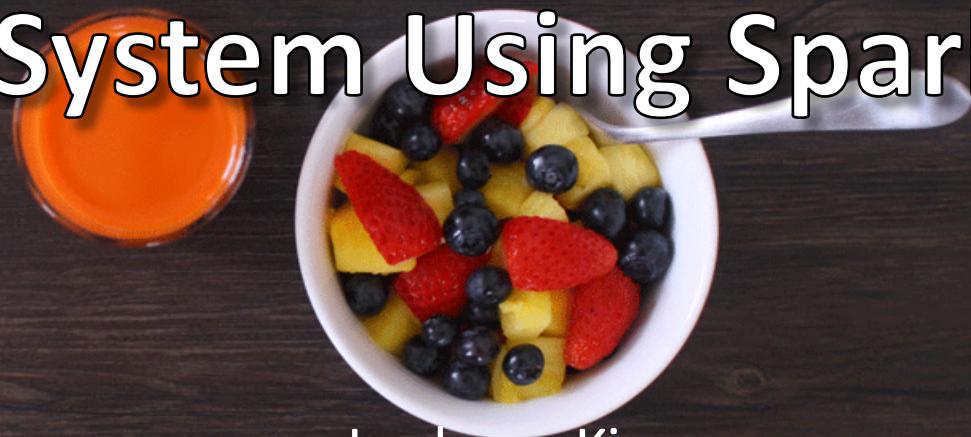




iRIS: A Large-Scale Food and Recipe Recommendation System Using Spark



Joohyun Kim

Sr. Data Scientist

MyFitnessPal – Under Armour Connected Fitness

myfitnesspal

POWERED BY
UNDER ARMOUR
CONNECTED FITNESS.



Who are we?





Under Armour = Apparel Company?

- <http://www.fool.com/investing/general/2015/06/07/how-under-armour-is-becoming-a-tech-company.aspx>

How Under Armour Is Becoming a Tech Company

By [Bradley Seth McNew](#) | [More Articles](#)

June 7, 2015 | [Comments \(0\)](#)

Under Armour (NYSE: [UA](#)) has grown its market share by leaps and bounds in recent years to become the second-largest athletic apparel seller in the U.S. as of 2014, behind only **Nike**.

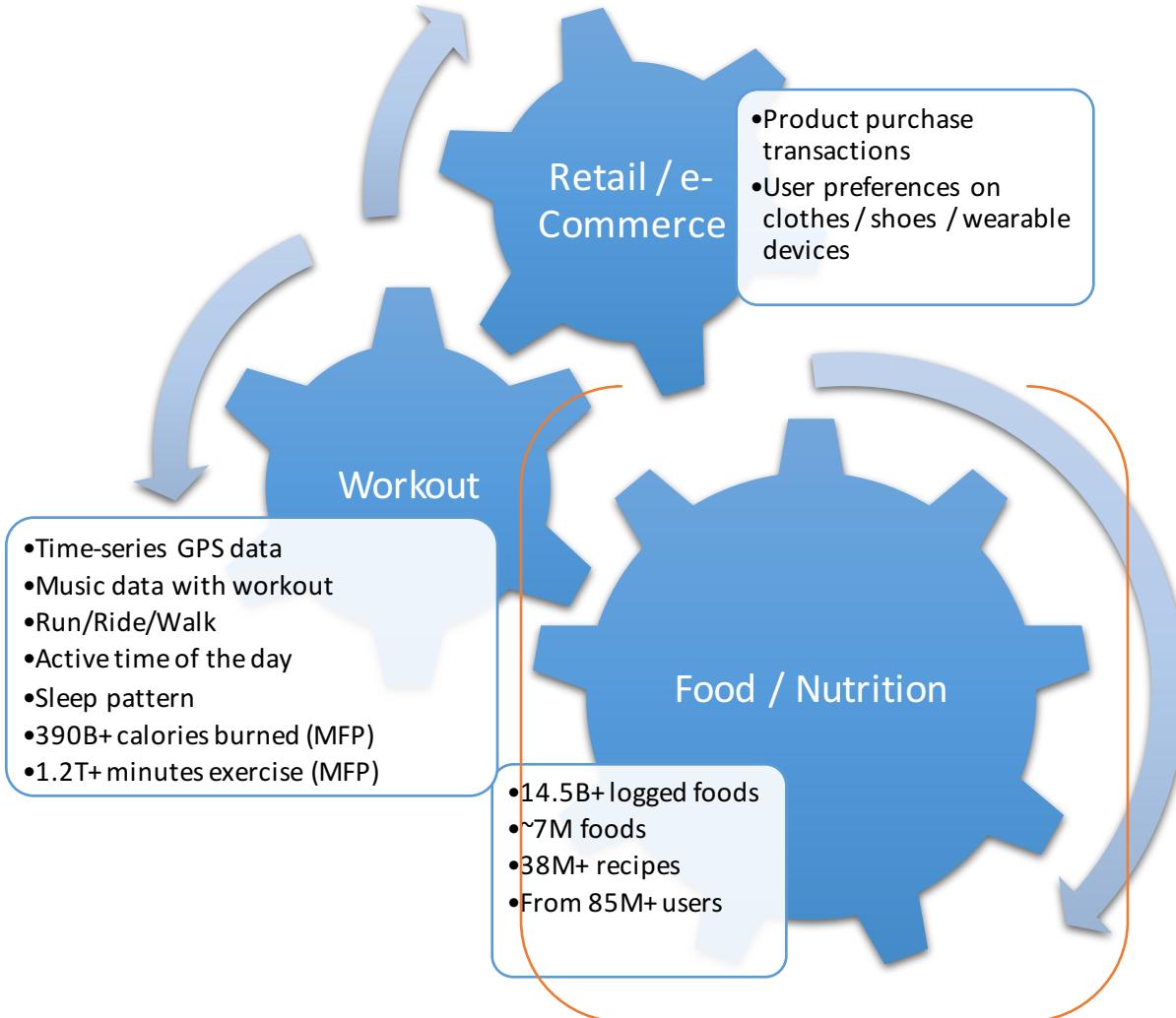
Even more exciting is the company's continued evolution. With recent smartphone app purchases and upgrades, a push into smart apparel, and now a new digital headquarters in Austin, Texas, Under Armour is quickly becoming a technology company.

Why Under Armour spent \$710 million for apps

Under Armour has made 2015 the year of connected fitness apps and devices. In Q1, the company purchased two fitness-tracking apps, MyFitnessPal and Endomondo. The company's MapMyFitness, bought in 2013, also was upgraded this year with a new premium service for serious fitness tracking. In total, Under Armour spent \$710 million on these apps. The company also released its own fitness app, called Record, in January.



Under Armour = Data Company!





Importance of Data



People You May Know



Yunchao Gong, Research Scientist at Facebook

Connect



Steve Ha, Global Data Center Operations

Connect



Vinod Marur, Engineer @ Google

Connect

[See more >](#)

Suggestions

Add people you know as Friends and connect with public Profiles you like.



Elmonis Kebre

Add as a friend



Latifah Akbar

Add as a friend



Agatha Renn

Add as a friend



Akoda Bushra

Add as a friend



Nausheen Adhikari

Add as a friend



Zainab Hossi

Add as a friend



Amritika Yarica

Add as a friend

Other Movies You Might Enjoy



Andha

Add

5 Not interested



Tu Mera Tashan

Add

5 Not interested



Dangal

Add

5 Not interested



Bajrangi Bhaijaan

Add

5 Not interested



Only Human

Add

5 Not interested



Sultan

Add

5 Not interested

[Close](#)



Google

News

U.S. edition

Search

Top Stories

News near you

Suggested for you

Alibaba and Amazon are headed for an epic showdown in India

The world's two largest e-commerce companies—America's Amazon and China's Alibaba—are poised to begin an epic battle for supremacy in one of the world's fastest growing online retail markets. Last year, India's health-care headquartered Amazon launched its...

[News about Amazon.com](#)

Apple Inc. (AAPL) Beats Google (GOOG) To Retain Top Spot As World's Most...

Last year Google (GOOG) and Apple (AAPL) were perceived to be the highest brand equity, and until recently the two companies were Apple Inc. (NASDAQ:AAPL) and Microsoft Corp.

Shown because you need to be updated about 2000.

Investment Analysts' Recent Ratings Changes for Dillard's (DDS)

The Legacy... 8 hours ago Dillard's had its "neutral" rating reaffirmed by analysts at JPMorgan Chase & Co. They now have a \$114.00 price target on the stock, down previously from \$126.00. 5/17/2015—Dillard's had its "neutral" rating reaffirmed by analysts at...

[News about Dillard's](#)

California triplets accepted to MIT

MIT's class of 172 hours ago

MIT accepted fewer than 5 percent of applicants accepted to MIT this year and

Clare, Edward and Christopher Goué were among them. Their grandfather taught at MIT, but all

three say they decided to attend the school independent of his ...

[News about Massachusetts Institute of Technology](#)

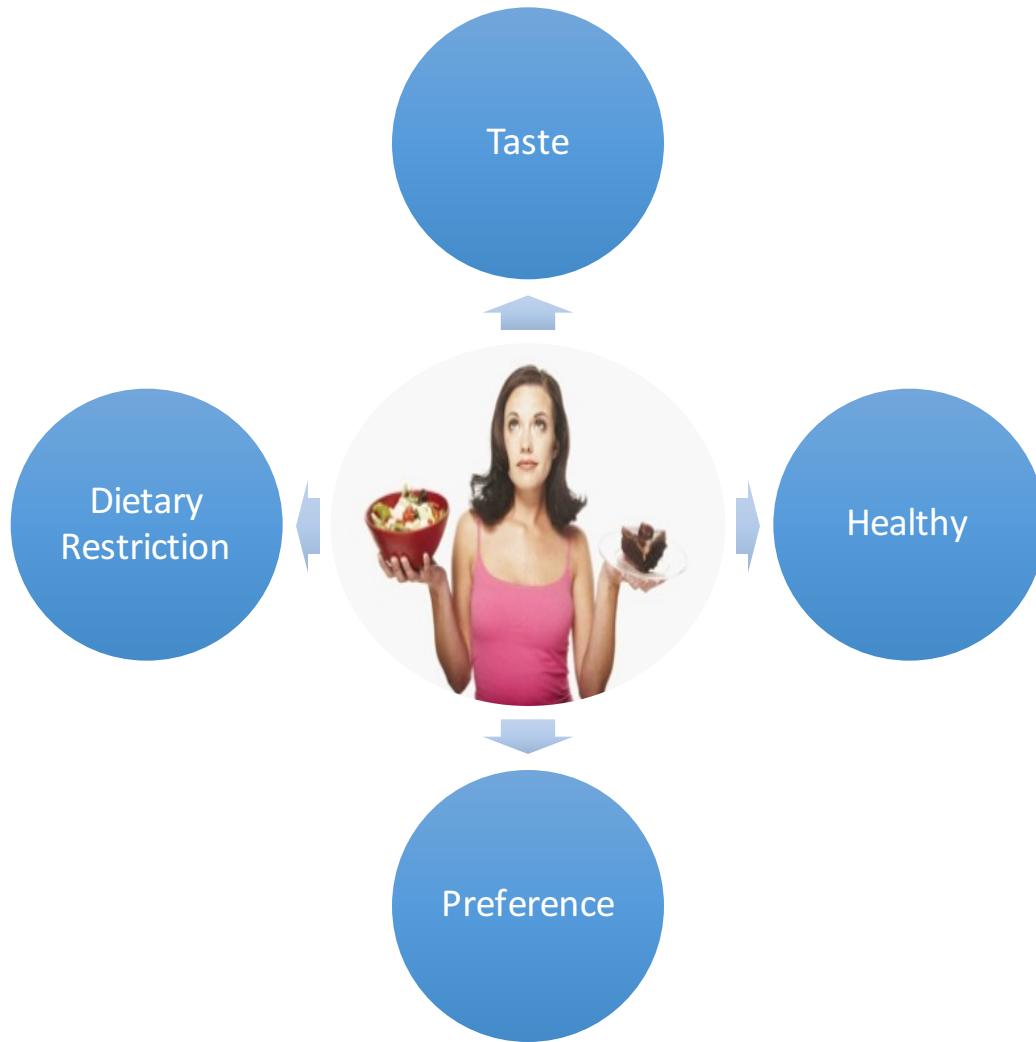


More Recommendations for You [See more](#)





Biggest Concern of Life: What to Eat?





Recommender System?



Typical ways of getting recommendation

Limited

Biased

Lack of Source



Collaborative Filtering

- Predict how a user may like a new item based on prior user behaviors with similar preference

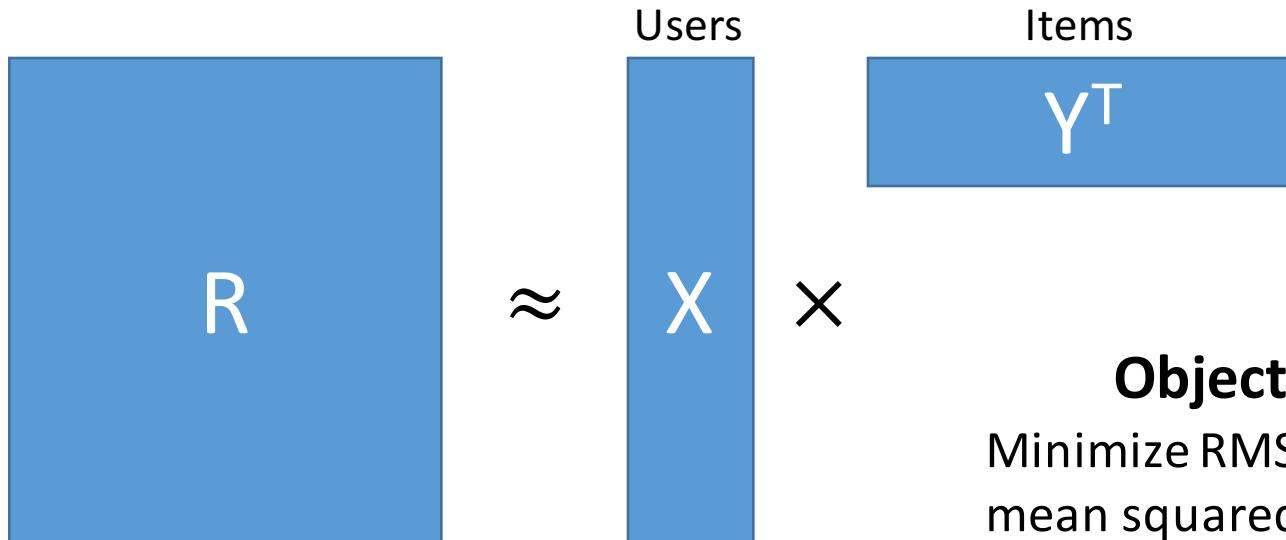
User – Food Logged Counts Table

	John	Mary	Mike	Jane
Banana	5	3	1	4
Blueberry	3	-	-	2
Apple	-	-	2	- (?)
Melon	1	-	-	- (?)



Matrix Factorization

- Filling in the missing entries in ratings (user-food logged counts) matrix
 - Formulate as low-rank matrix factorization
 - Factorize user-item matrix to user-feature and feature-item matrix (# features \ll # users or # items)



Objective

Minimize RMSE (Root mean squared error) between R and $X \times Y^T$



“Implicit” Matrix Factorization

- Explicit ratings are available for movies / songs
 - Typically 1~5 stars (ratings) given
- For MFP food logging events, there are only “logged” foods. No negative feedback
 - Can’t assume 0 count (no entry) as negative
 - Reference: Hu, Koren, and Volinsky, **Collaborative Filtering with Implicit Feedback Dataset**, ICDM 08
- Construct “binary” ratings matrix P , and factorize P instead of R (original ratings matrix)

$$\begin{matrix} R \\ \begin{bmatrix} 5 & 3 & ? & 3 & ? \\ ? & 2 & ? & ? & 4 \\ 1 & ? & ? & ? & ? \\ ? & ? & 1 & 2 & ? \\ ? & ? & 2 & ? & 1 \\ ? & 3 & ? & ? & 1 \end{bmatrix} \end{matrix} \Rightarrow \begin{matrix} P \\ \begin{bmatrix} 1 & 1 & ? & 1 & ? \\ ? & 1 & ? & ? & 1 \\ 1 & ? & ? & ? & ? \\ ? & ? & 1 & 1 & ? \\ ? & ? & 1 & ? & 1 \\ ? & 1 & ? & ? & 1 \end{bmatrix} \end{matrix} \approx \begin{matrix} X \\ \begin{bmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix} \end{matrix} \times \begin{matrix} Y \\ \begin{bmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix} \end{matrix}$$



Alternating Least Squares

- Optimizing X (user) and Y (item) at the same time is hard
- Fix X or Y \Rightarrow Solve for the other
 - Solve the system of linear equations
 - Take the derivative of objective function w.r.t X or Y, set 0, and solve
 - Starting with random initialization of Y
 - EM-like iterative process
- Iterate until the change is very small (or stop with fixed iteration number)



Scalability? \Rightarrow Parallelization!

- Rating matrix: 85M users \times 7M foods \cong 595T entries
 - Impossible to fit in a single machine
 - Sparse representation: billions of entries
- ALS can be easily parallelized with map-reduce framework
 - Sharding users and items vectors
 - Mapper on individual sub-matrix
 - Reducer on aggregation over users/items
- Spark MLLib
 - Parallelized version of ALS ready to use
 - Fast computation with DataFrame

```
val model = new ALS()  
  .setRank(20)  
  .setImplicitPrefs(true)  
  .setAlpha(40)  
  .setRegParam(0.1)  
  .setMaxIter(10)  
  .fit(ratings)
```



Food Recommendation Pipeline

Logged foods data (user / food)

Predict food preference by matrix factorization

Generate top K food recommendation



Generating Top K Recommendations

- We need to serve top recommended foods to users
 - With the trained factorized matrix model,
 - Predict top K foods for each user (in the order of their own preference)
- Seem trivial, but the computation is huge
 - For each user, retrieve food preference by $R = X \times Y^T$
 - Get top K per each user: $\min(O(Kmn), O(mn \log n))$
 - m: # of users, n: # of items
 - Same order of constructing whole ratings matrix
 - Major bottleneck of the entire pipeline
 - No easy way to get around the computation



Some Numbers

- Spark cluster
 - 72 nodes (1 master + 71 workers)
 - 2TB memory \Leftarrow **One of the largest clusters in production**
- Dataset
 - User : 85M+
 - Item (food) : \sim 7M
 - Rating (food log counts): 6.5B+ (aggregated per user/food)
- Time
 - ALS model training: 4 hours
 - Generating top K food recommendation for every user: 48 hours
 - **More than 20x speed improvement over Mahout in conventional Hadoop cluster**



Advantages Using Spark

- Faster development cycle
 - MLLib
 - Parallelization provided via RDD with abstraction
 - Easy to construct data pipeline with DataFrame
 - Easy to load / export data in and out of S3 / Redshift
- Faster model optimization
 - In-memory, distributed computation
 - Faster model training / testing
 - Significant reduction in parameter tuning / optimization on validation dataset
- Easy scalability
 - By launching more worker instances
- Enables frequent model updates
 - Reflect user preference change more often



Sample Food Recommendation

Logged Foods	Recommended Foods
Korean soy milk with high calcium	Cooked White Jasmine Rice
Coke 12oz	Steamed White Rice (Unenriched)
Fried rice	Pho
Korean mixed grain shake	Kimchi
Korean Rice Cake	Tofu - Fried
Sweet Soy Milk	Miso Soup With Seaweed and Tofu
Blackberries - Raw	Shrimp Dumplings
Blueberries - Raw	Miso Soup
Korean Melon (Chameh / 참외)	Salmon Nigiri
Japchae (Korean Stir-Fried Sweet Potato Noodles)	Sunny Side Up



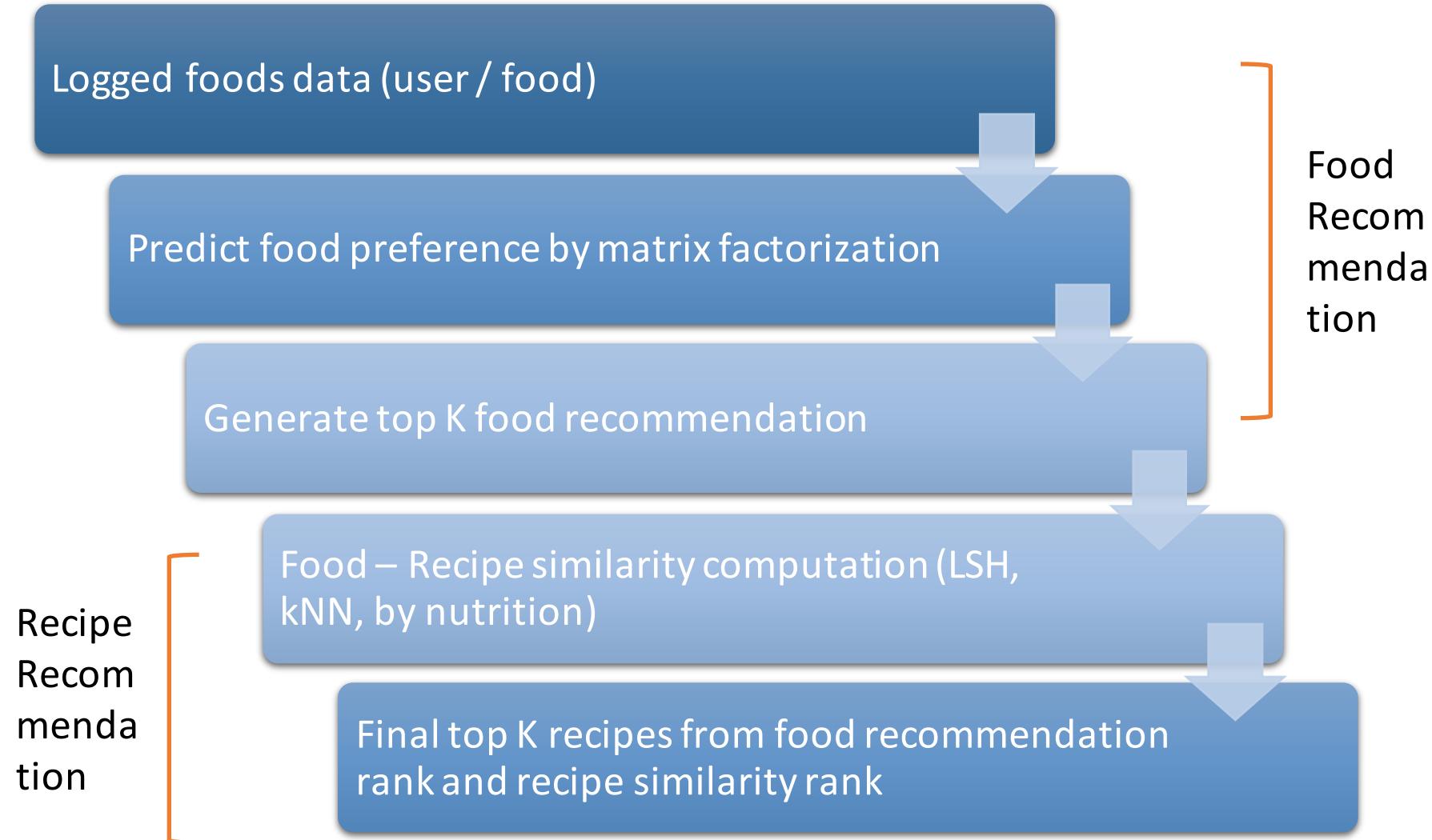
Extension: Recipe Recommendation

- Advantages of recommending recipes
 - Richer metadata (instructions, ingredients, cuisines, ...)
 - Complete food, home-cookable
 - Customizable with personal preference/restriction
- Recipes are not “public”
 - Currently, only foods are shared across different users
 - Recipes are “private” to individual user when created
 - Cannot construct standard user-item ratings matrix
- Solution: Recommend recipes using similarity with foods

The screenshot shows a user interface for a recipe card. At the top, there are navigation buttons: a left arrow, a 'Edit Ingredients' button, and a right arrow. Below this is a message: 'Please make sure we have all the ingredients.' The interface is divided into two main sections: 'Recipe' (which is highlighted in grey) and 'Ingredients'. The 'Ingredients' section lists the following items:
2 cups all-purpose flour
2 teaspoons baking soda
1/2 teaspoon salt
2 teaspoons ground cinnamon
3 large eggs
2 cups sugar
3/4 cup vegetable oil
3/4 cup buttermilk
2 teaspoons vanilla extract
2 cups grated carrot
1 (8-ounce) can crushed pineapple, drained
1 (3 1/2-ounce) can flaked coconut
1 cup chopped pecans or walnuts
Buttermilk Glaze
Cream Cheese Frosting



Extension: Recipe Recommendation





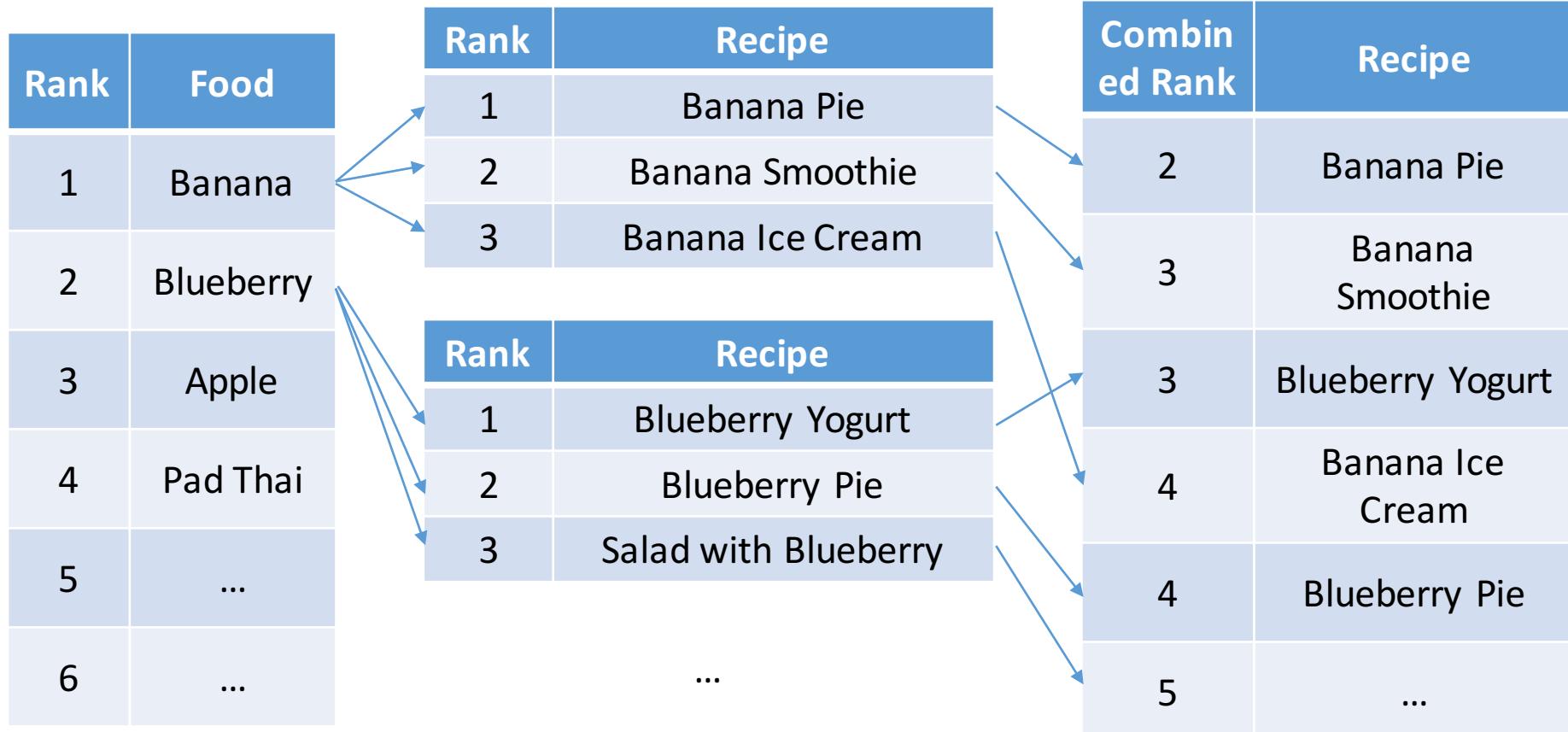
Nearest Neighbor with Locality Sensitive Hashing

- Naïve nearest neighbor computation: $O(n^2)$
 - Nearly impossible with 7M foods and 38M recipes
- Locality Sensitive Hashing (LSH)
 - Hashes similar items into the same buckets with high probability
 - Similarity metric: Euclidean distance on nutrition vector
 - NN computation much faster, look up within the bucket: $O(k^2)$ (k : max size of bucket)
- Spark implementation
 - <https://github.com/mrsqueeze/spark-hash>



Top K Recipe Recommendation

- Order by sum of food recommendation rank and recipe similarity rank





Sample Recipe Recommendations

Recommended Foods
Cooked White Jasmine Rice
Steamed White Rice (Unenriched)
Pho
Kimchi
Tofu - Fried
Miso Soup With Seaweed and Tofu
Shrimp Dumplings
Miso Soup
Salmon Nigiri
Sunny Side Up



Recommended Recipes
Noodle sauce
Stone Ground Dijon Mustard Marinade
Citrus Dijon Miso Dressing
Shirataki Noodle Soup
dumpling sauce
Dijon Miracle Whip
Paleo Vanilla Ice-Crème
Mable's Chili Burrito
cake batter milkshake
Pita pizza



Integrating with Taste Profiles

- Machine learning classifier that outputs probability distribution over 6 taste categories
 - Savory, Sweet, Sour, Spicy, Salty, Bitter
 - NN classifier performed over feature vector (semantic word vector + numeric nutritional value vector)
 - With small number (~1400) of labeled foods
 - 87% accuracy on separately labeled test set (~1200)
 - Works as additional metadata for foods
- Recommendation results can be further filtered / reordered by personal taste preferences
- With Spark, data integration with hash-join is much faster

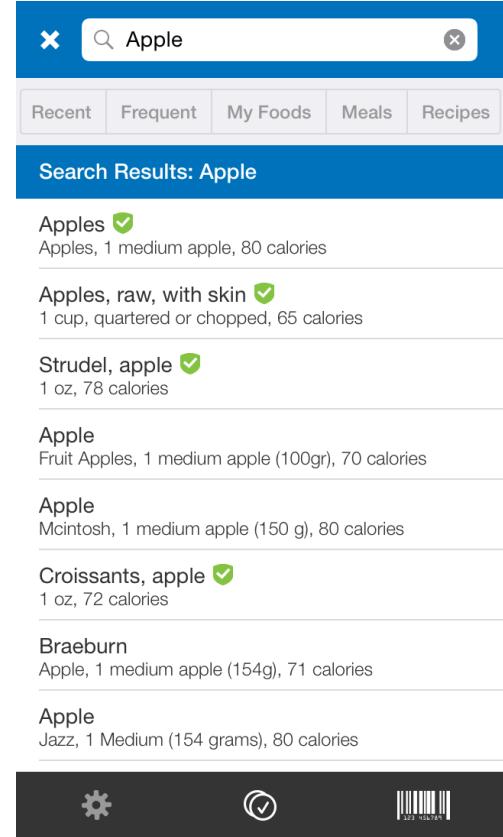


Problems

- Cold-start problem
 - New user (important)
 - New food item (not too much, since they may not be popular)
- Solution
 - Hybrid with content-based recommendation
 - Construct basic profile for new users to get baseline recommendation
 - May recommend new items based on feature similarity
 - Runs the pipeline more often
 - For new users, recommend top/popular foods
 - After a while, these users will get recommendation

Possible Extension

- Collaborative filtering over aggregated food clusters
 - User-generated foods contain (near) duplicates
 - Combine with verified food project
 - Spark-based data deduplication / processing pipeline
 - Construct “verified” foods out of thorough clustering
 - Recommend users with high-quality food
 - Recommend representative food from each cluster, instead of individual variation
 - Reducing the computation time due to reduced dimension



The screenshot shows the MyFitnessPal app interface. At the top, there is a search bar with the text "Apple". Below the search bar is a navigation bar with tabs: "Recent", "Frequent", "My Foods", "Meals", and "Recipes". The main area is titled "Search Results: Apple". The results list various types of apples, each with a green checkmark and a brief description of the food item and its calorie content. At the bottom of the screen are three icons: a gear for settings, a circular arrow for refresh, and a barcode.

Food Item	Description
Apples	Apples, 1 medium apple, 80 calories
Apples, raw, with skin	1 cup, quartered or chopped, 65 calories
Strudel, apple	1 oz, 78 calories
Apple	Fruit Apples, 1 medium apple (100gr), 70 calories
Apple	Mcintosh, 1 medium apple (150 g), 80 calories
Croissants, apple	1 oz, 72 calories
Braeburn	Apple, 1 medium apple (154g), 71 calories
Apple	Jazz, 1 Medium (154 grams), 80 calories



Application

- Recommend frequently paired foods
- Pairing foods within a single meal depends on
 - Individual user's own preference
 - Cultural difference (region, country)
- Simple way
 - Suggest popular foods based on co-occurrence stats per individual user / overall users
- Utilize this framework to capture better personalized preference

Carrier 2:10 PM Add Food

Sliced Havarti Cheese (Trader Joes)

Serving Size 1 Slice

Number of Servings 1

Nutrition Facts

Calories	110
Fat (g)	10
Carbs (g)	0
Protein (g)	6

[More Nutrition Facts](#)

Add Frequently Paired Foods

- Organic Brown Eggs
Kirkland Signature, 100 g (1 egg), 140 calories
- Stone Ground Corn Tortillas
Trader Jose's, 56 g, 100 calories
- Avocados - Raw
0.5 avocado, NS as to Florida or California, 161 calories



Summary

- Spark-powered machine learning pipeline for food/recipe recommendation system
 - Faster computation help reduce the time on development cycle
 - Help data scientists focus on core problems
 - Easy extension by attaching additional data processing steps with scalability
- Only scratched a surface
 - Food / recipe recommendation
 - Extensions with other data sources
 - Workout
 - Music
 - Retail
 - e-Commerce



UNDER ARMOUR.

Questions?