Ensemble Learning

Better Predictions Through Diversity

Todd Holloway

ETech 2008

Outline

Building a classifier (a tutorial example)

- Neighbor method
- Major ideas and challenges in classification

Ensembles in practice

- Netflix Prize

Ensemble diversity

- Why diversity?
- Assembling Classifiers
    - Bagging
    - AdaBoost

Further information

Supervised Learning

Learning a function from an attribute space to a known set of classes using training examples.

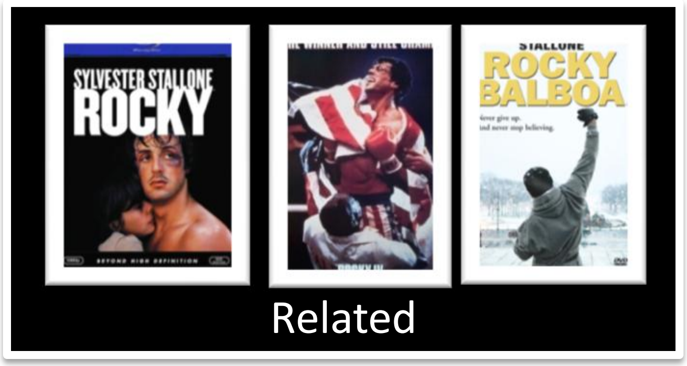
Ensemble Method

Aggregation of multiple learned models with the goal of improving accuracy.

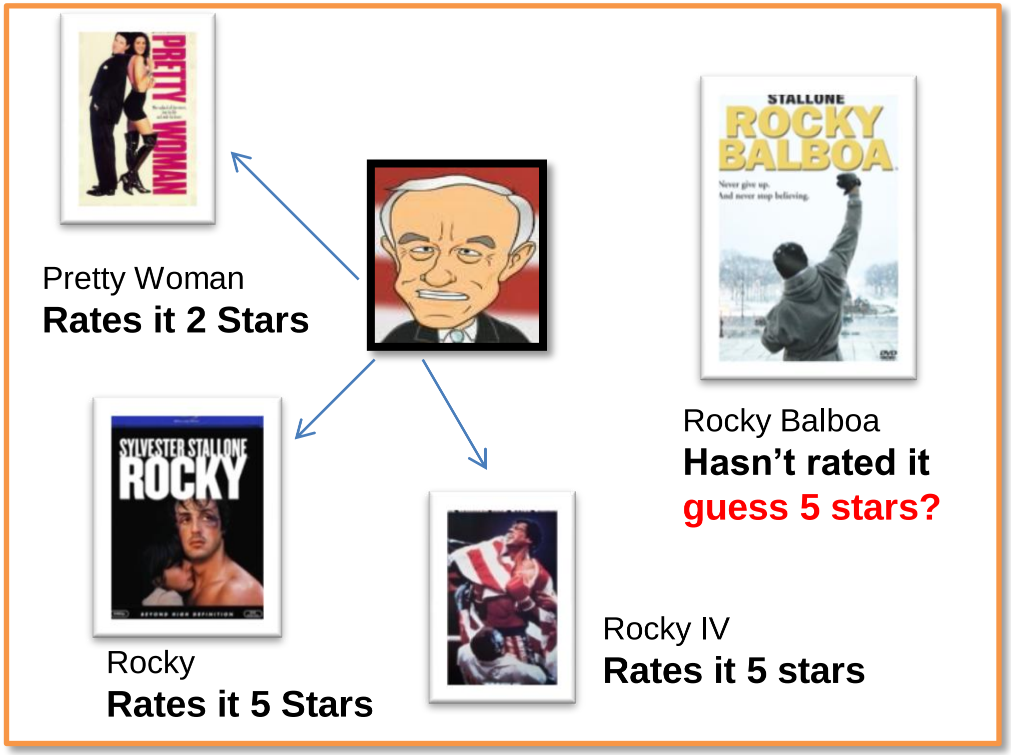
Tutorial: Neighbor Method

Idea

Related items are good predictors



Suppose the attributes are movie titles and a user's ratings of those movies. The task is to predict what that user will rate a new movie.



Pretty Woman
Rates it 2 Stars

Rocky
Rates it 5 Stars

Rocky IV
Rates it 5 stars

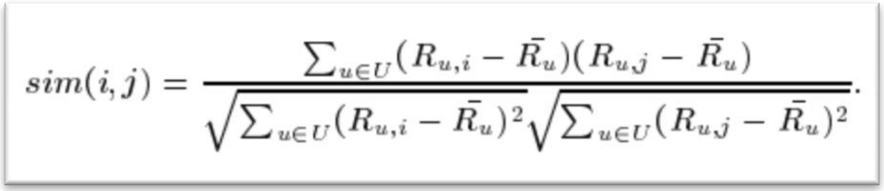
Rocky Balboa
Hasn't rated it
guess 5 stars?

Relatedness

The catch is to define ‘related’

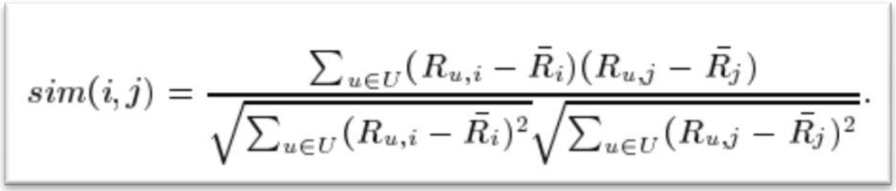
1. ‘Off the shelf’ measures

Adjusted Cosine



$$sim(i,j) = \frac{\sum_{u\in U}(R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u\in U}(R_{u,i} - \bar{R}_u)^2}\sqrt{\sum_{u\in U}(R_{u,j} - \bar{R}_u)^2}}.$$

Pearson Correlation



$$sim(i,j) = \frac{\sum_{u\in U}(R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{\sqrt{\sum_{u\in U}(R_{u,i} - \bar{R}_i)^2}\sqrt{\sum_{u\in U}(R_{u,j} - \bar{R}_j)^2}}.$$

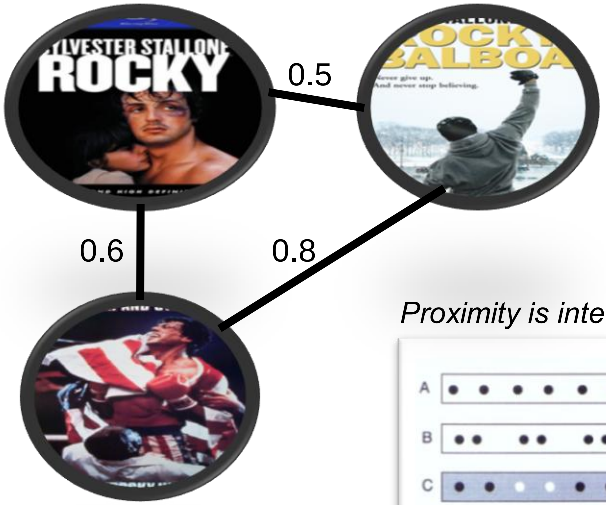
- Sarwar, et al. Item-based collaborative
filtering recommendation algorithms. 2001.

2. Tailor measure to dataset

How to begin to understand a relatedness measure?

Visualization of Relatedness Measure

1. Create a graph



Proximity is interpreted as relatedness…

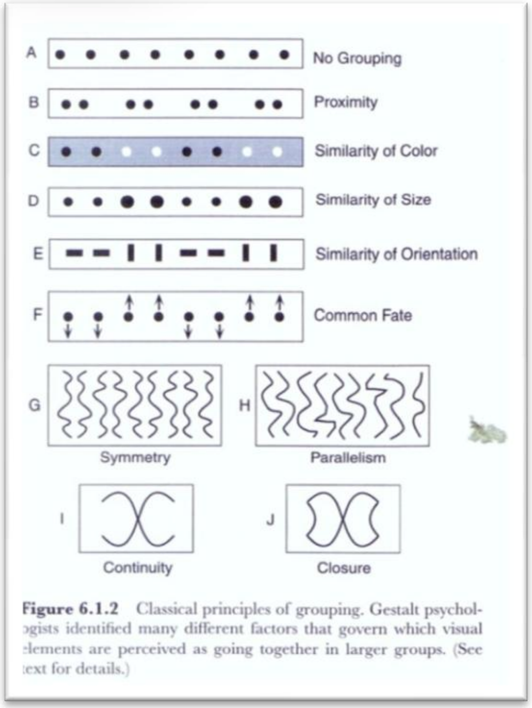


Figure 6.1.2 Classical principles of grouping. Gestalt psychologists identified many different factors that govern which visual elements are perceived as going together in larger groups. (See text for details.)

2. Arrange nodes

- Related nodes are close
- Unrelated are farther apart

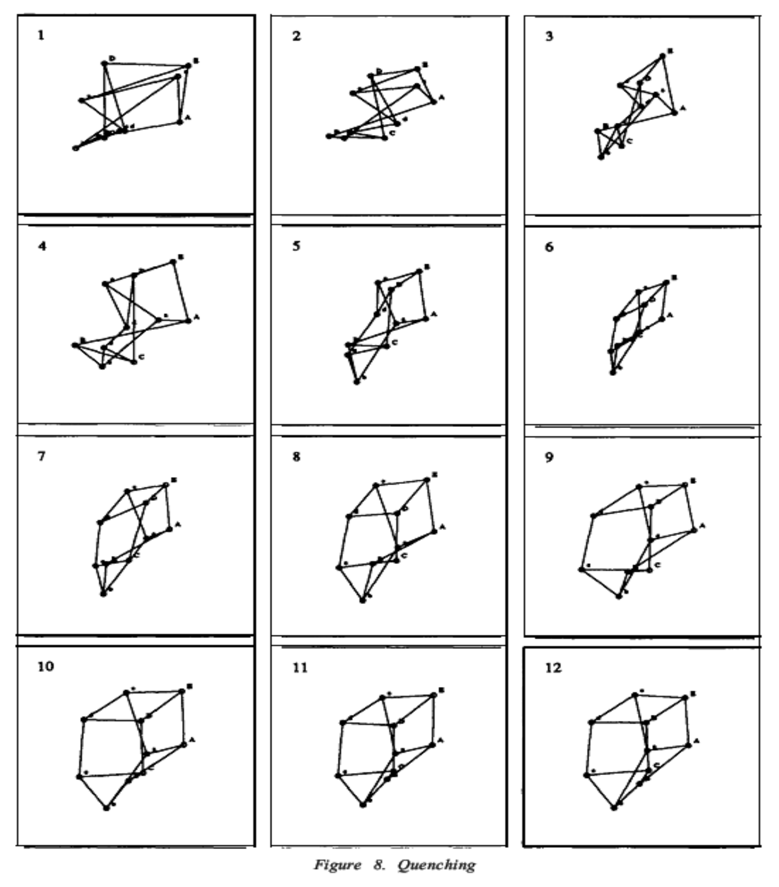
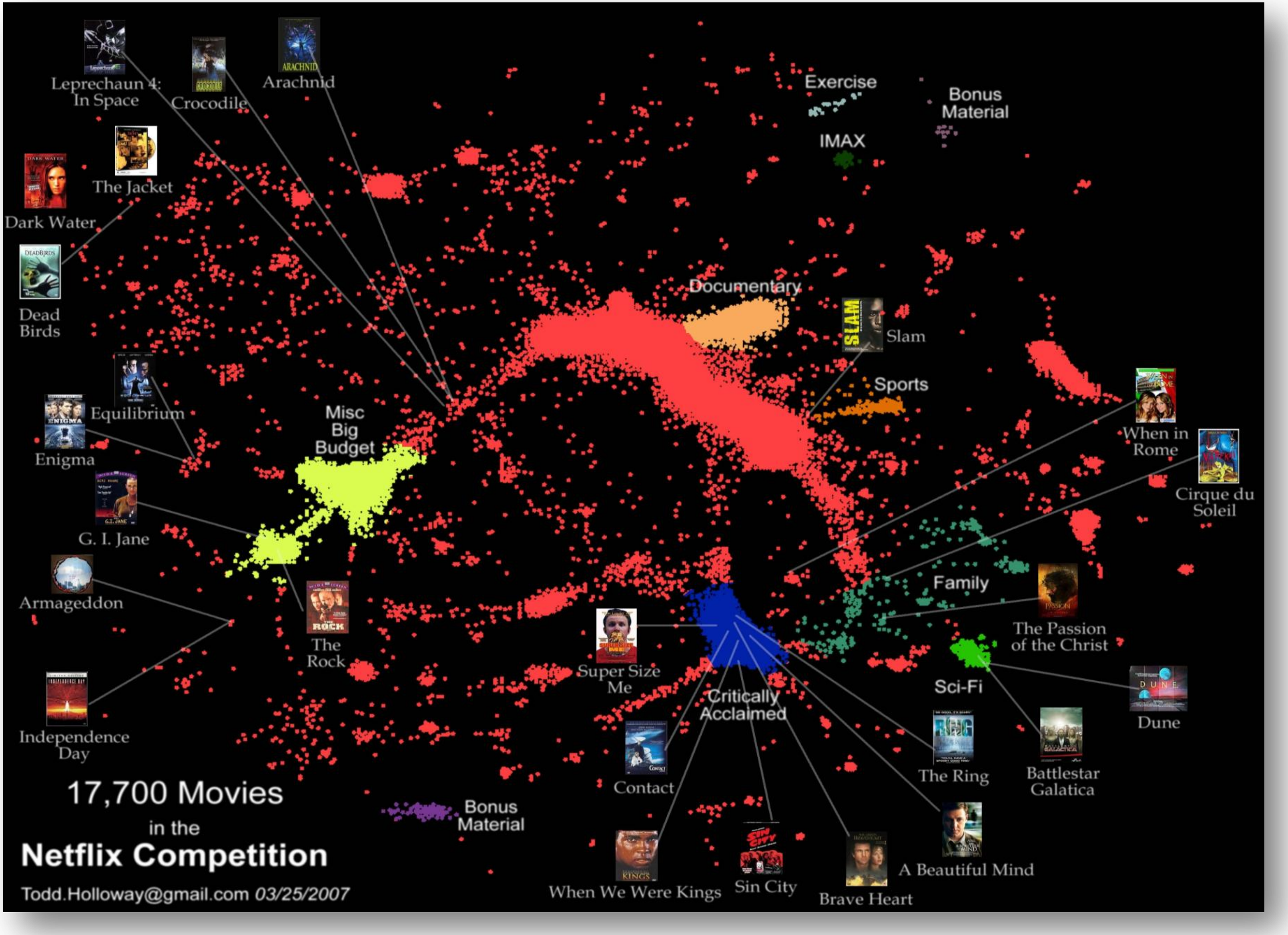


Figure 8. Quenching

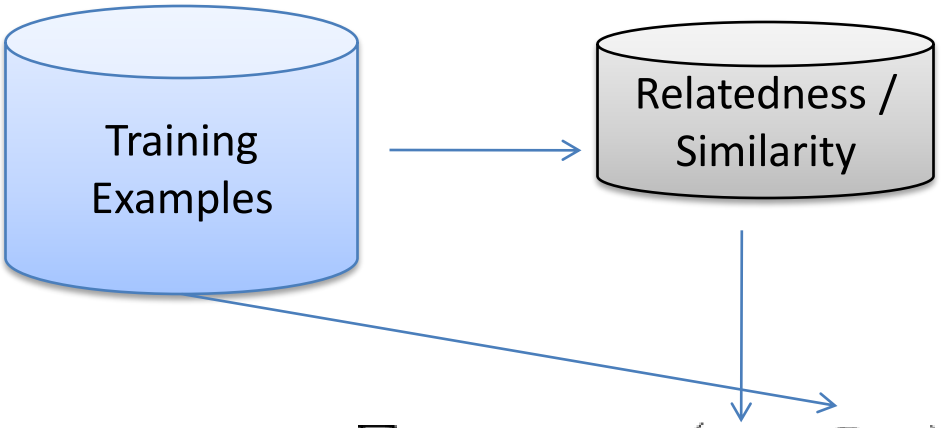
Fruchterman & Reingold. Graph drawing by Force Directed Placement. 1991.

Visualization of Relatedness Measure



What’s the big cluster in the center?

Assembling the Model



Training Examples

Relatedness / Similarity

$$P_{u,i} = \frac{\sum_{\text{all similar items, N}} (s_{i,N} * R_{u,N})}{\sum_{\text{all similar items, N}} (|s_{i,N}|)}$$

- Sarwar, et al. Item-based collaborative filtering recommendation algorithms. 2001.

This is similar to the approaches reported by Amazon in 2003, and Tivo in 2004.

- K. Ali and W. van Stam. Tivo: making show recommendations using a distributed collaborative filtering architecture. KDD, pages 394–401. ACM, 2004.
- G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 7(1):76–80, 2003.

Ensemble Learning in Practice: A Look at the Netflix Prize

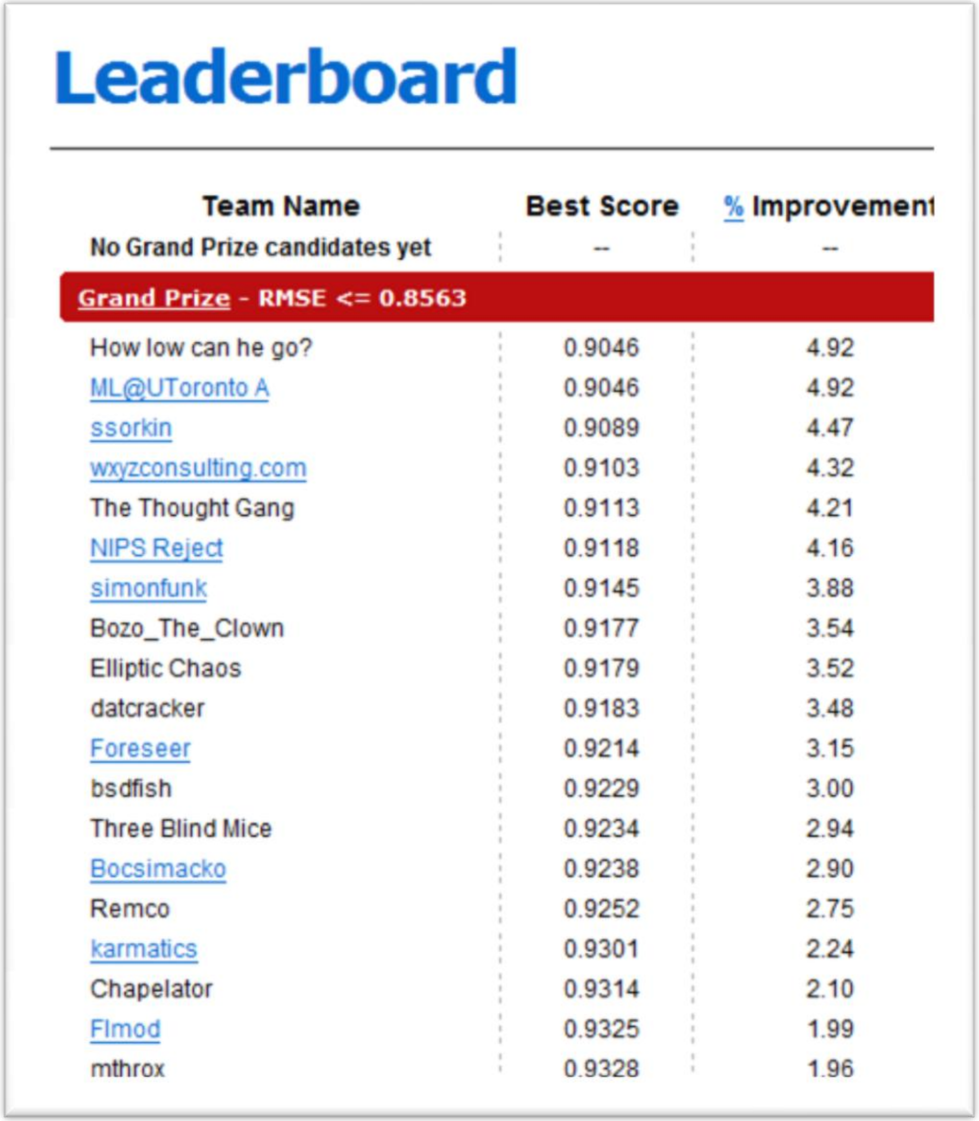
October 2006-present

- Training data is a set of users and ratings (1,2,3,4,5 stars) those users have given to movies.
- Predict what rating a user would give to any movie

- $1 million prize for a 10% improvement over Netflix's current method (MSE = 0.9514)

Just three weeks after it began, at least 40 teams had bested the Netflix method

Top teams showed about 5% improvement



Leaderboard

Team Name	Best Score	% Improvement
No Grand Prize candidates yet	--	--
Grand Prize - RMSE <= 0.8563		
How low can he go?	0.9046	4.92
ML@UToronto A	0.9046	4.92
ssorkin	0.9089	4.47
wxyzconsulting.com	0.9103	4.32
The Thought Gang	0.9113	4.21
NIPS Reject	0.9118	4.16
simonfunk	0.9145	3.88
Bozo_The_Clown	0.9177	3.54
Elliptic Chaos	0.9179	3.52
datcracker	0.9183	3.48
Foreseer	0.9214	3.15
bsdfish	0.9229	3.00
Three Blind Mice	0.9234	2.94
Bocsimacko	0.9238	2.90
Remco	0.9252	2.75
karmatics	0.9301	2.24
Chapelator	0.9314	2.10
Flmod	0.9325	1.99
mthrox	0.9328	1.96

From the Internet Archive.

However, improvement slowed and techniques became more sophisticated…

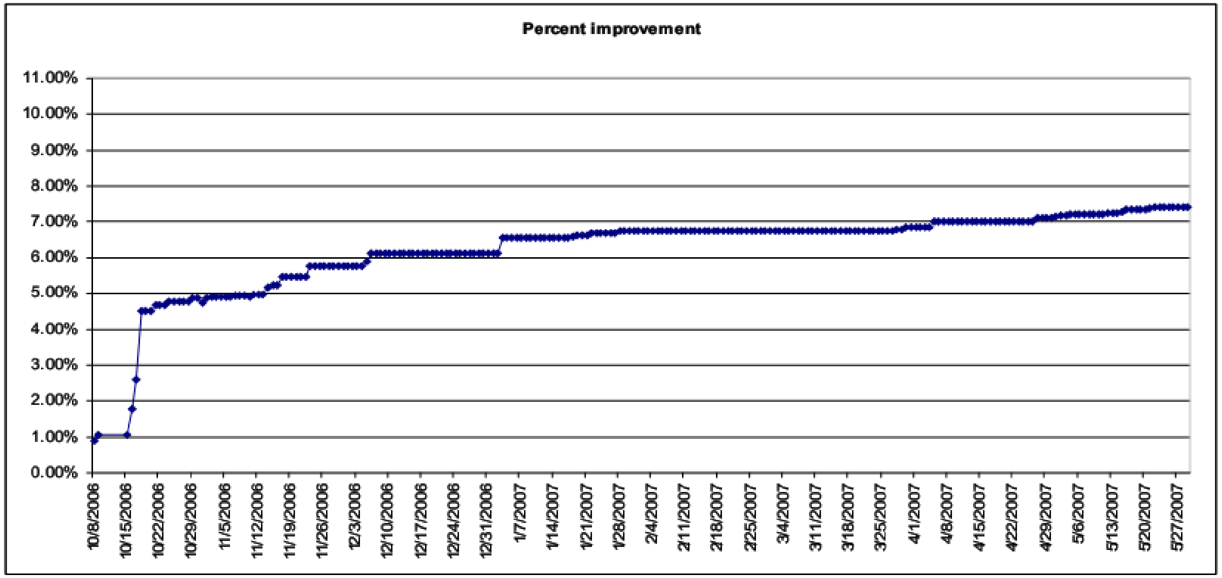


Figure 3: Aggregate improvement over Cinematch by time

Bennett and Lanning. KDCup 2007.

Techniques used…

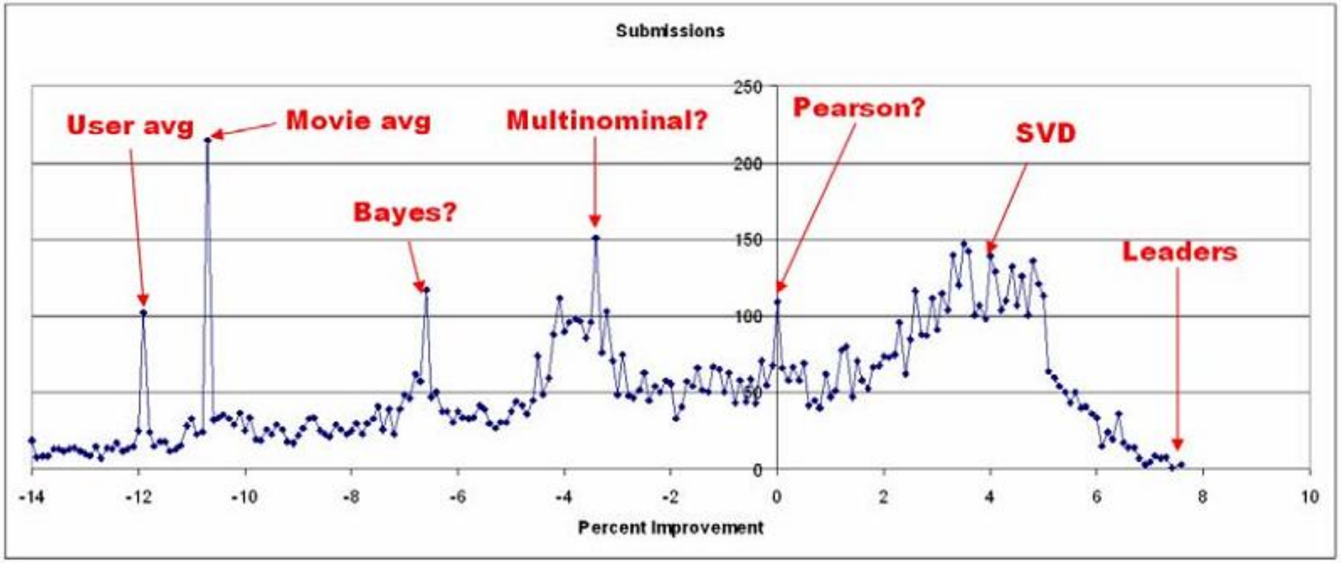
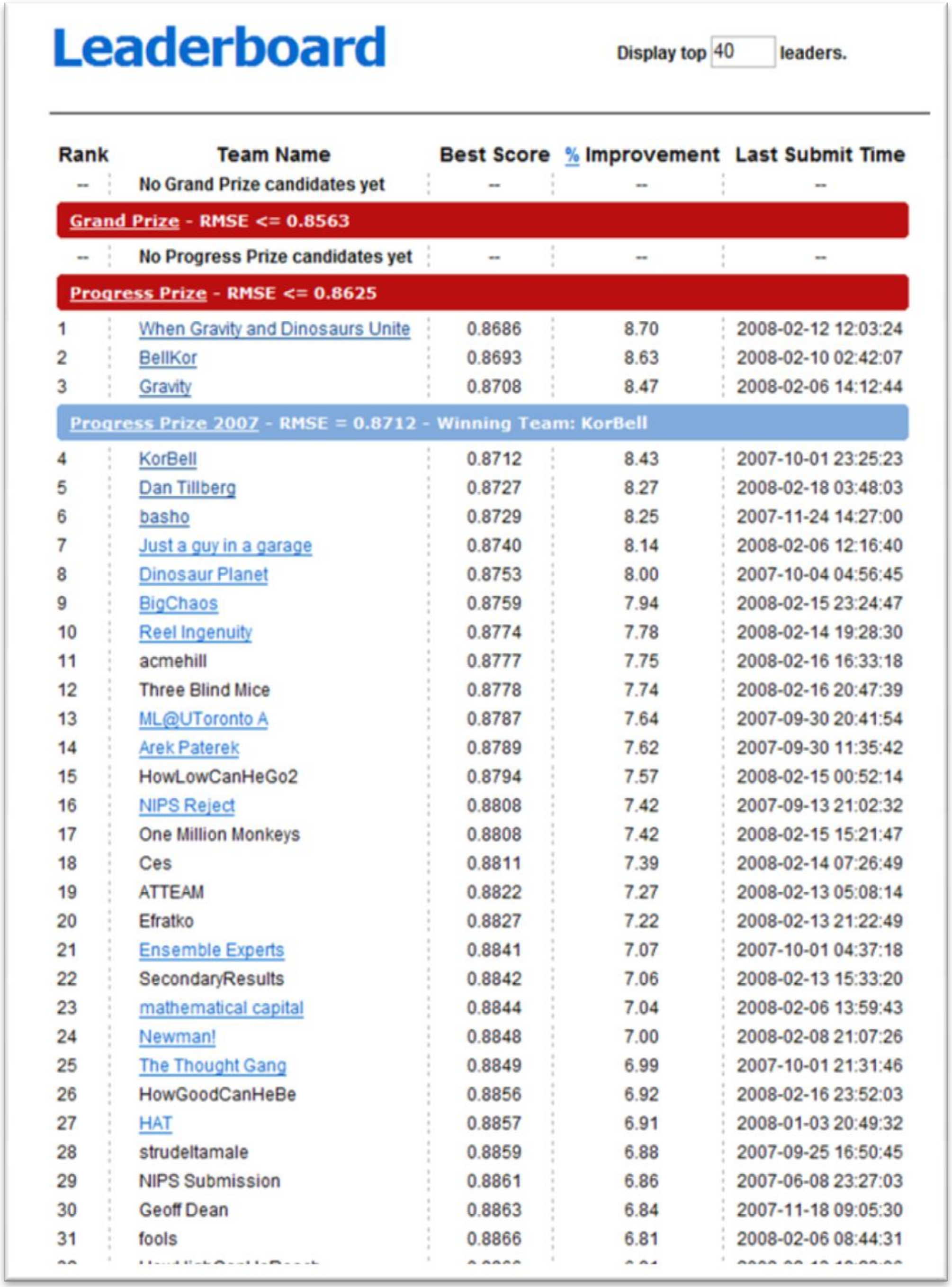


Figure 2: Detail of distribution of leading submissions indicating possible techniques

Bennett and Lanning. KDCup 2007.

Rookies (35)

“Thanks to Paul Harrison's collaboration, a simple mix of our solutions improved our result from 6.31 to 6.75”



Leaderboard

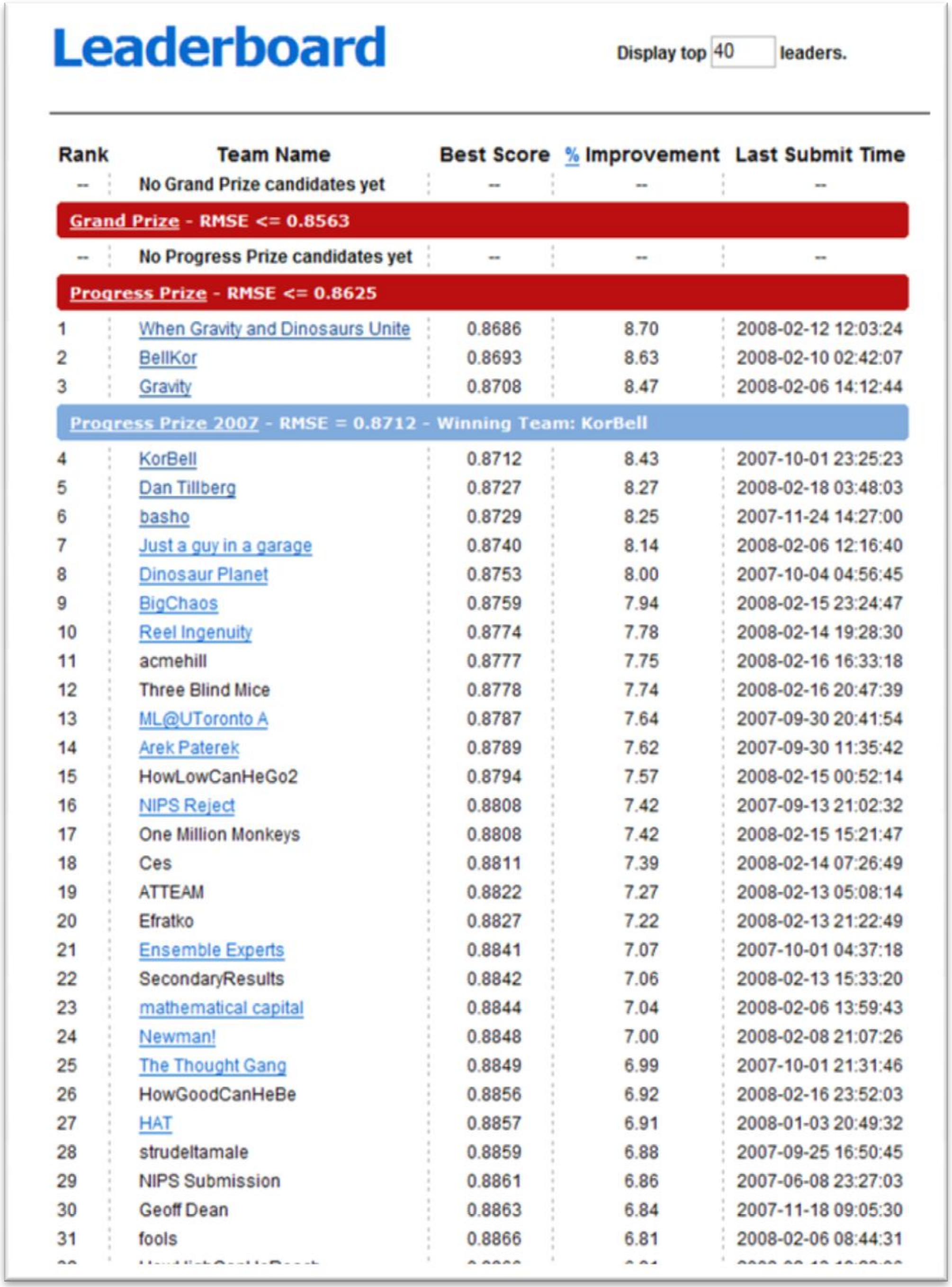
Display top 40 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
--	No Progress Prize candidates yet	--	--	--
Progress Prize - RMSE <= 0.8625				
1	When Gravity and Dinosaurs Unite	0.8686	8.70	2008-02-12 12:03:24
2	BellKor	0.8693	8.63	2008-02-10 02:42:07
3	Gravity	0.8708	8.47	2008-02-06 14:12:44
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell				
4	KorBell	0.8712	8.43	2007-10-01 23:25:23
5	Dan Tillberg	0.8727	8.27	2008-02-18 03:48:03
6	basho	0.8729	8.25	2007-11-24 14:27:00
7	Just a guy in a garage	0.8740	8.14	2008-02-06 12:16:40
8	Dinosaur Planet	0.8753	8.00	2007-10-04 04:56:45
9	BigChaos	0.8759	7.94	2008-02-15 23:24:47
10	Reel Ingenuity	0.8774	7.78	2008-02-14 19:28:30
11	acmehill	0.8777	7.75	2008-02-16 16:33:18
12	Three Blind Mice	0.8778	7.74	2008-02-16 20:47:39
13	ML@UToronto A	0.8787	7.64	2007-09-30 20:41:54
14	Arek Paterek	0.8789	7.62	2007-09-30 11:35:42
15	HowLowCanHeGo2	0.8794	7.57	2008-02-15 00:52:14
16	NIPS Reject	0.8808	7.42	2007-09-13 21:02:32
17	One Million Monkeys	0.8808	7.42	2008-02-15 15:21:47
18	Ces	0.8811	7.39	2008-02-14 07:26:49
19	ATTEAM	0.8822	7.27	2008-02-13 05:08:14
20	Efratko	0.8827	7.22	2008-02-13 21:22:49
21	Ensemble Experts	0.8841	7.07	2007-10-01 04:37:18
22	SecondaryResults	0.8842	7.06	2008-02-13 15:33:20
23	mathematical capital	0.8844	7.04	2008-02-06 13:59:43
24	Newman!	0.8848	7.00	2008-02-08 21:07:26
25	The Thought Gang	0.8849	6.99	2007-10-01 21:31:46
26	HowGoodCanHeBe	0.8856	6.92	2008-02-16 23:52:03
27	HAT	0.8857	6.91	2008-01-03 20:49:32
28	strudeltamale	0.8859	6.88	2007-09-25 16:50:45
29	NIPS Submission	0.8861	6.86	2007-06-08 23:27:03
30	Geoff Dean	0.8863	6.84	2007-11-18 09:05:30
31	fools	0.8866	6.81	2008-02-06 08:44:31

Arek Paterek (15)

“My approach is to **combine the results of many methods** (also two-way interactions between them) using linear regression on the test set. The best method in my ensemble is regularized SVD with biases, post processed with kernel ridge regression”

http://rainbow.mimuw.edu.pl/~ap/ap_kdd.pdf



Leaderboard

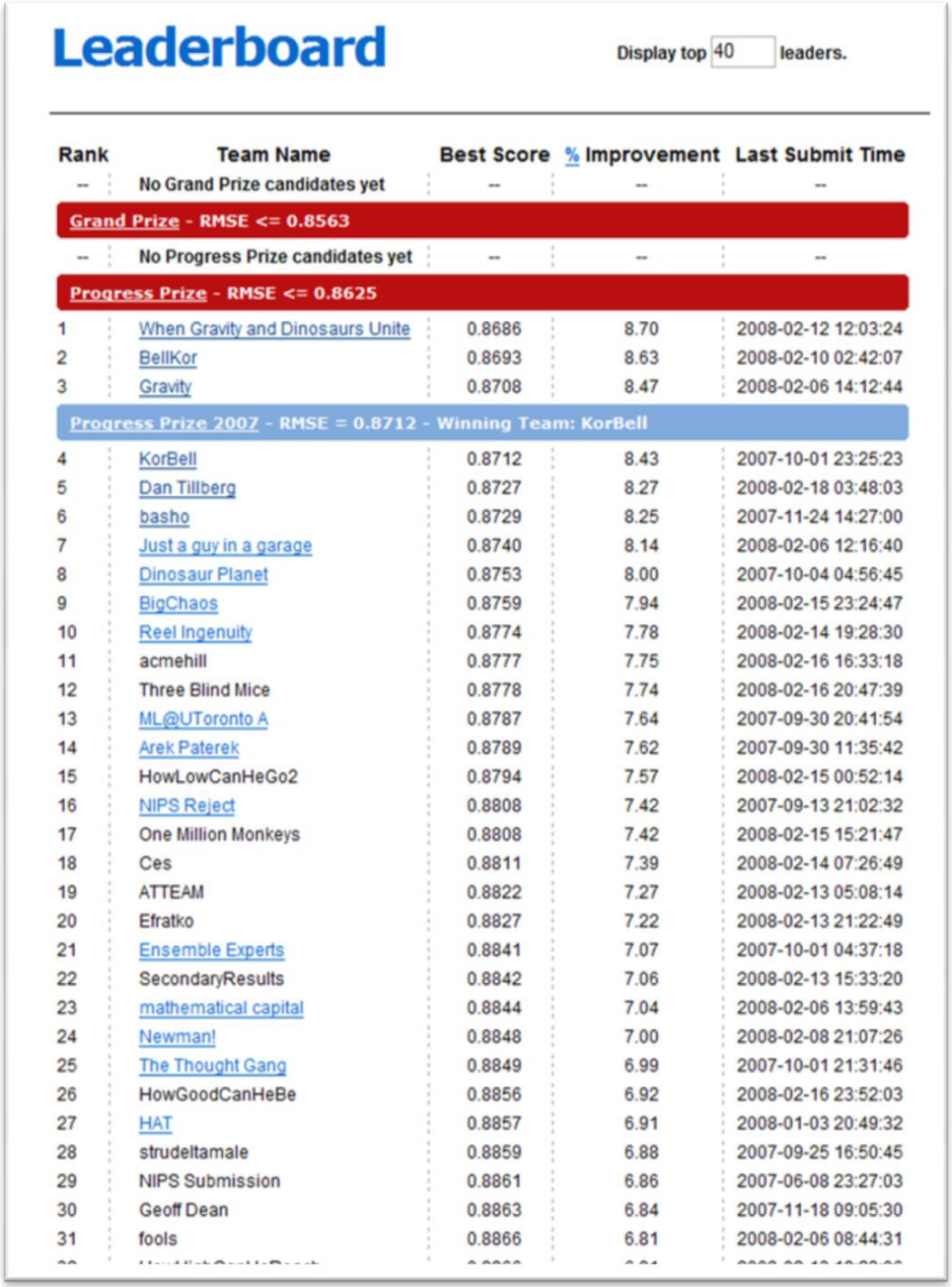
Display top 40 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
--	No Progress Prize candidates yet	--	--	--
Progress Prize - RMSE <= 0.8625				
1	When Gravity and Dinosaurs Unite	0.8686	8.70	2008-02-12 12:03:24
2	BellKor	0.8693	8.63	2008-02-10 02:42:07
3	Gravity	0.8708	8.47	2008-02-06 14:12:44
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell				
4	KorBell	0.8712	8.43	2007-10-01 23:25:23
5	Dan Tillberg	0.8727	8.27	2008-02-18 03:48:03
6	basho	0.8729	8.25	2007-11-24 14:27:00
7	Just a guy in a garage	0.8740	8.14	2008-02-06 12:16:40
8	Dinosaur Planet	0.8753	8.00	2007-10-04 04:56:45
9	BigChaos	0.8759	7.94	2008-02-15 23:24:47
10	Reel Ingenuity	0.8774	7.78	2008-02-14 19:28:30
11	acmehill	0.8777	7.75	2008-02-16 16:33:18
12	Three Blind Mice	0.8778	7.74	2008-02-16 20:47:39
13	ML@UToronto A	0.8787	7.64	2007-09-30 20:41:54
14	Arek Paterek	0.8789	7.62	2007-09-30 11:35:42
15	HowLowCanHeGo2	0.8794	7.57	2008-02-15 00:52:14
16	NIPS Reject	0.8808	7.42	2007-09-13 21:02:32
17	One Million Monkeys	0.8808	7.42	2008-02-15 15:21:47
18	Ces	0.8811	7.39	2008-02-14 07:26:49
19	ATTEAM	0.8822	7.27	2008-02-13 05:08:14
20	Efratko	0.8827	7.22	2008-02-13 21:22:49
21	Ensemble Experts	0.8841	7.07	2007-10-01 04:37:18
22	SecondaryResults	0.8842	7.06	2008-02-13 15:33:20
23	mathematical capital	0.8844	7.04	2008-02-06 13:59:43
24	Newman!	0.8848	7.00	2008-02-08 21:07:26
25	The Thought Gang	0.8849	6.99	2007-10-01 21:31:46
26	HowGoodCanHeBe	0.8856	6.92	2008-02-16 23:52:03
27	HAT	0.8857	6.91	2008-01-03 20:49:32
28	strudeltamale	0.8859	6.88	2007-09-25 16:50:45
29	NIPS Submission	0.8861	6.86	2007-06-08 23:27:03
30	Geoff Dean	0.8863	6.84	2007-11-18 09:05:30
31	fools	0.8866	6.81	2008-02-06 08:44:31

U of Toronto (13)

“When the predictions of **multiple** RBM models and **multiple** SVD models are linearly combined, we achieve an error rate that is well over 6% better than the score of Netflix’s own system.”

http://www.cs.toronto.edu/~rsalakhu/papers/rbmcf.pdf



Leaderboard

Display top 40 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
--	No Progress Prize candidates yet	--	--	--
Progress Prize - RMSE <= 0.8625				
1	When Gravity and Dinosaurs Unite	0.8686	8.70	2008-02-12 12:03:24
2	BellKor	0.8693	8.63	2008-02-10 02:42:07
3	Gravity	0.8708	8.47	2008-02-06 14:12:44
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell				
4	KorBell	0.8712	8.43	2007-10-01 23:25:23
5	Dan Tillberg	0.8727	8.27	2008-02-18 03:48:03
6	basho	0.8729	8.25	2007-11-24 14:27:00
7	Just a guy in a garage	0.8740	8.14	2008-02-06 12:16:40
8	Dinosaur Planet	0.8753	8.00	2007-10-04 04:56:45
9	BigChaos	0.8759	7.94	2008-02-15 23:24:47
10	Reel Ingenuity	0.8774	7.78	2008-02-14 19:28:30
11	acmehill	0.8777	7.75	2008-02-16 16:33:18
12	Three Blind Mice	0.8778	7.74	2008-02-16 20:47:39
13	ML@UToronto A	0.8787	7.64	2007-09-30 20:41:54
14	Arek Paterek	0.8789	7.62	2007-09-30 11:35:42
15	HowLowCanHeGo2	0.8794	7.57	2008-02-15 00:52:14
16	NIPS Reject	0.8808	7.42	2007-09-13 21:02:32
17	One Million Monkeys	0.8808	7.42	2008-02-15 15:21:47
18	Ces	0.8811	7.39	2008-02-14 07:26:49
19	ATTEAM	0.8822	7.27	2008-02-13 05:08:14
20	Efratko	0.8827	7.22	2008-02-13 21:22:49
21	Ensemble Experts	0.8841	7.07	2007-10-01 04:37:18
22	SecondaryResults	0.8842	7.06	2008-02-13 15:33:20
23	mathematical capital	0.8844	7.04	2008-02-06 13:59:43
24	Newman!	0.8848	7.00	2008-02-08 21:07:26
25	The Thought Gang	0.8849	6.99	2007-10-01 21:31:46
26	HowGoodCanHeBe	0.8856	6.92	2008-02-16 23:52:03
27	HAT	0.8857	6.91	2008-01-03 20:49:32
28	strudeltamale	0.8859	6.88	2007-09-25 16:50:45
29	NIPS Submission	0.8861	6.86	2007-06-08 23:27:03
30	Geoff Dean	0.8863	6.84	2007-11-18 09:05:30
31	fools	0.8866	6.81	2008-02-06 08:44:31

Gravity (3)

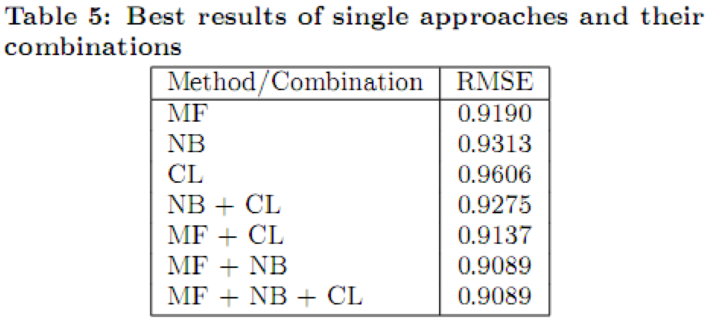
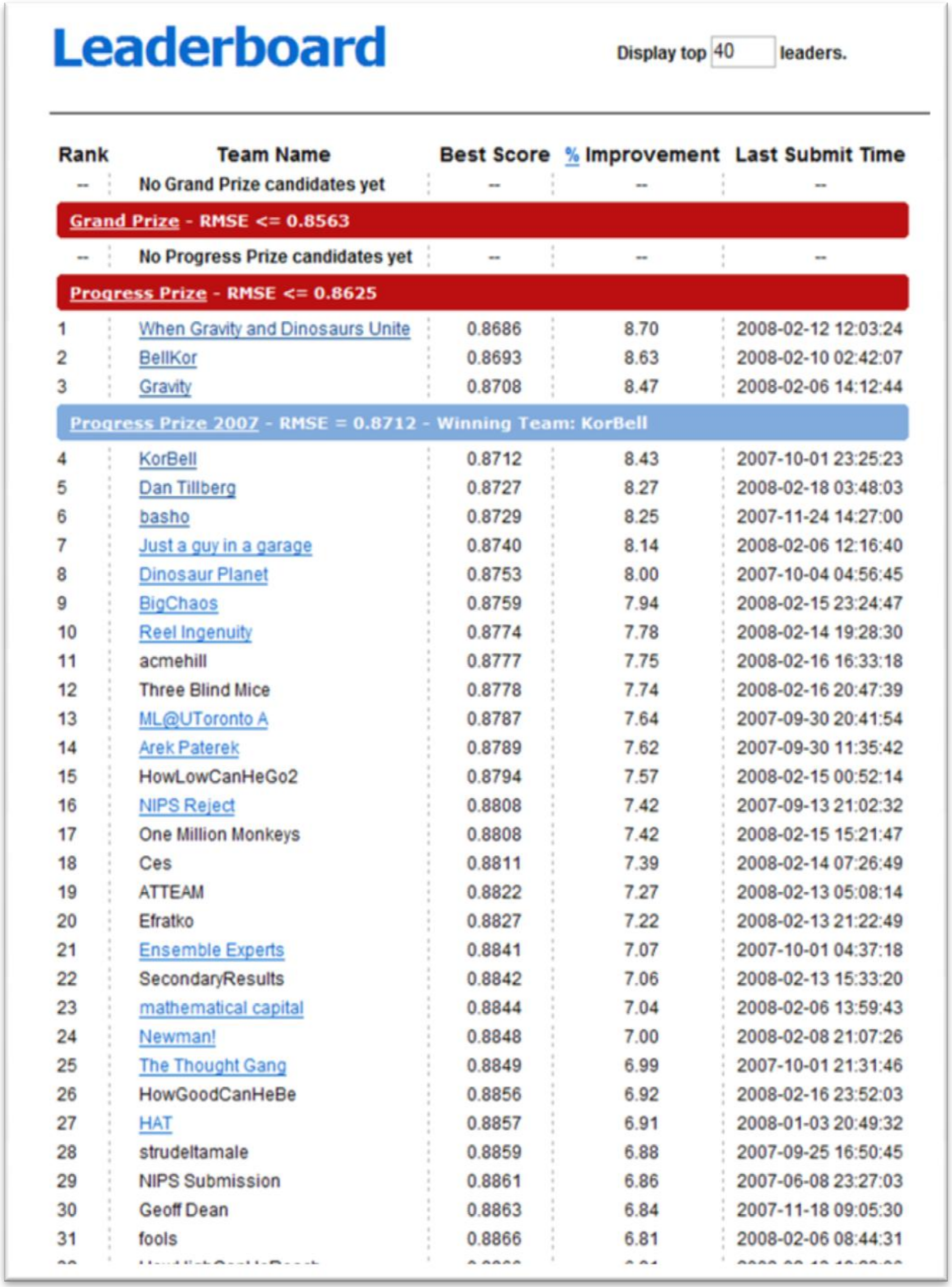


Table 5: Best results of single approaches and their combinations

Method/Combination	RMSE
MF	0.9190
NB	0.9313
CL	0.9606
NB + CL	0.9275
MF + CL	0.9137
MF + NB	0.9089
MF + NB + CL	0.9089

home.mit.bme.hu/~gtakacs/download/gravity.pdf



Leaderboard

Display top 40 leaders.

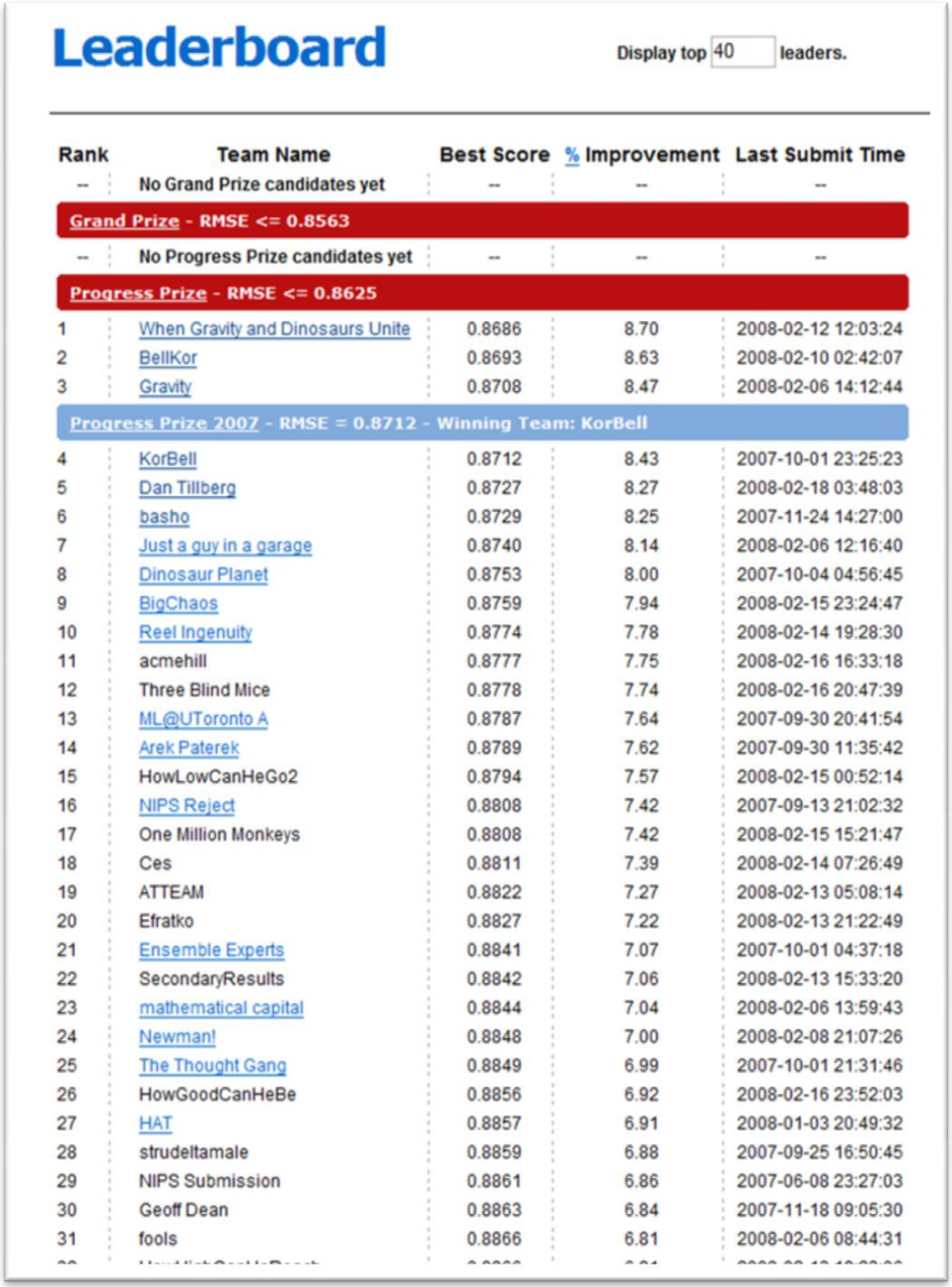
Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
--	No Progress Prize candidates yet	--	--	--
Progress Prize - RMSE <= 0.8625				
1	When Gravity and Dinosaurs Unite	0.8686	8.70	2008-02-12 12:03:24
2	BellKor	0.8693	8.63	2008-02-10 02:42:07
3	Gravity	0.8708	8.47	2008-02-06 14:12:44
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell				
4	KorBell	0.8712	8.43	2007-10-01 23:25:23
5	Dan Tillberg	0.8727	8.27	2008-02-18 03:48:03
6	basho	0.8729	8.25	2007-11-24 14:27:00
7	Just a guy in a garage	0.8740	8.14	2008-02-06 12:16:40
8	Dinosaur Planet	0.8753	8.00	2007-10-04 04:56:45
9	BigChaos	0.8759	7.94	2008-02-15 23:24:47
10	Reel Ingenuity	0.8774	7.78	2008-02-14 19:28:30
11	acmehill	0.8777	7.75	2008-02-16 16:33:18
12	Three Blind Mice	0.8778	7.74	2008-02-16 20:47:39
13	ML@UToronto A	0.8787	7.64	2007-09-30 20:41:54
14	Arek Paterek	0.8789	7.62	2007-09-30 11:35:42
15	HowLowCanHeGo2	0.8794	7.57	2008-02-15 00:52:14
16	NIPS Reject	0.8808	7.42	2007-09-13 21:02:32
17	One Million Monkeys	0.8808	7.42	2008-02-15 15:21:47
18	Ces	0.8811	7.39	2008-02-14 07:26:49
19	ATTEAM	0.8822	7.27	2008-02-13 05:08:14
20	Efratko	0.8827	7.22	2008-02-13 21:22:49
21	Ensemble Experts	0.8841	7.07	2007-10-01 04:37:18
22	SecondaryResults	0.8842	7.06	2008-02-13 15:33:20
23	mathematical capital	0.8844	7.04	2008-02-06 13:59:43
24	Newman!	0.8848	7.00	2008-02-08 21:07:26
25	The Thought Gang	0.8849	6.99	2007-10-01 21:31:46
26	HowGoodCanHeBe	0.8856	6.92	2008-02-16 23:52:03
27	HAT	0.8857	6.91	2008-01-03 20:49:32
28	strudeltamale	0.8859	6.88	2007-09-25 16:50:45
29	NIPS Submission	0.8861	6.86	2007-06-08 23:27:03
30	Geoff Dean	0.8863	6.84	2007-11-18 09:05:30
31	fools	0.8866	6.81	2008-02-06 08:44:31

BellKor (2)

“Predictive accuracy is substantially improved when blending multiple predictors. Our experience is that most efforts should be concentrated in deriving substantially different approaches, rather than refining a single technique. Consequently, our solution is an ensemble of many methods. “

“Our final solution (RMSE=0.8712) consists of blending 107 individual results. “

http://www.research.att.com/~volinsky/netflix/ProgressPrize2007BellKorSolution.pdf



Leaderboard

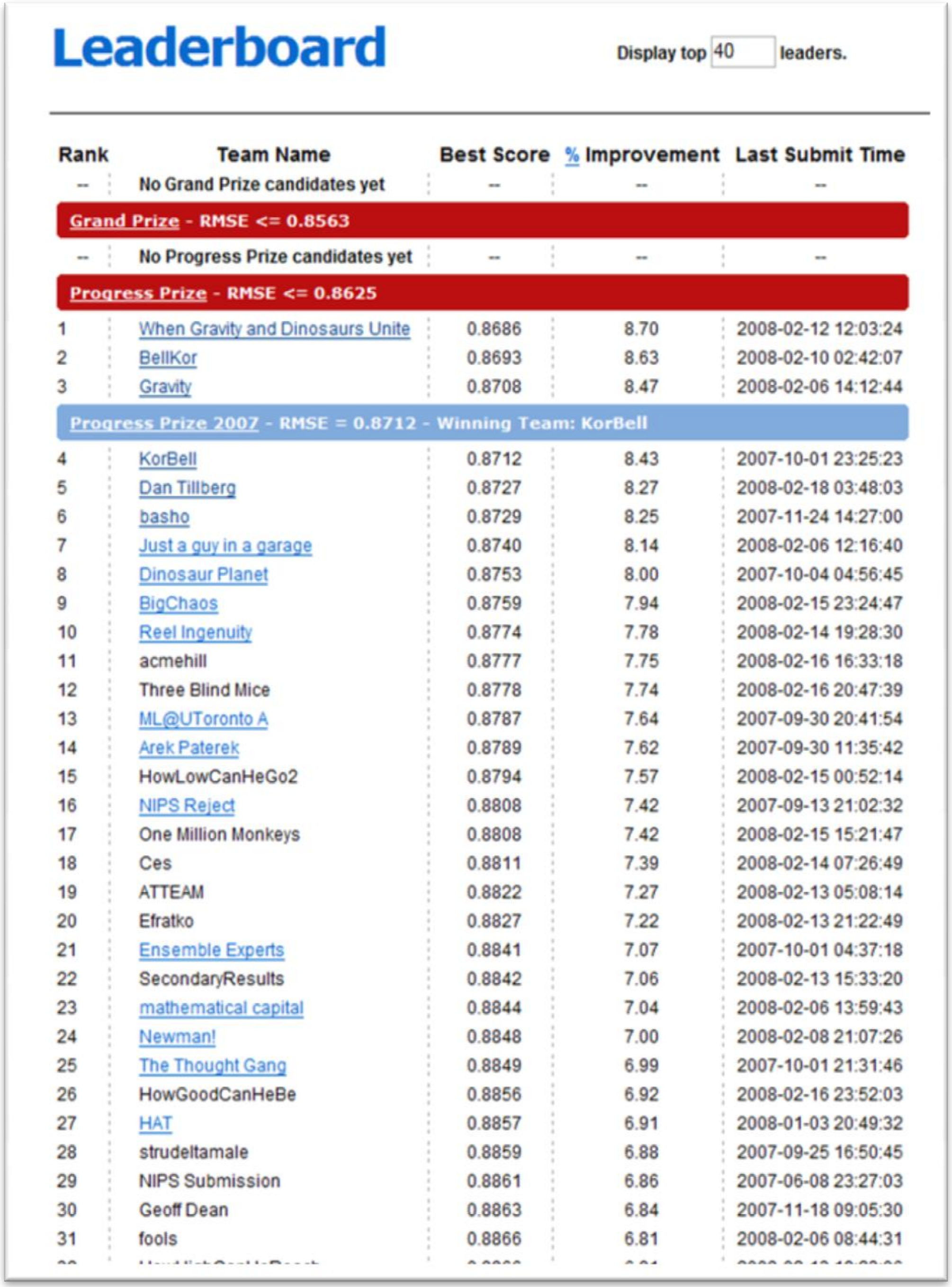
Display top 40 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
--	No Progress Prize candidates yet	--	--	--
Progress Prize - RMSE <= 0.8625				
1	When Gravity and Dinosaurs Unite	0.8686	8.70	2008-02-12 12:03:24
2	BellKor	0.8693	8.63	2008-02-10 02:42:07
3	Gravity	0.8708	8.47	2008-02-06 14:12:44
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell				
4	KorBell	0.8712	8.43	2007-10-01 23:25:23
5	Dan Tillberg	0.8727	8.27	2008-02-18 03:48:03
6	basho	0.8729	8.25	2007-11-24 14:27:00
7	Just a guy in a garage	0.8740	8.14	2008-02-06 12:16:40
8	Dinosaur Planet	0.8753	8.00	2007-10-04 04:56:45
9	BigChaos	0.8759	7.94	2008-02-15 23:24:47
10	Reel Ingenuity	0.8774	7.78	2008-02-14 19:28:30
11	acmehill	0.8777	7.75	2008-02-16 16:33:18
12	Three Blind Mice	0.8778	7.74	2008-02-16 20:47:39
13	ML@UToronto A	0.8787	7.64	2007-09-30 20:41:54
14	Arek Paterek	0.8789	7.62	2007-09-30 11:35:42
15	HowLowCanHeGo2	0.8794	7.57	2008-02-15 00:52:14
16	NIPS Reject	0.8808	7.42	2007-09-13 21:02:32
17	One Million Monkeys	0.8808	7.42	2008-02-15 15:21:47
18	Ces	0.8811	7.39	2008-02-14 07:26:49
19	ATTEAM	0.8822	7.27	2008-02-13 05:08:14
20	Efratko	0.8827	7.22	2008-02-13 21:22:49
21	Ensemble Experts	0.8841	7.07	2007-10-01 04:37:18
22	SecondaryResults	0.8842	7.06	2008-02-13 15:33:20
23	mathematical capital	0.8844	7.04	2008-02-06 13:59:43
24	Newman!	0.8848	7.00	2008-02-08 21:07:26
25	The Thought Gang	0.8849	6.99	2007-10-01 21:31:46
26	HowGoodCanHeBe	0.8856	6.92	2008-02-16 23:52:03
27	HAT	0.8857	6.91	2008-01-03 20:49:32
28	strudeltamale	0.8859	6.88	2007-09-25 16:50:45
29	NIPS Submission	0.8861	6.86	2007-06-08 23:27:03
30	Geoff Dean	0.8863	6.84	2007-11-18 09:05:30
31	fools	0.8866	6.81	2008-02-06 08:44:31

When Gravity and Dinosaurs Unite (1)

“Our common team blends the result of team Gravity and team Dinosaur Planet.”

Might have guessed from the name…



Leaderboard

Display top 40 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
--	No Progress Prize candidates yet	--	--	--
Progress Prize - RMSE <= 0.8625				
1	When Gravity and Dinosaurs Unite	0.8686	8.70	2008-02-12 12:03:24
2	BellKor	0.8693	8.63	2008-02-10 02:42:07
3	Gravity	0.8708	8.47	2008-02-06 14:12:44
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell				
4	KorBell	0.8712	8.43	2007-10-01 23:25:23
5	Dan Tillberg	0.8727	8.27	2008-02-18 03:48:03
6	basho	0.8729	8.25	2007-11-24 14:27:00
7	Just a guy in a garage	0.8740	8.14	2008-02-06 12:16:40
8	Dinosaur Planet	0.8753	8.00	2007-10-04 04:56:45
9	BigChaos	0.8759	7.94	2008-02-15 23:24:47
10	Reel Ingenuity	0.8774	7.78	2008-02-14 19:28:30
11	acmehill	0.8777	7.75	2008-02-16 16:33:18
12	Three Blind Mice	0.8778	7.74	2008-02-16 20:47:39
13	ML@UToronto A	0.8787	7.64	2007-09-30 20:41:54
14	Arek Paterek	0.8789	7.62	2007-09-30 11:35:42
15	HowLowCanHeGo2	0.8794	7.57	2008-02-15 00:52:14
16	NIPS Reject	0.8808	7.42	2007-09-13 21:02:32
17	One Million Monkeys	0.8808	7.42	2008-02-15 15:21:47
18	Ces	0.8811	7.39	2008-02-14 07:26:49
19	ATTEAM	0.8822	7.27	2008-02-13 05:08:14
20	Efratko	0.8827	7.22	2008-02-13 21:22:49
21	Ensemble Experts	0.8841	7.07	2007-10-01 04:37:18
22	SecondaryResults	0.8842	7.06	2008-02-13 15:33:20
23	mathematical capital	0.8844	7.04	2008-02-06 13:59:43
24	Newman!	0.8848	7.00	2008-02-08 21:07:26
25	The Thought Gang	0.8849	6.99	2007-10-01 21:31:46
26	HowGoodCanHeBe	0.8856	6.92	2008-02-16 23:52:03
27	HAT	0.8857	6.91	2008-01-03 20:49:32
28	strudeltamale	0.8859	6.88	2007-09-25 16:50:45
29	NIPS Submission	0.8861	6.86	2007-06-08 23:27:03
30	Geoff Dean	0.8863	6.84	2007-11-18 09:05:30
31	fools	0.8866	6.81	2008-02-06 08:44:31

Why combine models?

Diversity in Decision Making

- Utility of combining diverse, independent outcomes in human decision-making
  - Expert panels
  - Protective Mechanism (e.g. stock portfolio diversity)

- Suppose we have 5 completely independent decision makers…
  - If accuracy is 70% for each
    - $10\,(.7^3)(.3^2)+5(.7^4)(.3)+(.7^5)$
    - **83.7% majority vote accuracy**
  - 101 such classifiers
    - **99.9% majority vote accuracy**

A Reflection

- Combining models adds complexity
  - More difficult to characterize, anticipate predictions, explain predictions, etc.
  - But accuracy may increase

- **Violation of Ockham's Razor**
  - "simplicity leads to greater accuracy"
  - Identifying the best model requires identifying the proper "model complexity"

See Domingos, P. Occam's two razors: the sharp and the blunt. KDD. 1998.

Achieving Diversity

Diversity from different algorithms, or algorithm parameters

(as we’ve seen with the Netflix Prize leaders)

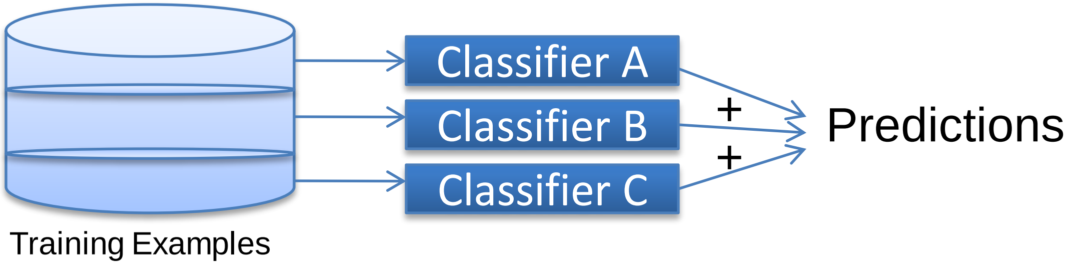
Examples

- 5 neighbor-based models with different relatedness measures
- 1 neighbor model + 1 Bayesian model

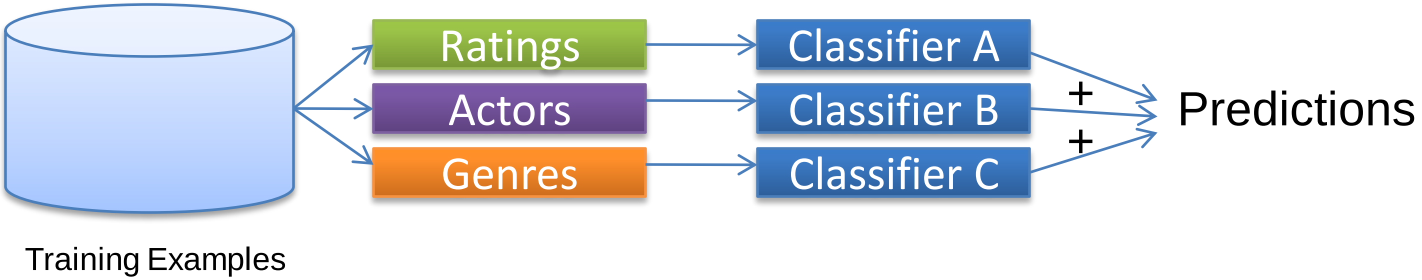
Achieving Diversity

Diversity from differences in inputs

1. Divide up training data among models



2. Different feature weightings



Two Particular Strategies

<u>Bagging</u>

- Use different subsets of the training data for each model

<u>Boosting</u>

- With each additional model, make misclassified examples more important (or less, in some cases)

Bagging Diversity

- Requirement: Need **unstable** classifier types
  - Unstable means a small change to the training data may lead to major decision changes.
  - Is the neighbor approach unstable? No, but many other types are.

Bagging Algorithm

For 1 to k,

1. Take a bootstrap sample of the training examples
2. Build a model using sample
3. Add model to ensemble

To make a prediction, run each model in the ensemble, and use the majority prediction.

Boosting

Incrementally create models using selectively using training examples based on some distribution.

AdaBoost (Adaptive Boosting) Algorithm

1. Initialize Weights
2. Construct a model. Compute the error.
3. Update the weights to reflect misclassified examples, and repeat step 2.
4. Finally, sum hypotheses…

AdaBoost Cont.

- Advantage
  - Very little code

- Disadvantage
  - Sensitive to noise and outliers. Why?

Recap

- Supervised learning
  - Learning from training data
  - Many challenges
- Ensembles
  - Diversity helps
  - Designing for diversity
    - Bagging
    - Boosting

Further Information…

Books

1. Kunchera, Ludmila. Combining Pattern Classifiers. Wiley. 2004.
2. Bishop, Christopher M. Pattern Recognition and Machine Learning. Springer. 2006.

Video

1. Mease, David. Statistical Aspects of Data Mining. http://video.google.com/videoplay?docid=-4669216290304603251&q=stats+202+engEDU&total=13&start=0&num=10&so=0&type=search&plindex=8
2. Modern Classifier Design. http://video.google.com/videoplay?docid=7691179100757584144&q=classifier&total=172&start=0&num=10&so=0&type=search&plindex=3

Artcles

1. Dietterich, T. G. Ensemble Learning. In The Handbook of Brain Theory and Neural Networks, Second edition, (M.A. Arbib, Ed.), Cambridge, MA: The MIT Press, 2002.
2. Elder, John and Seni Giovanni. From Trees to Forests and Rule Sets - A Unified Overview of Ensemble Methods. KDD 2007 http://Tutorial. videolectures.net/kdd07_elder_ftfr/
3. Polikar, Robi. Ensemble Based Systems in Decision Making. IEEE Circuits and Systems Magazine. 2006.
4. Takacs, et al. On the Gravity Recommendation System. KDD Cup Workshop at SIGKDD. 2007.