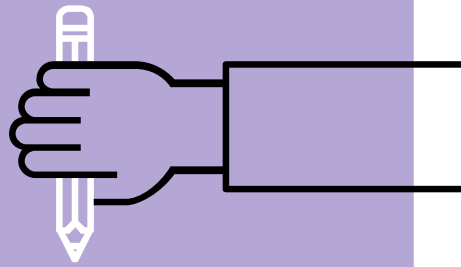
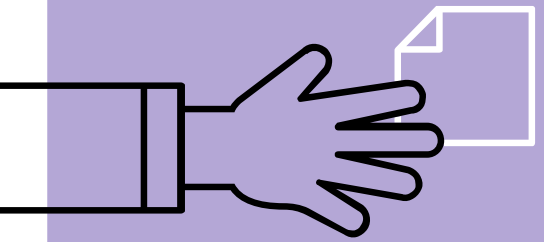
Heart Health and Recipe Recommendation



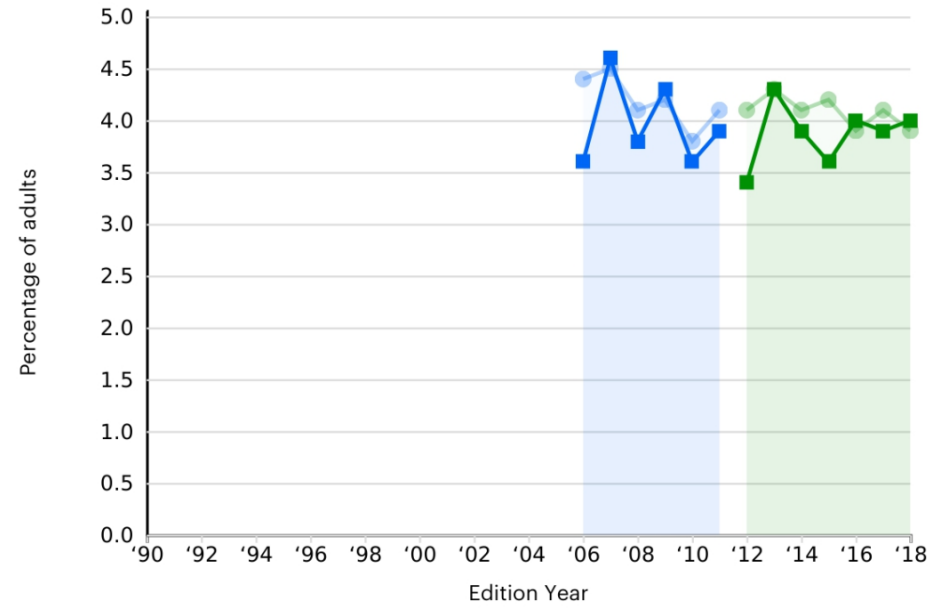


Business Case

- In the United States in 2019, coronary events are expected to occur in about 1,055,000 individuals, including 720,000 new and 335,000 recurrent coronary events. Benjamin EJ, Muntner P, Alonso A, et al. (2019 Jan). *Heart Disease and Stroke Statistics-2019 Update: A Report from the American Heart Association.* Retrieved from https://www.acc.org/latest-in-cardiology/ten-points-to-remember/2019/02/15/14/39/aha-2019-heart-disease-and-stroke-statistics
- The annual total cost of cardiovascular disease in the United States was estimated at $351.2 billion in 2014-2015, with $213.8 billion in direct cost, including 46% for inpatient care. Benjamin EJ, Muntner P, Alonso A, et al. (2019 Jan). *Heart Disease and Stroke Statistics-2019 Update: A Report from the American Heart Association.*.Retrieved from https://www.acc.org/latest-in-cardiology/ten-points-to-remember/2019/02/15/14/39/aha-2019-heart-disease-and-stroke-statistics
- A heart healthy diet can help prevent heart disease if such criteria are followed:
  - Eat more fruits and vegetables
  - Select whole grains
  - Limit trans and saturated fat intake
  - Choose low fat protein sources
  - Reduce sodium intake

Mayo Clinic Staff (2019 Jan). *Heart-healthy diet: 8 steps to prevent heart disease.* Retrieved from https://www.mayoclinic.org/diseases-conditions/heart-disease/in-depth/heart-healthy-diet/art-20047702

Trend: Heart Disease, Illinois, United States, 2018 Annual Report



SOURCE:

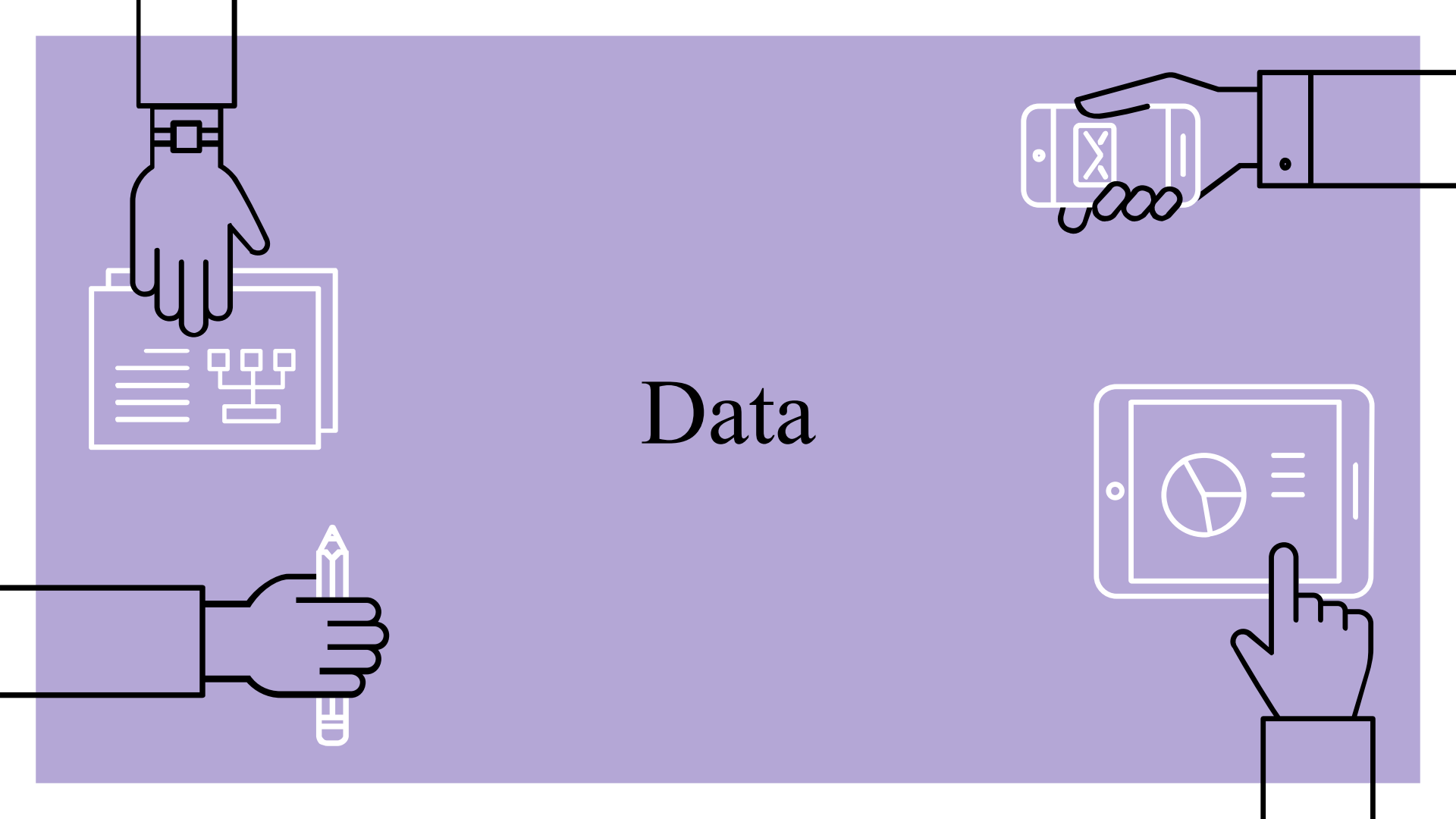
- CDC, Behavioral Risk Factor Surveillance System

[redacted]

Executive Summary

- Cardiovascular disease continues to be a leading cause of death in the U.S. but a healthy diet can help prevent heart disease
- The proposed model classifies if a recipe falls into a category which is likely to contribute to heart disease or if it is a healthy choice

Data



Data

- 3 separate Kaggle datasets were used:
  - Open Food Facts - 1 GB
    - Ingredient name and nutritional information columns
  - Epicurious Recipes - 12 MB
    - Recipe name and ingredient columns
  - BBC Good Food Christmas Recipes - 2.5 MB
    - Recipe name, description, author, ingredients and prep instructions
- Original data size was over 1 GB

Open Food Facts

	product_name	brands	countries_en	ingredients_text	serving_size	salt_100g	sodium_100g	vitamina_100g	vitaminc_100g
0	banana chips sweetened (whole)	nan	united states	bananas, vegetable oil (coconut oil, corn oil ...	28 g (1 onz)	0.0	0.0	0.0	0.021400001000000002
1	peanuts	torn & glasser	united states	peanuts, wheat flour, sugar, rice flour, tapio...	28 g (0.25 cup)	0.635	0.25	0.0	0.0
2	organic salted nut mix	grizzlies	united states	organic hazelnuts, organic cashews, organic wa...	28 g (0.25 cup)	1.22428	0.48200000000000004	2.0976923846153848e-05	0.000675000025

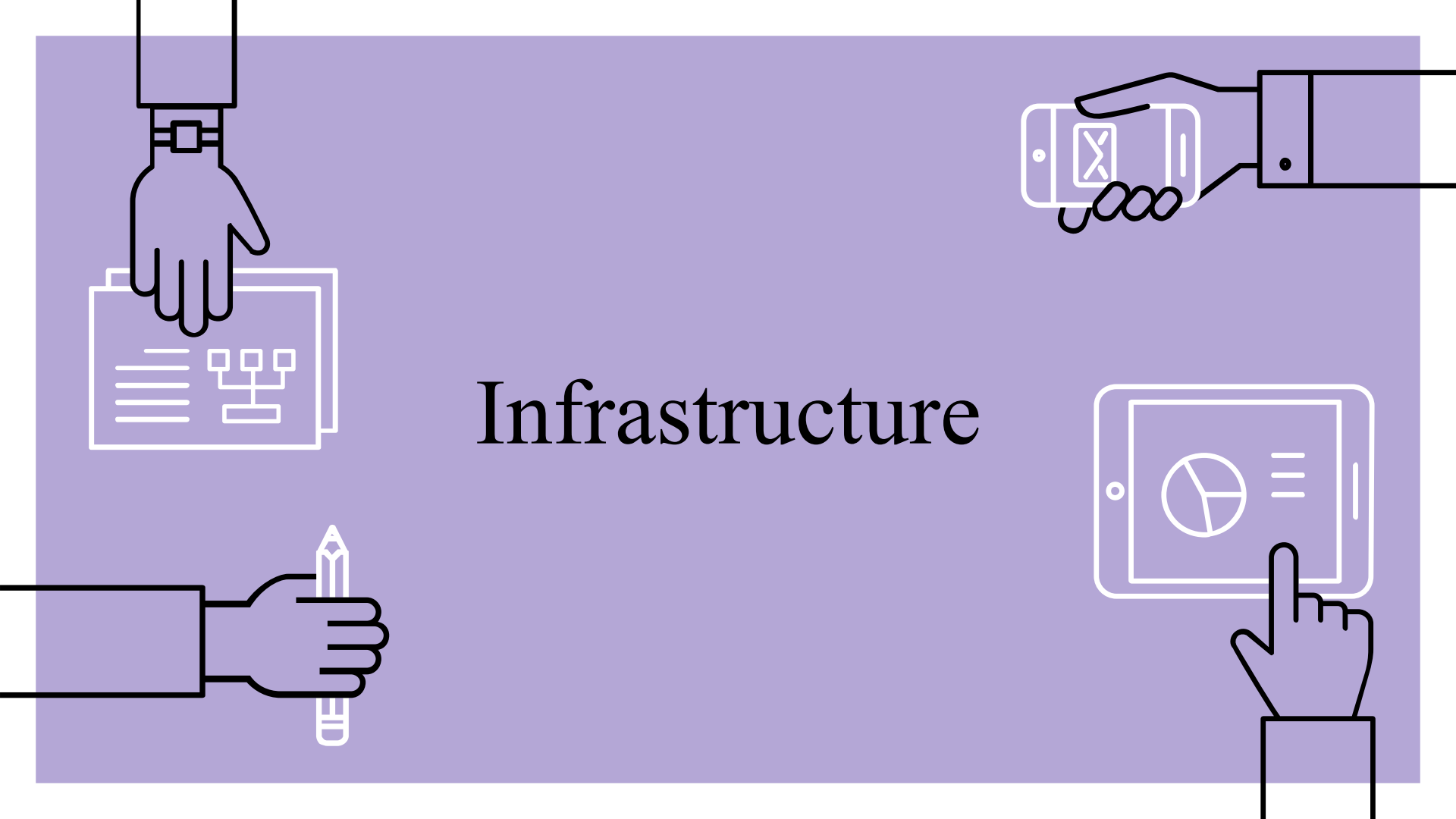
Epicurious Recipes

	title	Ingredients
0	Lentil, Apple, and Turkey Wrap	apple bean cookie fruit kid-friendly lentil le...
1	Boudin Blanc Terrine with Red Onion Confit	bake bastille day bon appétit chill dried frui...

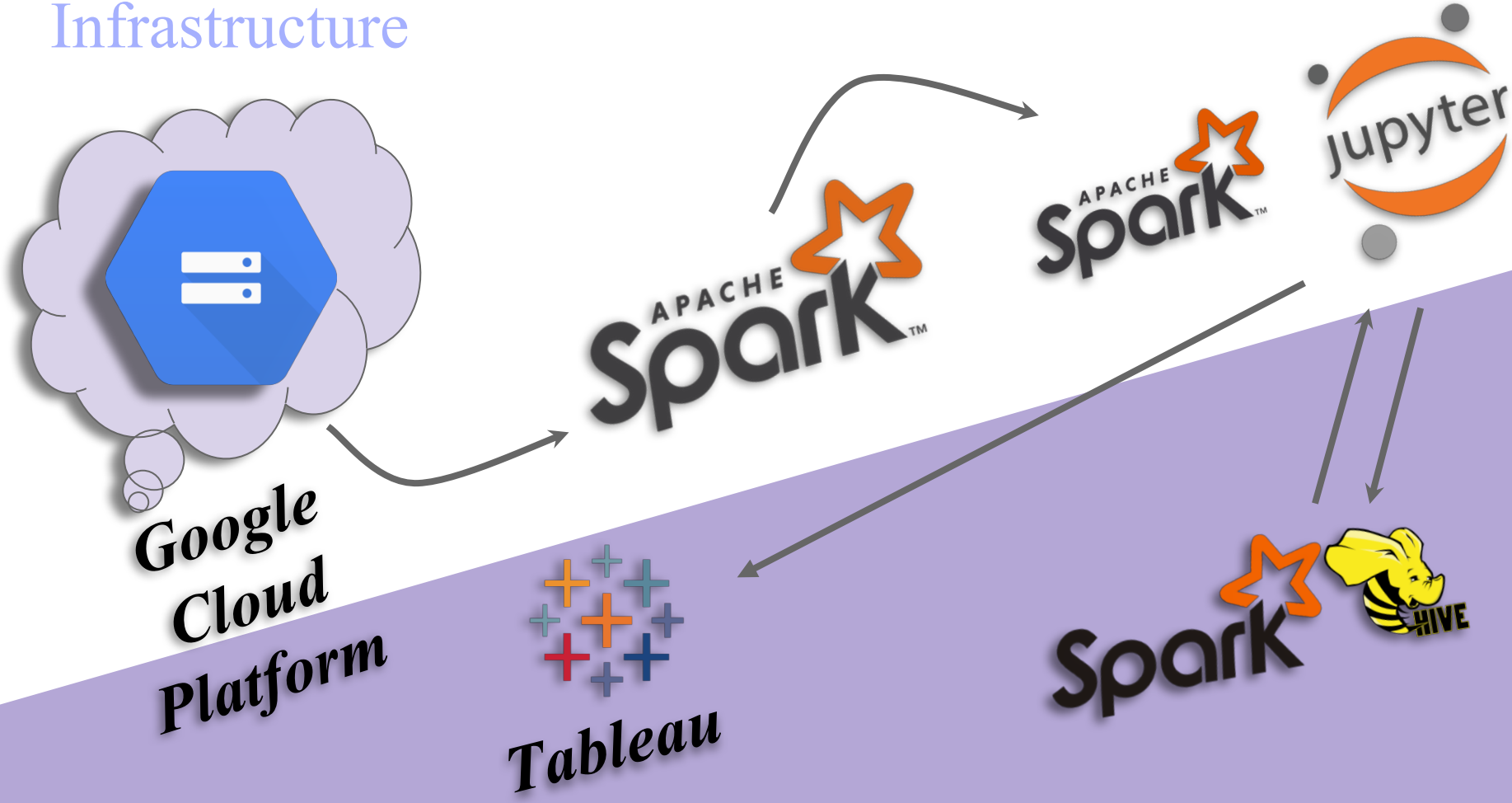
Christmas Recipes

	Author	Description	Ingredients	Method	Name	url
0	Mary Cadogan	Combine a few key Christmas flavours here to m...	[2 tbsp olive oil, knob butter, 1 onion, finel...	[Heat oven to 190C/fan 170C/gas 5. Heat 1 tbsp...	Christmas pie	https://www.bbcgoodfood.com/recipes/2793/chris...
1	Mary Cadogan	An easy-to-make alternative to traditional Chr...	[175g butter, chopped, 200g dark muscovado sug...	[Put the butter, sugar, fruit, zests, juice an...	Simmer-&-stir Christmas cake	https://www.bbcgoodfood.com/recipes/1160/simme...

Infrastructure



Infrastructure



- Started with :

  3 node Dataproc cluster

  1 master, 3 slaves

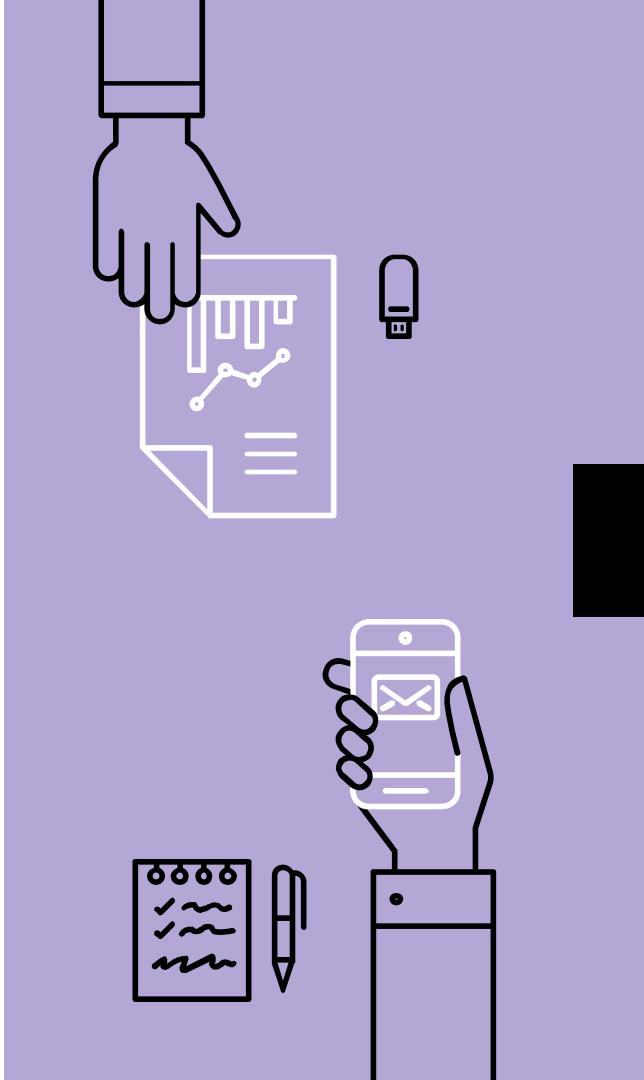
  2 vCPUs 7.5 GB memory

- Finished with:

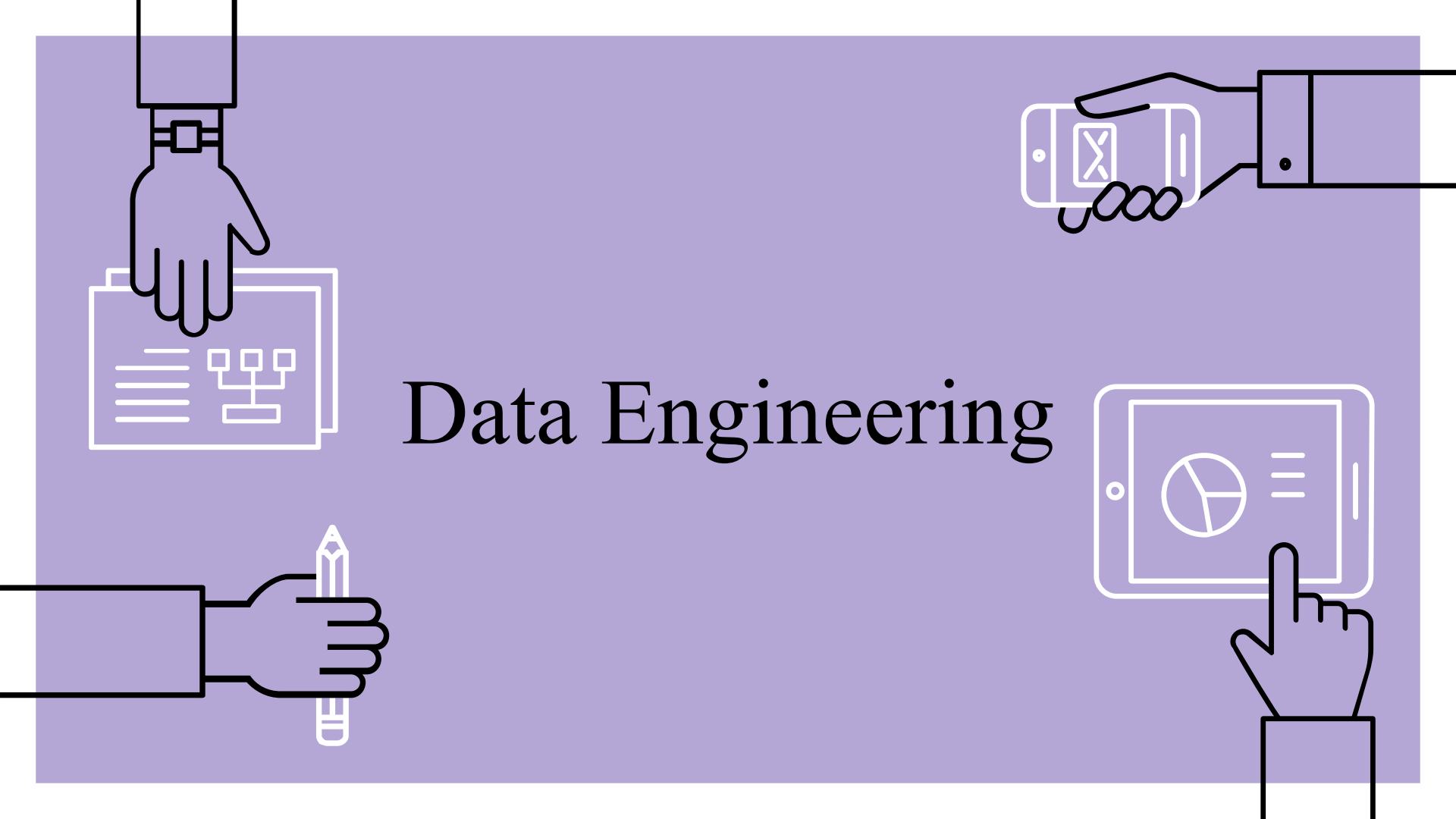
  4 node Dataproc cluster

  1 master, 4 slaves

  4 vCPUs 9 GB memory



Data Engineering

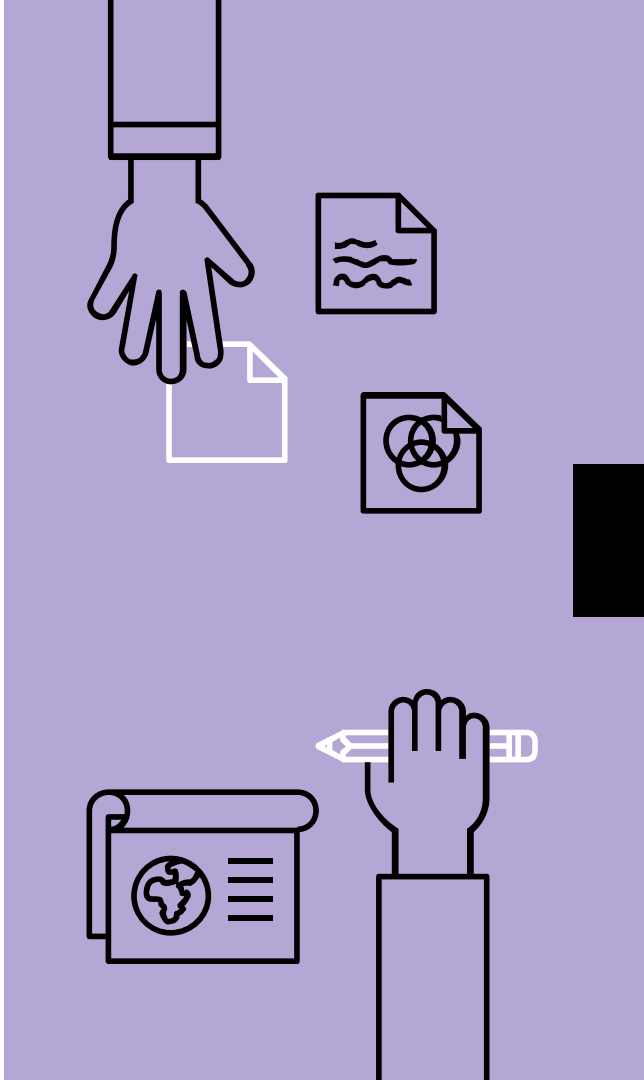


1. Cleaned data to normalize it as much as possible
    a. Removed unneeded columns
    b. Combined ingredient columns to have ingredient strings in both datasets
    c. Removed measurement information from ingredient column
    d. Removed hyphens, parentheses and other characters which may interfere with code

1. Exploded ingredient strings in both recipe datasets and created a list of top 20 ingredients to work with and use as keys

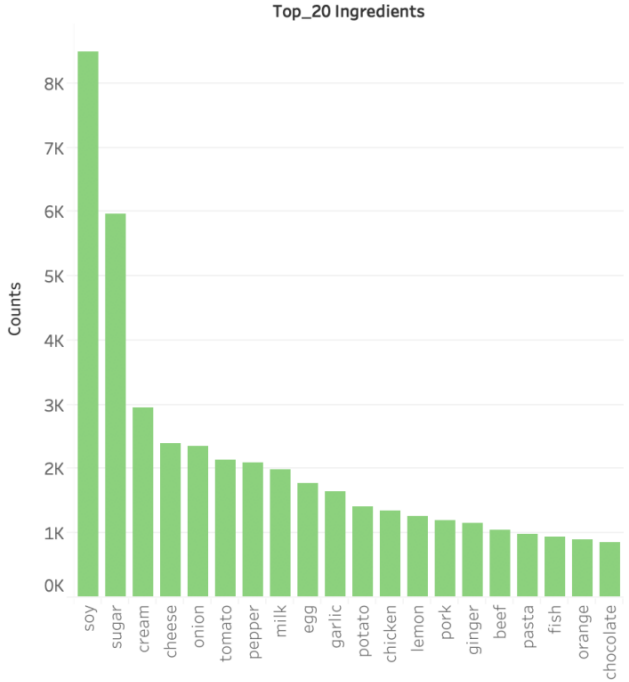
```
gen_rec_count=df_regrecipes.select(explode(split(col("Ingredients"), "\s+"))\
                           .alias("gen_rec_ing")).groupBy('gen_rec_ing').count()\
                           .orderBy("count", ascending=False)
```

1. Recipe and nutrient table ingredient names are not an exact match i.e. Recipe ingredient is “tomato”, while nutrient table ingredient is “heirloom tomato”

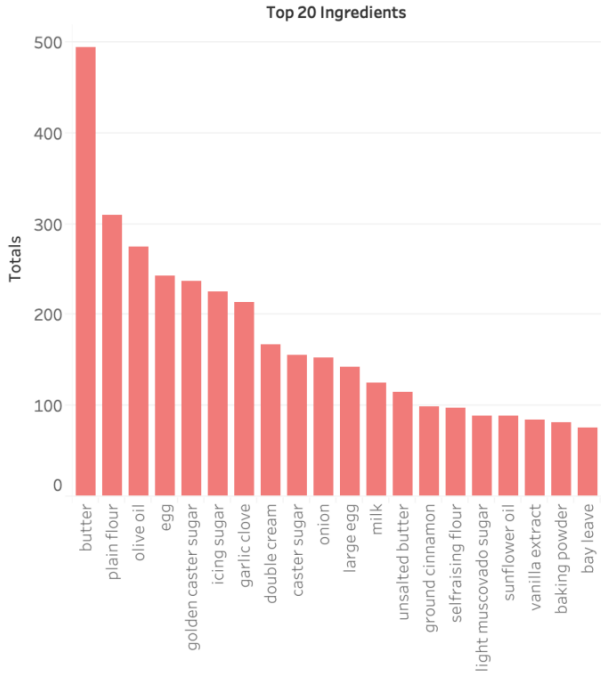


Top 20 Ingredients per Recipe Group

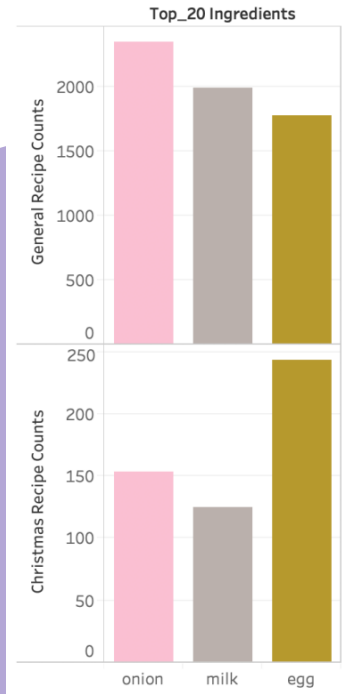
General Recipe Ingredients



Christmas Recipes Ingredients



Overlapping Ingredients



- Match appropriate key list to each recipe dataframe
- Match nutrients dataframe in Hive to 2 key sets, created 2 separate tables
- Match recipe dataframe and nutrient dataframe by new key columns

	title	Ingredients
0	Lentil, Apple, and Turkey Wrap	apple bean cookie fruit kid-friendly lentil le...
1	Boudin Blanc Terrine with Red Onion Confit	bake bastille day bon appétit chill dried frui...

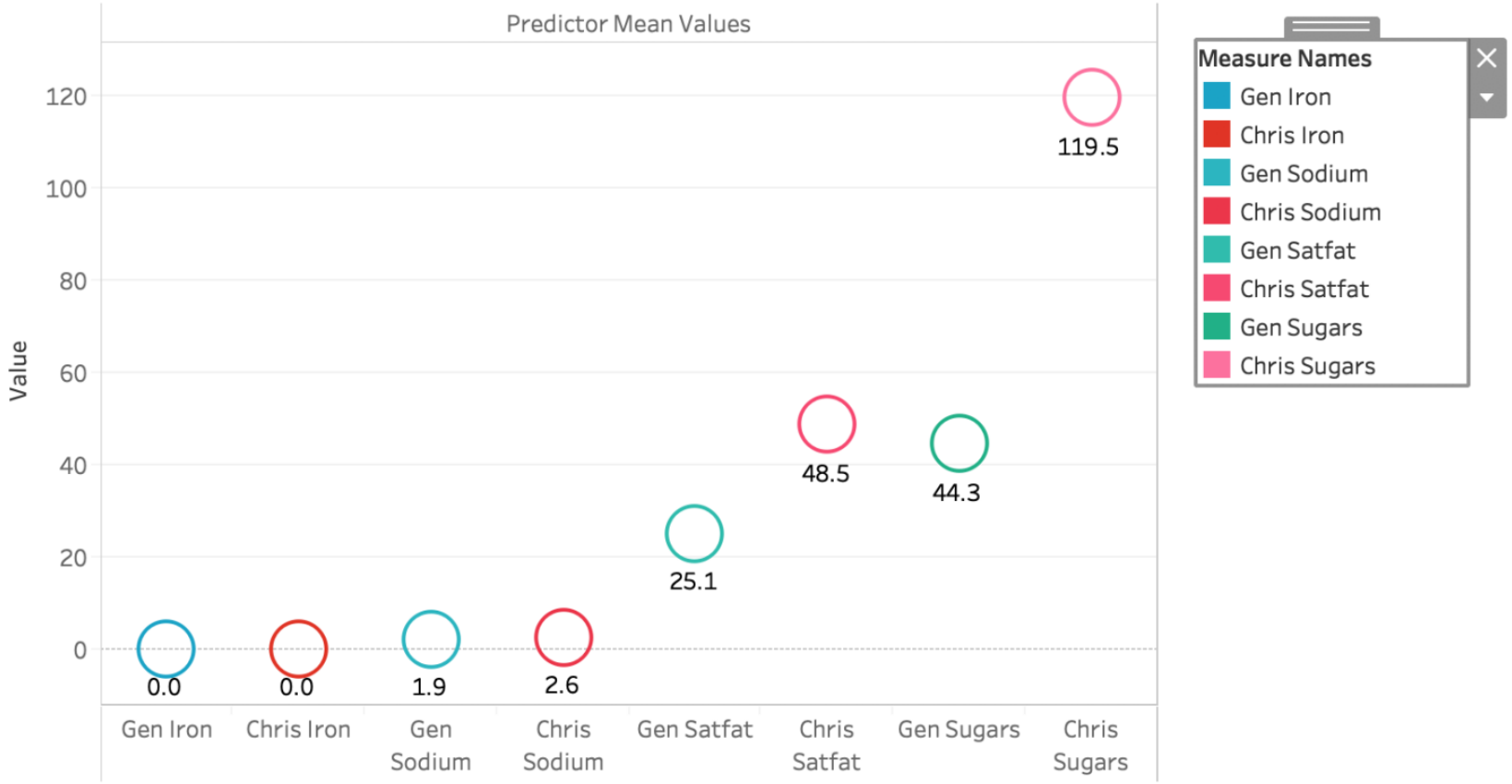
Key
apple
orange

	product_name	brands	countries_en	ingredients_text	serving_size	serving_quantity	num_additives	additives	ingredients_from_palm_oil_n
0	banana chips sweetened (whole)	nan	united states	bananas, vegetable oil (coconut oil, corn oil ...	28 g (1 onz)	28.0	0.0	[bananas -> en:bananas] [vegetable-oil -...	0.0
1	peanuts	torn & glasser	united states	peanuts, wheat flour, sugar, rice flour, tapio...	28 g (0.25 cup)	28.0	0.0	[peanuts -> en:peanuts] [wheat-flour -> ...	0.0

Key
banana
peanut

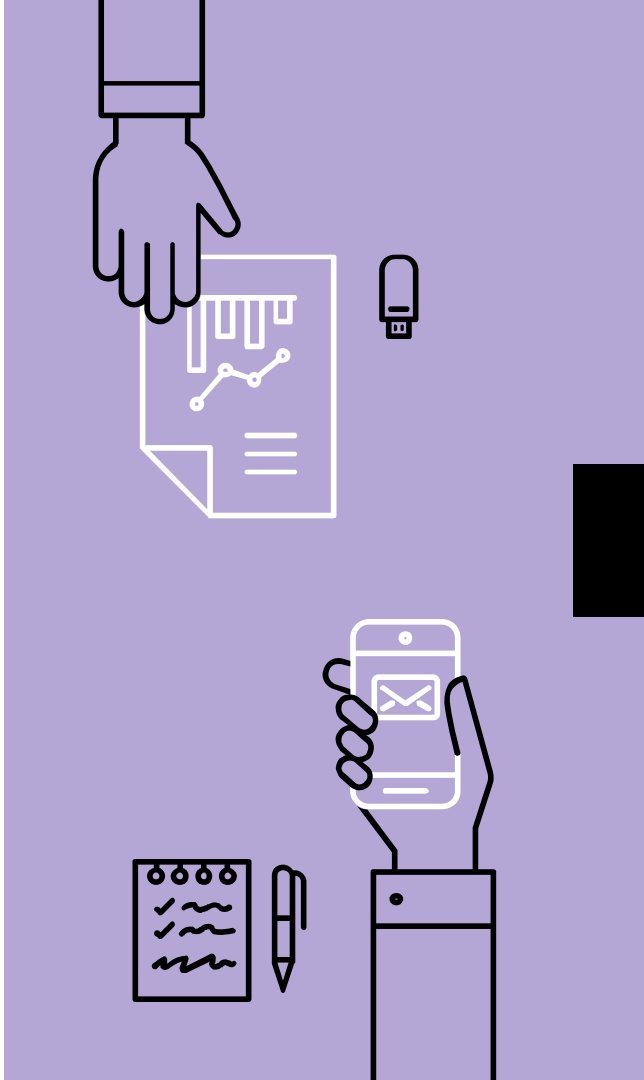
Name	energy	fat	saturatedfat	transfat	cholesterol	carbs	sugars	fiber	protein	salt	sodium	vit_a	vit_c	calcium	iron
Reindeer food	6619.0	100.0	21.247308	0.0	0.0075	188.33	155.00	8.242308	0.00	0.29718	0.117	0.000536	0.002131	0.094462	0.001957
Mashed peppered roots with toasted hazelnuts	3246.0	81.9	52.080000	0.0	0.2290	5.00	24.58	1.000000	3.33	1.76022	0.693	0.000895	0.031800	0.227667	0.005167

Summary Stats



Correlation with Most Important Feature

	Type	Correlation_with_Sugar
5	transfat	0.067062
9	vit_a	0.092847
2	fat	0.172141
8	sodium	0.210899
0	sugars	0.246940
11	calcium	0.248958
7	fiber	0.326225
3	cholesterol	0.393770
10	vit_c	0.402007
4	satfat	0.419790
1	protein	0.437222
12	iron	0.465473
6	carbs	0.797498



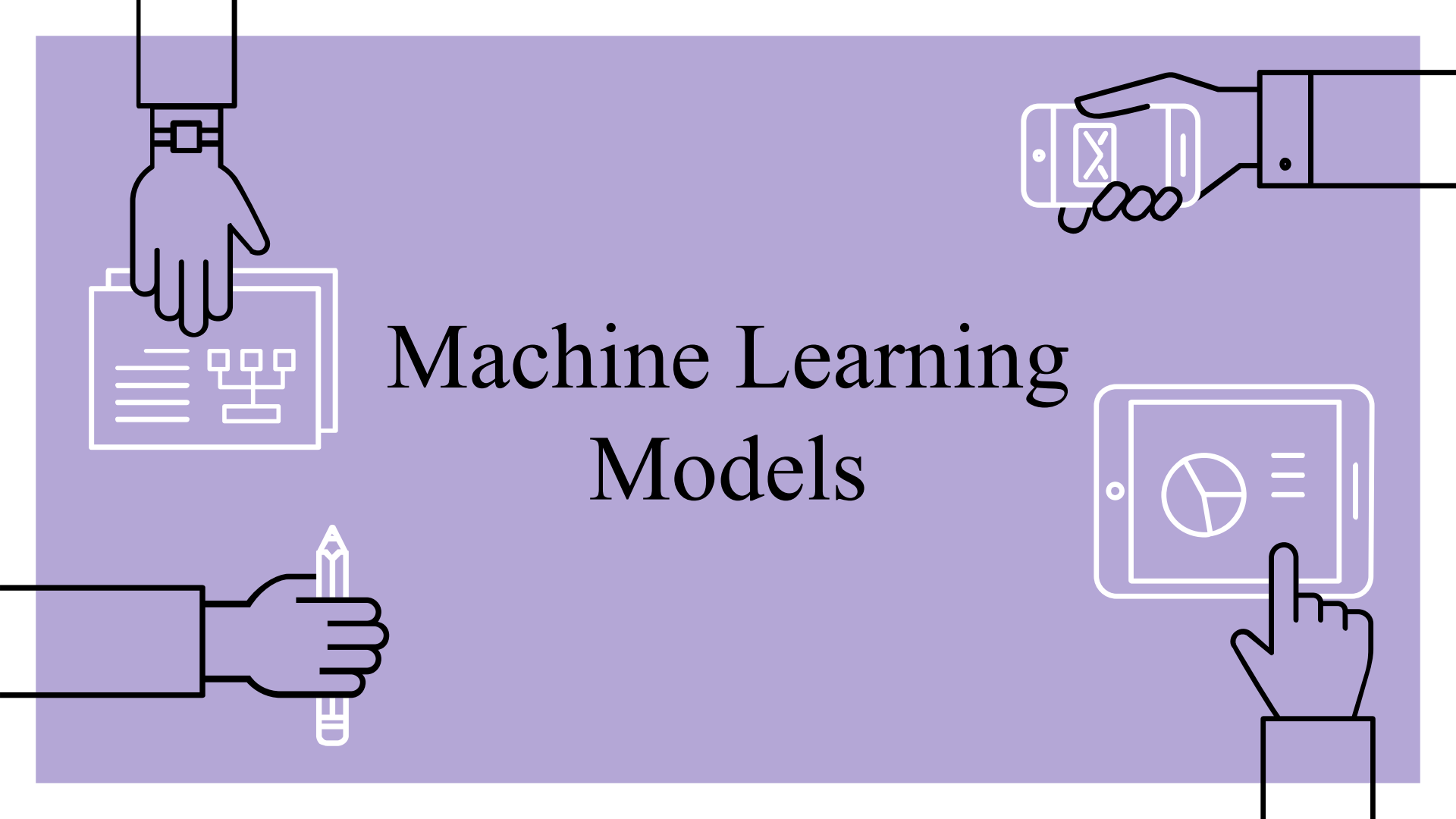
- Normalize each dataframe based on mean of each feature

	summary	fat	satfat	transfat	cholestorol	carbs	sugars
0	count	105	105	105	105	105	105
1	mean	242.53985817318872	27.23674109322684	0.46485713322957356	0.16049841784295582	71.57276186261858	41.048565383184524
2	stddev	467.71776299829077	9.120375225722576	1.999920711336166	0.06513498998330118	37.01094258914221	24.650778803474925
3	min	35.560001373291016	18.529998779296875	0.0	0.12800000607967377	31.25	18.415000915527344
4	max	2238.31005859375	56.029998779296875	18.75	0.421999990940094	204.8199920654297	144.0050048828125

- Combined recipe dataframes and new column to identify where the recipe is coming from

	title	energy	fat	satfat	transfat	cholesterol	carbs	sugars	fiber	protein	salt	sodium	vit_a	vit_c	calcium	iron	rec_type
0	Roast turkey & cranberry Wellington	11670.5	1	1	0	1	1	0	1	1	1	1	0	1	0	1	c
1	Roast turkey breast wrapped in bacon	4423.0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	c
2	Roast turkey with chestnut stuffing	9217.0	1	1	0	1	0	1	0	0	1	1	0	1	0	1	c
3	Roast turkey with citrus butter	8740.0	1	1	0	1	0	1	1	0	0	0	0	1	0	1	c
4	Roast turkey with lemon & garlic	8380.0	1	1	0	1	0	1	0	0	0	0	0	1	0	1	c

Machine Learning Models



Data Processing Pipeline

	title	energy	fat	satfat	transfat	cholesterol	carbs	sugars	fiber	protein	salt	sodium	vit_a	vit_c	calcium	iron	rec_type
0	Roast turkey & cranberry Wellington	11670.5	1	1	0	1	1	0	1	1	1	1	0	1	0	1	c
1	Roast turkey breast wrapped in bacon	4423.0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	c
2	Roast turkey with chestnut stuffing	9217.0	1	1	0	1	0	1	0	0	1	1	0	1	0	1	c
3	Roast turkey with citrus butter	8740.0	1	1	0	1	0	1	1	0	0	0	0	1	0	1	c
4	Roast turkey with lemon & garlic	8380.0	1	1	0	1	0	1	0	0	0	0	0	1	0	1	c

- Drop “title”, put “energy” into three buckets, convert recipe_type into numeric

	fat	satfat	transfat	cholesterol	carbs	sugars	fiber	protein	salt	sodium	vit_a	vit_c	calcium	iron	energy_category	general_recipe
0	1	1	0	1	1	1	1	1	0	0	0	1	0	1	2.0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0
2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0.0	1
3	0	1	0	0	1	0	0	0	1	1	0	0	0	0	1.0	0
4	1	1	0	1	1	1	1	1	1	1	0	1	0	1	2.0	0

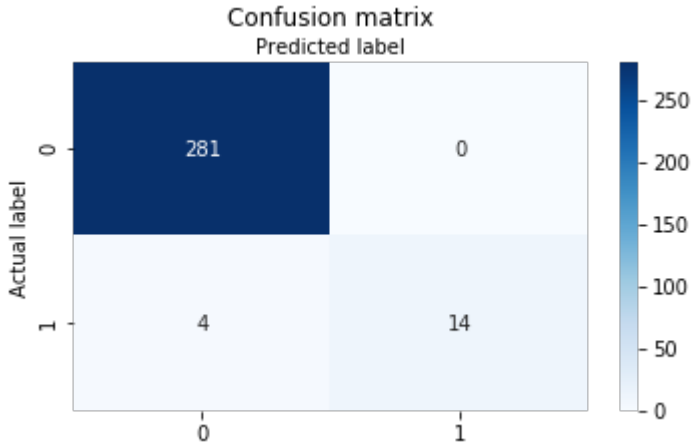
Data Processing Pipeline

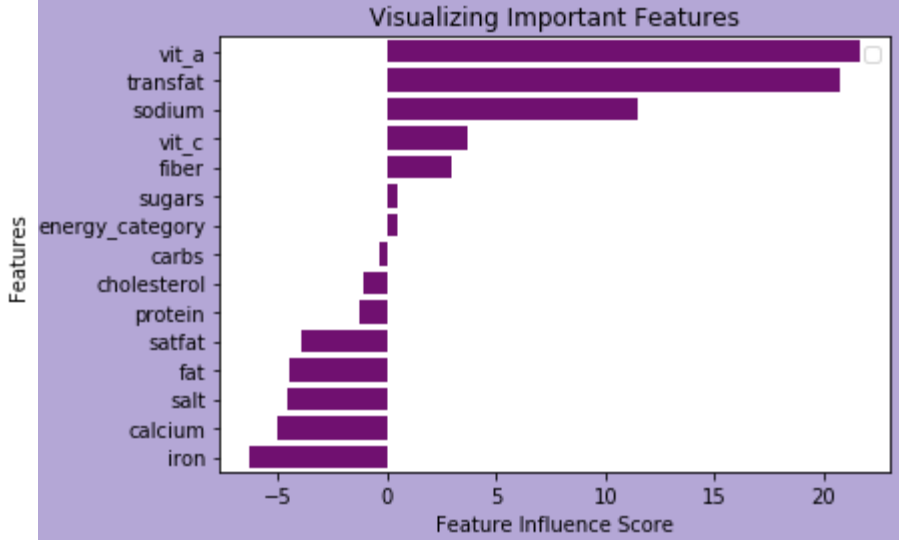
	title	energy	fat	satfat	transfat	cholesterol	carbs	sugars	fiber	protein	salt	sodium	vit_a	vit_c	calcium	iron	rec_type
0	Roast turkey & cranberry Wellington	11670.5	1	1	0	1	1	0	1	1	1	1	0	1	0	1	c
1	Roast turkey bre wrapped n bacon	4423.0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	c
2	Roast turkey with estnut ffing	9217.0	1	1	0	1	0	1	0	0	1	1	0	1	0	1	c
3	Roast turkey wit citrus b r	8740.0	1	1	0	1	0	1	1	0	0	0	0	1	0	1	c
4	Roast turkey lemon & g	8380.0	1	1	0	1	0	1	0	0	0	0	0	1	0	1	c

	fat	satfat	transfat	cholesterol	carbs	sugars	fiber	protein	salt	sodium	vit_a	vit_c	calcium	iron	energy_category	general_recipe
0	1	1	0	1	1	1	1	1	0	0	0	1	0	1	2.0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0
2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0.0	1
3	0	1	0	0	1	0	0	0	1	1	0	0	0	0	1.0	0
4	1	1	0	1	1	1	1	1	1	1	0	1	0	1	2.0	0

Logistic Regression

- Fit with all 15 predictors





Logistic

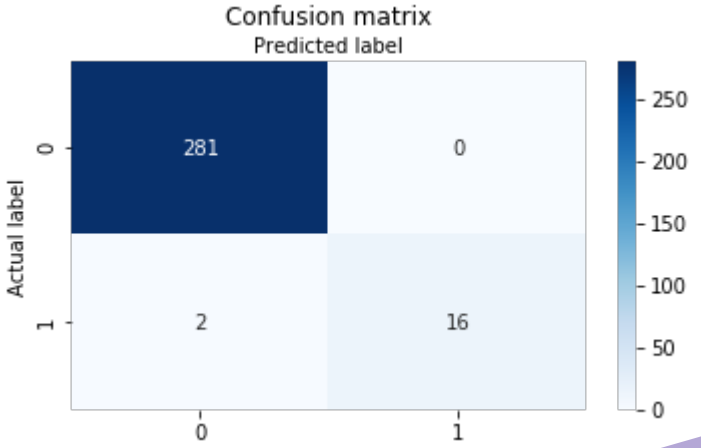
```
evaluator = MulticlassClassificationEvaluator(labelCol="general_recipe", predictionCol="prediction")

print(evaluator.evaluate(lg_predictions, {evaluator.metricName: "accuracy"}))
print(evaluator.evaluate(lg_predictions, {evaluator.metricName: "f1"}))
```

```
0.9866220735785953
0.985833225002068
```

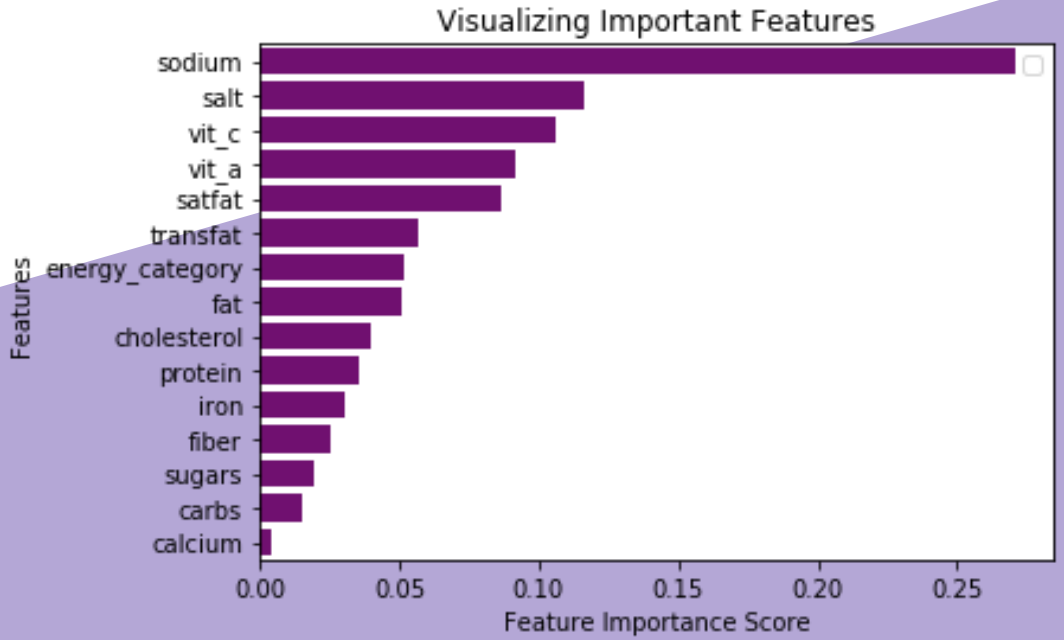
Random Forest

- Fit with all 15 predictors and feature exploration



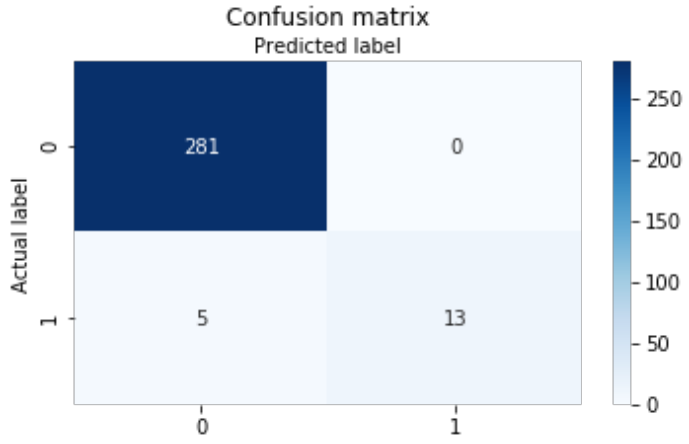

```
print(evaluator.evaluate(rf_predictions, {evaluator.metricName: "accuracy"}))
print(evaluator.evaluate(rf_predictions, {evaluator.metricName: "f1"}))
```

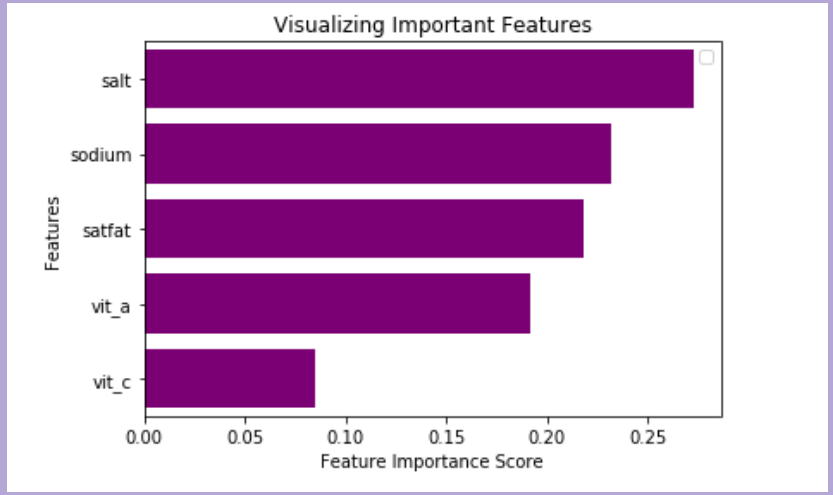
```
0.9933110367892977
0.9931261624410669
```



Random Forest

- Fit with top 5 predictors and feature exploration



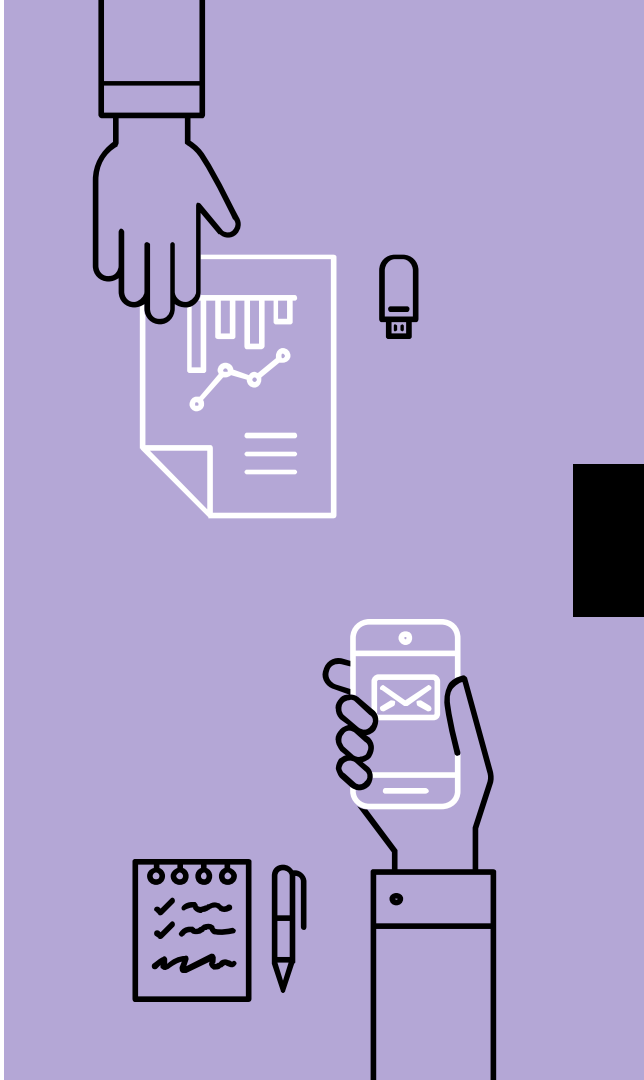



```
print(evaluator.evaluate(top_5_rf_predictions, {evaluator.metricName: "accuracy"}))
print(evaluator.evaluate(top_5_rf_predictions, {evaluator.metricName: "f1"}))
```

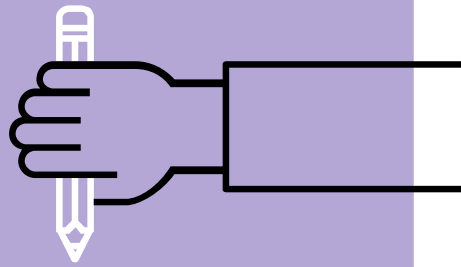
```
0.9832775919732442
0.9820027426385538
```

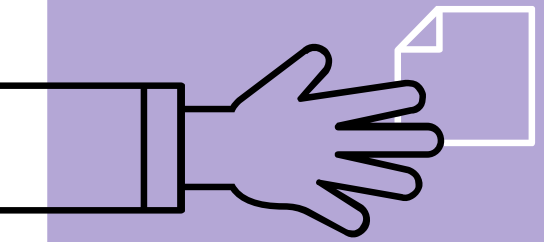
Classification Accuracy

Model	Number of Features	Accuracy	F1-Score
Logistic Regression	15	98.66	98.58
Random Forest Classifier	15	99.33	99.31
Random Forest Classifier	5	98.33	98.20



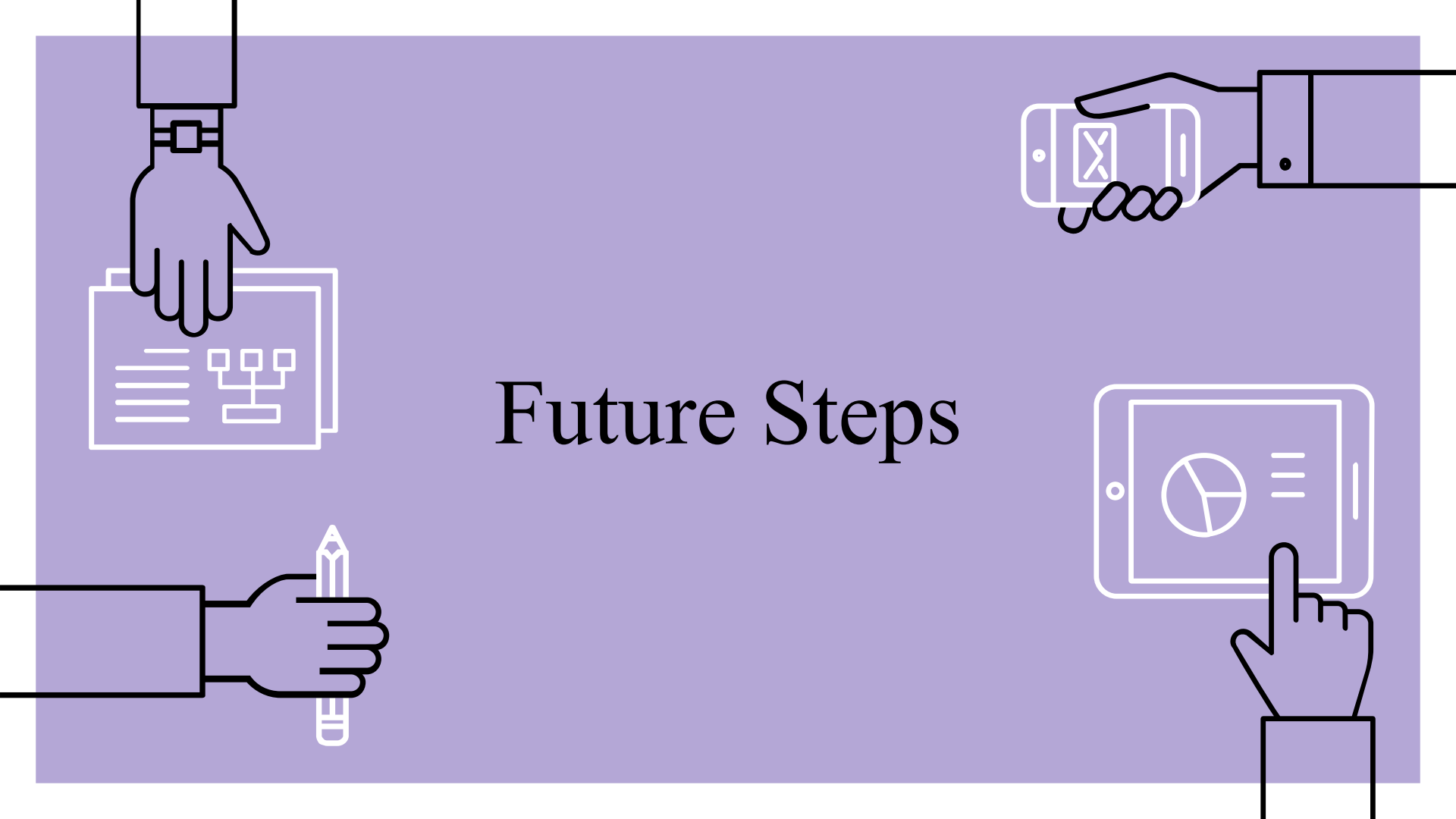
Recipe Classification Webpage



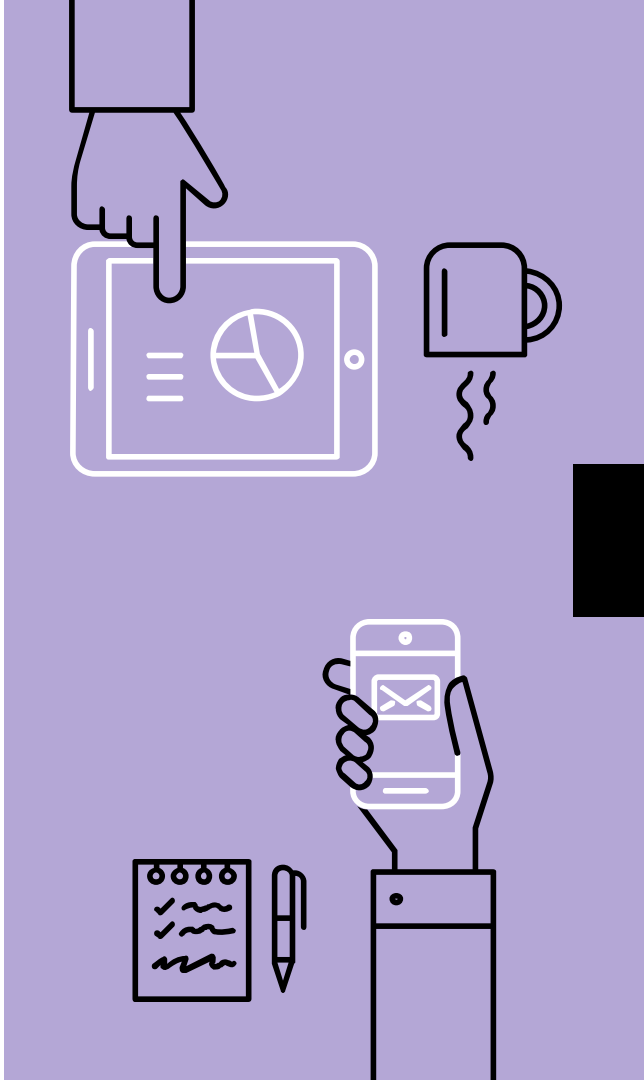


http://myprojecthome.x10host.com

Future Steps



- Recipe nutrition was calculated only based on top 20 ingredients appearing within the recipe dataset:
  - Calculate more accurate nutritional values by using all listed ingredients
- The top match to ingredient name was used from nutrition table:
  - Try for more exact matches to provide better nutritional value estimate
- Used only a general recipe and christmas datasets:
  - Acquire other holiday recipe datasets (ex. Thanksgiving) and see what kind of prediction difference this would make
- Link model to website interface to allow real-time prediction of new recipe type



References

Research:

https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/heart-disease-and-food
https://www.mayoclinic.org/diseases-conditions/heart-disease/in-depth/heart-healthy-diet/art-20047702
https://www.acc.org/latest-in-cardiology/ten-points-to-remember/2019/02/15/14/39/aha-2019-heart-disease-and-stroke-statistics
https://www.americashealthrankings.org/explore/annual/measure/CHD/state/IL

Data:

Open Food Facts
https://www.kaggle.com/openfoodfacts/world-food-facts
General Recipes
https://www.kaggle.com/hugodarwood/epirecipes
Christmas Recipes
https://www.kaggle.com/gjbroughton/christmas-recipes