ROBERT SCHUMAN CENTRE FOR ADVANCED STUDIES

EUI Working Papers

RSCAS 2012/56
ROBERT SCHUMAN CENTRE FOR ADVANCED STUDIES
Global Governance Programme-29

PERSONALIZATION AND THE FUTURE OF NEWS

Matthew Hindman

EUROPEAN UNIVERSITY INSTITUTE, FLORENCE
ROBERT SCHUMAN CENTRE FOR ADVANCED STUDIES
GLOBAL GOVERNANCE PROGRAMME

Personalization and the Future of News

MATTHEW HINDMAN

EUI Working Paper **RSCAS** 2012/56

This text may be downloaded only for personal research purposes. Additional reproduction for other purposes, whether in hard copies or electronically, requires the consent of the author(s), editor(s). If cited or quoted, reference should be made to the full name of the author(s), editor(s), the title, the working paper, or other series, the year and the publisher.

ISSN 1028-3625

© 2012 Matthew Hindman

Printed in Italy, October 2012
European University Institute
Badia Fiesolana
I – 50014 San Domenico di Fiesole (FI)
Italy
www.eui.eu/RSCAS/Publications/
www.eui.eu
cadmus.eui.eu

Robert Schuman Centre for Advanced Studies

The Robert Schuman Centre for Advanced Studies (RSCAS), created in 1992 and directed by Stefano Bartolini since September 2006, aims to develop inter-disciplinary and comparative research and to promote work on the major issues facing the process of integration and European society.

The Centre is home to a large post-doctoral programme and hosts major research programmes and projects, and a range of working groups and *ad hoc* initiatives. The research agenda is organised around a set of core themes and is continuously evolving, reflecting the changing agenda of European integration and the expanding membership of the European Union.

Details of the research of the Centre can be found on:
http://www.eui.eu/RSCAS/Research/

Research publications take the form of Working Papers, Policy Papers, Distinguished Lectures and books. Most of these are also available on the RSCAS website:
http://www.eui.eu/RSCAS/Publications/

The EUI and the RSCAS are not responsible for the opinion expressed by the author(s).

The Global Governance Programme at the EUI

The Global Governance Programme (GGP) is research turned into action. It provides a European setting to conduct research at the highest level and promote synergies between the worlds of research and policy-making, to generate ideas and identify creative and innovative solutions to global challenges.

The GGP comprises three core dimensions: research, policy and training. Diverse global governance issues are investigated in *research* strands and projects coordinated by senior scholars, both from the EUI and from other internationally recognized top institutions. The *policy* dimension is developed throughout the programme, but is highlighted in the GGP High-Level Policy Seminars, which bring together policy-makers and academics at the highest level to discuss issues of current global importance.The Academy of Global Governance (AGG) is a unique executive *training* programme where theory and "real world" experience meet. Young executives, policy makers, diplomats, officials, private sector professionals and junior academics, have the opportunity to meet, share views and debate with leading academics, top-level officials, heads of international organisations and senior executives, on topical issues relating to governance.

For more information:

www.globalgovernanceprogramme.eu

Abstract

Over the past two decades, much scholarship has theorized about how highly personalized news media might change the public sphere. But even as algorithmic content filtering has become widespread, social science research has lagged in understanding how such systems work, and how they have altered competitive dynamics between media outlets. Drawing on recent research into recommender systems, this paper examines the Netflix prize, as well as collaborative filtering algorithms deployed by Google and Yahoo. Content recommendation systems strongly advantage the very largest websites over small news outlets, with profound implications for the online news landscape.

Keywords

Future of journalism, online public sphere, recommender systems, news personalization, online news, digital journalism.

The Daily Me in Prophecy and Practice

In his 1995 book *Being Digital*, Nicholas Negroponte described a world in which everyone had a virtual newspaper entirely tailored to his or her personal taste. Negroponte proposed creating an intelligent, computerized "interface agent" that would "read every newswire newspaper and catch every TV and radio broadcast on the planet, and then construct a personalized summary":

> It would mix headline news with "less important" stories relating to acquaintances, people you will see tomorrow, and places you are about to go or have just come from. It would report on companies you know. In fact, under these conditions, you might be willing to pay the *Boston Globe* a lot more for ten pages than a hundred pages, if you could be confident that it was delivering you the right subset of information. You would consume every bit (so to speak). Call it *The Daily Me*. (153)

Negroponte's proposal was not wholly new, with antecedents (including some of Negroponte's own work) dating at least back to the 1970s. But Negroponte's vision of the Daily Me proved highly influential, partly because it arrived just as the Web was starting to transform the media landscape. The notion was endorsed by key technology industry leaders and top public policymakers (e.g. Gates 2000; Kennard 1999). Much subsequent scholarship focused on media self-selection as functionally equivalent to the Daily Me, with particular worry that the Internet would allow a partisan "echo chamber" (Sunstein 2001, 2009).

In recent years, improved filtering technologies and the emergence of social networking sites have produced something strikingly close to Negroponte's original vision. Google, Yahoo, Facebook, and Microsoft—the four firms that together receive one-third of Web visits, according to Experian Hitwise traffic data—all rely heavily on adaptive learning algorithms to match individuals with content they are likely to click on. Recommendation systems have long been a central part of online sellers such as eBay and Amazon.com (Schafer, Konstan, and Riedl 2001). And to the chagrin of some news editors and journalists, recommendation algorithms have become a central feature of online news outlets such as Yahoo news or CNN.com or Google News. Sites like Facebook have similarly endorsed such hyperpersonalization, with Facebook founder and CEO Mark Zuckerberg stating that "a squirrel dying in your front yard may be more relevant to your interests right now than people dying in Africa" (quoted in Pariser 2011). With the rise of the iPad and its imitators, Negroponte's idea that all of this personalized content would be sent to a thin, lightweight, "magical" tablet device has been partially realized too.

Recent scholarship such as Siva Vaidhyanathan's *The Googlization of Everything* (2011) and Joe Turow's *The Daily You* has viewed the trend toward personalized content and ubiquitious filtering as a part of a worrying concentration of corporate power. Eli Pariser's bestselling book *The Filter Bubble* voices similar worries. But for journalism scholarship as a whole, as Barbie Zelizer (2009) notes, there has been surprisingly little work to date on recommender systems. To the extent that algorithmic news filtering has been discussed at all, it has been unhelpfully lumped with a grab bag of different site features under the heading of "interactivity" (Bucy 2004; Deuze 2003; but see Stromer-Galley 2004). Recent research by Neil Thurman and Steve Schifferes has provided a taxonomy of different forms of personalization and chronicled their (mostly growing) deployment across different news sites (Thurman and Schifferes 2012; Thurman 2011). Even Thurman and Schifferes' work, however, says little about recommender systems because traditional news organizations have lagged in deploying them. Much work remains to be done.

This paper aims to advance scholarly understanding in two ways. First, it offers a more detailed examination of the inner workings of these recommendation algorithms than previous journalism scholarship. The mathematical and computer science literature on recommender systems has advanced substantially in recent years, but little of this new understanding has so far filtered into research on Web traffic, online news, or the future of journalism. In these realms, much of the writing on

recommender systems has been an assemblage of hypotheticals and what-ifs. Elaborate deductive conclusions have been built from false foundational assumptions.

Second, this paper examines the comparative impact of these technologies across news organizations, something that previous work has overlooked or misunderstood. Scholarship to date, where it exists, has focused on the impact of these technologies for an individual Web user or an adopting news organization. But there has been little exploration of the wholesale effects of these changes not only *within* a news organizations, but with regard to *competition between them*.

This paper begins with a detailed look at the Netflix prize, the first large-scale, open-submission machine learning contest. The Netflix Prize helped advance the state of the art significantly for recommender systems, and it led to several surprising insights into the principles behind them. Next, the paper examines how those principles apply to areas that are critical to the future of news. Two case studies are examined in more detail: Google News, the algorithmic news pioneer, and Yahoo!, which has recently revealed more about how its behavioral targeting ad technology works. Taken together, these cases tell us much about who will win, and who will lose, as recommender systems assume a growing role in the delivery of news.

Netflix and Content Recommendation

In October of 2006, movie-rental service Netflix kicked off the Netflix Prize, a worldwide competition to improve its video recommendation algorithm. The typical Netflix user signs up wanting to see a short list of movies, which she watches within a few months. Whether the subscriber stays or leaves thus depends on her ability to find new movies she wants to watch. Netflix has stated that three-quarters of the movies its viewers watch come directly from its recommendation system (Mayer-Schoenberger and Cukier 2013).

Netflix offered a $1 million prize to the first team that could beat CineMatch, its in-house recommendation engine. Even more remarkably, Netflix actually released its data. Once the contest started, anyone could download a real-world data set containing 100,480,500 one-to-five star ratings, from 480,189 anonymous users, for 17,770 movies.

The contest would end up running for more than two and a half years, engaging the efforts of more than 5,000 teams. In the process, it illuminated much that is usually hidden about the ways in which Web sites personalize the content that users see.

The central task of the contest was an example of *collaborative filtering*, using automated methods to infer a user's tastes from the preferences lots of other users. The key contest metric was root mean-squared error (RMSE)—a measure of how much, on average, a recommendation model misses the true value. If Joe Netflix Subscriber gives *The Empire Strikes Back* five stars, and the original CineMatch algorithm predicted that he would give it 4.5 stars, then the root squared error would be $\sqrt{(-.5 * -.5 = .5}$ (Note that squaring the errors, and then taking the square root, means that the RMSE is always positive.)

The contest hoped to drop the error as low as possible. Predicting that every user would give a movie its average rating produced an RMSE of 1.054—a typical error of more than a star in either direction. Netflix's CineMatch offered an initial RMSE of .9525, about a tenth of a star better. NetFlix described CineMatch as a relatively simple approach, "straightforward statistical linear models with a lot of data conditioning" (Netflix 2007). The contest winner, if any, would be the first time to drop the RMSE to .8563. Though this would still leave the model off by more than four-fifths of a star, it was nonetheless about twice the improvement that CineMatch had managed on its own.

The contest showed rapid progress out of the gate. Within a week, several teams had equalled CineMatch; within three weeks CineMatch had been bested by 3 percent. These efforts revealed that CineMatch was likely a variant of a K-nearest neighbor (KNN) algorithm. If we wanted to predict

Maria's rating for *Titanic*, for example, a KNN approach might start by finding the users who (1) saw *Titanic* and (2) agree with Maria's ratings of other movies—those who also hated *Gladiator* but gave five loving stars to *A Beautiful Mind*. Once this "neighborhood" of similar subscribers is found, Maria's predicted rating for *Titanic* is just a weighted average of her neighbors' ratings. If Alex, Becky, and Chris are the users most similar to Maria, and they gave *Titanic* 1, 4, and 5 stars, then Maria's predicted rating is just $\frac{1+4+5}{3} = 3.33$. KNN approaches dominated the early months of the contest.

The Netflix Prize attracted a wide range of participants from both industry and the academy, and even some members of the general public. Prominent entrants included machine learning faculty from the University of Toronto (the team ML@UToronto), and the Hungarian computer scientist and data mining expert Gábor Takács, who led the Gravity team. Dinosaur Planet, a team composed of Princeton undergraduates, also quickly rose into the top ranks of the public leader board. By late November, a team from AT&T Research Labs had also joined the competition. The team's key members were Yehuda Koren, a computer scientist and network visualization specialist, and Robert Bell, a statistician with a focus on machine learning. They were spurred on by their colleague Chris Volinsky, a statistician and the director of AT&T Labs Statistics Research Department. Bell and Koren called their team BellKor, and the duo would ultimately form the nucleus of the winning team.

One goal of the open competition was to attract and aggregate insights from a far broader and diverse group than otherwise possible. As Netflix had hoped, one of the largest single improvements came from an unlikely source. In early December 2006, participants were surprised to see the name Simon Funk jump to third place on the leaderboard. Simon Funk was the pseuodnym and pen name of Brandynn Webb, a computer scientist who had done previous professional work on artificial intelligence and pattern recognition.

While many teams were highly secretive about their methods—and even their membership—Funk explained his entire approach in a highly detailed blog post. Funk had applied a factor analysis technique called singular value decomposition (SVD) to the Netflix data. SVD is a type of latent factor model, in which many observed variables—i.e. the millions of movie ratings—are modeled as the sum of a (smaller) number of unknown variables. As Funk explained on his blog,

> The end result of SVD is essentially a list of inferred categories, sorted by relevance. Each category in turn is expressed simply by how well each user and movie belong (or anti-belong) to the category. So, for instance, a category might represent action movies, with movies with a lot of action at the top, and slow movies at the bottom, and correspondingly users who like action movies at the top, and those who prefer slow movies at the bottom" (Funk 2006).

While this claim is true in theory, interpreting factors can be difficult in practice (discussed below).

SVD had rarely been used with recommender systems, perhaps because the technique usually performs poorly on sparse data sets where most of the values are missing. The Netflix data was certainly sparse, with most users rating only a tiny fraction of the available movies. But Funk adapted the technique to ignore missing values. Taking inspiration from an incremental-SVM approach developed for language processing (Gorrell 2006), Funk found a way to implement the approach in only two lines of C code (Funk 2006). Funk even titled the blog post explaining his method "Try This At Home," encouraging other entrants to incorporate the SVD approach into their own models. Nearly all of the other top-ranked competitors did so. When the Netflix Prize was finally awarded, SVD-based methods provided the single largest component of the models on the winning and second-place teams.

Even so, it is unlikely SVD techniques on their own would have been powerful enough to win the competition. One of the more unexpected revelations of the Netflix competition was the big advantages of blending different learning techniques together. As BellKor reported at the end of the first year of the competition, "combining predictions from multiple, complementary models improved

performance, with one model's strengths compensating for the weaknesses of others" (AT&T 2009). While SVD might be the single best technique, it would often miss relationships that that would be obvious to a human observer, like recommending a sequel to a user who had liked the first movie in a series. KNN models were much better at finding clusters of closely related films. By the end of the contest, teams were using megablends of hundreds of different models. And while latent-factor models like SVD and nearest-neighbors models made up the largest portion of the solution, the final blends included a complex mishmash of different techniques, from principle component analysis to ridge regression to Restricted Boltman Machine neural network approaches. AT&T's Chris Volinsky said "I don't think when we started that anybody expected that that would be the way to win the competition" (AT&T 2010).

The same premium on diverse approaches also led, eventually, to a wave of mergers among teams. The Netflix Prize rules, in addition the $1 million grand prize, called for the awarding of $50,000 yearly "Progress Prizes" for the team currently closest to the goal, providing that there had indeed been substantial progress over the course of the year. The catch was that the winning Progress Prize team would be required to publish a full accounting of their techniques, allowing competitors to catch up.

As the end of the first year neared, Bellkor had led since March, with a narrow but stable edge over the second- and third-place Gravity and Dinosaur Planet teams. But with only a day left in the Progress Prize window, the fifth- and sixth-place teams combined their predictions, and the blended results vaulted them into second place. This unexpected move set off a flurry of last-minute activity (and forum debates about whether this new tactic was fair). The Dinosaur Planet and Gravity teams followed suit with a hasty merger of their own, and the merged team submitted a final score that edged out BellKor's previous best. The BellKor team worked through the night, submitting two last-minute entries that eeked out a narrow victory in the Progress Prize.

At the end of the first year BellKor had managed an 8.43 percent improvement over CineMatch. But most of the easy progress had already been made. Over the next year the pace of improvement would be far slower. In early 2008, after the publication of BellKor's methods, several new teams appeared in the top 20 sites. In February, When Gravity and Dinosaurs Unite passed BellKor to take the contest lead.

BellKor's next advance came from modeling temporal effects in the data. Some movies— for example, *The Big Lebowski*—grow more popular over time, while ratings for movies like *Revenge of the Transformers* decline (AT&T 2009). Individual users also changed their rating habits over time, becoming more or less stingy in their awarding of stars. Adding time-dependent effects added greatly to the models' complexity. Since the average user rated movies on forty days in the sample, for example, adding time-dependent effects resulted in a forty-fold increase in user factors. But it did provide a slight boost in model accuracy.

As the contest dragged on, it became clear that BellKor would be unable to win the prize on their own. So they, too, decided to seek the improvements that other teams had shown when they combined their efforts. As they later wrote, "Teams that collaborated always improved over their individual scores *provided each team brought different methods to the table*" (AT&T 2009, emphasis original)

Bob Bell suggested a method by which teams could compare their results, without giving away all of their secrets. By adding statistically-regular noise to their predicted ratings, teams could share their output and perform simple calculations to see how similar their approaches were. These results showed that the team BigChaos was the best merger candidate. BigChaos had a relatively low RMSE, but more importantly, its predictions were least correlated with the predictions of BellKor, suggesting a potential payoff to collaboration. After legal negotiations, the merger went through.

As it turned out, much of BigChaos's contribution came from using sophisticated neural networks to blend the results. As Bell and Koren later wrote, "Where BellKor used simple, linear blending

based on each model's individual performance, BigChaos had determined that the individual RMSE of a single method was not the best indication of how much that model would add to the blend" (AT&T 2009). After the merger, Bell and Koren would send their individual model results to the BigChaos members, who would blend them together for submission.

With the improvements from BigChaos, the combined team won the second Progress Prize in October 2008 by a comfortable margin. But progress soon stalled. Once again, the solution was to find another team to merge with. This time the candidate was Pragmatic Theory. Pragmatic Theory was particularly good at identifying unusual or even quirky predictors, like the length of a film's title, or users that rated films differently on Friday than they would on Monday. On their own, these sorts of features predict little about users' ratings. In the context of the full model, however, they did provide a small additional boost in accuracy.

The teams initially disguised the merger by continuing to post separately. By adding noise to their results, they could see how close they were to the 10.0 percent finish line without alerting other teams to their progress. By June 2009, the new merged team knew that they had reached their goal. On June 26, 2009, the new team went public, submitting a 10.05% result under the name BellKor's Pragmatic Chaos (BPC).

The hectic conclusion of the Netflix Prize proved a replay of the runup to the first progress prize. According to the rules, the posting of a result with better than 10 percent improvement triggered a final 30-day period for all teams to submit their final answers. With nothing to lose, many teams rushed to join forces. The main competitor to the BellKor-led effort was a super-group called The Ensemble, which ended up including 16 different original teams, including the previously-merged Dinosaur Planet and Gravity groups. The Ensemble improved rapidly in the last few weeks of the contest. In the final days, The Ensemble seemed to nudge past BPC on the public leaderboard. But since the leaderboard was based on public data, and the contest would be judged on a similarly-sized but unreleased private dataset, it was not clear who was really in the lead.

On September 21 2009, almost three years after the contest opened, BellKor's Pragmatic Chaos was publicly announced as the winner. Only later was it revealed that The Ensemble had achieved the exact same level of improvement: an RMSE of 0.8567. BPC had won because of a tie-breaker in the fine print of the contest: BPC's final results had been submitted 24 minutes earlier. After years of effort, the contest had ended in a photo finish.

What Digital Newsmakers Can Learn from Netflix

Why should those interested in online news care about the Netflix prize? One answer is that these recommender systems now have enormous influence on democratic discourse. In his recent book *The Filter Bubble* progressive activist Eli Pariser claims that posts from conservative friends were systematically excluded from his Facebook feed. This sort of filtering heightens concerns about partisan echo chambers (Sunstein 2009), and it might make it harder for citizens to seek out opposing views even if they are inclined to. Increasingly, learning algorithms are also replacing editorial judgment and longstanding news norms.

Yet recommender system should be of interest for an even more fundamental reason. Recommender systems do not just influence which articles users see, but also which sites they end up visiting in the first place.

Whether they are funded by advertising or subscriptions, news sites require traffic to succeed. Sites quite literally live or die based upon their stickiness—their ability to attract readers, to make those readers stay longer when they visit, and to convince them to return again once they leave. Even slight differences in site stickiness compound quickly, and rapidly create enormous differences in audience.

Recommendation systems are one of the most powerful tools available for sites to keep and grow their traffic, and those who cannot deploy them are at profound competitive disadvantage.

The key question is this: which sorts of news sites can build, and benefit most, from high-quality recommender systems?

The Netflix contest provides a partial answer. Moreover, the Netflix Prize is likely the only chance we will have in the near future to look at the inner workings of Web recommendation systems using a large, public, real-world dataset. Netflix initially planned a successor contest to the Netflix prize. However, facing a class-action lawsuit and an FTC investigation regarding user privacy, Netflix canceled the intended sequel (Hunt 2010). Given the legal complications that attended the contest's conclusion, it is currently unthinkable that another large Website would sponsor a similar contest or release a comparable dataset.

The Netflix prize is often discussed as an example of crowd-sourced problem solving. The results of the contest, however, suggest that the advantages of recommender systems will accrue highly unevenly. The very largest sites have been able to build excellent content recommendation systems; the smallest sites have not.

Recommender systems favor large, well-resourced organizations. In order to inspire such a high level of participation, Netflix had to be willing to write a $1 million check. Large teams won the competition, and the winning team needed to merge with two other teams to cross the finish line. Even stronger evidence for this point comes from The Ensemble's nearly successful last minute scramble to catch up. By combining the efforts of more than a dozen other teams, The Ensemble was able very quickly to equal the results of BellKor's Pragmatic Chaos. Indeed, The Ensemble would likely have won if the contest had continued just a few days longer.

Building successful algorithms is an iterative and accretive process. It benefits from larger groups and diverse approaches, and thus provides inevitable size advantages. As the competition evolved, the models produced became absurdly complex. Managing this complexity also requires expertise and substantial staffing.

Similarly, the sorts of individuals who rose up the contest leaderboard also suggest limits to crowdsourced problem solving in contests like the Netflix prize. Many who led the most successful teams were already prominent academic or corporate researchers. Even those who were initially unfamiliar names, such as Funk (Brandynn Webb) or the Princeton undergraduates who made up the Dinosaur Planet team, had formal training and professional experience closely related to the topic of the competition. The project may not have benefited much from the contributions of average citizens, but it certainly benefited from drawing on a broader and more diverse set of those with subject-area expertise. Netflix would never have been able to hire that level of expertise at five times the contest budget. But if the research and positive publicity was worth it, it nevertheless required a seven-figure investment.

Not only do big sites have a large edge in terms of resources, but they also have an even more crucial advantage: *more data*. Building an infrastructure to collect, store, organize, analyze, and constantly update data is an enormous investment. This is not something that a small startup could have done nearly as successfully, and not just because of the money, hardware and expertise required. Data comes from monitoring users, and startups do not have nearly as many users to monitor. As AT&T's team put it, "As the competition progressed, using more information almost always improved accuracy, even when it wasn't immediately obvious why the information mattered or how little the information contributed" (AT&T 2009).

The need for as much information as possible has broad implications. One thing often overlooked in the discussions of the Netflix prize is that Netflix *already* had reached to the overall level of accuracy they paid $1 million for. As the contest FAQ explained,

The RMSE experienced by customers on the Netflix site is significantly better than the RMSE reported for the training dataset. This is due both to the increase in ratings data but also to additional business logic we use to tune which of the large number of ratings to learn from... let's just say we'd be seriously in the running for a Progress Prize, if we were eligible. (Netflix 2007)

In other words, even at the very start of the competition, Netflix was able to do significantly better than the raw Cinematch results indicated. They did this both by adding more variables and by training on a larger data set. The same techniques used to extract more information from a simple list of users and movie ratings work even better with data from (for example) user demographics or browsing behavior.

Recent statements from Netflix indicate that they have gone even further in this direction. Much has changed in Netflix's business since 2006, as the company has gone from a DVD-by-mail model to a focus on video streaming over the Web. In a recent blog post detailing their followup to the Netflix Prize, they explain that they now operate as if "everything is a recommendation," and that they extract information from almost every aspect of user behavior (Amatriain and Basilico 2012). Most aspects of the site are now personalized based on this data. While they are deliberately cagy about details and metrics, Netflix nonetheless claims that optimized models and additional features now provide them with a five-fold improvement over ratings data alone (Amatriain and Basilico 2012).

Yet for learning algorithms more broadly, what constitutes *more information* is not always obvious. More data is not just about more variables. In the initial stages of the competition, several teams attempted to supplement the movie data with a host of other details about each movie: the director, actors, studio, genre, year, etc. In simple linear models, the inclusion of this data at first seemed to improve the results. But with more sophisticated latent factor models and nearest neighbor models, adding movie details did not improve the predictions *at all*. This is likely because the machine learning models had already implicitly included all of this information.

More information can also be found even without collecting more data, by transforming the existing data set in order to extract new features. Koren, in a lecture a few months after the contest's end, declared that "One thing that we discovered again and again [...] is that understanding the features in the data, or the character of the data, [...] is far more important than picking the right model or perfecting the model" (Koren 2009) The Netflix competition started off with a very limited feature set: just user, movie, rating, and day of rating. Jumps in accuracy involved taking that limited data and extracting new features, like temporal effects.

The moral here is somewhat paradoxical. Netflix released a massive data set in order to find the best algorithm, but the algorithms themselves proved less important than the data. Similar lessons have emerged in other, quite different realms of machine learning. In research on natural language processing, Microsoft researchers examined how accuracy improved across several different algorithms as the amount of training data increased. Although these algorithms showed dramatically different performance on tests of 1 million words, as the researchers scaled up the training set—to 10 million, 100 million, and finally 1 billion words—the algorithms' performance became more and more similar. As Banko and Brill concluded, "These results suggest that we may want to reconsider the trade-off between spending time and money on algorithm development versus spending it on corpus development" (Banko and Brill 2001).

The Netflix contest also highlighted several parts of the "black box problem." One disadvantage of complex learning techniques is that, when a model is performing well, it is often not clear *why*. The success of latent factor models in the competition emphasized this issue. In theory, one might think of latent factor models as revealing human-interpretable categories like "action movie vs. non-action movie," or "serious vs. escapist," or "male-focused vs. female-focused." Sometimes the results that latent factor models give do seem to map easily to categories that humans already understand or expect.

But that is not really what happened with the Netflix prize. The dimensions that popped out of the data do not map neatly to obvious predefined categories. Funk's first attempt at using an SVD model (Funk 2006) found that the most important dimension was anchored on one end by films like *Pearl Harbor* (2001), *Coyote Ugly* (2000), and *The Wedding Planner* (2001), while the other end of the scale was defined by films like *Lost in Translation* (2003), *The Royal Tenenbaums* (2001), and *Eternal Sunshine of the Spotless Mind* (2004). Obviously these are very different sorts of films, yet it is tough to articulate a concise description of what separates these groupings. As Koren (2009) later concluded, "It is very difficult to give names to these axes." And if one latent-factor model is tough to interpret, how much harder is it to interpret the final blend of more than 700 models—many of which were themselves blends of different component models?

In one way, however, Netflix's example calls into questions claims that filtering technologies will end up promoting echo chambers and eliminating serendipitous exposure. Such worries have been a centerpiece of scholarship on personalized news over the past decade (Sunstein 2001). One of Pariser's key claims about what he terms the "filter bubble" is that it is ostensibly invisible to users (Pariser 2011). Netflix, however, tries hard to make users aware of its recommendation system: "We want members to be aware of how we are adapting to their tastes. This not only promotes trust in the system, but encourages members to give feedback that will result in better recommendations" (Amatriain and Basilico 2012). Netflix also attempts to explain (in an oversimplified way) why specific movies are recommended, typically highlighting recommendations' similarity to movies the user has already rated.

Even more important, Netflix shows that there is a performance boost for recommending *diverse* content, not just for predicting ratings accurately. Partly, this is because Netflix subscriptions are often shared among members of a household who may have very different tastes. But as Netflix explains, "Even for a single person household we want to appeal to your range of interests and moods. To achieve this, in many parts of our system we are not only optimizing for accuracy, but also for diversity" (Amatriain and Basilico 2012). The biggest, most blended models that drew on the most varied features performed best overall in the Netflix Prize. In hindsight, it is perhaps unsurprising that recommending a highly diverse basket of movies also ends up improving performance. But given concerns about "filter bubbles" and online echo chambers, a performance bonus for diversity challenges conventional wisdom.

Google News

The Netflix experience demonstrated several features of recommender systems that are likely to persist across many different Websites and varied genres of online content. In recent years, several of the largest online Websites have been willing to release greater details about their recommender systems, and the algorithms with which they personalize content for their users. Even more so than with the Netflix prize, the information released by companies like Google and Yahoo! and Microsoft and Facebook is only an incomplete picture of the whole. These companies are undertandably wary about releasing information that would give their competitors an advantage.

Nonetheless, recent disclosures do provide key details about how recommender systems are being applied in practice, and how they benefit some organizations over others. In particular, the results of A/B testing provide compelling evidence of just how important personalized content is for improving site traffic. Recommendation systems dramatically increase stickiness for the largest websites in ways small sites cannot replicate.

Consider the case of Google News, one of the largest news sites on the Web, and a pioneer in replacing editorial judgment with algorithmic decision-making. In 2007, Google researchers released a paper detailing the company's internal work in news personalization (Das et al. 2007). In some ways recommending news stories is similar to recommending movies. Most users, most of the time, arrive at

news site without knowing which specific articles that they would like to see. As Google's researchers put it, user attitudes are dominated by the demands to "show us something interesting" (Das et al. 2007, 271)

Yet news targeting also presents a series of unique problems, too. First, news articles provide a particularly severe example of the "cold start" or "first rater" problem. All personalization algorithms perform well with lots of information on both the items to be recommended and individual user preferences. With movies, for example, the accuracy increases for a user as he or she rates more movies, and as each movie gets reviews from a larger number of Netflix subscribers. News content, however, shows an enormous amount of churn day to day, and even hour to hour. By definition news readers are most interested in content that is *new*, and therefore has relatively little training data. Making the matter worse, it is quite costly—in both time and computing power—to constantly rebuild or retrain the recommendation framework to offer predictions for the newest content. Because site speed is one of the most important parts of the user experience, all Google properties are subject to a strict response time requirements. Personalized results have to be returned to the user in no more than a couple hundred milliseconds.

The technical infrastructure Google News requires is daunting: numerous large scale data centers, more than one million server computers, enormous investments in fiber, even customized operating systems and file systems. Total development costs for this infrastructure, including both hardware and software components, likely exceeded $10 billion. Many recommendation algorithms are computationally costly to implement at scale, and some of Das et al.'s findings focus on achieving similar performance with less computation. The initial paper details several slightly different algorithms, all in the same general family of methods as the K-nearest neighbor algorithm described above.

The most dramatic results in the paper come from Google's testing of how much these personalized recommendations improve traffic numbers. Google benchmarked its initial algorithms against the baseline in which users were recommended the stories that were most popular at any given moment. By interleaving personalized results with results based just on popularity, Google was able to control for the fact that higher-ranked items get more attention.

The results were striking: overall, stories based on collaborative filtering had 38 percent more clicks than stories chosen just based on popularity (Das et al. 2007, 279).

These early methods have now been superseded by even more effective forms of targeting. In 2010 Google released a second report on its targeting algorithms in Google News (Liu, Dolan, and Pedersen 2010). Here Google distinguished between the collaborative filtering approaches, which were the basis of its earlier work, and content-based approaches. Collaborative filtering looks at the similarity between users and between items, whereas content-based methods use text analysis to match users with the types of stories they have favored in the past. Content-based models proved better at recommending brand new content, and they better allowed for user differences. For example, Google reported that its first-generation collaborative filtering mechanism recommended entertainment news stories to all users, even those who had never once clicked on entertainment news.

Liu et al. detail a hybrid model combining both collaborative and content-based approaches. When the recommendation system has few news clicks from user, its predictions rely on collaborative methods, which tend to focus on the current overall news popularity trends. Yet once the system records a significant amount of click data, recommendations are based more and more on users' past behavior and demonstrated interests.

This hybrid model shows dramatic improvements over collaborative filtering alone, which (again) was itself far better than simply recommending users whatever was popular. Compared to straight collaborative filtering, the hybrid model produced 31 percent more clicks on news stories, though this was largely the result of shifting traffic from interior sections of the site to recommended stories on the

front page. Even more importantly, over the course of the study users who saw the hybrid model had 14 percent more daily visits to the Google News site. This is a remarkably clear demonstration of just how much improved recommendation systems can boost daily traffic.

Other computer science researchers have replicated Google's results on additional news sites. Hewlett-Packard researchers Evan Kirshenbaum, George Forman, and Michael Dugan conducted an experiment that compared different methods of content recommendation on Forbes.com. A mixture of content-based and collaborative-filtering methods performed best. HP's hybrid model increased clickthrough rates by 37 percent compared to a popularity-only ranking system (Kirshenbaum, Forman, and Dugan 2012, 11). Here again, we see dramatic evidence that recommender systems increase the stickiness of news sites.

Yahoo! and Behavioral Targeting

If Google's results are potentially worrisome for traditional news organizations, recent research released by Yahoo! is perhaps even more dispiriting. Yahoo!, too, has been highly active in personalizing and targeting its news results. While Yahoo! itself has been circumspect about releasing details of its news targeting methods, journalistic accounts have similarly claimed big improvements in news traffic and clickthrough rates. One recent report claimed that personalized targeting increased clicks on Yahoo!'s "Today" box by 270 percent (Boyd 2011).

But if Yahoo! has been relatively discreet about its news targeting methods, recent research papers have pulled back the curtain on its targeted advertising abilities. The same technologies that provide users with the most clickable content also allow advertisers to match their messages to the most promising potential buyers. Understanding how this behavioral targeting works is crucial for understanding the political economy of online media.

There are three general types of online ad targeting. At the broadest level there is *property targeting*, in which ads are run on sites that feature related content or favorable demographics. Showing truck ads on an automobile site or a sports site is an example of property targeting. Second, there is a *user segment targeting*, which typically focuses on the age range and gender of the user: for example, showing ads for trucks to 25–40 year-old men across a wider variety of properties.

Both of these methods are crude compared to *behavioral targeting*. As the authors explain, "The key behind behavioral targeting is that the advertisers can show ads only to users within a specific demographic of high-value (such as people likely to buy a car) and combine that with a larger number of opportunities (places to show ads) per user." In this case the Yahoo! researchers used support vector machines, a common machine learning technique, to predict which users were likely to be good prospects. But it is almost certain that similar results would have been obtained with other learning techniques.

The key difference between the Yahoo! research and previous efforts (at least public ones) lies in the type of training data. Typically, behavioral targeting models have looked at clicks on an online ad as the key metric. Yahoo researchers instead trained their data on "conversions," sales that resulted directly from clicking on an online ad (Pandey et al. 2011).

Clicks on ads are uncommon, with typical click-through rates just a fraction of a percent. And if clicks are rare, conversions are only a tiny fraction of clicks. Increasingly, however, retailers have provided Web beacons that beam sales information back to advertising networks and/or partner sites. Still, only a handful of organizations have the detailed user behavior and conversion data necessary to target in this way.

Yahoo's research demonstrates just how much purchase data matters. Pandey et al. performed A/B testing between models trained on conversion data, and the same methods trained just on click data. In four tested real-world advertising campaigns, conversions increased between 59 and 264 percent. In

every case there was a dramatic drop in advertisers' cost per sale. Advertisers ultimate goal, of course, is to get the greatest number of sales for the least advertising cost. The bottom line, as the researchers conclude, is that "we can improve the number of conversions per ad impression without greatly increasing the number of impressions, which increases the value of our inventory" (Pandey et al. 2011, 3)

The research also suggests that the greatest improvements accrue to the largest advertising campaigns. Since conversion are rare, only the largest campaigns have enough sales data in order to train the models effectively. This is especially note-worthy given that Yahoo! is one of the largest sites on the Web, with an enormous online ad inventory. If only the largest campaigns on the largest sites are able to take advantage of these techniques, this has significant implications for the Web as a whole.

What do these results mean for news sites--and especially for newspaper websites? For starters, they show that standalone news organizations cannot perform behavioral targeting nearly as effectively as Yahoo! or Google. Many newspaper executives and print publishers have argued that local newspaper Websites are valuable because they (supposedly) reach a local audience. The problem is that location targeting through property targeting is crude and inefficient. Nearly everyone who visits local newspaper sites also visits the most popular sites. Potential local customers can be found more cheaply and efficiently on Yahoo! or Facebook than on the local newspaper Website.

Size matters for behavioral targeting. Even on Yahoo!, one of the largest online publishers, small advertising campaigns cannot be targeted as effectively as large campaigns can. Few if any newspapers even have conversion data, and no campaign on a mid-sized local news site has the scale that effective targeting requires. That means that newspapers must either partner with big sites or online ad networks—at substantial cost—or else subsist on substantially lower impression prices than their inventory would receive a larger Website. Neither alternative is attractive.

Conclusion

The rise of recommender systems as a key mechanism of news delivery is a tectonic shift in the online news landscape, on par with the arrival of the rotary press or the emergence of the Web itself two decades ago. Like these previous shifts, recommendation technology strongly favors some news organizations over others. In conclusion, we can discern seven broad, interrelated lessons about which types of news organizations are likely to win—and lose—in a world with ubiquitous algorithmic filtering.

First, and most important, *recommender systems can dramatically increase site traffic*. Web traffic is properly thought of as a dynamic, even evolutionary process. Recommender systems make sites stickier, and users respond by clicking more and visiting more often. Over time sites with recommender systems have grown in market share, while those without have shrunk.

Second, *recommender systems favor sites with lots of goods and content*. There is only value in matching if the underlying catalog of choices is large. Small sites benefit little: users do not need help sorting through the content of a news site that only produces 6 articles a day. In the same vein, sites that have a wide diversity of content benefit most from recommender systems. Publications with a narrower scope—say, sites that focus just on technology news or entertainment gossip—derive less value from recommender systems.

Third, *recommendation systems benefit sites with better hardware and more staff expertise*. Even when the underlying techniques are relatively standard for the industry, deploying them in a production environment still takes enormous time, energy, and effort. Moreover, targeting techniques are often expensive in terms of CPU cycles and computing resources. Smaller organizations are unlikely to have hardware and resources to deploy cutting-edge techniques.

The expertise and equipment needed to target content the can also be used to target advertising. There is now abundant evidence that personalization systems can provide dramatically better results for advertisers, providing more sales per dollar of advertising spending and while increasing the overall value of a site's ad inventory. As the Yahoo! research shows, some sites—and especially sites with certain kinds of data—are far better at targeting than others. Sites that make more money in online advertising can use that revenue to produce even more content or to improve their sites, further increased their advantages over competing organizations.

Fourth, *recommender systems benefit sites with more data, and more valuable kinds of data*. The most popular and most heavily used sites have a significant advantage in building recommender systems over sites that are currently less popular. More signals, and a greater diversity of signals, significantly improves performance.

Fifth, *recommender systems* do not *necessarily produce "echo chambers" or "filter bubbles."* For Netflix and for Google News, the best-performing algorithms recommend a broad range of content, and they intentionally balance accuracy with diversity. Ideological isolation can still happen through other means, of course, and more study is needed. But thus far, the best-documented examples of recommender systems do not support the worries of Sunstein, Pariser, and others.

Sixth, *personalization systems promote lock-in*, making switching between sites costly. Consider an occasional user of Google News who visits Yahoo! News for the first time. Initially, this user will see news content that is a significantly poorer match for her individual news tastes. Much of this apparent advantage is temporary, as time spent on the Yahoo! News site would provide more and more information for Yahoo's targeting algorithms. But from the user's perspective, personalization algorithms provide large initial barriers to switching from one provider to another.

Lastly, *recommender systems promote audience concentration*. This is the opposite of what previous scholarship has assumed. Negroponte concluded in 1995 that the Daily Me would be a powerful decentralizing and dispersive force: "The monolithic empires of mass media are dissolving into an array of cottage industries... the media barons of today will be grasping to hold onto their centralized empires tomorrow." (57–8)

While Negroponte's technological vision was prophetic, his economic logic was precisely backward. There is a long tradition in media scholarship that ties homogenized, broadcast media with media consolidation (see discussion in Neuman 1991). Mass broadcasting provided large economies of scale, where the same sitcom or news broadcast could be seen in hundreds of millions of homes simultaneously. But most observers have failed to understand that hyperpersonalization can produce the same result as broadcast standardization. One large Website, by learning its users tastes, can match users to their preferred content far more efficiently than hundreds of small "cottage industry" sites. Economies of scope—where the efficiencies come from providing a broad mix of different products—generate concentration just as surely as economies of scale do. The Daily Me provides media organizations with historically unprecedented economies of scope, and this reality continues to reshape the media landscape.

References

Amatriain, Xavier and Justin Basilico (2012). "Netflix Recommendations: Beyond the 5 Stars." Blog post. url: http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html.

AT&T (2009). "Statistics Can Find You a Movie." url: http://www.research.att.com/articles/featured_stories/2010_01/2010_02_netflix_article.html.

— (2010). "From the Lab: Winning the Netflix Prize." url: http://www.youtube.com/watch?v=ImpV70uLxyw.

Banko, M. and E. Brill (2001). "Scaling to Very Very Large Corpora for Natural Language Disambiguation." In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 26–33.

Boyd, E. B. (2011). "Brains And Bots Deep Inside Yahoo's CORE Grab A Billion Clicks." In: *Fast Company*. url: http://www.fastcompany.com/1770673/brains-and-bots-deep-inside-yahoos-core-grab-billion-clicks.

Bucy, E.P. (2004). "Second Generation Net News: Interactivity and Information Accessibility in the Online Environment." In: *International Journal on Media Management* 6.1-2, pp. 102–113.

Das, A.S. et al. (2007). "Google News Personalization: Scalable Online Collaborative Filtering." In: *Proceedings of the 16th international conference on World Wide Web*. ACM, pp. 271–280.

Deuze, M. (2003). "The Web and Its Journalisms: Considering the Consequences of Different Types of Newsmedia Online." In: *New media & society* 5.2, pp. 203–230.

Funk, Simon (2006). "Try This At Home." Blog post. December11.

Gates, B. (2000). *Business at the Speed of Thought: Succeed in the Digital Economy*. New York, NY: Warner Business Books.

Gorrell, G. (2006). "Generalized Hebbian Algorithm for Incremental Singular Value Decomposition in Natural Language Processing." In: *Proceedings of EACL*, pp. 97–104.

Hunt, Neil (2010). "Netflix Prize Update." Blog Post, March 23. url: http://blog.netflix.com/2010/03/this-is-neil-hunt-chief-product-officer.html.

Kennard, William E. (1999). "From the Vast Wasteland to the Vast Broadband." Speech to the National Association of Broadcasters. April 20. url: http://transition.fcc.gov/Speeches/Kennard/spwek914.html.

Kirshenbaum, E., G. Forman, and M. Dugan (2012). "A Live Comparison of Methods for Personalized Article Recommendation at Forbes. com." In: *Machine Learning and Knowledge Discovery in Databases*, pp. 51–66.

Koren, Yehuda (2009). "The Netflix Prizet: Quest for $1,000,000." Lecture, Rutgers University. url: http://www.youtube.com/watch?v=YWMzgCsFIFY.

Liu, J., P. Dolan, and E.R. Pedersen (2010). "Personalized News Recommendation Based on Click Behavior." In: *Proceedings of the 15th international conference on Intelligent user interfaces*. ACM, pp. 31–40.

Mayer-Schoenberger, Victor and Kenneth Cukier (2013). *Big Data*. New York: Pantheon.

Negroponte, N. (1995). *Being Digital*. New York, NY: Knopf. Netflix (2007). "Frequently Asked Questions." url: http://www.netflixprize.com/faq.

Neuman, W.R. (1991). *The Future of the Mass Audience*. Cambridge University Press.

Pandey, S. et al. (2011). "Learning to Target: What Works for Behavioral Targeting." In: *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, pp. 1805–1814.

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin.

Schafer, J.B., J.A. Konstan, and J. Riedl (2001). "E-commerce Recommendation Applications." In: *Data Mining and Knowledge Discovery* 5.1, pp. 115–153.

Stromer-Galley, J. (2004). "Interactivity-As-Product and Interactivity-As-Process." In: *The Information Society* 20.5, pp. 391–394.

Sunstein, C. (2001). *Republic.com*. Princeton, NJ: Princeton University Press.

— (2009). *Republic.com 2.0*. Princeton, NJ: Princeton University Press.

Thurman, N. (2011). "Making 'The Daily Me': Technology, Economics and Habit in the Mainstream Assimilation of Personalized News." In: *Journalism: Theory, Practice & Criticism* 12.4, pp. 395–415.

Thurman, N. and S. Schifferes (2012). "The Future of Personalization at News Web-sites: Lessons from a Longitudinal Study." In: *Journalism Studies*.

Turow, J. (2012). *The DailyYou: How the New Advertising Industry Is DefiningYour Identity and Your Worth*. New Haven, CT: Yale University Press.

Vaidhyanathan, S. (2012). *The Googlization of Everything: And Why We Should Worry*. Berkeley, CA: University of California Press.

Zelizer, B. (2009). "Journalism and the Academy." In: *The Handbook of Journalism Studies*. Ed. by K. Wahl-Jorgensen and T. Hanitzsch. New York: Routledge, pp. 29– 41.

Author contacts:

Matthew Hindman
The George Washington University
School of Media and Public Affairs
805 21st St NW
Washington, DC 20052
Email: hindman@gmail.com


European
University
Institute
European University Press