American Economic Journal: Microeconomics 2025, 17(1): 282–307
https://doi.org/10.1257/mic.20220248

Triplet Embeddings for Demand Estimation[†]

By Lorenzo Magnolfi, Jonathon McClure, and Alan Sorensen*

We propose a method to augment conventional demand estimation approaches with crowd-sourced data on the product space. Our method obtains triplets data (“product A is closer to B than it is to C”) from an online survey to compute an embedding—i.e., a low-dimensional representation of the latent product space. The embedding can either replace data on observed characteristics in mixed logit models, or provide pairwise product distances to discipline cross-elasticities in log-linear models. We illustrate both approaches by estimating demand for ready-to-eat cereals; the information contained in the embedding leads to more plausible substitution patterns and better fit. (*JEL* C45, C51, D11, D12, D21, L66)

Estimating demand systems in differentiated product markets is fundamental in empirical Industrial Organization (IO), and the toolkit of methods can be roughly divided into two approaches.[1] The *product space* approach assumes that consumers have preferences over products, and product-level demand comes from the aggregation of those preferences. This is perhaps the most natural way to conceptualize demand, and it has the advantage of yielding demand equations that are computationally simple to estimate (e.g., Christensen, Jorgenson, and Lau 1975; Deaton and Muellbauer 1980). The *characteristics space* approach, pioneered by Lancaster (1966) and McFadden (1974), instead treats products as bundles of characteristics and defines consumers’ preferences over these characteristics. Methods in this vein have their own advantages: they are based on theoretically grounded models of discrete choice; they have convenient analytical properties (e.g., closed-form solutions for firms’ predicted market shares); and with the inclusion of random coefficients on some characteristics (as suggested, for example, by Berry, Levinsohn, and Pakes 1995

* Magnolfi: Department of Eonomics, University of Wisconsin-Madison (email: magnolfi@wisc.edu); McClure: Department of Economics, Mitch Daniels School of Business, Purdue University (email: mcclur47@purdue.edu); Sorensen: Department of Economics, University of Wisconsin-Madison and NBER (email: sorensen@ssc.wisc.edu). Robin Lee was coeditor for this article. A previous version of this draft was circulated under the title “Embeddings and Distance-based Demand for Differentiated Products.” We thank Luis Armona, Steve Berry, Giovanni Compiani, Glenn Ellison, Phil Haile, Rob Porter, Chris Sullivan, Jeff Thurk, and seminar participants at University of Maryland, IO[2], IIOC, EC 2022 and NBER for helpful comments. The results reported below represent our own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are our own and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein. We received IRB approval from University of Wisconsin-Madison MRR (IRB I2020-0364).

[†] Go to https://doi.org/10.1257/mic.20220248 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

[1] See Berry and Haile (2021) and Gandhi and Nevo (2021) for recent surveys.

(henceforth, BLP) and McFadden and Train 2000), they allow for rich patterns of substitution between products.

Both of these approaches typically require data on the product space to yield credible estimates. Unless information about the product space is used to restrict substitution patterns, product space models quickly run into a curse of dimensionality: absent any restrictions, a market with J products will require estimation of separate parameters for each of the J^2 demand elasticities. And because a key rationale for the characteristics space approach is that it collapses preferences over J products down to a set of $K \ll J$ characteristics, it is essential that the demand-relevant characteristics are known to the researcher and observed in the data.

In this paper, we propose a pragmatic approach for obtaining complementary data that can be used when product characteristics are not observed, either to discipline the parameters in a product space model or to serve as (latent) characteristics in a characteristics space model. We first solicit product comparisons via an online survey to generate data of the form "product A is closer to B than it is to C"—commonly referred to as "triplets data" in the machine learning literature—and then apply the *t*-Distributed Stochastic Triplet Embedding (tSTE) algorithm proposed by Van Der Maaten and Weinberger (2012) to compute an *embedding* of the products in low-dimensional space. Distances between products are then easily calculated from this embedding, and cross-price elasticities in a product space model can be estimated as a function of these distances, as in Pinkse, Slade, and Brett (2002). Alternatively, the products' coordinates in the embedding can be treated as the characteristics in a conventional mixed logit demand model like BLP. As we explain below, these embeddings are easy to generate, and the data required to compute them are straightforward to obtain.

We illustrate the method by estimating demand for ready-to-eat breakfast cereals, augmenting the usual price and quantity data with survey data obtained from college students and Mechanical Turk workers. We chose this application—a common laboratory for evaluating demand estimation methods—because standard data on product characteristics are also available, which we can compare to the embedding. The triplets data from the survey lead to an embedding of products that appears quite sensible. We use the embedding to first show how it can be used in a product space model similar to that of Pinkse, Slade, and Brett (2002), where the (log) quantity demanded for any product is a linear function of its own (log) price and of all competing products' (log) prices, allowing the cross-price elasticity parameters to be functions of pairwise product distances computed from the embedding. Estimates of this model are computationally trivial to obtain, and they yield reasonable own- and cross-price elasticities—broadly similar to those reported in prior studies like Nevo (2001) and Backus, Conlon, and Sinkinson (2021). Importantly, we show that the distances computed from the embedding deliver more plausible estimates of substitution patterns than distances computed from observed product characteristics.

We then show how the coordinates from the embedding can be used in more conventional discrete choice models like BLP. If we treat the products' coordinates in the embedding as latent characteristics, essentially including them as the covariates in an otherwise standard BLP model, we obtain elasticity estimates that are comparable to those from a model that uses observable characteristics. This result is particularly encouraging because it suggests our method can deliver

credible estimates even in markets where demand-relevant characteristics are more elusive, such as fashion apparel, movies, or music. Using survey data to obtain an embedding is essentially a way of crowdsourcing data on product characteristics, which is a useful option when data on product characteristics are otherwise difficult to collect. This is a common and meaningful problem in empirical IO—for example, De Loecker, Eeckhout, and Unger (2020) and Syverson (2019) note that the lack of IO research on broad trends in markups stems largely from the difficulty of scaling methods that rely on market-specific product characteristics data.

Our paper contributes to an emerging literature that proposes new sources of data to estimate demand, and we are not the first to propose the use of embeddings. Bajari et al. (2021) use deep neural nets to generate numeric latent attributes (i.e., an embedding) from products' images and text descriptions, and then leverage those attributes to estimate a hedonic price function for apparel items on Amazon.com. This is a nice example where the demand-relevant information about a product—say, a woman's dress—cannot be easily summarized by a set of characteristics, even though humans can easily process and synthesize the relevant information from the product's image and/or text description. With similar motivation, Han et al. (2021) deploy a deep neural net to compute an embedding describing the product space for fonts. Armona, Lewis, and Zervas (2021) learn products' latent attributes from search data (consumers' web browsing histories) using Bayesian Personalized Ranking and apply their method to estimate demand for hotels. Compiani, Morozov, and Seiler (2023) use unstructured image and text data to generate an embedding and estimate a nested logit model of demand with overlapping nests. Related articles that use data on consumers' transactions and search to fit embeddings and estimate demand also include Ruiz, Athey, and Blei (2020); Kumar, Eckles, and Aral (2020); and Gabel and Timoshenko (2022). The primary distinction between our analysis and these prior studies is our use of survey data to compute the embedding. In essence we are relying on human respondents to describe the product space and then incorporating that information into standard methods for estimating demand.

We also view our approach as being similar in spirit to studies that employ auxiliary data to augment existing demand estimation methodologies. Berry, Levinsohn, and Pakes (2004) is a canonical study in which second-choice data from surveys are used to generate additional moments in the estimation of demand for automobiles. Petrin (2002) is an early example of combining demographic data with the usual price and quantity data to get richer estimates of substitution patterns. More recently, Conlon, Mortimer, and Sarkis (2022) show how demand estimates can be meaningfully improved by incorporating data on second-choice diversion ratios, in their case obtained from experimentally generated stockouts. They even show that the information contained in such data is powerful enough to enable the estimation of a semi-parametric model that imposes much lighter assumptions than conventional mixed logit.

I. Demand Estimation and Linear Embeddings

Consider a market, indexed by t, where firms offer a set $\mathcal{J}_t$ of differentiated products. Prices and quantities for each good j are denoted as p_{jt} and q_{jt}. The demand system that maps prices into quantities depends on two key sets of primitives: consumers'

preferences and demographics, and the product space. We assume that products can be represented by coordinates in the m-dimensional Euclidean space; thus, the product space in market t is a set of vectors $\mathbf{x}_t \equiv \{x_{1t}, \ldots, x_{Jt}\} \in \mathbb{R}^{m \times J_t}$. Hence, demand can be written as $q_{jt} = \sigma_j(p_t; \mathbf{x}_t, \theta)$ for some function σ_j, where θ is a vector of preference parameters.

The product space $\mathbf{x}_t$ is a key element of the empirical demand system under any estimation approach. In the *characteristics space* approach, demand is assumed to arise from discrete choices of individual consumers whose preferences are defined over the product space coordinates. Thus, x_{jt} enters consumers' indirect utility for product j interacted, with preference parameters. In the *product space* approach, the functions σ_j are estimated directly, with functional form restrictions imposed (typically based on either convenience or a representative consumer micro-foundation). The importance of the product space $\mathbf{x}_t$ is that it can play a role in disciplining the otherwise overabundant cross-elasticity parameters: as in Pinkse, Slade, and Brett (2002), cross elasticities of demand between products j and k can be modeled as a function of the distance $d_{jk}(\mathbf{x}_t)$ between the two products.

Within this framework, our method can be understood as a way of recovering $\mathbf{x}_t$ from auxiliary data as an *embedding* when product characteristics are not observable or are difficult to codify. The next subsections provide an overview of embeddings and how they can be incorporated into either of the two main approaches to demand estimation.

A. *Product Embeddings*

In machine learning, an *embedding* is a low dimensional, learned continuous vector representation of discrete variables.[2] In our case the discrete variables are just product indicators ("this is product j"), and the objective is to assign locations (real-valued vectors) to these products in a way that best satisfies the distance comparisons from a training dataset. As training data we use *triplets*—i.e., comparisons of the form "product A is closer to B than it is to C"—obtained from a survey that we describe in detail below in the context of our application.

Thus, given our set of products, we want to find a set of vectors $\mathbf{x} \equiv \{x_1, \ldots, x_J\}$ that represent the products in m-dimensional space, and we assume that this corresponds to the product space that enters the demand system. To learn the embedding from triplets data, we use the t-distributed Stochastic Triplet Embedding (tSTE) algorithm proposed by Van Der Maaten and Weinberger (2012). Letting $\mathcal{T}$ be the set of triplet comparisons in our data, each one indicating that some product i is closer to j than it is to k, tSTE solves

$$\max_{\mathbf{x} \in \mathbb{R}^{m \times J}} \sum_{(i,j,k) \in \mathcal{T}} \ln(\pi_{ijk}), \text{ where } \pi_{ijk} = \frac{\left(1 + \frac{\|x_i - x_j\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}{\left(1 + \frac{\|x_i - x_j\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}} + \left(1 + \frac{\|x_i - x_k\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}},$$

[2] Common uses of embeddings in machine learning include image classification and natural language processing. For example, Google's Word2Vec algorithm uses a neural network to assign vector representations to words so that the cosine similarity between any two words' vectors can be used as a measure of their semantic similarity. Embeddings are also commonly used for visualizing high-dimensional data: collapsing to two or three dimensions allows for simple plots in which clusters and other patterns are easy to see.

and α is the degrees of freedom parameter for the underlying Student-t kernel. To gain intuition about this program, note that to fit one single triplet the embedding would assign equal coordinates to products i and j and infinitely far coordinates to products i and k, so that $\ln(\pi_{ijk})$ diverges. As we introduce further triplet comparisons, the solution is more intricate: the embedding has to fit more complex patterns because the same products are involved in multiple comparisons. The tSTE objective function is also similar to the log-likelihood of a binary logit model, where agents respond to "is i closer to j than it is to k" according to utilities that depend on the distances in the (unknown) product space and an additive error.[3] To draw another comparison with discrete choice, just like some models of multinomial choice can be estimated from data on a subset of the choices available (McFadden 1978; Fox 2007), in tSTE triplet comparisons are sufficient to pin down the full product space $\mathbf{x}$. The functional form based on the heavy-tailed Student-t distribution is preferred to GEV distributions because it performs well without overreacting to noisy triplets that violate the broader consensus in the data.

In our empirical application to ready-to-eat breakfast cereal, we have $J = 86$ products, so if we choose to fit a six-dimensional embedding ($m = 6$), then the above program is a numerical optimization problem with 516 free variables. We found that ordinary gradient descent algorithms solve this optimization problem in a matter of minutes.[4]

The choice of the hyperparameters α and m deserves some discussion. Following the literature (Van der Maaten and Hinton 2008; Van Der Maaten and Weinberger 2012), we set $\alpha = m - 1$ and find that our embedding is not very sensitive to this parameter. The choice of m is more critical, and while various rules of thumb have been proposed in the machine learning literature, there seems to be no widely accepted criterion. We propose a pragmatic approach that selects m by iteratively adding dimensions until additional embedding characteristics do not have a meaningful impact, i.e., produce very similar distances among products. Additional details on this procedure, including two alternative ways of implementing it (based on Frobenius distances and PCA, respectively), are in Supplemental Appendix B.

B. *Using Embeddings in Demand Models*

We now describe how to use embedding data when estimating demand for differentiated products. First, we consider how embeddings can enhance product space approaches to demand estimation. Models in this class are often deemed unsuitable for applications to markets with differentiated products, because in even the simplest specifications (e.g., linear) the number of parameters grows exponentially with the number of products.[5]

[3] We thank Steve Berry for noticing this similarity.

[4] We used a version of the MATLAB code provided by Laurens Van der Maaten: https://lvdmaaten.github.io/ste/Stochastic_Triplet_Embedding.html.

[5] Other reasons include the difficulties in incorporating heterogeneity across consumers and in evaluating the demand for new products (Gandhi and Nevo 2021).

Various solutions to this problem have been devised.[6] In this paper, we adopt the method proposed by Pinkse, Slade, and Brett (2002), who note that when competition among firms is spatial (i.e., it depends on some topology of the product space), the parameters that govern substitution between products can be projected on a flexible function of their distances. When products have an observable location in the physical space, as in the application of Pinkse, Slade, and Brett (2002) or in Houde (2012), distances are straightforward to measure. When spatial competition is only figurative, as in the case of Pinkse and Slade (2004)'s study of the UK beer market, distance can instead be modeled as a function of observable product characteristics.

Using an embedding computed from triplets data as described above, we can obtain a map of the product space even when the products' characteristics are difficult to observe or quantify, and the distances between products in the embedding can be used in the framework of Pinkse, Slade, and Brett (2002). In the empirical exercise below, we estimate, with product-level data, the log-linear demand model

$$\ln(q_{jt}) = \alpha_j + \beta_j \ln(p_{jt}) + \sum_{k \neq j} f(d_{jk}; \gamma) \ln(p_{kt}) + \epsilon_{jt}, \tag{1}$$

where α_j, β_j and γ correspond to preference parameters θ, and ϵ_{jt} is a consumer product-specific unobservable. The function f is a real-valued transformation of the pairwise distances among products we compute from the embedding; we discuss specific parameterizations of this function below.

The log-linear formulation we adopt is convenient because the coefficients on log prices can be interpreted directly as elasticities: β_j is the own-price elasticity for product j, and the cross elasticity between products j and k is a function of their distance d_{jk}. This functional form is restrictive, as it rules out that cross elasticities between two products depend on the availability of other close substitutes. We adopt it here because of its simplicity and obvious computational advantage: elasticities can be obtained from simple linear or nonlinear regressions once a functional form for f has been chosen and suitable identifying assumptions have been made.[7] This is in contrast with state-of-the-art implementations of discrete-choice demand models, which instead require computationally intensive nonlinear optimization routines.

Although the log-linear specification is convenient for showcasing our method, other specifications of the model that incorporate distances are possible and may be preferable depending on the application at hand. First, from an econometric perspective, while the log-linear specification models demand as a regression—with one structural error per equation—this is a strong restriction that is relaxed in more flexible classes of models (Berry and Haile 2021). Embedding data could however be used to discipline flexible models of *inverse* demand. Second, as the log-linear model lacks economic structure, it may be preferable to use a specification corresponding to a micro-founded demand system—to enable welfare analysis or to enforce certain theoretical properties. With this in mind, we discuss in Supplemental

[6] For example, the researcher can restrict substitution across categories of goods by modeling choice as a multistage budgeting problem, as in Deaton and Muellbauer (1980).

[7] Identification and estimation of the model are discussed in Section IIIA.

Appendix A.2 an alternative specification based on AIDS (Deaton and Muellbauer 1980).

Embeddings can also be used to estimate characteristics-space demand models. The natural way to use an embedding in a conventional logit-style demand model (like BLP) is to treat the products' coordinates in the embedding as characteristics (i.e., x variables in the consumer's indirect utility function). If an m-dimensional embedding is computed, then each of the m dimensions can be treated as a characteristic. Because each dimension of the embedding enters the model separately, this approach is more flexible than the one described above for the product space model; it allows the data to determine which dimensions of the embedding are most relevant to substitution.[8]

A disadvantage of this approach is that these are latent characteristics without any natural interpretation. However, we expect that in many cases the latent characteristics from a crowd-sourced embedding will give a better overall description of the products and their relationships to one another than could be obtained from observable characteristics. For example, the 2020 Toyota Camry and the 2020 MINI Clubman are very similar cars based on horsepower, fuel efficiency, passenger volume, and curb weight,[9] but we suspect consumers would not identify the two cars as being near each other in product space. In our cereal application, our survey appropriately indicates that Cocoa Pebbles are closer to Cocoa Krispies than Tootie Frooties, even though Tootie Frooties are closer based on sugar, fiber, and calories from fat.

II. Empirical Application: Embeddings and Data

We illustrate our method by estimating demand for breakfast cereals. Fairly rich data on cereals' nutritional and other characteristics are available, and this product category has been the subject of important studies on demand estimation (e.g., Nevo 2001). Given the nature of our method, we cannot perform a Monte Carlo exercise: it is not possible to recover from survey responses a synthetic product space. Instead, we use the existing characteristic data in this application as a yardstick to measure the usefulness of embedding data. In this section, we first describe the survey we used to collect the triplets data and then summarize the embedding that we compute from those data.

A. *Survey*

To obtain the triplet data needed to learn the embedding of cereal products, we conducted an online survey that asked respondents to make a series of product comparisons. Each page showed a reference product along with eight comparison

[8] In Section IIIA we discuss how to add similar flexibility to the log-linear model.

[9] The HP, MPG, volume, and weight specifications for the Camry (Clubman) are 203 (189), 34 (29), 100 (93), and 3241 (3235).

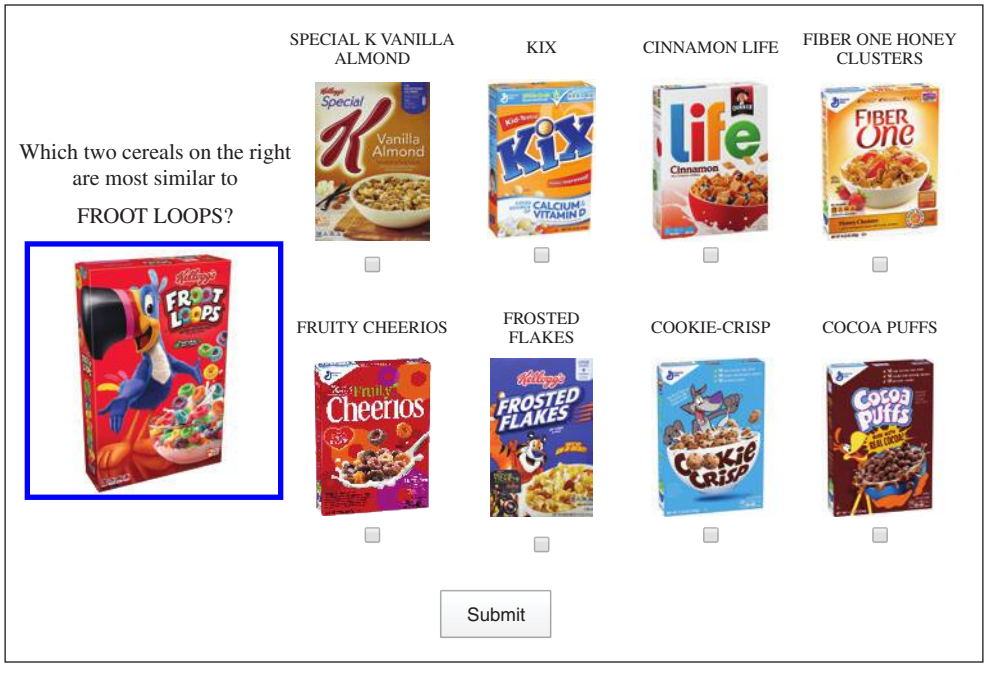

FIGURE 1. SAMPLE SURVEY PAGE

Note: The figure shows a sample page from our online survey.

products, and the respondent was asked to indicate which two were most similar to the reference product.[10] Figure 1 shows a sample page from the survey.

Each comparison page thus yields 12 triplets: each of the 2 checked products is considered closer to the reference product than the 6 unchecked products. Survey respondents were asked to complete up to 20 comparison pages, so each respondent generated as many as 240 triplet comparisons.

The survey respondents included 456 undergraduate students at the University of Wisconsin and 220 workers from Amazon's Mechanical Turk platform. Respondents were first asked to indicate how often they eat cereal and how many different cereals they have tried (see Figure 5 in Supplemental Appendix E), and were then shown the sequence of comparison pages. We found that embeddings based on Turk workers' responses versus undergraduate students' responses were similar,[11] so we pooled their responses when computing the embedding used in the rest of the paper. We discarded data from a very small percentage of respondents who indicated no prior experience with breakfast cereal, but this has little

[10] This approach to obtaining triplet comparison is similar to Wilber, Kwak, and Belongie (2014)—see further discussion in Supplemental Appendix C.

[11] If we compute embeddings separately for the two samples, the resulting product distances have a correlation of 0.88.

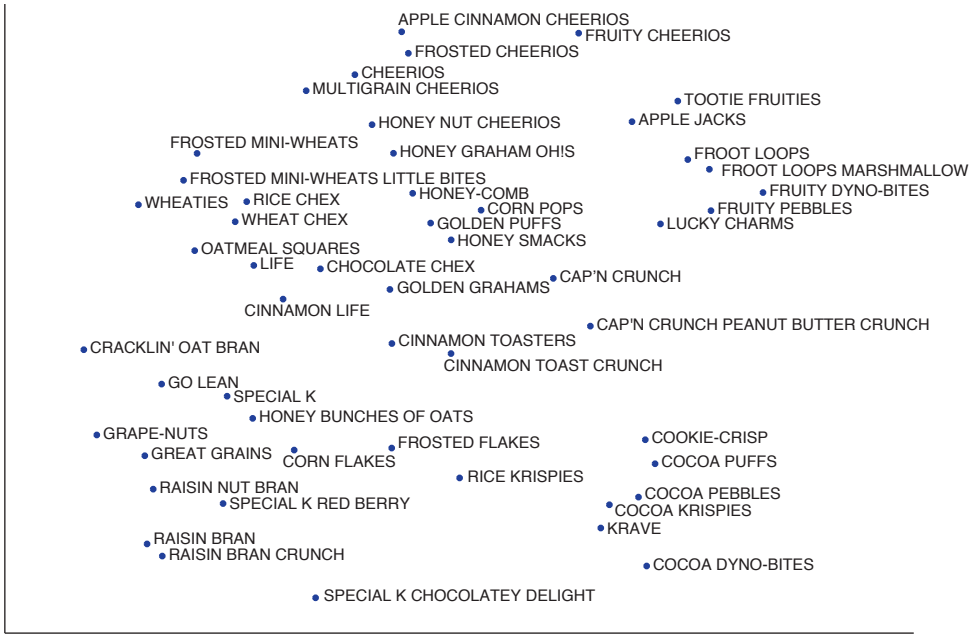

FIGURE 2. PLOT OF TWO-DIMENSIONAL EMBEDDING

Notes: The figure shows a two-dimensional embedding for ready-to-eat cereals estimated from the triplets data. To avoid overlapping text, not every cereal in the sample is plotted; plots showing all 86 cereals can be found in Supplemental Appendix E.

impact on the computed embedding. The final sample includes 175,116 triplet comparisons (Magnolfi, McClure, and Sorensen 2021).

B. *Computed Embedding*

For the demand estimation below we use six-dimensional embeddings,[12] but for purposes of visualization Figure 2 shows a two-dimensional embedding computed from the same triplets data. Even with only two dimensions, the algorithm neatly organizes the products into reasonable clusters—for example, sugary fruity cereals (clustered in the northeast region of the figure) and sugary chocolatey cereals (clustered in the southeast).

We report descriptive statistics for the six embedding characteristics in Table 1, panel A. The characteristics are mean-zero but do not have unit variance, and display some correlation. Based on distances from the six-dimensional embedding, Table 2 lists the two nearest cereals to some of the highest revenue brands in our sample. In general, the embedding appears to be correctly identifying the most similar products. This should not be surprising, since identifying similar cereals is not

[12] Supplemental Appendix B provides a detailed discussion of how we chose the number of dimensions.

Table 1—Embedding Characteristics and Characteristics Data

Variable	Mean	SD	Minimum	Maximum
Panel A. Summary statistics				
x_1	0.000	1.589	−3.219	3.385
x_2	0.000	0.894	−1.372	2.849
x_3	0.000	1.107	−2.514	1.998
x_4	0.000	1.083	−1.998	2.347
x_5	0.000	1.171	−3.059	2.280
x_6	0.000	0.826	−1.700	2.123

	Sugar	Fiber	Cal. from fat	Kids	All-family
Panel B. Pairwise correlations					
x_1	−0.098	0.588	0.265	−0.694	0.647
x_2	−0.009	−0.176	−0.089	0.273	−0.272
x_3	−0.209	0.164	−0.125	−0.332	0.314
x_4	−0.192	0.225	0.138	−0.338	0.343
x_5	−0.458	0.155	0.028	−0.507	0.491
x_6	0.495	−0.302	0.177	0.418	−0.404

Notes: Panel A reports summary statistics for embedding vectors, denoted x_1 through x_6. Panel B reports the pairwise correlation between observable characteristics (column variable) and embedding characteristics (row variable).

Table 2—Examples of Nearby Brands Based on Six-Dimensional Embedding

Brand	Nearest brand	Second-nearest brand
GM Honey Nut Cheerios	GM Honey Nut Cheerios Medley Cr.	Post Honey Graham Oh's
Kellogg's Frosted Flakes	Malt-o-Meal Frosted Flakes	Kellogg's Corn Flakes
GM Cinnamon Toast Crunch	GM French Toast Cr.	Malt-o-Meal Cinnamon Toasters
Kellogg's Froot Loops	Malt-o-Meal Tootie Fruities	Kellogg's Apple Jacks
Kellogg's Raisin Bran	Kellogg's Raisin Bran Cr.	Post Raisin Bran
Kellogg's Rice Krispies	GM Kix	Kellogg's Corn Pops
GM Cocoa Puffs	Kellogg's Cocoa Krispies	GM Reese's Puffs

Note: The table reports, for the sample of ready-to-eat cereal brands in the first column, the nearest and second-nearest brands in the six-dimensional embedding.

difficult for a human, and our procedure is essentially synthesizing thousands of comparisons made by humans.

Our interpretation of the computed embedding deserves some discussion. A first natural question is whether embedding characteristics correspond to intuitive dimensions of product differentiation. We discuss this aspect further below, after having introduced data on observable characteristics. Another question concerns whether the product distances that come from the survey are the right distances for demand estimation. It is important to note that the distances themselves are not intended to be measures of substitution. Like ordinary product characteristics in conventional discrete-choice methods, they are *inputs* into the demand estimation, which uses price and quantity data to measure substitution patterns. Ideally, we want the demand estimation to use these inputs as flexibly as needed to deliver the true substitution patterns—much as allowing for random coefficients on product characteristics allows for flexible substitution in the discrete-choice framework—so it may not be enough to simply use Euclidean distances. When estimating the demand model in subection A below we discuss how to incorporate the distances more flexibly.

C. *Observed Characteristics Data*

For all cereals in our sample, we have data on nutritional characteristics from the Nutritionix database (Nutritionx 2021). These include sugar (grams per serving), fiber (grams per serving), and calories from fat (per serving).[13] This set of characteristics is similar to Nevo (2001), only omitting his "mushy" characteristic, which was author generated and not recoverable from nutritional data. We also construct a categorical variable to reflect differences in the target demographic, labeling each brand as targeted to "Kids," "Adult," and "All family." We emphasize that these data are not necessary to estimate demand with our method. However, they provide a useful benchmark: we use them here to help interpret embedding characteristics and in Section III to estimate demand systems that can be compared with those that use embedding characteristics.[14]

To compare observed and embedding characteristics, we show pairwise correlations in Table 1, panel B. Embedding characteristics (x_1 through x_6) are partially picking up variation in observables: for example, x_6 is the embedding characteristics most positively correlated with sugar, and is also positively correlated with calories from fat and kids, while negatively correlated with fiber and all-family. x_1 is instead highly correlated with fiber, fat, and all-family, while negatively correlated with kids. Beyond pairwise comparisons, we analyze how the full set of embedding characteristics is related to the set of observed characteristics by computing canonical correlations.[15] The first two canonical correlations between the (six) embedding dimensions and (five) observed characteristics are 0.87 and 0.67,[16] and the canonical loadings indicate that the first canonical variate is highly correlated with the category indicators (kids and all family) while the second is highly correlated with sugar and calories from fat. Finally, we run the full set of embedding and observable characteristics through PCA: retaining at least 95 percent of the variation requires 8 components versus the 9 total continuous dimensions. Taken together, these results indicate that while embedding vectors may be reflecting some product features that are observable, they are also encoding additional information.

D. *Price and Quantity Data*

Our data on prices and quantities come from Nielsen's Retail Scanner data from the year 2017 (Nielsen IQ 2024b). The unit of observation in our analysis is a UPC-retailer-DMA-week. Our sample of UPCs consists of the highest selling UPCs for the 86 brands that together account for 80 percent of total sales in the breakfast cereal category. We focus on large markets with many competing products, limiting the sample by keeping product-market combinations that appear in all 52 weeks of the data, keeping markets with at least 50 UPCs, and keeping UPCs that appear in at

[13] We rescale the sugar, fiber, and calorie measures to have mean zero and unit variance.

[14] We can also combine observed characteristics with embedding data—see Section C..

[15] The first canonical correlation between the matrix of embedding characteristics $\boldsymbol{x}$ and observables $\boldsymbol{y}$ is $\max_{a\in\mathbb{R}^6, b\in\mathbb{R}^4} \operatorname{corr}(a'\boldsymbol{x}, b'\boldsymbol{y})$. See Härdle and Simar (2019) for an overview of canonical correlation analysis.

[16] The remaining canonical dimensions are not statistically significant.

Table 3—Price and Quantity Summary Statistics

	Mean	SD	Percentiles 10th	50th	90th
Cereal products ($N = 86$)					
Average price	3.58	0.83	2.50	3.51	4.77
Average weekly sales	216.16	649.63	9	55	480
Number of retailers	153.06	33.27	105	165	189
Retailer-DMA pairs ($N = 189$)					
# of cereal products carried	69.65	8.20	57	72	79
Avg. weekly cereal revenues (000)	45.30	77.48	3.71	17.15	117.46

Note: The table reports summary statistics for the 86 cereal UPCs and 189 retailer-DMA pairs we use for demand estimation.

least 50 markets. This results in a sample of 684,476 UPC-retailer-DMA-week observations, containing 43 retailer chains, 111 DMAs, and 189 unique retailer-DMA pairs across 52 weeks. Table 3 shows some basic summary statistics for the 86 products in the sample, as well as for the 189 retailer-DMA pairs.

III. Empirical Application: Demand Estimation

In this section we provide details of the demand estimation for both the product and the characteristics space models. In each case, we emphasize the comparison to demand estimates from the same model *without* the use of an embedding—i.e., either using pairwise product distances computed from observable characteristics in the product space model, or using observable characteristics as the x variables in the discrete choice model.

A. *Log-Linear Demand Estimates*

We use the Nielsen price and quantity data to estimate the linear model shown in equation (1), including fixed effects at the week, DMA, retailer, and product level. We obtain the pairwise product distances d_{jk} from the embedding data. Theory suggests that the function $f(d_{jk};\gamma)$ should be monotonically decreasing in d_{jk}, since more distant products should have lower substitution. While certain functional forms such as $f(d_{jk};\gamma) = \gamma/(1 + d_{jk})$ can easily incorporate this, there are important reasons to estimate $f(d_{jk};\gamma)$ flexibly. First, it allows estimated substitution patterns to be driven more by the sales data than by the embedding. Second, if we are unsure whether the embedding is returning reasonable product distances, a flexible distance function provides a validation method. If the estimated distance function is non-monotonic or flat, this suggests the embedding is doing a poor job of capturing product attributes that are relevant to substitution.

We experimented with flexible approaches, including sieves and b-splines, but found that a simple cubic polynomial in scaled distances worked well:

$$f(d_{jk};\gamma) = \gamma_0 + \gamma_1 d_{jk} + \gamma_2 d_{jk}^2 + \gamma_3 d_{jk}^3. \tag{2}$$

If the distances d_{jk} are Euclidean, the log-linear model described by equations (1) and (2) can be estimated by OLS. However, a more flexible approach is to compute distances in a way that allows different dimensions of the embedding to have different weights. For instance, the substitutability of two products may depend weakly on how close they are in the first dimension of the embedding while depending strongly on how close they are in the second dimension. Since the dimensions of the embedding do not have natural interpretations, we may want to let the data determine which dimensions matter most for substitution. With this consideration in mind, we estimate the log-linear model with the same cubic polynomial distance function as in equation (2), but defining pairwise product distances as

$$\tilde{d}_{jk} = \left[\sum_m \omega_m \left(x_{jm} - x_{km}\right)^2\right]^{\frac{1}{2}}, \tag{3}$$

with ω_1 (the weight on the first dimension) normalized to one, and the remaining ω_m coefficients left as parameters to be estimated for all other embedding dimensions m. A disadvantage of this modification is that the regression is no longer linear in the parameters, so it must be estimated by nonlinear least squares. This increases the computational burden, but can still be done with a single line of code (e.g., using Stata's nls command). Estimating the ω weights does seem to matter: the distances in two dimensions of the embedding are estimated to be more important than the others (ω_4 and ω_5 are above 1.5), and one dimension seems to hardly matter at all (ω_3 is near zero).[17]

As noted in Berry and Haile (2021), the identification of demand for differentiated products is complicated by two fundamental challenges: price endogeneity, and codependence of the demand for each product on the latent demand shocks for all other products in the market. For the purpose of showcasing the embedding data and identifying substitutes in the simplest possible context, we impose strong assumptions to set aside these challenges. In particular, we assume that prices and product distances are uncorrelated with the unobservable ϵ_{jt}. Coupled with the restrictions implied by the specification of equation (1), the assumption of exogenous prices allows us to estimate the model straightforwardly with either OLS or NLS, depending on whether we estimate weights in the distance function as in equation (3).

Alternatively, a researcher can use instruments to identify the log-linear model. In our specification, we need instruments not only for own price, but also for the prices of all other products. In principle, Hausman or BLP instruments can be used in this context.[18] To simplify estimation, we proceed with the assumption that prices are exogenous, leaving the discussion of an IV specification to Supplemental Appendix A.1. If the researcher wants to estimate a log-linear (or similar) specification using instruments, having distances from embeddings limits the number of parameters to be estimated and thus makes it easier for the exogenous variation provided by the instruments to pin down the parameters of the model.

[17] Appendix Table 2 of Supplemental Appendix E reports the full set of estimated parameters, along with a comparison to the estimates from a model that uses Euclidean distances. Using Euclidean distances yields similar elasticity estimates but a worse overall fit.

[18] However, in a setting like ours with high-frequency data, these instruments may not generate estimators with good sampling properties, as argued by Rossi (2014).

In evaluating our results, the main comparison we want to make is to an alternative specification that relies only on observable characteristics to compute pairwise product distances. That is, we estimate the same log-linear model (1) and compute non-Euclidean product distances d_{jk} as in equation (3), but using the nutritional characteristics introduced in Section C. We use the same cubic polynomial distance function as in (2) but also add a term to reflect different categories of cereals: letting $G_j \in \{\text{Kids, Adult, All Family}\}$ denote the category of cereal j, the modified distance function is

$$f(d_{jk};\gamma) = \gamma_0 + \gamma_1 d_{jk} + \gamma_2 d_{jk}^2 + \gamma_3 d_{jk}^3 + \gamma_4 \mathbf{1}\{G_j \neq G_k\}.$$

We expect γ_4 to be negative, as two products in different categories should be less substitutable than two in the same category. As with the benchmark model based on the embedding, we estimate weights on different characteristics when computing the pairwise product distances d_{jk}.

In the absence of product characteristics, we could have compared our method with a standard log-log product space demand specification, which requires in this application estimating $J^2 = 7{,}896$ elasticity parameters. Even with our large weekly dataset, this specification produces imprecise and implausible results, with only a small share of cross-elasticity parameters being positive and statistically significant.[19]

Comparison between Embeddings and Characteristics.—Figure 3 summarizes the distributions of estimated own- and cross-price elasticities from the two specifications. The left panel shows kernel density estimates of the own-price elasticities. These estimates fall in a reasonable range (all negative, mostly between -1 and -4) and are similar between the two specifications. The similarity is not surprising, as own-price elasticity estimates are driven largely by the price and quantity data. Where the two models differ is in the estimated substitution patterns, which depend on the estimated distance functions.

We show the estimated distance functions in the right panel of Figure 3. When product distances are computed from the embedding, this function has the expected monotonically decreasing shape: nearby products are estimated to have larger cross-price elasticities. When distances are computed from observable characteristics, the estimated distance function is non-monotonic and overall relatively flat, implying that cross-price elasticities for the nearest products are hardly different from those for the most distant products. This feature results in systematically higher diversion to closer products when using embedding data, as illustrated in Appendix Figure 6. In Supplemental Appendix D we assign cereals to intuitive groupings (e.g., chocolate-flavored, high-sugar cereals; low-sugar cereals; etc.) and show that embedding-based estimates predict diversion to be higher within group than across groups, which is what we would intuitively expect. Estimates based on observed characteristics do not generate this pattern.

[19] Additionally, this specification has heavy computational requirements, as it uses 142.6GB of RAM and takes approximately 12 hours to run on a Linux server.

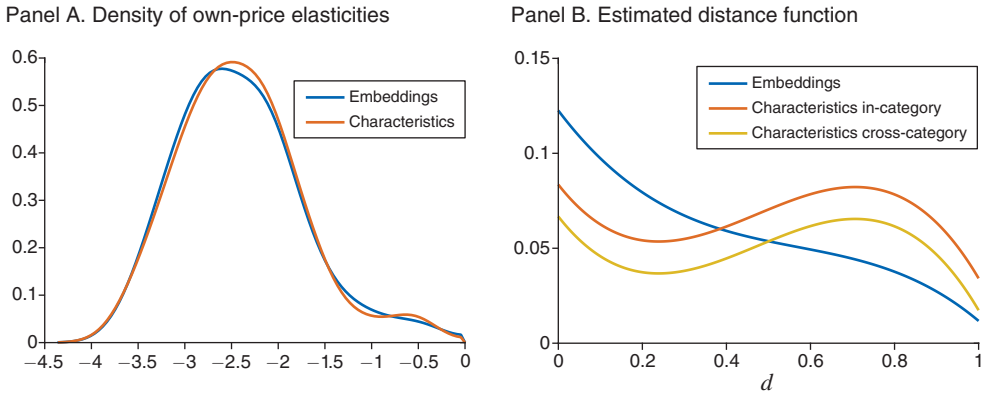

FIGURE 3. ELASTICITY ESTIMATES FOR THE LOG-LINEAR MODEL

Notes: Panel A shows the density of own-price elasticities β_j for the log-linear model (equation (2)). Panel B shows $f(d)$ of equation (3) implied by the estimated γ parameters. The two parallel distance functions for the model based on observed characteristics represent estimated distances for products in the same versus different categories (Kids, Adult, All Family).

The two specifications differ not only in the amount of diversion to close products but also in how they identify close substitutes. To illustrate this point, Table 4 shows estimated cross-price elasticities for two pairs of very similar cereals. The model that uses the embedding delivers relatively high cross elasticities between Honey Nut Cheerios and Honey Graham Oh!s (0.090) and between Cocoa Pebbles and Cocoa Krispies (0.107), and relatively low cross elasticities between dissimilar pairs (e.g., 0.054 between Honey Nut Cheerios and Cocoa Pebbles). By contrast, the model that uses observed characteristics produces cross elasticities in a narrow range (0.041 to 0.068 for the example products in the table).

In addition to generating more plausible elasticity estimates, the specification based on the embedding also delivers superior fit out-of-sample. We use the coefficients estimated using the 2017 Nielsen data sample to predict sales for the same markets in 2018.[20] When compared to the characteristics specification, the embedding specification produces log sales predictions that are closer to the data: the RMSEs corresponding to the two models are 1.227 versus 1.058, respectively. For reference, the RMSE of a constant prediction is 1.487.

B. *BLP Demand Estimates*

To show how the embedding can be used in a characteristics space model, we estimate a standard BLP demand system similar to Nevo (2001) and Backus, Conlon, and Sinkinson (2021). We refer the reader to those articles for a full description of the micro-foundations of the model. In this section, we briefly describe the details of our implementation. Our specification includes retailer-DMA-week (market)

[20] We also apply the retailer-DMA-week fixed effect from 2017 to the same week in 2018 (i.e., the FE for 2017w3 in a given retailer-DMA is applied to 2018w3 in the same retailer-DMA); this aims to capture seasonality.

TABLE 4—COMPARISON OF ELASTICITIES BETWEEN SIMILAR PRODUCTS—LOG-LINEAR MODEL

Cereal		1	2	3	4
Honey Nut Cheerios	1	−2.950	0.090	0.054	0.055
		−2.899	0.039	0.063	0.058
Honey Graham Oh!s	2	0.090	−1.807	0.052	0.053
		0.039	−1.637	0.041	0.045
Cocoa Pebbles	3	0.054	0.052	−3.322	0.107
		0.063	0.041	−3.283	0.065
Cocoa Krispies	4	0.055	0.053	0.107	−2.527
		0.058	0.045	0.065	−2.432

Notes: In each cell the table reports elasticities e_{jk} corresponding to the row model j and the column model k. Cells contain elasticities for the log-linear model of equation (1), with distances based on the embedding (on top) or on observed characteristics (on bottom).

and product fixed effects as variables that enter consumers' indirect utility linearly. These fixed effects capture some unobserved determinants of utility, as in Nevo (2001). Variables that have a nonlinear impact on demand are price, the constant, and product characteristics: either the nutritional variables (in the case of observable characteristics), or the coordinates from a six-dimensional embedding calculated from the survey triplets. For convenience, we will refer to the former model as Characteristics BLP and the latter as Embedding BLP. Aside from the different characteristics (x variables), everything in the two specifications is identical.

The effect of nonlinear variables on demand is modeled via random coefficients in the indirect utility of a household i. These coefficients are $\beta_i \sim N(\beta + \Pi D_i, \Sigma)$, where β and Π are vectors of parameters, and D_i are demographic characteristics of household i. We estimate the diagonal elements in Σ corresponding to each nonlinear variable.[21] The model includes demographic interactions ΠD_i for prices and the nutritional variables (or embedding coordinates), with log household income and an indicator for the presence of children in the household as the included demographics.[22] We estimate a log-normal income distribution with/without kids and a binomial distribution for the presence of kids from the households in the Nielsen Consumer Panel data (Nielsen IQ 2024a). Values of D_i correspond to 200 Halton draws per market from these distributions.

Instruments are needed to identify and estimate this model. To this aim, we create the quadratic differentiation IVs of Gandhi and Houde (2019). For $\delta_{jk}(l) = x_{jl} - x_{kl}$, given characteristic l and products j, k, define

$$z_{jt}^{quad} = \left\{ \sum_k \delta_{jk}^2(l), \sum_k \delta_{jk}(l) \times \delta_{jk}(\ell) \right\} \quad \forall (l, \ell), \tag{6}$$

where l, ℓ are the nonlinear characteristics (price and observable characteristics or embedding coordinates). We then follow Backus, Conlon and Sinkinson (2021) in interacting these variables with moments of the demographics in each market, taking

[21] In the Characteristics BLP some values of Σ were consistently estimated to be near zero, so in the final specification we set them to zero to aid convergence.

[22] We exclude the interaction of demographics and the constant as this largely drives outside shares, and we have already calibrated market size at the market level.

the tenth, fiftieth, and ninetieth percentile incomes for households with and without children as well as the percentage of households with children, obtaining thus a total of 168 instruments. After using these to estimate the model using 2-step GMM, we then adopt the approximation to the optimal instruments of Reynaert and Verboven (2014). To estimate a discrete-choice demand model we also need to specify market size. We follow Backus, Conlon, and Sinkinson (2021) in estimating the market size as the number of individuals entering the retailer, using variation in purchases of staple products (milk and eggs) as predictors.

To keep computation manageable, we estimate this model on a subsample of our data. We limit the sample successively to (i) the top 15 DMAs by market sales, (ii) the top 15 retailers within that set of DMAs, and (iii) a random set of 20 weeks. Our final subsample for BLP estimation contains 32,385 observations, with 540 unique retailer-DMA-week markets.

Comparison between Embeddings and Characteristics.—We compute parameter estimates and standard errors (conditional on the embedding)[23] with pyBLP (Conlon and Gortmaker 2020) and report them in Table 5. The two specifications deliver similar results in most respects: price coefficients are negative and significant, the interactions of price and income are positive and significant, and the random coefficients on the constant and on price are statistically significant. More importantly, the implied elasticities are similar in magnitude and positively correlated: the median own-price elasticity in the Characteristics BLP is -2.453, versus -2.477 in the Embedding BLP;[24] and the correlation between own-price elasticities is 0.972. For cross-price elasticities, the medians are 0.014 and 0.010 (respectively), and the correlation is 0.762.

Table 6 shows own- and cross-price elasticity estimates for the same examples as in Table 4 above. As with the log-linear product space model, the embedding specification delivers more plausible substitution patterns. For similar cereals, cross-elasticities from the Embedding BLP are higher than from the Characteristics BLP (e.g., 0.275 versus 0.132 for the cross-elasticity between Honey Graham Oh!s and Honey Nut Cheerios) and for dissimilar cereals, they are lower (e.g., 0.030 versus 0.037 for the cross-elasticity between Honey Graham Oh!s and Cocoa Krispies). Both specifications show evidence of logit-style substitution patterns, with generally higher diversion to products with high market shares (e.g., Honey Nut Cheerios), but less so for the Embedding BLP.

Supplemental Appendix Figure 7 shows that using embedding data generates slightly higher diversion to close substitute products. Also, the exercise in Supplemental Appendix D shows that when cereals are assigned to intuitively defined groups (e.g., chocolate-flavored high-sugar), the Embedding BLP predicts higher diversion within group than across groups, whereas the Characteristics BLP does not.

[23] It is, in theory, possible to bootstrap both the embedding and BLP estimation. However, this procedure is expensive computationally and requires costly human oversight to check the convergence of pyBLP in each bootstrap sample. Moreover, the loss of precision in the BLP coefficients due to sampling of the survey triplets is likely to be small: embeddings are very similar when we bootstrap the triplets data–see Appendix C for more details.

[24] For comparison, Backus, Conlon, and Sinkinson (2021) get a median own-price elasticity of -2.665.

TABLE 5—ESTIMATED COEFFICIENTS OF BLP MODEL

Parameter	Variable	Characteristics	Embeddings
β	Price	−2.667	−3.099
		(0.363)	(0.300)
Σ	Constant	3.765	4.199
		(1.271)	(0.454)
	Price	0.820	0.947
		(0.036)	(0.037)
	x_1	–	0.015
			(0.192)
	x_2	0.016	0.000
		(0.026)	(0.402)
	x_3	–	0.000
			(0.095)
	x_4	0.090	0.842
		(0.099)	(0.195)
	x_5	–	0.000
			(0.193)
	x_6	–	1.573
			(0.186)

Π		Income	Kids	Income	Kids
	Price	0.121	−20.950	0.142	−0.097
		(0.035)	(0.000)	(0.027)	(0.064)
	x_1	−0.169	−0.948	0.139	−0.081
		(0.018)	(0.000)	(0.015)	(0.021)
	x_2	0.135	–	0.058	−0.087
		(0.019)		(0.027)	(0.034)
	x_3	0.003	–	−0.139	0.072
		(0.019)		(0.023)	(0.033)
	x_4	0.060	–	0.029	−0.151
		(0.140)		(0.021)	(0.039)
	x_5	0.104	–	0.057	−0.107
		(0.137)		(0.018)	(0.025)
	x_6	–	–	−0.153	0.226
				(0.033)	(0.044)

Nonlinear variables	Observables	6D embedding
Median own-price elasticity	−2.453	−2.477
Median outside diversion	0.285	0.310

Notes: The table reports estimates (on top) and standard errors (below) for the parameters of the BLP model. Observable characteristics x_1 through x_5 refer to sugar, fiber, calories from fat, and indicators for whether the cereal is for kids or an all-family cereal. $n = 32{,}385$.

The Embedding BLP thus delivers elasticity estimates that are arguably more plausible—and at the very least similar—to what we obtain from observed characteristics. Additionally, the Embedding BLP specification has lower out-of-sample RMSE as compared to the Characteristics BLP; when predicting 2018 market shares, the fit is better for both inside products (0.0290 versus 0.0294) and outside share (0.0527 versus 0.0750). We take this as an encouraging result because it means we can obtain credible estimates of demand *even when observable characteristics are unavailable*—a challenge we believe is inherent to many markets of interest.

Table 6—Comparison of Elasticities between Similar Products—BLP

Cereal		1	2	3	4
Honey Nut Cheerios	1	−2.483	0.042	0.014	0.023
		−2.541	0.020	0.018	0.030
Honey Graham Oh!s	2	0.275	−2.879	0.029	0.030
		0.132	−2.802	0.038	0.037
Cocoa Pebbles	3	0.076	0.024	−2.589	0.063
		0.099	0.031	−2.586	0.027
Cocoa Krispies	4	0.081	0.016	0.040	−2.303
		0.108	0.020	0.017	−2.449

Notes: In each cell the table reports elasticities e_{jk} corresponding to the row model j and the column model k. Cells contain elasticities from the Embedding BLP model on top, and Characteristics BLP on the bottom.

IV. Discussion and Extensions

A. *Log-Linear Demand versus BLP*

For various reasons we noted above, mixed logit models like BLP have become the standard for estimating rich demand systems in differentiated product markets. However, in some contexts—most notably, in analyses of antitrust cases conducted by the DOJ or FTC—researchers need to obtain demand estimates more simply and more quickly than is feasible within the BLP framework. The results of our empirical exercise are encouraging in this regard. The log-linear (product space) model that uses distances from the embedding delivers estimates of substitution patterns that, for some pairs of products, are arguably even more plausible than those from BLP in our application. To illustrate this point, we report in Table 7 the two closest substitute brands for each of the ten most popular cereals in our sample, according to the log-log model and according to the BLP model, both estimated with embedding data. The simple log-linear model, augmented with crowd-sourced data on products' locations, does a good job of recovering sensible patterns of price substitution,[25] even though it is substantially easier to estimate than state-of-the-art BLP demand systems.[26]

B. *An Embedding Based on Purchase Correlations*

The ideal scenario for a researcher aiming to estimate substitution patterns is to have price and quantity data paired with actual data on consumers' second (and third and fourth …) choices (see, e.g., Berry, Levinsohn, and Pakes 2004). Such

[25] The log-linear specification is instead ill-suited to predict substitution in characteristics.

[26] Estimates of the log-linear demand models took less than 20 seconds to compute on a Windows desktop. Estimates of the BLP models took over 100 times longer, even when using the limited sample and running on a powerful Linux server. But differences in computation time understate the overall difference in time and complexity between the two approaches, since arriving at reliable BLP estimates requires considerable back-and-forth on things like start values, scaling, etc., even with the aid of helpful software packages like pyBLP.

Table 7—Examples of Nearby Brands (New)

Brand	Model	Closest brand	Second-closest brand
Honey Nut Cheerios (HNC)	Log-log	HNC Medley Crunch	Honey Graham Oh!s
	BLP	Cheerios	Lucky Charms
Kellogg's Frosted Flakes	Log-log	MoM Frosted Flakes	Honey Bunches Of Oats
	BLP	Honey Nut Cheerios	Kellogg's Raisin Bran
Cinnamon Toast Cr.	Log-log	French Toast Cr.	Cinnamon Toasters
	BLP	Reese's Puffs	Honey Nut Cheerios
Cheerios	Log-log	Multigrain Cheerios	HNC Medley Crunch
	BLP	Honey Nut Cheerios	Rice Krispies
Honey Bunches Of Oats	Log-log	Special K Oat and Honey	Kellogg's Frosted Flakes
	BLP	Honey Nut Cheerios	Cinnamon Toast Cr.
Lucky Charms	Log-log	Apple Jacks	Froot Loops
	BLP	Honey Nut Cheerios	Cheerios
Frosted Mini-Wheats (FMW)	Log-log	Oatmeal Squares	Cinnamon Life
	BLP	Raisin Bran	FMW Little Bites
Froot Loops	Log-log	Apple Jacks	Tootie Fruities
	BLP	Honey Nut Cheerios	Frosted Flakes
Kellogg's Raisin Bran	Log-log	Raisin Bran Cr.	Post Raisin Bran
	BLP	Frosted Mini-Wheats	Frosted Flakes
Rice Krispies	Log-log	Golden Puffs	Golden Crisp
	BLP	Cheerios	Honey Nut Cheerios

Note: The table reports, for the ten ready-to-eat cereal brands with the highest market share (in the first column), the nearest and second-nearest brand substitutes according to the log-log demand system and to the BLP demand system estimated with embedding data.

second-choice data can be used to generate additional moments that, when combined with the BLP moment conditions, discipline the estimates to better predict actual patterns of substitution. While we do not have second-choice data for our empirical application to cereal, we can borrow an idea from Atalay et al. (2022) that uses Nielsen's Consumer Panel data to learn which products households consider to be substitutes.

Atalay et al. (2022) use the Consumer Panel data to determine sets of products that are ever purchased by the same household across a large number of shopping trips, and then gauge the substitutability of a given pair of products by how commonly the two products are purchased by the same household. The underlying premise is that if individuals within each household have preferences over products' characteristics and these preferences are stable over time, then temporary changes in relative prices (e.g., due to periodic sales or stockouts) will induce consumers to occasionally purchase substitutes for their preferred product. In our case, if a household sometimes purchases Frosted Flakes and sometimes purchases Froot Loops, but never purchases Raisin Bran, the implication is that Froot Loops is a closer substitute to Frosted Flakes than Raisin Bran for that household.

This idea is formalized by constructing a dissimilarity matrix $\mathbf{D}$ with $1 - \rho_{jk}$ as its (j,k)-th element, where ρ_{jk} is the pairwise purchase correlation between products j and k—i.e., a measure of how likely a household is to have ever purchased product k conditional on having ever purchased product j. An embedding can then be computed based on this dissimilarity matrix; we do this using the tSNE algorithm

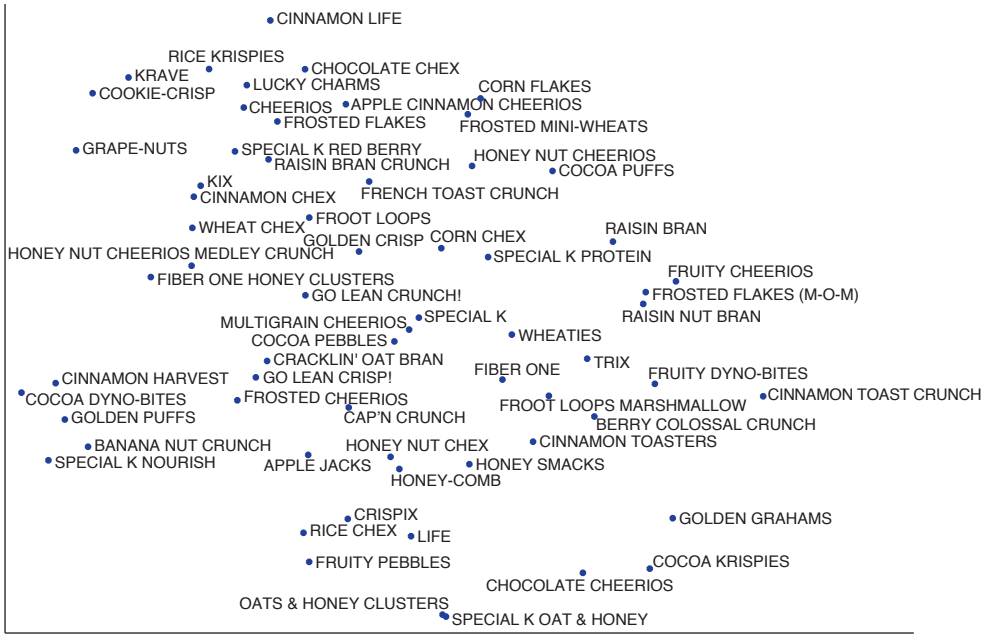

FIGURE 4. TWO-DIMENSIONAL EMBEDDING BASED ON CONSUMER PANEL

Notes: The figure shows a two-dimensional embedding for ready-to-eat cereals estimated from the Consumer Panel micro-data. To avoid overlapping text, not every cereal is included in the plot.

Van der Maaten and Hinton (2008).[27] A two-dimensional embedding is shown in Figure 4. As with the embedding based on the survey triplets, it mostly clusters similar products together, such as sugary cereals in the northwest quadrant.

If we estimate our product-space demand model using distances from this alternative embedding, we get reasonably similar estimates of products' own- and cross-price elasticities. The magnitudes are similar,[28] and more importantly, they are positively correlated with the elasticities we estimate using the embedding based on survey triplets: the correlation of the own-price elasticities is 0.979, and of cross-price elasticities is 0.491.

Thus, the distances from the survey-based embedding deliver results similar to those that would result from an embedding computed from micro-data on consumers' actual choices. We interpret this as further validation of our approach. When data that directly reflect consumers' substitution choices are available (e.g., second-choice survey data as in Grieco, Murry, and Yurukoglu 2021 or household panel data as in Atalay et al. 2022), it certainly makes sense to use those data. But in the absence of such data, our method is a viable alternative.

[27] tSNE (*t*-distributed stochastic neighbor embedding) is analogous to tSTE, except that instead of triplets it uses feature data or (in our case) data on products' distances or dissimilarities to compute the embedding.

[28] The mean own-price elasticity is −2.42 and the mean cross-price elasticity is 0.053, identical to when we use the embedding based on survey triplets.

C. *Using an Embedding in Combination with Observed Characteristics*

For the purposes of this study, we have emphasized comparisons between estimates based on product embeddings and estimates based on observed characteristics alone. In practice, however, when data on observed characteristics are available we might want to use them in combination with an embedding. Survey triplets could be used in such cases to learn additional latent characteristics beyond the characteristics that are already observed.

The most straightforward way to combine the two approaches is to compute what we call a mixed embedding. If we have data on K product characteristics, we can compute a $(K + m)$-dimensional embedding with our triplets data by solving the optimization program (1) with the first K columns of the embedding matrix fixed at the known data values. This forces the algorithm to find the remaining m columns to rationalize the triplets data *after* accounting for the K observed characteristics. We implemented this approach for our empirical example and found that it works well, mostly delivering elasticity estimates close to those that come from using the embedding alone. Detailed results are shown in Appendix Table 3 in Supplemental Appendix E.

D. *Relation to Existing Approaches*

Three combined elements distinguish our approach: the use of survey data; considering the product space x_t as unobserved or latent; and a sequential procedure whereby we first recover x_t in a model selection step, and then estimate preferences θ. We discuss these aspects in turn, noting how they relate to existing approaches.

Our use of surveys is reminiscent of two other common uses of survey data in demand analysis. First, conjoint surveys (see Allenby, Hardt, and Rossi 2019, for an overview) ask respondents to rate the desirability of each product in a set of hypothetical offerings, and response data are then used to estimate preferences for the products' observed attributes. Second, a well-known literature has shown the value of using second-choice data from surveys to generate additional moments for estimating conventional discrete-choice models like BLP (Berry, Levinsohn, and Pakes 2004; Conlon, Mortimer, and Sarkis 2022; Grieco, Murry, and Yurukoglu 2021). In both of these cases, surveys are used *together with* information and/or data the researcher already has about product characteristics. By contrast, our survey aims to learn about the product space in situations where the demand-relevant product characteristics are difficult to identify and/or quantify.

Thus, our approach is not a substitute for demand estimation with second-choice data. In situations where the researcher already has the data necessary to estimate a discrete choice model based on observed characteristics, getting additional second-choice data from a survey designed to ask about preferences is an excellent idea, and recent studies have shown that in some contexts it can be easy and inexpensive to do so (Conlon and Gortmaker 2023). If instead data on products' characteristics are not available, our method can recover a representation of the latent product space, which can then be used to estimate preferences. In principle, a researcher may use survey data on stated preferences (e.g., product rankings) or choice-based surveys even when product characteristics are unobserved, to either

recover preferences directly or additionally estimate the product space. There is an intuitive trade-off: asking about products may elicit low-quality responses when individuals are asked about products they almost never choose, but on the other hand it may be hard to learn about the less popular products when asking consumers about their personal rankings of top (or frequently purchased) products. In general, the relative performance of these approaches is an open question that we leave for future research.

Finally, our study is related to earlier work by Goettler and Shachar (2001), who also aim to recover both product attributes and preferences. To do so, they rely on panel data on consumers' television viewing choices in combination with a bliss-point model of demand and simultaneously estimate television shows' latent attributes along with consumers' preferences for those attributes. Though similar in spirit to our exercise, we highlight a few key differences with our study. First, our method requires aggregate data on prices and quantities in addition to auxiliary survey data on the product space, as opposed to micro data on individual choices. Second, we implement a sequential procedure in which we keep the two data sources separate, instead of a joint estimation procedure.[29] While intuitively this may suggest some loss of econometric efficiency, we believe it comes with significant advantages. Our method is computationally simple and highly portable, as we rely on proven ML algorithms to recover the high-dimensional product space and then leverage the researcher's method of choice for demand estimation. Moreover, the method does not require significant econometric advances (such as characterizing sampling properties of tSTE or related algorithms) because we can estimate demand conditional on the product space (in the spirit of, e.g., Bonhomme, Lamadon, and Manresa 2022).

V. Conclusion

The demand estimation toolkit available to empirical researchers in industrial economics has seen many advances in the last few decades. In particular, we have learned how to specify, identify, and estimate more and more flexible models. In this paper we propose complementing these modeling innovations with a new source of data: triplet comparisons obtained from an online survey, which can be used to compute an embedding of the latent product space. To showcase the usefulness of the data, we use the embedding in conjunction with data on prices and quantities to estimate two specifications: a simple log-linear model of demand, and a BLP model. In an application to the ready-to-eat cereals market, our method produces estimates that compare favorably with those obtained using standard data on product characteristics.

Beyond our illustrative application, embeddings will be particularly valuable in empirical settings where characteristics are hard to observe or measure, thus making standard demand models hard to estimate. For example, our method could be used to estimate demand in the market for mobile apps. Recovering credible substitution

[29] This is in line with the approach in Armona, Lewis, and Zervas (2021), who use search data to estimate latent characteristics in a first step, and then estimate demand with price and quantity data.

patterns in this market is essential to answer policy-relevant questions about market power and the effects of consolidation, but conventional discrete-choice methods are hard to apply because demand-relevant characteristics of mobile apps are difficult to define and measure. Our method promises to be a useful alternative in this setting.

A common finding in machine learning is that small amounts of human input can yield large improvements in model performance (see, e.g., Ouyang et al. 2022). This has led to many "human-in-the-loop" approaches to ML modeling, of which crowdsourced labeling is but one example (see Mosqueira-Rey et al. 2023 for a survey.). Our paper makes a similar point in the context of estimating structural demand models: a little human input from a small-scale survey can go a long way in disciplining the model and improving its predictions. Since online tools have made it quite easy to design and implement surveys, we believe approaches like ours could become a standard component of the structural empirical researcher's toolkit.

REFERENCES

Allenby, Greg M., Nino Hardt, and Peter E. Rossi. 2019. "Economic Foundations of Conjoint Analysis." In *Handbook of the Economics of Marketing*, Vol. 1, edited by Jean-Pierre Dubé and Peter E. Rossi, 151–92. Amsterdam: Elsevier.

Armona, Luis, Gregory Lewis, and Georgios Zervas. 2021. "Learning Product Characteristics and Consumer Preferences from Search Data." Unpublished.

Atalay, Enghin, Erika Frost, Alan Sorensen, Christopher Sullivan, and Wanjia Zhu. 2022. "Large Scale Estimation of Demand and Markups." Unpublished.

Backus, Matthew, Christopher Conlon, and Michael Sinkinson. 2021. "Common Ownership and Competition in the Ready-to-Eat Cereal Industry." NBER Working Paper 28350.

Bajari, Patrick L., Zihao Cen, Victor Chernozhukov, Manoj Manukonda, Jin Wang, Ramon Huerta, Junbo Li, et al. 2021. "Hedonic Prices and Quality Adjusted Price Indices Powered by AI." Unpublished.

Berry, Steven T., and Philip A. Haile. 2021. "Foundations of Demand Estimation." In *Handbook of Industrial Organization*, Vol. 4, edited by Kate Ho, Ali Hortaçsu, and Alessandro Lizzeri, 1–62. Amsterdam: Elsevier.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63 (4): 841–90.

Berry, Steven, James Levinsohn, and Ariel Pakes. 2004. "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market." *Journal of Political Economy* 112 (1): 68–105.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa. 2022. "Discretizing Unobserved Heterogeneity." *Econometrica* 90 (2): 625–43.

Christensen, Laurits R., Dale W. Jorgenson, and Lawrence J. Lau. 1975. "Transcendental Logarithmic Utility Functions." *American Economic Review* 65 (3): 367–83.

Compiani, Giovanni, Ilya Morozov, and Stephan Seiler. 2023. "Demand Estimation with Text and Image Data." Unpublished.

Conlon, Christopher, and Jeff Gortmaker. 2020. "Best Practices for Differentiated Products Demand Estimation with Pyblp." *RAND Journal of Economics* 51 (4): 1108–61.

Conlon, Christopher, and Jeff Gortmaker. 2023. "Incorporating Micro Data into Differentiated Products Demand Estimation with PyBLP." Unpublished.

Conlon, Christopher, Julie Holland Mortimer, and Paul Sarkis. 2022. "Estimating Preferences and Substitution Patterns from Second Choice Data Alone." Unpublished.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2020. "The Rise of Market Power and the Macroeconomic Implications." *Quarterly Journal of Economics* 135 (2): 561–644.

Deaton, Angus, and John Muellbauer. 1980. "An Almost Ideal Demand System." *American Economic Review* 70 (3): 312–26.

Fox, Jeremy T. 2007. "Semiparametric Estimation of Multinomial Discrete-Choice Models Using a Subset of Choices." *RAND Journal of Economics* 38 (4): 1002–19.

Gabel, Sebastian, and Artem Timoshenko. 2022. "Product Choice with Large Assortments: A Scalable Deep-Learning Model." *Management Science* 68 (3): 1808–27.

Gandhi, Amit, and Jean-François Houde. 2019. "Measuring Substitution Patterns in Differentiated-Products Industries." NBER Working Paper 26375.

Gandhi, Amit, and Aviv Nevo. 2021. "Empirical Models of Demand and Supply in Differentiated Products Industries." In *Handbook of Industrial Organization*, Vol. 4, edited by Kate Ho, Ali Hortaçsu, and Alessandro Lizzeri, 63–139. Amsterdam: Elsevier.

Goettler, Ronald L., and Ron Shachar. 2001. "Spatial Competition in the Network Television Industry." *RAND Journal of Economics* 32 (4): 624–56.

Grieco, Paul L. E., Charles Murry, and Ali Yurukoglu. 2021. "The Evolution of Market Power in the US Auto Industry." NBER Working Paper 29013.

Han, Sukjin, Eric H. Schulman, Kristen Grauman, and Santhosh Ramakrishnan. 2021. "Shapes as Product Differentiation: Neural Network Embedding in the Analysis of Markets for Fonts." Unpublished.

Härdle, Wolfgang K., and Léopold Simar. 2019. *Applied Multivariate Statistical Analysis*. Cham, Switzerland: Springer Nature.

Houde, Jean-François. 2012. "Spatial Differentiation and Vertical Mergers in Retail Markets for Gasoline." *American Economic Review* 102 (5): 2147–82.

Kumar, Madhav, Dean Eckles, and Sinan Aral. 2020. "Scalable Bundling via Dense Product Embeddings." arXiv: 2002.00100.

Lancaster, Kelvin J. 1966. "A New Approach to Consumer Theory." *Journal of Political Economy* 74 (2): 132–57.

Magnolfi, Lorenzo, Jonathon McClure, and Alan Sorensen. 2021. "Triplet Comparisons Data for Cereals." Unpublished.

Magnolfi, Lorenzo, Jonathon McClure, and Alan Sorensen. 2025. *Data and Code for "Triplet Embeddings for Demand Estimation."* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E199922V1.

McFadden, Daniel. 1974. "The Measurement of Urban Travel Demand." *Journal of Public Economics* 3 (4): 303–28.

McFadden, Daniel. 1978. "Modeling Choice of Residential Location." In *Spatial Interaction Theory and Planning Models*, edited by A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull, 75–96. Amsterdam: North Holland.

McFadden, Daniel, and Kenneth Train. 2000. "Mixed MNL Models for Discrete Response." *Journal of Applied Econometrics* 15 (5): 447–70.

Mosqueira-Rey, Eduardo, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes- Bascarán, and Ángel Fernández-Leal. 2023. "Human-in-the-Loop Machine Learning: A State of the Art." *Artificial Intelligence Review* 56: 3005–54.

Nevo, Aviv. 2001. "Measuring Market Power in the Ready-to-Eat Cereal Industry." *Econometrica* 69 (2): 307–42.

Nielsen IQ. 2024a. *Consumer Panel Data*. Chicago, IL: Nielsen IQ. https://www.chicagobooth.edu/research/kilts/research-data/nielseniq (accessed February 2024).

Nielsen IQ. 2024b. *Retail Scanner Data*. Chicago, IL: Nielsen IQ. https://www.chicagobooth.edu/research/kilts/research-data/nielseniq (accessed February 2024).

Nutritionix. 2021. *Cereal Nutritional Information*. Nutritionx. https://www.nutritionix.com/ (accessed October 2021).

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." In *Advances in Neural Information Processing Systems*, Vol. 35, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, 27730–44. Long Beach, CA: Neural Information Processing Systems Foundation, Inc.

Petrin, Amil. 2002. "Quantifying the Benefits of New Products: The Case of the Minivan." *Journal of Political Economy* 110 (4): 705–29.

Pinkse, Joris, and Margaret E. Slade. 2004. "Mergers, Brand Competition, and the Price of a Pint." *European Economic Review* 48 (3): 617–43.

Pinkse, Joris, Margaret E. Slade, and Craig Brett. 2002. "Spatial Price Competition: A Semiparametric Approach." *Econometrica* 70 (3): 1111–53.

Reynaert, Mathias, and Frank Verboven. 2014. "Improving the Performance of Random Coefficients Demand Models: The Role of Optimal Instruments." *Journal of Econometrics* 179 (1): 83–98.

Rossi, Peter E. 2014. "Even the Rich Can Make Themselves Poor: A Critical Examination of IV Methods in Marketing Applications." *Marketing Science* 33 (5): 655–72.

Ruiz, Francisco J. R., Susan Athey, and David M. Blei. 2020. "Shopper: A Probabilistic Model of Consumer Choice with Substitutes and Complements." *Annals of Applied Statistics* 14 (1): 1–27.

Syverson, Chad. 2019. "Macroeconomics and Market Power: Context, Implications, and Open Questions." *Journal of Economic Perspectives* 33 (3): 23–43.

Van der Maaten, Laurens, and Geoffrey Hinton. 2008. "Visualizing Data Using t-SNE." *Journal of Machine Learning Research* 9 (86): 2579–2605.

Van der Maaten, Laurens, and Kilian Weinberger. 2012. "Stochastic Triplet Embedding." In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, 1–6. Santander, Spain: IEEE.

Wilber, Michael J., Iljung S. Kwak, and Serge J. Belongie. 2014. "Cost-Effective Hits for Relative Similarity Comparisons." In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 2, edited by Jeffrey P. Bigham and David Parkes, 227–33. Pittsburgh, PA: Association for the Advancement of Artificial Intelligence.