The Story of the Netflix Prize: An Ensembler's Tale

Lester Mackey

University of California, Berkeley

In the Beginning

- Oct. 2, 2006: Slashdot article

Build a Better Netflix, Win a Million Dollars?

Posted by **CmdrTaco** on Monday October 02 2006, @10:56AM from the trumps-our-redesign-contest dept.
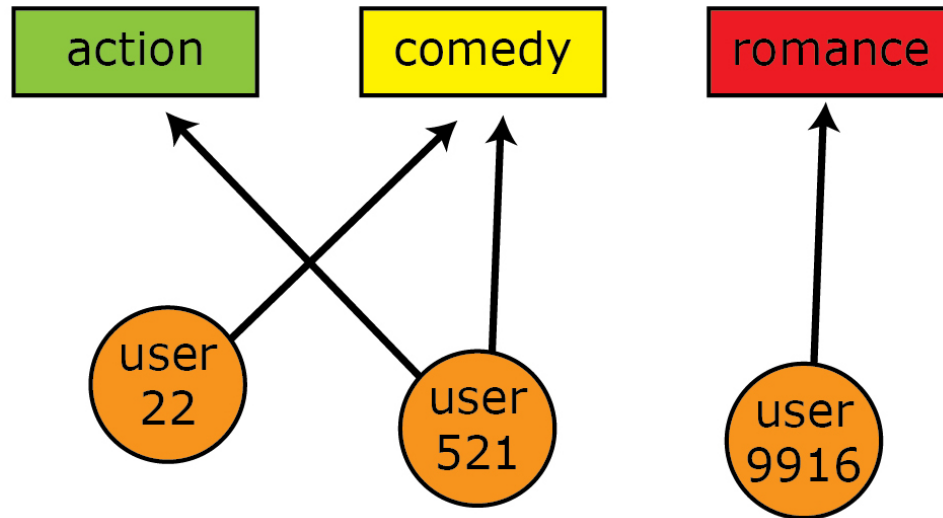
- Days later: Dinosaur Planet founded

= David Lin + David Weiss + LM

  - Named after the first movie in the Netflix dataset

A First Approach: Clustering

- Divide users (or movies) into groups based on similarities

- Use group information to predict user ratings
  - e.g. The average action-lover gives Indiana Jones a 5
- Hard clustering: each user belongs to a single cluster
- Soft clustering: each user fractionally belongs to all clusters

Clustering with Missing Data

- Centroid-based clustering
  - Represent user by incomplete ratings vector, r_u
  $$r_u = (1, 5, ?, ?, 3, ?, 4)$$
  - Represent cluster by centroid vector, c_k
    - Typically, c_k is average of user vectors in cluster k
  - Minimize (estimated) distance between users and their cluster centers
- Result: -0.3% improvement over Cinematch

Matching Cinematch

- Incorporate prior information
  - Positive ratings {3,4,5} vs. negative ratings {1,2}
    - Estimate $E[r|r \geq 3], \; E[r|r < 3], \; P(r < 3)$ and combine
  - Ordinal nature of rating data
    - Estimate $P(r < t) \text{ for } t \in \{2, 3, 4, 5\}$ and combine
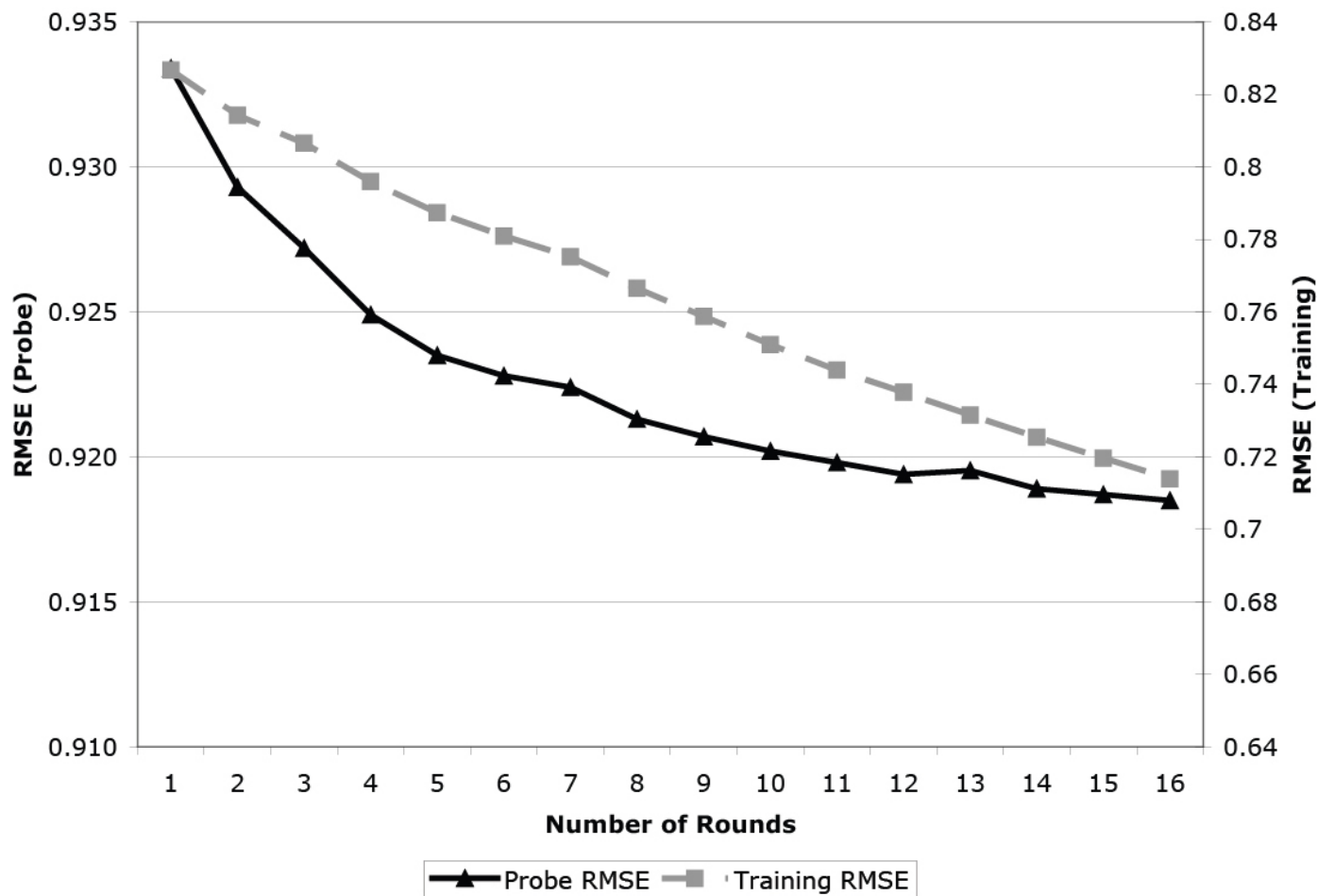 - Result: 0.5% improvement over Cinematch

Training on Errors

Recurring theme

- Train one model to predict and hence correct the errors of another model

- Long history in statistics and machine learning
  - Tukey's twicing (1977)
  - Boosting (Schapire, 1990)
  - Gradient boosting (Friedman, 1999)

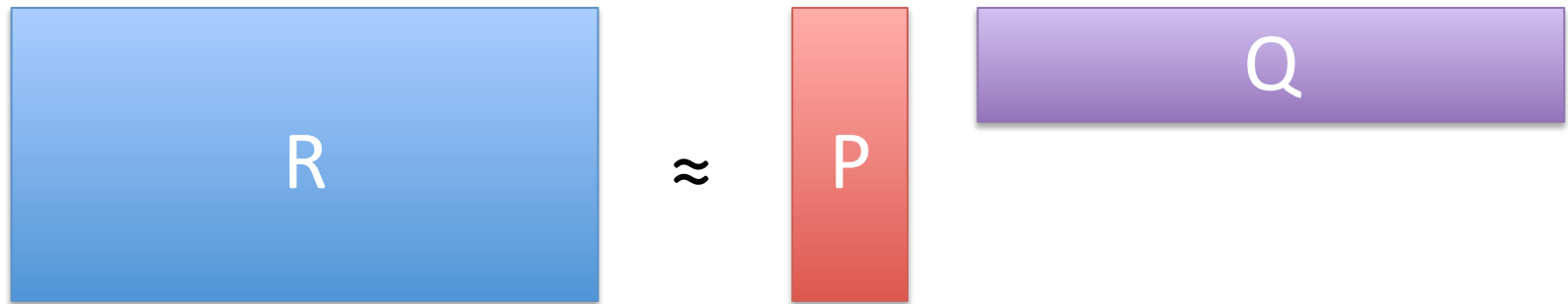- e.g., Cluster on errors of clustering predictions

Clustering on Errors

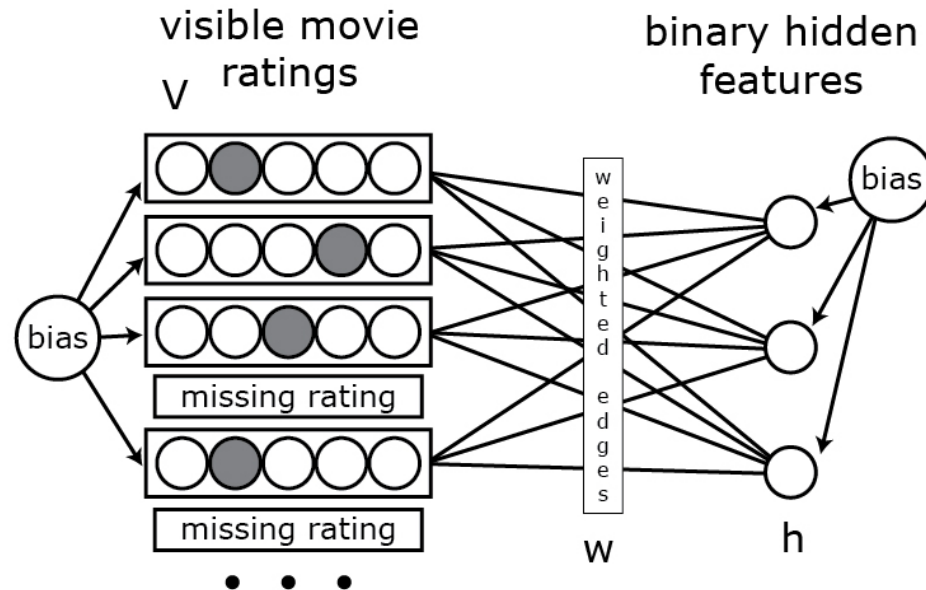Result: 3.0% improvement over Cinematch

The Three Pillars

- Matrix factorization

R ≈ P Q

- Alternating least squares
- Online/Stochastic gradient descent
  - Dec. 2, 2006, Simon Funk (Brandyn Webb)

- Typical improvement: 4%

The Three Pillars

- Restricted Boltzmann machines
  - May 2007, Salakhutdinov and Mnih

- Typical improvement: 5%

The Three Pillars

- Nearest neighbor methods
  - Classical KNN: 0.5% improvement
  - KNN with learned weights:
    - Aug. 2007, Bell, Koren, Volinsky
- Typical improvement: 4.6%
- Often **trained on errors** of MF or RBM

A Bug's Life

Finding Nemo

Shrek

Image courtesy of Chris Hefele
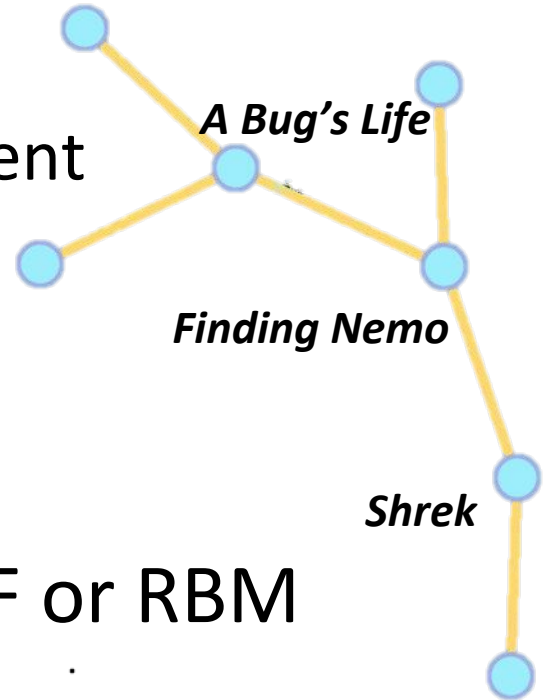
Milestones

- Spring 2007: Dinosaur Planet enters "Top 10"
- June 2007: DP graduates from college
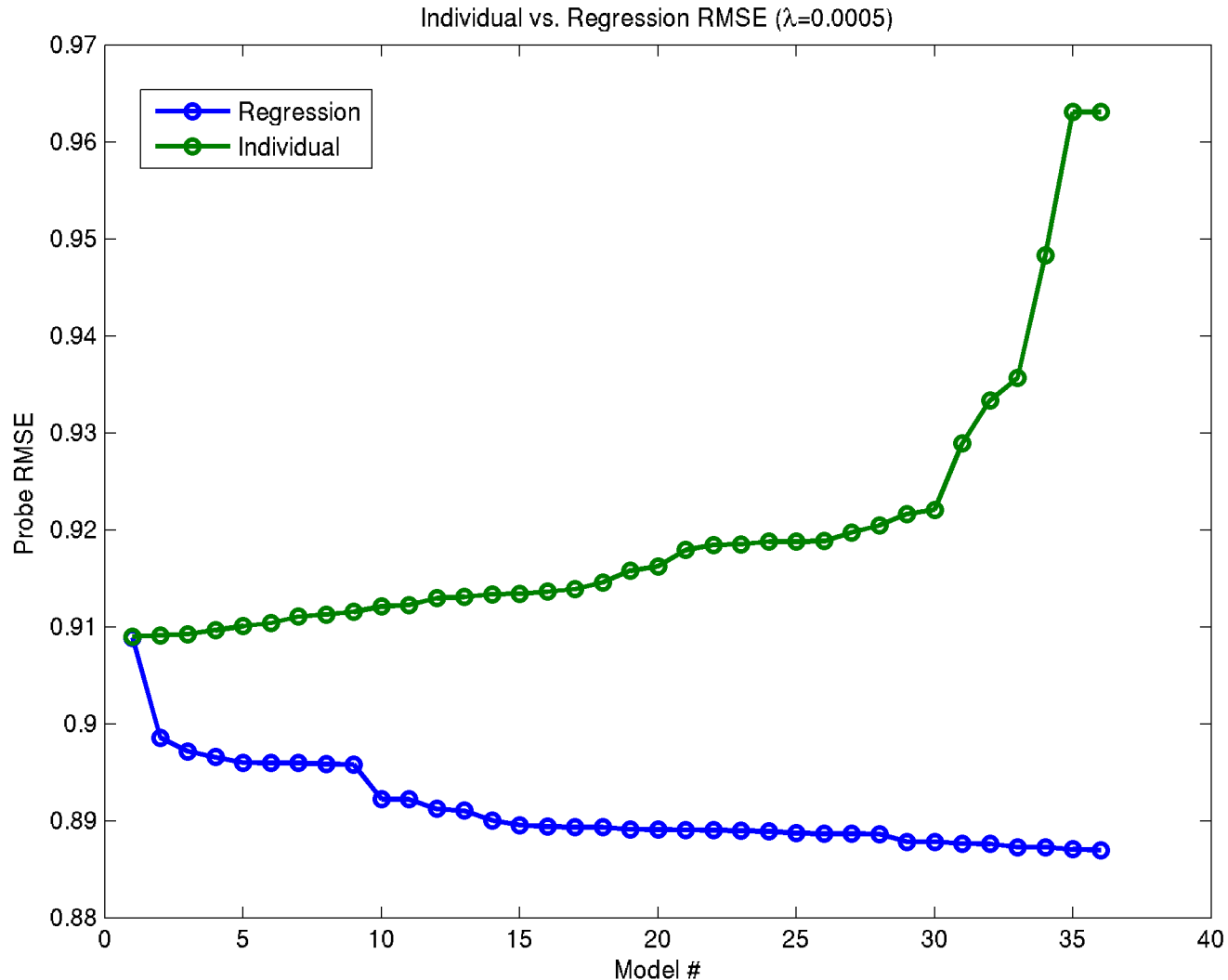
Model Ensembling

Recurring theme

- Combining the predictions of multiple models to yield improved performance
- Motivation:
  - Diminishing returns from optimizing a single algorithm
    - Best single model improvement: 8.24% (Aron Miller)
    - The Ensemble's final improvement: 10.09%
  - Different models capture different aspects of the data
    - Global commonalities of MF vs. Local similarities of KNN
  - Variance reduction from uncorrelated inputs

Model Ensembling

- Stacked linear regression (Wolpert, Breiman)
  - Target = held-out ratings, r
  - Covariates = model predictions, P
  - Tikhonov regularization to reduce overfitting

$$\min_{\beta} \|\mathbf{r} - \mathbf{P}\beta\|^2 + \lambda\|\beta\|^2$$

Model Ensembling

Individual vs. Regression RMSE (λ=0.0005)

Result: 2.0% improvement over best model

Model Ensembling Variations

- Add user, item, and date features as covariates
  - User rating count
  - Date of rating
  - Average inverse user rating count per movie
- Sparse regression: L1 regularizer or nonnegativity constraints
- Regress on pairwise interactions
  - Greedy selection or bagging with random subsets
- Result: 7.96% improvement over Cinematch

Progress Prize 2007

- Sept. 3, 2007
  - Dinosaur Planet takes first place (from reigning champion BellKor)
- One hour later
  - BellKor takes back first place ← **Recurring theme**
- Sept. 19, 2007
  - Gravity contacts DP about potential collaboration ← **Recurring theme**
  - Gabor Takacs, Istvan Pilaszy, Bottyan Nemeth, Domonkos Tikk
- Oct. 1, 2007
  - When Gravity and Dinosaurs Unite overtake BellKor with 8.38%
- 76 seconds later
  - BellKor ties with 8.38%
- Oct. 2, 2007
  - KorBell wins the first $50,000 progress prize with 8.43% improvement

Progress Prize 2008

- Sep. 30, 2008
  - Top 2 teams (BellKor and BigChaos) combine to become the only qualifier for the progress prize

- Jan. 2009: Grand Prize Team founded

Grand Prize Team

- Anyone could join
- Share of $1 million grand prize proportional to improvement in team score
- Many heeded the call
  - Joe Sill, Ces Bertino, Dan Nabutovsky, Bill Roberts, Wojtek Kulik, Willem Mestrom, David Purdy
- And brought new techniques with them
  - Gaussian missing data model
  - Feature-Weighted Linear Stacking

Gaussian Missing Data Model (Roberts)

- Assume each vector of user ratings drawn from common multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$
  - Incomplete vector of observed ratings drawn from marginal distribution
- Choose (μ, Σ) to maximize likelihood
  - Expectation-Maximization or gradient ascent
- Predict missing ratings as conditional expectation given observed ratings
- Result: 6.38% improvement

Feature-Weighted Linear Stacking (Sill, Takacs, Mackey, Lin)

- An adaptive approach to stacked linear regression
- Allow model ensembling weights to depend linearly on known features of the user, movie, and date
  - Did the user rate more than 3 movies on this date?
  - Log number of times the movie has been rated
  - Log number of distinct dates on which a user has rated
  - Log of average correlation between movies rated by user and movie to be predicted
- Result: 8.82% ➡ 9.46% improvement for GPT

The Last Call

- June 26, 2009
  - Top 3 teams (BellKor, BigChaos, and Pragmatic Theory) combine to pass the Grand Prize threshold
  - Initiates 30 day last call period for $1 million grand prize
- June 30, 2009
  - GPT begins deeper collaboration
  - Message board to share ideas, server to share code and predictions
- July 5, 2009
  - Vandelay Industries! contacts GPT about potential collaboration
- July 7, 2009
  - Opera Solutions joins Vandelay Industries!
- July 20, 2009
  - The Ensemble is born

The Ensemble

- Grand Prize Team
  - Gravity
    - Gabor Takacs, Istvan Pilaszy, Bottyan Nemeth, Domonkos Tikk
  - Dinosaur Planet
    - David Lin, Lester Mackey, David Weiss
  - Joe Sill
  - Ces Bertino
  - Dan Nabutovsky
  - William Roberts
  - Wojtek Kulik
  - Willem Mestrom
  - David Purdy

- Vandelay Industries!
  - Greg McAlpin
  - Bill Bame
  - Bo Yang
  - Chris Hefele
  - Jeff Howbert
  - Xiang Liang
  - Larry Ya Luo
  - Aron Miller
  - Steve Pagliarulo
  - Opera Solutions
    - Bruce Deng, Peng Zhou, Priyanka Rastog, Arvind Gangadha, Jacob Spoelstra
  - Craig Carmichael
  - Mike Linacre
  - Edward de Grijs
  - Clive Gifford
  - Feeds2
    - Nicholas Ampazis, George Tsagas

Learn more at http://the-ensemble.com/
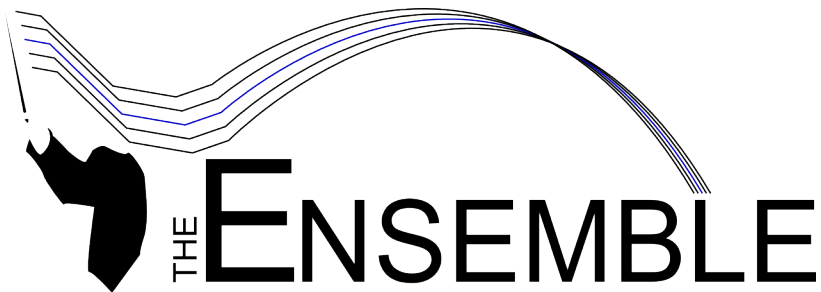
The Road to the Grand Prize

- July 25, 2009
  - The Ensemble submits
  - Quiz RMSE 0.8554 (10.09% improvement)
- July 26, 2009, 6:18pm
  - BellKor's Pragmatic Chaos responds
  - Quiz RMSE 0.8554 (10.09% improvement)
- July 26, 2009, 6:38pm
  - The Ensemble makes its final submission
  - Quiz RMSE 0.8553 (10.10% improvement)
- July 26, 2009, 6:42pm: Contest closes

The Other Road to the Grand Prize

- July 25, 2009
  - The Ensemble submits
  - Test RMSE 0.8568 (10.05% improvement)
- July 26, 2009, 6:18pm
  - BellKor's Pragmatic Chaos responds
  - Test RMSE 0.8567 (10.06% improvement)
- July 26, 2009, 6:38pm
  - The Ensemble makes its final submission
  - Test RMSE 0.8567 (10.06% improvement)
- Tie breaker: Time of submission

The End

And then there were two…

Teams shown from first appearance in top 20.

Copyright AT&T