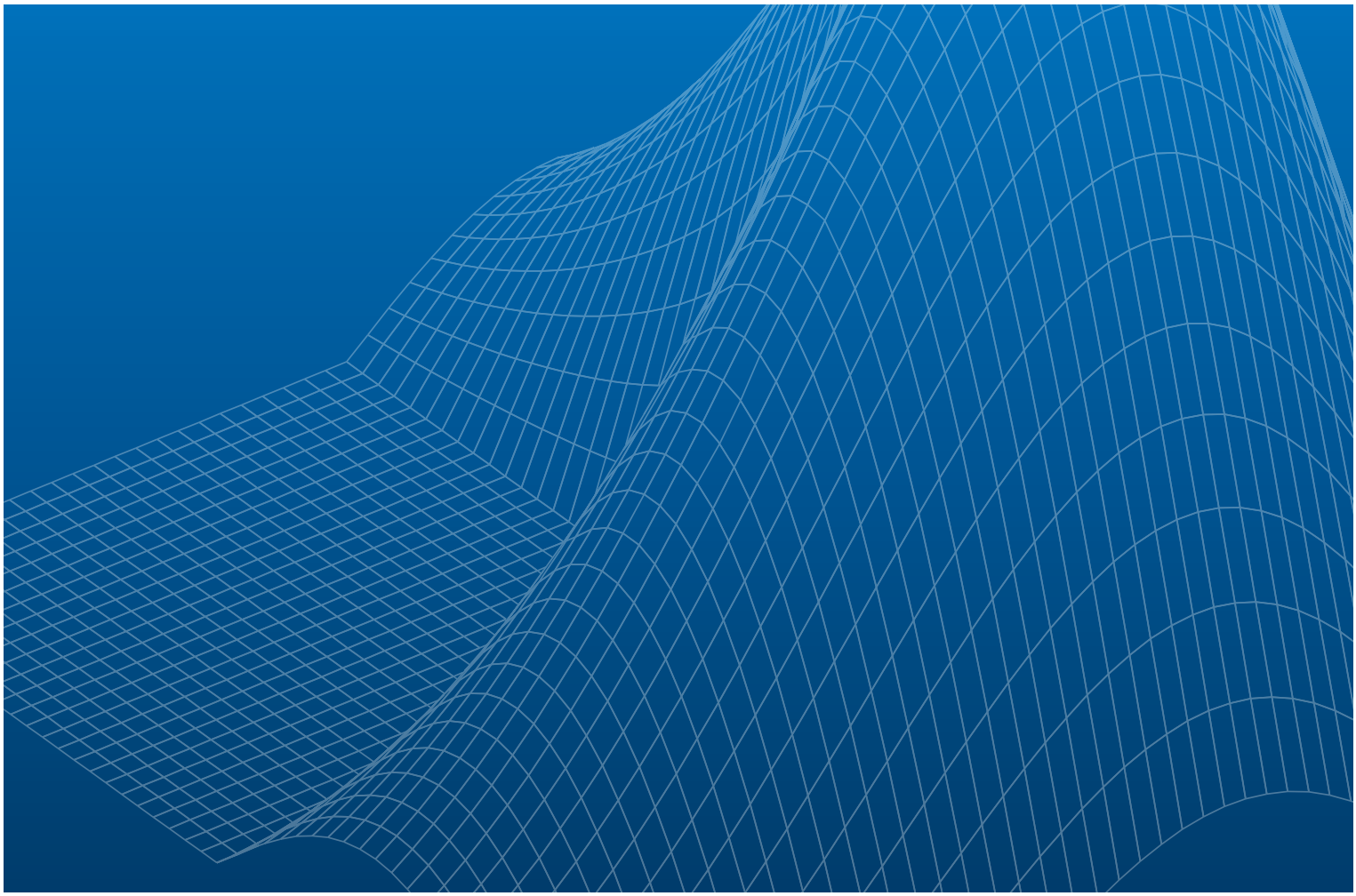


# The Netflix Prize and Production Machine Learning Systems: An Insider Look

WHITE PAPER

MathWorks®

In 2006, Netflix announced a famed machine learning and data mining competition called "Netflix Prize," with a $1 million award that was claimed in 2009.

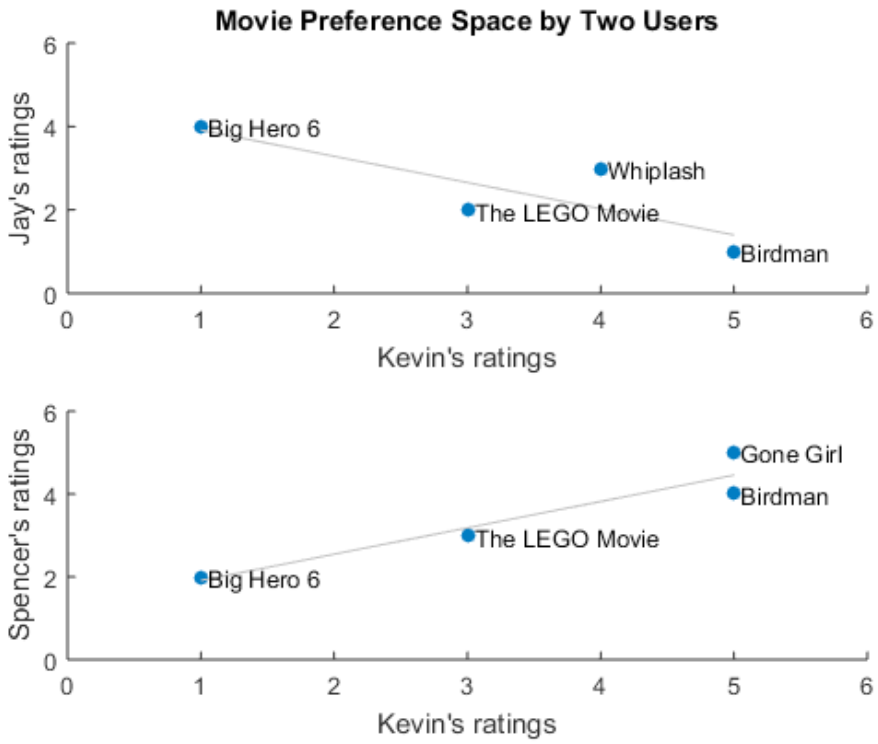
With all the publicity and media attention, whatever happened to the winning solutions in the end? Was it adopted in production? If not, then why? The Netflix blog post *Netflix Recommendations: Beyond the 5 stars* reveals practical insights about what matters, not just for recommender systems, but also for any real world commercial machine learning applications. This paper takes a closer look at the lessons learned.

## Contents

## The Netflix Competition

The goal of the Netflix Prize was to crowd source a movie recommendation algorithm that delivers 10%+ improvement in prediction accuracy over the existing system. If you use Netflix, you see movies listed under "movies you may like" or "more movies like this", etc. This algorithm powers personalized user experience on Netflix.

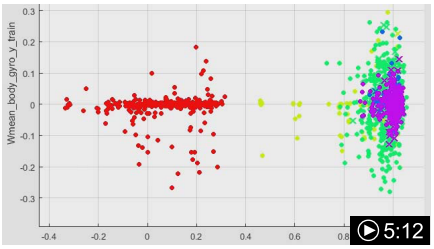
Here is a quick example of how the competition worked. *Collaborative filtering (CF)* is one of the fundamental algorithms for recommender systems, and it is based on an idea that you can use the ratings from users who share similar tastes to predict five-star ratings for unrated items. In this fictitious example, ratings by two users are compared over movies they both rated. When you plot a best-fit line, the line slopes up if user ratings are similar, and down if ratings are different.



Collaborative filtering can predict ratings of unrated items by using this correlation, and by combining existing ratings for other items among similar users.

To determine the prize winner, Netflix used metrics called ***root mean square error (RMSE)*** (which takes the difference between the actual ratings such as four stars and predicted ratings such as 1.4965 stars) and averages them to produce a single number. If the actual ratings and predicted ratings match completely, then RMSE would be zero.

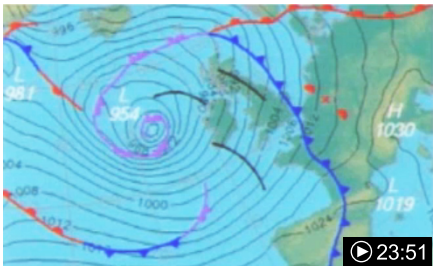
However, recommendations are typically delivered as a top-N list on the website – not as raw predicted ratings such as 1.4965 stars. Is RMSE really a meaningful metric then? For a competition, Netflix had to select a single metric to determine the winner.




*Classify Data Using the Classification Learner App*

## What Netflix Did with the Winning Solutions

There were two Netflix Prize submission standouts: the winning team that managed a 10% improvement, and a second team that designed an 8.43% improvement. Netflix adopted the solution that provided an 8.43% increase in accurate recommendations. Netflix did not adopt the solution that achieved a 10% improvement because the benefit of the additional accuracy gains did not outweigh the engineering effort needed to bring it into a production environment.




*How Weather and Pricing Affect Sales: Using MATLAB to Improve Tesco's Supply Chain*

In addition, the Netflix business model changed from DVD rental to streaming, which in turn changed the way data is collected and recommendations are delivered.

It is interesting to consider why the additional 1.57% in accuracy gains may not have been worth the effort. For example, you could improve the RMSE by closing the prediction gaps in lower ratings. However, Netflix would not have shown those low-rated movies on the user interface anyway. In a production system, improvements that support **scalability** and **agility** have a greater impact than RMSE.

## Lessons Learned: New Metrics

Even with the solution adopted, Netflix needed to overcome additional engineering challenges because the RMSE metric encouraged participants to focus more on accuracy than scalability and agility.

- The number of ratings in the competition dataset was 100 million, but the actual production system had over 5 billion
- The competition dataset was static, but the number of ratings in the production system keeps growing (4 million ratings per day, at the time the blog post was written)

When Netflix talks about their current system, it is notable what they highlight.

- "75% of what people watch is from some sort of recommendation"
- "Continuously optimizing the member experience and have measured significant gains in member satisfaction"

What Netflix finds most important is usage, user experience, user satisfaction, and user retention, which all aligns with their business goals better than RMSE. The second bullet point refers to A/B testing that Netflix is conducting on their live production system. That means they are constantly changing that system...

## Lessons Learned: System Architecture

"Coming up with a software architecture that handles large volumes of existing data, is responsive to user interactions, and makes it easy to experiment with new recommendation approaches is not a trivial task," write Netflix bloggers ***Xavier Amatriain*** and ***Justin Basilico*** (to learn more, see this article ***System Architectures for Personalization and Recommendation.***)

One of the techniques used in the Netflix Prize winning solutions was an ensemble method called linear stacking. Netflix adopted a form of a linear stacking technique to combine the predictions from multiple predictive models, in order to produce final recommendations. You can set up multiple subsystems to run different predictive models, and then combine the output of those systems to produce the final result. This is a very flexible architecture because you keep adding more to the combination as you develop new algorithms.

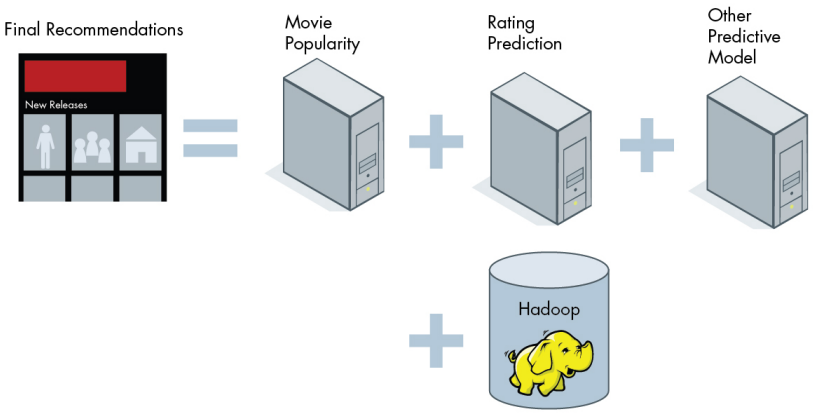


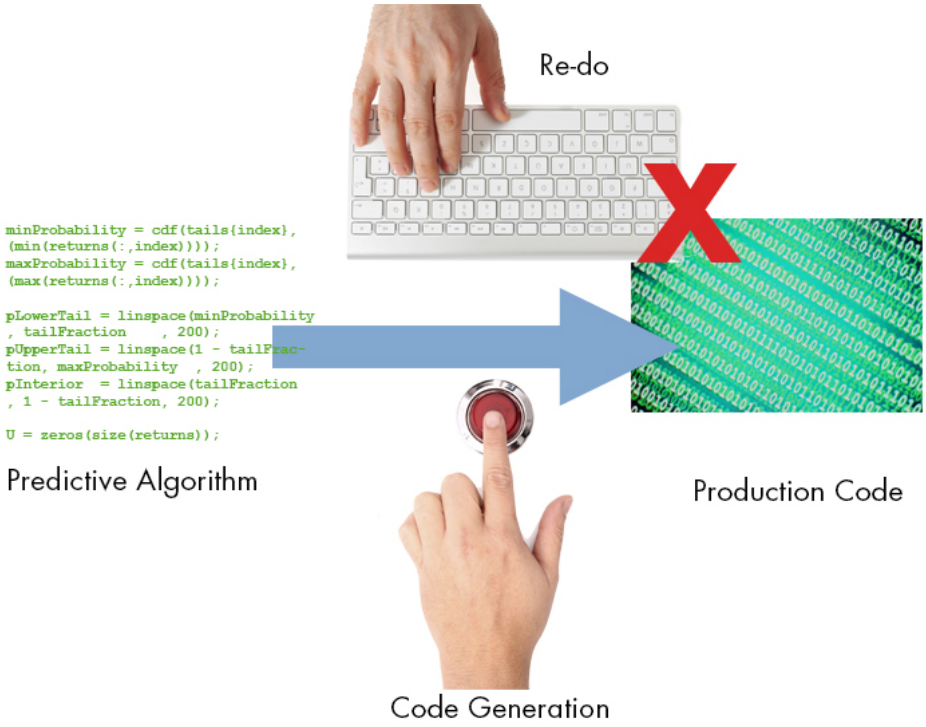

*Figure 1: Linear Stacking.*

Netflix uses three layers of service: offline, nearline, and online.

- Offline to process data – precomputes the time-consuming steps in a batch process
- Nearline to process events – bridges the two subsystems by precomputing frequently performed operations and capturing the result before active user actions
- Online to process requests – responds to user action instantaneously by taking advantage of the offline and nearline outputs


## Avoid Dual Implementations

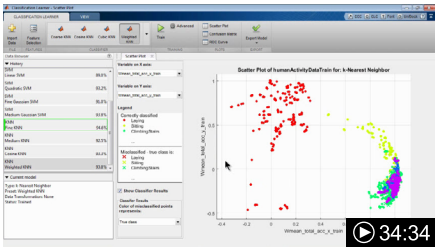
Netflix uses Hadoop to run this offline process, and the algorithm used on the small dataset must be rewritten in MapReduce. This is an example of what Netflix calls the ***dual implementation problem*** (see slide 20), and it's not limited to the offline process, either. It's important to develop and validate machine learning algorithms in specialized tools, and re-implement algorithms in another language to scale them on large systems. This limits the scalability and agility because this is a very time-consuming process. Netflix proposes that sharing as many components as possible between the development system and production system eases this pain point.



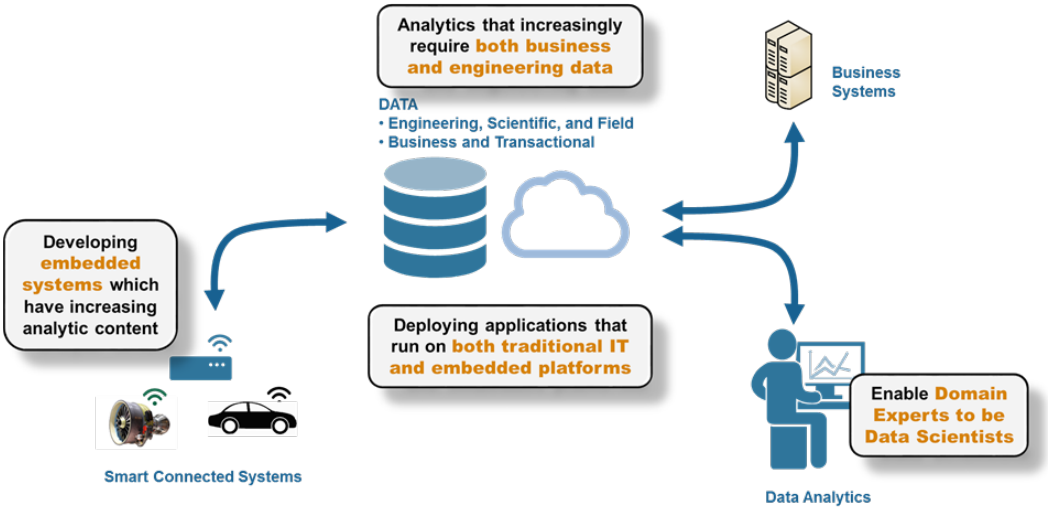
*MATLAB* addresses this pain point by enabling direct deployment of MATLAB code to production systems through various deployment options.

For example, algorithms developed in MATLAB can be compiled directly into various deployment targets.

- You can update the algorithms in MATLAB; updates are immediately deployable to production systems
- If you switch your production systems, you can simply use another deployment option suitable to your new environment
- It is easy to compare and validate the output from the prototype you develop in MATLAB from the production systems because they are based on the same source code



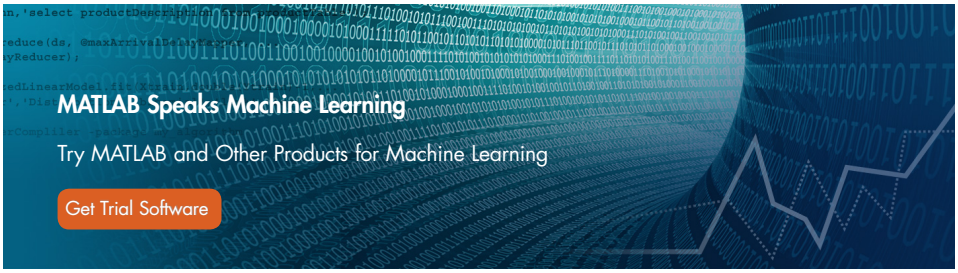

*Machine Learning Made Easy*



To learn more about MATLAB capabilities, refer to the following resources.

- *MATLAB MapReduce and Hadoop*
- *MATLAB Production Server*
- *Machine Learning*

## Machine Learning Examples with MATLAB


## About MathWorks

MathWorks® is the leading developer of mathematical computing software. Engineers and scientists worldwide rely on its products to accelerate the pace of discovery, innovation, and development.

MATLAB®, the language of technical computing, is a programming environment for algorithm development, data analysis, visualization, and numeric computation. The company produces nearly 100 additional products for specialized tasks such as data analysis and image processing.

Learn more at *mathworks.com*





92959v00 01/16