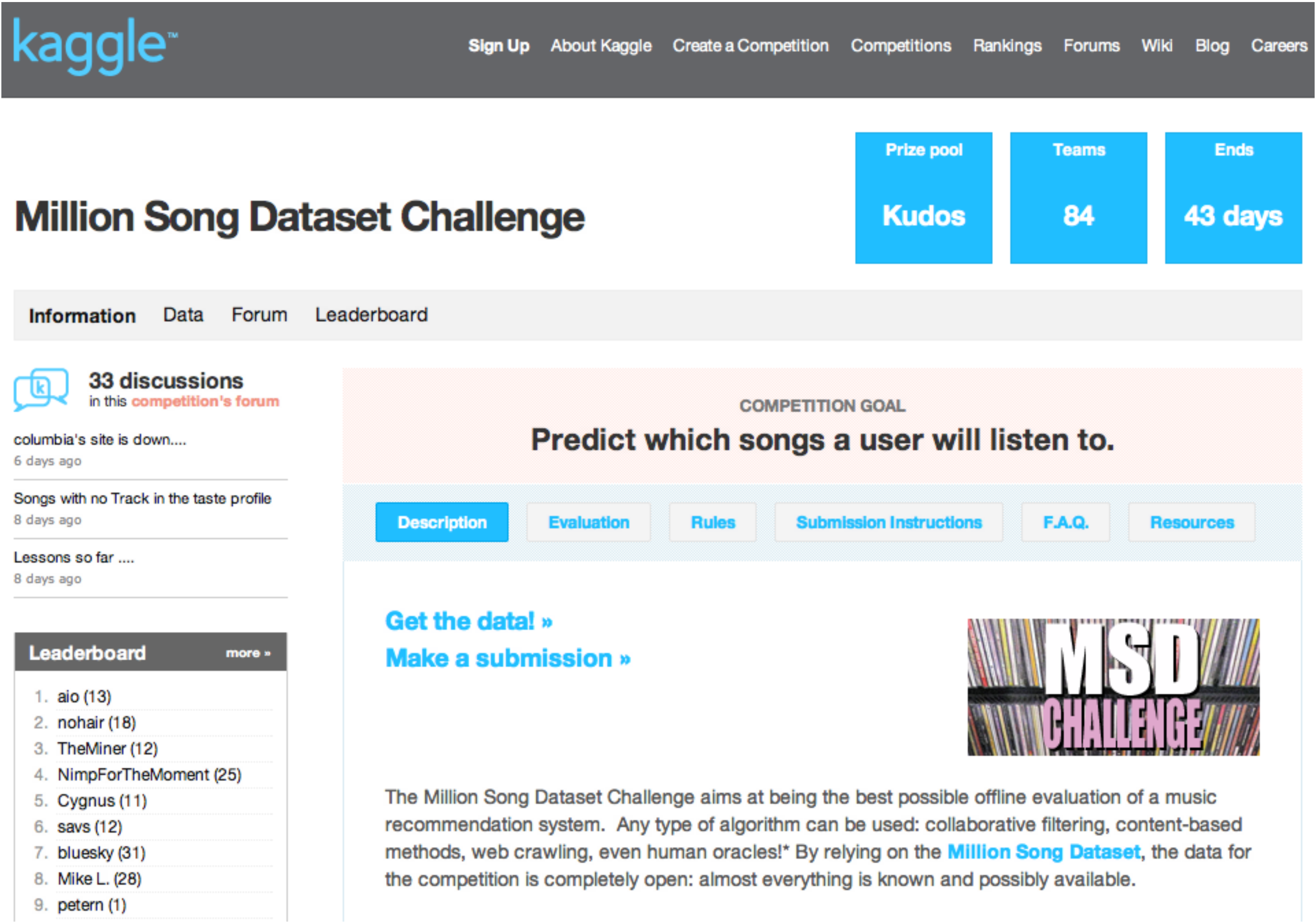
Data Mining with Pig

Andrew B. Clegg

Hadoop User Group UK

@andrew_clegg

Image: [alasam @ flickr]()



Million Song Dataset Challenge

Prize pool	Teams	Ends
Kudos	84	43 days

33 discussions
in this competition's forum

columbia's site is down....
6 days ago

Songs with no Track in the taste profile
8 days ago

Lessons so far ....
8 days ago

Leaderboard

1.	aio (13)
2.	nohair (18)
3.	TheMiner (12)
4.	NimpForTheMoment (25)
5.	Cygnus (11)
6.	savs (12)
7.	bluesky (31)
8.	Mike L. (28)
9.	petern (1)

COMPETITION GOAL

Predict which songs a user will listen to.

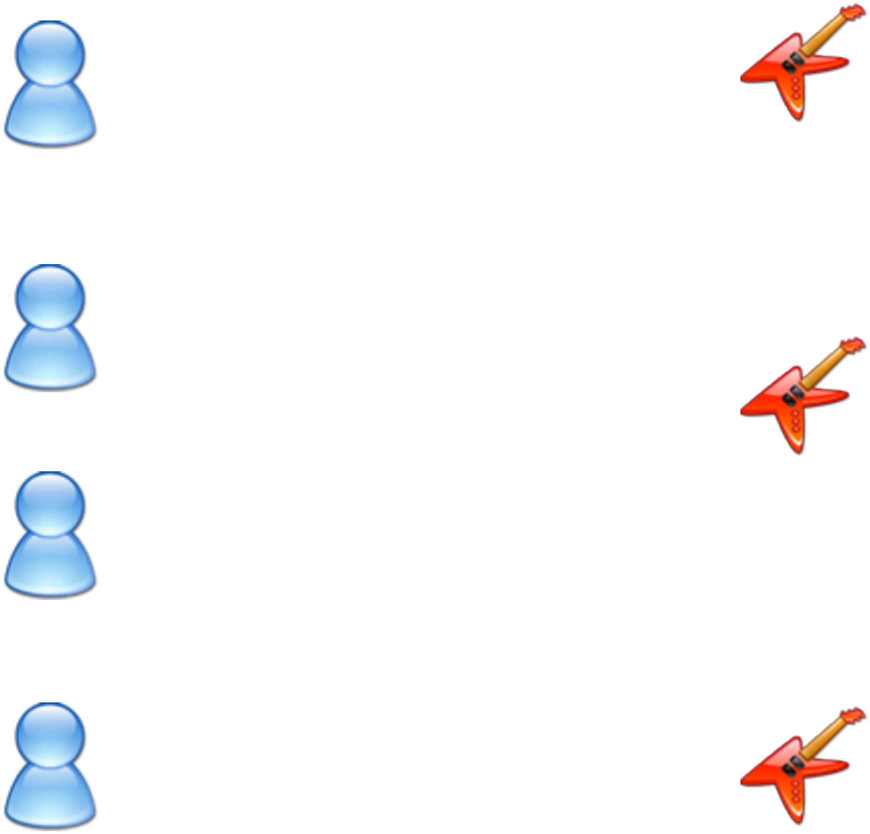
Get the data! »

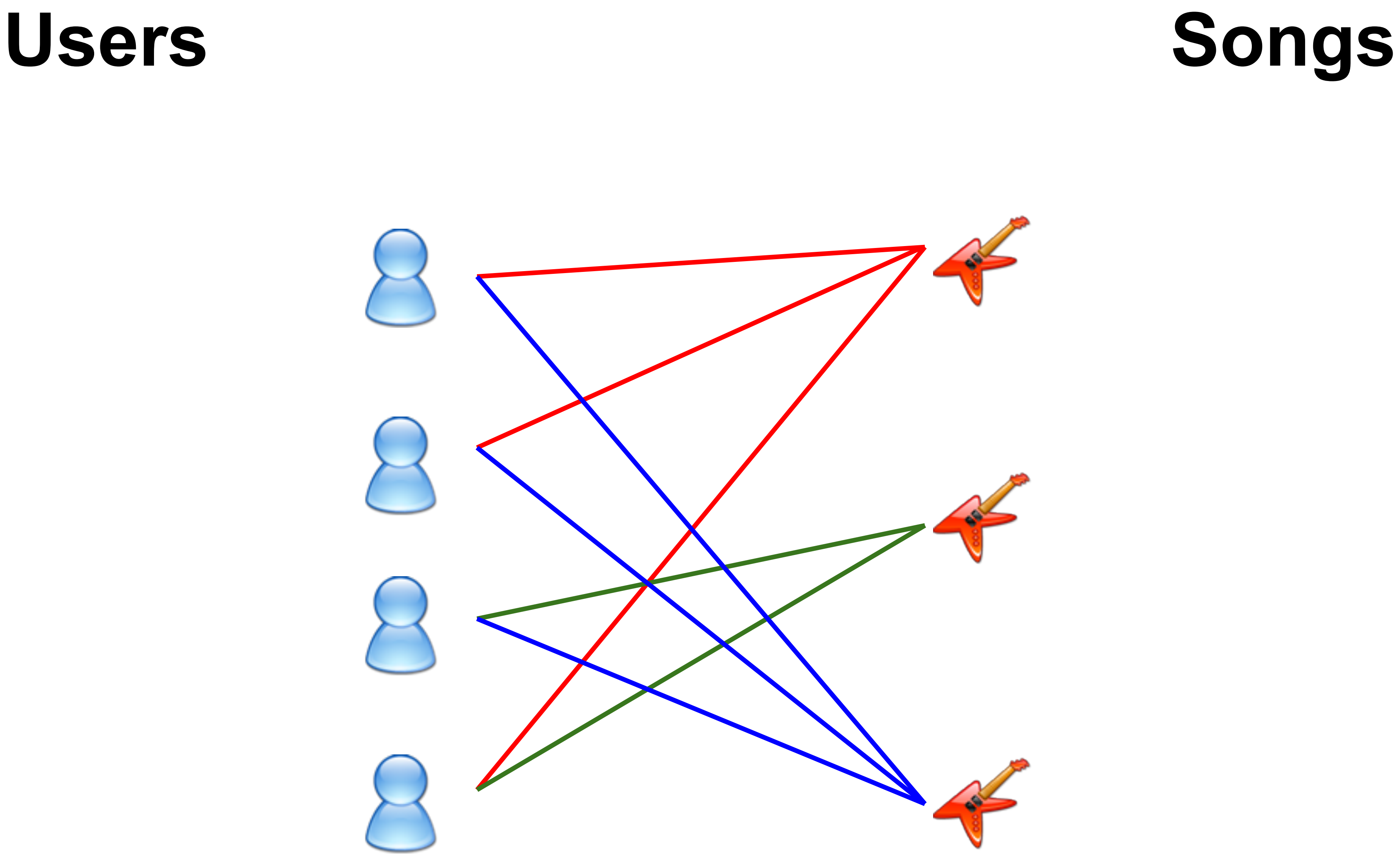
Make a submission »

The Million Song Dataset Challenge aims at being the best possible offline evaluation of a music recommendation system. Any type of algorithm can be used: collaborative filtering, content-based methods, web crawling, even human oracles!* By relying on the Million Song Dataset, the data for the competition is completely open: almost everything is known and possibly available.

Users

Songs

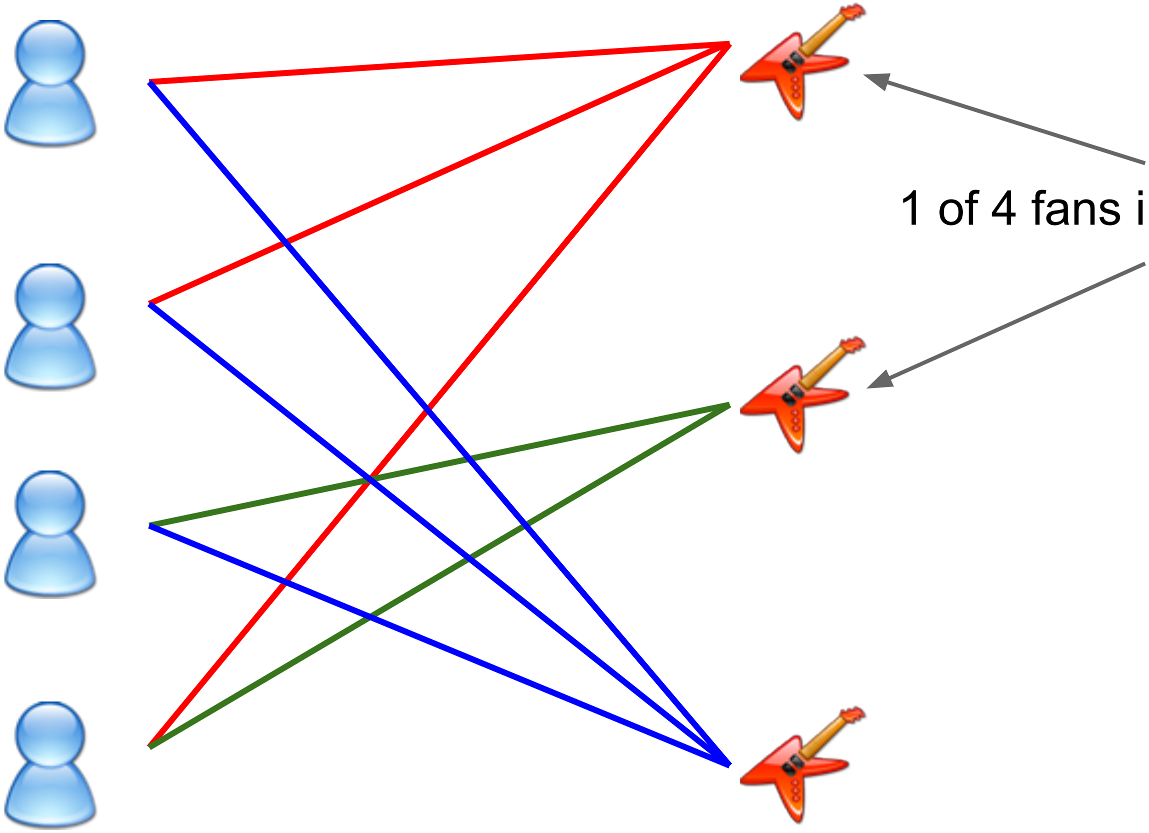




Users
Songs

Users

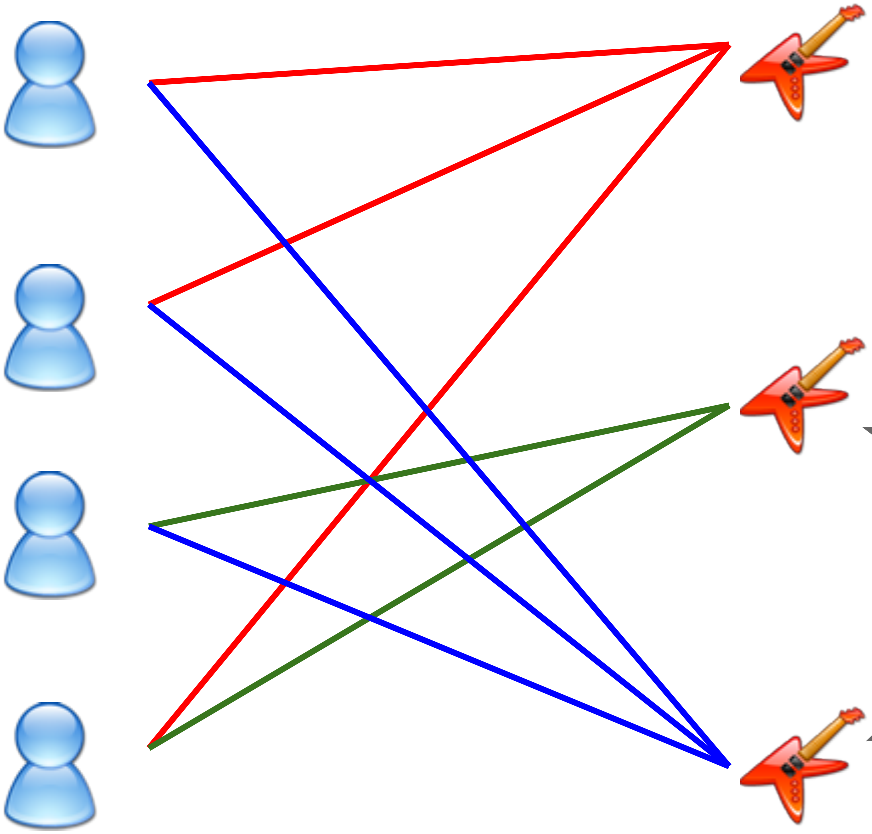
Songs



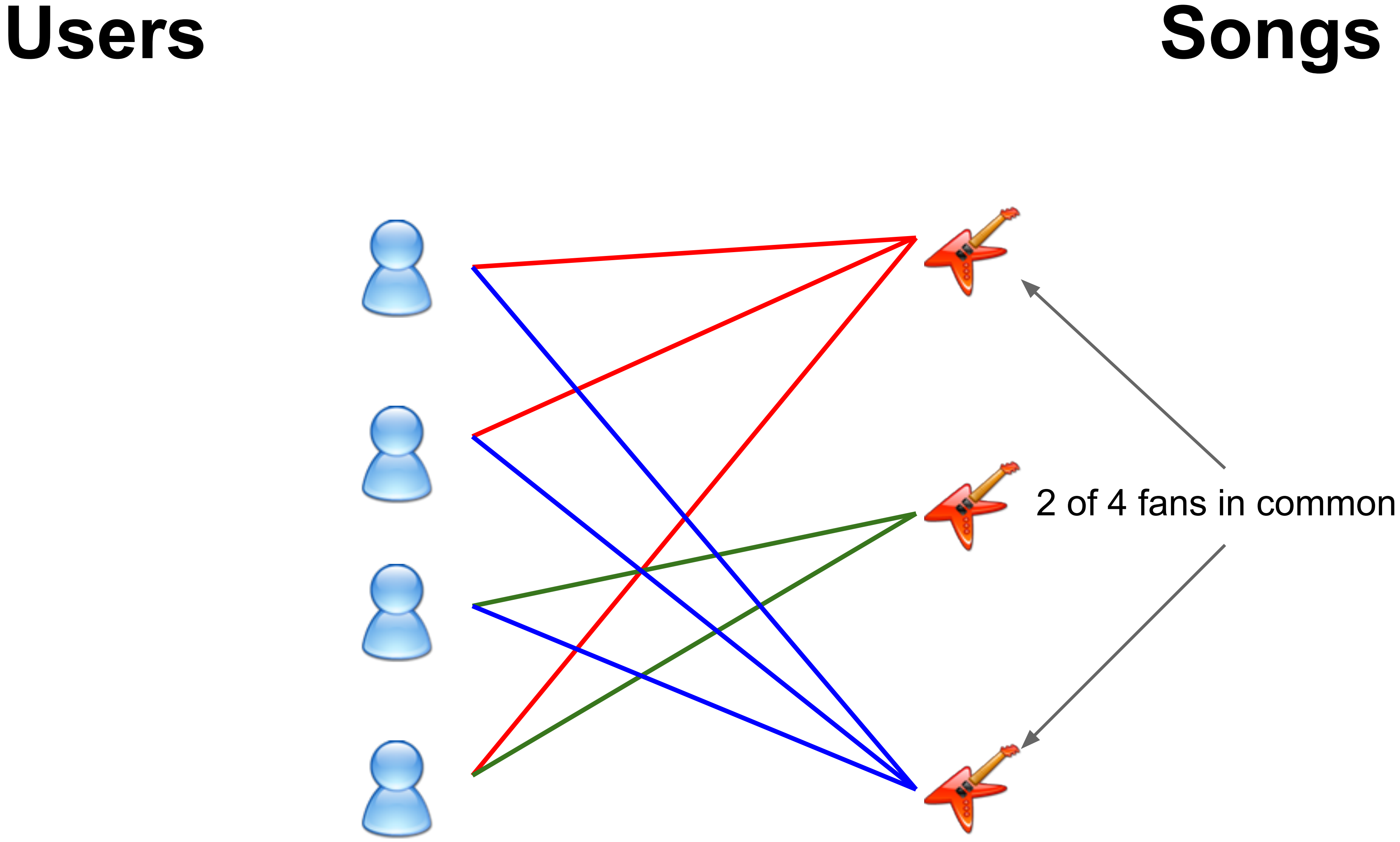
1 of 4 fans in common

Users

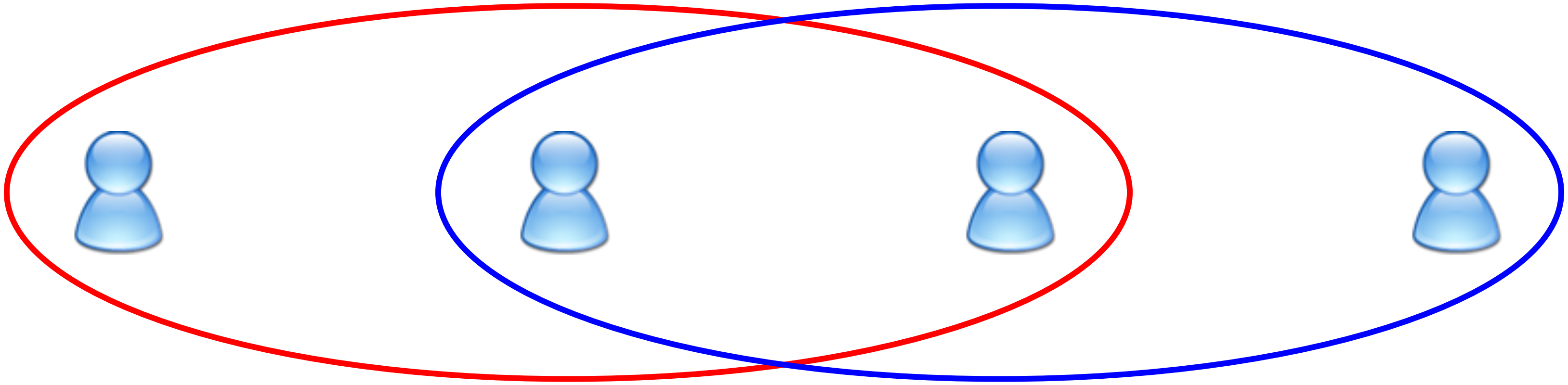
Songs



1 of 4 fans in common


Users
Songs
2 of 4 fans in common

Jaccard similarity



$$\frac{\text{Size of intersection (2)}}{\text{Size of union (4)}}$$

Live demo time!

https://github.com/andrewclegg/pig-data-mining-talk

Image: all over the internets. Who knows.

Hints and tips

Use short numeric IDs to reduce data transfer

Hash the values if assigning IDs is impractical

Replicated joins are *way* more efficient
(for joining a small dataset to a larger one)

Use log-probabilities to avoid floating-point underflow (*when applicable*)

Approximate similarity methods

MinHash -- generates similar hashes for sets with similar members

Finding similar items reduces to comparing the hashes of all the sets

This is a kind of *locality-sensitive hashing*...

... a subject for another talk.

More resources

Jacob Perkins' Data Recipes blog

DataFu from LinkedIn

pignlproc by Olivier Grisel

pig-vector by Ted Dunning

Large-Scale Machine Learning at Twitter