# Fantastic Embeddings and How to Align Them: Zero-Shot Inference in a Multi-Shop Scenario

Federico Bianchi*
Bocconi University
Milano, Italy
f.bianchi@unibocconi.it

Jacopo Tagliabue*†
Coveo Labs
New York, NY
jtagliabue@coveo.com

Bingqing Yu*
Coveo
Montreal, Canada
cyu2@coveo.com

Luca Bigon‡
Coveo
Montreal, Canada
lbigon@coveo.com

Ciro Greco§
Coveo Labs
New York, NY
cgreco@coveo.com

## ABSTRACT

This paper addresses the challenge of leveraging multiple embedding spaces for multi-shop personalization, proving that zero-shot inference is possible by transferring shopping intent from one website to another without manual intervention. We detail a machine learning pipeline to train and optimize embeddings *within shops* first, and support the quantitative findings with additional qualitative insights. We then turn to the harder task of using learned embeddings *across shops*: if products from different shops live in the same vector space, user intent - as represented by regions in this space - can then be transferred in a zero-shot fashion across websites. We propose and benchmark unsupervised and supervised methods to "travel" between embedding spaces, each with its own assumptions on data quantity and quality. We show that zero-shot personalization is indeed possible at scale by testing the shared embedding space with two downstream tasks, event prediction and type-ahead suggestions. Finally, we curate a cross-shop anonymized embeddings dataset to foster an inclusive discussion of this important business scenario.

## CCS CONCEPTS

• **Information systems** → **Recommender systems**; *Query suggestion*; • **Theory of computation** → *Unsupervised learning and clustering*.

## KEYWORDS

neural networks, product embeddings, product recommendation, transfer learning, zero-shot learning

*Main authors, contributed equally to ideation and execution of this research project and are listed alphabetically.
†Corresponding author.
‡Author was responsible for data ingestion and data engineering.
§Author was responsible for analysis of the downstream NLP task.


*SIGIR eCom'20, July 30, 2020, Virtual Event, China*


**ACM Reference Format:**
Federico Bianchi, Jacopo Tagliabue, Bingqing Yu, Luca Bigon, and Ciro Greco. 2020. Fantastic Embeddings and How to Align Them: Zero-Shot Inference in a Multi-Shop Scenario. In *Proceedings of ACM SIGIR Workshop on eCommerce (SIGIR eCom'20).* ACM, New York, NY, USA, 11 pages.

## 1 INTRODUCTION

Inspired by the similarity between words in sentences and products in browsing sessions, recent work in recommender systems re-adapted the NLP CBOW model [20] to create *product embeddings* [17], i.e. low-dimensional representations which can be used alone or fed to downstream neural architectures for other machine learning tasks. Product embeddings have been mostly investigated as static entities so far, but, exactly as words [10], products are all but static. Since the creation of embeddings is a stochastic process, training embeddings for similar products in different digital shops will produce embedding spaces which are not immediately comparable: how can we build a unified cross-shop representation of products? In *this* work, we present an end-to-end machine learning pipeline to solve the transfer learning challenge in digital commerce, together with substantial evidence that the proposed methods – even with no supervision – solve effectively industry problems that are otherwise hard to tackle in a principled way (e.g. zero-shot inference in a multi-shop scenario).

We summarize the main contributions of *this* paper as follows:

- we extensively investigate product embeddings in both the *within-shop* and *cross-shop* scenarios. Since this is the *first* research work to tackle *cross-shop* inference by aligning embedding spaces, Section 2 explains the use cases at length. While *within-shop* training is not a novel topic *per se* (Section 3), we report detailed quantitative results since we could *not* replicate previous findings in hyperparameter tuning; we also improve upon existing pipelines by proposing a qualitative validation as well;
- we propose, implement and benchmark several aligning methods, varying the degree of supervision and data quality required. We provide quantitative and qualitative validation of the proposed methods for two downstream tasks: a "next event prediction" and a type-ahead personalization task, in


**Figure 1: Cross-shop use case: a user browsing "basketball-related" products on Shop A and then continuing the session on Shop B with similar products.**

which aligned embeddings are used as input to a conditional language model;
- curate and release in the public domain a cross-shop product embeddings dataset[1] to foster reproducible research on this topic. With practitioners in the industry in mind, we also detail our cloud architecture in Appendix A.

Our analysis of product data from several stores found that product embeddings, while superficially similar to word embeddings, have their own peculiarities, and data assumptions need to be assessed on a case-by-case basis. Moreover, our benchmarks confirm that the proposed methodology is of great interest when a single SaaS provider can leverage cross-client data, or when a multi-brand/multi-regional group can use data from one store to improve performance on another.

## 2 USE CASES FROM THE INDUSTRY

Shoppers are likely to browse in multiple related digital shops before making the final purchase decision, as most online shopping sessions (as high as 99% [13]) do not end with a transaction. The cross-shop scenario depicted in Fig. 1 is therefore very common: the shopper starts browsing on **Shop A** for basketball products and ends up continuing his session on **Shop B**.

Providing relevant content to unknown shoppers is of paramount importance to increase the probability of a conversion, considering that e-commerce websites tend to have high bounce rates (i.e. average percentage of users who leave after a single interaction with the page ranges between 25% and 40% [27]) and low ratios of recurring customers (<9% in our dataset). Moreover, there is vast consensus in the industry on the importance of personalization [26] in boosting the quality of the shopping experience and increasing revenues: but how is it possible to personalize the experience of a user that has never been on the target site?

The rationale for *this* research work is thus the importance of providing personalized experiences *as early as possible* and with *as little user data as possible*: generally speaking, we propose to leverage the aligned product embedding space to model shopper's intent during a session – if cross-shop browsing is, so to speak, a walk through the (aligned) product space, we can feed users's position to downstream neural systems to capture their shopping intent.

[1] At the time of drafting *this* paper, discussions within the legal team of *Coveo* are still ongoing to settle on a final license for the data; as such, dataset details may change before final publication: feel free to reach out to us for any update.

**Table 1: A sample of multi-brand retailers from Fortune 500.**

| Group | Rev. (M$) | Brands | Examples |
|---|---|---|---|
| TJX | 41.717 | 7 | HomeSense, Marshalls |
| Nike | 39.117 | 4 | Converse, Nike |
| Gap | 16.383 | 9 | Gap, Old Navy |
| VF | 13.870 | 19 | Eastpack, Napapijri |
| L Brands | 12.914 | 3 | Victoria Secret, Pink |
| Hanesbrands | 6.966 | 29 | Champion, Playtex |

There are two types of players which would naturally benefit from cross-shop personalization. The first is retail groups who own and operate multiple brands and shops (e.g. Gap Inc owns and operates Gap, Old Navy, etc). To give an idea of the size of this market share, the combined revenues generated by *Fortune 500* retail groups with these characteristics is more than 130 billion dollars (see Table 1). For these retailers, a portion of the user base consistently shops across different websites of the same group and it would be therefore beneficial to them to implement optimization strategies across multiple websites. Given the size of the market, it is easy to see how the implementation of successful personalization strategies across shops would translate into remarkable business value. At the same time, most of these groups are "traditional" retailers (as opposed to digitally native companies e.g. Amazon). Therefore, even if they would be benefiting the most from a unified view of their customers across different digital properties, in practice they are more likely to experience roadblocks related to technology. To this extent, the immediate value of the present work is to show for the first time that personalization across shops can be achieved even with minimal data tracking, no meta-data and no human intervention. The traditional nature of these retailers may also explain why cross-shop behavior is a niche use case in the research community, whose agenda is mostly set by tech companies – by publishing our findings we wish the community would join us in tackling this important use case.

The second type of players are multi-tenant SaaS providers who provide AI-based services. For these companies the main challenge is to scale quickly within the verticals and minimize the friction in deployment cycles: being able to leverage some kind of "network effect" to transfer knowledge from one client to another would certainly be a distinctive competitive advantage. Recently, AI SaaS providers for e-commerce have received great attention from venture capitalists. As an indication of the size of the market opportunity, only in 2019 and only in the space of AI-powered search and recommendations, we witnessed *Algolia* raising USD110M [32], *Lucidworks* raising USD100M [34] and *Coveo* raising CAD227M [33]. While a full cross-shop data strategy depends on many non-technical assumptions (see Section 4 for a discussion of legal constraints), it is important to realize that some multi-property retail groups turn to external providers for certain AI services. While our methods do not assume any common meta-data between target shops (e.g. the two shops can be even in *different* language), we expect our models to work better with catalogs that have significant "semantic overlap" (e.g. two shops selling sport apparel, Section 4).

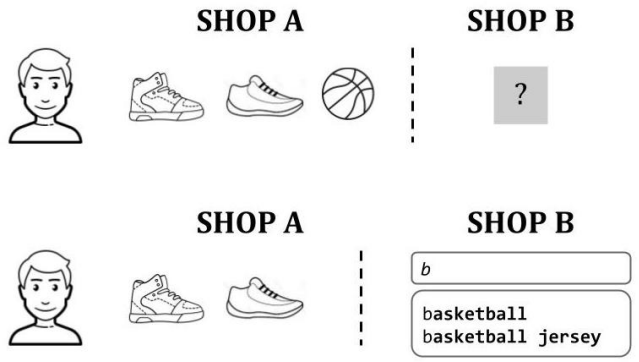


**Figure 2: Zero-shot prediction tasks can be solved by transferring shopper's intent to the target website leveraging aligned embeddings; in the first task, the system predicts product interactions on Shop B from user browsing products on Shop A; in the second task, products on Shop A are used to personalize type-ahead suggestion on Shop B – since the session is basketball-themed, we expect the system to promote (*ceteris paribus*) basketball queries.**

We show several effective methods to achieve transfer learning across shops, each making different assumptions about data quantity and quality available. As discussed at length in Section 5, a distinguishing feature of this use case is that we make no assumption *at all* about catalog overlap (i.e. the shops involved can have 0 items in common), making it much more challenging that the typical (and well-studied) retargeting use case (i.e. a shopper sees ads on *Site X* for the same product she was viewing on *Site Y*). Our most interesting result is proving that even without any cross-shop data, personalization on the target shop can be achieved successfully in a pure zero-shot fashion. To showcase the possibilities opened up by cross-shop embeddings, we demonstrate the effectiveness of the aligned space tackling two prediction tasks, as depicted in Figure 2: if the system observes the behavior of a user on **Shop A**, can it predict what she is going to browse/type on **Shop B**? As we shall see, the answer is "yes" for both use cases.

## 3 RELATED WORK

The work sits at the intersection of several research topics.

**Product Embeddings**. Word2vec [22] was introduced in 2013 as a neural method to generate vector representations of words based on co-occurrences; soon, the model was adapted to the product space, where it found immediate use in recommender systems [12]. [35] introduced *Meta-Prod2Vec*: given our focus on cross-shop learning, we decided to *not* use product information as there is no guarantee that two shops will have comparable metadata. [7] first studied the role of hyperparameters in recommendation quality: we extensively investigate hyperparameters as well, but we improve upon their validation procedure and add a qualitative evaluations of the embedding spaces.

**Aligning Embedding Spaces**. The problem of learning a mapping between spaces has been widely explored in NLP. In fact, Alignment is important for language translation [3, 18], to study language change [4, 10, 14, 30, 38]. However, as explained in Section 5, the availability of unequivocal pairs of matching items in two spaces (e.g. *uno* and *one* in language translation) make vector space alignment in NLP significantly different from our use case. [5] is a recent work on zero and few shots prediction in a recommender setting across multiple "spaces": their problem is phrased as a meta-learning task over graphs representing different cities, while our work is focused on behavioral-based embeddings and sessions across multiple spaces. Possibly because of the maturity of data ingestion required to rebuild session data and the difficulty in finding suitable datasets for experimentation, *this* work is the first to our knowledge to extensively study product embeddings across multiple spaces.

**Deep Learning in Type-ahead Systems**. Suggest-as-you-type is a well studied problem in the IR community [6]. Recent works have embraced neural networks: [24] introduces a char-based language model, [37] applies RNN to a noisy channel model (but the inner language model is not personalized like our proposed method). Specifically in e-commerce, [15] uses *fastText* to embed previous queries and then re-ranks suggestions accordingly: our personalization layer does not require linguistic resources or previous queries, as the vast majority of sessions (>90% in our network) for mid-size shops do *not* contain search queries. [39] is the first exploration of cross-shop type-ahead systems, obtaining transfer learning by placing products in the same space through shared image features. The proposed *prod2vec* embeddings significantly outperform image-based representations to produce accurate conditional language models (18% MRR improvement over the same shop).

**Table 2: Descriptive stats for Shop A and Shop B**

| Shop | Sessions (events) | SKUs | 25/50/75 pct |
|---|---|---|---|
| A | 3M (10M) | 23k | 3, 5, 7 |
| B | 11M (32M) | 42k | 3, 5, 7 |

## 4 DATASET

*Coveo* is a Canadian SaaS provider of search and recommendation APIs with a global network of more than 500 customers, including several *Fortune 500* companies. For *this* research, we leverage behavioral data collected over 12 months from two mid-size shops (revenues >10M and <100M) in the same vertical (sport apparel); we refer to them as **Shop A** and **Shop B**. Data is sessionized by the pipeline after ingestion: *prod2vec* embeddings are trained on product interactions that occur within each recorded shopper session (Section 5.1). In the interest of practitioners in the industry, we share details on our cloud design choices in Appendix A.

Catalogs from **A** and **B** were also obtained to perform a qualitative check on our validation strategy and test semi-supervised approaches. After cleaning user sessions from bot-like behavior and sampling, descriptive statistics for the final product embedding dataset can be found in Table 2; even if **A** and **B** differ in catalog size and traffic, they have <9% of *recurring* customers (i.e. shoppers with more than 3 sessions in 12 months).

We believe it is important to explicitly address two potential legal concerns about the underlying dataset of *this* research:

**Table 3: Hyperparameters and their ranges.**

| Gensim Parameter | Tested Values |
|---|---|
| *min_count* | 2, 3, 5, 10, 15, 30 |
| *window* | 2, 3, 5, 10, 15 |
| *iter* | 5, 10, 20, 30, 50 |
| *ns_exponent* | -1.0, -0.5, 0.0, 0.75, 1.0 |

- end-user *privacy*: data collected is fully anonymized, in line with GDPR adequacy; data tracking required to produce aligned embeddings is *significantly less* than other standard e-commerce use cases (e.g. re-targeting);
- data *ownership*: the possibility to use aggregate (embeddings-based) data across websites depends on case-by-case legal constraints and specific contractual clauses. Websites operated by the same group have generally no issue in sharing data to improve overall performance. On the other hand, websites operate by different companies may see each other as competitors. In our experience, the answer is not clear-cut: mid-size shops (like **A** and **B**) tend to be less protective and more focused on the upside of a system that is aware of industry trends; bigger players, on the other side, seem to be more defensive; interestingly, the latter are more likely to have multi-brand deployment, making the methods here developed still relevant for many use cases.

Finally, a sample of browsing sessions for distinct users with cross-shop behavior was obtained to benchmark different methods on the downstream prediction tasks: it is worth remembering that several proposed methods for cross-shop inference (Section 5) do *not* rely on cross-shop data, which is used in the unsupervised and semi-supervised case as gold standard only.

## 5 METHODS

The cross-shop inference is built in two phases. First, the system learns the best embeddings for **A** and **B** *separately*, second, it learns a mapping function from one space to the other, implicitly aligning the two embedding spaces and enabling cross-shop predictions.

### 5.1 Learning optimal product embeddings

Product embeddings are trained using CBOW with negative sampling [20, 22], by swapping the concept of words in a sentence with products in a browsing session; for completeness we report a standard formulation [23]. For each product $p \in \mathcal{P}$, its center-product embedding and context-product embedding are $d$-dimensional vectors in $\mathbb{R}$, $\mathcal{U}[\mathrm{p}]$ and $\mathcal{V}[\mathrm{p}]$: embeddings are learned by solving the following optimization problem:

$$\max_{\substack{\mathcal{U}:\mathcal{P}\to\mathbb{R}^d \\ \mathcal{V}:\mathcal{P}\to\mathbb{R}^d}} \sum_{(p,c)\in\mathcal{D}^+} \log\sigma\left(\mathcal{U}[p]^\top\mathcal{V}[c]\right) + \sum_{(p,c)\in\mathcal{D}^-} \log\sigma\left(-\mathcal{U}[p]^\top\mathcal{V}[c]\right) \quad (1)$$

where $D^+/D^-$ are positive/negative pairs in $D$, and $\sigma()$ is the standard sigmoid function. Following the findings in [7], we performed extensive tuning on the most important hyperparameters (Table 3) and develop both quantitative and qualitative protocols to evaluate the quality of the produced embedding space.

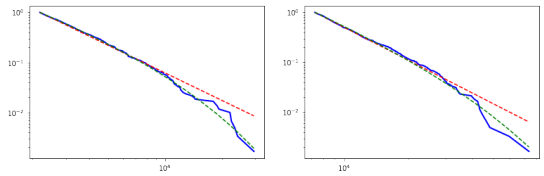

**Figure 3: Shop A (left) and Shop B (right) log plots for product views: *empirical* distribution is in blue, *power-law* in red and *truncated power-law* in green. Truncated power-law is a better fit than standard power-law for both shops ($p < .05$), with $\alpha = 2.32$ for A and $\alpha = 2.72$ for B. Power-law analysis and plots are made with [1].**

*5.1.1 Quantitative validation.* We focused on a *Next Event Prediction* (NEP) task to evaluate quantitatively the quality of the embeddings: given a session $s$ made by events $e_1$, ... $e_n$, how well $e_1$, ... $e_{n-1}$ can predict $e_n$?

To address the NEP, we propose to use the entire session preceding the target event, by constructing a session vector averaging the embeddings for $e_1$, ... $e_{n-1}$ and then apply a *Nearest Neighbors* classifier to predict $e_n$. Our choice is in contrast with what proposed by [7], which conducts hyperparameter tuning using kNN with just one item, $e_{n-1}$, as seed: from our experience in digital commerce, buying preferences are indeed multi-faceted, and important information about user intentions may be hidden at the start of the session ([8, 39])[2]. Both **H@10** and **NDCG@10** were calculated for each trained model, but **NDCG@10** was primarily used for evaluation:

$$Discounted\ CG_k = DCG_k = \sum_{i=1}^{k} \frac{rating(i)}{log_2(i+1)} \quad (2)$$

$$Ideal\ DCG_k = IDCG_k = \sum_{i=1}^{|REL|} \frac{rating(i)}{log_2(i+1)} \quad (3)$$

$$NDCG_k = \frac{DCG_k}{IDCG_k} \quad (4)$$

where $|REL|$ is the list of ground truth target events, up to $k$, and $rating(i)$ is the binary relevance value, which means $rating(i) = 1$ if event $i$ is found in the ground truth target events; otherwise, $rating(i) = 0$. Best and worst models, with parameters and score, can be found in Table 4. It is interesting to remark that our extensive validation could *not* confirm many generalizations put forward in [7]: negative exponent was *not* found to be a consistent factor in improving embeddings quality and **Shop A** and **Shop B** best parameter combinations are very similar, despite the underlying distribution being different (Figure 3); moreover, the gap between best and worst models was found to be significant, but not as wide as [7] indicated.

[2]We also used LSTM as an alternative algorithm for validation, with similar results. We opted to report only kNN since a simpler model allows our valuation to be focused on the quality of the embeddings themselves, not so much the algorithm.

**Table 4: Best and worst parameter settings by shop, with validation score.**

| Model | Min Count | Window | Iter. | Exp. | NDCG@10 |
|---|---|---|---|---|---|
| **A - Best** | 15 | 10 | 30 | 0.75 | 0.1490 |
| **A - Worst** | 2 | 15 | 10 | -0.5 | 0.1058 |
| **B - Best** | 15 | 5 | 30 | 0.75 | 0.2452 |
| **B - Worst** | 5 | 10 | 30 | -0.5 | 0.1881 |

*5.1.2 Qualitative validation.* The evaluation of word embedding models is intrinsically built on human-curated analogies such as *boy* : *king* = *women* : *?* [25] as both a quantitative check ("how many analogies can be solved by the vector algebra in the given space?") and a qualitative one ("can we confirm, as humans, that the semantic properties captured by the space are indeed close to our linguistic intuitions?"). While analogies are indeed potentially meaningful in the product spaces for specific use cases (e.g. what is the Nike's "air jordan shoes" equivalent for Adidas?), compiling a list for validation would be time-consuming and involving arbitrary choices.

To have an independent qualitative confirmation that the NEP task is enforcing meaningful distinctions between spaces trained with different parameters, we sampled a model from the top 5 and one from bottom 5 in the NEP ranking, and leverage domain experts to classify products into sport activities (soccer, basketball, tennis, etc., for a total of $N$=10 activities). We use t-sne [19] to project embeddings into two-dimensions and color-code the products with labels: as shown in Figure 4, better embeddings form sharper clusters with homogeneous coloring. To confirm the visual results, we train a Multilayer Perceptron (MLP) with the objective of predicting the activity from the embeddings[3]. Confirming the visual inspection, the accuracy score was 0.95 for the high-performing model and 0.32 for the low-performing one.

## 5.2 Crossing the (shop) chasm

Cross-embedding learning in the NLP space takes place in a continuum of supervision: from thousands of "true" pairs [21], to dozen of them [2], to no pair at all [18]. However, it should be emphasized that aligning word spaces and aligning product spaces are *not* the same task:

(1) given two languages, both will contain the same "semantic regions" (e.g., general topics like *places*, *animals*, *numerals*, etc.) and, within those regions, several overlapping tokens (e.g. *dog* is *cane* in Italian, *one* is *uno*, *lake* is *lago*, etc.); however, given even shops in the same vertical such as **Shop A** and **Shop B**, there is no guarantee they will both contain products for, say, *climbing*;
(2) given two languages, there are linguistic resources mapping items from one to the other non-arbitrarily; however, given shops in the same verticals, finding exact duplicates is non-trivial and there are many cases in which mapping is arguably undetermined.

[3]The MLP has two dense layers with *relu* activation, a softmax layer for prediction, *dropout* of 0.5 between layers, *SGD* as optimizer.

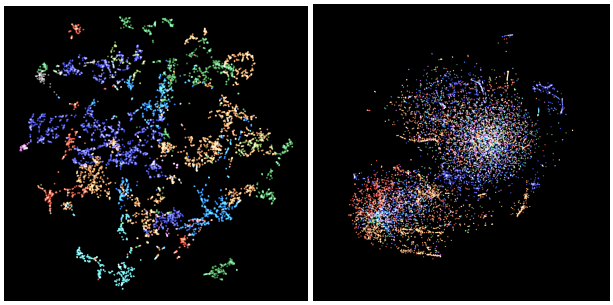

**Figure 4: 2-dimensional projections (t-sne) of high-scoring (left) and low-scoring (right) models according to the NEP task. Each point is a product in Shop A embedding space, color-coded by sport activity through catalog meta-data: it is easy to notice that high-scoring models produce sharper clusters in the embedding space. Projections are obtained with following parameters: *perplexity=25, learning rate=10, iterations=500*.**

It is also important to stress that *no product is assumed to be the same across the two shops*: while we know **Shop A** and **Shop B** have comparable catalogs in terms of *type* of items (e.g. they both sell sneakers, boots, etc.), we make no assumption about them having the same *tokens* (i.e. we don't know if they both sell a specific pair of shoes, *Air Zoom 95*), and we make no use of textual meta-data.[4].

Considering those differences, we built and tested a wide range of unsupervised and supervised models to address the *cross-shop* challenge:

- *image-based model* (IM), a completely unsupervised model using weak similarity signals derived from image vectors to build a "noisy" seed for a self-learning framework [3]. In particular, we sample images from **Shop A** and **Shop B** full catalogs and run through a pre-trained VGG-16 network [28] to extract features from the *fc2* layer; PCA is then applied to reduce the feature dimensions from 4096 to $d$ dimensions; K-means is then used to group the vectors for **Shop A** into $k$ clusters: 2 points closest to the centroids of each cluster are the "sample points"; for each of these points, we use kNN to retrieve the closest image from **Shop B**. The seed dictionary built in this fashion is used to bootstrap the self-learning framework, and iteratively improve the mapping and the dictionary until convergence. It is worth noting that the alignment results reported below are achieved even if the seed dictionary is indeed noisy (as verified manually by sampling the quality of the pairings), witnessing the robustness of the proposed procedure. Different values for $d$ (5, 10, 20, 40, 60, 100) and $k$ (15, 30, 50, 70) were tested, but we report the scores for the best combination ($d$ = 20 and $k$ = 50). This method is both completely unsupervised *and* fully "**zero**-shot" in the *cross-shop* scenario, as no data on *cross-shop* sessions is ever showed to the model during training;

[4]Assuming user and/or attribute overlap is the typical setting for cross-domain recommender systems [11]: for this reason, they are not a meaningful baseline for the scope of *this* work.

- *user-based model* (UM), a fully supervised model leveraging *directly* users browsing on the two target shops. In particular, given the last product seen on the source shop and the first on the target shop, we learn to map the two products using linear regression that is then generalized to map all the embeddings of the source shop to the target shop.
- *user-translation model* (TM), a fully supervised model that is using *directly* shoppers browsing on the two target shops as if it was a bi-lingual parallel corpus. In particular, task is modelled after a different NLP architecture, sequence-to-sequence networks for machine translation [29]: the intuition is that **Shop A** and **Shop B** behave quite literally as different languages and deep neural nets are well suited to learn how to encode in one space and decode in the other the latent intent of the shopper. We use the sequence to sequence model provided by the OpenNMT tool [16] that comes with 2-layer LSTM with 500 hidden units on both the encoder and the decoder layer. We initialize the embeddings of the layers using our product embeddings. The model is trained to translate the sequence of products seen in **Shop A** into a sequence of product seen in **Shop B**.

By proposing and testing methods with different degrees of supervision, we provide substantial evidence that aligning embedding is possible in a variety of business scenarios: in particular, insofar as data tracking and technological capabilities vary across retailers, purely unsupervised methods (*IM*) are particularly interesting as they make almost no assumption about available data and existing *cross-shop* data points. On the other hand, supervised models (*TM*) provide "natural" upper bounds for unsupervised counterparts, and can be deployed in business contexts where advanced data ingestion and data practices are already present. In general, our own experience is that these models can satisfy complementary business scenarios: for example, if historical fine-grained data is unavailable at day one (as it often is), aligning product embeddings with no cross-shop data is crucial to deliver personalization without advanced tracking capabilities.

# 6 EXPERIMENTS

We apply alignment methods to two downstream tasks: the first one is a straightforward extension to two shops of *NEP*, as presented in Section 5.1 – by aligning different product spaces, we hope to prove we can reliably guess shopper interactions with products on the target shop by transferring her intent from the first shop; the second task is an NLP-related task, in which aligned embeddings are used to build a conditional language model that can provide personalized suggestions to shoppers arriving at the target site [31]: the query suggestion task is useful both to establish that *prod2vec* transfer learning is superior to the image-based one [39], and to prove that intent vectors are not just useful for recommendations, but also for a variety of personalization tasks in NLP.

It is important to highlight that our focus is to establish for the *first* time that aligning product embeddings allow to transfer shopper intent between shops in scalable and effective ways; for this reason, we picked architectures which are straightforward to understand, in order to make sure the variation in the results are due to the quality of the learned embeddings and not to the implementation of the downstream task models – while more sophisticated options are detailed in Section 7, our benchmarks show that aligned embeddings are indeed an extremely promising area of exploration.

Finally, it is important to stress that given the novelty of the setting (as discussed in Section 2) and the differences with cross-space tasks in NLP settings, *prima facie* plausible baselines are actually not good candidates for the scenarios at hand. For example, even in the presence of high-quality cross-shop tracking, joint embeddings cannot be trained on cross-shop sessions due to data sparsity; as another example, recent alignment techniques that are successful for word spaces (e.g. [10]) rely on the assumption that either many labeled pairs are available, or that the vast majority of the embedding space is comprised by pairs of identical items; other interesting ideas, such as using product titles for a similarity metrics, would require uniformity in meta-data, which is an assumption that no proposed models make. Framed as a zero-shot inference, *multi-shop* predictions are a relatively new challenge and we hope our work (and dataset) to be a long-lasting contribution to the community.

## 6.1 Next Event Prediction across shops

For the cross-shop prediction task, we sampled 12510 browsing sessions over a month (*not included in the training set*) for distinct users that visited **Shop A** and **Shop B** within the same day.

*6.1.1 Quantitative evaluation.* We benchmark the cross-shop methods from Section 5.2 against three baselines of increasing sophistication:

- *popularity model* (PM): while trivial to implement, leveraging product popularity is by far the most common heuristic in the industry for the zero-shot scenario, and it has been proven to be surprisingly competitive in many e-commerce settings against statistical and neural approaches [9]; also, given that popular products are more likely to be on display and generate a classic "rich get richer dynamics", quantitative results for *PM* are likely to overestimate its efficacy and therefore raising the bar for other methods;
- *activity-based model* (AM): a *semi-supervised* model, inspired by evidence from NLP literature in which *some* supervision goes a long way in helping with the alignment process [2]; in particular, the model leverages domain knowledge (sport activity for each product) that is however not directly related to the mapping we are trying to learn. We randomly sample 20 products from **Shop A** of category $S$ and from **Shop B** within the same category, using activities as "known similar regions", and we then we learn a mapping function using standard linear regression from the centroid of the sampled products from the two spaces;
- *iterative alignment model* (NM): state-of-the-art unsupervised method from [3], originated in the NLP literature: the model is quite sophisticated and its performances in this scenario shed interesting insights on how peculiar the task of aligning product embeddings is (as compared to word embeddings); in a nutshell, *NM* leverages the structure of embedding spaces to build an initial weak dictionary; the dictionary is then used to bootstrap a self-learning process,

**Table 5: NDCG@10 for supervised and unsupervised models in the *First Item Prediction* (FIP) and *Any Item Prediction* (AIP) tasks: best results per type are highlighted in bold.**

| Model | Type | FIP | AIP |
|---|---|---|---|
| PM | Unsupervised | 0.00232 | 0.00297 |
| NM | Unsupervised | 0.00097 | 0.00112 |
| IM | Unsupervised | **0.01506** | **0.01628** |
| AM | Semi-supervised | 0.00108 | 0.00121 |
| UM | Supervised | 0.02741 | 0.02854 |
| TM | Supervised | **0.03786** | **0.04501** |

which iterates through mapping and dictionary optimization, until convergence is reached.

Table 5 reports **NDCG@10** for all models for two prediction tasks: *First Item Prediction* (FIP) and *Any Item Prediction* (AIP). *FIP* is the ability of the proposed model to guess the first product in the target shop, while *AIP* is the ability to guess *any* product found in the session in the target shop. Unsurprisingly, fully supervised models outperform all other methods; among unsupervised models, the *IM* model we propose is the best one, resulting in a 549% increase over the industry baseline and even significantly *beating the semi-supervised baseline AM*[5]; the performance gap between *IM* and *NM* highlights that straightforward implementation of SOTA models from NLP does not guarantee the same results in the product scenario. Among supervised models, *TM* outperforms *UM* on *FIP* and provides a 1530% increase over the industry baseline; to test if *TM* improves significantly with data quantity, we ran an additional test on a separate cross-shop dataset from our network of clients: *TM* results on this second set for *FIP*/*AIP* are 0.066/0.071, and 0.021/0.023 for *UM*, showing that indeed the seq2seq architecture may be the best option for use cases in which significant amount of cross-shop behavior has been tracked already.

In the spirit of ablation studies, we generated predictions on the same *cross-shop* dataset using *IM* but employing instead *low-scoring* embedding spaces, to assess whether picking optimized vs non-optimized spaces make a difference in the zero-shot prediction task: the reported **NDCG@10** for this setting is 0.005, which is *significantly lower* than the reported best score obtained with the optimized embeddings.

*6.1.2 Qualitative evaluation.* Given the novelty of the experimental settings, a qualitative evaluation is important as well to interpret the outcome of the benchmarks above: is the alignment of the two spaces capturing important human-level concepts? We devised two additional tests to answer these questions. First, we test the aligned embeddings in a "cross-shop activity prediction" task: using the same setup from Section 5.1.2, we train an MLP on **Shop A** *aligned* embeddings and use it without additional training on **Shop B** *aligned* embeddings. The mean accuracy for activity prediction over 5 runs is $\mu = 0.73$ ($SD = 0.002$), confirming that the alignment process can effectively transfer learning from **A** to **B**.

[5]Generally speaking, *AM* seems to overfit on common categories and turns out to be worse than the simple *PM* model.

Second, we perform *error analysis* on several misclassified cases. Our exploration highlights that pure quantitative measures - such as **NDCG@10** - are great at capturing high-level patterns of efficacy for the chosen models, but cannot capture important differences in particular cases of *cross-shop* predictions. If we think about the particular task of zero-shot recommendation, **NDCG@K** is asking the model to pick *the* one correct product out of *several thousands*, which is likely to underestimate the practical efficacy of the proposed recommendations. Instead of just computing an hit/miss ratio for **NDCG@K**, we ran the *IM* model on the test set recording, for every "miss", the distance in the shared embedding space between the target product and the predicted one; we then order these wrong predictions according to the magnitude of the error, and analyze sessions from the top and bottom of the distribution. Interestingly enough, sessions with a small recorded error are the ones that looks coherent to a human observer, as in **Session A** in Figure 6, where running shoes from *Brooks* manufacturer are confused by the model with running shoes from *Mizuno* manufacturer; when error margin gets big, situations like **Session B** are more common: products in the same *cross-shop* session are very different, since the shopper intent may have drifted between the two visits - the prediction of the model is significantly off (wrong object, wrong manufacturer, wrong sport activity). To try and quantify the proportion of "reasonable" mistakes, we train an MLP mapping the target and the predicted product to a sport activity (as in Section 5.1.2), and comparing the first predicted activity versus the ground truth: this model achieves *zero-shot* accuracy of 0.44, which raises to 0.66 if we consider just sessions whose error distance is below the median (i.e. sessions with more "stable" intent).

All combined, these findings suggest that models are successfully transferring shopping intent and they are likely to perform well in practice for all the sessions in which intent across shop is consistent, even when the predicted item is not *exactly* a match (e.g. **Session A** in Figure 6; cases like **Session B** are unlikely to be solvable anyway).

## 6.2 Personalized Type-Ahead across shops

As a second, less direct application of aligned embedding spaces, we propose to exploit product embeddings in a conditional language model, to provide personalized type-ahead suggestion to incoming users on a target shop (Fig. 2). We deploy the same type-ahead framework we proposed in [39], in which an encoder-decoder architecture is employed to first encode user intent, and then use an LSTM-powered char-based language model to sort query completions by their probability (please refer to the paper for architectural details): as illustrated by Fig. 7, if the user's session is basketball-themed (**1**), we expect completions like *basketball jersey* for prefix *b*; if it is tennis-themed (**2**), the same prefix may instead trigger a tennis brand like *babolat*.

*6.2.1 Quantitative evaluation.* Table 6 shows the results of our quantitative benchmarks for the *cross-shop* scenario, comparing a non-personalized baseline to models performing transfer learning. For the personalized predictions, we train a conditional language models on the target shop first. At prediction time, we feed to the target shop model the *aligned* embeddings from the source

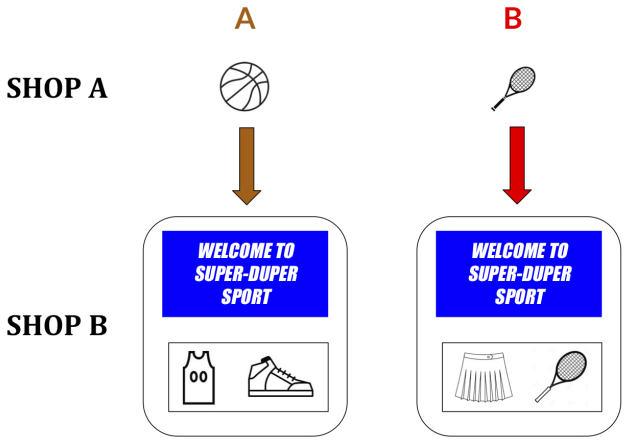


**Figure 5: Landing pages can be customized in real-time by transferring intent from previous shops to the current one: by focusing on the general activity, instead than the exact product, we make the task easier for the model and unlock more use cases for the clients. In this example, Shop B presents a basketball-themed page to User A and a tennis-themed page to User B.**


**Figure 6: Two sample sessions from the *cross-shop* portion of the dataset: Session A is a session with stable shopping intent (i.e. "running") and model prediction is wrong but plausible; Session B is made of two disconnected intents and model prediction is significantly wrong.**

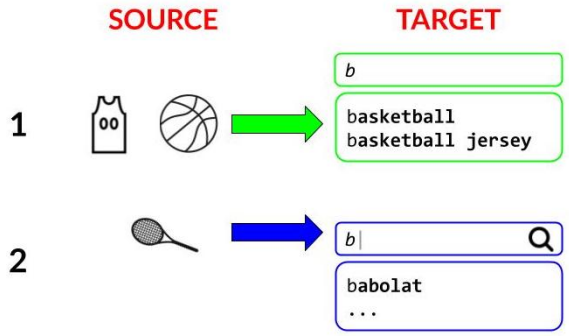


**Figure 7: Two sessions illustrating cross-shop personalization for type-ahead suggestions: the same prefix "b" on the target website triggers different completion depending on intent transferred from the source shop.**

**Table 6: MRR@5 in the *cross-shop* scenario, for different seed length (SL), for shoppers going from A to B and issuing a query there.**

| Model | SL=0 | SL=1 |
|---|---|---|
| PM | 0.001 | 0.045 |
| Vec2Seq+IM | 0.005 | 0.050 |
| Vec2Seq+UM | 0.003 | 0.055 |
| Vec2Seq+TM | **0.007** | **0.062** |

shop, perform average pooling in the encoder [39], and read off the decoder conditional probabilities of the target query suggestions.

We use *Mean reciprocal rank* (**MRR**) as our main metric, as a standard in the auto-completion literature: **MRR@k** is **MRR** measured by retrieving from the model the first $k$ suggestions. In our experiments, $k$ is set to 5 to mimic the target production environment:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i} \tag{5}$$

where $\text{rank}_i$ is the position of the first relevant result in the $i$-th query and $Q$ is the total number of queries.

The best supervised models provide up to 600% uplift, but even the *purely unsupervised* model significantly outperforms the non-personalized model, establishing that transferring intent is *significantly* better than treating all incoming shoppers as *new*; for mid-size and large retailers, capturing the interest of even a small percentage of these users may provide significant business benefits.

*6.2.2 Qualitative evaluation.* Quantitative benchmarks provide empirical evidence on the overall efficiency of personalization, but as discussed, *cross-shop* sessions "in the wild" sometime show drifting intent across sites. To specifically test how much the transferred intent is able to capture *semantic similarity* across the two aligned spaces, we devise a small user study. We recruited 20 native speakers, whose age ranged between 22 and 45; subjects (Figure 8) were presented with a product image from S-Shop (1), a seed character (2) and were asked to pick the most relevant completion among 5 candidates (3). The <*product image*, *seed*> pairs are taken from representative queries from the *cross-shop* set, for a total of 30 stimuli for each subject; five candidate queries are chosen by first retrieving the top 35 candidates from the unconditioned model, and then sampling without replacement. By collecting semantic judgment directly, our prediction is that the performance gain from personalization will be higher, since the study should eliminate the popularity bias implicit in search logs.

*PM*, *IM* and *TM* are tested against the collected dataset, resulting in a **MRR@5** of, respectively, 0.076, 0.123 and 0.138; *TM* accuracy with $SL = 1$ is 81% higher than *PM*, supporting our hypothesis that the aligned embeddings successfully transfer user intent in the zero-shot scenario.

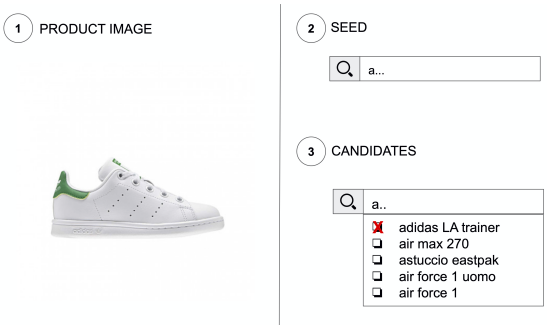


**Figure 8: Example of a stimulus in the qualitative user study.**

## 7 (VECTOR) SPACE, THE FINAL FRONTIER: WHAT'S NEXT?

In *this* work we detailed a machine learning pipeline for behavioral data ingestion finalized to train *prod2vec* models, i.e. generate neural product representations for several downstream prediction tasks. In the first part, we focused on training the best embeddings as judged by quantitative and qualitative validation. Product representations have been found to be increasingly useful in many e-commerce scenarios [31], but the understanding of them in realistic industry scenarios is still incomplete; on this point, it is telling that several findings of a recent hyperparameter study ([7]) could *not* be replicated in our context. For this reason, we believe that the *within-shop* training portion of our pipeline can provide a useful assessment for production systems in the industry, starting from our validation best practices and engineering considerations. Prompted by the industry need for few-shots and scalable personalization and practical deployment concern of our growing client network, the second part of *this* work was focused on generalizing product spaces to address the *cross-shop* scenario depicted in Fig. 2. We devised and tested several models with varying degrees of supervisions, and, again, supplemented our quantitative benchmarks with additional qualitative tasks to gain a better understanding of model performances in this new scenario. All in all, the evidence provided is a strong argument in favor of our initial research hypothesis, i.e. that embedding spaces from two shops can be successfully aligned, so that zero-shot predictions can be performed in a principled way.

While the theoretical and engineering foundations of the platform have proven to be solid and crucial in solving retail problems at scale, our roadmap is focused on taking these ideas even further. Broadly speaking, we can classify open issues in two categories, *research* and *product* improvements:

- **research**: since i) there is independent demand for general purpose *prod2vec* models, ii) universal tracking is still available in a limited fashion, we did not test end-to-end learning by using *cross-shop* predictions as the optimization task *directly*; as more data becomes available, it is a natural extension to the methods proposed in *this* work. Moreover, as highlighted in Section 6, significant optimization can be made to neural architectures for downstream tasks now that *this* study first established the viability of aligned embeddings to capture user's intent across shops;
- **product**: as discussed in Section 2, online retailers are facing increasing pressure to deliver relevant experiences to incoming customers; the question is not *whether* personalization should be done, but *how soon* into the shopper journey it can be done. We are actively working with several fashion groups to deploy cross-shop models and perform live A/B testing of the proposed methods; in our growing SaaS network of retailers, we believe more and more global multi-shop opportunities will soon benefit at scale from our research.

On a final note, we hope that curating the first dataset of its kind will help drawing increasing attention from industry and academic practitioners to these important business scenarios. SaaS providers with an extensive network of clients are ideally suited to leverage transfer learning techniques, including the alignment of embeddings here introduced; at the same time, some of the biggest traditional retailers in the world are indeed *multi-brand* groups, and they could "transfer knowledge" between their brands to provide personalization in an hyper-competitive, data-driven market.

In a time characterized by growing concerns on long-term storage of personal data [36], we *do* believe that small-data learning will be a distinctive feature for successful players in this space.

## ACKNOWLEDGMENTS

Thanks to the anonymous reviewers and Piero Molino for comments on previous versions of *this* work. Thanks to Caterina *Caterí* Vernieri, who brought LaTeX magic and stellar T-factor into our paper and our lives (not in this order). Special thanks to Andrea Polonioli for his support.

## REFERENCES

[1] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. 2014. powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions. *PLoS ONE* 9, 1 (Jan 2014), e85777. https://doi.org/10.1371/journal.pone.0085777

[2] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *ACL*. Association for Computational Linguistics, Vancouver, Canada, 451–462. https://doi.org/10.18653/v1/P17-1042

[3] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*. Association for Computational Linguistics, Melbourne, Australia, 789–798. https://doi.org/10.18653/v1/P18-1073

[4] David Bamman, Chris Dyer, and Noah A Smith. 2014. Distributed representations of geographically situated language. In *ACL*. 828–834.

[5] Avishek Joey Bose, Ankit Jain, Piero Molino, and William L. Hamilton. 2019. Meta-Graph: Few shot Link Prediction via Meta Learning. *ArXiv* abs/1912.09867 (2019).

[6] Fei Cai and Maarten de Rijke. 2016. *A Survey of Query Auto Completion in Information Retrieval.* Now Publishers Inc., Hanover, MA, USA.

[7] Hugo Caselles-Dupré, Florian Lesaint, and Jimena Royo-Letelier. 2018. Word2vec applied to Recommendation: Hyperparameters Matter. In *Proceedings of RecSys '18.* https://doi.org/10.1145/3240323.3240377

[8] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16).* Association for Computing Machinery, New York, NY, USA, 191–198. https://doi.org/10.1145/2959100.2959190

[9] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. In *RecSys (RecSys '19).* ACM, New York, NY, USA, 101–109. https://doi.org/10.1145/3298689.3347058

[10] Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. 2019. Training Temporal Word Embeddings with a Compass. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6326–6334.

[11] Ignacio Fernández-Tobías, Iván Cantador, Marius Kaminskas, and Francesco Ricci. 2012. Cross-domain recommender systems: A survey of the State of the Art. *Proceedings of the 2nd Spanish Conference on Information Retrieval* (01 2012).

[12] Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. 2015. E-commerce in Your Inbox: Product Recommendations at Scale. In *Proceedings of KDD '15*. https://doi.org/10.1145/2783258.2788627

[13] N. Gudigantala, P. Bicen, and M. Eom. 2016. An examination of antecedents of conversion rates of e-commerce retailers. *Management Research Review* 39 (2016), 82–114. https://doi.org/10.1108/MRR-05-2014-0112

[14] William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1489–1501.

[15] Manojkumar Rangasamy Kannadasan and Grigor Aslanyan. 2019. Personalized Query Auto-Completion Through a Lightweight Representation of the User Context. *arXiv preprint arXiv:1905.01386* (2019).

[16] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proc. ACL*. https://doi.org/10.18653/v1/P17-4012

[17] Thom Lake, Sinead A Williamson, Alexander T Hawk, Christopher C Johnson, and Benjamin P Wing. 2019. Large-scale collaborative filtering with product embeddings. *arXiv preprint arXiv:1901.04321* (2019).

[18] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

[19] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[20] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013).

[21] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation. *ArXiv* abs/1309.4168 (2013).

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *NIPS*. Curran Associates Inc., USA, 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959

[23] Cun Mu, Guang Yang, and Zheng Yan. 2018. Revisiting Skip-Gram Negative Sampling Model with Regularization. *CoRR* abs/1804.00306 (2018). arXiv:1804.00306 http://arxiv.org/abs/1804.00306

[24] Dae Hoon Park and Rikio Chiba. 2017. A neural language model for query auto-completion. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1189–1192.

[25] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[26] Timo Schreiner, Alexandra Rese, and Daniel Baier. 2019. Multichannel personalization: Identifying consumer preferences for product recommendations in advertisements across different media channels. *Journal of Retailing and Consumer Services* 48 (2019), 87 – 99. https://doi.org/10.1016/j.jretconser.2019.02.010

[27] SimilarWeb. 2019. *Top sites ranking for E-commerce And Shopping in the world.* Retrieved December 23, 2019 from https://www.similarweb.com/top-websites/category/e-commerce-and-shopping

[28] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

[29] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *NIPS*.

[30] Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers)*. 448–453.

[31] Jacopo Tagliabue, Bingqing Yu, and Marie Beaulieu. 2020. How to Grow a (Product) Tree: Personalized Category Suggestions for eCommerce Type-Ahead. In *Proceedings of The 3rd Workshop on e-Commerce and NLP*. Association for Computational Linguistics, Seattle, WA, USA, 7–18. https://www.aclweb.org/anthology/2020.ecnlp-1.2

[32] Techcrunch. [n.d.]. *Algolia finds $110M from Accel and Salesforce.* https://techcrunch.com/2019/10/15/algolia-finds-110m-from-accel-and-salesforce-for-its-search-as-a-service-used-by-slack-twitch-and-8k-others/

[33] Techcrunch. [n.d.]. *coveo-raises-227m-at-1b-valuation.* https://techcrunch.com/2019/11/06/coveo-raises-227m-at-1b-valuation-for-ai-based-enterprise-search-and-personalization/

[34] Techcrunch. [n.d.]. *Lucidworks raises $100M to expand in AI finds.* https://techcrunch.com/2019/08/12/lucidworks-raises-100m-to-expand-in-ai-powered-search-as-a-service-for-organizations/

[35] Flavian Vasile, Elena Smirnova, and Alexis Conneau. 2018. Meta-Prod2Vec - Product Embeddings Using Side-Information for Recommendation. In *Proceedings of RecSys '16*. https://doi.org/citation.cfm?doid=2959100.2959160

[36] Paul Voigt and Axel von dem Bussche. 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide* (1st ed.). Springer Publishing Company, Incorporated.

[37] Po-Wei Wang et al. 2018. Realtime Query Completion via Deep Language Models.. In *eCOM@SIGIR (CEUR Workshop Proceedings)*, Jon Degenhardt, Giuseppe Di Fabbrizio, Surya Kallumadi, Mohit Kumar, Andrew Trotman, Yiu-Chang Lin, and Huasha Zhao (Eds.), Vol. 2319. CEUR-WS.org.

[38] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh acm international conference on web search and data mining*. 673–681.

[39] Bingqing Yu, Jacopo Tagliabue, Ciro Greco, and Federico Bianchi. 2020. An Image is Worth a Thousand Features: Scalable Product Representations for In-Session Type-Ahead Personalization. In *Companion Proceedings of the Web Conference*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3366424.3386198


## A DATA PIPELINE WITH PAAS SERVICES

For practitioners in the same industry, Figure 9 gives a high-level sketch of how the chosen PaaS services fit together in the pipeline:

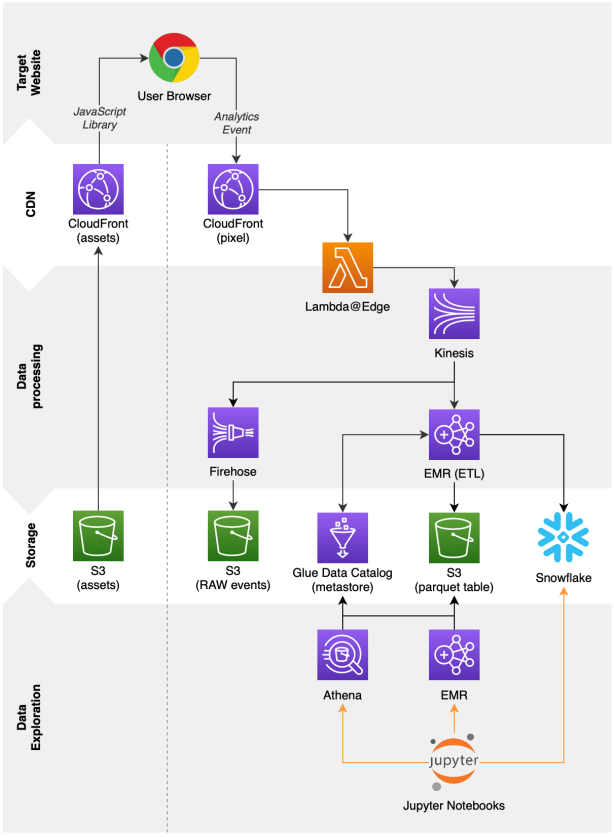


**Figure 9: Cloud-based data ingestion pipeline.**

- the Javascript library is stored on S3 and globally distributed through AWS CloudFront[6];
- the pixel endpoint is reachable through AWS CloudFront, to ensure high performances;
- incoming events are processed by an AWS Lambda@Edge[7] and streamed to internal consumers by using AWS Kinesis[8];

[6] https://aws.amazon.com/cloudfront/
[7] https://aws.amazon.com/lambda/edge/
[8] https://aws.amazon.com/kinesis/

- AWS Firehose[9] is used to persist all the RAW events in S3[10] for future re-processing;
- the ETL processing is done in an AWS EMR[11] Cluster; normalized and sessionized events are then stored on S3 in a Parquet format;
- tables metadata are stored in AWS Glue Data Catalog[12]; data are made querable with Spark-SQL on EMR and AWS Athena[13].
- data are also stored in Snowflake[14] as part of our project for a future simplification of our data warehouse practices.

[9] https://aws.amazon.com/kinesis/data-firehose/
[10] https://aws.amazon.com/s3/
[11] https://aws.amazon.com/emr/
[12] https://aws.amazon.com/glue/
[13] https://aws.amazon.com/athena/
[14] https://www.snowflake.com/