Rapid Exploitation of Human Language in a Low-Resource Environment

A case study in low data high accuracy prediction

Alex K. Rojas

Overview

Problem Statement

Intro to Advance in AI Algorithms

Testing and Use Case

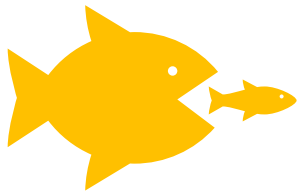
- -Use Case
- -Data Sources
- -Solution Architecture
- -Methodology Used
- -Algorithm
- -Deliverables

Findings

Problem Statement

No Data, No AI

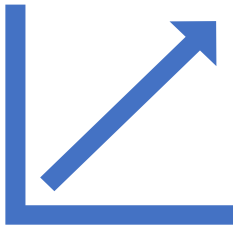
Bad Data, Bad AI



Little Data, Little Chance

Problem Statement

YES, Data quality, consistency and integrity remain a **common challenge** across Artificial Intelligence use cases (especially in Natural Language Processing)



YES, Most algorithms and state of the art machine learning methods **rely on large numbers of training examples to provide high accuracy percentages**

Is all lost? **NO!**

Intro to Advancements in AI Algorithms

HO-LRL™ (Higher-Order Low-Resource Learning™) is a data transformation for Machine Learning in low-resource settings – e.g., lack of sufficient training data in real-time

HO-LRL™ supports both generative and discriminative learning including **Natural Language Processing (NLP) with deep learning networks and latent embeddings**

HO-LRL™ as used in this case study is part of Intuidex's Watchman for Defense™ (W4D) Product for multi-INT fusion and alerting, including pattern of life and anomaly detection

HO-LRL™ Technology Profile

Work in the relational learning field has shown the value of leveraging associations in the form of latent relations in an entity-relation graph (Nelson, Keiler, & Pottenger, Modeling Microtext with Higher Order Learning, 2013) (Ganiz, Lytkin, & Pottenger, Leveraging Higher Order Dependencies between Features for Text Classification, 2009) (Ganiz, Pottenger, & Yang, 2006) (Li, et al., 2007) (Menon & Pottenger, 2009) (Nelson & Pottenger, Nuclear Detection Using Higher Order Learning, 2011) (Nelson, Pottenger, Keiler, & Grinberg, 2012) (US Patent No. 8,572,071, 2013) (USA Patent No. 8,301,768, 2012)

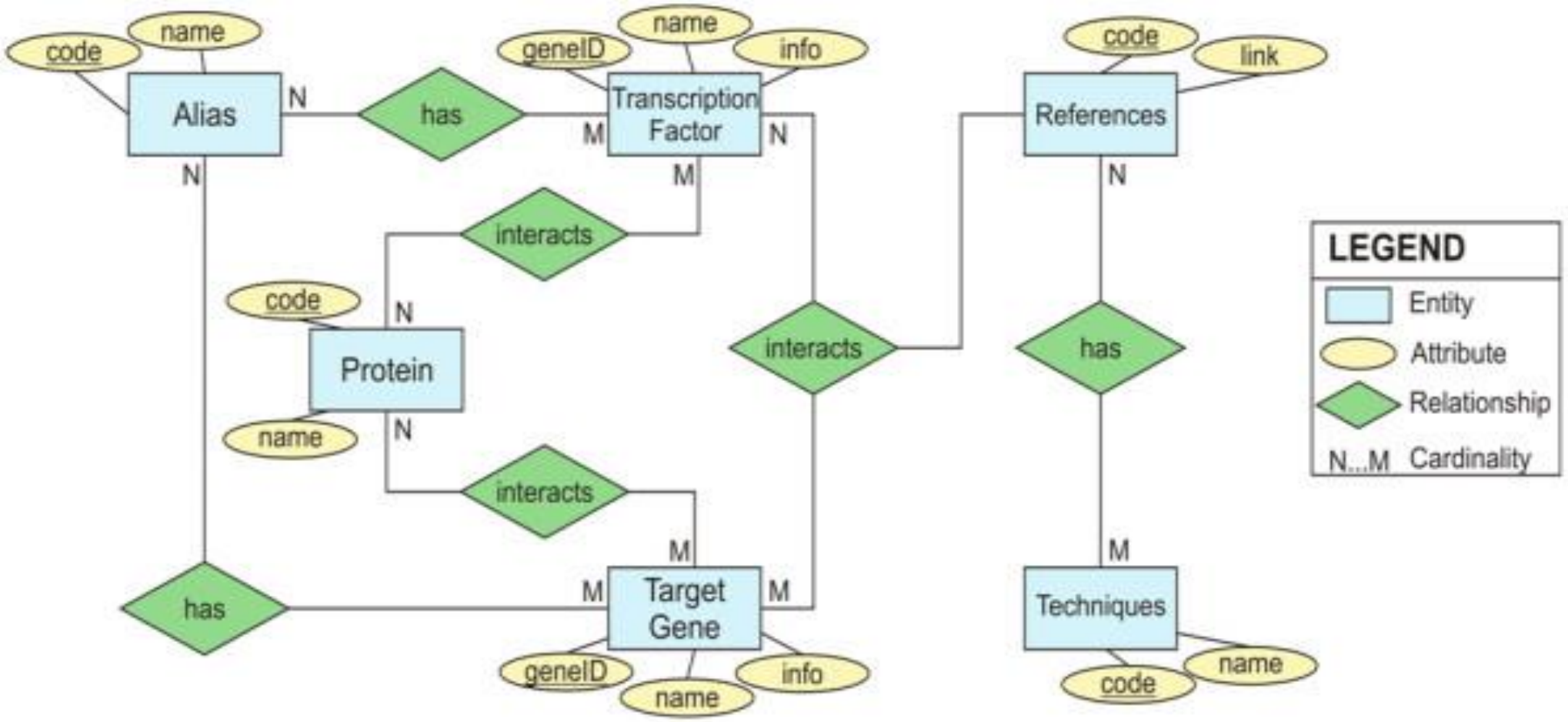
Dr. W. Pottenger ‘s research as the principal investigator and inventor of HO-LRL™ includes the discovery that Latent Semantic Indexing/Analysis (LSI/A) is dependent on the latent relations between entities in an entity-relation graph

HO-LRL™ supports data transformation in both generative and discriminative learning while significantly outperforming traditional approaches. Using HO-LRL™, models that generalize well can be built from very small amounts of data

Technology

High Order Low Resource Learning (HO-LRL™)

A general multi-attribute entity-relation graph has nodes that are people, places and things, etc.



Benefits of HO-LRL™ Technology

Wide Application of Uses: **Leveraging HO-LRL™ improves the performance of several algorithms and applications including threat detection in streaming message traffic, anomaly and threat detection in network traffic, classification of radio-nuclear signals, deep-learning based information extraction and video captioning, as well as applications in ecommerce, law enforcement and counterterrorism**

Smaller Training Times: Latent embedding-based architecture allows for substantially smaller training times. In one experiment with deep-learning information extraction input embeddings, HO-LRL™ training time was 1/10th of leading competitor BERT and 1/20th of RoBERTa.

Less Training Data Required: In the aforementioned experiment, training data of only 100K tokens of Publicly Available Information (PAI) was used to train the HO-LRL™ embeddings.

Performance: **In the aforementioned experiments, HO-LRL™ achieved 13 points greater F-beta, the harmonic mean of precision and recall, in the information extraction task from PAI**

Testing and Use Case – Problem Statement

Modeling from a previous predictive analytics projects Serco performed for our customers, Serco and Intuidex partnered to develop a use case to better address **low resource** environments

The purpose of this analytics project was to test on time delivery of military assets based on uses cases modeled on our knowledge of maintenance availabilities.

Testing and Use Case – Problem Statement (Continued)

In the original military asset delivery use cases, AI/machine learning products were to learn from condition and maintenance planning data in order to predict schedule overruns and the factors responsible for their predicted overrun. Currently these are jobs performed by teams of people within the military who produce Availability Duration Scorecard predictions .

The overall conclusion at the end of the Serco Projects was that the baseline prediction solution provided prediction results similar or better to the current human expert-driven ADS approach (14% versus 12.45 % mean absolute percent error) and **it could get better**.

Table 1: Original Data Sources

Data Source Name	Raw Format as Delivered	Pertinent Years Received	Size after Delivered	Number of Entries	Number of Fields
NMD: **Navy Maintenance Data** **February Extracts** **May Extracts**	Multiple Structured Text Files	 2007-2016 2014-2016	 2.3GB 1.6GB	 11K-91K 3K-275K	15-159
OARS: **Open Architectural Retrieval System**	Multiple Structured Text Files	2007-2016	3GB	-2.9M	118
NDE: **Navy Data Environment**	Single table .xlsx	2007-2016	1MB	6K	21

Table 2: Additional Data Sources

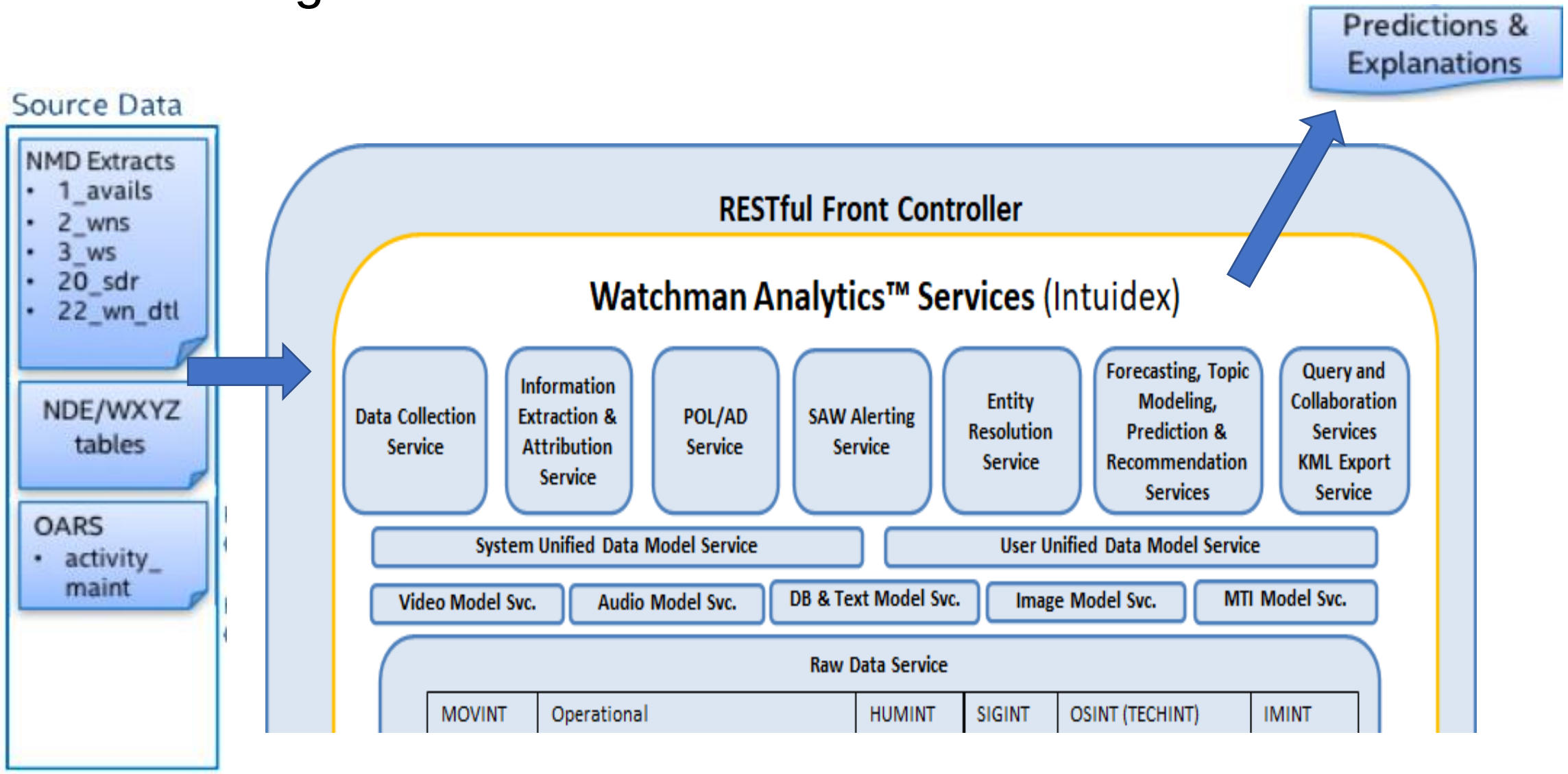
Data Source Name	Raw Format as Delivered	Years	Size after Delivered	Number of Entries	Number of Fields
RMC Weekly Report	1308 semi-structured .doc, .docx, .pdf files	2014-2016	230MB	680	18
PMR Weekly Report	1555 semi-structured .doc, .docx	2011-2016	341MB	1546	24
Letter of Authorization	1208 .pdf with structured enclosure	2007-2016	151MB	58993	21

Testing and Use Case - Sample Data Sources

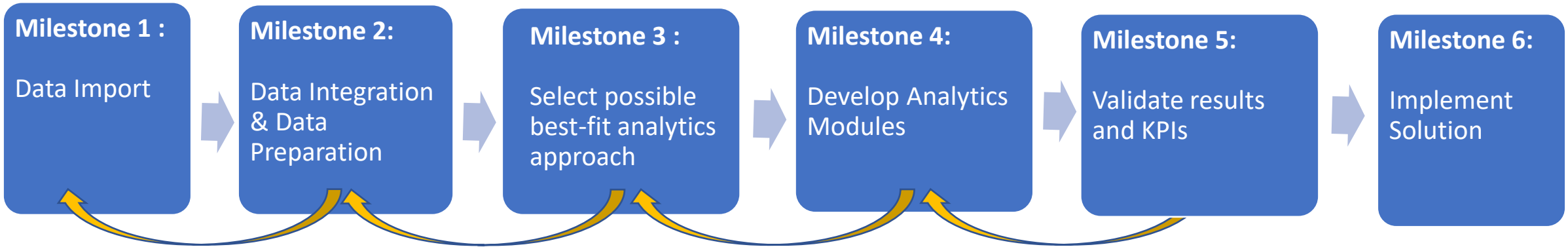
- Large variety of data sources
- Sparse, conflicting and incomplete data
- Unique identifiers inconsistent
- Descriptions, dates, and types of avail inconsistent
- Missing data, 20% of available
- Inconsistent spellings and numerical data, along with a framework and infrastructure not optimized for collection of such data

- Using the profile of these data sources along with the documented issues, a program was run to auto generate data similar in character along with specific data quality issues. The amount of data used was **0.2%** (6500) of the original data volume (3,158,219 records)

Testing and Use Case - Solution Architecture



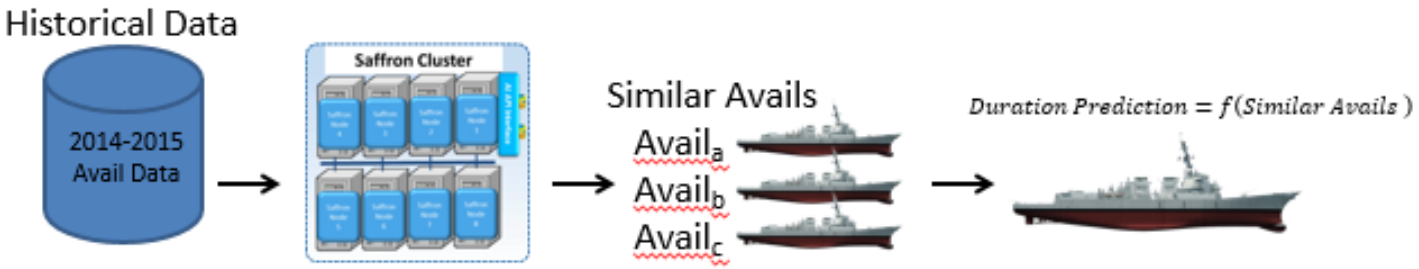
Testing and Use Case – Methodology Used



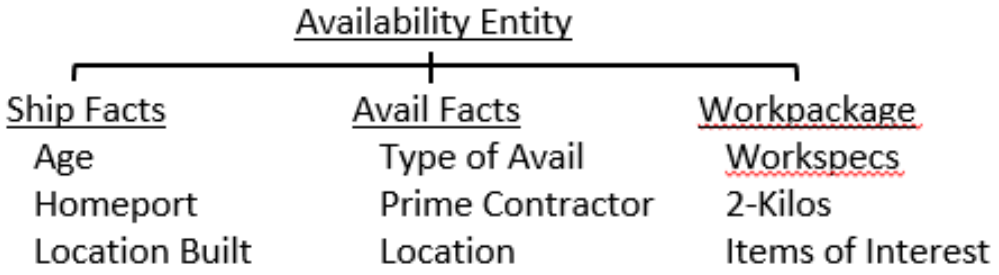
Testing and Use Case – Sample Original Algorithm

- **Base Algorithm was based on similarity from historical overruns.**
- Used an average weighted by similarity on the duration of the nearest neighbor availabilities found for the overrun prediction.
- Originally this was done for A-60 (60 days) and extended to A-720 (2 years)
- Prediction was better than human error rate but could be better!

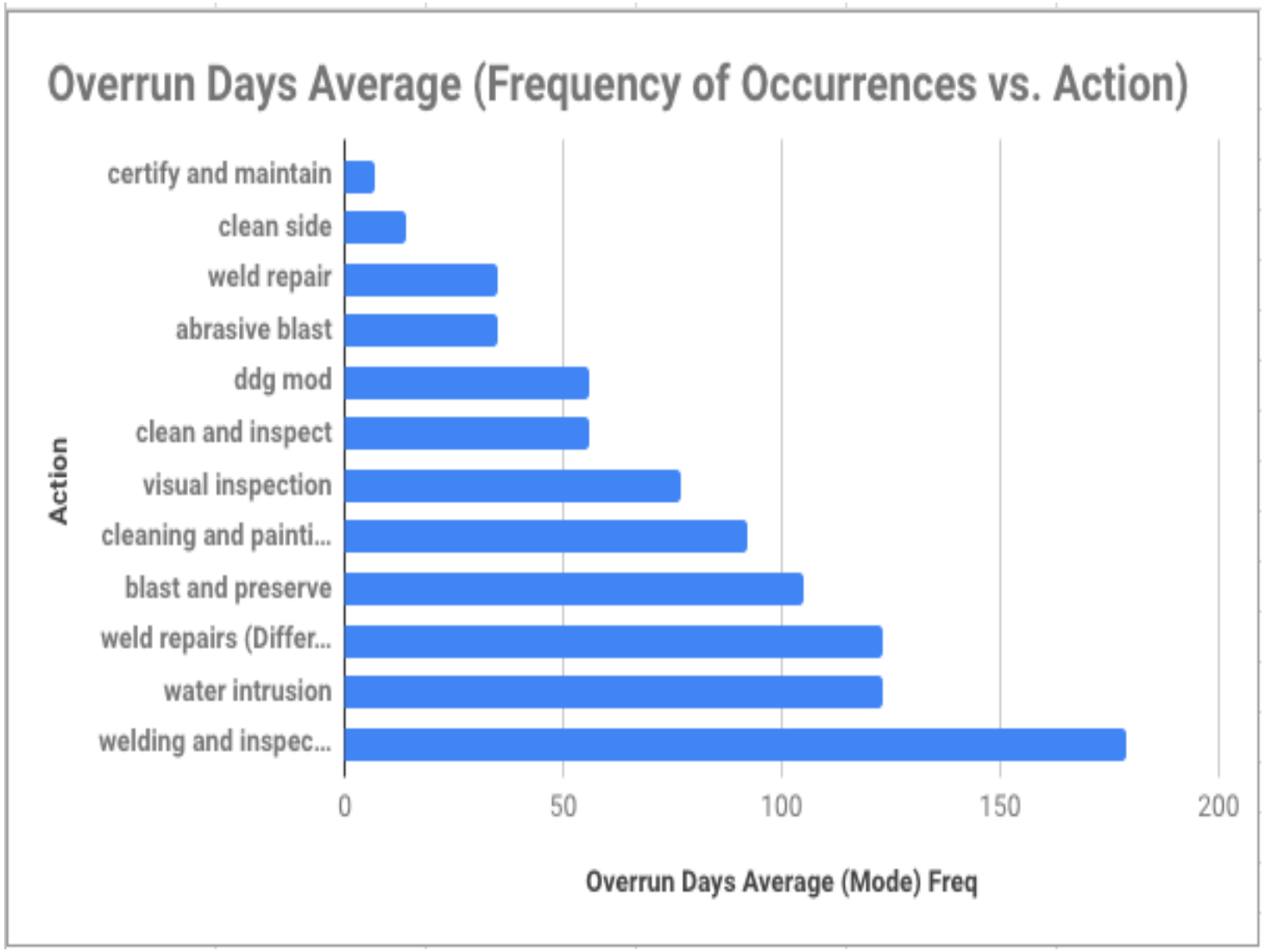
❑ Duration Predictions based on similar Avails



❑ Avail entity comprises three categories

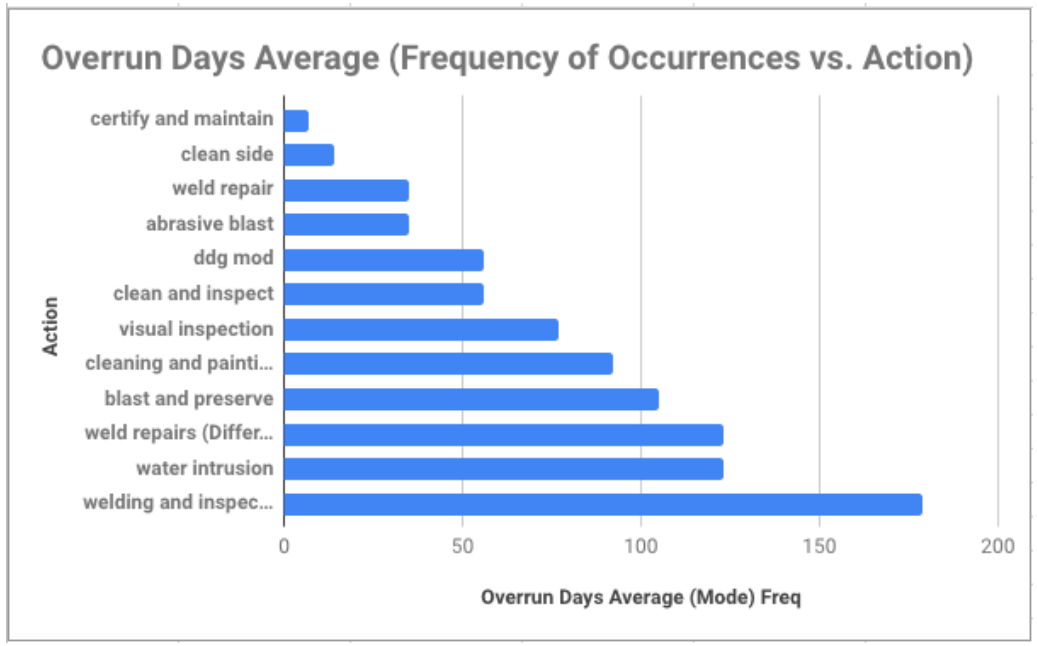


Multi-Modal Models and Data Fusion



HO-LRL™ can be applied by taking the NLP features that were previously identified as correlated to the overrun class that's being predicted, and compute a HO-LRL™ transform (which is part of Watchman for Defense™). Also allows to add other correlated features such as ship age, work location, etc.

Testing and Use Case - Algorithm Advance and Upgrade



- Incorporate NLP aspects to the solution based on project findings: **n-gram text correlations to target condition**
  - 8.2% of records had specific n-grams which had a 75% chance to cause the target condition; further subdivided to exact number of days for condition.
- Incorporate **Ship Age**, and **location of work** as a factor into the digital signature

- Incorporate HO-LRL™ technique to transform the data in a machine learning algorithm to focus on important discriminators for target conditions

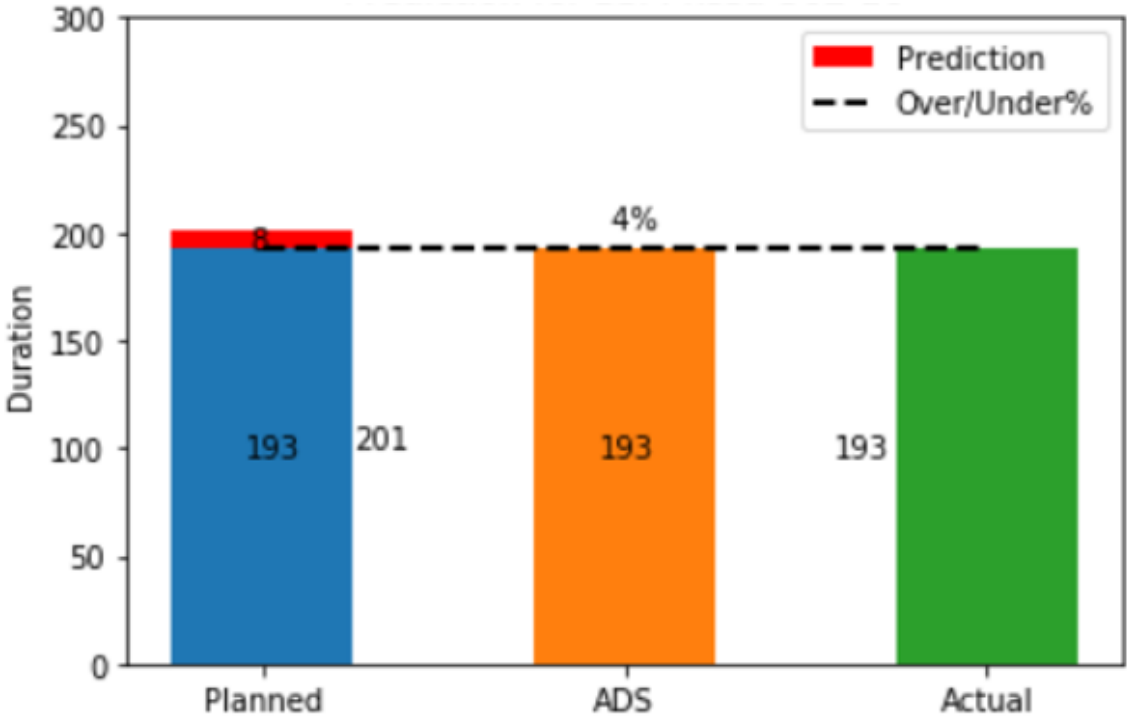
Testing and Use Case – Deliverables Samples

- Duration predictions for future availabilities
- Identification of ships most likely to have overruns
- Explanations of predictions
- Comparison between planned, human expert prediction, AI/ML prediction, and actual

Predicted End: 3/15/2020
Predicted Overrun: 75 days

Potential Drivers of Growth/New Work

Description	# Avails w/ Similar Work	Growth
MPDE Repairs	7 Avails w/ Similar Work	6 Had high growth
LMA Intake/Uptake Work	3 Avails w/ Similar work	3 had high growth



Findings

HO-LRL™ Benefit for Predictive Maintenance Analytics		
Metric	**HO-LRL™**	**Original Implementation**
$F_{\beta=0.25}$ / Accuracy	**>99%**	86%
Error / MAPE (Mean Absolute Percentage Error)	**<1%**	14%
Training Time	<20 seconds (avg. depends on data set size and resources)	1:300 hrs. ratio
Training Data	6,500 records (0.2% of original data)—Least Minimum: 180 samples, 30 samples overrun and 150 non-overrun samples	3,158,219 records
Time to Implement	2 weeks	Few months when data available

Conclusions

- PERFORMANCE:
  - Using HO-LRL™ for NLP yielded >**99% $F_{\beta=0.25}$ based on our test using only one column of text (Block 35) for the Avail Overrun Condition Prediction.**
  - **13% Increase** from original using an ensemble of methods and classifiers

- TRAINING TIME: **<20 seconds (avg. depends on data set size and resources)!**

- TRAINING DATA: **The amount of data used was 0.2% (6500) of the original data volume (3,158,219 records). The least minimum amount of data required was 180 samples. 30 with the overrun condition and 150 with non-overrun. HO-LRL beat standard SVM by several points with >99% confidence (F-beta)!!!**

Conclusions (Continued)

- For many areas/applications data availability is limited. Vocabulary different from generic vocabulary such as that of Wikipedia. Generating a model with high quality in such a low-resource setting is highly desirable in these applications.

- Work based on a suite of technologies that facilitates learning models on both dynamic and static data by transforming data representations into a “higher order” format based on latent relationships between data items.

- In related previous work, it was shown that leveraging higher order dependencies improves the performance of several different algorithms and applications including threat detection in streaming message traffic, anomaly and threat detection in network traffic, classification of radio nuclear signals in border protection, as well as applications in ecommerce, law enforcement and counterterrorism

- **Take-away**: What is remarkable is that this represents a >99.99999% reduction in training data, subsequent training time, with increased accuracy!

Q&A

Thank You!