Using Machine Learning to Predict NAPCS Codes from Product Descriptions

Economic Statistical Methods Division

Carma Hogue —ESMD Division Chief

Andrea Roberson —ESMD Mathematical Statistician

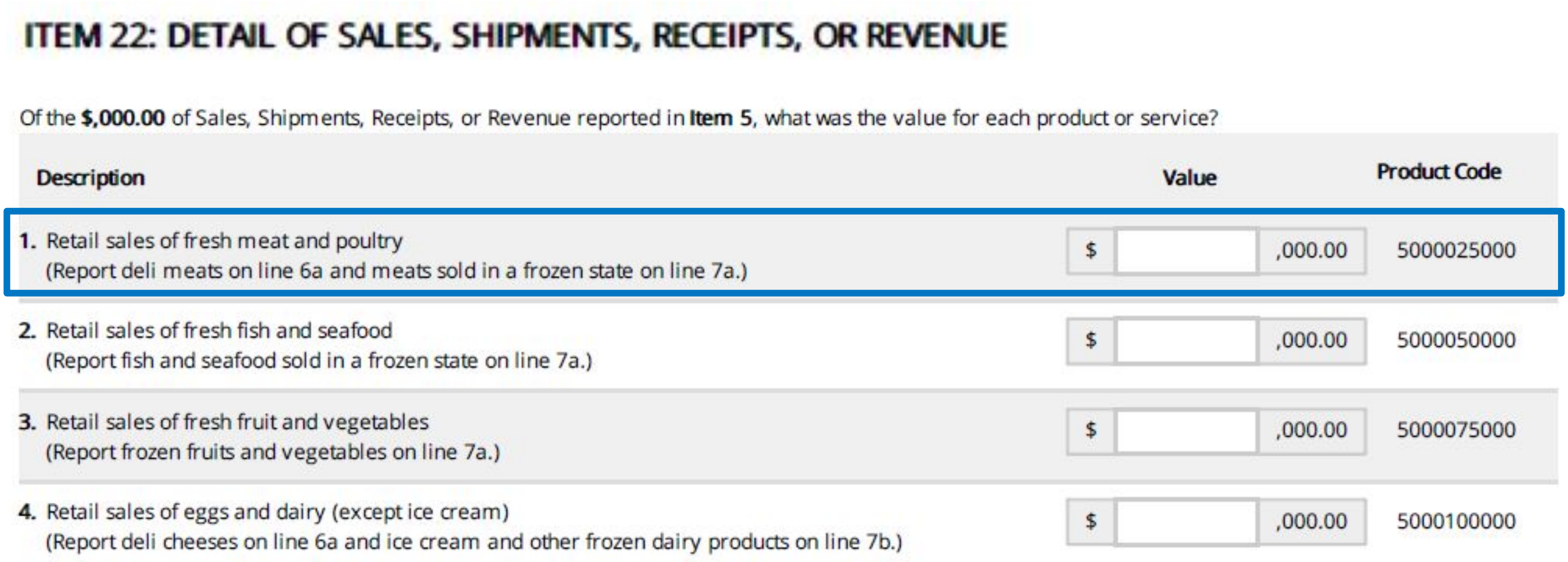
CIVIC DIGITAL FELLOWSHIP

United States™ Census Bureau

Beverly Bendix &
Carnegie Mellon University
M.S Public Policy & Management

Fatima Koli
Columbia University
M.S Data Science

The project was motivated by item 22 of the Economic Census



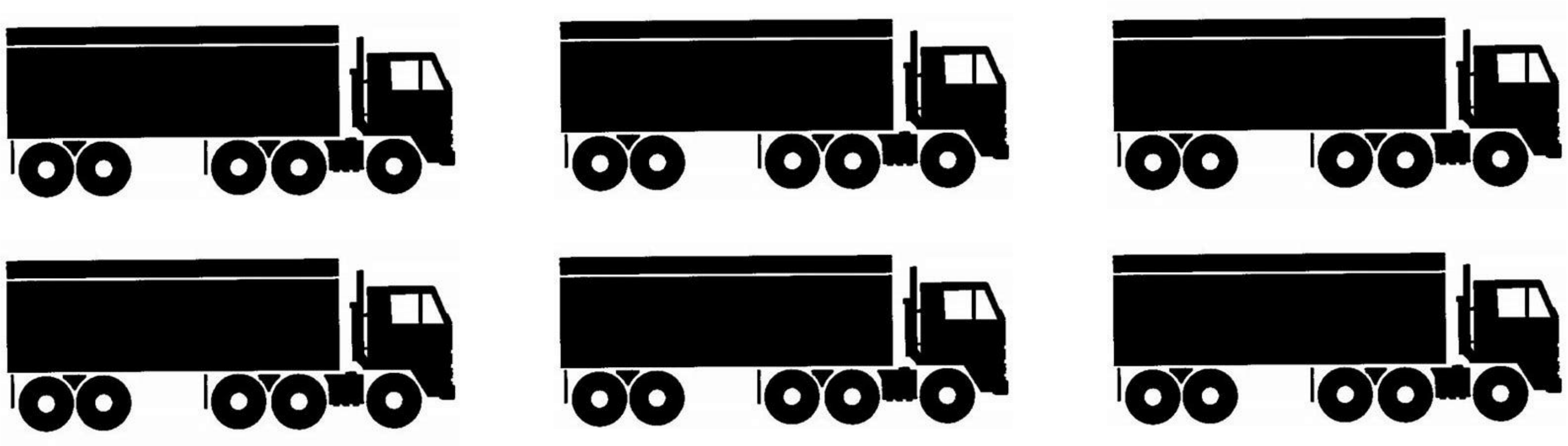
ITEM 22: DETAIL OF SALES, SHIPMENTS, RECEIPTS, OR REVENUE

Of the **$,000.00** of Sales, Shipments, Receipts, or Revenue reported in **Item 5**, what was the value for each product or service?

Description	Value	Product Code
1. Retail sales of fresh meat and poultry (Report deli meats on line 6a and meats sold in a frozen state on line 7a.)	$,000.00	5000025000
2. Retail sales of fresh fish and seafood (Report fish and seafood sold in a frozen state on line 7a.)	$,000.00	5000050000
3. Retail sales of fresh fruit and vegetables (Report frozen fruits and vegetables on line 7a.)	$,000.00	5000075000
4. Retail sales of eggs and dairy (except ice cream) (Report deli cheeses on line 6a and ice cream and other frozen dairy products on line 7b.)	$,000.00	5000100000

3.7 Million Businesses are Impacted

...equal to 6 tractor trailers of notification letters



http://www.tecnopneumatic.gr/Uploads/PhotosFiles/analosima_voulkanismou_WEB/tech_icons/Tractor_Trailer2.jpg

The 1,200+ NAPCS codes are complex & can be ambiguous

Which category?

- Retail sales of candy, prepackaged cookies, and **snack foods**

- Retail sales of food **dry goods** and other foods purchased for future consumption (Include flour, sugar, fats and oils, coffee, honey, jams and jellies, pasta, and crackers.)

- Retail sales of fresh fruit and vegetables

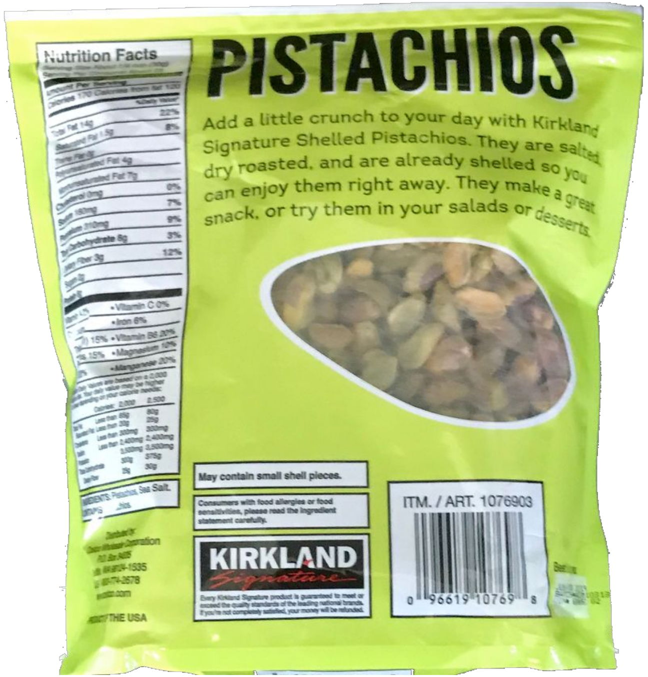

PISTACHIOS
Add a little crunch to your day with Kirkland Signature Shelled Pistachios. They are salted dry roasted, and are already shelled so you can enjoy them right away. They make a great snack, or try them in your salads or desserts.
Nutrition Facts
May contain small shell pieces.
Consumers with food allergies or food sensitivities, please read the ingredient statement carefully.
KIRKLAND Signature
ITM. / ART. 1076903

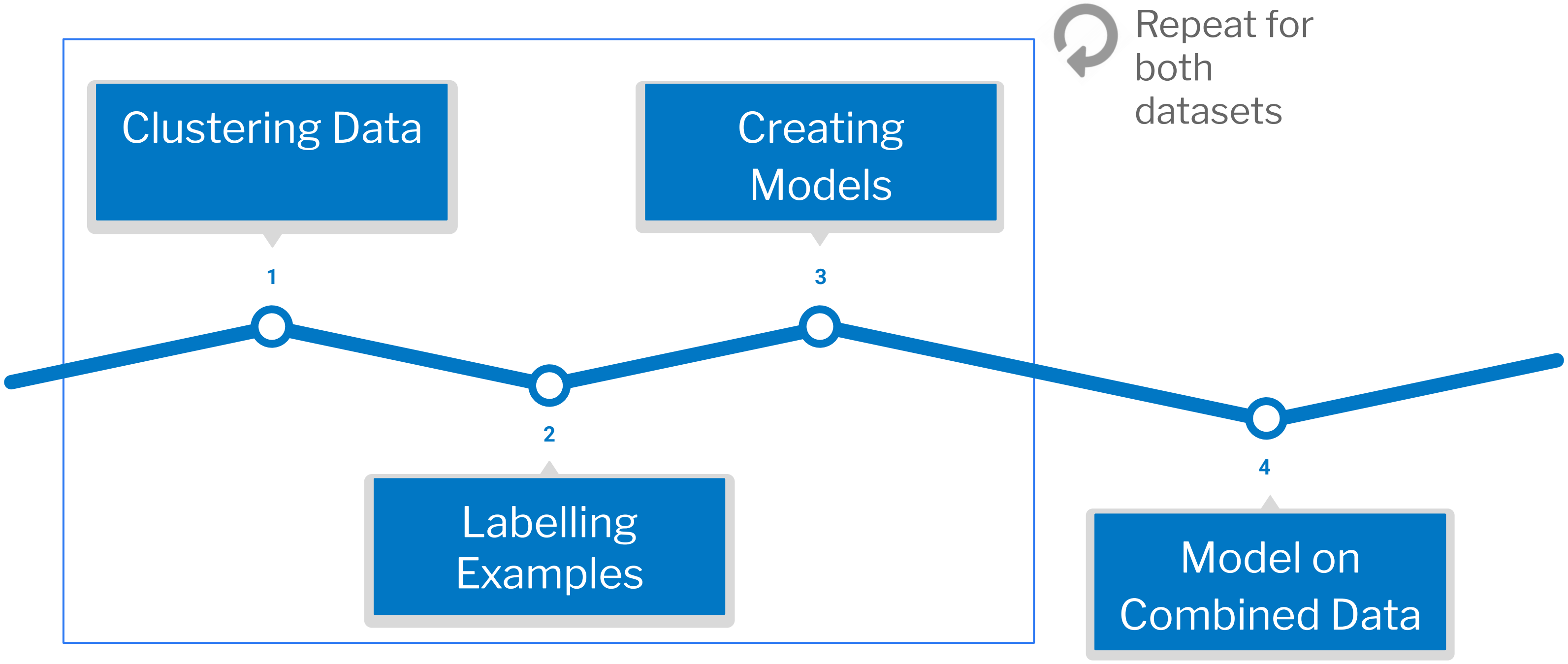
Photo credit: Jeff Meisel

This is a **text classification** problem

Using the UPC product descriptions, can we predict the NAPCS code?

Process

2 Datasets:
- Grocery store
- Drugstore



1 Clustering Data

2 Labelling Examples

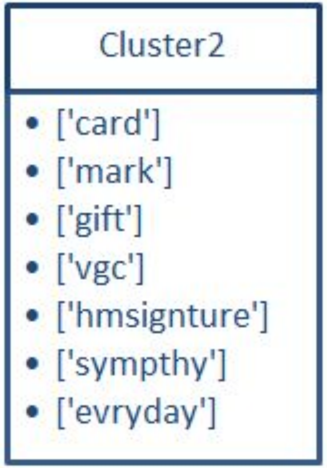
3 Creating Models

4 Model on Combined Data

Repeat for both datasets

Clustering Uncovers Patterns in the Data

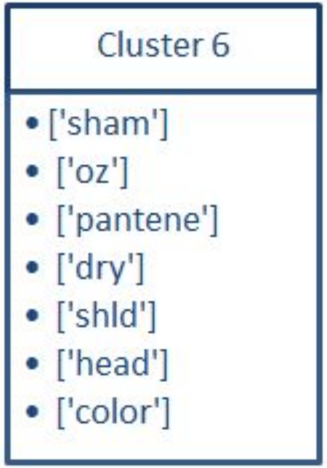
GREETING CARDS



Cluster2

- ['card']
- ['mark']
- ['gift']
- ['vgc']
- ['hmsignture']
- ['sympthy']
- ['evryday']

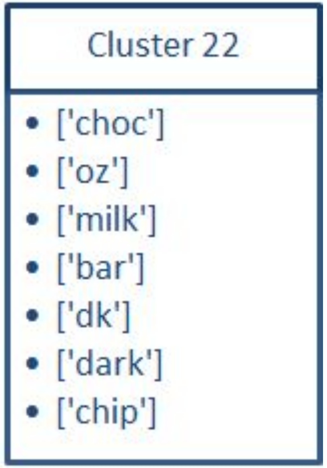
HAIRCARE



Cluster 6

- ['sham']
- ['oz']
- ['pantene']
- ['dry']
- ['shld']
- ['head']
- ['color']

CHOCOLATE



Cluster 22

- ['choc']
- ['oz']
- ['milk']
- ['bar']
- ['dk']
- ['dark']
- ['chip']

… But Clustering Results Can be Misleading

Descriptions closest to “MEOW MIX ORIGINAL CAT FOOD 3.15LB”

Description	Similarity
FRISKIES PARTY MIX ORIGINAL 20OZ	.976
CHEEZ-IT SNACK MIX ORIGINAL 4.5Z	.958
CORN NUTS ORIGINAL 4OZ	.948
CHIPS AHOY! ORIGINAL COOKIES 13OZ	.944

UPC Descriptions are Used to Predict NAPCS Codes

UPC Description	→	NAPCS Code
SALMON KING HEAD WILD FRESH		Fresh fish and seafood
NO NONSENSE XSPORT NO SHOW SOCK ASTD		Women’s socks or Men’s socks?
MINI TUB OF FUN GENERIC		?

Label, Label, Label!

Dataset:	Drugstore	Grocery Store
# Obs. :	14K	597K

3 difficult weeks later:

Labeled:	14K	36.7K
% of Dataset	100%	6.1%
NAPCS Codes:	44	45

CIVIC
DIGITAL
FELLOWSHIP

Accuracy of Our Models:

95% on drugstore data

90% on grocery store data

Tested Model's Ability to Generalize

Training Data	Grocery Store (GS)	GS + 10% Drugstore	GS + 20% Drugstore	GS + 30% Drugstore
Accuracy on Drugstore Test Set	52%	75%	81%	85%

This will significantly modernize the Economic Census

BEFORE	AFTER
High respondent burden	**Low** respondent burden
Poor data quality	Automated classification is more **consistent** and **accurate**
Collected **every 5 years**	Could be done **annually!**

Recognition

Thank you to **Andrea Roberson** for her machine learning acumen and guidance and to **Carma Hogue** for taking us onto her team!

CIVIC DIGITAL FELLOWSHIP

Questions? Contact us at beverly.bendix@gmail.com and fatimazkoli@gmail.com

Talk to us on Twitter! @fatima_koli