

# RepoRecs

Cait Riggs

[caitriggs.com](http://caitriggs.com)

[github.com/caitriggs/github-collaborator](https://github.com/caitriggs/github-collaborator)

Are you still watching “The Office”?


Continue watching

Exit

Are you still committing to “Linux”?

Continue contributing

Exit

The background of the landing page is a dark, stylized illustration of a futuristic tunnel or space station interior. A large, circular opening in the center reveals a bright blue sky with a white contrail. In the foreground, a small, dark, cat-like robot with a yellow visor and a backpack is seen from behind, looking out through the circular opening. The overall aesthetic is sci-fi and tech-oriented.

RepoRecs

## Find your match on GitHub.

See repo recommendations based on your activity.

madnury

Get started

## Data Acquisition

GHTorrent data dump  
to MySQL tables

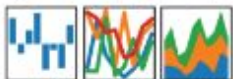
59m Repos  
19m Users  
700m Commits



16m User to  
Repo Ratings

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



MySQL®

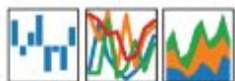
## Data Acquisition

GHTorrent data dump  
to MySQL tables

59m Repos  
19m Users  
700m Commits

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



MySQL

## Modeling

User-Item Collaborative  
Filtering using ALS

16m User to  
Repo Ratings

amazon  
web services

APACHE  
Spark

Top  
Repos per  
User

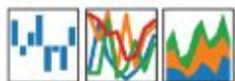
## Data Acquisition

GHTorrent data dump  
to MySQL tables

59m Repos  
19m Users  
700m Commits

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



MySQL

## Modeling

User-Item Collaborative  
Filtering using ALS

16m User to  
Repo Ratings

amazon  
web services

APACHE  
Spark

## Recommendations

Recommend Top Repos  
for a GitHub User

Top  
Repos per  
User

Explore  
RepoRecos!



Flask  
web development,  
one drop at a time

# Collaborative Filtering

	Repo 1	Repo 2	Repo 3	Repo m
User 1	4		2	4
User 2		4		
User 3	4		2	
User n	2		4	4

- Explicit approach
  - Starred = 2
  - Owned/Forked = 4





# Collaborative Filtering

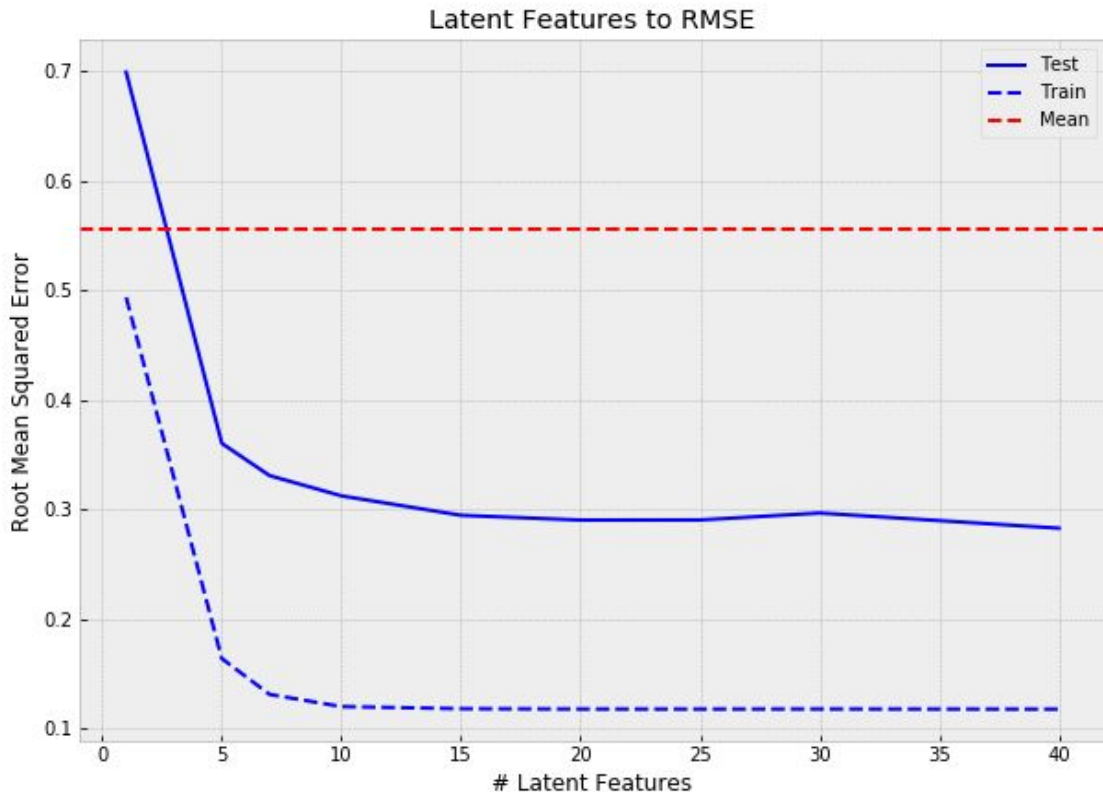
	Repo 1	Repo 2	Repo 3	Repo m
User 1	.67			2.34
User 2		1.32		
User 3	1.01			
User n			3.33	1.17

- Explicit approach
  - Starred = 2
  - Owned/Forked = 4
- Implicit approach
  - $\text{Log}_{10}(\text{\#commits/repo})$

🔒 2,166 commits

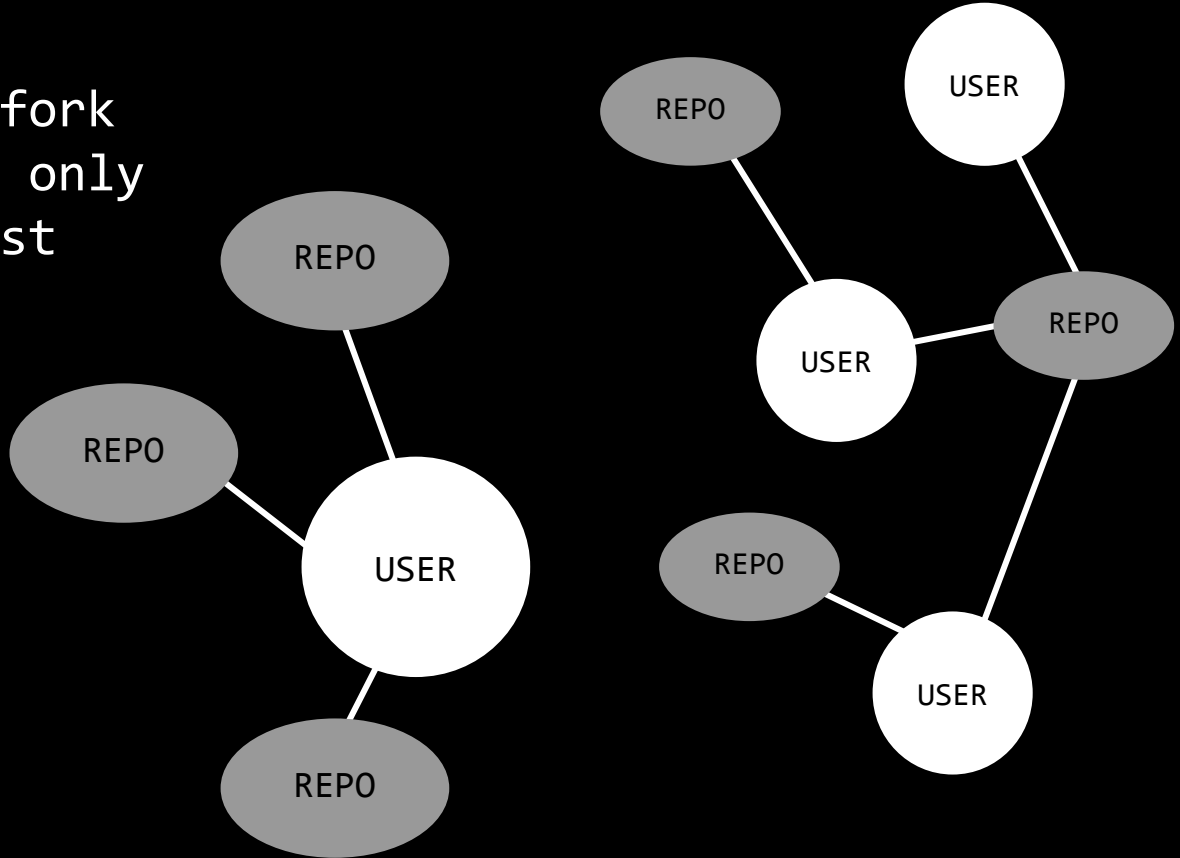
# ALS Model Evaluation

- Used **RMSE** as a check for overfitting and rank selection
- **Precision@k** for recommendation quality metric



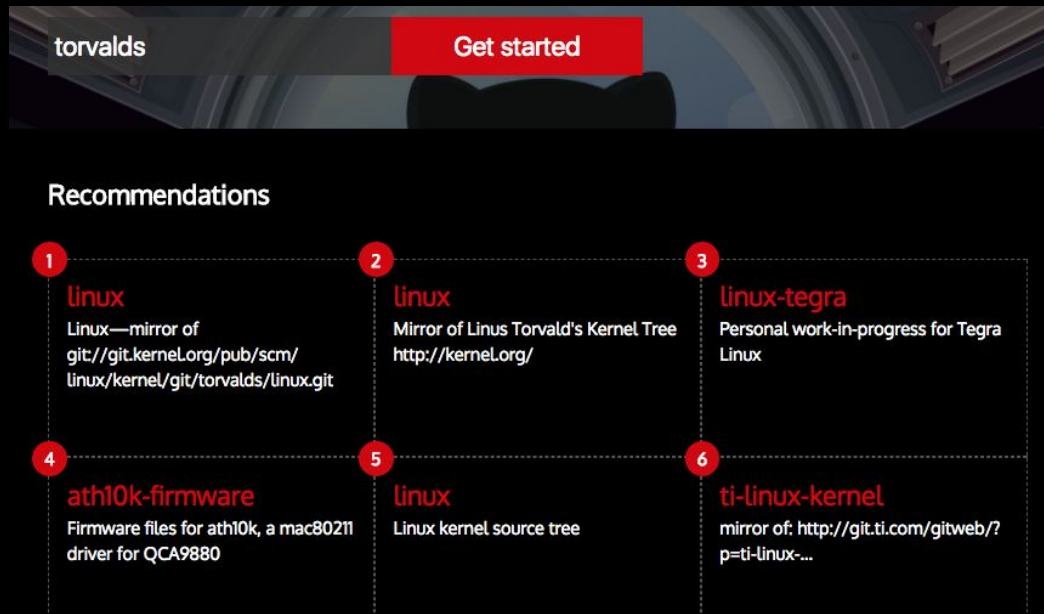
# Collaborative Pitfalls

- Users who don't fork other repos, and only use GitHub to host their projects



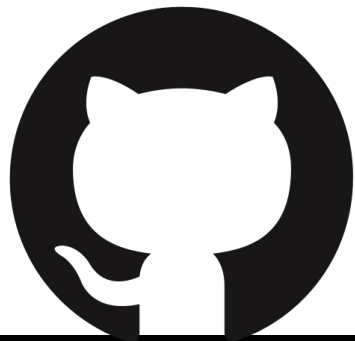
# Collaborative Pitfalls

- Users who don't fork other repos, and only use GitHub to host their projects
- Model picks up on the users who fork their projects (e.g. Linus Torvalds)



# Next Up

- Further explore recs:
  - Content-based filtering
  - Weighting recs by repo 'quality'
- Live web app!



# RepoRecs

Cait Riggs

[caitriggs.com](http://caitriggs.com)

[github.com/caitriggs/github-collaborator](https://github.com/caitriggs/github-collaborator)