Shopping in the Multiverse: A Counterfactual Approach to In-Session Attribution

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

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

ABSTRACT

We tackle the challenge of in-session attribution for on-site search engines in eCommerce. We phrase the problem as a causal counterfactual inference, and contrast the approach with rule-based systems from industry settings and prediction models from the multi-touch attribution literature. We approach counterfactuals in analogy with treatments in formal semantics, explicitly modeling possible outcomes through alternative shopper timelines; in particular, we propose to learn a generative browsing model over a target shop, leveraging the latent space induced by *prod2vec* embeddings; we show how natural language queries can be effectively represented in the same space and how "search intervention" can be performed to assess causal contribution. Finally, we validate the methodology on a synthetic dataset, mimicking important patterns emerged in customer interviews and qualitative analysis, and we present preliminary findings on an industry dataset from a partnering shop.

CCS CONCEPTS

• **Computing methodologies** → *Learning latent representations*; *Neural networks*; **Causal reasoning and diagnostics**; • **Applied computing** → **Online shopping**.

KEYWORDS

neural networks, search attribution, causal inference

ACM Reference Format:
Jacopo Tagliabue and Bingqing Yu. 2020. Shopping in the Multiverse: A Counterfactual Approach to In-Session Attribution. In *Proceedings of ACM SIGIR Workshop on eCommerce (SIGIR eCom'20).* ACM, New York, NY, USA, 9 pages.

1 INTRODUCTION

> "Futures not achieved are only branches of the past: dead branches." – Italo Calvino, *The Invisible Cities*.

Simon searches for "running shoes" on the sport apparel eCommerce *Balls&Things*: he clicks on a pair of shoes, does not love them, goes back to the running section and starts browsing in other categories: socks, t-shirts, headbands... finally, Simon finds a pair of shorts that he really likes, adds them to cart and completes the transaction. Later that day, Alice, *Balls&Things* director of digital marketing, is looking at *Google Analytics* conversion dashboard: Simon's session is not just a win (i.e. *conversion*), it is a win for *on-site search*.

Is it, though? In this case, it seems pretty obvious that the search behavior and the conversion event are only mildly related, but some other cases are subtler: did our search engine lead Simon to the purchase, or would he have bought shorts anyway? *On-site attribution* is the task of determining the value of each on-site customer touchpoint, such as on-site search, that leads to a conversion. Addressing the challenge in a principled way is important to many stakeholders: Alice, who needs to allocate budget depending on conversion signals; John, *Balls&Things* CTO, who needs to measure search performance; Jean, director of UX, who is redesigning the site experience. In this *short* paper, we present preliminary results in taking a causal approach to on-site search attribution. In particular:

- we phrase the on-site attribution challenge as a *causal* inference: did the search interaction cause the conversion? In turn, we model causality in a counterfactual fashion: would Simon have bought shorts anyway without that interaction? We propose a novel assessment of counterfactual statements in a dense, high-dimensional space, drawing from the notion of *possible world* popularized by modal logic [16] – if we could observe a parallel timeline in which Simon does not search for "running shoes", we could verify what happens to the purchase event;
- we present a deep learning framework allowing for the probabilistic generation of alternative timelines; our solution has minimal barriers to adoption, it requires nothing but essential data tracking and it is straightforward to extend to other functionalities (e.g. recommendation, category listing);
- we test the proposed method on synthetic and industry datasets; we show how to simulate "search interventions", highlight interesting cases and compare it with industry tools.

We believe our findings are interesting to a broad audience: big eCommerce sites with home-grown APIs in need of more accurate measuring tools; mid-size eCommerce sites evaluating external providers and strategic initiatives; multi-tenant SaaS providers that need to measure and communicate the ROI brought by adoption.

*Authors contributed equally to this research project and are listed alphabetically.
†Corresponding author.

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

Figure 1: Bob's session on the sport apparel shop: his journey starts with a query (*1*) and a click (*3*) from the SERP (*2*); he then goes up one level to the category page for *basketball shoes* (*4*), browses another pair of shoes (*5*), finally adds the product to the cart (*6*) and buys it (*7*).

2 GOING BEYOND RULES: AN INDUSTRY PERSPECTIVE

Consider now Bob The Shopper, and his customer journey in Fig. 1: conversion happens after a mix of browsing (*4, 5*) and search behavior (*1, 2, 3*). When Alice looks at Bob's behavior through the lens of industry standard tools – such as *Google Analytics* (henceforth *GA*) or *Adobe Omniture* - Bob's session features in conversion reports as a search win[1], as *GA* just considers whether *any* search interaction was present at some point in a converting session. Since shoppers using search are arguably more motivated to buy in the first place, framing attribution in this way risks conflating search efficiency and prior intent into one measure: if Bob The Shopper searches for "nba shoes" and then buys a headband after extensive browsing on the website, how much credit should we attribute to the search engine?

In Bob's case in Fig. 1, it can be argued that search results containing basketball shoes prompted the user to explore first a particular product (*3*), then the general category page (*4*) and finally after some browsing, purchasing basketball shoes (*7*). The intuitive reasoning behind this example is at the heart of the current proposal: i) *attribution* is a causal relation: search interaction is important only *insofar* as it is causally involved in the conversion; ii) *attribution* is not a binary concept, but a matter of degree; iii) in-session *attribution* is crucially linked to shopper intent: if Bob is looking for basketball-related things, and search points Bob - so to speak - in the "right direction", we are inclined to attribute the final win to the search engine. The model we propose in Section 5 is our attempt of making this type of reasoning more precise.

From an industry perspective, it is important to note that calculating causal attribution via continuous A/B testing is not generally feasible, as switching off the search box for a significant portion of users will bring a harmful impact on business revenue; moreover, when *multiple* providers are involved, A/B tests are harder to manage as more players need to be coordinated to assure the validity of the final results. Generally speaking, our findings support an *holistic* view of A.I. services within a retailer. Search contribution to revenues depends on many factors: UX, site structure, number of SKUs, etc. Deciding how much to invest in Information Retrieval technologies should be part of a general strategy of digital growth more than just a local optimization. *Coveo* is a multitenant SaaS provider with more than 500 clients, ranging from mid-sized business to Fortune 500 companies: *this* paper is therefore part of a larger project combining research and product development.

[1] "Conversions are calculated for sessions that include at least one search on your website.", from https://support.google.com/analytics/answer/1032321?hl=en.

3 RELATED WORK

This short contribution sits at the intersection of multiple fields: we briefly survey existing literature by main topic.

Channel Attribution. The problem of attributing conversion to known prior events is a well studied problem in marketing [8] and, more recently, machine learning [23]; recent work in deep learning on attribution includes LSTM-based frameworks such as [17, 35]. While marketing campaigns could be included in our model, we measure in-session attribution focusing on fine-grained user interactions. There is no action taken outside the target website, and no feature collection is required for user identification or segmentation. It should also be stressed that predicting a conversion event *per se* (as in [35]) may fail to detect the subtle differences among interactions occurring within a session, and thus lead to a biased estimation of causal influence; for example, if session intent is very clear (Fig. **2.1**), search interactions may be predictive but still somehow not very *influential*: by taking a counterfactual perspective, we are trying to isolate conversions that would have *not* happened anyway.

Browsing Models for eCommerce. LSTM-based neural networks [7, 31] have recently been proven to be effective in modelling long-range dependencies in eCommerce browsing. Dense representations for products have been popularized by *prod2vec* models [11, 32]: to the best of our knowledge, *this* is the first work analyzing in-session trajectories through a unique representational space, which includes product embeddings and possibly infinite linguistic behavior.

Causal Inference. Inferring causal relations from observational data is a well-studied topic in philosophy [26], econometrics [13] and classical machine learning [22]. In an eCommmerce setting, [24] estimates the causal importance of recommendations by exploiting instantaneous shock in traffic: data constraints, both in times – nine months – and in traffic – 2M unique users, challenge the widespread applicability of the method to other shops. Matching methods are another common substitutes for randomized trials [27], especially in the medical literature: however, finding two shoppers with, say, 5 matching interactions in a shop with 25k products is almost impossible – in other words, while some of the ideas below are philosophically similar to matching methods, traditional approaches are hard to scale to the dimensions of our space. In the deep learning tradition, [21] proposes a CNN-based technique for multivariate time-series modelling; however, their interest is mostly in understanding how randomly changing a real number affects predictability, which is a much narrow counterfactual than the ones needed for our use case – our proposal crucially relies on taking seriously the multi-dimensional space of possibilities in front of a

shopper that lands on a target shop. Standard packages aimed at democratizing causal reasoning [1] are geared towards traditional statistical methods (regression, random forests etc.) and they cannot scale easily to handle the high-dimensional space of products and linguistic interactions. Finally, for (mostly) nostalgic reasons, it is worth remembering that linking causation to counterfactuals was first done by David Hume[2] and further improved by David Lewis [15]: our method owes a great debt to their revolutionary ideas.

4 METHODOLOGY

Since gold labels of "true attributions" are not available, we proceed to validate our methodology in two steps: first, we analyze hundreds of converting search sessions from partnering *Shop X*, extract important patterns and produce a synthetic dataset with known proportions of "causally relevant interactions". Second, once we confirm that our methodology successfully detects interesting patterns, we apply it to *Shop X* test set and compare the results with other attribution strategies.

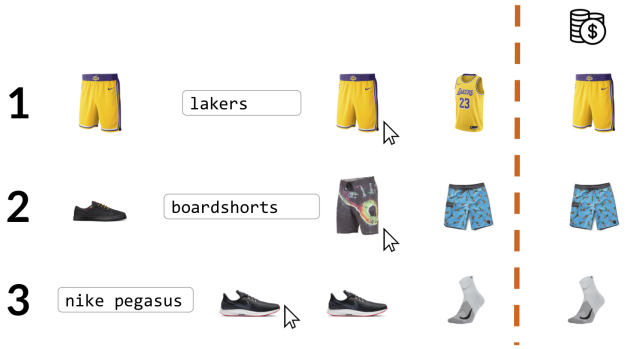

Figure 2: Sample converting sessions from *Shop X*, illustrating important patterns: *Session 1* is a single-intent session; *Session 2* is a multi-intent session, showing how the search engine effectively allows the shopper to move across the website; *Session 3* shows how search interactions may sometime be successful – the user clicks on a product – but the final purchase is unrelated.

4.1 The "Attribution Zoo"

We sampled and analyzed 500+ converting sessions from *Shop X* (Section 4.2) *including search interactions* (query issued by user is followed by one or more products clicked by the shopper). We classify major patterns according to two independent dimensions:

(1) *relevance of interaction for purchase*: some interactions are *directly* linked to a purchase, as the shopper buys exactly the product found through search (Fig. **2.1**); on the other side of the spectrum, some interactions are completely irrelevant: a shopper searches for "sneakers" but then buys socks (Fig. **2.3**). In the middle, cases like Fig. 1 – search interactions is semantically related to the purchase, but product is not exactly the same;

(2) *breadth of shopping intent*: most sessions are single-intent sessions, such that browsing, searching and buying happens within a somewhat specific category of products (e.g. *sneakers*, *lakers merchandise*, etc. – Fig. **2.1**). Other sessions are *multi*-intent: the shopper starts with an intent, e.g. sneakers, and then uses the search box to quickly move to another intent, swimsuit (Fig. **2.2**). Since our model is based on the product space induced by user browsing (Section 5.1), a broader range of intent corresponds to the physical equivalent of shoppers visiting several aisles of a shop. In this analogy, search interactions are a very handy way of "travelling" a great distance with a single action.

The result of the extensive qualitative analysis was validated through interviews with Customer Success Managers and target clients, confirming our intuitions: *ceteris paribus*, more *relevant* interactions should be considered more causally influential for conversion; *ceteris paribus*, the search interaction in Fig. **2.2** has greater influence on conversion than Fig. **2.1**, since in Fig. **2.2** the search engine is instrumental in answering the shopper's inquiry, moving the shopper to a very different region of the product space. These important observations and customer insights are used to build a synthetic dataset out of real shopping events.

4.2 Datasets

Our first dataset is a synthetic dataset (**SD**) built by composing sub-sessions from the available data on *Shop X* – please see Appendix B for the details. **SD** has a total of 1,075,000 sessions, 375,000 with search; sessions with search reflect the main patterns from Section 4.1 in known proportions (175,000 relevant, 200,000 irrelevant); single-intent vs multi-intent sessions are generated with specific code paths, to highlight the qualitative differences within the *relevant* bucket. Our second dataset is an industry dataset (**ID**), obtained by collecting anonymized sessions from June 2019 to September 2019 from our partnering *Shop X* – we use 912,884 sessions for *prod2vec* training (Section 5) and we then run our benchmarks on a total of 10K sampled converting sessions with search. *Shop X* is a mid-size eCommerce, has 26k distinct SKUs, with Alexa Ranking >150k. *Shop X* was an ideal first candidate for this project, since it doesn't leverage recommendation or listing APIs, but only an external search provider.

5 THE SHOPPER MULTIVERSE

The main intuition behind our model is that Bob's session (Fig. 1) is best conceptualized as a path in a vector space, more than a Markov-like sequence of actions [4]. At a first approximation, we can describe our strategy for calculating attribution as two-fold: first, we build a *shared* vector space in which both product browsing and search interactions can live – if events are points in a multi-dimensional space, converting sessions are specific *paths* in that space; second, leveraging the full differentiability of shopping paths, we train a deep recurrent neural network over historical sessions to build a simulation model of browsing: when evaluating a session, we run the model with selected perturbations to assess how much the outcome (i.e. the conversion) would change if events (i.e. the interactions with search) had been different. We now proceed to specify the technical components in detail.

[2]More precisely, in *An Enquiry Concerning Human Understanding*, Section VII.

5.1 Building a shared vector space

We prepare dense vectors for all the products in the target shop. Product embeddings are trained with the CBOW negative sampling [19, 20], by swapping the concept of words in a sentence with products in a browsing session (from *word2vec* to *prod2vec*) [11]; ETL and hyperparameter optimization follows the guidelines presented in [6]; for this task, we use interaction-specific embeddings [37], such that a product with $SKU = xyz$ is embedded in the space as xyz_{detail}, xyz_{click}, $xyz_{purchase}$. Given a product p and interaction I, we denote the associated embedding as $V(p_I)$. Starting from this "shop space", we use the same intuition behind *Search2Prod2Vec* [30] as our query embedding strategy: for a query q, we take the top N items returned by the engine, $p1_{detail}$, $p2_{detail}$, ...pn_{detail} and then create a *deep set* [25] by computing the average of $V(p1_{detail})$, $V(p2_{detail})$, ...$V(pn_{detail})$; in other words, we take the engine response as the *denotation* [29] of the issued query, so that the *meaning* of q is a function from q to a set of products falling under that concept [9][3].

5.2 Training a generative browsing model

LSTMs have obtained SOTA results in language [18] and browsing [7] modelling. In our use case, we exploit both the sequential and generative [28] nature of *LSTMs*: in particular, since the trained network models the conditional probability of a product given the previous ones, $P(x_n|x_0, ...x_{n-1})$, we will use this distribution to guide interventions in a principled way (as opposed to pure random permutation [21]). Our model is trained using *Shop X*'s usage logs, by feeding our pre-trained *prod2vec* embeddings as input at each timestep. For training, we use teacher forcing to pass the vector of the current timestep's target, offset by one position, as the input for the next timestep [33]. Hence, once trained, the model can start with any given input sequence and use autoregression sequence generation to predict the tokens for the next timesteps [2]. In addition to human inspection, we also obtain a $HR@1 = 0.16$ over a hold-out dataset, which confirms LSTM ability to capture user behaviour, as compared to similar models in the product embeddings literature [32]. In order to use the LSTM model to generate our alternative timelines, we experiment with different sampling methods that have proven to work well for non-deterministic sequence generation: *Top-K sampling* [10], *Top-K sampling with temperature* and *Top-P sampling* [12]. Since all these decoding methods are non-deterministic, we generate T samples per alternative timeline (Section 5.3), to reduce the effect of random fluctuations. Preliminary experiments prove that simulation results were consistent across different methods, therefore we only report results obtained with optimal settings: *Top-K* sampling, with $K = 3$ and $T = 100$.

When fully trained, our browsing model implicitly captures two important dimensions: first, how latent user intent shapes the unfolding of a shopping session; second, how site structure implicitly constraints what product is reachable from what (i.e. two pairs of sneakers may be only two clicks away, sneakers and a basketball four). It is the combination of these two latent properties - what the shopper wants and how easy is to get it just by browsing - that makes the model so appealing for counterfactual reasoning.

[3]Empirically, we determined $N = 20$ to be a good threshold.

5.3 Assessing causal influence

C causing an event E is understood in a counterfactual framework as "E would not have happened, if C had not been the case" [15]; in turn, counterfactuals in formal semantics and modal logic are understood through the notion of *possible world* [16], alternative realities in which events are different: what we need to find is a possible world w_1, which is exactly like the actual world $w_@$ up until C occurred, and verify if E happens there without C. Since outer worlds are hard to come by, counterfactual causation is often framed through *interventions* [34], i.e. hypothetical changes made to an upstream variable in a causal chain to verify the downstream effect on the target. In Fig. **2.2**, an *intervention* would be to redirect the shopper to the sneakers category after the first item interaction, instead of letting the shopper search for "boardshorts": a key appeal of the model is its geometrical intuitive interpretation – the more the timeline *post-intervention* ends up far in the shop space from the original purchase, the more the target interaction causally explains the outcome.

More precisely, to measure the change brought by our intervention, we take inspiration from the *difference in difference* approach (*DD*) for time-series, popular in econometrics [3, 14]. In particular, the causal influence of the search interaction is a function of the relative position in the product space of the ending spot for three timelines (Fig. 3): i) the original converting session including a search interaction (@); ii) an alternative timeline, generated by keeping constant all events until *after* the interaction (click on product), and then letting the model unfold into the future (w_1); iii) an alternative timeline, generated by keeping constant all events *before* the query, and then letting the model unfold into the future (w_2). Since shoppers' paths live in a dense product space, an easy way to quantify the *DD* is by making use of the cosine distance in the embedding space. After timeline generation, we obtain two cosine distances by comparing the original purchase spot to the ending spots of our two alternative timelines. To constrain the length of generated samples and allow faster computation, we let the decoding process to stop once the original purchase timestep is reached. As mentioned in Section 5.2, we use *Top-K* sampling to generate 100 samples for each type of alternative timelines, and take the average of their ending spots to represent the ending points for w_1 and w_2. It is the relative ending position of @, w_1 and w_2 that captures the counterfactual importance of search – for Fig. **2.1**, all paths will end in a similar region, for Fig. **2.2**, @ and w_1 will be closer, with w_2 far apart. Several examples of actual and generated paths are collected in Appendix A.

6 EXPERIMENTS

In the absence of golden labels, we first run the proposed multiverse method (**MV**) on our synthetic dataset **SD**, and then run it on actual data comparing the computed attribution with other industry methods.

6.1 Results on SD

We run **MV** on **SD** as a double sanity check: *quantitatively*, we are looking at evidence that the proposed method is able to distinguish relevant vs irrelevant search interactions; *qualitatively*, we are

Figure 3: Causal influence is a function of three timelines in the shop space: the actual converting session (@, *orange*); a *blue* session (for simplicity, shown here as a single session and not an average across many simulations), where the future diverges *after* the interaction; a *green session*, where the future diverges immediately *before* the query is issued.

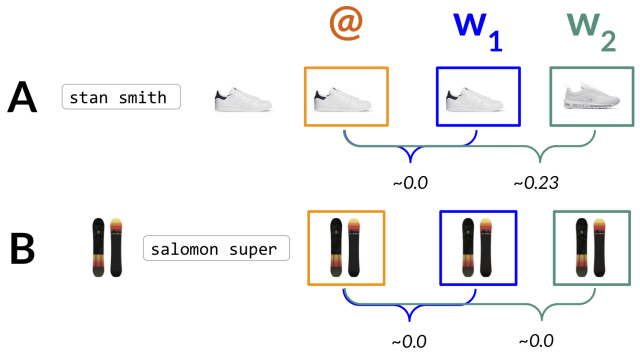

Figure 4: Two converting sessions with search attribution. Session A starts with a query and ends up with purchasing a pair of shoes (*orange*) – products from alternative timelines (average) are shown in *blue* and *green*, with cosine distance below; session B starts with product browsing and ends in a snowboard purchase after a search query – the alternative timelines (on average) all converge to the same product.

looking at evidence that the proposed method is capturing counterfactual nuances we care about, e.g. the difference between Fig. **2.1** and Fig. **2.2**. To obtain an accuracy measure, we first considered writing a decision module with manual rules that would encode the geometrical intuitions of Appendix A; however, it turned out to be more practical to train a small multilayer perceptron to learn the decision boundary starting from the three timelines, instead of coming up with hard-coded rules. After training the binary classifier on 80% of **SD**, we test **MV** using hold-out samples and compare predictions with the synthetic labels, obtaining as accuracy 0.93, recall 0.93 and F1 0.93. Outside of the binary context, we also report a "geometrical interpretation" of the timelines generated by **MV**, for the four types of causal patterns to be found in **SD**. The *Distance Score* (*DS*) is calculated starting from the cosine distance (d) between the last event in w_1 and @, and between w_2 and @: $DS = d(w_2, @)/d(w_1, @) * (1/d(w_1, @))$; in other words, *DS* captures how far from the actual events the generated timelines are (using $d(w_1, @)$ as a normalizing factor). The *DS* average values – after re-scaling and log-normalization – reported in Table 1 show from a non-binary perspective **MV**'s ability to not just distinguish relevant vs irrelevant interactions (positive vs negative *DS*), but also to make subtler distinctions and assign appropriate causal weights. We perform error analysis on misclassified samples and identified three patterns for future improvement:

(1) *low quality embeddings*: when SKUs and/or issued queries suffer from data sparsity, timelines become more erratic; several options for "cold-start" embeddings exist in the *prod2vec* literature, with varying degrees of engineering disruption [32], and it is certainly an important insight for future work;
(2) *regression to the mean*: when a somewhat irrelevant search is issued in a session happening in the most crowded region of the space – i.e. *shoes* –, all timelines still converge, mimicking a relevant pattern;
(3) *sensitivity to pre-search intent*: when many browsing interactions happen before a *relevant* search, **MV** tends to mark search as irrelevant; while technically those queries were created to be causally connected to the purchase, this behavior is a very interesting side-effect of embracing counterfactual reasoning. In particular, we verified that shortening the sequence of events before search makes **MV** predict the correct (*relevant*) label: the more shopping intent is displayed before the query, the more the model thinks conversion would have happened *anyway*.

Table 1: Distance Score for four distinct causal patterns: search unrelated to conversion (*CU*), single intent with exact match (*SE*), multi-intent (*ME*), single intent with related-but-not-exact match (*SR*).

Pattern	Distance Score
CU	-1.062
SE	2.315
ME	3.359
SR	2.262

From a qualitative perspective, we manually inspect sample sessions belonging to different patterns. In particular, after quantitatively validating the overall accuracy of **MV**, we are now interested to see how counterfactual inference is represented in the model. Consider **A** and **B** in Fig. 4: they are both converting sessions accurately attributed to the search interaction. However, search plays a different role in them: in **A**, the shopper lands on the website and search is causally responsible to guide her immediately to fulfill her intent; in **B**, the shopper is browsing Salomon boards even before issuing a query. **MV** is able to capture the difference between those two cases thanks to the w_2 timeline: in **B**, search is less important as the intent is clear enough to make us believe that the shopper would have purchased anyway that board. Please refer to Appendix A for the full charts of the patterns in Fig. 2.

Figure 5: A converting session, featuring browsing and add-to-cart event *before* any query is issued: for SS the query is predictive of the purchase; MS, on the other hand, thinks the shopper was going to buy the shoes *anyway*.

6.2 Results on ID

Experimental results are presented in Table 2, reported as percentage of conversions for sessions with search that are actually *attributed* to search interactions by the selected methodology. First, we compare **MV** with industry standard heuristics, as customary in the multi-touch attribution literature [35]; *GA*-style attribution is a natural upper bound as *any* converting session with search will be counted positively and it is therefore likely to over-estimate the causal impact. Second, we report another common heuristics, a "click-then-buy" rule (**CB**) to the effect that products count as causally related only when purchased items are among after-search clicks; third, we implement a "semantic similarity" method (**SS**), which measures the distance in the embedding space between the target interaction and the purchased product. To simplify the comparison with industry standards, for **MV** and **SS** we report the results of the binary classifier, trained for both as described in Section 6.1.

We sample interesting sessions from the test set (i.e. sessions where models disagree on the causal interpretation) and make three main observations:

(1) all *non-GA* models reach the same conclusion as [24]: once industry heuristics are replaced by stricter rules or more sophisticated inference, the magnitude of direct attribution is less than what was naively thought; it is interesting to note that search impact seems to be significantly bigger than what the literature estimates for recommendations;
(2) **MV** and **SS** are more flexible than **CB**, which relies on a deterministic match that excessively penalizes interactions that are still valuable, such as Bob's query in Fig. 1;
(3) **MV** and **SS** tend to agree on the overall score, but **SS** ignores entirely the pre-search intention, and therefore it is behaving as a *prediction* method ("how closely related are interaction and purchase?"), more than a *causal* one ("would Bob have bought those shoes anyway?"). Fig. 5 shows a session where predictions from **SS** and **MV** diverge: since the shopper issued a query *after* browsing and adding to cart the target product, **MV** marks the search interaction as weakly relevant, since most alternative timelines would end in the same way. For the reasons explained at length in Section 2, we believe that the ability to capture this type of causal dynamics is what makes **MV** so appealing when discussing in-session attribution.

Based on the quantitative performance on **SD**, and the qualitative analysis on both **SD** and **ID**, we conclude that we have strong *prima facie* reasons to consider attribution judgments by **MV** an accurate representation of the underlying causal dynamics. As we stress in the ensuing section, extending the validation to other shops / verticals / services is a natural next step to confirm the generalizability of these findings.

Table 2: Search attribution on ID for all the methods.

Method	% Search Attr. **ID**
GA	100
CB	47
SS	77
MV	75

7 CONCLUSIONS AND FUTURE WORK

In *this* paper we presented preliminary results towards a more sophisticated understanding of A.I. services in the context of attribution; by leveraging the link between causality and counterfactual reasoning, on one hand, and counterfactuals and generative models, on the other, we were able to frame in-session attribution as a question of model behavior under perturbation – contrary to approaches to user browsing (and attribution) where state-space is constrained by artificially coarse-grained models, we did it by taking seriously the high-dimensional nature of the underlying space, in which thousands of products and *unbounded* linguistic interactions co-exist. The proposed method, while now applied to on-site attribution for eCommerce search, is general enough to be applicable to any situation in which a generative model can be successfully trained to represent the space of possible actions for the target user. Our findings support the pre-theoretical intuition that *relevance* with respect to the underlying intent is crucial for conversion.

While results are very encouraging, our claims are still limited in scope, as we extensively validated only one shop. Our roadmap starts with adding use cases: *Shop X* was an ideal first candidate as we could test search interactions in isolation, but *recommendation*, *category listing* and even marketing actions are straightforward extensions once the multiverse is generalized to multiple services per session. It is also important to highlight that *in-session* attribution, by design, treats all incoming users as *distinct*: no long-term pattern is considered, and all shoppers are treated in the same way; this simplification was justified by *Shop X* statistics – <9% of users visited the site for more than 2 times in a year –, but more nuanced treatments are possible.

It should be obvious that our counterfactuals are reliable only insofar as the timeline generating model is; there may be better generative models for specific cases, such as, for example, sequence-based GANs [36]; furthermore, it should be possible to engineer a "human-in-the-loop" evaluation to collect relevance judgments to better align model objectives with human intuitions – going from a purely binary concept of attribution to a continuous one is in itself an interesting product / UX challenge. Finally, while *do*-calculus [22] has been historically developed for a discrete, low-dimensional world, it would be interesting to draw explicit and formal connections between our geometric model in high-dimensional spaces and the basic theory of causal inference.

While we do recognize that controlling for all the causal effects in our method is more challenging [5] than in graphical models, we also value tangible *practical progress*: having a fully unbiased estimate of search importance *may* not be possible at scale, but we can (and should) still strive to improve our attribution strategy over heuristics we know *for sure* to be partial.

ACKNOWLEDGMENTS

Thanks to Andrea Polonioli and Federico Bianchi for the usual thoughtful comments, and to our reviewers for greatly helping improving the paper. We had great discussions with Ciro Greco, Giovanni Cassani, Eric Savoie and Charles Fortier on previous versions of this work, and we would like to thank Mattia Pavoni for gathering valuable client feedback. Finally, thanks to Bingqing's mother for giving us a much needed shopper's perspective – and for great dumplings.

REFERENCES

[1] 2019. DoWhy: A Python package for causal inference.
[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2015).
[3] Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. 2002. *How Much Should We Trust Differences-in-Differences Estimates?* NBER Working Papers 8841. National Bureau of Economic Research, Inc. https://ideas.repec.org/p/nbr/nberwo/8841.html
[4] Dimitris J. Bertsimas, Adam J. Mersereau, and Nitin R. Patel. [n.d.]. *Dynamic Classification of Online Customers.* 107–118. https://doi.org/10.1137/1.9781611972733.10 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611972733.10
[5] Timothy Besley and Anne Case. 1994. *Unnatural Experiments? Estimating the Incidence of Endogenous Policies.* Working Paper 4956. National Bureau of Economic Research. https://doi.org/10.3386/w4956
[6] Federico Bianchi, Luca Bigon, Jacopo Tagliabue, Bingqing Yu, and Ciro Greco. 2020. Fantastic Embeddings and How to Align Them: Zero-Shot Inference in a Multi-Shop Scenario. *eCOM@SIGIR* (2020).
[7] Luca Bigon, Giovanni Cassani, Ciro Greco, Lucas Lacasa, Mattia Pavoni, Andrea Polonioli, and Jacopo Tagliabue. 2019. Prediction is very hard, especially about conversion. Predicting user purchases from clickstream data in fashion e-commerce, In 4th International Workshop on Fashion and KDD (Anchorage, USA). *ArXiv*.
[8] John Chandler-Pepelnjak. 2009. Measuring roi beyond the last ad. *Atlas Institute Digital Marketing Insight* (2009), 1–6.
[9] Gennaro Chierchia and Sally McConnell-Ginet. 2000. *Meaning and Grammar (2nd Ed.): An Introduction to Semantics.* MIT Press, Cambridge, MA, USA.
[10] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *ACL*.
[11] 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
[12] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. *ArXiv* abs/1904.09751 (2020).
[13] Michael Lechner. 2011. The Estimation of Causal Effects by Difference-in-Difference Methods. *Foundations and Trends® in Econometrics* 4, 3 (2011), 165–224. https://doi.org/10.1561/0800000014
[14] Michael Lechner. 2011. The Estimation of Causal Effects by Difference-in-Difference Methods. *Foundations and Trends(R) in Econometrics* 4, 3 (2011), 165–224. https://EconPapers.repec.org/RePEc:now:fnteco:0800000014
[15] David Lewis. 1973. Causation. *The Journal of Philosophy* 70, 17 (1973), 556–567. http://www.jstor.org/stable/2025310
[16] David K. Lewis. 1973. *Counterfactuals.* Blackwell.
[17] Ning Li, Sai Kumar Arava, Chen Dong, Zhenyu Yan, and Abhishek Pani. 2018. Deep Neural Net with Attention for Multi-channel Multi-touch Attribution. *ArXiv* abs/1809.02230 (2018).
[18] Gábor Melis, Chris Dyer, and Phil Blunsom. 2018. On the State of the Art of Evaluation in Neural Language Models. *ArXiv* abs/1707.05589 (2018).
[19] Tomas Mikolov et al. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (Lake Tahoe, Nevada) *(NIPS'13).* Curran Associates Inc., USA, 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959
[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] Meike Nauta, Doina Bucur, and Christin Seifert. 2019. Causal Discovery with Attention-Based Convolutional Neural Networks. *Machine Learning and Knowledge Extraction* 1, 1 (Jan 2019), 312–340. https://doi.org/10.3390/make1010019
[22] Judea Pearl. 2000. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, USA.
[23] Xuhui Shao and Lexin Li. 2011. Data-Driven Multi-Touch Attribution Models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Diego, California, USA) *(KDD '11).* Association for Computing Machinery, New York, NY, USA, 258–264. https://doi.org/10.1145/2020408.2020453
[24] Amit Sharma, Jake M. Hofman, and Duncan J. Watts. 2015. Estimating the Causal Impact of Recommendation Systems from Observational Data. In *EC '15.*
[25] Maximilian Sölch, Adnan Akhundov, Patrick van der Smagt, and Justin Bayer. 2019. On Deep Set Learning and the Choice of Aggregations. In *ICANN.*
[26] Peter Spirtes. 2010. Introduction to Causal Inference. *J. Mach. Learn. Res.* 11 (Aug. 2010), 1643–1662.
[27] Elizabeth A. Stuart. 2010. Matching Methods for Causal Inference: A Review and a Look Forward. *Statist. Sci.* 25, 1 (02 2010), 1–21. https://doi.org/10.1214/09-STS313
[28] Ilya Sutskever, James Martens, and Geoffrey Hinton. 2011. Generating Text with Recurrent Neural Networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning* (Bellevue, Washington, USA) *(ICML'11).* Omnipress, Madison, WI, USA, 1017–1024.
[29] Jacopo Tagliabue and Reuben Cohn-Gordon. 2019. Lexical Learning as an Online Optimal Experiment: Building Efficient Search Engines through Human-Machine Collaboration. *ArXiv* abs/1910.14164 (2019).
[30] 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
[31] Arthur Toth, Louis Tan, Giuseppe Di Fabbrizio, and Ankur Datta. 2017. Predicting Shopping Behavior with Mixture of RNNs.
[32] 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
[33] Ronald J. Williams and David Zipser. 1989. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks.
[34] James Woodward. 2003. *Making Things Happen: A Theory of Causal Explanation.* Oxford University Press.
[35] Dongdong Yang, Kevin F. Dyer, and Senzhang Wang. 2020. Interpretable Deep Learning Model for Online Multi-touch Attribution. *ArXiv* abs/2004.00384 (2020).
[36] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. *ArXiv* abs/1609.05473 (2017).
[37] Xiaoting Zhao, Raphael Louca, Diane Hu, and Liangjie Hong. 2020. The Difference Between a Click and a Cart-Add: Learning Interaction-Specific Embeddings. In *Companion Proceedings of the Web Conference* (Taipei, Taiwan). Association for Computing Machinery, New York, NY, USA.

A VISUALIZATION OF SAMPLE SESSIONS

Embedding product interactions and linguistic behavior in a unified dense space (Section 5.1) allows quite literally to model shopping sessions as *paths* in the underlying space, in which products and queries that are semantically related live close by (Fig. 6).

The attribution model put forward in Section 5.3 is based on the relative position of the last item in three timelines: the actual converting session we are evaluating, and two sessions generated by simulation through the trained deep neural network. One of the greatest strength of the counterfactual approach we propose is its intuitive representation, and the ability to make principled distinctions between different types of search influence (great, mild, almost nonexistent). We collect here four projections in the product space of representative timelines:

(1) Fig. 7 shows the ending states of the timelines corresponding to *Session 1* and *Session 2* in Fig. 2. The model labels both sessions as sessions in which search was relevant to the purchase, but the geometric interpretation of the underlying timelines highlights the big difference between the two

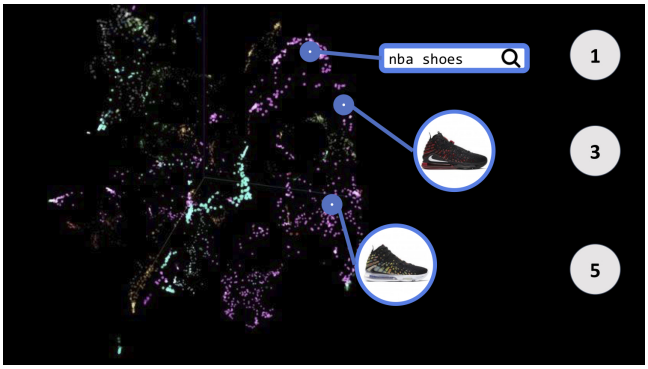

Figure 6: Bob's journey (Fig. 1) in the shop space: by building a shared vector space for browsing interactions and services (search/recommendation), we can model conversions as paths in a dense space. In the image, *prod2vec* embeddings from the target shop are represented through T-SNE; products are color-coded by sport activity (basketball, running, soccer etc.).

cases: in the "lakers" case, search is relevant but the shopper was already interested in that part of the space – in other words, she was likely to found that product anyway; in the "boardshorts" case, the original (*orange*) *and* the *after search* timelines (*blue*) are very close, while the *no search* (*green*) timeline ends up in a very different region, since that future diverges quickly as the model predominantly explore the sneaker portion of the space. In this case, search influence on the final conversion is larger.

(2) Fig. 8 shows the ending states of the timelines corresponding to *Session 3* in Fig. 2 plus a multi-intent, not related session (not shown for brevity). The model correctly labels both sessions as sessions in which search was *not* relevant to the purchase – also in this case, the geometric interpretation provides an intuitive understanding of the model behavior: the three timelines in both cases are very far from each other in the product space, with no immediate correlation between the target points (compare with Fig. 7).

B GENERATING A SYNTHETIC DATASET

Generating synthetic datasets with known dependencies is a common strategy in the causal literature, since labels on real-world datasets are impossible to obtain. We detail here the process of generating a synthetic dataset starting from an existing *Shop S*, by dividing the final dataset in four major subsets: non-converting sessions NC, converting session without search CW, converting sessions related to a search query CS, converting session unrelated to search interaction CU. To capture the qualitative subtleties analyzed in Section 4, CS sessions fall in turn in three buckets: single-intent, "exact match" queries SE; multi-intent, "exact match" queries ME; single-intent, "related match" queries SR. The parameters k and r control the cardinality $C(D)$ of the resulting synthetic dataset D, as specified below:

- $C(CW) = k$ and $C(NC) = k * 2$;

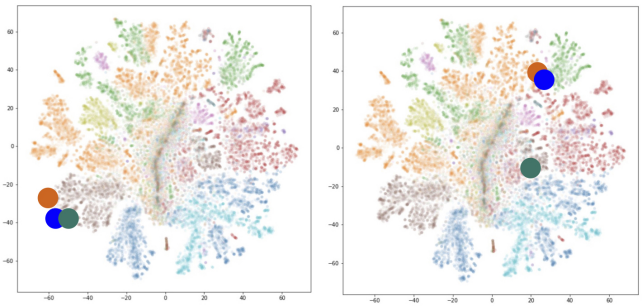

Figure 7: Single-intent (*left*) and multi-intent (*right*) sessions in which the search interaction is causally relevant for the final purchase. The picture shows the ending states of three timelines for the cases in Fig. 2.1 and Fig. 2.2: *orange* is the actual world (@ in Fig. 3), *blue* is the timeline generated after the search interaction (w_1), *green* is the timeline generated before the search interaction (w_2).

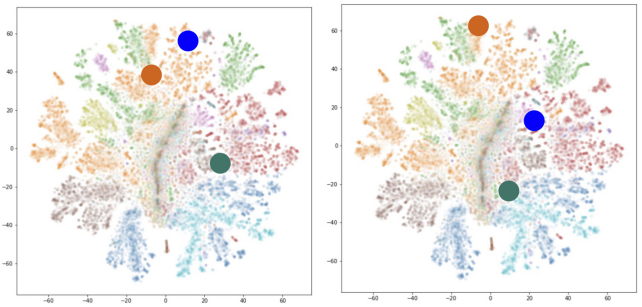

Figure 8: Single-intent (*left*) and multi-intent (*right*) sessions in which the search interaction is *not* causally relevant for the final purchase. The picture shows the ending states of three timelines for the cases in Fig. 2.3 and an additional case, in which search is not relevant and the shopper travels extensively in the underlying space: *orange* is the actual world (@ in Fig. 3), *blue* is the timeline generated after the search interaction (w_1), *green* is the timeline generated before the search interaction (w_2).

- $C(CU) = k$;
- $C(SE) = k * r$ where $0.0 \leq r \leq 1.0$;
- $C(ME) = k * r/2$;
- $C(SR) = k * r/4$.

We generate NC, CS and CU in the planned cardinality with these methods:

- **NC**: we sample $C(NC)$ non-converting sessions from a real *Shop S*;
- **CU**: we sample $C(CU)$ converting sessions *without* search interactions; we inject in each session a search interaction randomly sampled from the dataset; the insertion position is sampled proportionally to empirical frequencies in converting sessions with search queries;
- **SE**: we sample $C(SE)$ non-converting sessions with search interaction involving a click on product Px; sampling from

the underlying shop space in proportion to vector similarity, we generate up to $e = 3$ random product interactions between Px_{click} and a synthetic event Px_{add}, and do the same for $i = 2$ simulated actions between Px_{add} an $Px_{purchase}$.

- **ME**: we sample a random starting product Px uniformly from the product space and generate up to $y = 4$ product interaction repeatedly sample new items in proportion to vector similarity. We then sample $n = 20$ search interactions from search sessions, and select the query+click pair which is the farthest from the current product (to simulate a "jump" in the product space as in Fig. **2.2**). After the click, we proceed as for **SE** above: we generate up to $e = 3$ random interactions between Px_{click} and a synthetic event Px_{add}, and do the same for $i = 2$ simulated actions between Px_{add} an $Px_{purchase}$
- **SR**: we sample $C(SR)$ non-converting sessions with search interaction involving a click on product Px; we use the underlying space to select a target purchase of nearby product Rx; we then proceed to generate up to $e = 3$ random interactions between Px_{click} and a synthetic event Rx_{add}, and do the same for $i = 2$ simulated actions between Rx_{add} an $Rx_{purchase}$.

i, e and y above are chosen after some empirical tests in order to produce shopping sequences which represent realistic browsing for shoppers while at the same time maintain a uniform "underlying intent" to test the model ability in picking-up the relevant long-range behavioral dependencies.

It is important to note that the methodology is pretty general, insofar as it relies on *prod2vec* as a proxy for semantic product proximity: if more/different qualitative patterns of causal influence are needed, it is straightforward to extend the generating procedure to include them as well.