Building a Nutritional Knowledge Graph to recommend customized food choices

Presented by:

Niranjan Navaneethan
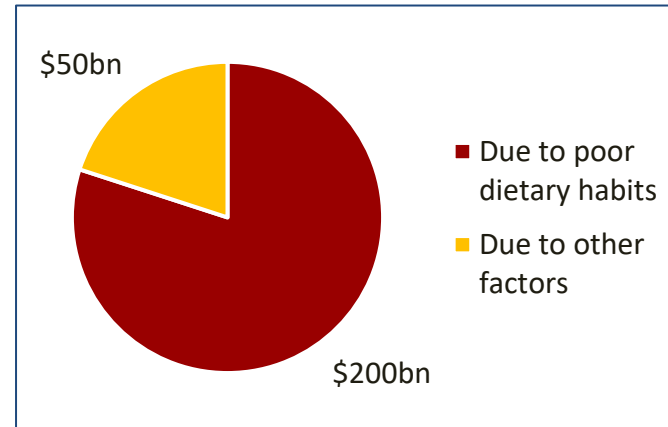
Parth Shah

Motivation

How many dollars are we losing a year towards heart diseases, strokes, and diabetes?

Total cost in healthcare = $250bn

$50bn

$200bn

- Due to poor dietary habits
- Due to other factors

What is the most common reason stated for difficulty following diets?

Lack of convenience in accessing the right knowledge & information: 78% of people tend to get confused about food choices due to conflicting pieces of information

Objective

Build a Nutritional Knowledge Graph (Nutri-KG) to marginally change dietary options by providing smart recommendations in terms of nutrition and flavor

Why is a Knowledge Graph helpful to solve this?

Nutri-KG will allow us to find combinations of food that share similar flavors and identify relationships with their nutritional information, thereby enabling the user to make incremental health choices.
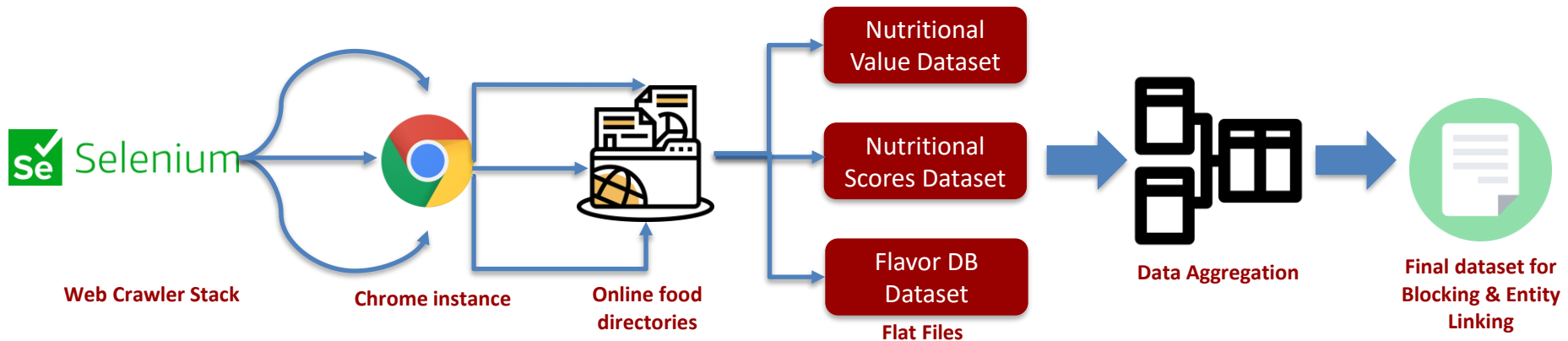
Solution Design
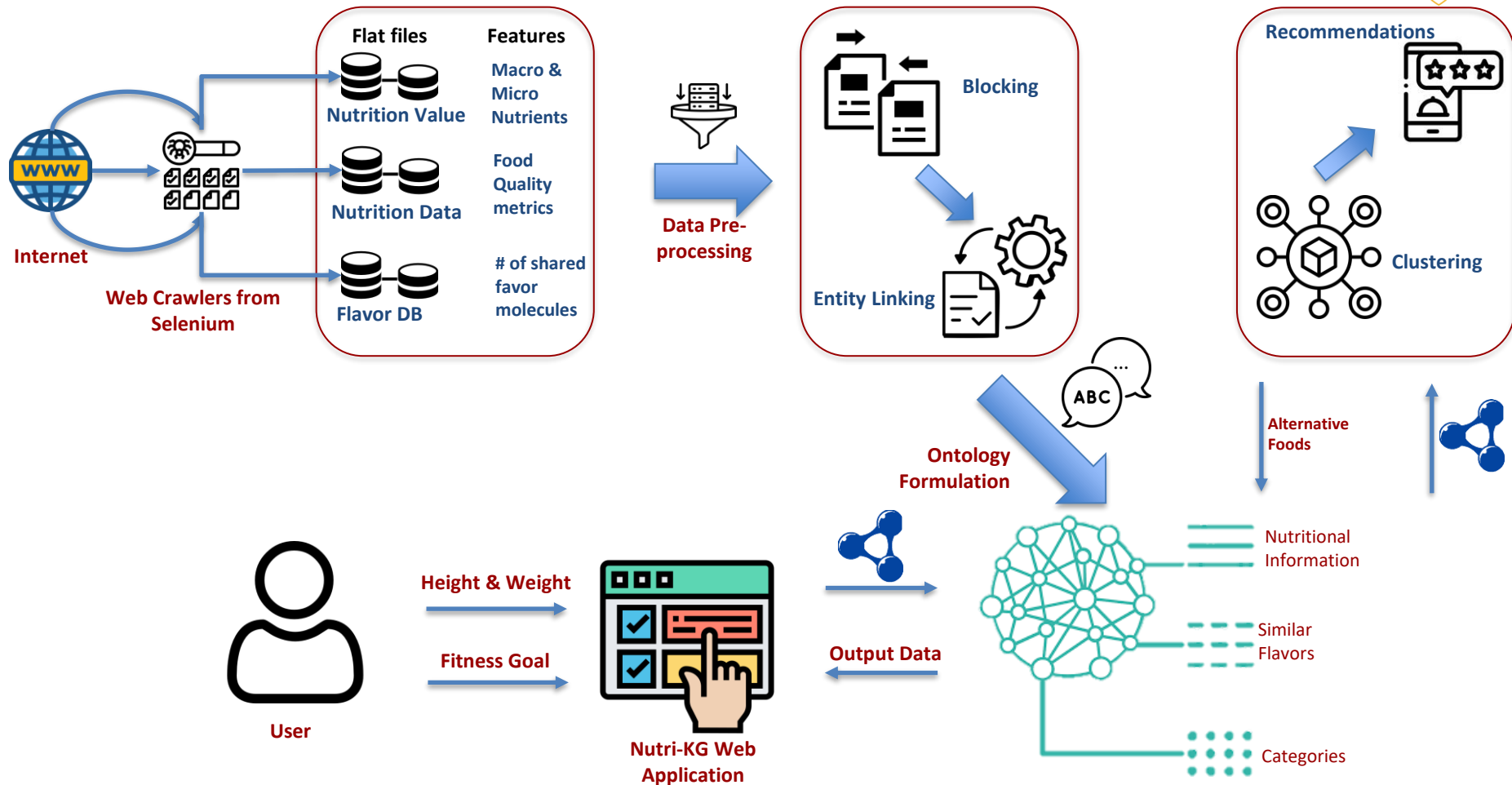
Data Extraction – Sources & Process Flow

Sources	Data Format	Features	Application for Extracted Data
Nutrition value	Dynamic webpage with semi-structured data	- Category - Protein - Carbs - Fats - Vitamin A,D,C - Sodium	This will serve us as base fields to be used for the recommender system based on user input
NutritionData know what you eat	Unstructured Data stored deep in the food directory	- Glycemic Load - Completeness Score - Amino Acid Score	This provides enriching features which improve recommendations as some foods might have good macros but affect the glycemic load on the body
FlavorDB	1000 webpages with multi-page structure data	- Combination of Food - Number of shared Flavor molecules	This will serve us as the links between similarly flavored foods and allow us to suggest healthier options with a similar flavor

Process Flow

Web Crawler Stack — Selenium

Chrome instance

Online food directories

Nutritional Value Dataset

Nutritional Scores Dataset

Flavor DB Dataset

Flat Files

Data Aggregation

Final dataset for Blocking & Entity Linking

Data Extraction - Issues, Solution and Summary

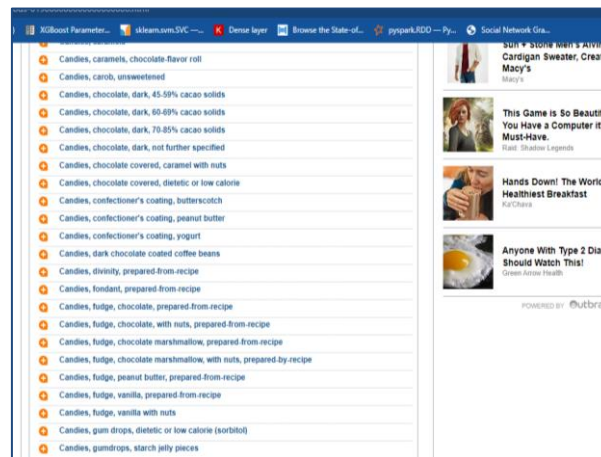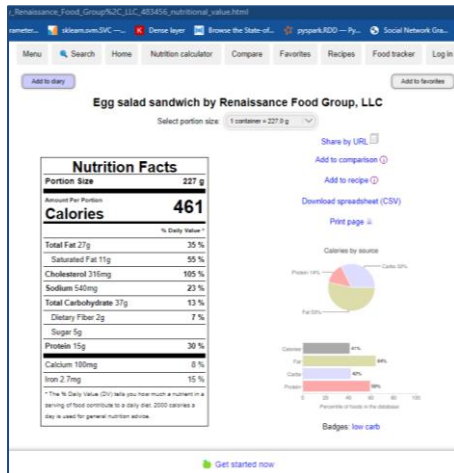| Issue	Description	Solution
Website Blocks Crawlers and reports 503	- Highly dynamic website with pop-ups. No patterns in URLs as well - Website had server request limits which were regularly triggered	- Dynamic Crawlers to handle all pop-ups and link navigation - Random downtimes were introduced to mimic human behavior
Data Source KG Inaccessible	- The flavor data was meant to be retrieved from Flavor KG, which was not maintained and the data was not accessible	- Alternate data source of Flavor DB was used as it met the requirements and the data was regularly updated

Summary on Extracted Data

Source	Total # of webpages accessed	Total # of data points	Average Missing Value % per Feature
Nutrition Value	75k	56k	<5%
Nutrition Data Self	40k	36k	≈30%
Food DB	≈1000	950	<5%

USC Viterbi
School of Engineering

University of Southern California

Data Extraction - Snapshots

How did it start?

What did we end up with?

Entity Linking – Process Flow

What did we take into account for our dataset before linking them?

Issue	Description	Implication
Missing data	**Imperfect scraping or source leads to missing data in macro/micro nutrient information**	**Incorrect macro/micro nutrient information**
Large data	**Two sources yielded 50K+ lines each**	**Large runtimes due to O(n^2) algorithm and large dataset**
Ambiguity	**Food names are close, may differ just by portion size. (Eg, Apple Juice 32oz)**	**Incorrect linking of entities**

Process Flow

Processed data from multiple sources

Set of Food combinations
Food A, Food B
Food C, Food D
Food B, Food D

Jaccard Similarity

Threshold = 0.2

Text Processing: Cleaning, Removal of numbers

Levenshtein Similarity — **Threshold = 0.85**

String Matching Score — **Threshold = 0.8**

Similar Entity: Score > 0.85? — **Yes** → **Link Entities**
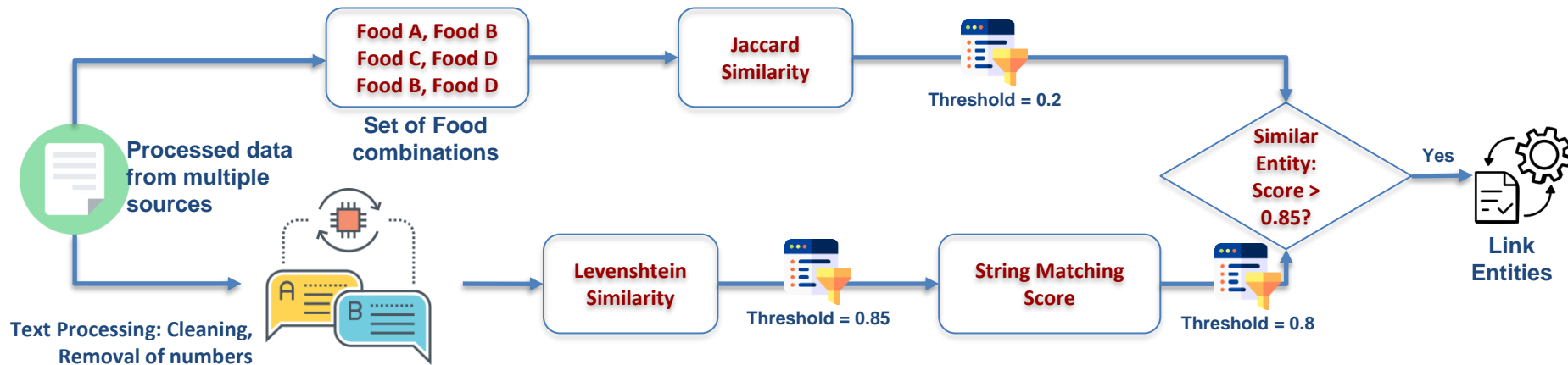
Entity Linking – Challenges and Solution

Approach to Challenges

Issue	Solution	Example
Missing data	**Take data from other source after verifying portion size**	`doritos cool ranch 8z by frito lay,,19.88,150.0,18.0,,8.0,0.0,15.96,,2.0,1.0,180.04,,0.08,,,0.0,,,,,,`
Large data	**Blocking with name length; label created with +/- 2 of length (Eg. 4-8)**	**doritos cool ranch 8z by frito lay of length 7 is labelled 5-9**
Ambiguity	**Keep numbers in name, run range check on a common field**	**doritos cool ranch 8z by frito lay doritos cool ranch 16z by frito lay**

Evaluation of Linking

Issue	Basic	Improved
Missing data	113K across 19 fields in 50K entities	85K across 19 fields in 50K entities
Large data	18 mins (10K entities)	2 mins (10K entities) x9 improvement!
Ambiguity	2K entities not linked	475 entities not linked

Ontologies and RDF Construction

What are the different ontologies used and their purpose?

Ontology Source

https://dsci558.org/Flavor

https://dbpedia.org/ontology/Food

https://schema.org/NutritionInformation

https://dsci558.org/FoodCategory
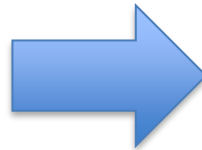
Properties & Relations

flavorOf, numOfMolecules

approximateCalories, glycemicIndex

calories, carbohydrateContent, proteinContent, fatContent...

has

RDF Snapshot

Flavor Triple

```
food:candies-hersheys-skor-toffee-bar flavor:flavorOf food:candies-preparedfromrecipe-toffee,
        food:crunch-n-munch-buttery-toffee-popcorn-with-peanuts-6-oz-by-conagra-brands .
```

Nutritional Triple
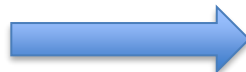
```
food:babyfood-junior-squash-vegetables a food:title ;
    food:name "babyfood junior squash vegetables"^^xsd:string ;
    nutriinfo:aminoAcidScore "NaN"^^xsd:float ;
    nutriinfo:calories "3.8"^^xsd:float ;
    nutriinfo:completenessScore "76.0"^^xsd:float ;
    nutriinfo:fat "0.032"^^xsd:float ;
```

Category Triple
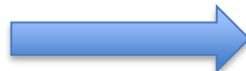
```
foodcategory:Foundation-Foods a foodcategory:title ;
    foodcategory:has food:beef-jellied-luncheon-meat-cured,
        food:beef-smoked-cured-chopped,
        food:beets-solids-and-liquids-canned-pickled,
```

Construction of Nutri-KG – Process Flow

Process Flow

Text Processing Module: Cleaning, Removal of numbers

CSV Data from Source 1

CSV Data from Source 2

Flavor DB

RLTK Records

Blocking

Record Linkage

Inverted Index of Flavor Data

Triples: .ttl file

Nutritional Information

Similar Flavors

Categories

Defined namespaces and available ontologies
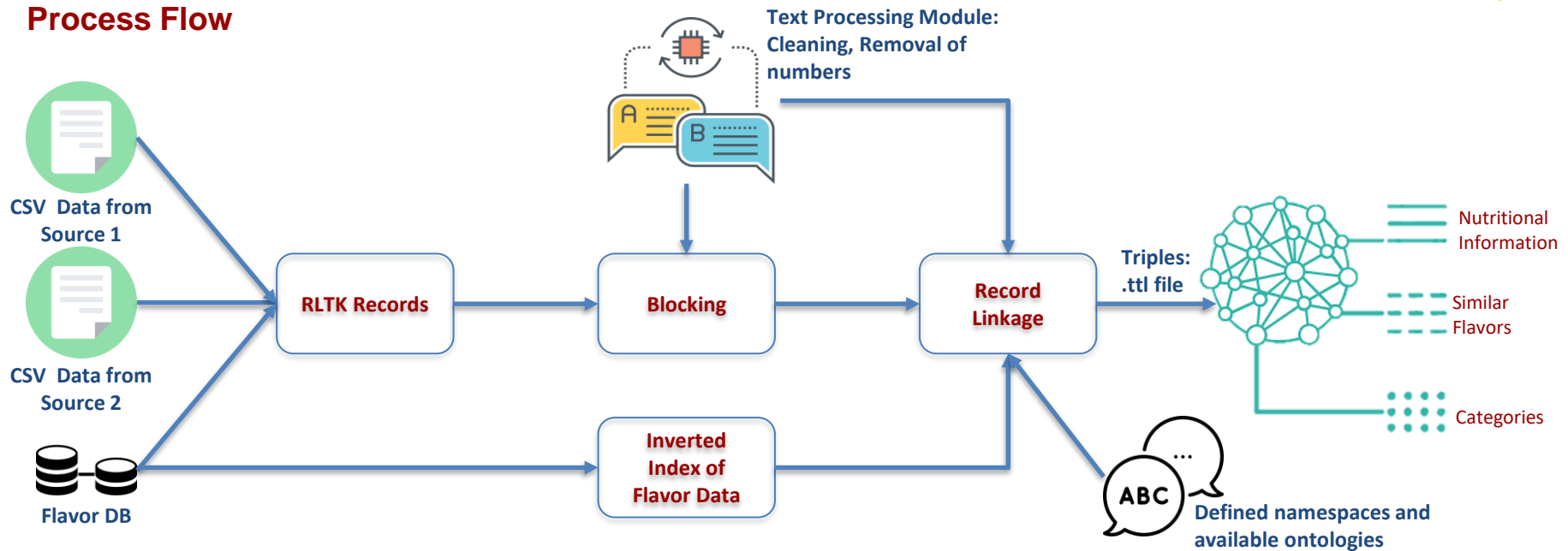
Nutri-KG construction environment using KG-LAB
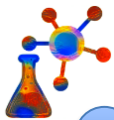
KG Lab provides a simple abstraction layer in Python for building Knowledge Graphs. This allows us to scale the solution comfortably as well. Given below are the reasons -

1 pandas NumPy — Data science workflows – close integration with the python data analysis stack; pandas, numpy, scikit-learn, matplotlib

2 kafka Spark — Distributed systems infrastructure – integrate with big data tools for data engineering/cloud computing infrastructure

Construction of Nutri-KG – Evaluation & Snapshot

Quality of Nutri-KG

- Manually checked clearly defined use cases based on different fitness goals. Results were similar tasting foods and conformed to nutritional information constraints.
- Checked nutritional information of select packaged foods
- Data scraped from .org websites to ensure data is reliable with complete information since two different sources were used
- Foods are commonly seen in supermarkets and therefore grounded in the real world

Nutri-KG snapshot

USC Viterbi
School of Engineering

University of Southern California
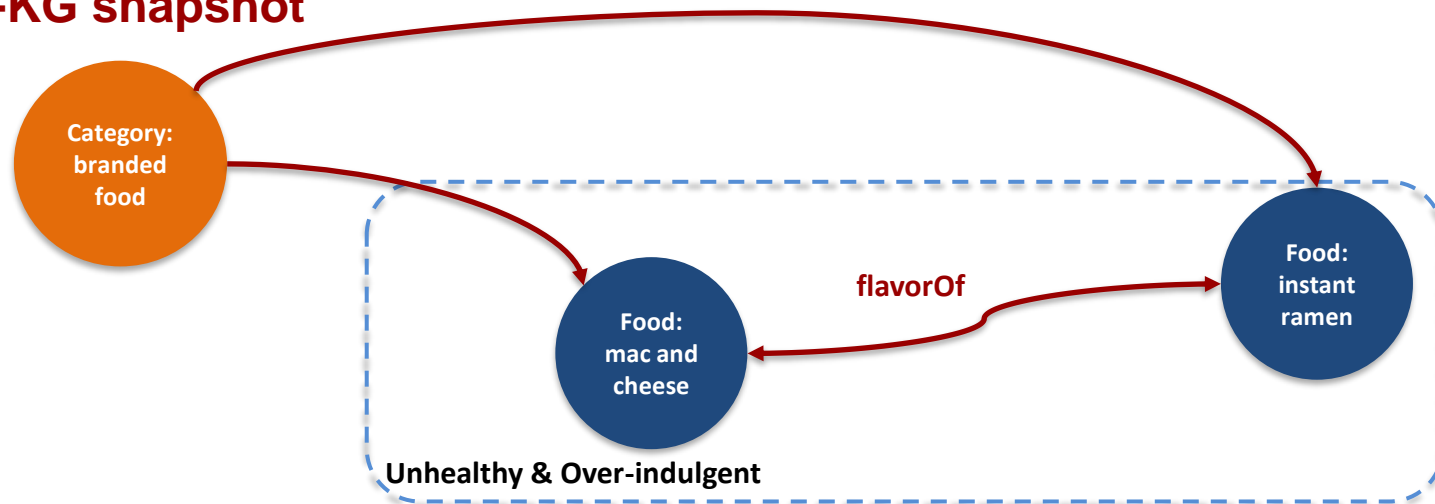
Nutri KG Recommendation System – Motivation and Process Flow

Motivation for Recommendation System

Suggestion of similar tasting foods

Nutri-KG enables us to provide similar tasting, healthier alternatives that will ensure the user does not quit their fitness journey because of drastic changes

Ability to customize diet plan according to fitness goals

The recommender system is designed to adjust its suggestions based on the user's fitness goals entered during sign-up

Process Flow

Nutritional Information

Similar Flavors

Categories

Finding Best K

Analyzing Clusters

Creating Food Cohorts

Customized Recommendation Block

Fitness Goals

Height, Age

Recommended Healthier Alternative
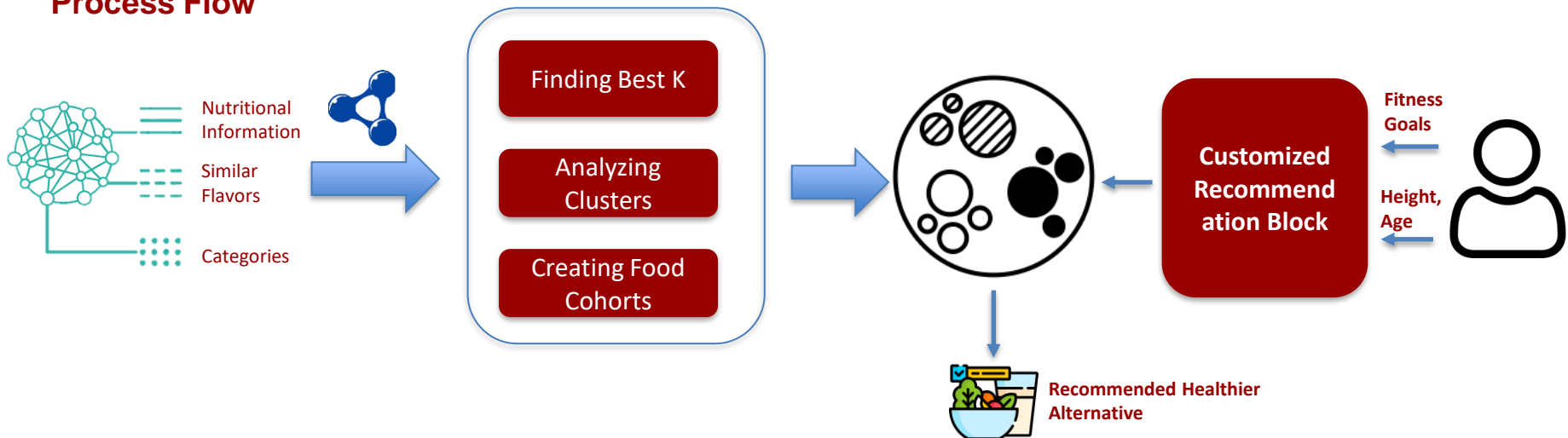
Nutri KG – Results, Challenges, and Solution

Food Option Cohorts

Confectionaries

> Avg. Calories — **137**
> Avg. Carbs — **14**
> Avg. Fat — **5**
> Avg. Protein — **8**
> Avg. Sodium — **150**
> Avg. Glycemic Load — **8**

Whole Nutrients

> Avg. Calories — **106**
> Avg. Carbs — **12**
> Avg. Fat — **5**
> Avg. Protein — **3**
> Avg. Sodium — **10**
> Avg. Glycemic Load — **8**

Frozen & Preserved

> Avg. Calories — **83**
> Avg. Carbs — **6**
> Avg. Fat — **4**
> Avg. Protein — **8**
> Avg. Sodium — **604**
> Avg. Glycemic Load — **7**

Packaged Snacks

> Avg. Calories — **93**
> Avg. Carbs — **17**
> Avg. Fat — **2**
> Avg. Protein — **2**
> Avg. Sodium — **46**
> Avg. Glycemic Load — **11**

Keto & Strength

> Avg. Calories — **158**
> Avg. Carbs — **0.6**
> Avg. Fat — **8**
> Avg. Protein — **20**
> Avg. Sodium — **60**
> Avg. Glycemic Load — **1**

Unhealthy & Over-indulgent

> Avg. Calories — **170**
> Avg. Carbs — **19**
> Avg. Fat — **7**
> Avg. Protein — **8**
> Avg. Sodium — **395**
> Avg. Glycemic Load — **15**

- To be consumed on occasions
- Healthy Foods
- To be avoided
- Metrics to watch out for

Challenges with the recommendation engine and solutions

Lack of User Data

Effective recommendations are obtained from algorithms like Collaborative Filtering. We resorted to item-based recommendations as we do not have extensive user data

New Data added to the system

The clusters will require re-adjustment when extensive data is added. This is currently overcome by classifying new foods into existing clusters
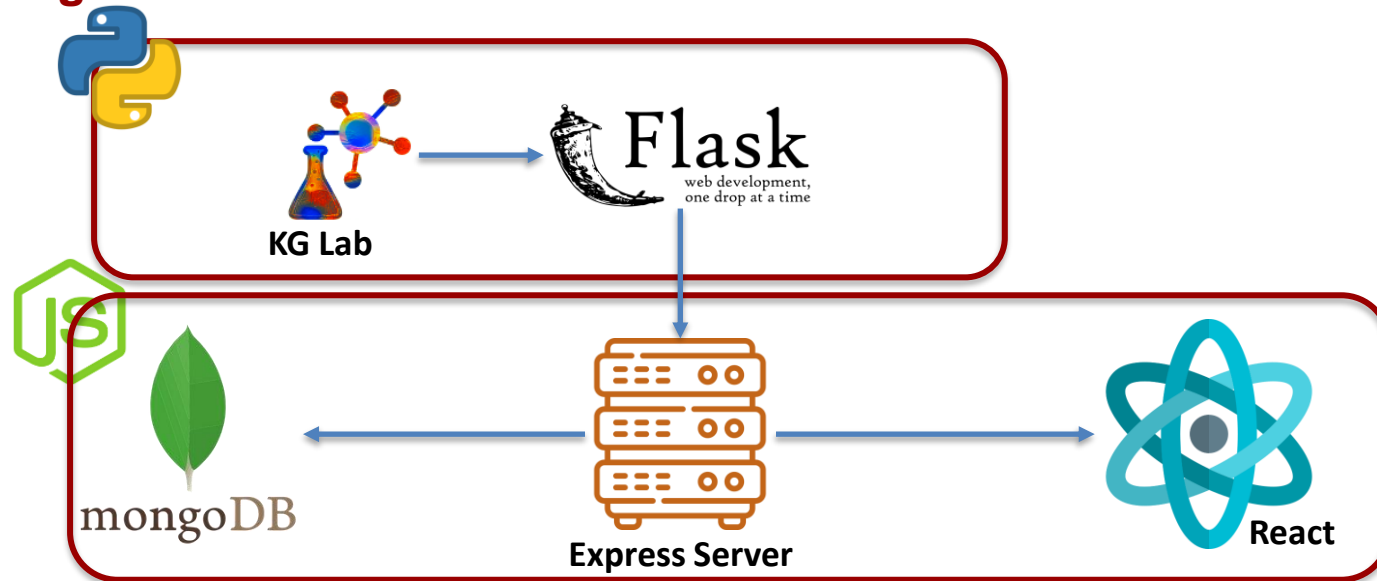
Nutri KG – User Interface Design

How will the user interface enable the user to interact with Nutri-KG?

- Login
- Profile for each user
- Search foods and view recommendations
- Save meals and view recommendations

UI Design – Process Flow

Nutri KG – User Interface Snapshots – Part 1

Homepage

User Profile Page

Nutri KG – User Interface Snapshots – Part 2

NutriKG   Explore   Meals   Niranjan   Profile ▾

Explore

Select a food and learn more about it. Also see other foods that are similar in flavor but more aligned with your goals!

Name: Niranjan
Fitness Goal: Bulk up

🔍 restaurant lasagna with meat italian   ✕

Selected food   Recommended food

restaurant lasagna with meat italian [Unhealthy & Over-indulgent]

Calories: 845
Carbohydrate: 51.92
Fat: 48.853
Protein: 49.49
Saturated Fatty Acids: 22.58

turkey roasted cooked meat only breast retail parts [Keto & Strength]

Calories: 1174
Carbohydrate: 0
Fat: 17.001
Protein: 254.67
Saturated Fatty Acids: 5.696

---

restaurant lasagna with meat italian [Unhealthy & Over-indulgent]

Calories: 845
Carbohydrate: 51.92
Fat: 48.853
Protein: 49.49
Saturated Fatty Acids: 22.58
Sodium: 2129.62
Category: Branded Foods

turkey roasted cooked meat only breast retail parts [Keto & Strength]

Calories: 1174
Carbohydrate: 0
Fat: 17.001
Protein: 254.67
Saturated Fatty Acids: 5.696
Sodium: 983.82
Category: Branded Foods

USC Viterbi
School of Engineering

University of Southern California

Nutri KG – SPARQL Querying & Analysis

Nutri-KG Application

1. Get similar tasting foods to user-provided query food

2. Get user fitness goals (and in the future, allergy information)

3. Filter foods based on nutritional information (Eg. lower calorie food if the goal is to eat healthily)

4. Order foods by nutritional information and display them as alternatives

SPARQL Queries to get simple but effective results from Nutri-KG

```
sparql = """
SELECT ?o ?c2 ?c1
WHERE {
    food1:{userInput} flavor1:flavorOf ?o .
    food1:{userInput} nutriinfo1:calories ?c1 .
    ?o nutriinfo1:calories ?c2 .
    FILTER(?c2 < ?c1)
}
    ORDER BY ?c2
"""
```

SPARQL Query
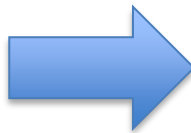
	o	c2	c1
0	food1:act-ii-fat-free-popcorn-32593-oz-by-cona...	130.0	150.0
1	food1:act-ii-butter-lovers-popcorn-275-oz-by-c...	140.0	150.0
2	food1:chesters-flamin-hot-popcorn-875-ounces-b...	140.0	150.0
3	food1:cracker-jackd-sweet--salty-salted-carame...	140.0	150.0

Query Output

USC Viterbi
School of Engineering

University of Southern California

Nutri KG – Analysis Results

Interesting insights about our food from Nutri-KG!
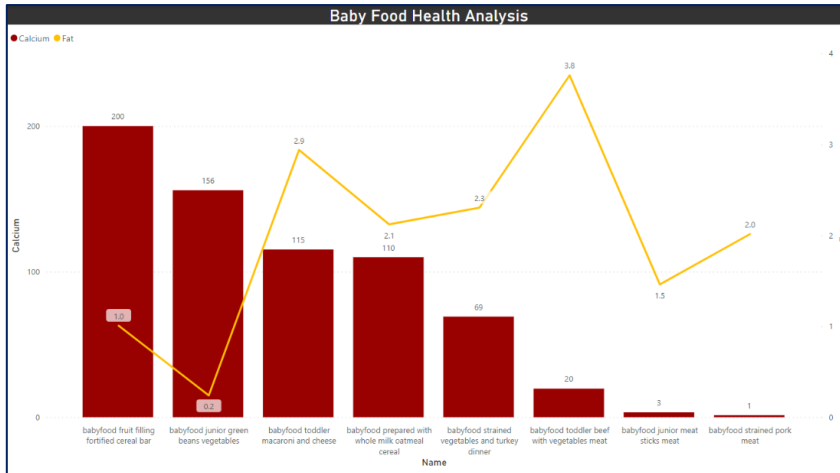
Case: Users want to understand what are the right options for their baby

- A balanced diet of fats and calcium is crucial for a baby. We can provide insights into foods that can provide this.

- Whole milk oatmeal & cereal, and toddler mac & cheese provide the perfect balance between these two macros
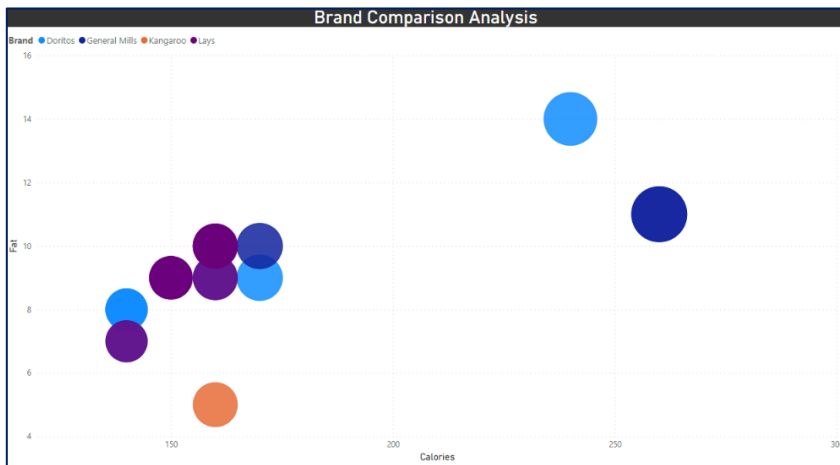
Case: What is the healthiest option amongst all chip brands

- Nutri-KG has access to branded foods and we have analyzed the category of chips to compare the brands

- Clearly, brands like Kangaroo and Lays are relatively healthier options to consume

Nutri KG – Recommendation Results and Opportunities

You can make a meal and we'll predict if that's good for you and offer alternatives aligned with your fitness goals

Nutri KG – Looking Forward
This is just the beginning. We have a lot of ideas!

User

- Weight
- Height
- Age
- Gender

- Fitness Goals
- Diet Preferences

- Usual Foods consumed

Knowledge Graph Data Sources

Food Categories: Macro/Micro Nutrients

Cereals

Packaged Food

Fast Food

Healthy Food

Flavor KG

SPICE

SWEET

SOUR

Nutri KG

NUTRI-KG
Enhance your Life

Construction

Recommendations

User Choice

Bacon

Steak

Ice Cream

Eggs & Turkey Bacon

Chicken Salad

Frozen Yoghurt

1) **Include Allergy conscious recommendations**

2) **More complex flavor information; use regional information**

3) **User data: Meals that work and those that don't -> feedback loop**

4) **Add pricing and availability information of recommended foods**

USC Viterbi
School of Engineering

University of Southern California