# Putting the Cure into Strategic Procurement

NLP soft weighted voting ensemble multi-class classifier to predict spend categories using Government of California's 2012 -2015 purchase order data

Dene Stalk

BrainStation Data Science Diploma Capstone Project

March 2020

# Putting the Cure into Strategic Procurement

## DENE STALK

Data Science Diploma
BSc, MBA

Work Experience
Management Consulting, Internal Strategy, Tourism

## Importance of conducting a spend analysis

- Identify cost saving opportunities
- Facilitate strategic procurement decision making and sourcing

## Capstone Overview

- Predict purchase order spend categories
- 6 weeks
- Government of California's 2012-2015 purchase orders
- NLP weighted voting ensemble using supplier name, item name and item description

Tools: Python, Pandas, Numpy, Sklearn, Tableau, MySQL

### Correctly Classified

**77% of line items**

**92% of Total Spend**

# Overview of Business Importance, Challenges and Machine Learning's value in procurement

## Why a Spend Analysis?

- Full spend visibility
- Identify cost saving opportunities
- Data-driven sourcing

## Challenges

- Disconnect across regions, languages and business units
- Labour intensive
- Prone to human error

## ML's Procurement Value

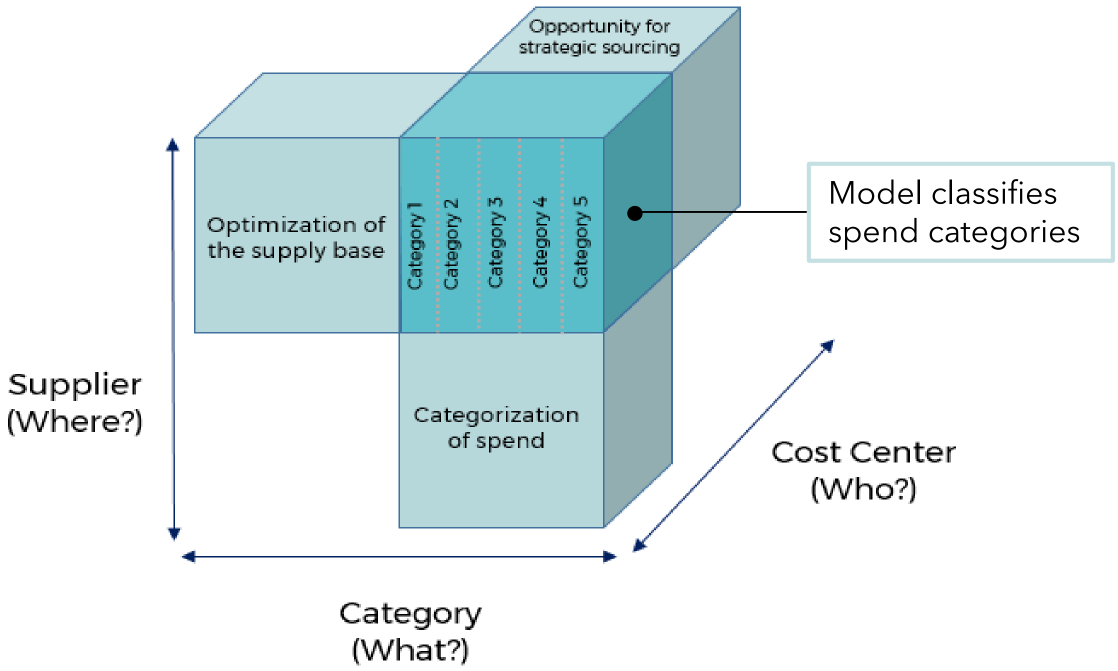
- Reduces need for manual intervention
- Timely strategic decision making
- Better inventory and supplier management

# Importance of Correctly Classifying Spend Categories

Purchase Order (PO): document issued by a buyer committing to purchasing products or services

Takes an average of 1 year to identify and correct 100k misclassified line items

Increased spend visibility leads to lower Total Cost of Ownership[1]



Improved buying power through strategic sourcing

Optimizing inventory management processes

Managing risk exposure to maverick spending[2]

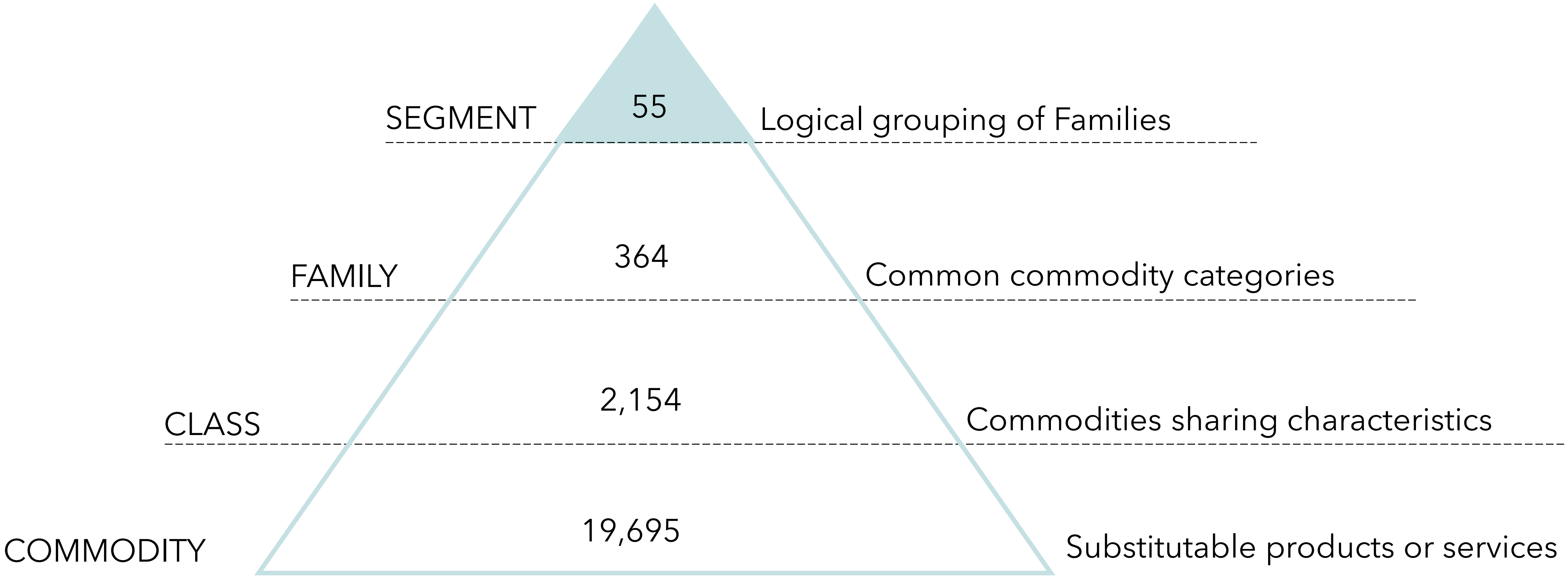
[1] - Total Cost of Ownership is a combination of purchasing and carrying costs

[2] - Maverick spending refers to purchases made outside of agreed contracts

Source: Sievo, Zycus

# United Nation’s Classification System

**UNSPSC** : United Nations Standard Products & Services Code



SEGMENT 55 Logical grouping of Families

FAMILY 364 Common commodity categories

CLASS 2,154 Commodities sharing characteristics

COMMODITY 19,695 Substitutable products or services

Source: Open Canada

# Data Collection and Description

CA.GOV CALIFORNIA OPEN DATA PORTAL

**Purchase Order Data 2012-2015**

Government of California merged information from various procurement systems, then mapped the UNSPSC taxonomy

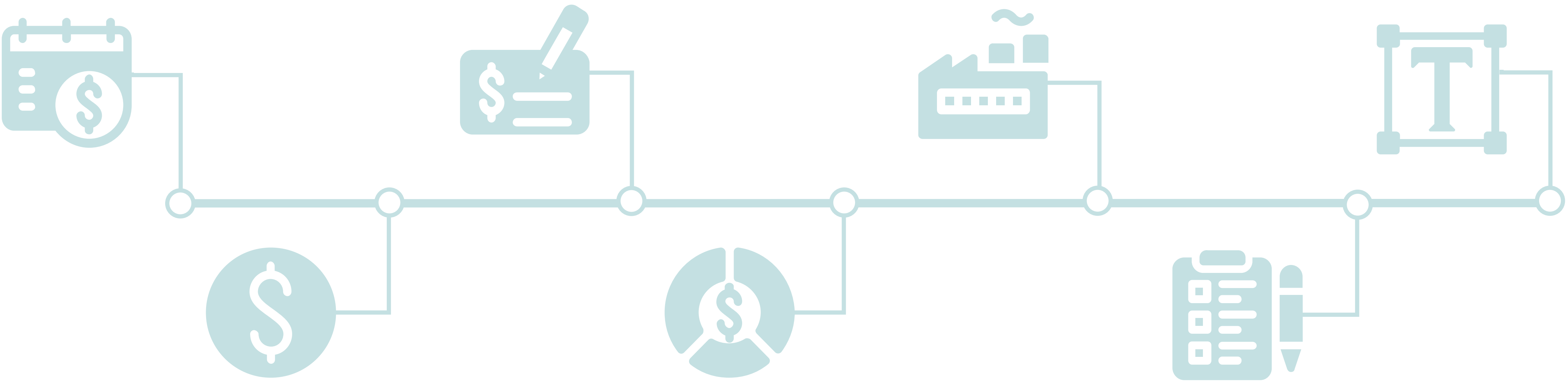
## Cleaned Dataset Description

3 fiscal years

326k line items

24k suppliers

3.6m words



$147bn in
Total Spend

26 Spend categories[1]

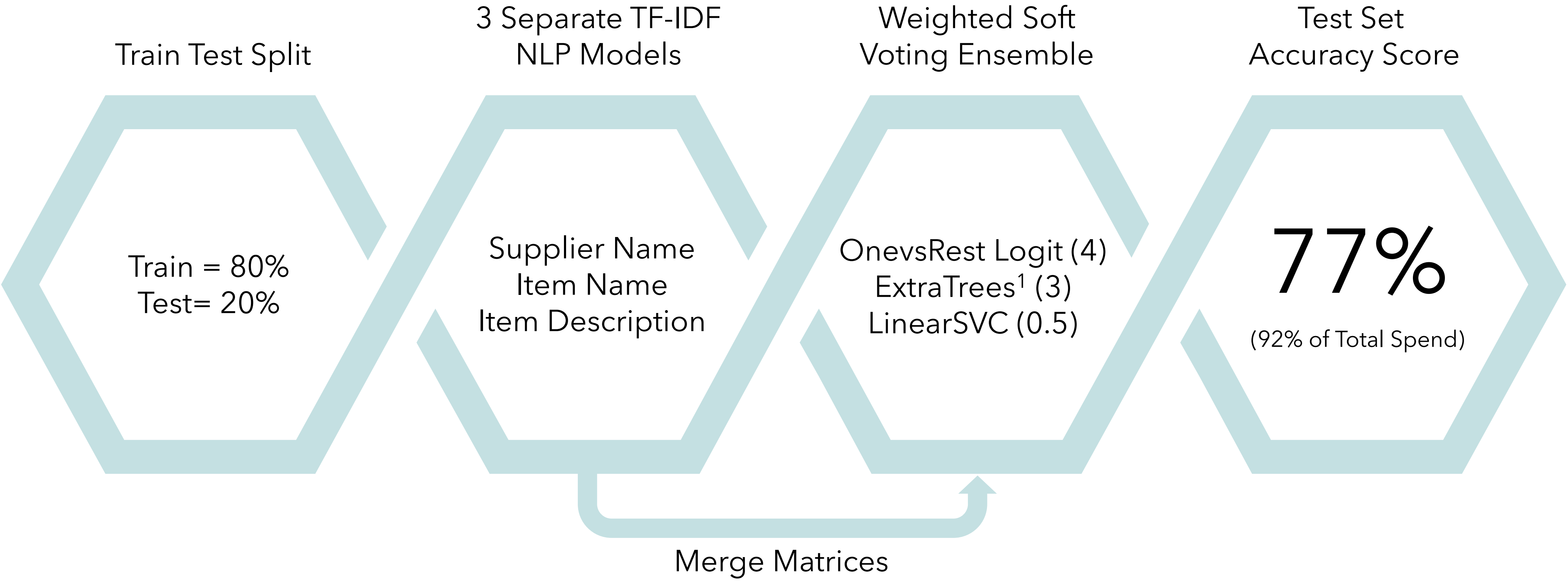
171k unique item names
208k unique item descriptions

[1] – Bottom categories in terms of items lines were placed into either Other Services or Other Goods

Detailed data cleaning process available at github.com/denestalk

# Modelling Methodology

NLP weighted soft voting ensemble predicted 65k line items of unseen data with 77% accuracy

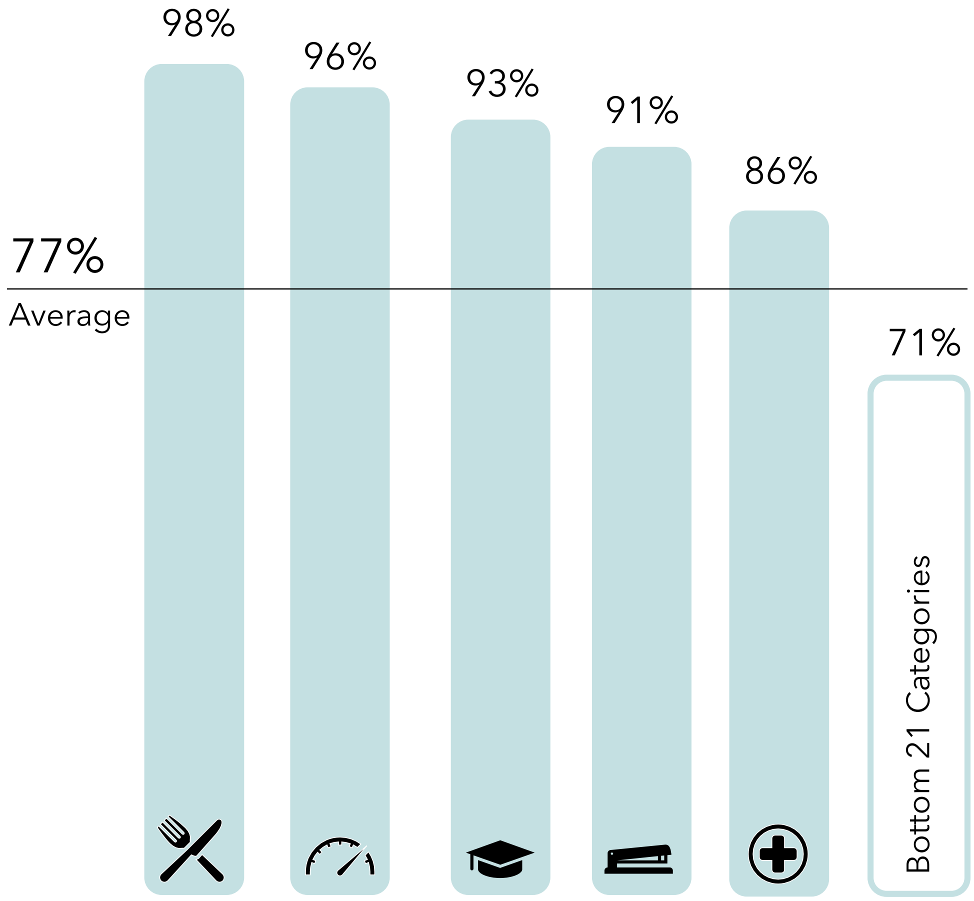


[1] - ExtraTrees differs from RandomForest by splitting nodes randomly and does not resample observations

Detailed modelling process available at github.com/denestalk

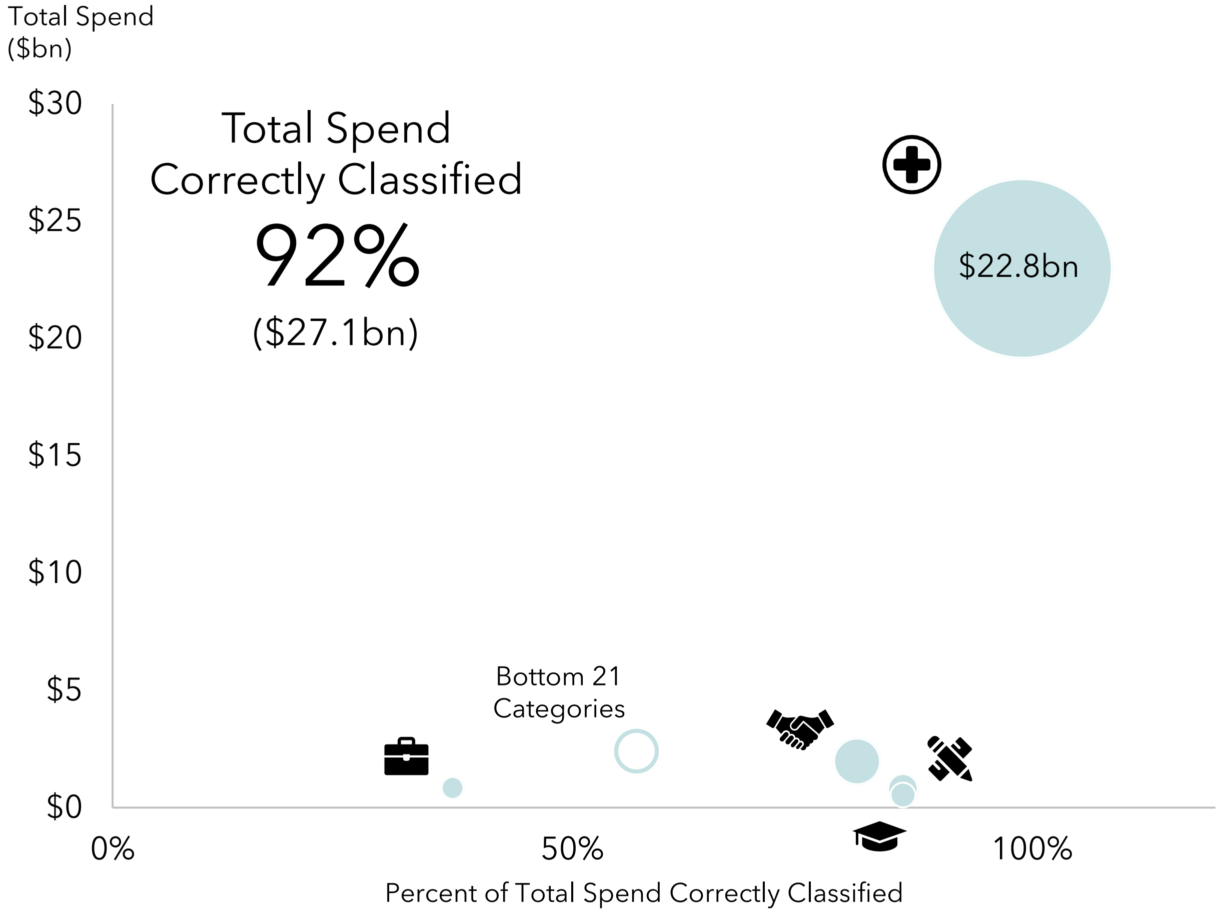
# Results

## Top 5 Categories by Percent of Line Items Correctly Classified



## Top 5 Categories by Total Spend

(bubble size indicates Total Spend correctly classified)



Food & Beverage Products
Fuel & Lubricants
Health Services
Office Supplies
Education & Training Services
Professional Services
Other Services
Engineering Services

Category names have been abbreviated
Full breakdown available at github.com/denestalk

# Conclusion

## Model Strengths

- Relatively strong predictive value using only 3 free text columns
- Correctly classified the vast majority of spend
- 17 categories' line items correctly classified with over 70% accuracy
- Model is scalable

## Model Weaknesses

- Not tested on corporate data
- Health Services over represented in terms of Total Spend
- 3 categories < 50% of line items correctly predicted

## Further Analysis

- Meta-model incorporating quantity and unit price
- Deeper analysis on strategic value in relation to supplier management
- Test transferability on new Government of California data and corporate data
- Create an interactive dashboard

Category names have been abbreviated

# Thank you

**Dene Stalk**
denemstalk@gmail.com
416-573-5784

linkedin.com/in/denestalk
github.com/denestalk