Erwin Deng

Antonio Santamaria

Lost in Translation

MIT SLOAN · FORD MOTOR COMPANY · CAPSTONE 2026

Mapping Entities Across Undocumented Plant Data

Faculty Advisor: Prof. Retsef Levi | **Ford Mentors:** Chaitra Hedge · Ishan Patel · Al Hogan · Max Dsouza · Balpreet Singh

Project Overview

Problem

Ford can't scale analytics

Every data source has to be stitched together by hand.

50+ PLANTS

5 million CARS / YEAR

10+ DATASETS

0 LINKS BETWEEN THEM

Why

Same entity, different names

Every dataset has its own naming convention for the same physical thing. **Dataset 1** might call a line "L1," "L2," "LA", while **Dataset 2** calls the same lines "Line1," "Line2," "LineA."
Multiply that across 10+ datasets, 50+ plants and multiple hierarchy levels, and manual matching by a subject-matter expert becomes a bottleneck that **doesn't scale**.

Dataset 1

Area A

Department

Line 1	Workstation	Workstation	Workstation
Line 2	Workstation	Workstation	Workstation

Department

Line 1	Workstation	Workstation	Workstation
Line 2	Workstation	Workstation	Workstation

Dataset 2

Area A

Line

Station · Station · Station · Station · Station

Line

Station · Station · Station · Station · Station

OUR APPROACH Our program surfaces the most probable matches between any two selected columns — so the SME **confirms** matches instead of searching for them.

Matching Methods · From Simple to Sophisticated

MODEL 1

Fuzzy match

String-overlap on raw names (RapidFuzz). Simple, deterministic, clears the near-identical names.

MODEL 2

Embeddings

Serialize each entity to a sentence, embed it and rank candidates by meaning.

MODEL 3

Line-anchored

2 stage model: identify first the line, and then predict only the entities in that line.

Serialization

Each entity becomes one sentence: **"Manufacturing node 'Trim 1 MainDrive' at station 'Trim 1' on line 'Trim' in area 'Final'."**

We rank the top 5 candidates **recall@1** / **recall@5**

Results · Matching Is Easy at Higher Levels, Harder Deeper in the Hierarchy

Match Assistant — COLUMN 1 Node → COLUMN 2 Line

SELECTED · DATASET 1 | Node **"Trim 1 MainDrive"** · Area **"Final"**

RANKED CANDIDATES · DATASET 2

Line **#100 Principal** · Dept. **Trim I & II**, Area **Final Assembly**	85%
Line **#200 Principal** · Dept. **Trim I & II**, Area **Final Assembly**	80%
Line **#300 Principal** · Dept. **Chassis**, Area **Final Assembly**	10%

SME picks two columns; the program ranks the most probable matches for human confirmation

VALIDATION To measure how well the program performs, we tested it against manually-verified ground truth at the Hermosillo Assembly Plant.

Solved **Node → Line**

Coarse levels, easy match

Fuzzy string overlap

recall@1	35%
recall@5	44%

Embeddings semantic

recall@1	74%
recall@5	96%

Hard **Operation → Node**

Deeper levels, lack of information

Embeddings semantic

recall@1	2%
recall@5	4%

Line-anchored predict line first

recall@1	7%
recall@5	36%

Why It Matters & What's Next

From disconnected data to root-cause analysis

The system proposes and ranks candidate matches with evidence for an expert to confirm. This mapping will allow machine metrics to be attached to the processes they measure, so a defect can be traced to its equipment, bottleneck, or workload.

Build once. Scale everywhere. The same method generalizes across levels and plants.

Then

Manual matching by a subject-matter expert.

Now

A program that ranks matches by probability and speeds up matching for the SME, validated against ground truth.

Next

Fold the program into an app scaling across plants; gather feedback and iterate.