Hiscox University Data Challenge 2018

Xuhui Li, Stewart Hutchins, Jonny Powell, Will Bennett, Courtney Elmy

Given the task to identify the key triggers and trends which lead to a product recall, we aimed to process and analyse data to do exactly that. Throughout the project, we considered how to best present our findings in a clear and concise manner.

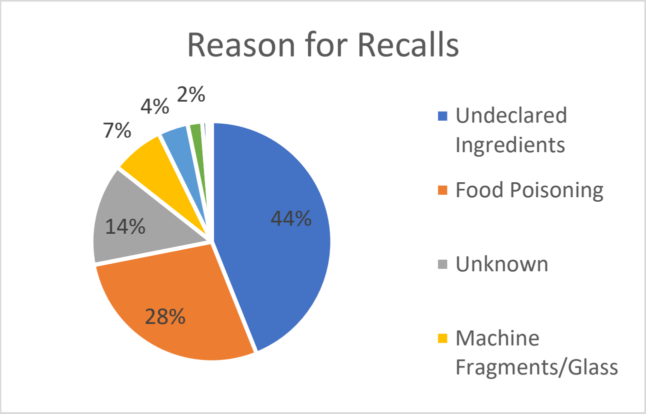
We aimed to make the most of the FDA datasets. Although not all analytical ideas and techniques were fully exploited, we found what we believe to be an accurate representation of the data and the most appropriate solutions to the problem. Furthermore, we hope any currently unpursued ideas and techniques can either be implemented in section 2 of the project, or show promise for other future applications. Moreover we believe these unique ideas show merit and a deeper understanding of statistical, computational and data science concepts.

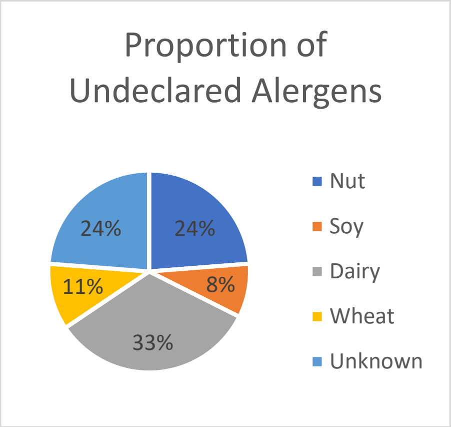
Key Insights

What are the most common reasons for a recall and what controls should be prioritized?

Using FDA data [1], we found there were three common triggers, which account for just under 80% of all recalls:

- undeclared ingredients, (44%),
- food poisoning, (28%)
- physical contamination, (7%)

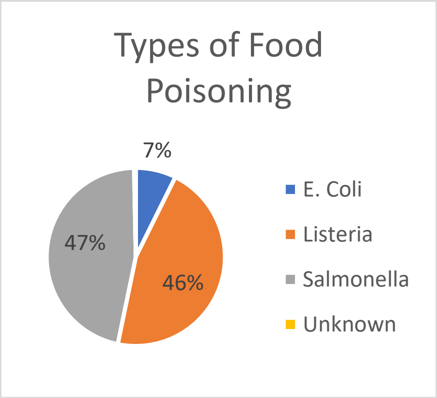




Dairy contributes the most to undeclared allergens, representing 33% of undeclared allergens, and is therefore responsible for an astounding 15% of all product recalls.

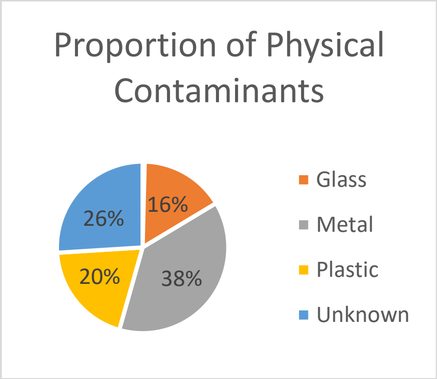
Food poisoning, the second most likely reason for a recall, contributes to 28% of all recalls; the salmonella and listeria pathogens combined account for 93%, and the other 7% comprises of E. Coli.

Considering all recalls, listeria and salmonella are the sources for 13% of recalls each, the equivalent of at least 500 million units over 10 years. (Due to the method for determining the number of units, as explained later, the true figure could be at 2 or 3 times this figure). This equates to approximately 50 million units per year; as [2] states 48 million Americans get sick from food poisoning every year, this figure is not particularly out of the question.

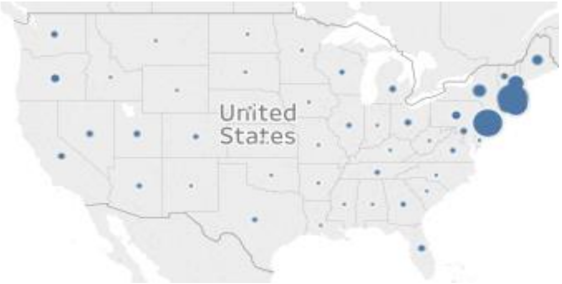


From the dataset [3], we can see dairy, sandwich and vegetable products are at a higher risk of listeria contamination than other products, while nut and powdered drink products are more susceptible to salmonella contamination. This is concurrent with [4], which states listeria is found mostly in smoked fish, meats and cheeses.

Although not as much of an issue as undeclared allergens or food poisoning, physical contaminants, such as glass, metal and plastic, are a significant cause of product recall.

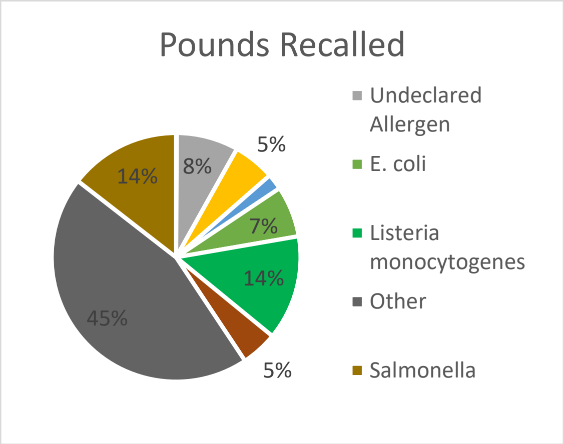


It is possible that location should be considered as a factor contributing to recalls. When observing a heat map, we see California has the highest proportion of recalls (17.7%) with New York (7.6%), New Jersey (7.2%) and Texas (6.1%) second, third, and fourth; however, it is imperative to consider the amount of food produced by each state. [5] If we assume income produced from food or amount of farmland per state is a good measure of food produced (i.e. a linear relationship), we see Alaska and North Easterly States, such as New Jersey and Massachusetts, produce a relatively high number of recalls.



When exclusively examining meat products [6] we used data from the Department of Agriculture Food Safety and Inspection Service (USDA) from 2008 to 2017. Undeclared allergen is the highest proportion of pounds recalled from 2008 to 2017 (34%). Therefore, it is highly expected to account for the largest percentage in the future, in comparison to with extraneous material (11%) and microbiological contaminations such as listeria (11%), E. Coli (11%) and Salmonella.

It is also indicated that the number of pork and poultry recalls, due to extraneous material, will increase in the future. Overall recalls of poultry and pork are very likely to rise, while that of the red meats, beef and lamb, will remain steady.



Food recalls can be a serious threat, causing not only financial repercussions as a direct result, but also a damaged reputation leading decreased sales. The data seems to highlight the underlying causes: undeclared allergens, bacterial contamination and physical contamination (a full list of proportions for individual products can be found at [7]).

As priorities, manufacturers should implement:

- Regular and thorough cleaning and sanitation processes, to reduce risk of microbiological contamination.
- A compliance audit of all packaged food, to reduce labeling errors or undeclared allergens.
- An accurate and comprehensive record of allergens within ingredients and allergens which appear within the manufacturer's walls.
- A program of frequent and accurate testing for microbial presence, as well as testing the functions of equipment. Microbial problems occur from under-processing or post-processing contamination; a lower frequency of microbes may not guarantee consistency, however ensuring equipment reaches the correct temperatures mitigates the issue.

We believe companies should also invest in developing new, innovative visual controls, from existing and emerging techniques, such as x-ray or computer vision and recognition, to reduce the risk of contamination from foreign objects.

With this guidance, the chance and impact of a food recall should decrease. However, there are often recalls which occur due to unforeseen circumstances, outside of the 'usual reasons'. The largest recall in U.S. history, 'The Westland/Hallmark Beef Recall' [8], did not occur for any of the above reasons, but instead was the result of inhumane animal treatment; a video showing sick cows, unable to walk onto the processing line that were not removed from the manufacturing process, was released. This damaged the business's reputation; as a result, they were forced to recall 143 million pounds of beef (2 years of product) and ordered to pay $500 million in settlements.

In addition to educating and ensuring enforcement of humane animal practices, we suggest that businesses should take a more active role in the manufacturing processes. If the manufacturer and brand are the same entity, a rigid code of conduct should be implemented, with internal and external inspectors ensuring proper regulation of said processes. If the manufacturer is a third party, the business should inspect the premises regularly, ensuring adherence to good manufacturing practices.

Methodology

When looking at the food-enforcement data, we noticed there were many similar recalls. A set of recalls could occur on the same day, have the same manufacturing firm, and be for the same reason; however, as their product description would slightly differ, the recalls would count as separate events. It was the consensus that a set like this should only count as one event.

We used the recall initiation date and firm name as unique identifiers, processing the data such that if multiple records had the same recalling firm and initiation date, only one event in the new data set would be recorded (source code [9]). This raises the issue of mapping all of the product descriptions and product quantities to a single event. Due to the similarity of product descriptions, it wouldn't be unrealistic to assume that analysis could map similar descriptions to the same food type, e.g. raspberry and blueberry ice cream could still map to an 'ice cream' or 'frozen desert' food type. Therefore, it wouldn't matter which description, in the condensing process, was used. Product quantity values were concatenated for later processing, to prevent loss of information.

Observing the "product quantity" variable, we can see some patterns in language used. Quantities were often described as 'x units' or 'x cases' followed by the number of units per box or pounds per unit. By looking for key words such as "units", "bottles", "cans", "tins", "cases", "ct", etc. we could detect the number of units/ cases in a single record. The same process was used for boxes, but this was a less prevalently used measure, so did not yield a useful number of results.

During the repeat removal process, strings were concatenated, if they did not match a previously added line; therefore, the number of units in a recall single was the sum of numbers preceded by a key word (source code [10]). Due to the variations in language, the process wasn't perfect, it is currently unable to use weight as a measurement and unable to obtain the desired figures for some events. Although imperfect, obtaining a number of units/cases per recall allows a magnitude to be applied to each recall event.

The next step was the 'meat of the matter': to standardize each food and reason for recall into short, easily readable products or reasons, as opposed to free-form strings. First, we created a program (source code: [11]) to count the frequency of each word in the data set, from this we attempted to pick common words and phrases which would identify a food or reason for recall. We also used an FDA document [6], to guide the reason for recall selection process.

A text file, 'Food Types.txt' [12] defines lists of foods. The structure of a line, in the config file, is a series of key words or foods associated with a more general food type, e.g. ice, juice, soda, etc. followed by "drink". A text file was chosen to input food types, to allow a user to search for food types or ingredients, without alteration to the source code. It should be noted: the system will categorize a food based on the first word it finds from the configuration file. As such, the precedence of words should be carefully considered; e.g. if the word 'butter' were searched for before the words 'buttermilk' or 'butter bean', a buttermilk or butter bean product would be mapped to the butter food type.

Once we had a vocabulary of key words there were 2 approaches to finding the reason for a recall:

- Examine the entire string, from beginning to end, to find a key word.
- Examine specific sections of the string, expanding the length of the substring from a specific point; if we knew the reason were 'Undeclared Allergen', it would be more appropriate to look for the allergen name directly before or after the key words 'undeclared', 'allergen', 'may contain', or 'residue'.

Consider the examples 'undeclared allergen: peanuts' or 'product may contain peanut residue'. We can see how the 2nd approach may be more appropriate in these instances.

The process also created two separate text documents, containing all the food descriptions and recall reasons which couldn't be classified [13, 14], so reasons or foods could be added to the system.

Challenges and What's Next?

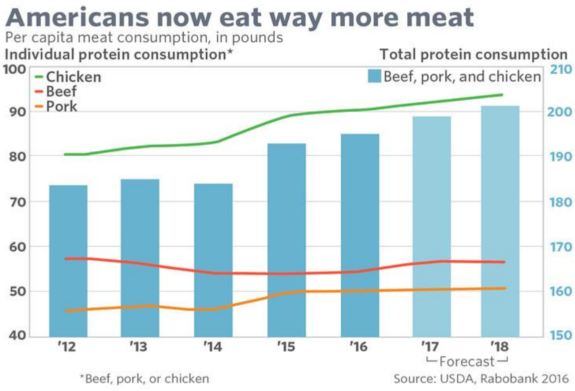
Ideally, we would have liked to extract the number of units or cases, or weight for all recalls. However, as mentioned briefly before, there is little standardization of this variable; one event may say '12 thousand total pints', while another says '115 industrial cases, 15oz/ct, (20ct/box)', and another may say '14,000'. Although creating a system to derive the number of units/cases for all records is not impossible, it is difficult and extended beyond the scope of the 1st question. However, an attempt to solve this problem was made, it entailed looking at events where both units and cases were given. By plotting a scatter graph of these points, the gradient of the trendline would give the average number of units in a case. Unfortunately, there was not sufficient data to find the relationship. (Raw data can be found at [15]).

A better solution may be to separate the different linguistical ways of expressing a volume, weight, or number of units/ cases to break down the problem. Then, for each linguistical expression, derive which combination of operators and numbers produce the correct measure.

As previously explained, when configuring the food types text document, the order of a list must be carefully considered. Furthermore, because the English language is confusing, classification of some products is difficult. Note the example of '*ham, lettuce, and tomato on wheat*'; although to a reader, this is clearly a sandwich, if we only used the key word 'sandwich', the system wouldn't work as intended. As a result, we have to use phrases like "in a bagel", "on a bun", "on wheat", etc. to ensure all sandwiches are detected. One way to fix this issue may be to use a neural network. In theory, the combination of words, e.g. 'on' and 'wheat', in labelled data would be recognised as significant to a sandwich product type. We believe it would be best to derive this labelled dataset from the given FDA event dataset [16]. Both a product description is given, as well as an industry code. As such words from the product description can be associated with each industry/ generic food type.

Although not a priority, a further classification problem is, "do we care if a '*bacon, lettuce, and tomato*' is specifically a sandwich?"; would it be better to classify this as a meat, vegetable, or wheat product? The solution to this may be to implement a fuzzy or neuro-fuzzy system - a system typically used to classify continuous data into more abstract, discrete bins. A classic example is 'at what point does a cold temperature become warm or hot?'. By the same logic 'at what point does a sandwich stop becoming a wheat product and start becoming a meat or vegetable product?'.

To continue the project, we aim to compare the frequency and volume of a recall to its demand over time. Earlier in the project we noted the red meats, beef and lamb, saw little growth in their recall volumes, while other meats, such as pork and poultry, recalls were increasing year on year. From the graph right [18], we see an increase in poultry recalls may be explained by its increase in demand, however the consumption of pork has remained relatively consistent, perhaps indicating a bad investment opportunity. We aim to perform statistical analysis to these observations, to provide accurate figures and recommendations.


Americans now eat way more meat

Source: USDA, Rabobank 2016

We also aim to analyse the volatility of a product. Given a product or reason for recall, we would expect the number of units recalled to follow a normal distribution pattern (this may not be true in reality). Products with a particularly high standard deviation, would be a particularly bad investment, due to the decreased predictability of the size of a recall. If a reason for recall's units/ cases were found to follow a normal distribution pattern, we would need to look to the 1st part of the project and identify which products were particularly affected by each reason [7]. This idea may require the implementation of a better unit extraction strategy.

As an extension, we may apply the above neural network solution, in order to determine a more accurate classification technique, than the current key word search. [17] creates a sample of 200 events (or less, if altered), which can be manually checked to calculate the accuracy of each classification strategy, thus we can determine the better of the two classification strategies.

References:

[1] FDA, ‘Downloads [/food/event]’ Last Updated: Feb 2018, Last Download: Dec 2018, [Online] Available: https://open.fda.gov/downloads/

[2] Fortune, ‘America’s food industry has a $55.5 billion safety problem’, May 2016, Accessed: Feb 2018, [Online] Available: http://fortune.com/food-contamination/

[3] Team Probable Octo Palm Tree, ‘Graphs & Data for key insights’, Feb 2018, https://github.com/jonnyowenpowell/probable-octo-palm-tree/blob/master/Data/Graphs%20%26%20Data%20for%20Key%20Insites.xlsx

[4] European Food Safety Authority, ‘Listeria’, Accessed: Feb 2018, [Online] Available: https://www.efsa.europa.eu/en/topics/topic/listeria

[5] Team Probable Octo Palm Tree, ‘Events per State’, Jan 2018, [Online] Available: https://github.com/jonnyowenpowell/probable-octo-palm-tree/blob/master/Data/Events%20per%20State.xlsx

[6] United Sates Department of Agriculture, ‘Summary of Recall Cases in Calendar year 20XX’, Jan 2018, Accessed: Jan 2018, [Online] Available: https://www.fsis.usda.gov/wps/portal/fsis/topics/recalls-and-public-health-alerts/recall-summaries/recall-summaries-2017

[7] Team Probable Octo Palm Tree, ‘Suplementry_FDA_Food_Workbook.twb’, Jan 2018, https://github.com/jonnyowenpowell/probable-octo-palm-tree/tree/master/Suplementry%20Tableau

[8] Nat Berman, ‘The 10 Biggest Food Recalls in U.S. History’, Jan 2017, Accessed: Feb 2018, [Online] Available: http://moneyinc.com/biggest-food-recalls-u-s-history/

[9] Team Probable Octo Palm Tree, ‘Remove Repeats’, Jan 2018, https://github.com/jonnyowenpowell/probable-octo-palm-tree/tree/master/FileHandling/Remove%20Repeats

[10] Team Probable Octo Palm Tree, ‘Unit Extraction’, Jan 2018, https://github.com/jonnyowenpowell/probable-octo-palm-tree/tree/master/FileHandling/Unit%20Extraction

[11] Team Probable Octo Palm Tree, ‘Key Word Search Reason’, Jan 2018, https://github.com/jonnyowenpowell/probable-octo-palm-tree/tree/master/FileHandling/Key%20Word%20Search%20Reason

[12] Team Probable Octo Palm Tree, ‘Food Separation and Causation’, Jan 2018, https://github.com/jonnyowenpowell/probable-octo-palm-tree/blob/master/FileHandling/FoodSeparation%20%26%20Causation/Food%20Types.txt

[13] Team Probable Octo Palm Tree, ‘UnknownReasons.txt’, Jan 2018, https://github.com/jonnyowenpowell/probable-octo-palm-tree/blob/master/FileHandling/FoodSeparation%20%26%20Causation/UnknownReasons.txt

[14] Team Probable Octo Palm Tree, ‘UnknownFood.txt’, Jan 2018, https://github.com/jonnyowenpowell/probable-octo-palm-tree/blob/master/FileHandling/FoodSeparation%20%26%20Causation/unknownFood.txt

[15] Team Probable Octo Palm Tree, ‘Units Vs Case.csv’, Jan 2018, https://github.com/jonnyowenpowell/probable-octo-palm-tree/blob/master/FileHandling/Units%20Vs%20Case.csv

[16] FDA, ‘Downloads [/food/event]’ Last Updated: Feb 2018, Last Download: Dec 2018, [Online] Available: https://open.fda.gov/downloads/

[17] Team Probable Octo Palm Tree, ‘Create Sample’, Jan 2018, https://github.com/jonnyowenpowell/probable-octo-palm-tree/tree/master/FileHandling/Create%20Sample

[18] Catey Hill, ‘This chart proves Americans love their meat’, Dec 2016, Accessed Feb 2018, [Online] Available: https://www.marketwatch.com/story/this-chart-proves-americans-love-their-meat-2016-08-15