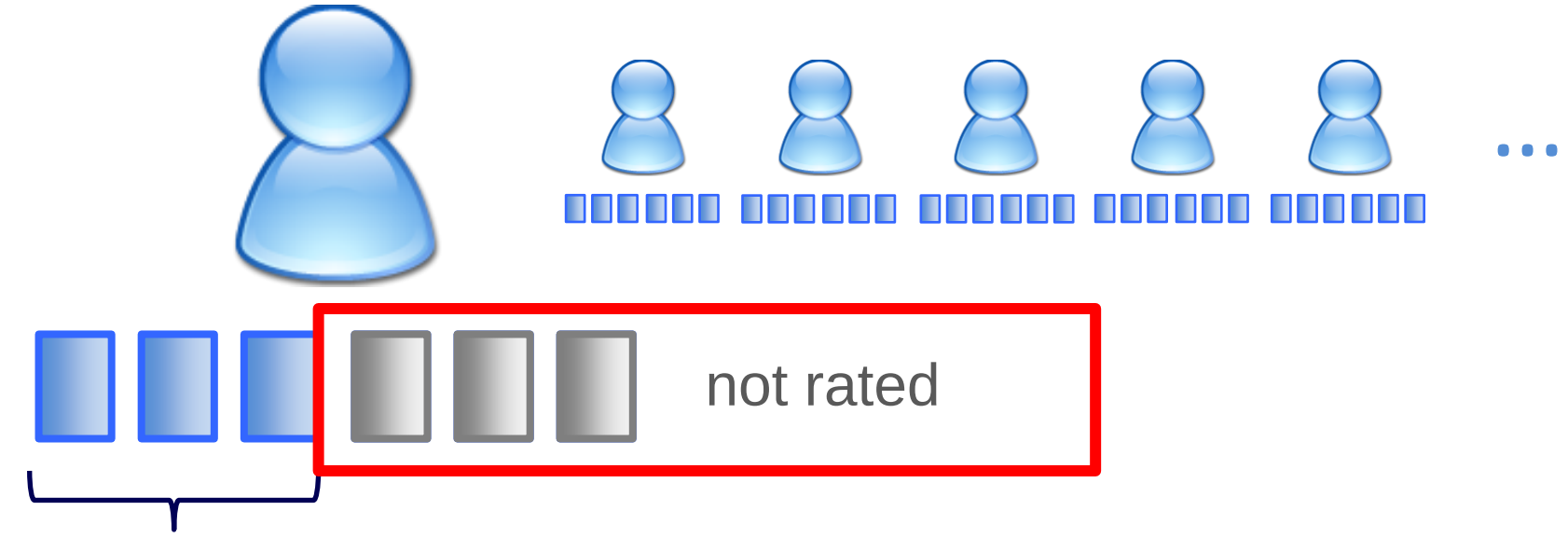
Track 2

Sebastian Stange,
Martin Köppelmann,
Caroline Fetzer

Track 2 – Assignment 2011

#User	#Items	#Ratings	#Ratings to predict
249.012	296.111	61.944.406	607.032



First Results

- random predictions
  -> **49,9723%**, as expected
- base prediction = number of ratings for track
  -> **42,8856%**

Results till now

10.127% error rate

How did we manage that?

MACHINE LEARNING!

Strategy

- Create own sample as subset of the training set
- Chose relevant attributes by hand
- train classifier on sample
- create prediction for each user-track-pair in the yahoo test set

Sampling own testsets

As a reminder:

- Our base prediction was **42,8856%** predicting ratings for yahoo sample

Now using our sampling:

- chosen unrated tracks randomly
  -> base prediction resulted in **4.485%** error rate

- users chosen from yahoo sample; tracks chosen proportional to their high rating count
  -> base prediction resulted in **41.524%** error rate

Sampling own testsets finished

Next Step

- define attributes
- have a deeper look at the provided data
  - generate some statistics using attributes

Attributes

user attributes
number of ratings
number of rated genres out of the 50 most rated genres / 50
number of high rated genres out of the 50 most rated genres / 50
average rating of the user
RMSE of users ratings

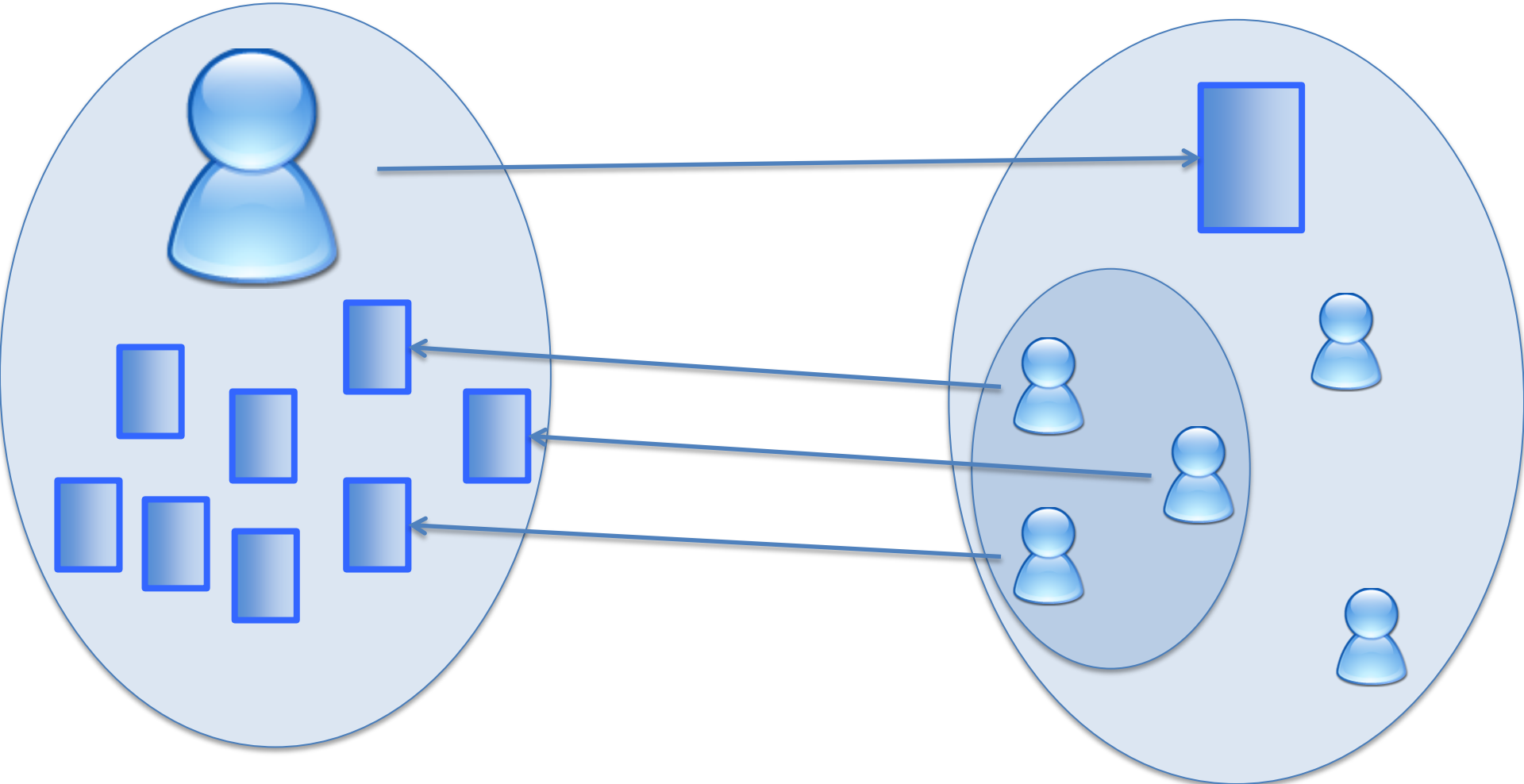
Attributes

track attributes
number of ratings
number of ratings >= 80
number of genres that are within the 50 most rated genres / number of genres
number of ratings missing to 20
6 more....

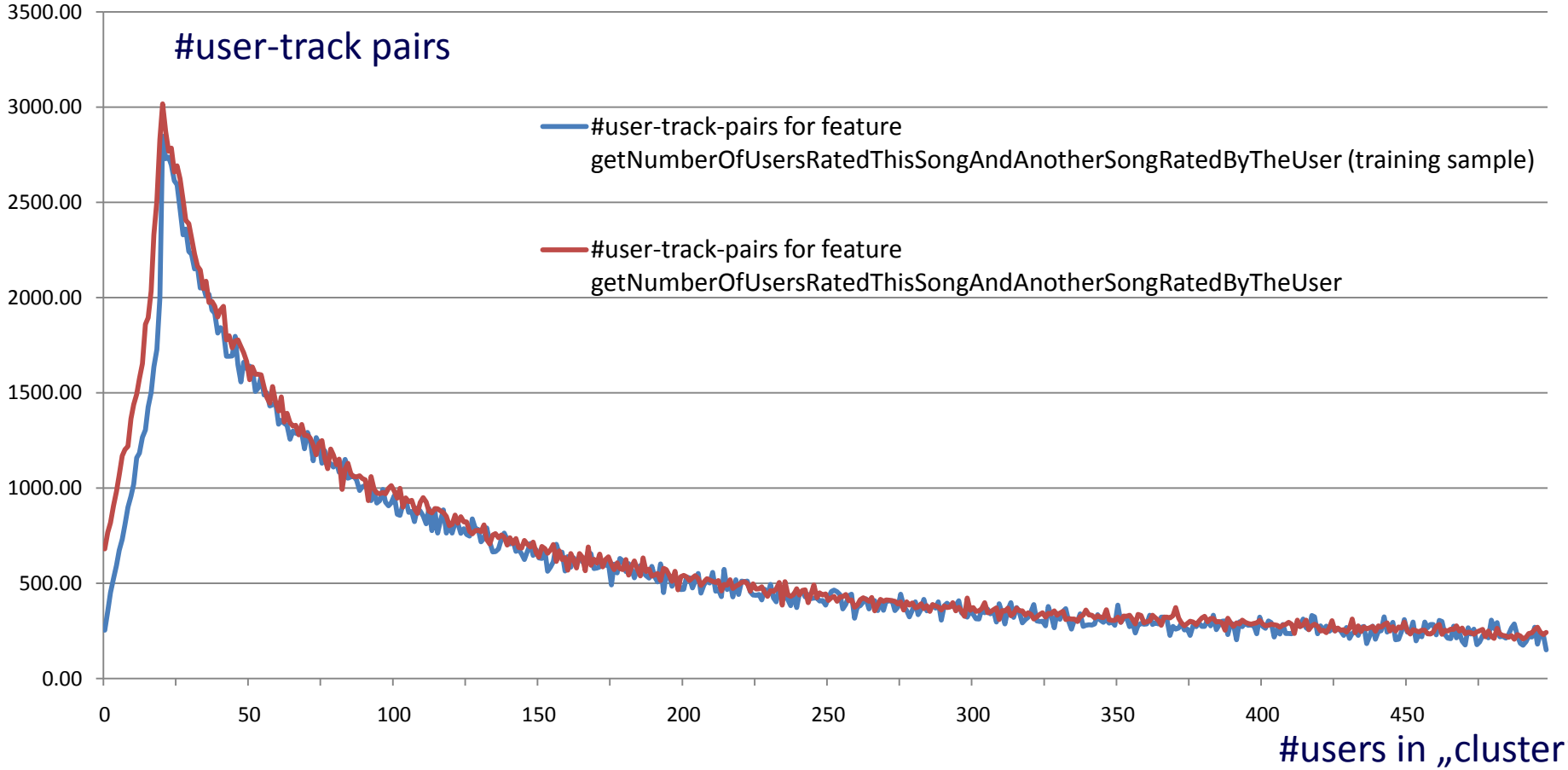
Attributes

user/item attributes:
number of tracks rated of the user from the same album
number of tracks rated of the user from the same artist / #tracks of this artist
rating for genre of track
number of users rated this song and another song rated from the given user (CF)
6 more….

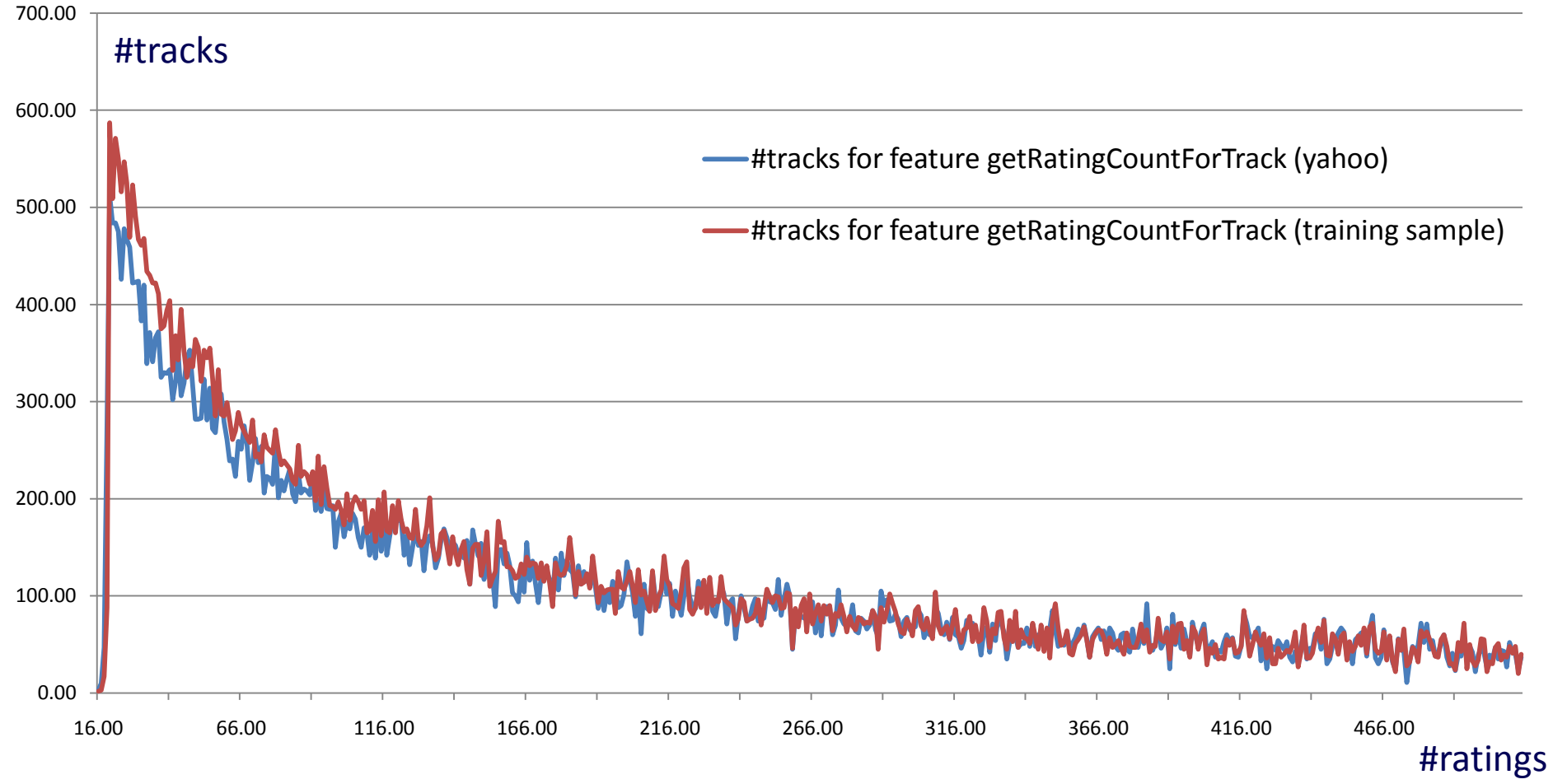
4 Attributes – CF attribute in detail



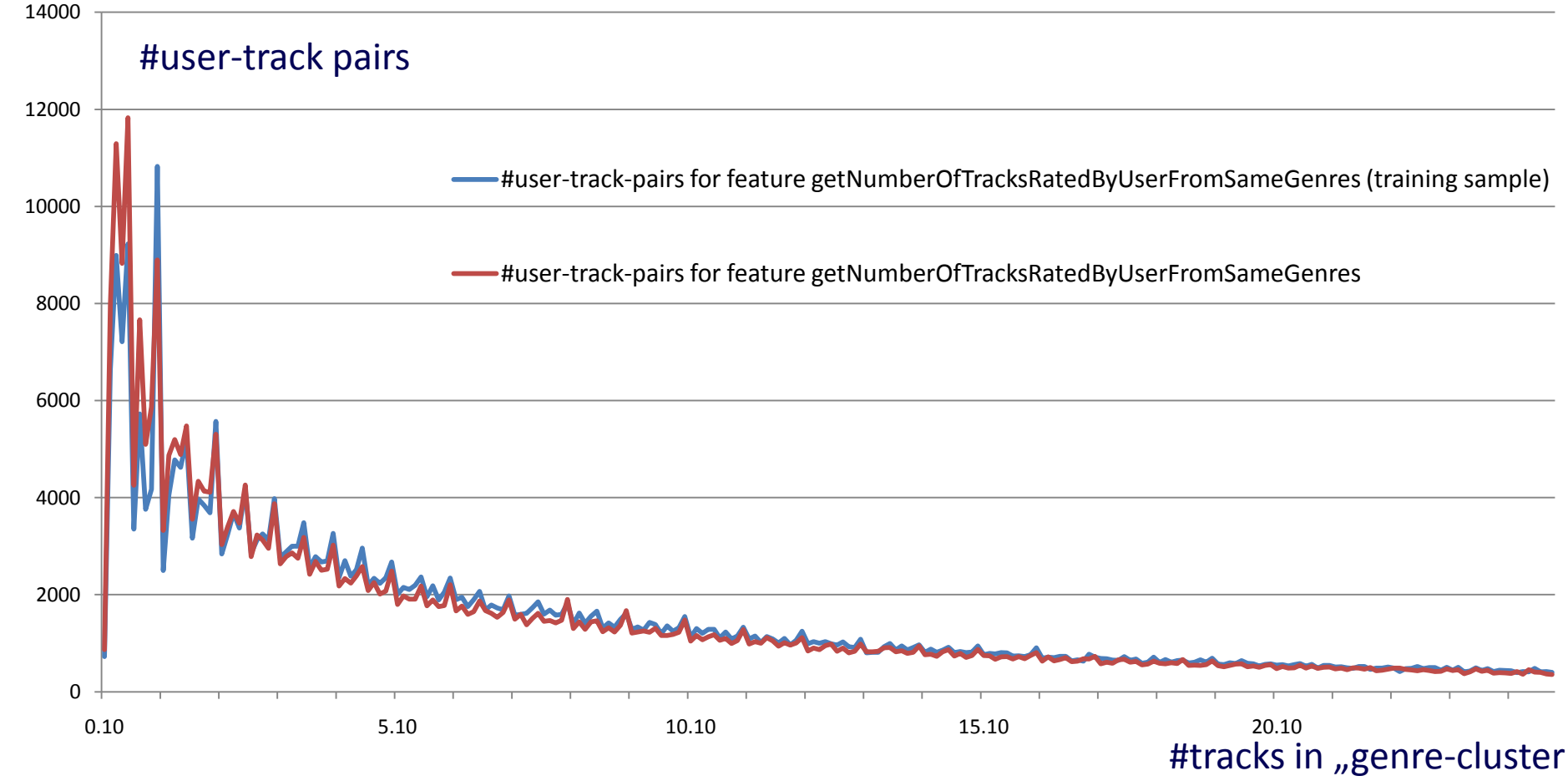
Sampling own testsets contd.



Sampling own testsets contd.



Sampling own testsets contd.



Results for different Classifiers

Classifer	Own Validation	Yahoo
Naive Bayes	17,68%	-
Logistic	15,92%	-
REPTree	14,82%	14,52%
J48	14,37%	14,61%
JRip	10,52%	20,43%
SMO	6,26%	24,21%
PART	10,47%	10,13%

Lets dive into…

…Naïve Bayes

LOOK INTO FILE

6

Lets dive into…

…Logistic

LOOK INTO FILE

Lets dive into…

…PART

getNumberOfTracksRatedByUserFromSameGenresNorm > 7.500337 AND
getNumberOfTracksRatedByUserFromSameAlbumNorm > 0.012987 AND
getGenreCountForTrack > 0 AND
getNumberOfTracksRatedByUserFromSameAlbumNorm > 0.209302 AND
getNumberOfTracksRatedByUserFromSameGenres > 3.993825 AND
getNumberOfTracksRatedByUserFromSameGenresNorm > 25.722462: 1 (20824.25/134.01)

=> ~0.65% error rate for this rule

Timetable

- vector attributes (cosine similarity)
- cf attributes

- delete outlier
- discretize attributes
- tweak & optimize

Leaderboard

Rank	Team Name	Best Score (Error Rate %)	Last Submit Time
1	NoSleepNoMercy	2.7561	2011-06-05 23:41:10
2	Lemon	2.7951	2011-06-08 00:46:06
3	acacam	2.8043	2011-06-08 00:18:16

•
•
•

75	seric	9.4751	2011-06-08 01:09:20
76	cheng	9.6837	2011-06-08 01:02:59
77	PIZZA	9.8824	2011-05-18 17:33:35
78	Adwocs	10.0382	2011-05-24 00:35:29
79	Eurystheus	10.1279	2011-06-07 14:08:13
80	glouppe-ulg	10.3068	2011-04-11 02:26:16
81	hidamari	10.4415	2011-05-03 07:58:31
82	testcode	10.4474	2011-05-28 02:28:38
83	Omega	10.4646	2011-05-29 23:23:58
84	samurai	10.4719	2011-05-04 03:14:30

Sampling own testsets contd.