Evaluating Collaborative Filtering Over Time
Neal Lathia, Stephen Hailes, Licia Capra
University College London, UK
SIGIR IREVAL 2009

What is a good recommendation?

A Recommender System

Is Evaluated As

- Split Dataset
- Train Algorithm
- Predict Test

Are the predictions accurate?
Are the top-N recommendations good?
(mean error, coverage, precision,...)

Is Deployed As

- Users Rate
- Train Algorithm
- Predict Ratings
- Recommend

Does accuracy improve over time?
How does the ranking change?
(time-averaged metrics, and more...?)

We modify the evaluation process to include the reality of temporal updates and explore:

Similarity
Does like-mindedness persist?

Temporal Accuracy
Is one algorithm consistently more accurate?

Temporal Diversity
Are users offered the same recommendations again and again?

What about...?
Novelty, serendipity, niche vs popular content, context, delay,...

What makes a recommender system good for you?