Comparison of Machine Learning Algorithms to Build a Predictive Model for Classification of Survey Write-in Responses

Dr. Andrea Roberson
Justin Nguyen

Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Annual Capital Expenditures Survey (ACES) Research-Overview

- Provides national-level estimates of annual capital investment in new and used buildings, structures, machinery, and equipment by U.S. non-farm businesses

ITEM 2 **CAPITAL EXPENDITURES**

Report the following domestic capital expenditures data for the entire company.
(Refer to page 4 of Instructions)

Example: if figure is $1,179,125,628.00 report → Bil. 1 | Mil. 1 7 9 | Thou. 1 2 6

Row	CAPITAL EXPENDITURES *(Refer to Page 2 of Instructions)*	Structures (1)			Equipment (2)			Other *(Describe in Item 3)* (3)			Total *(Add columns 1+2+3)* (4)		
		Bil.	Mil.	Thou.	Bil.	Mil.	Thou.	Bil.	Mil.	Thou.	Bil.	Mil.	Thou.
20	Capital expenditures for NEW structures and equipment ***(Include major additions, alterations, and capitalized repairs to existing structures)***												
21	Capital expenditures for USED structures and equipment												
22	**TOTAL capital expenditures** *(Add Rows 20 + 21)*												

Total should equal Item 1A, Row 11

ITEM 3 **List the items included in "Other."** Report in thousands of dollars. **Furniture and fixtures, computers, capitalized computer software, and motor vehicles** should be reported as equipment. **Leasehold improvements** should be considered new structures or new equipment based on what is being improved.

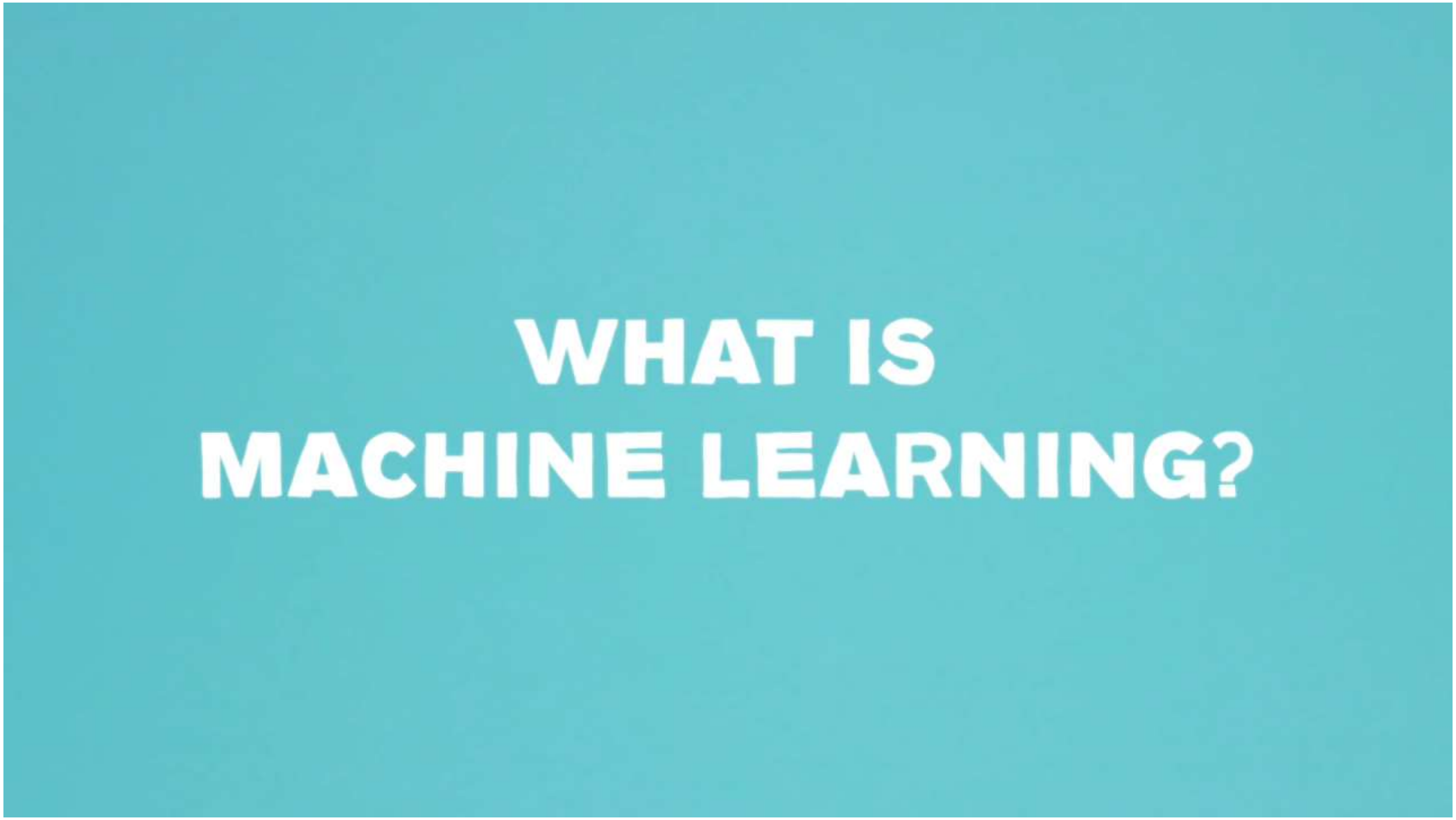
Row	(1) Description of Capital Expenditures	(2) Bil.	Mil.	Thou.
30	lending **money** to **commercial** banks (fabricated response)			

Source: https://www.census.gov/programs-surveys/aces.html

Purpose of the Research

- Prior U.S. Census Bureau studies identified areas of improvement in our editing processes in order to improve the timeliness and quality of our estimates while reducing cost
- A U.S. Census Bureau Economic Edit Reduction team identified edits and processes that can be automated
- Suggestions included automating the manual examination of ACES survey write-ins
- The use of Machine Learning (ML) classifiers was recommended to successfully predict the correct class of capital expenditures

What is Machine Learning (ML)



Source: https://www.oxfordsparks.ox.ac.uk/

Modernization of Statistical Production

- National Statistics Offices should all explore the use of ML (Chu and Poirier, 2015)

- Applications
  - Decision Trees (Portugal): Detection of errors in foreign trade transaction data.
    - Reduced manual examination of records

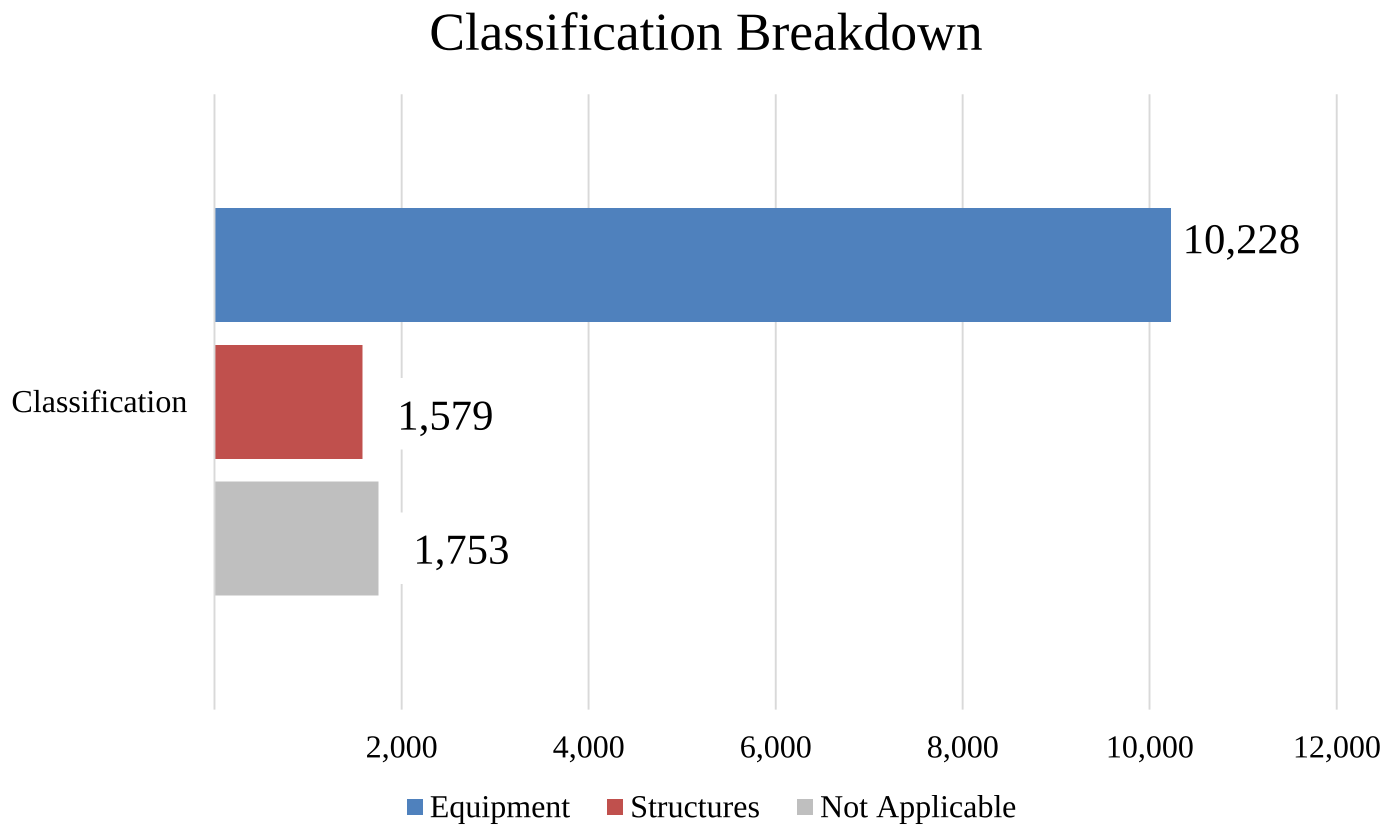
ML Techniques for Write-In Classification

- Logistic regression [statistics]
  - Training data: Binary response (0:1) and predictors
  - Maximum likelihood leads to model parameters
  - Resulting model is used to predict responses

- Support Vector Machines [non-statistics]
  - Training data: Binary response (0:1) and predictors
  - Hyperplanes in the space of predictors separate responses
  - SVM optimization problem comes from geometry

Data

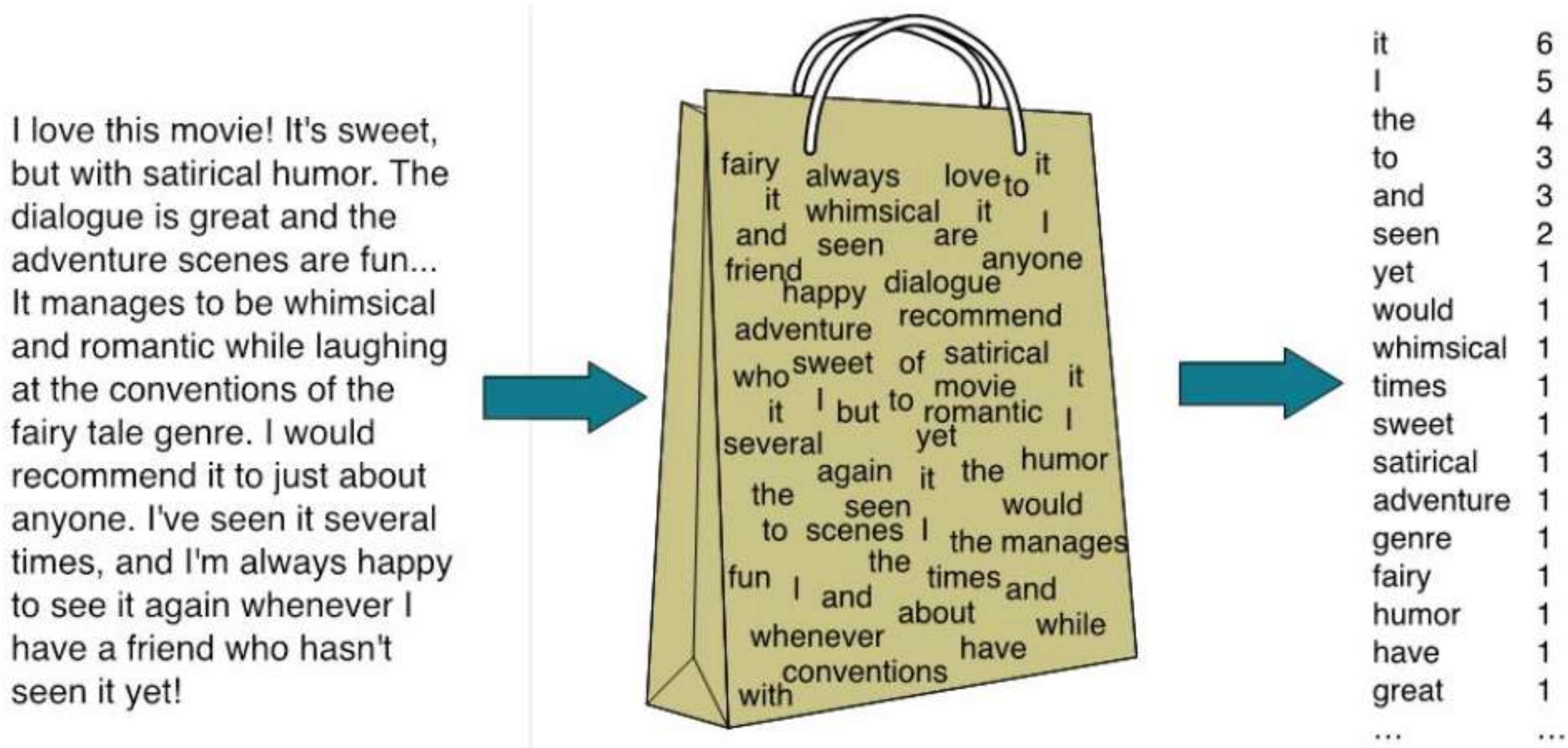
- **2015 and 2016 ACES Write-in Data**



Source: U.S. Census Bureau, 2015 and 2016 Annual Capital Expenditures Survey

Text Classification Overview

- **Bag of words model**



Source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

Term-Document Matrices

Write-in 1 **Class: Equipment**	The first **Bass** I remember catching was when I was six, my Dad had a performance of Strange Brew by **Cream** on the boat. Recollections indeed: an extremely fashionable **Cream**, with serious sideburns all round. I remember wondering why bassist Bruce was playing a **guitar**. Lost in my thoughts, it happened, there was a beautiful **Bass** hooked to my fishing rod.
Write-in 2 **Class: Equipment**	When weekend **fishermen** looking to unwind come to Rock Harbor, they take out their rods and reels and angle for striped **bass** one by one. When **commercial fishermen** like Mike Abdow go out on their boats to earn a living, they catch **bass** the same way. But right now, the waters are rough between people who fish **bass** for a living and those who angle for pleasure. Big-**money** sporting interests are trying to stop small-time **commercial fishermen** from pulling in any morestriped **bass**.
Write-in 3 **Class: Structures**	**Bank** Negara will change the way it calculates the cost of lending **money** to **commercial** banks and financial institutions, the central **bank** said in a release. In a statement, the **bank** said that as of Nov. 1, the base lending rate, at which **commercial** banks can borrow from the central **bank**, will become more responsive to movements in **money**-market rates. Several weeks ago, the **bank** said it would change the way it calculates the base rate.

Term-Document Example

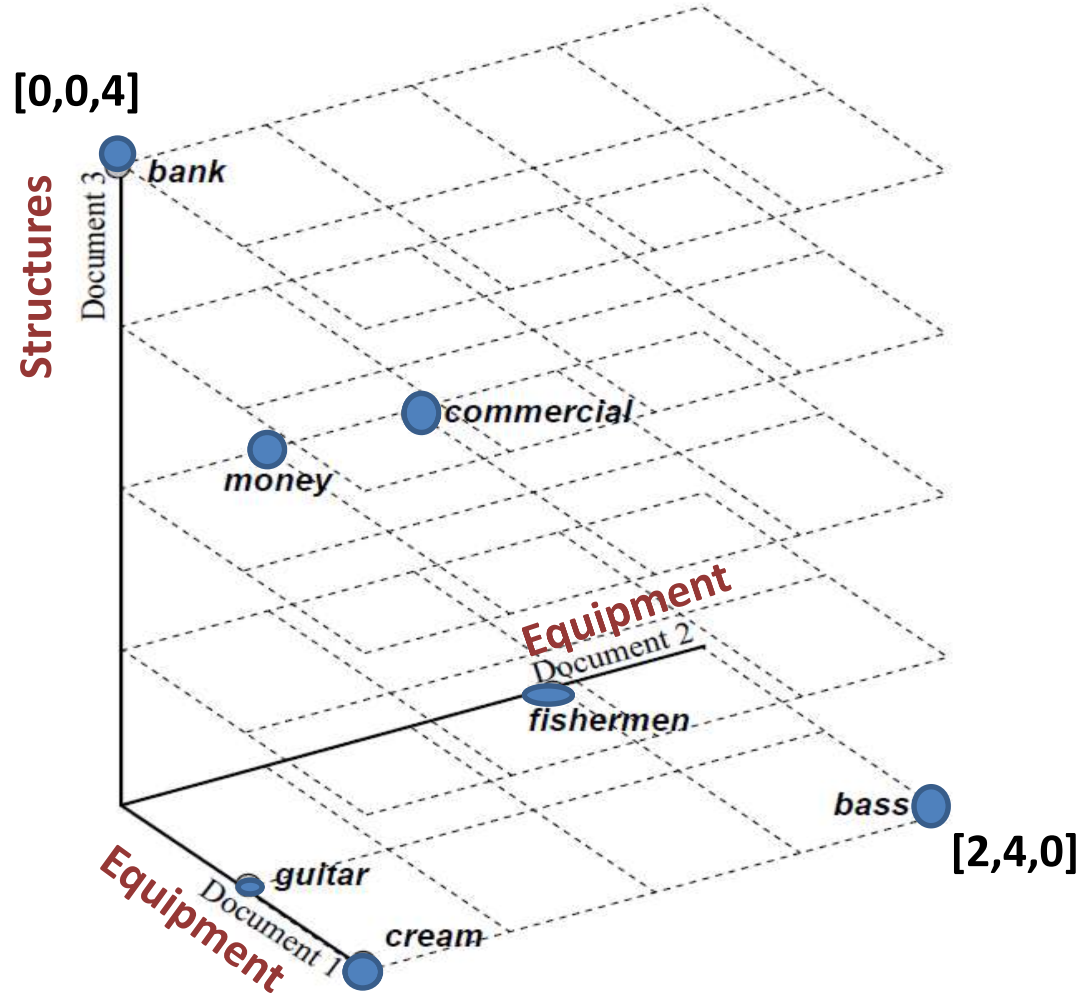
Vocabulary						
bank	bass	commercial	cream	guitar	fishermen	money

A Word Vector →

	Write-in 1	Write-in 2	Write-in 3
bank	0	0	4
bass	2	4	0
commercial	0	2	2
cream	2	0	0
guitar	1	0	0
fishermen	0	3	0
money	0	1	2
CLASS	Equipment	Equipment	Structures

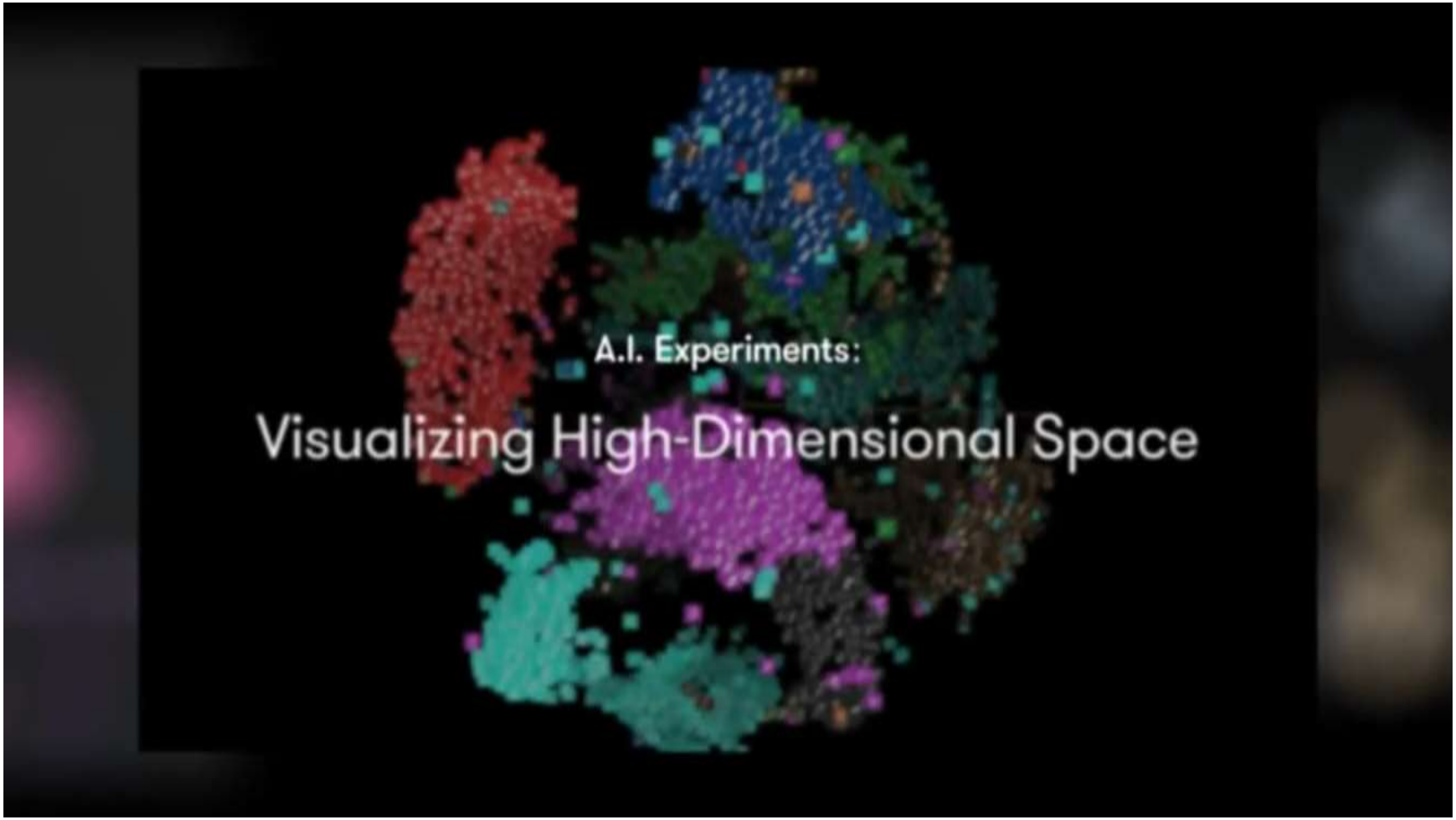
We count the number of times each word is used in each of our write-ins.

Documents in Term Space



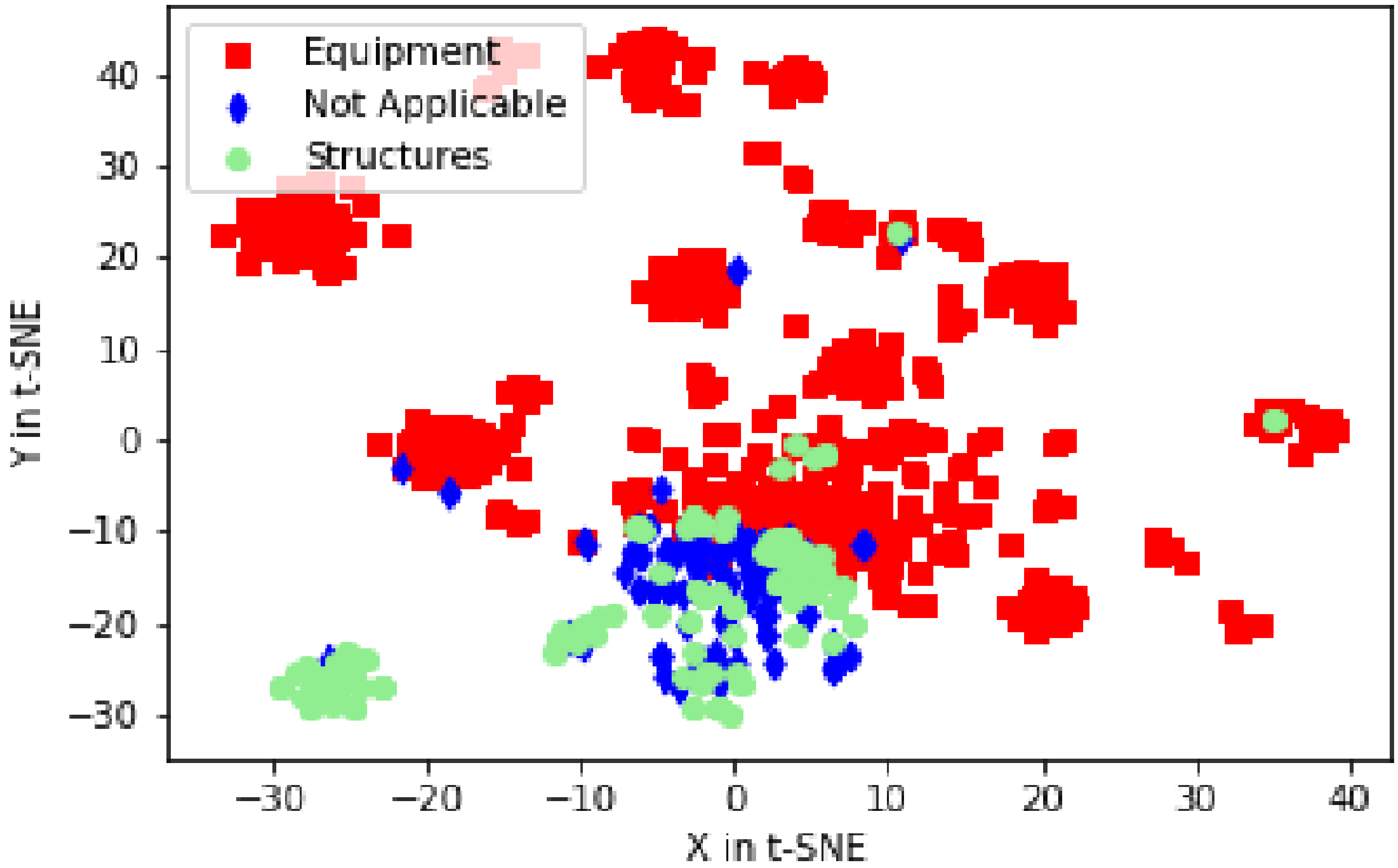
Visualizing the word vectors as points in three dimensional space

A.I. Experiments: Visualizing High Dimensional Space



Methodology
Prepare the Data

t-SNE visualization of test data



Logistic Regression (Fit the Model)

- Grid Search- LR

```
import sklearn
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import Pipeline
from sklearn.model_selection import GridSearchCV

pipeline = Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()),
                     ('clf', LogisticRegression())])
parameters = {
        'vect__ngram_range': ((1, 1) , (1,2),(1, 3)),  # unigrams or bigrams or trigrams
        'clf__penalty': ('l2', 'l1'),
        'clf__C': (1, 10,12, 14, 15, 20, 25, 30, 35, 40, 45, 50, 55,70,75,100) }

grid_search = GridSearchCV(pipeline, parameters, n_jobs=-1, verbose=1, refit=True,
                  cv=15, scoring='accuracy')
```

Performance Measures

Table 1: Confusion Matrix (Rueda and Diaz-Uriarte, 2007)

True Class	Predicted Class			
	equipment	structures	not applicable	Total
Equipment	Ee	Es	En	E.
Structures	Se	Ss	Sn	S.
Not Applicable	Ne	Ns	Nn	N.

- The entries in the confusion matrix have the following meaning in the context of the study:
  - ***Ee*** is the number of **correct** predictions that a write-in is Equipment**.**
  - ***Es*** is the number of **incorrect** predictions that a write-in is a **Structure,** when in fact it is **Equipment.**
  - ***En*** is the number of **incorrect** of predictions that a write-in **is Not Applicable,** when in fact it is **Equipment.**

Statistics Used to Evaluate Performance

Correct Classification Rate

$$CCR = \frac{Ee+Nn+Ss}{E.+N.+S.}$$

False Discovery Rate

$$FDR = \frac{Es+En}{En+Nn+Sn+Es+Ns+Ss}$$

Specificity

$$Specificity = \frac{Ee}{Ee+Es+En}$$

Sensitivity

$$Sensitivity = \frac{Nn+Ss}{N.+S.}$$

Results

The performance statistics for the compared methods on the test data.

Model	SVMs	Logistic Regression
CCR	.9789	.9794
FDR	.0076	.0076
Specificity	.9978	.9978
Sensitivity	.9146	.9171

Pros and Cons of SVM

Pros

- Can deal with very high dimensional data.
- SVMs work very well in practice, even with very small training sets

Cons

- Non-Probabilistic: SVMs do not directly provide probability estimates

Pros and Cons of LR

Pros

- Wide spread industry comfort for logistic regression solutions → trusted
- Convenient probability scores
- Quick to train

Cons

- Logistic regression tends to underperform when there are non-linear decision boundaries.

Conclusions

- LR had a slightly higher Correct Classification rate than SVM.
- LR achieved the highest sensitivity.
- LR was the overall best performing method.
- Recommend ACES staff deploy a LR model into a production system.

References

Bewick, V. Cheek L and Ball J. 2005. *Statistics review 14: Logistic regression*. Critical Care, London, England.

Chanawee Chavaltada, Kitsuchart Pasupa, David R. Hardoon. 2017. *A Comparative Study of ML Techinques for Automatic Product Categorisation.*, ISNN(1):10-17.

Kim, W., Gordon, D., Sebat, J., Ye, K. Q., and Finch, S. 2008. *Computing power and sample size for case-control association studies with copy number polymorphism: Application of mixture-based likelihood* ratio test PLoS ONE, 3, e3475. https://doi.org/10.1371/journal.pone.0003475

Thorsten Joachims, 2001. A statistical learning model of text classification for support vector machines. In *Proceedings of the 24th Annual International ACM SIGIR conference on Research and Development in Information Retrieval).* Association for Computing Machinery, pages128–136. https://doi.org/10.1145/383952.38974.

Marin Vuković, Krešimir Pripužić, and Hrvoje Belani. 2009. An Intelligent Automatic Hoax Detection System. In *Proceedings of the 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Lecture notes in Computer Science, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04595-0_39.

Sida Wang and Christoper Manning. 2012. Baselines and bigrams: simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 90-94.

Questions?

Contact information:

Dr. Andrea Roberson

andrea.roberson@census.gov

Justin D. Nguyen

justin.d.nguyen@census.gov

Thank you for your attendance and attention!