Machine Learning Assisted Data Extraction and Normalization

Dr. Jonathan Brown

Mr. Devin Geraghty

Naval Surface Warfare Center, Dahlgren Division

Presented at:

2019 ICEAA Workshop Tampa, FL

14 May 2019

The Leader in Warfare Systems Development and Integration

NAVSEA WARFARE CENTERS DAHLGREN

NAVAL SURFACE WARFARE CENTER
DAHLGREN DIVISION
DAHLGREN | DAM NECK

100 anniversary 1918-2018 DAHLGREN A CENTURY OF INNOVATION

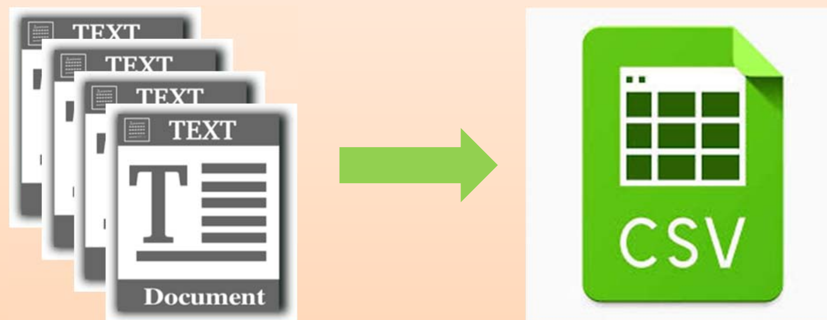
Agenda

- Problem:35,000 records to label
- Solution: Machine Learning
- Machine Learning Types
- Machine Learning Experiment
  - Process
  - Define Labels and Features
  - Process Data
  - Results Summary
- Next Steps

Problem

- ❑ Amount of data available for analysis has increased dramatically in recent years
- ❑ Time-prohibitive to clean, normalize, analyze larger datasets using traditional methods
  - ❑ Forced to analyze only part of the data
  - ❑ Label useless because too hard to unwire
- ❑ Alternative methods are required to more quickly process data for use

Specific Data Example



Task planning sheets capture information primarily used in planning, documenting, and communicating between government program offices and executing agencies:

- Task Title
- Task Descriptions
- Deliverables List
- Funding received and executed
- Responsible organizations, etc.

By the numbers	
Number of records	35,044
Number of entries per record	38
Types	Strings, Floats

How can we efficiently clean, normalize, and map this data into a usable form?

Solution: Machine Learning?

What is machine learning?

- ❑ Definition: A method of data analysis, using algorithms, where systems learn on their own
  - – Application examples: filter email spam, refine search engine results, traffic predictions, fraud detection, object recognition, text classification
- ❑ Alternate definition: the science of getting computers to act without being explicitly programmed (Coursera)

Why use machine learning?

- ❑ Manually mapping ~35,000 lines is time-intensive
  - – Instead, pass a few examples to a machine learning algorithm and get a mapping in less time
- ❑ Manually reviewing and formatting text is time-intensive
  - – Instead, use tools like Python™* to handle large amounts of data (i.e., normalizing)

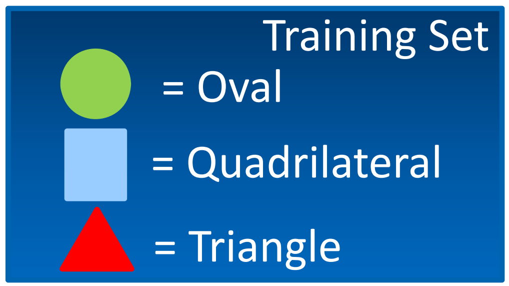
"Supervised" and "Unsupervised" are the primary methods of machine learning

* "Python™" is a trademark of Python Software Foundation

Machine learning replicates human learning but can handle much larger amounts of data more quickly

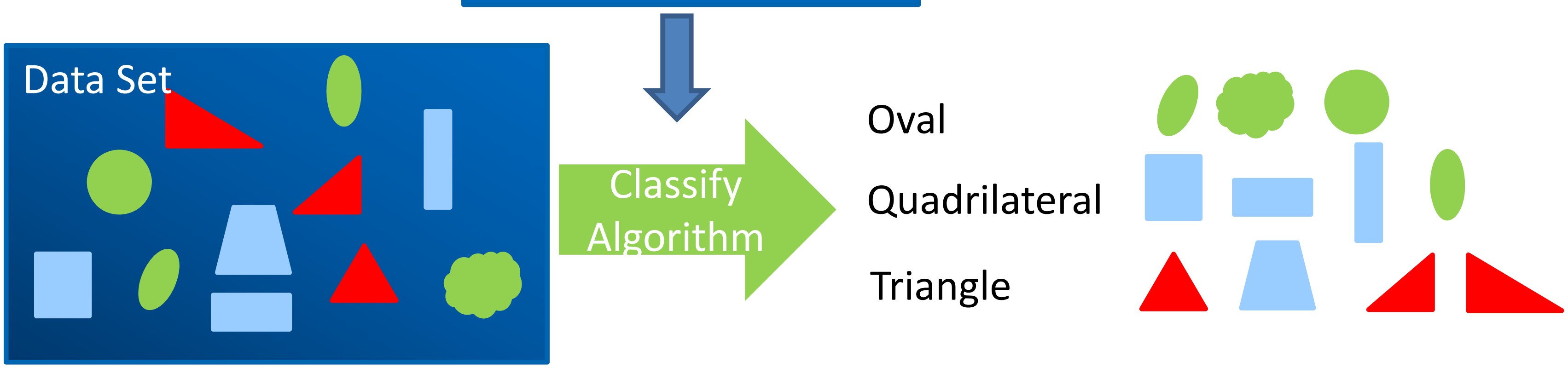
Machine Learning Algorithm Types: Supervised

- **In supervised learning, the data scientist acts as a guide for the machine learning algorithm by providing examples**, using a training set of data, where the correct answers are known and labeled.



Common Supervised Algorithm Types:

- Classification
- Regression



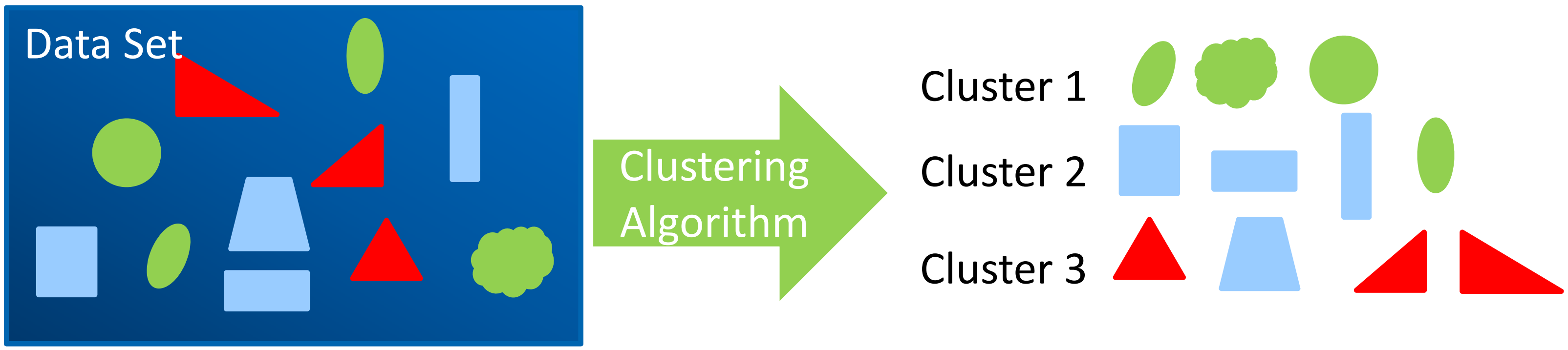
Learn with guidance from the data scientist

Machine Learning Algorithm Types: Unsupervised

- Unsupervised learning is closer to "true" artificial intelligence methods. **In unsupervised learning, the computer learns without guidance from the data scientist.** These methods are usually more complex but can tackle questions humans cannot or when the correct answer is unknown. Unsupervised learning identifies structure in data.

Common Unsupervised Algorithm Types:

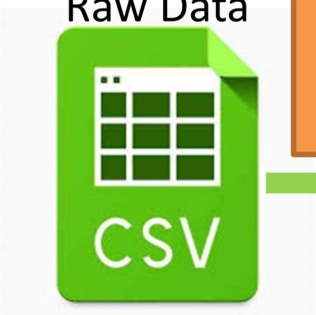
– Clustering



Learn without guidance from the data scientist

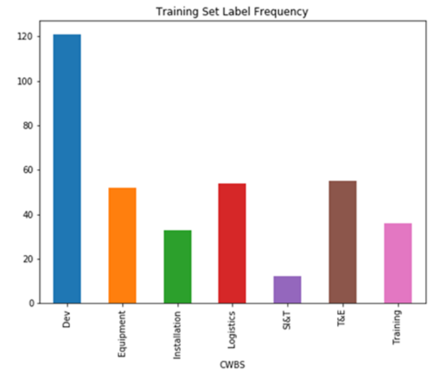
Experiment #1 Supervised Learning Process

Raw Data

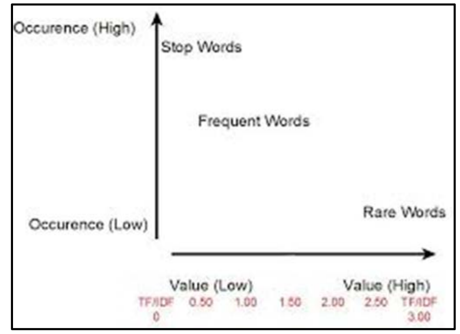


Define Labels and Features

Create Training and Test Sets

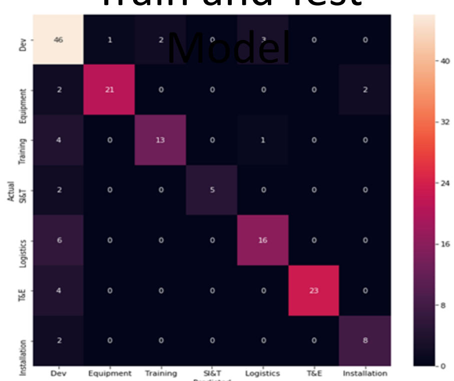


Process Data

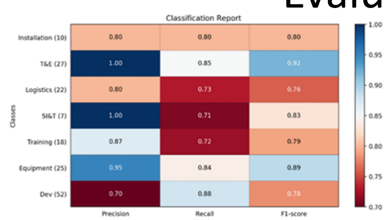


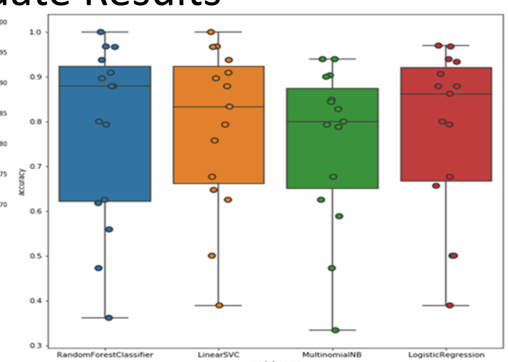
Text Classification: Random Forest, Feature: Task Title

Train and Test Model



Evaluate Results





Use the Random Forest Classifier to map dataset into predefined categories

Experiment #1 Define Labels and Features

Features → ML Algorithm → Labels

- Labels Selected
  - Align with cost work breakdown structure (CWBS)

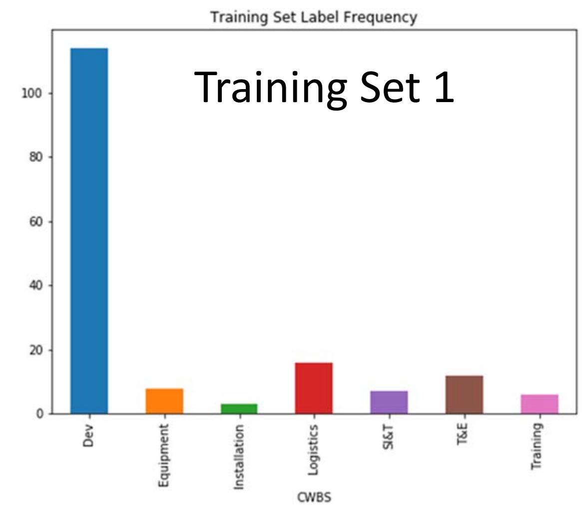
Development
Equipment
Installation
Logistics
Ship Integration and Test (I&T)
Test and Evaluation (T&E)
Training

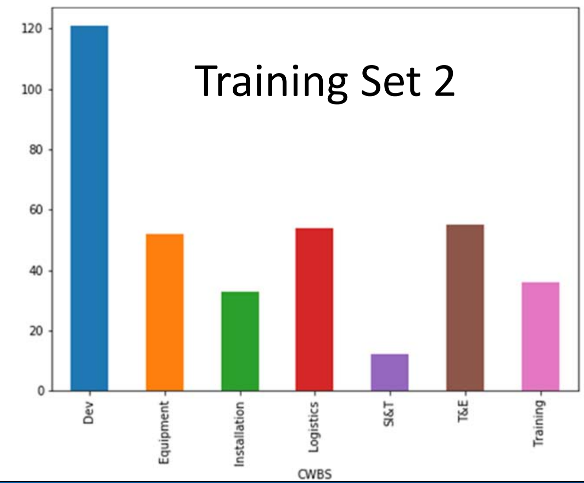
- Feature Selected
  - Task title

Task Title	Performer	Task Description	Effort	Task Summary

Experiment #1 Create Training and Test Sets

- Manually Labeled Example Sets
  - Example Set 1, 232 records
  - Example Set 2, 485 records
- Training Sets
  - 67% of the Example Sets used as Training Set
  - Training Set 1 & 2, 155, 325 records
- Test Sets
  - 33% of the Example Sets used as Training Set
  - Test Set 1 & 2, 76, 163 records





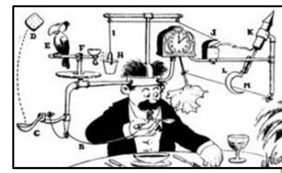
Initial results from Training Set 1 poor, Major improvements with Training Set 2

Experiment #1 Process the Data

Normalization Steps Remove:	Example Text
Remove duplicates	Install 2 radar systems on a mast and perform testing on the interfaces.
Rows without data	Install 2 radar systems on a mast and perform testing on the interfaces.
Words with < 3 characters	Install 2 radar systems mast and perform testing the interfaces.
Punctuation and standalone numbers	Install radar systems mast and perform testing the interfaces
English function words	Install radar systems mast perform testing interfaces
Lemmatization	Install radar system mast perform test interface
Convert to numeric (tf-idf)*	1.45, 0.51, 0.65, 0.08, 1.42, 1.61, 0.34 (example only)

*Term Frequency-Inverse Document Frequency

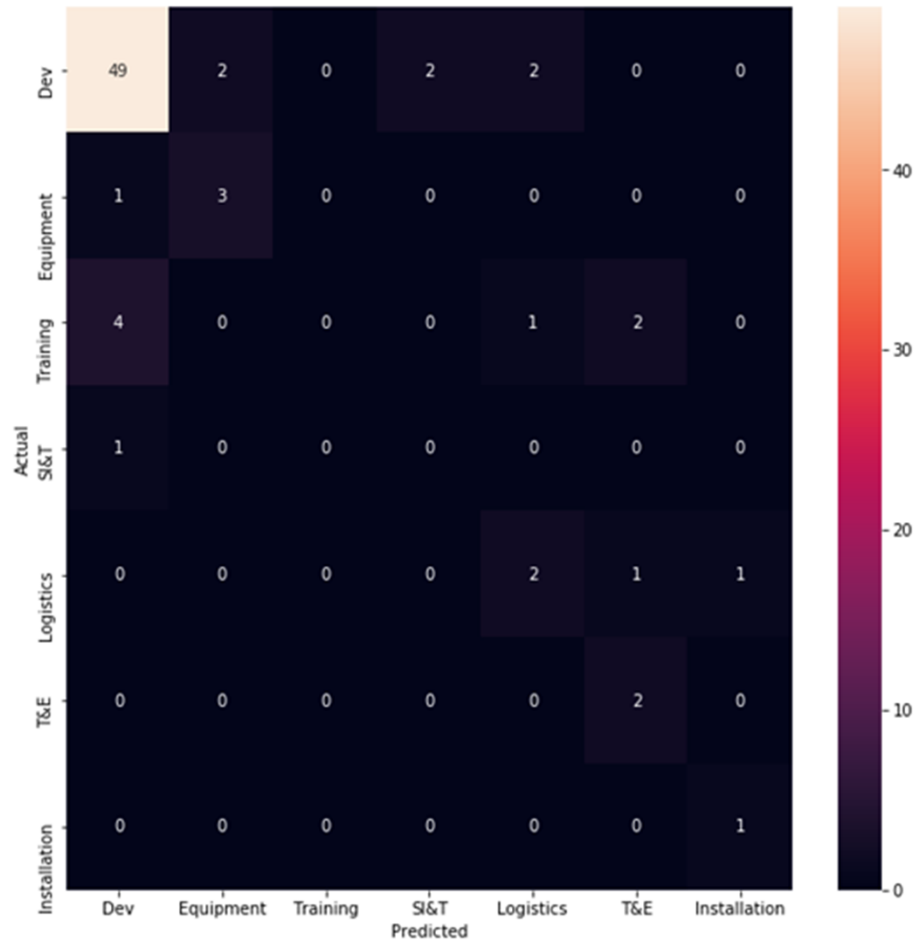
Used Python to convert text to numeric values for analysis



Label	Highest tf-idf
Dev	1) radar
	2) isea
Equipment	1) eqpt
	2) equipment
Installation	1) combatsystem
	2) install
Logistics	1) integrated
	2) logistics
SI&T	1) leadership
	2) test
T&E	1) evaluation
	2) interoperability
Training	1) engineering/ils/training
	2) training

Experiment #1 Results Summary Test Set 1

- Initial model was trained using Training Set 1 and tested using Test Set 1
- All algorithms performed poorly
  - Accuracy of model was 50% or less for all but Dev and Equipment
  - Small numbers of other categories in test set
- Hypothesized training set distribution was cause

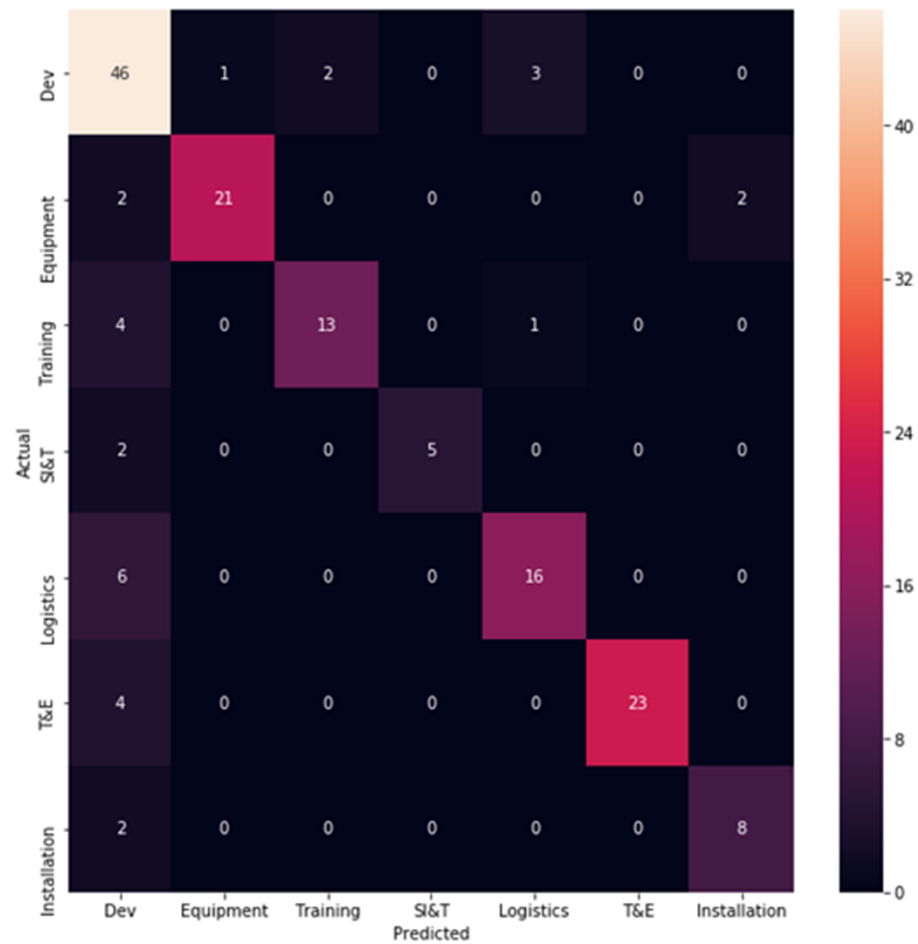


Initial training set unbalanced distribution (development heavy) negatively impacted algorithm performance

Distribution A: Approved for Public Release. Distribution Unlimited

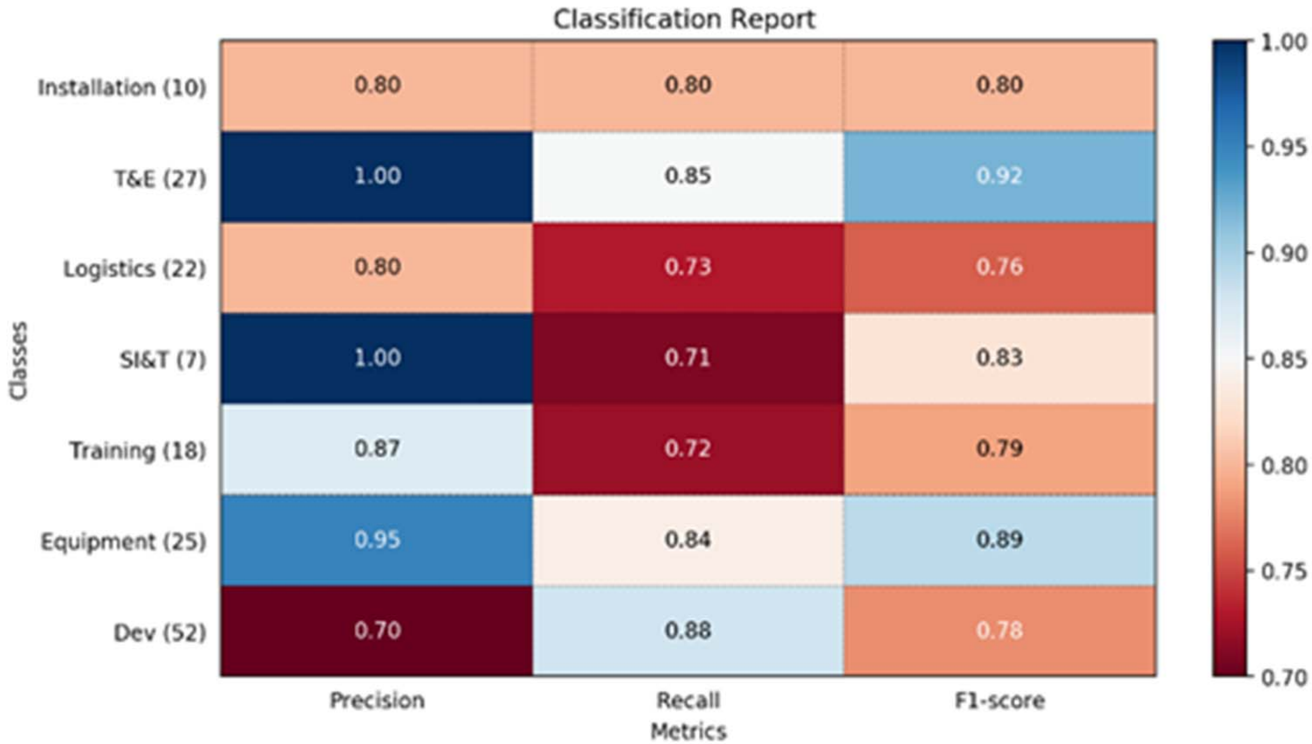
Experiment #1 Results Summary Test Set 2

- Updated model was trained using Training Set 2 and tested using Test Set 2
- Improved all algorithm performance
  - Diagonal heat map visually shows improvement
  - Still some over prediction for the development category



Second balanced training set performed significantly better

Experiment #1 Results Summary



<u>Three primary metrics</u>

- ❑ Precision = Percentage of predicted positives that are actually correct
- ❑ Recall = Percentage of actual positive that are predicted correctly
- ❑ F1-score = Average of the two

Machine learning could accurately classify the labels with an F1 0.76-0.92

Next Steps

- Reduce missed "easy wins"
  - Increase size of training set to refine algorithms
  - Improve initial data normalization
    - Bi and Tri grams
    - Alternate word divides "/"
- Use alternate features "Task Title" + "Task Description"
- Optimize code
- Apply developed algorithms to map entire data set
- Apply methodology to other data sets
  - Expand beyond task orders
  - Expand beyond cost data
  - Expand beyond text classification

Label	Highest tf-idf
Dev	1) radar
	2) isea
Equipment	1) eqpt
	2) equipment
Installation	1) combatsystem
	2) install
Logistics	1) integrated
	2) logistics
SI&T	1) leadership
	2) test
T&E	1) evaluation
	2) interoperability
Training	1) engineering/ils/training
	2) training

Example #	Task Title	Predicted Label	Actual Label
1	Development Engineering/Training Support	Dev	Training
2	Leadership	Dev	SI&T
3	Common Acq Logistics	Dev	Logistics
4	Tech Refresh Procure/Install Support	Dev	Installation
5	Training SME Support	Training	Dev

Machine learning can be applied to cost normalization problems

NAVSEA WARFARE CENTERS DAHLGREN

Thank You

anniversary
1918-2018
DAHLGREN
A CENTURY OF INNOVATION