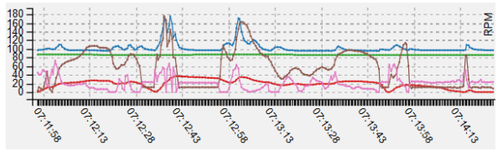
general motors

An ontology-based text mining and semantic similarity system for knowledge discovery in the automotive domain

Dnyanesh Rajpathak, Ph.D.
Staff Researcher
Advanced Analytics Center of Excellence,
Chief Data Analytics Office
Email: dnyanesh.Rajpathak@gm.com

Different stages of vehicle life cycle and data collection

DFMEA / Requirements	Station to Station End of line	Pre-launch Verbatim Telemetry	Verbatim Telemetry	Verbatim Telemetry



"Ignition On/Ignition Off"

Diagnostic Trouble Codes (DTCs)
P0300 | **B**0001 | **U**1040 | **C**0550

Telemetry

~ Hundred of thousands/year

Design	Manufacturing	Pre-launch	Launch	Warranty

Key challenges associated with the text data

- Heterogenous data sources
  - Technical, non-technical, and survey
- Lean language
  - ‘Engine Control Module’ vs ‘Powertrain Control Module’
- Surface representation consists of limited/no overlap
  - ‘Vehicle Loss of Power’ vs ‘PCM P0300’
- High volume data (hundred of millions)
  - Curse of dimensionality
- Different types of noises
  - Misspelling, run-on-word, additional whitespaces, abbreviations, etc.

Data cleaning pipeline is developed to pre-process the data

Pre-warranty

Warranty

External (e.g., National Highway Traffic Safety Administration)

~ Pre-launch verbatim

Warranty Repair Verbatim

Technician Assistance Verbatim

Customer Assistance Verbatim

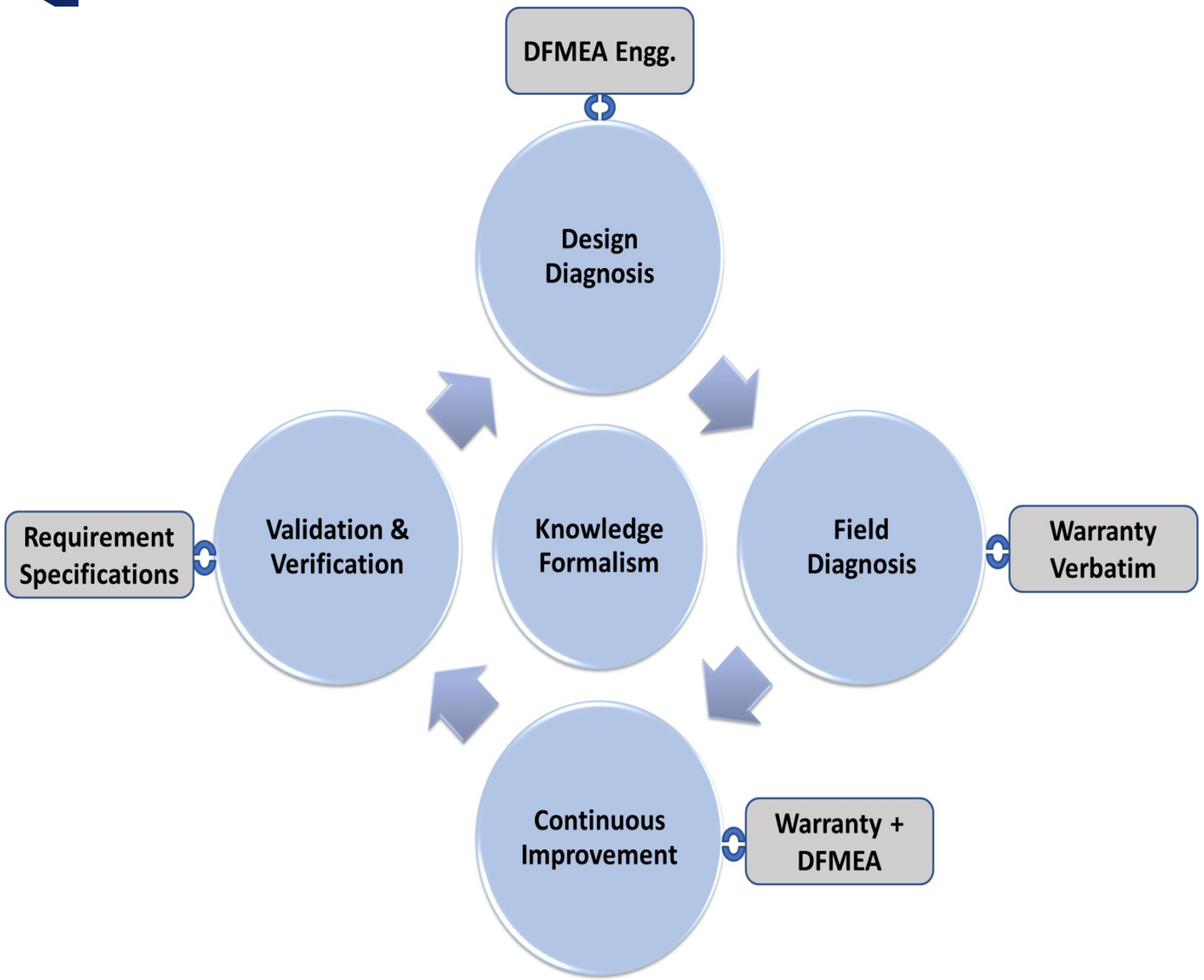
Vehicle Owners Questionnaire NHTSA

Misspelling Correction → Extra White Spaces → Word2Vec

Run On Word Correction → Abbreviation Disambiguation

Cleaned Unstructured Data

Business problems



- **Knowledge formalism**
  - Automatic ontology learning from warranty data
- **Validation & Verification**
  - Automatic linking of requirements using semantic similarity
- **Diagnosis & Prognosis**
  - Automatic development of fault dependency matrix
  - Text mining for knowledge discovery from the warranty data
- **Continuous Improvement**
  - Augmentation of DFMEAs considering new symptoms and failure modes discovered from the warranty data

An ontology-based text mining system for knowledge discovery

"CS Vehicle with loss of power. Detected leaks in cooler line, it was loose, repaired"

Warranty Data

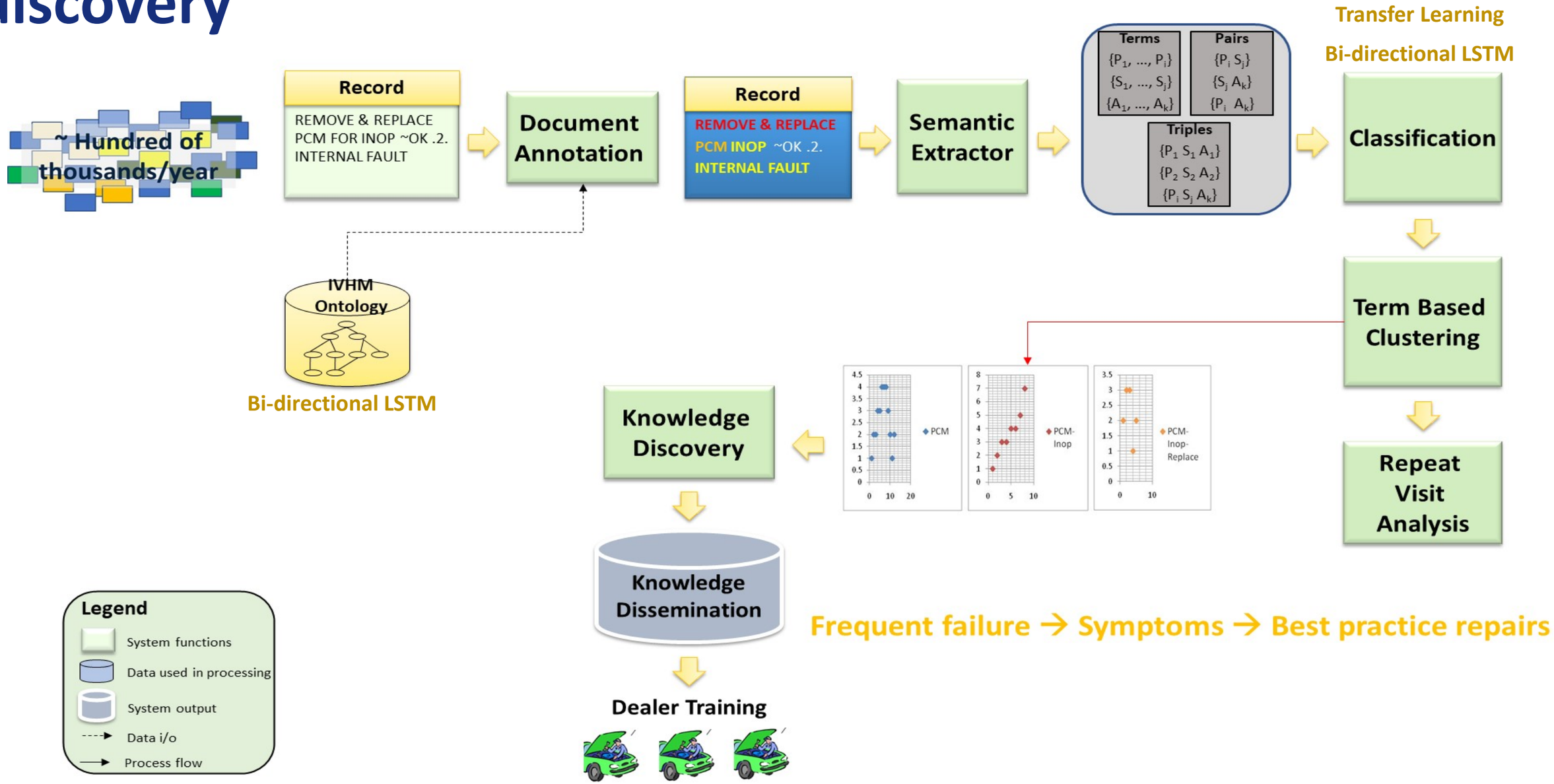
~ Hundred of thousands/year

SCAN TOOLS

Service/Quality

- Frequent Failures → Best Practice Repairs
- Legacy Issues
- Dealer Training Need

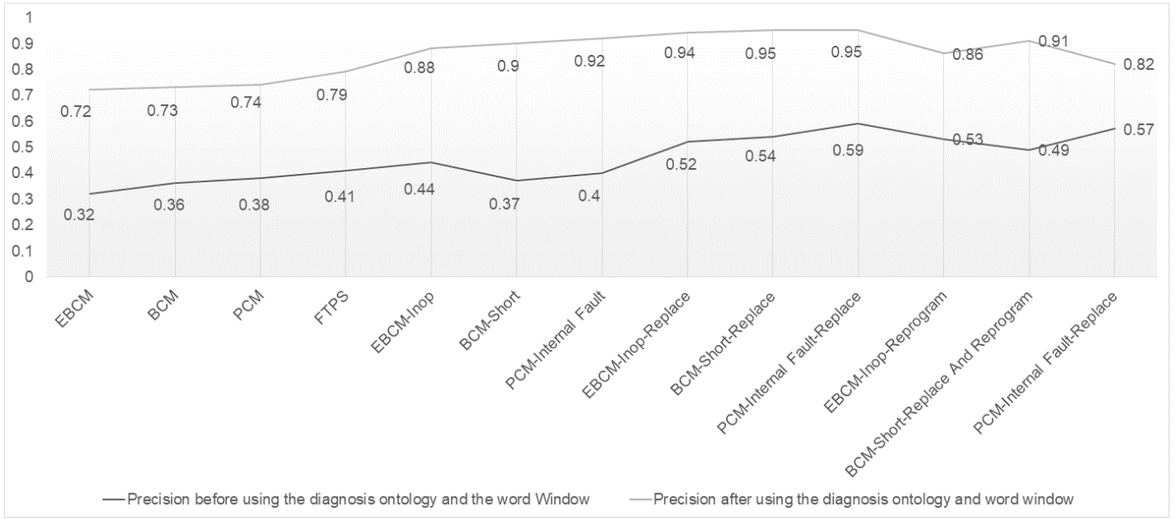
An ontology-based text mining system for knowledge discovery



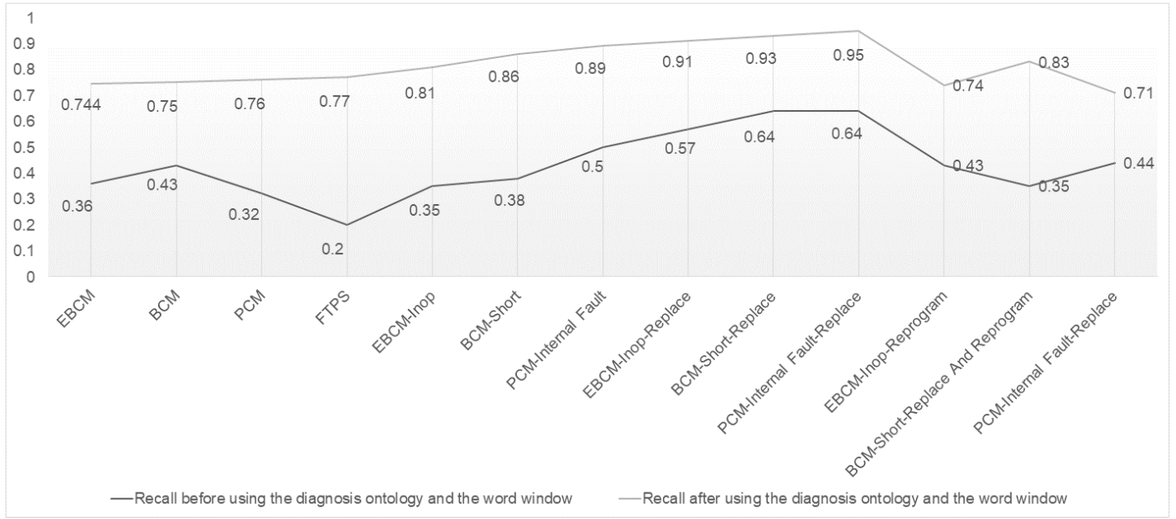
Frequent failure → Symptoms → Best practice repairs

Results

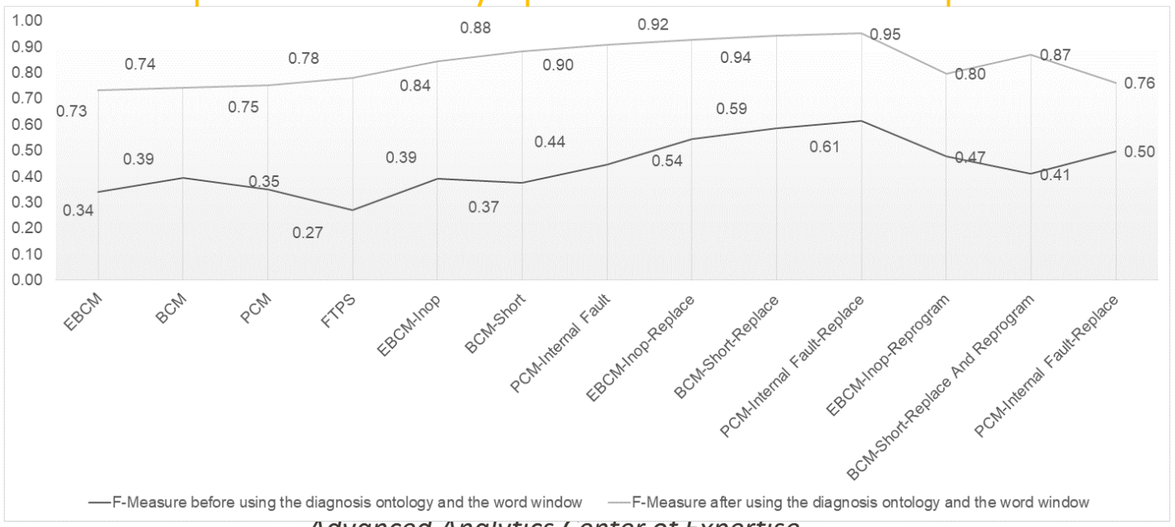
Precision: Frequent Failure → Symptoms → Best Practice Repairs



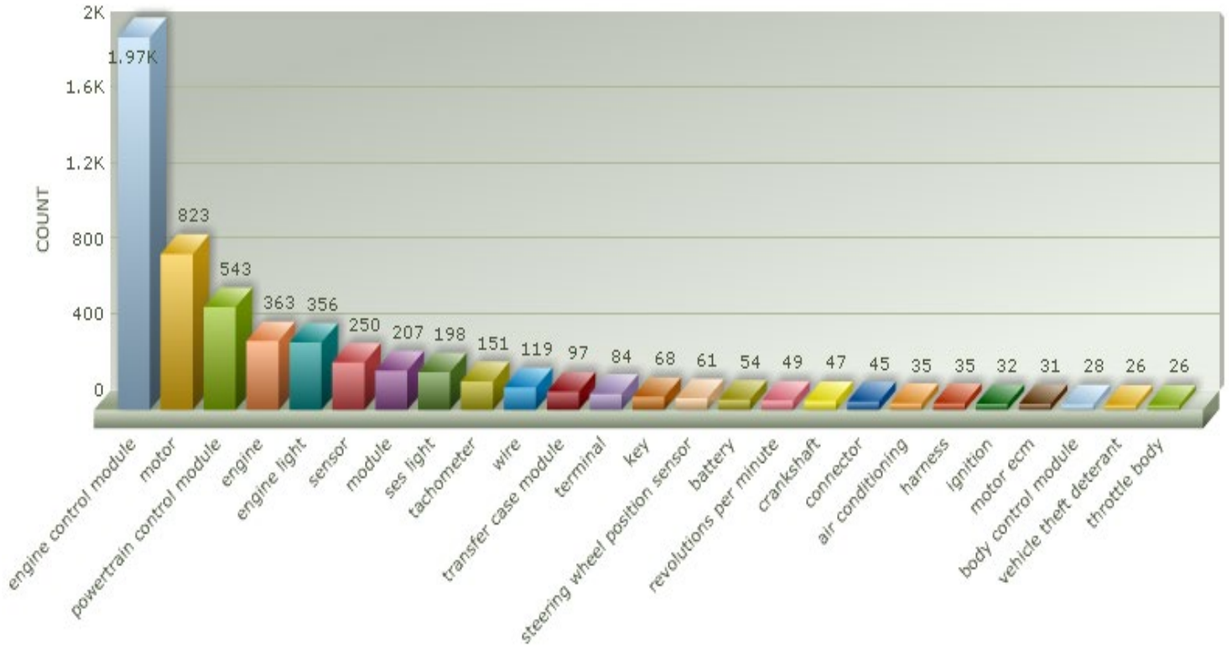
Recall: Frequent Failure → Symptoms → Best Practice Repairs

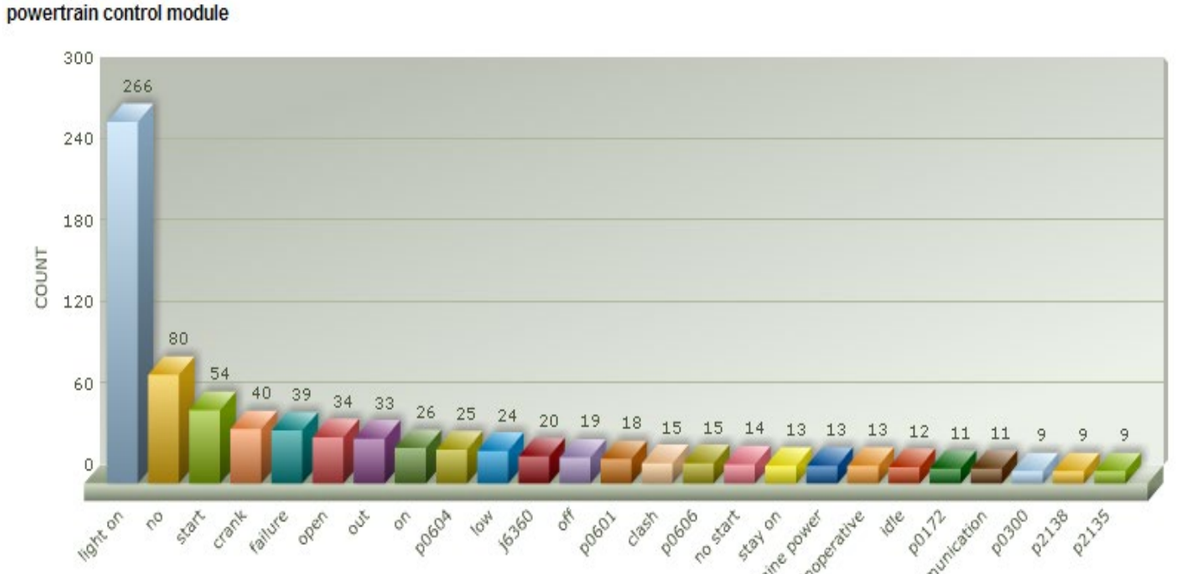


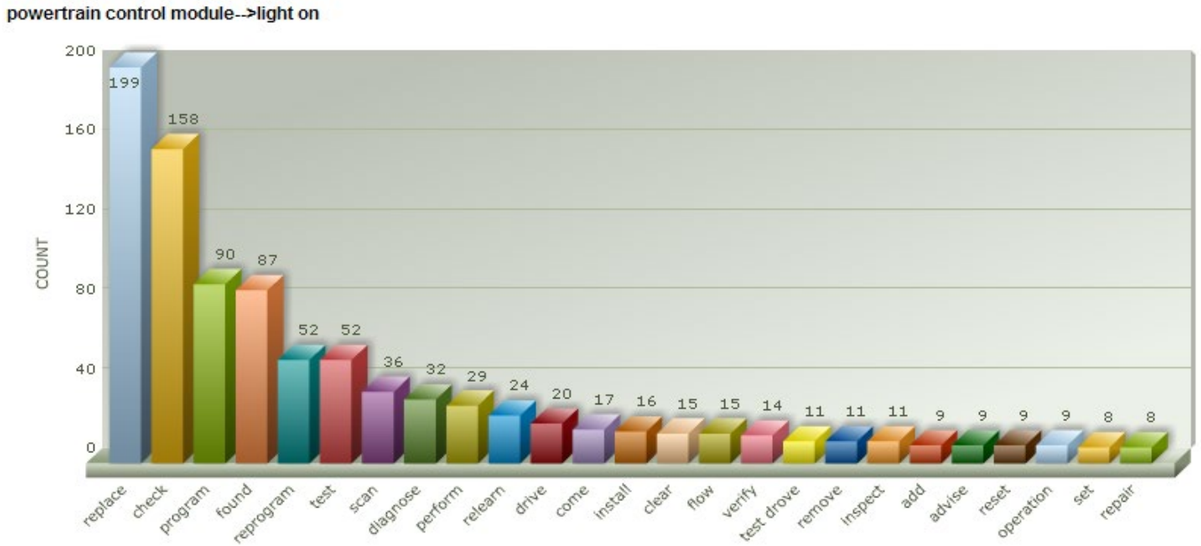
F1: Frequent Failure → Symptoms → Best Practice Repairs



Drill down analysis for knowledge discovery







- ❑ Suppliers
- ❑ Manufacturing plants
- ❑ Platform level analysis

Automatic development of fault dependency matrix

1. DFMEA/ Engineering Data

Components, Symptoms, Failure modes

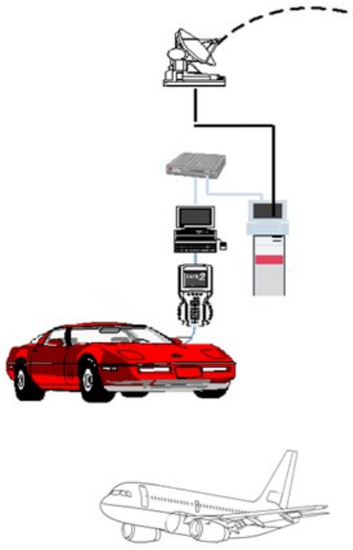
OEM's Database

Components, Symptoms, Failure modes

Context Vector, Syntactic Vector, Semantic Vector

Semantic Similarity Model

5. Enhance DFME/Engineering data with newly discovered knowledge



3. Data is collected via diagnostic tools

2. Field Diagnosis is performed

4. Develop text-based Fault D-Matrix

Filaure Source: 9, Test: 9	P0451	P0446	P0452	P0454	Loss Of Power	Engine Stalled	Engine Will Not Start	Unusual Guage Reading	Engine Runs Rough
FTP Sensor_Circuit Short And Pulling Down	1	0	0	0	1	0	0	0	0
FTP Sensor_Circuit Short To Ground	1	1	1	1	1	1	1	0	0
FTP Sensor_Faulty	1	1	0	0	1	1	1	1	0
FTP Sensor_Out Of Range	1	0	0	0	0	0	0	0	0
FTP Sensor_Open Circuit	1	1	0	0	0	0	0	0	0
FTP Sensor_Short Circuit in 5V Ref	0	0	0	0	1	0	0	0	0
FTP Sensor_Crack	1	0	0	0	0	0	0	0	0
FTP Sensor_Moisture	1	0	0	0	0	0	0	0	1
FTP Sensor_Stuck	1	0	0	0	0	0	0	0	0

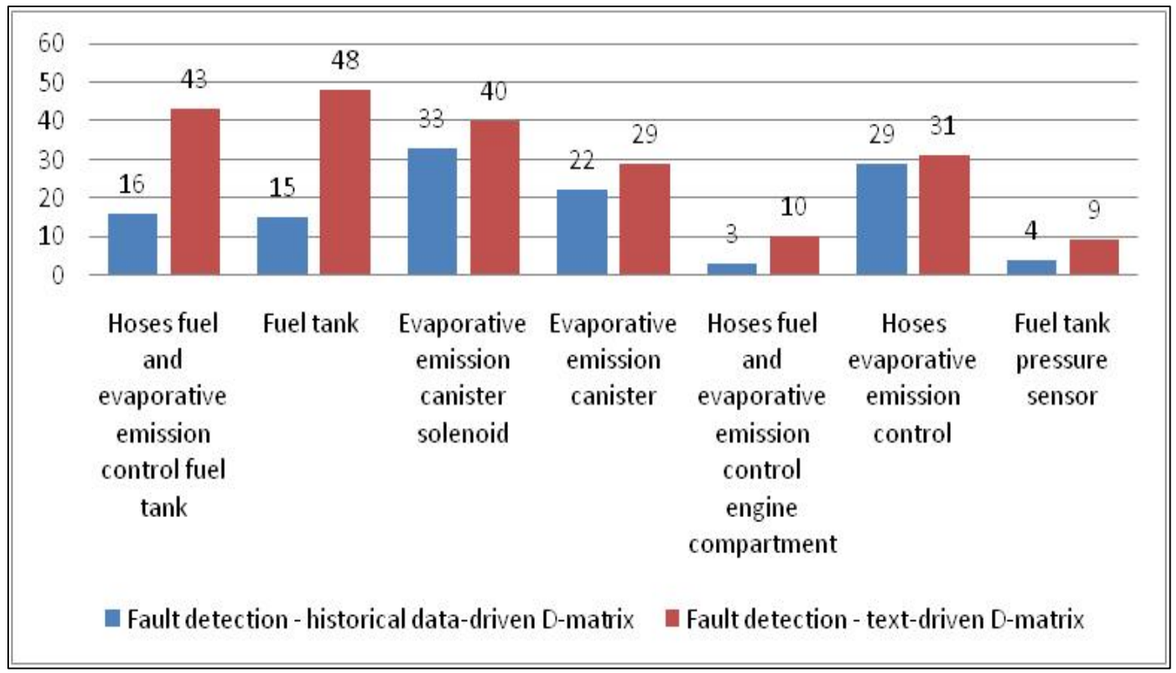
Results

$$P_D = \frac{[\sum_{i=1}^{m} D_i]}{m}$$

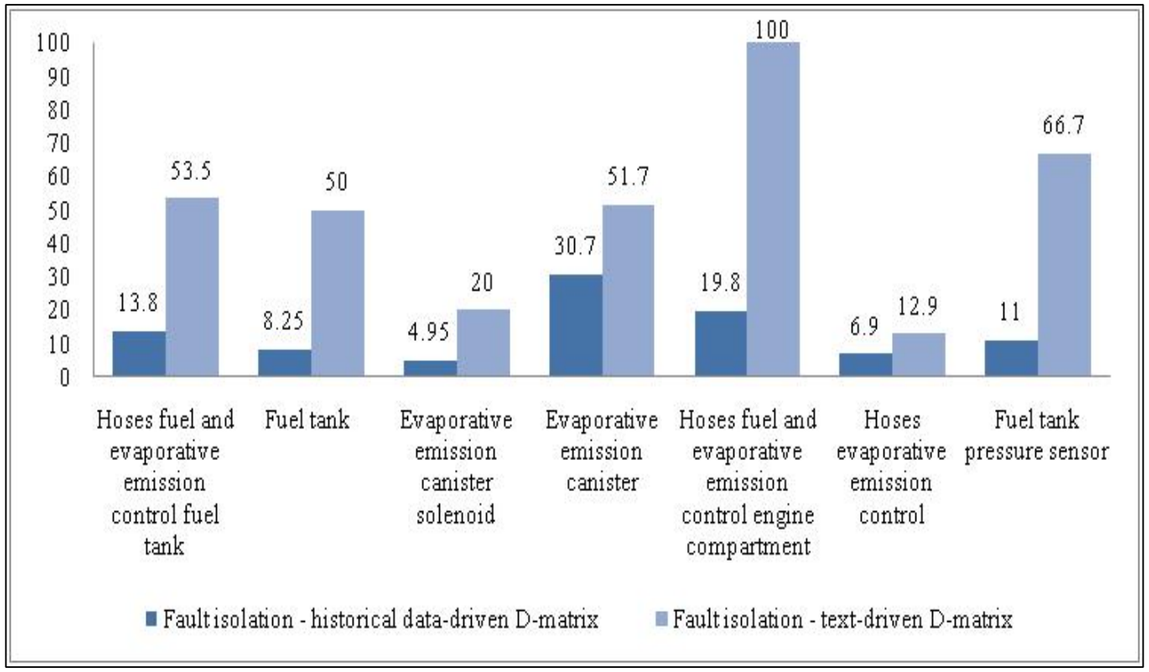
The ***fault detection*** (P_D) is defined as the percent of faults detected by the symptoms by observing the failure modes associated with a system.

$$F_I = \frac{[\sum_{i=1}^{m} ISO_i]}{m}$$

The ***fault isolation*** (F_I) is the probability that the symptoms uniquely isolate the faults of a system for the failure modes associated with a system.

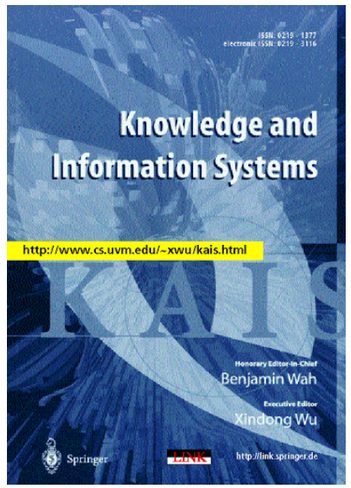


Comparison of fault detection between historical Data-driven D-matrix vs. Text-driven D-matrix

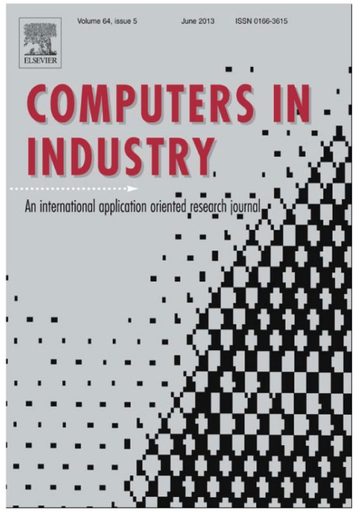


Comparison of % fault isolation between historical Data-driven D-matrix vs. Text-driven D-matrix

Publications

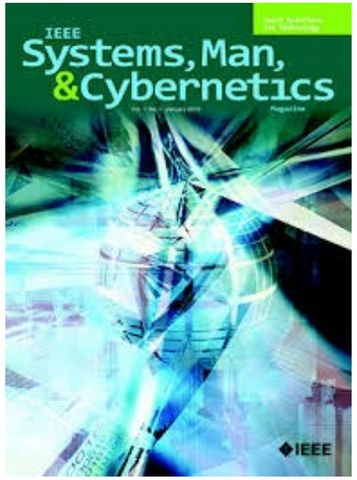


Rajpathak, D., Chougule, R., and Bandyopadhyay, P. (2012). A domain specific decision support system for knowledge discovery using association and text mining. Knowledge and Information Systems, vol. 31, pp. 405-432



Rajpathak, D. G. (2013). An Ontology-Based Text Mining System for Knowledge Discovery from the diagnosis data in the automotive domain. Computers in Industry, vol. 64 (5), pp. 565-580

Publications



Rajpathak, D. G. and Singh, S. (2014). An Ontology-Based Text Mining Method to Develop D-Matrix From Unstructured Text. IEEE Transactions on Systems, Man, and Cybernetics, vol. 44 (7), pp. 966-977

Rajpathak, D., Xu, Y., and Gibbs, I. (2020). An integrated framework for automatic ontology learning for unstructured repair text data for effective fault detection and isolation in automotive domain. Computers in Industry, vol. 123, pp. 103338

Thank you