Escaping the McNamara Fallacy: Towards more Impactful Recommender Systems Research

Dietmar Jannach[1] and Christine Bauer[2]

[1]University of Klagenfurt, dietmar.jannach@aau.at
[2]Utrecht University, c.bauer@uu.nl

This is the accepted version of the paper appearing in the AI Magazine. The final publication is available at: https://ojs.aaai.org/index.php/aimagazine/article/view/5312

Jannach, Dietmar & Bauer, Christine (2020). Escaping the McNamara Fallacy: Toward More Impactful Recommender Systems Research. AI Magazine, 41(4), pp 79–95. DOI: 10.1609/aimag.v41i4.5312

Abstract

Recommender systems are among today's most successful application areas of AI. However, in the recommender systems research community, we have fallen prey of a McNamara fallacy to a worrying extent: In the majority of our research efforts, we rely almost exclusively on computational measures such as prediction accuracy, which are easier to make than applying other evaluation methods. However, it remains unclear whether small improvements in terms of such computational measures actually matter a lot and whether they lead us to better systems in practice. A paradigm shift in terms of our research culture and goals is therefore needed. We cannot focus exclusively on abstract computational measures any longer, but must direct our attention to research questions that are more relevant and have more impact in the real world. In this work, we review the various ways of how recommender systems may create value; how they, positively or negatively, impact consumers, businesses, and the society; and how we can measure the resulting effects. Through our analyses, we identify a number of research gaps and propose ways of broadening and improving our methodology in a way that leads us to more impactful research in our field.

1 Introduction

Whenever we visit our favorite media streaming site, check for updates on social media, or shop online, it is very likely that the content we see is personalized and tailored to our interests and needs. Recommender systems are the technology behind this automated adaptation and personalization, and they are among the most successful applications of AI in practice. The broad successful commercial use of modern recommender systems dates back to the late 1990s [Schafer et al., 1999]. Amazon.com was among the early adopters, realizing that there is an enormous potential value in providing customers with automated recommendations. Specifically, they reported vastly improved click-through and conversion rates with personalized recommendations compared to situations where they presented unpersonalized content [Linden et al., 2003]. Nowadays, recommendations have become an ubiquitous component of our online user experience, e.g., on e-commerce sites, video and music streaming platforms, and on social networks.

The huge success of recommender systems in practice has led to a continuously growing academic interest in this area and recommender systems have become their own research field over the past twenty years. Today, also boosted by the recent boom in machine learning, academic research on recommender systems mainly focuses on the continuous improvement of the algorithms. A large number of papers are published each year that propose new algorithms that are used to filter and rank the content that is presented to the consumer, claiming to be better than the state-of-the-art

in a certain dimension. The most important dimension for researchers is being able to accurately predict the relevance of individual items to consumers, with the goal of presenting the assumedly most relevant ones as recommendations.

To provide evidence that a new algorithm is better than an existing one, the community has developed a standardized research approach. This research method, broadly speaking, in most cases consists of comparing different algorithms in terms of their ability to predict preference information contained in a held-out test set. We outline the principles of such a typically used "matrix completion" research operationalization in Figure 1.

Although this research approach has several advantages—like being repeatable and independent of a specific application domain—it can represent a severe over-simplification of the underlying problem. Being able to predict the relevance of an item for a consumer with high confidence is, without a doubt, an important ingredient for any successful recommender system. However, even the most accurate prediction can be worthless or even lead to bad recommendations or other undesired effects, e.g., when the consumer's context or the intended purpose of the recommender system are not taken into account. For example, even a perfect prediction of the consumer's interest in a shopping item on an e-commerce site can be of little value for the company in case the customer would have bought this item anyway. Even worse, recommending such items can—even in cases where we are absolutely sure they will be liked by the customer—lead to missed sales opportunities for other items [Bodapati, 2008].

A fundamental problem of our research, thus, lies in the fact that—unlike in other application domains of machine learning, e.g., in automated translation or image recognition—higher prediction accuracy does *not necessarily* lead to a better (e.g., more effective) system. In fact, there are a number of studies that indicate that the results from offline experiments are *not* indicative of the effectiveness of an algorithm in practice, see, e.g., the case of Netflix [Gomez-Uribe and Hunt, 2015] or the results from a number of other studies [Rossetti et al., 2016, Cremonesi et al., 2012, Garcin et al., 2014, Maksai et al., 2015, Beel and Langer, 2015, Ekstrand et al., 2014, McNee et al., 2002].

Despite this evidence, we observe patterns of a "leaderboard chasing" culture in algorithms research, where the main or only research goal is to outperform other algorithms in terms of prediction accuracy by a few percent, usually without being based on theory or a specific research hypothesis. In some ways, we therefore seem to have fallen prey to a "McNamara fallacy". This fallacy refers to decision-making based solely on quantitative measures and in particular on measures that are easy to take. It is named after US Secretary of Defense Robert McNamara, who is said to have relied too much on such measures during the Vietnam war.

To analyze the extent of this problem, we scanned the proceedings of major conference series for papers on recommender systems. In this process, we for example identified 117 relevant papers that were published at AAAI and IJCAI in 2018 and 2019. Looking at the methodological approach in these papers, it turned out that over 92% of the papers relied *exclusively* on offline experiments. Only a handful of papers combined offline experiments with a user study, and another small set of papers very briefly reported outcomes of a controlled field experiment (A/B test). Papers that were published at ACM RecSys in the same years are more diverse in terms of the methodological approach, in particular because user-centric research is explicitly mentioned in the topics of interest. Still, even at ACM RecSys almost three of four papers solely use offline experimentation.

As a result of the known limitations of this predominant research approach, it remains unclear how much impact our academic work has in practice. Currently, our machine learning models become increasingly complex, but ultimately we cannot be sure that these claimed innovations matter in real-world applications. Even worse, there exist indications that at least some improvements in accuracy were only obtained because too weak or non-optimized baselines were chosen [Lin, 2019, Rendle et al., 2019, Ferrari Dacrema et al., 2019, Makridakis et al., 2018]. At this point we however want to emphasize that we in no way argue that complex models would not be useful or effective in practice. In fact, a number of reports on successful deployments of complex models based on matrix factorization or deep learning recommenders exist, e.g., for YouTube [Covington et al., 2016]. However, in this latter case and in similar works on real-world deployments, the success is measured in terms of particular application-specific key performance indicators (KPIs). Unfortunately, such works typically provide little information about the compared baselines, the absolute size of the improvements, and how the algorithms perform in an offline evaluation.

Overall, we therefore argue that we require a paradigm shift in how we conduct research on recommender systems. One main ingredient of future, more impactful research is to move beyond our sometimes over-simplifying problem abstractions and to consider the various ways that recommender systems have effects on their consumers, businesses, or the society. With this paper, we

Recommendation as a Matrix Completion Problem see also Jannach et al. 2015

The usual input for offline experiments in recommender systems research is a sparse user-item interaction matrix M, describing, e.g., how users rated items or whether they purchased a certain item. The following matrix shows the ratings that were given by four users on five items.

M

Users/Items	Item-1	Item-2	Item-3	Item-4	Item-5
User-1	3		4	5	
User-2	2	4		3	3
User-3	4	5			
User-4	3		5	3	4

x	Known Ratings
	Missing Ratings

For the experiment, a subset of the known entries in the matrix are withheld, leading to matrix M'.

M'

Users/Items	Item-1	Item-2	Item-3	Item-4	Item-5
User-1	3		4	?	
User-2	2	4		3	?
User-3	?	5			
User-4	3		?	3	4

x	Known Ratings
	Missing Ratings
?	Withheld Ratings

The matrix M' subsequently serves as a basis to train a machine learning model that can be used to predict values for the empty cells of the matrix, thus the term *matrix completion*. The quality of such a model can be quantitatively assessed by comparing the predicted values and the withheld ones (i.e., the ground truth). A common performance measure is the Root Mean Squared Error (RMSE). Given a vector of hidden true ratings $\vec{h}$ and vector of predicted ratings $\vec{h}^*$ of length N, RMSE is

$$\sqrt{\frac{1}{n}\sum_{i=1}^{N}(\vec{h}_n^* - \vec{h}_n)^2}$$

Alternatively, various classification and ranking measures, e.g., precision, recall, or the mean reciprocal rank, are applied in the literature.

Figure 1: Overview of the Matrix Completion Research Operationalization

contribute an analysis and categorization of the different forms of such effects and how those effects can be measured. Based on this analysis, we derive how we should extend or adapt our research practice to deliver findings that have an impact in the real world.

Next, we elaborate why evaluating recommender systems can be very challenging and we point out a number of research gaps. Building on this, we put forward a number of specific directions how we can improve our research practices. With this, our work both synthesizes previous insights from [Jannach and Adomavicius, 2016, Jannach and Jugovac, 2019, Abdollahpouri et al., 2020, Bauer and Zangerle, 2019] and provides a forward-looking perspective on recommender systems research.

2 Impact of Recommender Systems: Purpose, Value, and Risks

In the literature, recommender systems are commonly characterized as tools that help consumers find items of interest in situations of information or choice overload. Such a definition matches our standard research approach very well, where *(i)* the system's task is to predict the relevance of the items for individual consumers and where *(ii)* we equate higher prediction accuracy with better recommendation quality and better user experience.

Although relevance prediction is a central problem for any recommender system, the conceptualization and understanding of what relevance connotes is rather narrow in recommender systems research. One underlying assumption of its conceptualization is, for example, that the recommendations are exclusively optimized to match the end consumer's interests. In reality, however, the goals of other stakeholders, in particular those of the service providers, may be equally or even more relevant. Likewise, the intended purpose of the system (i.e., helping the consumer find relevant content) is monodimensional. Recommender systems can in fact serve various purposes, both for consumers and providers, and they correspondingly may create value for the involved stakeholders in different ways [Abdollahpouri et al., 2020].

In the following sections, we will, as the first contribution, provide a more multi-faceted picture

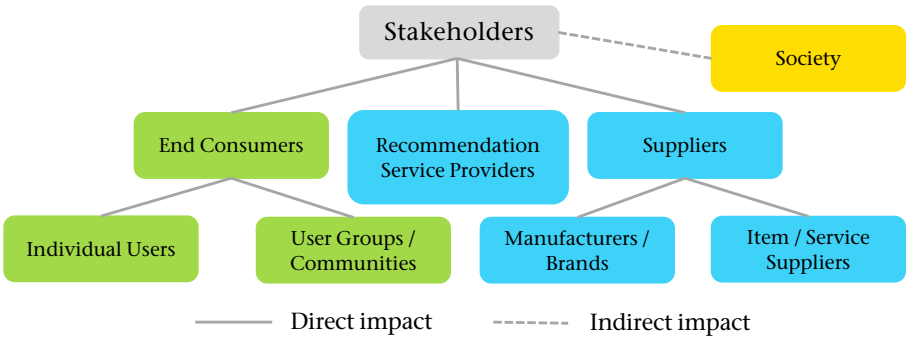

Figure 2: Possible Stakeholders of a Recommender System

of how recommender systems have an impact on various stakeholders, what the recommenders' purposes are, how they may create value, and which risks they might bear. Most of these areas are, unfortunately, largely underexplored. While individual works can be found in the literature that address some of the issues, we find that major research gaps remain, which we point out in this section.

2.1 The Multiple Stakeholders of Recommender Systems

Most research focuses on the value for the *end consumer* of a recommendation service, e.g., consumers on an e-commerce site or users of a media streaming service. Multiple other stakeholders are, however, affected by the existence of a recommendation service. The observed impact can furthermore depend on the particular way the system is configured, e.g., whether the system optimizes for the platform provider or aims to achieve a win-win situation. Figure 2 categorizes possible stakeholders of a recommender system, with stakeholders that represent businesses or organizations shown in blue boxes.

The main stakeholders can be characterized as follows:

- *(End) Consumers*: These are the persons who receive the recommendations. Besides individual consumers, recommender systems can also be designed to support decision-making processes of groups, leading to a group recommendation problem [Masthoff, 2015]. Finally, a system has also an impact on an entire community of consumers through its recommendations, e.g., when it reinforces behavioral patterns in the collective behavior of consumers through collaborative filtering techniques.
- *Recommendation Service Providers*: These are the organizations that provide a recommendation service as part of their business or, more generally, to support their organization's goals. These providers are typically the ones that are in control of the used recommendation algorithms and their configurations. Examples for such service providers are online retailers such as Amazon, streaming media services such as Spotify or Netflix, social media sites such as Facebook, or news portals such as Google News.
- *Suppliers*: These are businesses or organizations that create or provide the items that are recommended to consumers through the recommendation service. Depending on the domain, these are, for example, hotel chains who market their offerings through booking platforms, or manufacturers of items that are sold on an e-commerce platform. Suppliers may also be retailers by themselves who use a larger platform such as Amazon as a sales channel. In some cases the recommendation service providers might also be the suppliers themselves.
- *Society*: Ultimately, if the recommendation service is prominent enough (e.g., on a social media or global news site), the recommendations can even have an indirect impact on the society as well, e.g., by creating filter bubbles or echo chambers.

We may certainly assume that there are many situations where optimizing the recommendations for the consumers' experience will directly or indirectly benefit the provider's goals. This is, for instance, the case when more useful recommendations lead to more sales or higher consumer engagement. There are, however, also many situations, where there are potential trade-offs between the goals of the different stakeholders.

Consider the example of a hotel chain that markets its property through a booking platform, and a consumer who searches the platform for a hotel. The consumer's goal is usually to find a hotel that matches her preferences, e.g., in terms of the price or location. The hotel chain, on the other hand, is interested in being listed as a recommendation on the booking site even if the match with the given consumer preferences is not exact. The chain's interest might furthermore be to present those hotels more prominently where they have an overcapacity. The main business model of the booking platform, finally, might consist of charging a booking commission on a percentage basis to the hotel chain. This, as a result, might seduce booking platforms to promote hotels with a higher commission. At the same time, however, long-term relationships with consumers as well as hotel chains are important. Table 1 summarizes the different and potentially conflicting stakeholder goals.

Stakeholder	Goal
Consumer	Searches for a hotel with an acceptable price close to the city center; already a potential trade-off.
Hotel Chain	Wants to be recommended even if it is not a perfect match; may be interested in getting rid of overcapacity.
Booking Platform	Wants to maximize commission; but also interested in long-term relationships with the various other stakeholders.

Table 1: Potentially conflicting stakeholder goals in the tourism domain.

Overall, the underlying optimization problem for a recommendation system can involve multiple objectives that have to be considered in parallel. The research literature on multi-objective optimization is rich [Deb, 2014]. Research on multi-stakeholder settings is, however, still limited, both from the perspective of algorithm design [Abdollahpouri et al., 2020] and from the perspective of how to properly evaluate such recommender algorithms considering the multiple perspectives [Bauer and Zangerle, 2019]. Furthermore, when more stakeholders are considered, additional questions regarding fairness and ethics may arise, which represents another important research gap.

2.2 Purpose and Value of Recommender Systems

Most published research in our field does not explicitly mention the intended purpose of the recommender system or algorithms it seeks to improve. The underlying, implicit and very reasonable assumption often is that more accurate algorithms lead to better item rankings, which ultimately make it easier for consumers to find what they are interested in. The implicit purpose and value of such an improved system mostly is that it makes it easier for consumers to "*find good items*," as it is termed in the seminal work by Herlocker et al. [2000]. However, as pointed out in the previous section, in reality it is not always clear what a good (or: relevant) item actually is. The relevance of an item, as mentioned, can depend on various factors, including the consumer's current goals, situational context, and the specific purpose of the recommender from the viewpoints of different stakeholders.

Our predominant research operationalization, which is based on optimizing accuracy measures on historical datasets, seems too narrow for being able to capture the value of a recommender system. From a platform provider's perspective, a recommender may actually serve a multitude of purposes and, correspondingly, create value for the various stakeholders in different ways. Jannach and Adomavicius [2016] therefore developed a purpose-oriented framework for the evaluation of recommender systems, where they considered the purpose and value both from the perspective of consumers and providers as shown in Figure 3.

In the following, we will give various examples of how a recommender system can create value and emphasize that for many of the value dimensions we still need to develop appropriate and standardized means for assessing them.

2.2.1 Consumer Value

The probably most researched and discussed consumer-related purpose of a recommender system is to "help users find objects that match their long-term preferences" [Jannach and Adomavicius, 2016]. In fact, most of the research that is operationalized as a *matrix completion* problem formulation can be considered as implicitly focusing on this purpose. There are indeed a number of cases where

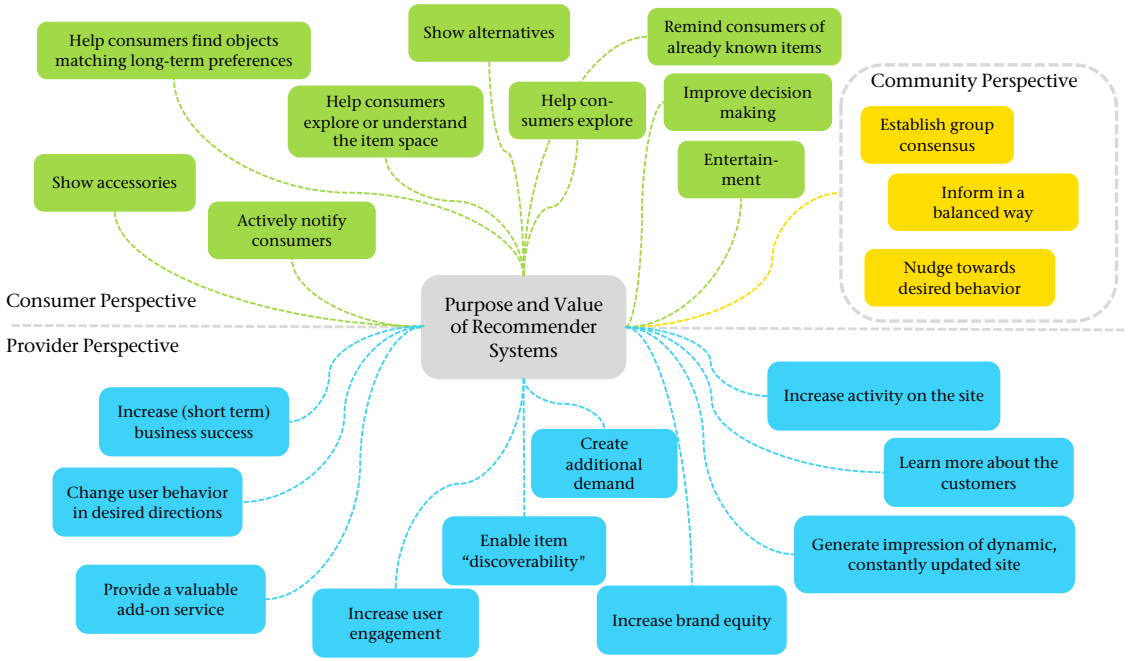

Figure 3: Purpose and Value of Recommender Systems, see also [Jannach and Adomavicius, 2016]

recommendations based solely on the long-term preferences are helpful. As an example, consider the landing pages of e-commerce sites or media services after users have logged in. In such situations, long-term preference models are particularly valuable, as no information about the consumer's current intent or contextual situation is yet available.

However, as shown in Figure 3, there are many other ways in which recommenders can create value for consumers and other stakeholders, and where finding items that are *generally* relevant for the consumer is not sufficient to create effective recommendations. In many of these cases, it is the consumer context and the current consumer's intention that matters. At the beginning of the 2010s, we have seen a number of papers being published on the topic of *context-aware* recommender systems [Adomavicius and Tuzhilin, 2015]. While the interest in taking explicit context information into account has flattened out since then, we observe that considering the interactional context in terms of the consumer's last activities in an ongoing session received more attention in recent years. However, also research in this area, called session-based recommendation [Quadrana et al., 2018], almost exclusively focuses on offline experiments and relies on abstract accuracy measures such as Precision or Recall and only very few user studies have been published.

Again, the main problem is that such accuracy measurements do not explicitly take into account in which ways the recommender aims to support the consumer. Consider Amazon's "Customers who bought . . . " recommendations. The helpfulness of a given set of recommendations in the context of a currently viewed item can largely depend on the consumer's decision phase. In an early decision phase, the best value of the recommender might result from *showing alternatives* to the presently viewed item. In later phases, however, the recommender might focus on a smaller set of rather similar options or even start to present accessories. As a result, depending on the decision phase, entirely different sets of items should be considered by the recommender. As another example, consider a music streaming site that automatically creates a playlist from a track selected by the consumer. Also in this case, it is important to understand the user's intentions—e.g., relax or being motivated during exercises, listen to familiar tracks or discover new things—to make purposeful recommendations.

Most of the above-discussed ways in which a recommender may create value for consumers are currently not investigated in much depth and thus represent important research gaps to be tackled. Furthermore, for several of them, it seems to be very difficult to assess the effectiveness of algorithms based on offline experiments, because they abstract too much from the given problem settings and entice us to use only those measures that are easy to take.

2.2.2 Organizational Value

The potential value for providers—recommendation service providers and item suppliers alike—is less investigated in the academic literature than consumer value. The underlying implicit assumption in academic research is often that more accurate relevance predictions, and thus more relevant recommendations, directly or indirectly lead to increased value for the organizational stakeholders. In other words, the assumption is again that accuracy measures, which are easy to take, are good proxies also for this side of the value perspective. Whether or not this is indeed the case for a given application is however unanswered.

In fact, the intended purpose and value of a recommender for service providers may be manifold, as shown in Figure 3. One prominent goal is to *increase (short-term) business success*, e.g., promoting certain items through recommendations. These could, for example, include items with higher profit or overstocked items. A recommendation system may also be used as a means to *change consumer behavior in desired directions*. In particular, a recommender can be helpful to point consumers to certain areas of the catalog (e.g., to increase sales of long-tail items or to help consumers to *discover* items they have not been aware of before), thereby stimulating *cross-sales* and *additional demand*. Another, more indirect effect of a recommender system is that it can help to increase consumers' *engagement* with the website or application, or generally increase the *activity on the site*. This, in turn, can lead to higher re-subscription rates or a higher rate of consumers upgrading from a free to a paid service.

Beyond the increase of sales or re-subscription numbers, a recommender system may also serve strategic purposes. Most importantly, good recommendations can be a *valuable add-on service* that attracts customers when competitors do not provide such a personalized service. Customers who use the service over longer periods of time might also be more *hesitant to switch* to an alternative provider once they receive valuable recommendations and *develop trust* when the system already knows their preferences as they perceive high switching costs.

Overall, academic literature considering the organization-oriented value of recommenders is scarce. One main reason lies in the fact that most research relies on offline evaluation and today's datasets that are used for such evaluations rarely contain business-related information. One of the few exceptions is, for instance, the work by Jannach and Adomavicius [2017] investigating profitability aspects of recommender systems in offline experiments using fictitious profit values. Research considering organizational value, thus, usually follows a research design that is typical for information systems research with consumers in the loop and typically using multiple types of measures to determine the potential effects of recommenders [e.g., Adomavicius et al., 2018].

2.2.3 Group, Community, and Societal Value

In our overview of recommendation purposes in Figure 3, we highlight cases where the recommender does not only have an impact on individual consumers but on entire groups or communities. The probably best-researched area in that context is the one of group recommendation. In this line of research, the recipient of the recommendation is not an individual, but a group of consumers. The particular problem is that the members of the group might have diverging preferences. The purpose and value of a group recommender system therefore is to support the group in making a joint decision.

In a number of mostly earlier technical approaches to group recommendations, one main goal was to find a good or the best strategy to aggregate the preferences of the group members. A simple technique is to compute relevance predictions for each item, e.g., for a movie to be watched together, and then to compute the average prediction. Other strategies are based on social choice theory [Arrow, 1951] and partly follow more sophisticated computation patterns. Given the complexity and social dynamics of group decisions, it soon became evident that offline experiments do not sufficiently inform us about the true value of a group recommender. This led to more informative setups involving, for example, simulated group decision experiments [Delic et al., 2017, Bauer and Ferwerda, 2020].

Beyond group decision settings, recommenders may also be used to influence entire user communities or the society, e.g., in the context of health, environment, or energy. Karlsen and Andersen [2019], for example, envision future systems that use *digital nudges* in a personalized way, i.e., in the form of a recommender, to entice desired user behavior. Beyond smaller groups, recommender systems might serve even larger communities or an entire society. They could, for example, be used to inform a society in a fair and balanced way on social media or news sites.

2.3 Risks of Recommender Systems

So far, we focused on the potential value of recommender systems. Recommenders may, however, also have undesired effects on different stakeholders. Most of our measurement approaches focus exclusively on the positive effects (e.g., by measuring the accuracy of the predictions), but very limited research exists on understanding or quantifying the negative effects. Figure 4 shows examples of potential risks of recommender systems.

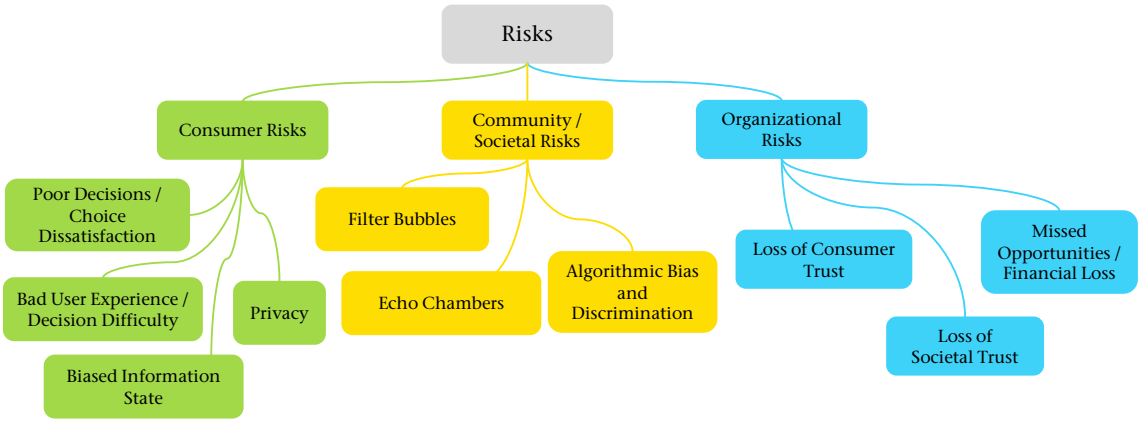

Figure 4: Selected Risks of Recommender Systems

From the perspective of the *end consumer*, one main effect of a poorly working or even malfunctioning recommender system could be that consumers—as a result of being incited by the system—make poor decisions, ultimately leading to low satisfaction with their choices. Such poor choices might have direct financial consequences or just lead to a waste of time. A poorly designed recommender system may furthermore have an effect on the user experience in different ways. If, for example, major parts of a platform are based on personalized recommendations, as in the case of many media streaming sites, consumers might have difficulties finding what they want or need because they keep stuck in their information bubble. Furthermore, the specific selection of recommended items within a recommendation list may increase the choice difficulty for consumers. This can, for instance, happen in case of too many, too few, or excessively similar items in a recommendation list. Finally, the selection of the recommended items may leave consumers in a state where they only have a limited perspective on the information state, e.g., the space of options on an e-commerce site or the spectrum of opinions on social media.

This latter aspect of being in a biased information state can easily expand to *community-related risks* of recommender systems. Filter bubbles and echo chambers [Pariser, 2012] are probably the best known and usually undesired effects that recommendation and information filtering may incur. Such phenomena may emerge when a recommender system, e.g., on a social media site, has a tendency (or: bias) to mostly present information that is in line with a user's existing beliefs or past preferences. As a result, a recommender may thereby, for example, reinforce political views, potentially leading to a stronger polarization or extremism in a society. In that context, algorithms that tend to recommend trending or popular items furthermore run the risk of being misused for the over-proportional dissemination of certain content, for example fake news.

Generally, algorithmic biases can be part of the intentional design of a system (e.g., to recommend popular or profitable items) or implicitly emerge when the system is learning from data that is unevenly distributed or biased in the first place. In either case, the recommendations from such a system may be considered unfair, e.g., because they favor the majority, or discriminate certain parts of the society, see Ledford [2019] for a recent case of algorithm-induced discrimination.

Finally, there are also a number of risks from the *provider* perspective. An ineffective recommender system can, first of all, be seen as a missed opportunity to, for example, generate more sales on an e-commerce site or keep customers engaged on a media site. Although this seems a modest risk at first glance, this situation can easily lead to a competitive disadvantage on the market, in case that competitors are able to create consumer or organizational value through their recommender systems. A poorly working or malfunctioning recommender can lead to a loss of trust by the consumers, which may happen at least at two different levels. First, consumers might cease to consider the recommendations in their decision processes, which makes the recommender less effective. Second,

consumers might even lose trust in the recommendation provider, in particular when they feel that the recommendations are unfair or biased. This can further lead to a loss of societal trust towards the provider, in case that such practices are questioned and become public.

3 Rethinking our Research Approach

Given today's predominant research approach—improving accuracy metrics in offline experiments—the real-world impact of most of our research output might be much more limited than we think. Even though there is no doubt that academic research in recommender systems has led to algorithmic innovations that have been very successfully picked up by industry, most of the important value perspectives discussed above cannot be investigated with our most common research approach at all. Interestingly, when discussing with peers at conferences, we often find that we are all well aware of the mentioned limitations. When discussing the reasons why we continue to rely on this very limiting research approach, typical statements include "we would need a real system to evaluate this" or "user studies are difficult."

Given these difficulties, it seems we often prefer to measure what can be easily measured, despite the unclear value and the sometimes limited insights that we can obtain from such measurements. An underlying additional problem certainly is that research based on offline experiments can sometimes be easier to publish. In contrast to cases where a novel research design has to be developed and defended against reviewers, papers using standard offline evaluation procedures are usually much less questioned regarding methodological aspects. Generally, this leaves us in a very unsatisfactory situation. There is huge academic interest in the field of recommender systems, with a huge number of papers published each year. Still, the impact of this research is often unclear. At the same time, there are many interesting and relevant questions in this area, which are often only addressed by a small number of research groups.

We, as a community, should therefore re-think how we do research and—in the spirit of the work by Wagstaff [2012]—also focus more on problems "that matter." In the following sections, we will first elaborate on the importance of keeping the goals and the purpose of a recommender in mind when evaluating it, and then review viable ways of measuring the effectiveness of recommenders in a more impact-oriented way.

3.1 Choosing Evaluation Designs with Goal and Purpose in Mind

Any meaningful evaluation of the effectiveness of a recommender system or algorithm requires us to have a clear idea about its intended goal, purpose, and value. Without a precise understanding of these aspects, it is impossible to decide on the research approach and in particular on the metrics that we should use in the evaluation.

In the research literature, "helping the consumer find relevant items" is, as detailed above, the most common (implicitly assumed) system purpose. The corresponding performance metrics are based on prediction accuracy, e.g., Precision, Recall or Root Mean Squared Error (RMSE). While this combination of purpose and metric is certainly plausible, it relies on the assumption that higher prediction accuracy consistently leads to better recommendations. In reality, however, we do not know if higher accuracy leads to recommendations that are more useful for consumers, if consumers will find more interesting things, or if they will be persuaded by the recommendations to make more purchases.

Ultimately, it is important—both in academic and industrial settings—that we ensure to use a combination of research design and evaluation measures that are suitable to validate our claims and goals. Jannach and Adomavicius [2016] proposed a layered conceptual framework as a guidance to align: *(i)* the overarching (organizational) goals, *(ii)* the specific purposes of the recommender system in this context, *(iii)* the corresponding computational tasks, and *(iv)* the evaluation approach.

Let us consider the example of a music streaming service as illustrated in Table 2, where the overarching goal for using a recommender is to ensure long-term profitability of the whole service through a high rate of renewed subscriptions. The specific purpose of the recommender given such a goal could be to increase user engagement with the service. At the computational level, high engagement can probably be achieved by balancing two factors. First, it is important to estimate, with high accuracy, whether a consumer will like a certain recommendation. Second, the system should also help the consumer discover something new (e.g., a new artist) from time to time to the extent that the respective consumer enjoys discovery. All these considerations then determine how

Framework Layer	Specific Example
Overarching Goal	Ensure long-term profitability of the service
↓	↓
Purpose of the Recommender	Increase user engagement
↓	↓
Computational Task	Recommend mix of familiar and novel items assumed to be liked by the consumer
↓	↓
Evaluation Approach	*Offline*: Accuracy, Novelty *User Study*: Adoption, Intention-to-Reuse *Online*: Streaming Activity, Session Lengths

Table 2: Example of Using the Framework by Jannach and Adomavicius [2016].

we should measure. In the described case, accuracy measures can embody one of the components to assess the system's effectiveness in the computational task. Because discovery of novel items is an integral part of the computational task, additional measurements, e.g., regarding novelty, are required at this level as well. However, these measures are not able to inform us about the effectiveness of a system in terms of user engagement, and even less about re-subscription rates. Therefore, additional measurements are required. With the help of user studies, one could assess how many of the recommendations are adopted by the participants, if they found the recommendations helpful and if they would use a similar system again in the future. At the topmost level, measuring the effects of a recommender on re-subscription rates can be difficult as well in practice [Gomez-Uribe and Hunt, 2015], and one can for example resort to measure the activity on the site when a new system is A/B tested. As an alternative, one can assess the participants willingness-to-pay in a user study; for an overview of measurement methods see, e.g., Breidert et al. [2006].

Generally, a framework like the proposed one may be a helpful guide both in industry and academia. For industry, the framework is designed as an aid to establish a clear vision and shared understanding of the intended goals of the recommendation service among the involved organizational units, from the executive level to product managers to data scientists and engineers. It furthermore helps to choose or design suitable operational measurements that can then be aggregated or mapped to organization-oriented KPIs. Finally, it can entice us to think more about specific purposes of a recommender in a given application domain. This, in turn, might point us to a need for novel experimental designs and metrics especially for cases where the intended value, e.g., user satisfaction, cannot be assessed with our predominant research instruments and measurement methods.

3.2 What to Measure – Focusing on Relevant Questions

Next, we review possible ways of measuring the effectiveness of recommenders, emphasizing the variety of possible measures for assessing recommenders in a more impact-oriented way. We structure our review by *organization-oriented* and *consumer-oriented* measures.

Organization-oriented Measures Generally, the choice of the performance measures and KPIs in practice does not only depend on the intended purpose of the recommender, but also on the specific operational model of the organization. In a recent literature survey on articles that report on real-world deployments of recommender systems, Jannach and Jugovac [2019] identified the following five types of measurements that are commonly used in A/B tests (Figure 5):

1. *Click-Through Rate (CTR)*: CTR measures how many clicks a recommendation garners. This metric is frequently used in the context of news recommendation. However, optimizing for CTR can be misleading because—except for certain business models, e.g., ones based on ad impressions—a higher CTR does usually not translate to increased organizational value in the long run. Short-term increases in CTR can, for example, be achieved by recommending generally popular items, through click-bait headlines, or better positioning of the recommendations [Garcin et al., 2014].

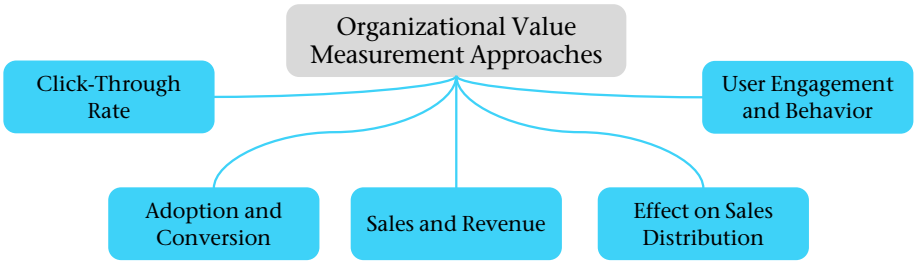

Figure 5: Measuring the Organizational Value of Recommenders [Jannach and Jugovac, 2019]

2. *Adoption and Conversion Measures*: This type of measurements goes beyond simply recording clicks. Media streaming companies such as YouTube or Netflix, for example, only consider a recommendation a success, if a certain fraction of the video was watched. Similarly, many providers compute conversion rates that, for example, measure if a recommendation resulted in a purchase. Depending on the domain, various types of conversion rates are feasible, e.g., to determine how many job recommendations led to a contact between a job seeker and an employer on a business network.

3. *Sales and Revenue*: These are the most direct measurements that can be determined in field tests, i.e., if a recommender led to improved KPIs, e.g., by promoting certain items, through cross-sales effects, or by stimulating consumers to explore additional areas of the catalog.

4. *Effect on Sales Distribution*: In some cases, providers are interested in understanding or influencing *what* their customers purchase or consume. A typical goal could be to use a recommender to point consumers to the long-tail of the item space, assuming that such item suggestions lead to discovery effects and longer-term organizational value, both in terms of sales and user engagement.

5. *User Engagement*: User engagement is a frequently used proxy of organizational value, in particular for providers that offer flat-rate subscription models as most media streaming services do. User engagement is commonly measured through interaction-based metrics such as the time spent on the platform, the number of visits, or the length of the interaction sessions.

The literature review in Jannach and Jugovac [2019] has shown that recommender systems can be effective for any of these value dimensions, leaving no doubt about the broad success of recommenders in practice. The reported gains in terms of the different metrics however varied across domains and application scenarios, e.g., from around 1% to over 500% in increased sales. The main reasons for these differences probably are the baselines that were used for the comparison. Sometimes, an existing recommender was fine-tuned; in other cases, there was no previous recommendation functionality at all. Interestingly, in almost all of the investigated real-world cases in [Jannach and Jugovac, 2019], comparisons were made between algorithms that were quite different in nature, e.g., a complex method is compared against a popularity-based baseline. This stands in strong contrast to what is typically measured in the academic evaluations, where research is sometimes based on making subtle changes to a complex algorithm, e.g., by using a different loss function when optimizing.

Unfortunately, when results of field tests are reported, this part is often comparably shallow, where only a few paragraphs or a subsection within a longer technical paper provide information on the field test. Often, limited information is provided about the baseline system. Sometimes not even the KPI to be optimized is revealed and statistical significance results are almost never provided. Nonetheless, these reports from real-world settings are helpful for us as academic researchers to understand in which ways recommenders create value in practice and how this value is measured.

Consumer-oriented Measures Commonly used accuracy metrics such as RMSE or Precision and Recall help assessing how good an algorithm is at predicting whether or not a consumer will like or consume an item. Given the known limitations of the aforementioned metrics in terms of assessing the utility of the resulting recommendations for consumers, researchers have developed a number of additional consumer-oriented *offline* metrics. These metrics are designed to characterize other potentially desired quality factors of recommendations for consumers, the most prominent ones

being diversity, novelty, and serendipity [Gunawardana and Shani, 2015]. Various alternative ways of computing these metrics were proposed, and considerable research was devoted to algorithms that aim at balancing these often competing quality factors; see also [Kaminskas and Bridge, 2016] for a recent overview on such "beyond-accuracy" metrics.

Such metrics can be very useful to analyze certain characteristics of different algorithms, e.g., to check whether they have a tendency to recommend niche items or rather popular items, which can be important from a provider's perspective as well. However, these purely computational metrics are not able to tell us about the consumers' perception of the recommendations [Ekstrand et al., 2014]. In fact, for most of the proposed novelty and diversity measures in the literature, there is little or no evidence that the computational approaches correlate with consumer perceptions. Nonetheless, we use these measures and to some extent probably do so simply because these measurements are easy to make.

A more promising approach is to rely on controlled user studies when it comes to make consumer-oriented assessments of the usefulness and value of a recommender system. In the recommender systems research literature, user studies are not uncommon, but far less frequent than pure offline experiments. Most often, user studies are used when the goal is to explicitly investigate aspects of the human-computer interaction, user experience, human decision-making, or consumer behavior. Only in a few cases, effects of using different algorithms on user perceptions are investigated [e.g., Ekstrand et al., 2014, Kamehkhosh and Jannach, 2017].

Today, two comprehensive frameworks for the user-centric evaluation of recommendations, proposed by Pu et al. [2011] and Knijnenburg et al. [2012] and partly inspired by the Technology Acceptance Model [Benbasat and Barki, 2007], are commonly used. These frameworks define sets of general quality factors for recommender systems, outline possible relationships between the factors, and propose indicators to assess the effectiveness of a recommender. The data for the statistical analyses can be collected both by observing and recording the actions and decisions of the study participants or with the help of questionnaires. The concrete questions being asked and participant actions being recorded depend on the specific research questions and the underlying hypotheses.

Figure 6 shows the measurement constructs (variables) considered in the ResQue framework by Pu et al. [2011]. Depending on the research question, only a subset of the constructs, for which also corresponding questionnaire items are proposed, might be relevant. The research hypotheses correspond to suspected relationships between the constructs and paths in the model.

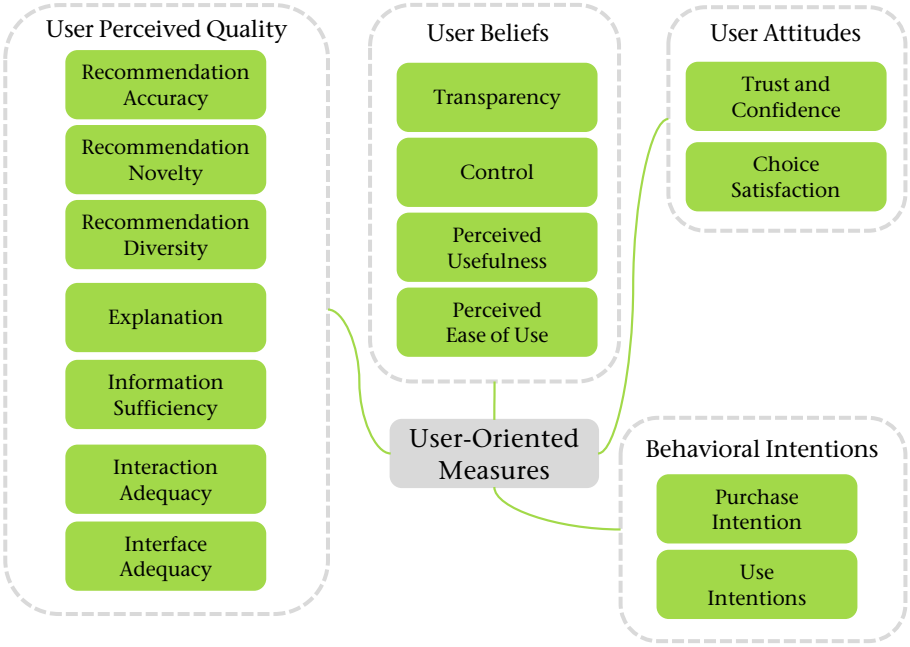

Figure 6: Variables in the ResQue Framework, adapted from Pu et al. [2011]

Given specific research questions, additional variables might be relevant. Some studies, for example, use "effectiveness" or "fun" as subjective variables, others consider "willingness to pay" or "willingness to give feedback" as ultimate outcome variables.

The framework by Knijnenburg et al. [2012] covers many of the facets identified by Pu et al. [2011] as well. The organization into groups is, however, slightly different. Furthermore, Knijnenburg et al. [2012] include a number of additional independent variables in their framework, including user characteristics or the contextual situation, and consider the possible two-way interactions between subjective variables and logged behavioral data.

Generally, both frameworks acknowledge that many factors other than accuracy can have an impact on the effectiveness of a recommender system, i.e., that users trust the system and adopt their recommendations [Xiao and Benbasat, 2007]. Furthermore, different to most algorithmic research, user studies are usually based on theoretical considerations or at least an explicit research hypothesis, e.g., that a certain variation of the user interface will make the system easier to use and, as a result, that users intend to rely on the recommendation feature more often in the future.

Still, also controlled user studies have some limitations. The typical reservations include that no real system was used, that the situation for the participants is artificial, or that the participants are not representative for the general population of such a system. Academic research in various scientific fields that rely on controlled user experiments has, however, developed a number of established research practices and sophisticated statistical analyses that aim to minimize or at least quantify some of these risks.

3.3 How to Evaluate – Ways Forward

In this section, we outline the ways in which we—as a community—should broaden our research methodology in order to obtain more impactful insights from our research in the future.

Improved Offline Evaluations Our current offline evaluation procedures have several known limitations, most importantly that they are often not able to inform us whether a new algorithm leads to better recommendations "in the wild." Although this is a major limitation, it does not entail that we should entirely give up offline experiments. Yet, we could reconsider which kinds of experiments and analyses we can reliably do without having the consumer in the loop. And we need to be considerate in the claims we make from these experiments and analyses.

Multi-Faceted Evaluation with Validated Metrics: One step forward is to consider more and more informative metrics. Currently, higher accuracy is still the "holy grail". The increased use of additional measures that consider the diversity, novelty, or serendipity of the recommendations is a positive development. Future research would benefit significantly if such additional metrics would be considered even more often and if more standardized reporting schemes would be established that consider those metrics. However, whenever claims about improved recommendations are made based on such alternative metrics, it is necessary that the used metrics are validated for the given domain. This would probably require to execute a controlled user study that, for example, shows that the chosen diversity metric is correlated with the diversity level perceived by consumers and that higher diversity improves the user experience.

More Analytical Research: In general, both academia and industry would benefit if a more analytical—in contrast to a predictive—approach would be adopted more often. With offline experiments, we can, in fact, analyze a multitude of general characteristics of different algorithms. We may, for example, assess whether an algorithm has a higher tendency to recommend more popular items than another one. We may also investigate various forms of coverage, or if an algorithm has a bias to recommend almost the same set of items to everyone. In that context, we may furthermore analyze how stable an algorithm is in the various metrics when hyper-parameters are slightly changed.

Overall, one crucial point is that in such analyses often no "winner" exists. It may, for example, depend on the application domain and the business model whether the recommendation of mostly popular items is desirable or not. Lee and Hosanagar [2019], for example, found in a field test that a given collaborative filtering algorithm led to increased sales for long-tail items. But the additional profit that was created through the recommendation of already popular items was even higher. Generally, such analytical research approaches would allow to derive more actionable insights that help researchers in academia and industry making better-informed decisions about the selection of the approach or algorithm for their particular problem.

Investigating Long-Term and Indirect Effects Using Simulation: Current research mostly focuses on the short-term, direct effects of recommender systems. This is the case for both, research based on offline experiments and user studies. Recommender systems do, however, also have long-term and indirect effects. For instance, Dias et al. [2008] found that a recommender on an e-commerce site might not

necessarily lead to an increase of sales of the recommended items, but to a generally higher purchase volume as consumers discover new item categories in the shop. Similarly, Kamehkhosh et al. [2019] observed inspirational effects of a recommender in a music application.

In other research fields, the investigation of longer-term and emergent effects is often performed using agent-based simulation approaches, see e.g., Wall [2016] for an overview in the domain of managerial sciences. In recommender systems literature, research on longitudinal effects is scarce. An exception is the recent work by Zhang et al. [2020], who identified a longitudinal *performance paradox* of recommender systems, where an increased reliance of consumers on the system's recommendations seems to make the system less useful in the longer run. While such simulations are usually based on several abstractions and simplifications, they allow us to investigate a multitude of configurations at low cost that would not be feasible to analyze in field tests.

Multi-Stakeholder Evaluation: Despite the fact that a recommender system involves various potential stakeholders, researchers usually focus on the consumer value [Bauer and Zangerle, 2019]. Only in recent years, multi-stakeholder recommendation problems received considerable research interest [e.g., Said et al., 2012, Abdollahpouri et al., 2020]. While a few works [e.g., Azaria et al., 2013] exist that focus on price- and profit-ware recommendation approaches, research in this area is still scattered, see Jannach and Adomavicius [2017] for an overview. Following the above discussion, simulation approaches that are common in other domains may be employed to analyze possible longer-term effects of different provider strategies. Alternatively, the problem of balancing the different interests of the stakeholders can be modeled, including side constraints such as consumers' budgets, as a mathematical optimization problem [Wang and Wu, 2009].

Multi-Modal Evaluation Today's often narrow research approach based on offline experimentation calls for a much richer methodological repertoire than we use today. In many cases, it might be advantageous or even required to combine multiple methods and approaches to evaluate a recommender solution. Such multi-modal evaluations should give us a much more comprehensive picture than, e.g., an isolated analysis of prediction accuracy.

Only a small number of offline–online comparisons of algorithms have been published. In many cases, these comparisons led to very informative and partially unexpected results. Garcin et al. [2014], for example, analyzed the performance of different news recommendation strategies both on an online portal and through offline experiments. They found that recommending the most popular items was the best strategy in an offline setting, whereas a more adaptive method was much better in a real-world environment. Other studies [e.g., Cremonesi et al., 2012, Beel and Langer, 2015, Rossetti et al., 2016] found that algorithms with higher offline accuracy do not necessarily lead to recommendations that are perceived being of higher quality in user studies.

Besides offline–online contrasts with respect to accuracy, combining computational experiments and user studies allows to investigate other quality factors of recommendations, e.g., how consumers perceive the novelty or diversity of recommendations of different algorithms [Ekstrand et al., 2014]. Furthermore, it may turn out that even algorithms with extremely low offline accuracy can lead to a satisfying user experience, e.g., when a music recommendation service is very strong in helping consumers discover new items [Ludewig and Jannach, 2019]. Ultimately, there are various additional ways in which such comparison studies may be helpful. They can, for example, help validate that an employed method is truly effective, e.g., when proposing a diversification algorithm [Ziegler et al., 2005]. Furthermore, a comparison of offline experiments, user study, and field test, such as done in Jannach et al. [2016], might reveal that the assumptions made for the offline simulation protocol are not realistic. Ultimately, such comparison studies can build the basis for designing novel and better metrics to predict the "online success" from offline experiments [Maksai et al., 2015].

Generally, it is not only important to adopt more comprehensive evaluation approaches, but also to consider alternative ways of conducting research. There are various ways in which *qualitative* research approaches could help understand and characterize certain phenomena or explore new research directions. Possibly helpful methods include interviews, focus groups, case studies, and various other types of observational and phenomenological research methods. At the same time, our research could also be more often guided by *theory*. Most algorithmic research on recommender systems, for example, comes without *hypothesis development*, which is, in contrast, very common in fields like information systems. We often simply assume that, e.g., "higher diversity is better," but do not provide any pointers to underlying theory, e.g., from psychology, that supports such an assumption. Frequently, we do not consider the specific application domain either. As a result, because our computational measures are also not validated, we might end up with sophisticated technical approaches that optimize the wrong measures and goals. Simulation studies, as mentioned

above, may therefore be an interesting middle-ground between qualitative and theory-guided research, which we believe has not reached its full potential for our research field yet.

About Domain-Specifics and General Models Our discussion of the potential value of recommenders showed that whether a recommendation is good or not depends on the particular domain or even application. In the news domain, for example, it is important to take the recency of the items into account. In the music domain, recommenders are often considered useful when they support discovery. In other domains, like tourism, the geographical vicinity might be very relevant. In e-commerce, finally, profitability considerations may play a role for the provider as well. To be useful and effective in the real-world, recommender systems have to take such domain-specifics into account. It is very pleasing to see that our field has developed a variety of techniques that consider such particularities.

Clearly, however, as academic researchers we are interested in generalizable solutions, i.e., we are typically not interested in designing algorithms that work well for only one particular scenario, e.g., the recommendation of a specific type of fashion products. As a result, our community has put forward and consistently improved domain-agnostic algorithms, including collaborative filtering methods based on nearest neighbors or matrix factorization techniques, which are nowadays widely used in industry.

Given the broad adoption of such methods, it is very attractive for researchers to try to improve such general-purpose methods, as being successful at this task promises high impact. However, such improvements are then often only demonstrated for a very specific experimental configuration of datasets (domains), evaluation measures, and baselines, which does not inform about their generalizability. We therefore argue that researchers should more often focus on domain- and application-specific aspects and aim to develop novel solutions for certain types of problem settings. Given the insights from these specific problems, we can then more reliably build solutions that generalize beyond a given domain. For this, it is however important to acknowledge that there is no "best model," which is an assumption that has led us to our current leaderboard-chasing culture in different subfields of recommender systems research.

4 Summary

The success of recommender systems in practice has led to a tremendous academic interest in this area and recommender systems—which may be considered one of the most visible applications of machine learning and artificial intelligence—have become their own research field. However, it seems that in this research community, we have fallen prey of a McNamara fallacy to a worrying extent: We have developed a research culture that overly relies on quantitative measures in offline experimentation and particularly on measures that are easy to take. As a result—despite the huge number of papers that are published on recommender systems every year—, it remains unclear how much impact our research actually has in practice.

In this work, we call for a paradigm shift with the hope that our work raises awareness in our community that many of our research efforts might lead to a dead end, as long as we do not focus on the relevant questions that matter in the real world, and refine and broaden our research approach and instruments accordingly.

References

Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, 2020. doi: 10.1007/s11257-019-09256-1.

Gediminas Adomavicius and Alexander Tuzhilin. Context-Aware Recommender Systems. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 191–226. Springer, Boston, MA, USA, 2nd edition, 2015. doi: 10.1007/978-1-4899-7637-6_6.

Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley, and Jingjing Zhang. Effects of online recommendations on consumers' willingness to pay. *Information Systems Research*, 29(1):84–102, 2018. doi: 10.1287/isre.2017.0703.

Kenneth J. Arrow. *Social Choice and Individual Values*. John Wiley, New York, NY, USA, 1951.

Amos Azaria, Avinatan Hassidim, Sarit Kraus, Adi Eshkol, Ofer Weintraub, and Irit Netanely. Movie recommender system for profit maximization. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 121–128, New York, NY, USA, October 2013. ACM. doi: 10.1145/2507157.2507162.

Christine Bauer and Bruce Ferwerda. Conformity behavior in group playlist creation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–10, 2020. doi: 10.1145/3334480.3382942.

Christine Bauer and Eva Zangerle. Leveraging multi-method evaluation for multi-stakeholder settings. In Oren Sar Shalom, Dietmar Jannach, and Ido Guy, editors, *Proceedings of the 1st Workshop on the Impact of Recommender Systems*, ImpactRS '19, 2019. arXiv:2001.04348.

Jöran Beel and Stefan Langer. A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In *Research and Advanced Technology for Digital Libraries: Proceedings of the International Conference on Theory and Practice of Digital Libraries*, TPDL '15, pages 153–168, Cham, Germany, 2015. Springer. doi: 10.1007/978-3-319-24592-8_12.

Izak Benbasat and Henri Barki. Quo vadis TAM? *Journal of the Association for Information Systems*, 8 (4), 2007. doi: 10.17705/1jais.00126.

Anand V Bodapati. Recommendation systems with purchase data. *Journal of Marketing Research*, 45 (1):77–93, February 2008. doi: 10.1509/jmkr.45.1.077.

Christoph Breidert, Michael Hahsler, and Thomas Reutterer. A review of methods for measuring willingness-to-pay. *Innovative Marketing*, 2(4):8–32, December 2006.

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 191–198, New York, NY, USA, 2016. ACM. doi: 10.1145/2959100.2959190.

Paolo Cremonesi, Franca Garzotto, and Roberto Turrin. Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study. *Transactions on Interactive Intelligent Systems*, 2(2):1–41, 2012. doi: 10.1145/2209310.2209314.

Kalyanmoy Deb. Multi-objective optimization. In Edmund K. Burke and Graham Kendall, editors, *Search Methodologies*, pages 403–449. Springer, Boston, MA, USA, 2014. doi: 10.1007/978-1-4614-6940-7_15.

Amra Delic, Julia Neidhardt, Laurens Rook, Hannes Werthner, and Markus Zanker. Researching individual satisfaction with group decisions in tourism: Experimental evidence. In Roland Schegg and Brigitte Stangl, editors, *Proceedings Information and Communication Technologies in Tourism 2017*, pages 73–85, Cham, Germany, 2017. Springer. doi: 10.1007/978-3-319-51168-9_6.

M. Benjamin Dias, Dominique Locher, Ming Li, Wael El-Deredy, and Paulo J.G. Lisboa. The value of personalised recommender systems to e-business: A case study. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, pages 291–294, New York, NY, USA, 2008. ACM. doi: 10.1145/1454008.1454054.

Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 2014 ACM Conference on Recommender Systems*, RecSys '14, pages 161–168, New York, NY, USA, 2014. ACM. doi: 10.1145/2645710.2645737.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, pages 101–109, New York, NY, USA, 2019. ACM. doi: 10.1145/3298689.3347058.

Florent Garcin, Boi Faltings, Olivier Donatsch, Ayar Alazzawi, Christophe Bruttin, and Amr Huber. Offline and online evaluation of news recommender systems at swissinfo.ch. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 169–176, New York, NY, USA, 2014. doi: 10.1145/2645710.2645745.

Carlos A. Gomez-Uribe and Neil Hunt. The Netflix recommender system: Algorithms, business value, and innovation. *Transactions on Management Information Systems*, 6(4), 2015. doi: 10.1145/2843948.

Asela Gunawardana and Guy Shani. Evaluating recommender systems. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 265–308. Springer, Boston, MA, USA, 2nd edition, 2015. doi: 10.1007/978-1-4899-7637-6_8.

Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, pages 241–250, New York, NY, USA, 2000. ACM. doi: 10.1145/358916.358995.

Dietmar Jannach and Gediminas Adomavicius. Recommendations with a purpose. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 7–10, New York, NY, USA, 2016. ACM. doi: 10.1145/2959100.2959186.

Dietmar Jannach and Gediminas Adomavicius. Price and profit awareness in recommender systems. In *Proceedings the 1st International Workshop on Value-Aware Multi-Stakeholder Recommendation Workshop at ACM RecSys 2017*, VAMS '17, August 2017. arXiv:1707.08029.

Dietmar Jannach and Michael Jugovac. Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems*, 10(4), December 2019. doi: 10.1145/3370082.

Dietmar Jannach, Michael Jugovac, and Lukas Lerche. Supporting the design of machine learning workflows with a recommendation system. *ACM Transactions on Interactive Intelligent Systems*, 6 (1), February 2016. doi: 10.1145/2852082.

Iman Kamehkhosh and Dietmar Jannach. User perception of next-track music recommendations. In *Proceedings of the 2017 Conference on User Modeling Adaptation and Personalization*, UMAP '17, pages 113–121, New York, NY, USA, 2017. ACM. doi: 10.1145/3079628.3079668.

Iman Kamehkhosh, Geoffray Bonnin, and Dietmar Jannach. Effects of recommendations on the playlist creation behavior of users. *User Modeling and User-Adapted Interaction*, 2019. doi: 10.1007/s11257-019-09237-4.

Marius Kaminskas and Derek Bridge. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems*, 7(1), December 2016. ISSN 2160-6455. doi: 10.1145/2926720.

Randi Karlsen and Anders Andersen. Recommendations with a nudge. *technologies*, 7(45), 2019. doi: 10.3390/technologies7020045.

Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22: 441–504, 2012. doi: 10.1007/s11257-011-9118-4.

Heidi Ledford. Millions of black people affected by racial bias in health-care algorithms. *Nature*, 574 (7780):608–609, October 2019. doi: 10.1038/d41586-019-03228-6.

Dokyun Lee and Kartik Hosanagar. How do recommender systems affect sales diversity? a cross-category investigation via randomized field experiment. *Information Systems Research*, 30(1): 239–259, 2019. doi: 10.1287/isre.2018.0800.

Jimmy Lin. The neural hype and comparisons against weak baselines. *SIGIR Forum*, 52(2):40–51, January 2019. doi: 10.1145/3308774.3308781.

Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003. doi: 10.1109/MIC.2003.1167344.

Malte Ludewig and Dietmar Jannach. User-centric evaluation of session-based recommendations for an automated radio station. In *Proceedings of the 2019 ACM Conference on Recommender Systems*, RecSys '19, pages 516–520, New York, NY, USA, September 2019. ACM. doi: 10.1145/3298689.3347046.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3), 2018. doi: 10.1371/journal.pone.0194889.

Andrii Maksai, Florent Garcin, and Boi Faltings. Predicting online performance of news recommender systems through richer evaluation metrics. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, pages 179–186, New York, NY, USA, 2015. ACM. doi: 10.1145/2792838.2800184.

Judith Masthoff. Group recommender systems: Aggregation, satisfaction and group attributes. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 743–776. Springer, Boston, MA, USA, 2nd edition, 2015. doi: 10.1007/978-1-4899-7637-6_22.

Sean M. McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. On the recommending of citations for research papers. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, CSCW '02, pages 116–125, New York, NY, USA, 2002. ACM. doi: 10.1145/587078.587096.

Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Books, London, 2012.

Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 157–164, New York, NY, USA, 2011. ACM. doi: 10.1145/2043932.2043962.

Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. Sequence-aware recommender systems. *ACM Computing Surveys*, 51(4), July 2018. doi: 10.1145/3190616.

Steffen Rendle, Li Zhang, and Yehuda Koren. On the difficulty of evaluating baselines: A study on recommender systems. 2019. arXiv:1905.01395.

Marco Rossetti, Fabio Stella, and Markus Zanker. Contrasting offline and online results when evaluating recommendation algorithms. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 31–34, New York, NY, USA, 2016. ACM. doi: 10.1145/2959100.2959176.

Alan Said, Domonkos Tikk, Klara Stumpf, Yue Shi, Martha Larson, and Paolo Cremonesi. Recommender systems evaluation: A 3D benchmark. In *Proceedings of the Workshop on Recommendation Utility Evaluation: Beyond RMSE at ACM RecSys '12*, RUE '12, pages 21–23, September 2012.

J. Ben Schafer, Joseph Konstan, and John Riedl. Recommender systems in e-commerce. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, EC '99, pages 158–166, New York, NY, USA, 1999. ACM. doi: 10.1145/336992.337035.

Kiri L. Wagstaff. Machine learning that matters. In *Proceedings of the Twenty-Ninth International Conference on Machine Learning*, ICML '12, pages 529–534, 2012.

Friederike Wall. Agent-based modeling in managerial science: an illustrative survey and study. *Review of Managerial Science*, 10(1):135–193, January 2016. doi: 10.1007/s11846-014-0139-3.

Hsiao-Fan Wang and Cheng-Ting Wu. A mathematical model for product selection strategies in a recommender system. *Expert Systems with Applications*, 36(3, Part 2):7299–7308, April 2009. doi: 10.1016/j.eswa.2008.09.006.

Bo Xiao and Izak Benbasat. E-commerce product recommendation agents: Use, characteristics, and impact. *MIS Quarterly*, 31(1):137–209, 2007.

Jingjing Zhang, Gediminas Adomavicius, Alok Gupta, and Wolfgang Ketter. Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. *Information Systems Research*, 31(1), 2020. doi: 10.1287/isre.2019.0876.

Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 22–32, New York, NY, USA, 2005. ACM. doi: 10.1145/1060745.1060754.