Identifying whole food allergens using Machine Learning

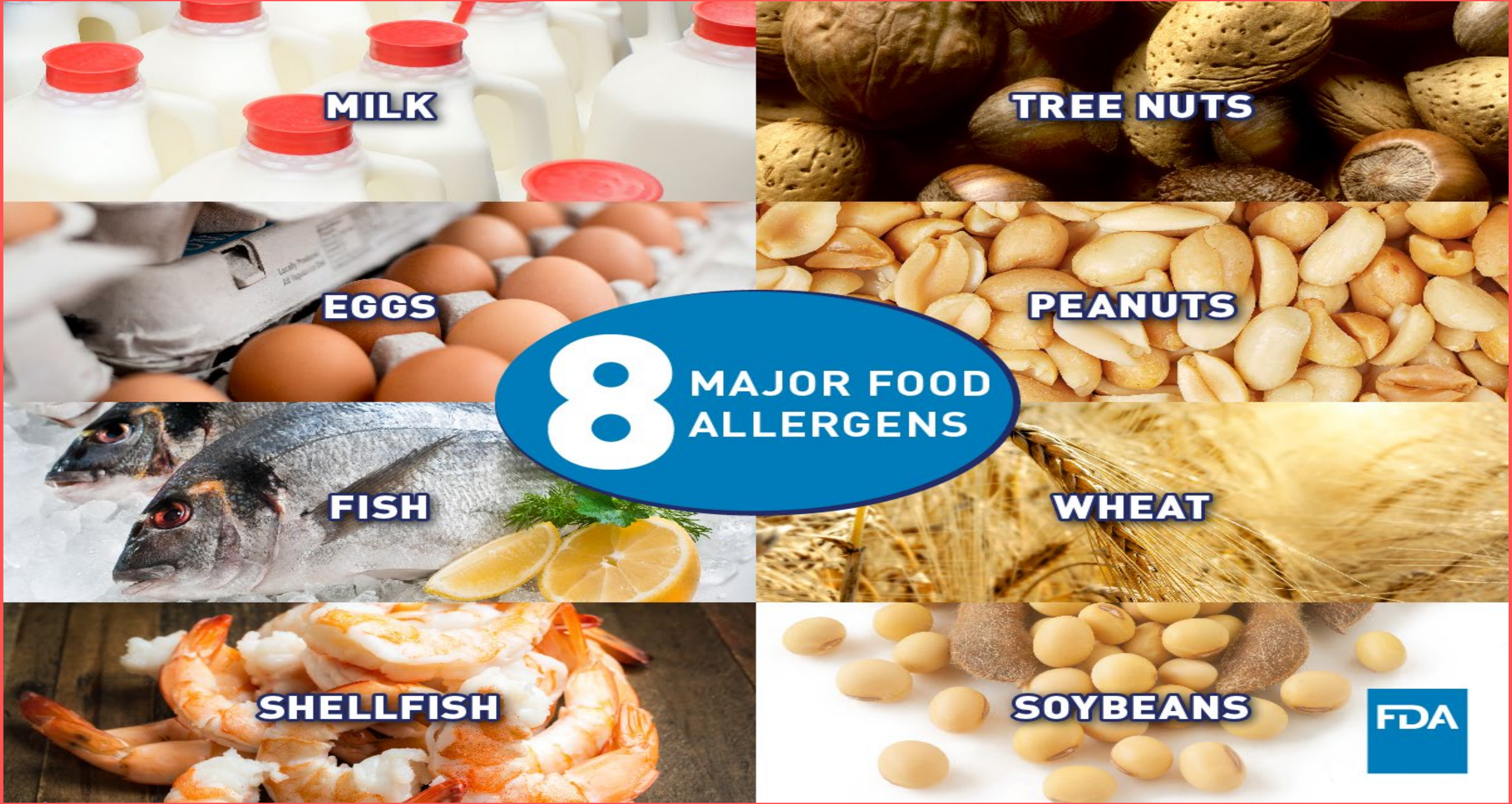
Yuthika Shekhar and Abhijeet Talaulikar

32 million

Americans have food allergies.

1 in 10 Adults

1 in 13 Children



Source: https://www.fda.gov/food/food-labeling-nutrition/food-allergies

Objective

- The problem with food labels
- Raw foods vs whole foods
- Impossible to make exhaustive list of allergens
- Absence of hierarchical allergen ontologies

This would assist the FDA in regulating the labels on food products.

Data Sources

Composition of Foods Raw, Processed, Prepared USDA National Nutrient Database for Standard Reference

The U.S. Department of Agriculture (USDA) National Nutrient Database for Standard Reference, contains data on 7,793 food items and up to 150 nutrients.

Open FDA Substance Data

Get Unique Ingredient Identifier (UNII) codes for substances and their synonyms based on substance’s molecular structure. This is generated through a joint effort of FDA and GSRS.

FALCPA food allergen list

This contains data of 8 major food allergens and their derivatives.

Methodology

Data Collection & Cleaning

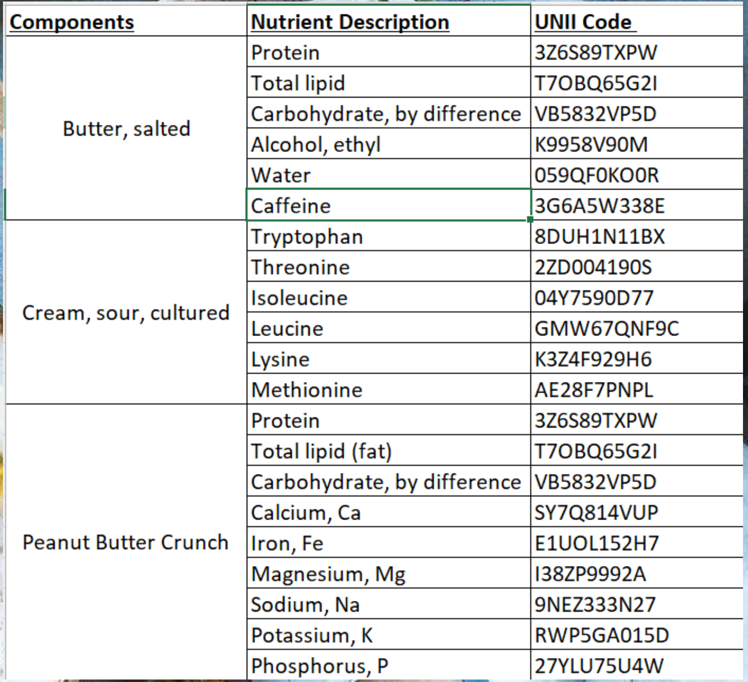
Data Annotation

Modeling

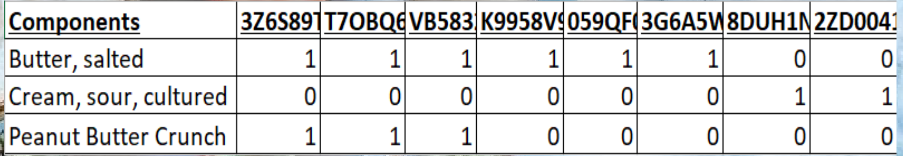
1

Data collection & cleaning

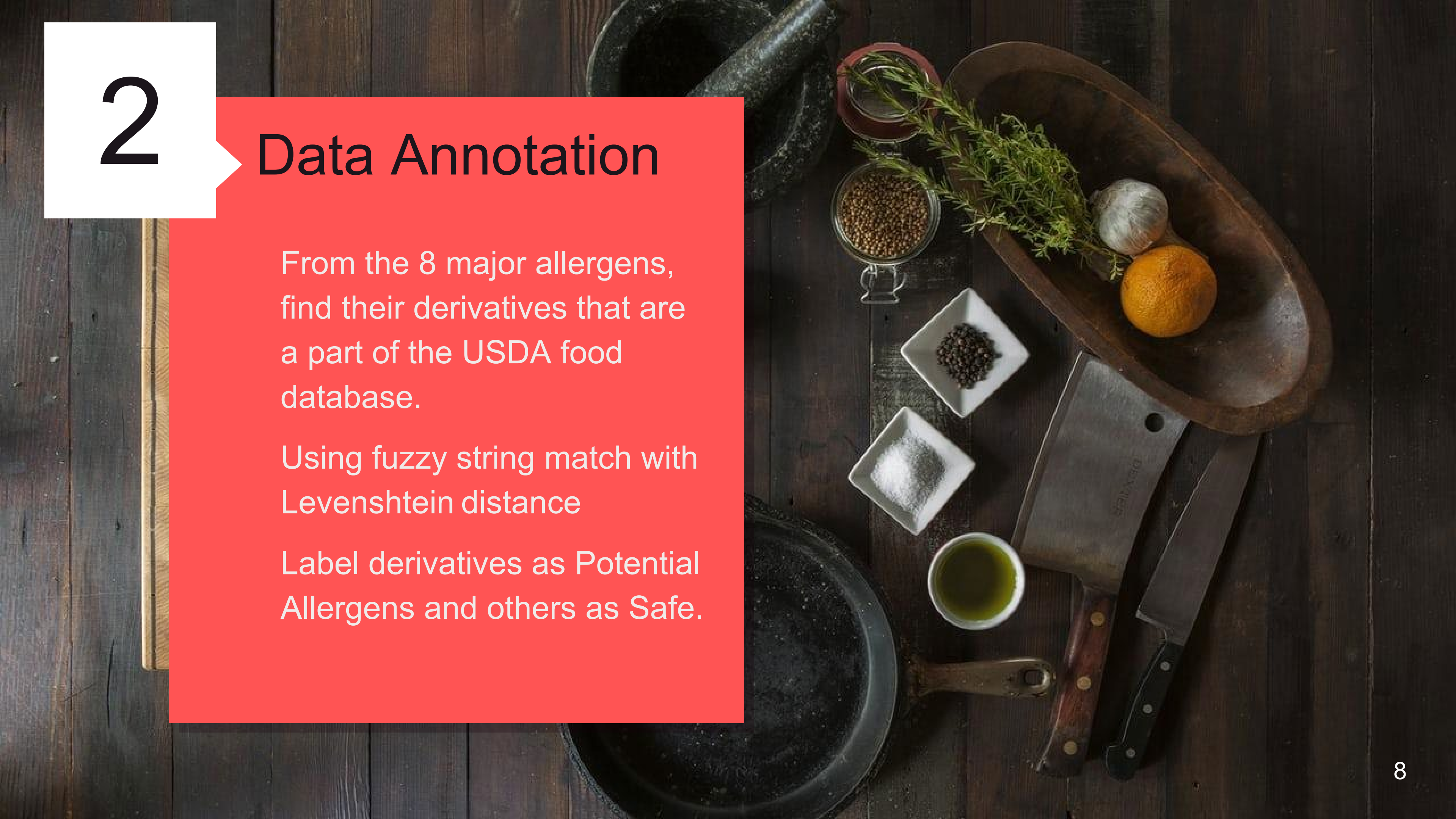
- Map constituents to their UNII codes in USDA nutrient database.
- For 150 nutrients, the dataset has 150 one-hot encoded columns for presence of ingredients.



Components	Nutrient Description	UNII Code
Butter, salted	Protein	3Z6S89TXPW
	Total lipid	T7OBQ65G2I
	Carbohydrate, by difference	VB5832VP5D
	Alcohol, ethyl	K9958V90M
	Water	059QF0KO0R
	Caffeine	3G6A5W338E
Cream, sour, cultured	Tryptophan	8DUH1N11BX
	Threonine	2ZD004190S
	Isoleucine	04Y7590D77
	Leucine	GMW67QNF9C
	Lysine	K3Z4F929H6
	Methionine	AE28F7PNPL
Peanut Butter Crunch	Protein	3Z6S89TXPW
	Total lipid (fat)	T7OBQ65G2I
	Carbohydrate, by difference	VB5832VP5D
	Calcium, Ca	SY7Q814VUP
	Iron, Fe	E1UOL152H7
	Magnesium, Mg	I38ZP9992A
	Sodium, Na	9NEZ333N27
	Potassium, K	RWP5GA015D
	Phosphorus, P	27YLU75U4W



Components	3Z6S89T	T7OBQ6	VB583	K9958V	059QF	3G6A5W	8DUH1N	2ZD004
Butter, salted	1	1	1	1	1	1	0	0
Cream, sour, cultured	0	0	0	0	0	0	1	1
Peanut Butter Crunch	1	1	1	0	0	0	0	0



2 Data Annotation

From the 8 major allergens, find their derivatives that are a part of the USDA food database.

Using fuzzy string match with Levenshtein distance

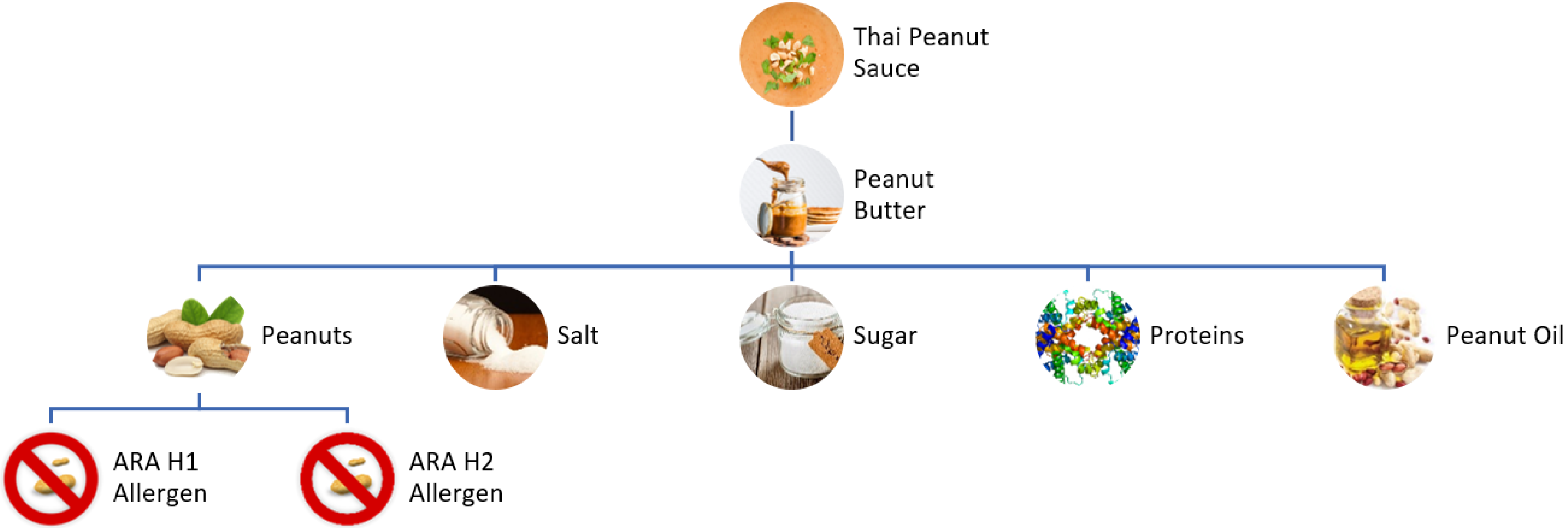
Label derivatives as Potential Allergens and others as Safe.

3

Modelling

- With around 8000 observations, 150 features and 1 binary label, we can use a classifier like SVM or gradient boosting to label new foods as potential allergens or safe.
- Find multidimensional association rules with high support and confidence that link nutrients to the label.

Expected Outcomes: Allergen Ontologies



Thai Peanut Sauce

Peanut Butter

Peanuts

Salt

Sugar

Proteins

Peanut Oil

ARA H1 Allergen

ARA H2 Allergen

Focus Areas

We address the following areas with the solution

- Product Safety Surveillance
- Artificial Intelligence
- Empowering patients and consumers to make better-informed decisions.

References

- https://data.nal.usda.gov/dataset/composition-foods-raw-processed-prepared-usda-national-nutrient-database-standard-reference-release-27
- https://www.usda.gov/media/blog/2019/11/21/new-nutrient-content-information-now-online
- https://open.fda.gov/apis/other/substance/
- https://www.fda.gov/food/food-labeling-nutrition/food-allergies
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4954633/
- https://precision.fda.gov/uniisearch