# Semantic Nutrition: Estimating Nutrition with Mobile Assistants

Joshua D'Arcy[1,2,3], Sabrina Qi[3], Dori Steinberg, Ph.D., R.D.[4,5], Jessilyn Dunn, Ph.D.[2,3]
*[1]Duke University School of Medicine [2]Department of Bioinformatics and Biostatistics, Duke University Health System*
*[3]Duke University Pratt School of Engineering [4]Duke School of Nursing [5]Duke Global Digital Health Science Center*

**Background.** Food logging is effective for improving dietary intake and clinical outcomes, but notoriously difficult in clinical and research settings. Determining the relationship between clinical outcomes and nutritional intake depends on accessing accurate nutrition data. Common food logging practices include handwritten food diaries or manual entry of foods into a smartphone application. Handwritten food diaries require significant effort to link the underlying nutritional information for each entry and transcribe it into a digital format. Smartphone applications introduce a need for technological literacy for food logging, but users still need to manually log their food diaries. Furthermore, it is uncommon to find perfect matches of a logged food item in a nutrition database. We propose a process that uses voice transcription via a mobile assistant (Siri on iOS or Google Assistant on Android) for low-barrier food entry, natural language processing (NLP) for nutrition estimation based on semantic relationships, and rapid indexing to digitize this data to improve research and clinical care. The digitization of dietary data will serve as a foundation for efficient and scalable food logging.

**Methods.** We tested a mobile platform that uses virtual assistants to transcribe natural language to food diary entries. We obtained a database of expertly annotated nutrition profiles for 50,000 food items in 12-dimensional space (Table 1). Given the diversity of food, food log entries will frequently lack perfect matches to this database. In order to estimate the nutritional values for these unmatched food log entries, we designed a word embedding method to map individual food items to their most likely nutrition profile. We applied a word2vec model trained on the Google News corpus to our food database to convert each food item to a 300-dimensional embedded vector representation. We combined multi-word food entries into one vector by taking their element-wise sum. Our model enables nutrition profile estimates of food log entries that do not have an equivalent entry in our food database based on the cosine distance nearest neighbor in the food database.

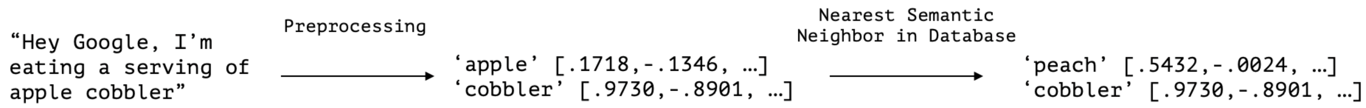


**Figure 1**: Natural language converted to an embedded vector representation (using NLP).

**Table 1**: Food dataset with reference column of embedded vector representations.

| food_name | carbohydrates | sugar | total_fats | sat_fats | protein | fiber | calories | cholesterol | potassium | sodium | calcium | magnesium | vector representation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| avocado roll | 50.70 | 0.29 | 10.98 | 1.60 | 6.54 | 4.72 | 321.41 | 0.0 | 422.22 | 34.26 | 13.0 | 37.0 | [0.05908203125, 0.455078125, -0.014404296875, ... |
| fried tofu | 2.51 | 0.77 | 5.72 | 0.83 | 5.34 | 1.11 | 76.55 | 0.0 | 41.39 | 4.54 | 105.0 | 17.0 | [-0.326171875, 0.05706787109375, 0.232421875, ... |
| orzo | 63.51 | 2.27 | 1.28 | 0.24 | 11.09 | 2.72 | 315.54 | 0.0 | 189.66 | 7.30 | 19.0 | 45.0 | [-0.0810546875, 0.059814453125, -0.0537109375,... |
| greek salad | 7.00 | 5.07 | 8.68 | 3.69 | 3.82 | 1.37 | 119.25 | 18.5 | 203.96 | 400.08 | 119.0 | 14.0 | [-0.16455078125, -0.0166015625, -0.015625, 0.7... |
| coffee | 0.00 | 0.00 | 0.05 | 0.00 | 0.28 | 0.00 | 2.37 | 0.0 | 116.13 | 4.74 | 4.0 | 7.0 | [-0.1611328125, -0.13671875, -0.373046875, 0.6... |

**Future Directions.** In this abstract, we present a novel method to digitize nutrition data and improve food logging using voice transcription and NLP. Ongoing NLP work includes training our own word embeddings using a corpus more representative of the semantic relationships between foods (i.e. online food blogs versus the Google News corpus). We believe that it is critical to develop efficient and scalable methods to access and analyze nutrient intake from food logs in order to conduct robust randomized controlled clinical nutrition trials and incorporate dietary information into clinical practice.