Managing/analyzing the Netflix data

Russ Lenth

Department of Statistics & Actuarial Science
The University of Iowa

22S:295 HPC Seminar
October 25, 2007

The Netflix prize

- For details: `www.netflixprize.com`
- $1 million prize for beating *Cinematch* program for predicting movie ratings by 10%
- Annual progress prize of $50K.
- Cinematch RMSE is 0.9525; $1M goal 0.8572
- Contest begins October 2, 2006 and continues through at least October 2, 2011
- Current leaders (as of Oct. 19): "BellKor" team (Bob Bell, Yehudi Koren, AT&T Research), RMSE = 0.8709

The data

- Training data variables: Movie ID, Customer ID, Date, Rating (1–5)
- About 18,000 movies, 480,000 customers, and over 100 million observations
- Packaged as 17,770 separate text files, one for each movie
- These files are saved (gzip format) and available to all in `/space/yoyo/data/Netflix/training_data`

mv_0012345.txt

```
0012345:
0365262 5 2005-05-04
1076294 3 2005-03-07
.   .   .
2209921 4 2006-12-23
```