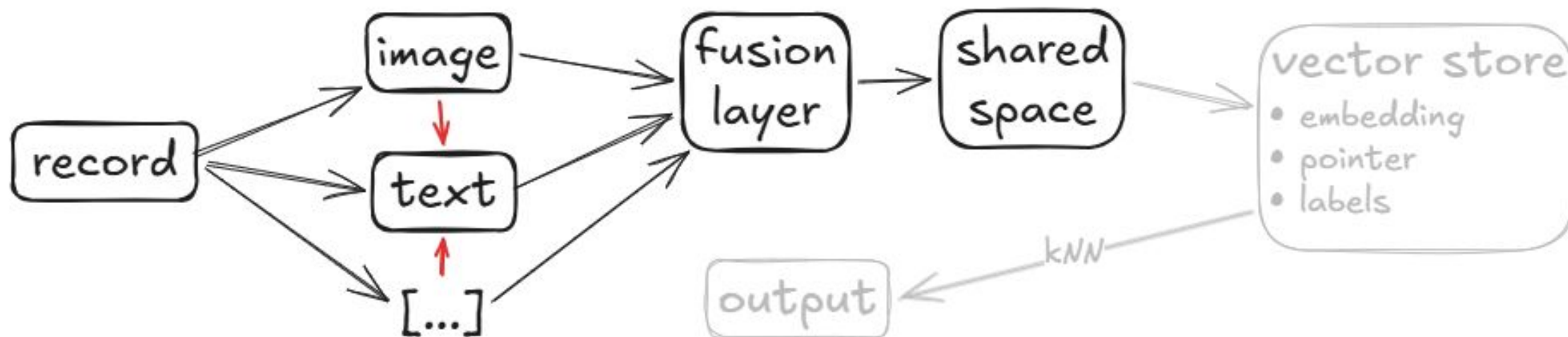


System design: Representation



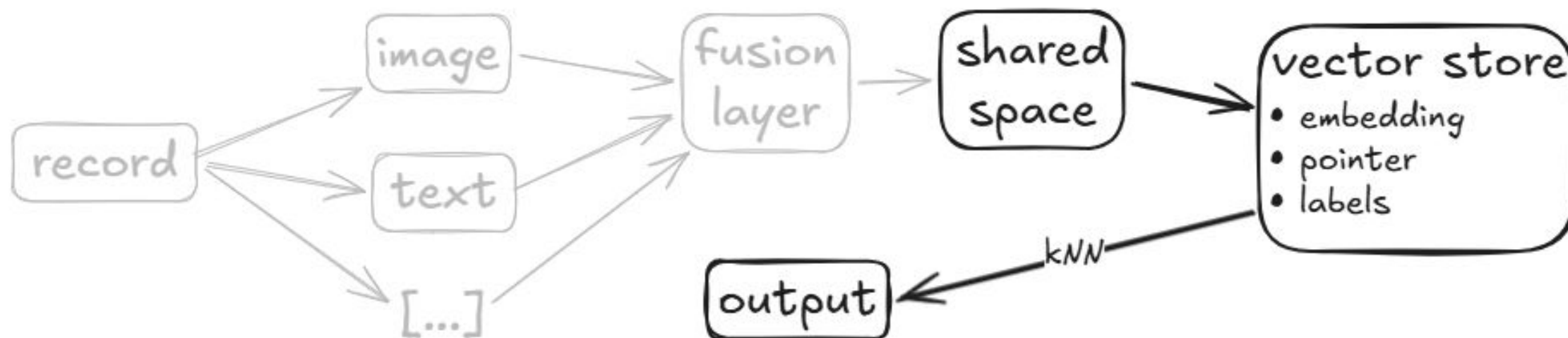
- Embed each modality "chunk" of a record
- Fuse modality embeddings into one shared space
- Goal: cross-record comparability without joins

System design: Training



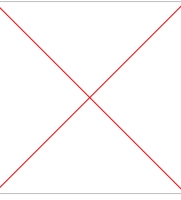
- Train using intra- and inter-record modality pairs
- Text acts as hub for alignment
- Learned fusion of modality embeddings

System design: Inference



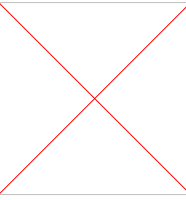
- Retrieve nearest neighbors in vector store
- Aggregate neighbor labels for prediction
- Use pointers to fetch related records (multi-hop)

Supported tasks



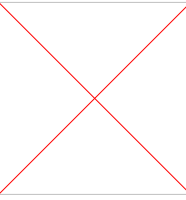
- **Classification & regression**
 - Similarity-weighted voting over kNN labels
- **Anomaly detection**
 - Max cosine similarity below a threshold
- **Related records**
 - Nearest neighbors across types
 - Useful as RAG within AI agent
- **Weak labeling**
 - KNN classifications + human review

Governance



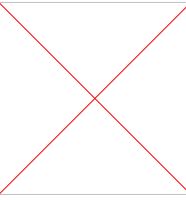
- **Pointer-based original record retrieval**
 - Vector store is a map, not warehouse
 - No duplication of data into "super silo"
 - Obey data-level authorization
- **Case-based decision support**
 - Auditability
 - Traceability

Evaluation



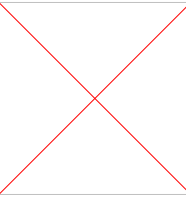
- End-to-end evaluation across tasks
- Task-appropriate metrics
 - classification, regression, relatedness
- Record-level UUID splits prevent leakage

■ Implications: Representation as infrastructure



- **"Semantic join"**
 - Text-as-hub, leverage existing text encoders
- **Case-based reasoning**
 - Enables traceability & auditability
 - Emergent context-based reasoning
- **Representation + retrieval**
 - Works well in initial tests
 - Evaluating quality can be a challenge

■ Implications: Operating & scaling the system



- **Latency**
 - Function app with warm infra
 - Parallel encoding and hash & cache
- **Scalability**
 - Add new record type or modality
 - Modularity mitigates some retraining pain
- **Other challenges**
 - Data imbalances in training and late fusion
 - Correlation-to-causation
 - E2E evaluation

Thank You!



William Dyer

Global AI Tech Lead

Terumo Blood and Cell Technologies

william.dyer@terumo.com

www.linkedin.com/in/william-dyer/

