




Yelp Recommender System:

Connecting people with restaurants of their choice and taste

Reporter: Ildar Abdrashitov
Mentor: Nemanja Radojković

Other Movies You Might Enjoy Close

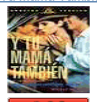


[Amelie](#)

Add

★★★★☆

Not Interested




[Y Tu Mama Tambien](#)

Add

★★★★☆

Not Interested

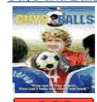


Eiken has been added to your Queue at position 2.

This movie is available now.

[Move To Top Of My Queue](#)

[Continue Browsing](#) [Visit your Queue](#)




[Guys and Dolls](#)

Add

★★★★☆

Not Interested




[Mostly Martha](#)

Add

★★★★☆

Not Interested




[Only Human](#)

Add

★★★★☆

Not Interested



[Russian Dolls](#)

Add

★★★★☆

Not Interested

Close

amazon.com

Recommended for You

Amazon.com has new recommendations for you based on [items](#) you purchased or told us you own.



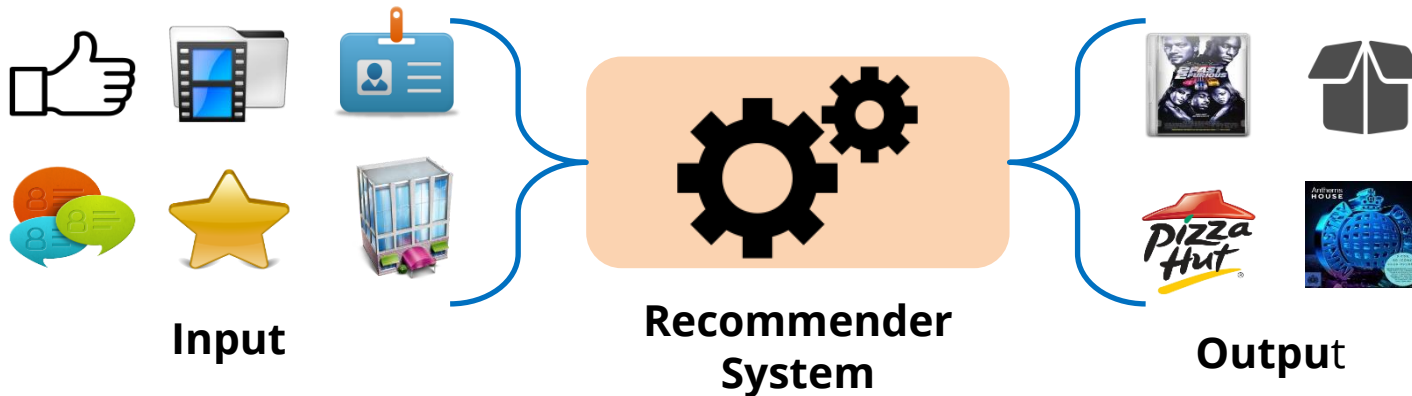
[Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop](#)



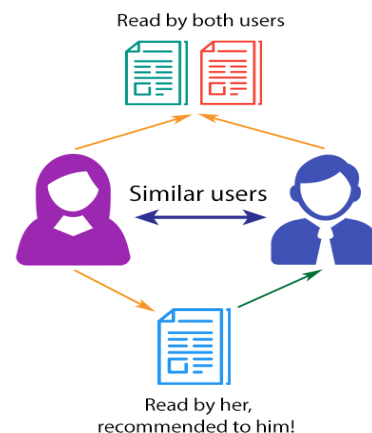
[Google Apps Administrator Guide: A Private-Label Web Workspace](#)



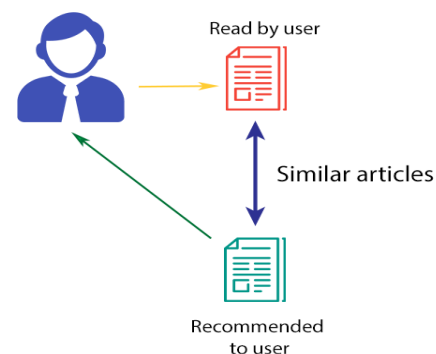
[Googlepedia: The Ultimate Google Resource \(3rd Edition\)](#)



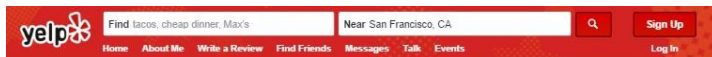
- Collaborative filtering:



- Content-based filtering:



- Hybrid: Collaborative + Content-based



Yelp Dataset Challenge

Round 8 Of The Yelp Dataset Challenge: Now With Photos!

We've had 7 rounds, over \$45,000 in cash prizes awarded, [hundreds of academic papers written](#), and we are excited to see round 8.

Our dataset has been updated for this iteration of the challenge - we're sure there are plenty of interesting insights waiting there for you. This set includes information about local businesses in 10 cities across 4 countries.

This round also includes a new type of data - photos! These photos nicely complement reviews, business attributes, check-ins, and tips, and open the door to even more exciting research. An auxiliary file has been provided for download (see the "Get the Data" link on this page), containing 200,000 pictures from 85,901 businesses described in the main dataset. The photo archive includes a json file linking each photo to its corresponding business in the dataset, and listing its caption (if any), and type of content as determined by our [image classifier](#) (we currently only list labels for some restaurants).

This treasure trove of local business data is waiting to be mined and we can't wait to see you push the frontiers of data science research with our data.



The Challenge Dataset:

- 2.7M reviews and 649K tips by 687K users for 86K businesses
- 566K business attributes, e.g., hours, parking availability, ambience.
- Social network of 687K users for a total of 4.2M social edges.
- Aggregated check-ins over time for each of the 86K businesses
- 200,000 pictures from the included businesses

[Get the Data](#)

Cities:

- U.K.: Edinburgh
- Germany: Karlsruhe
- Canada: Montreal and Waterloo
- U.S.: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison

5 Data Sets:

Businesses

Check-ins

Reviews

Tips

Users

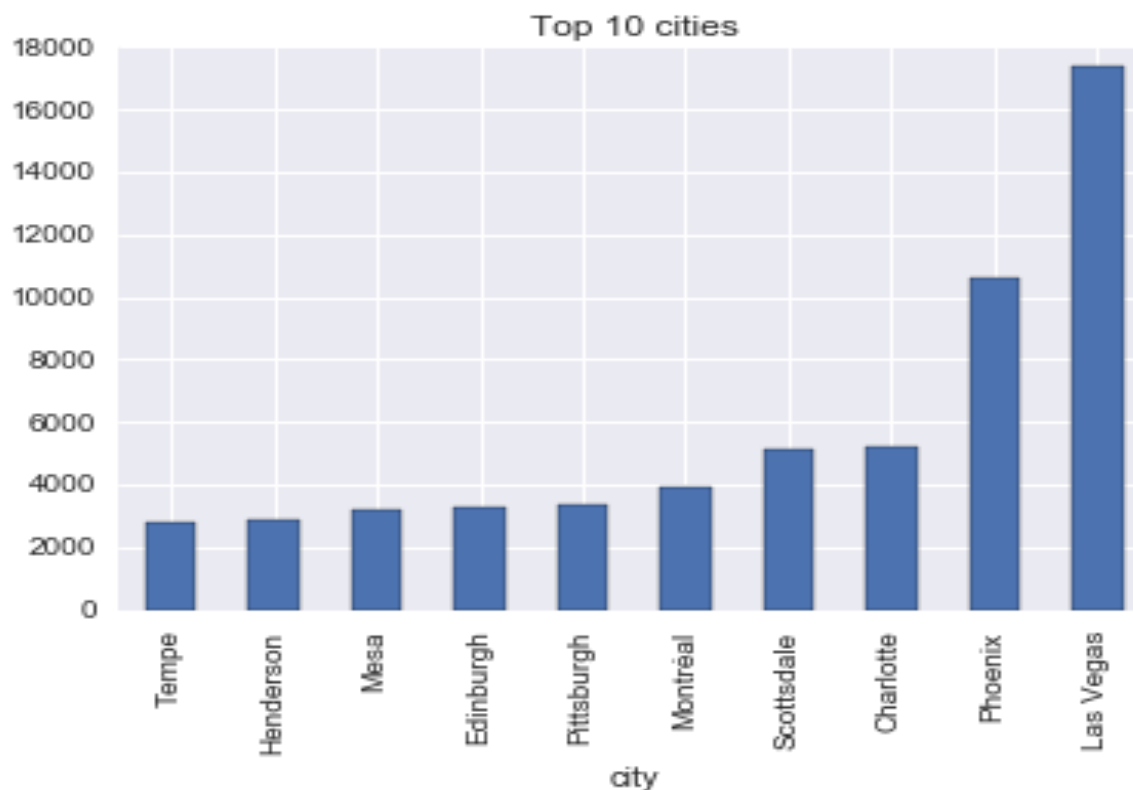
Data Format: JSON

Overall Size: ~ 2.2 Gb

Data Source: https://www.yelp.com/dataset_challenge

Businesses:

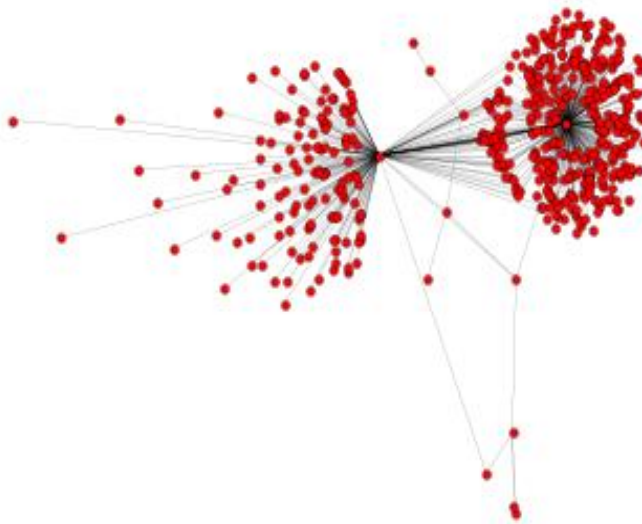
- 77445 unique businesses - 31.6 avg. review count
- 412 cities - 3.6 avg. stars



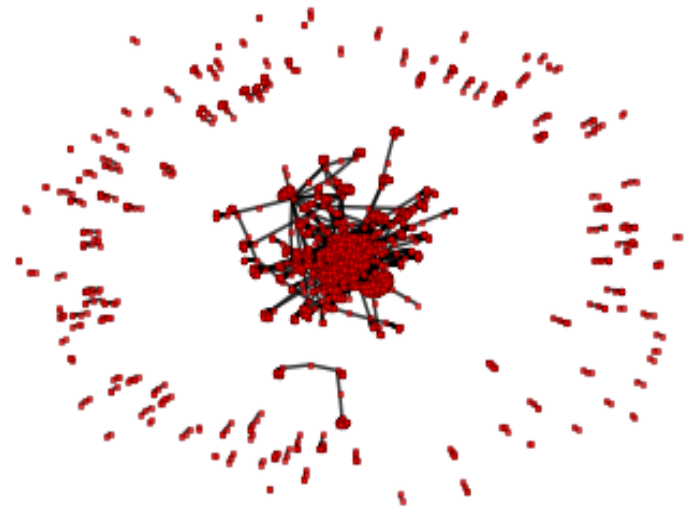
Users:

- 552 339 unique users
- 3.7 average stars
- 3 812 max n of friends
- 0 min n of friends (most common => 302 899 users)

5 users SN graph

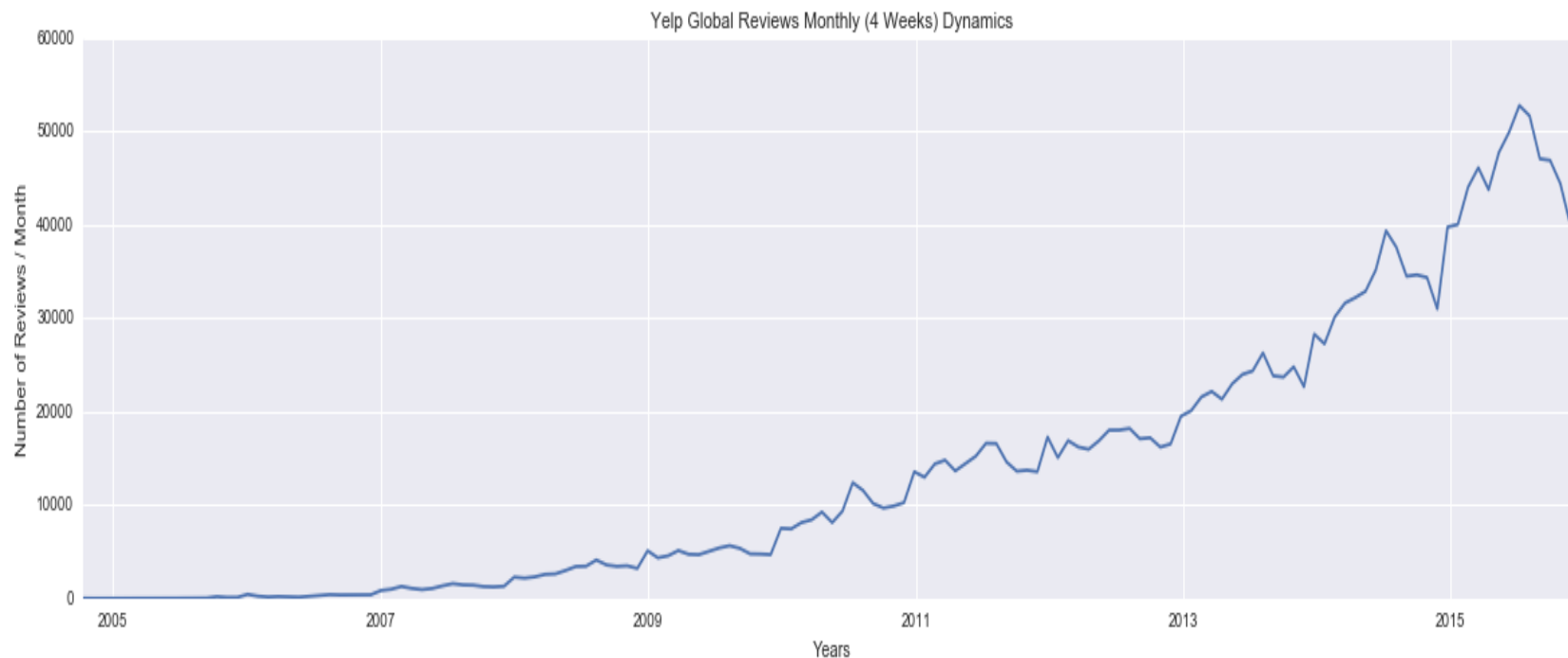


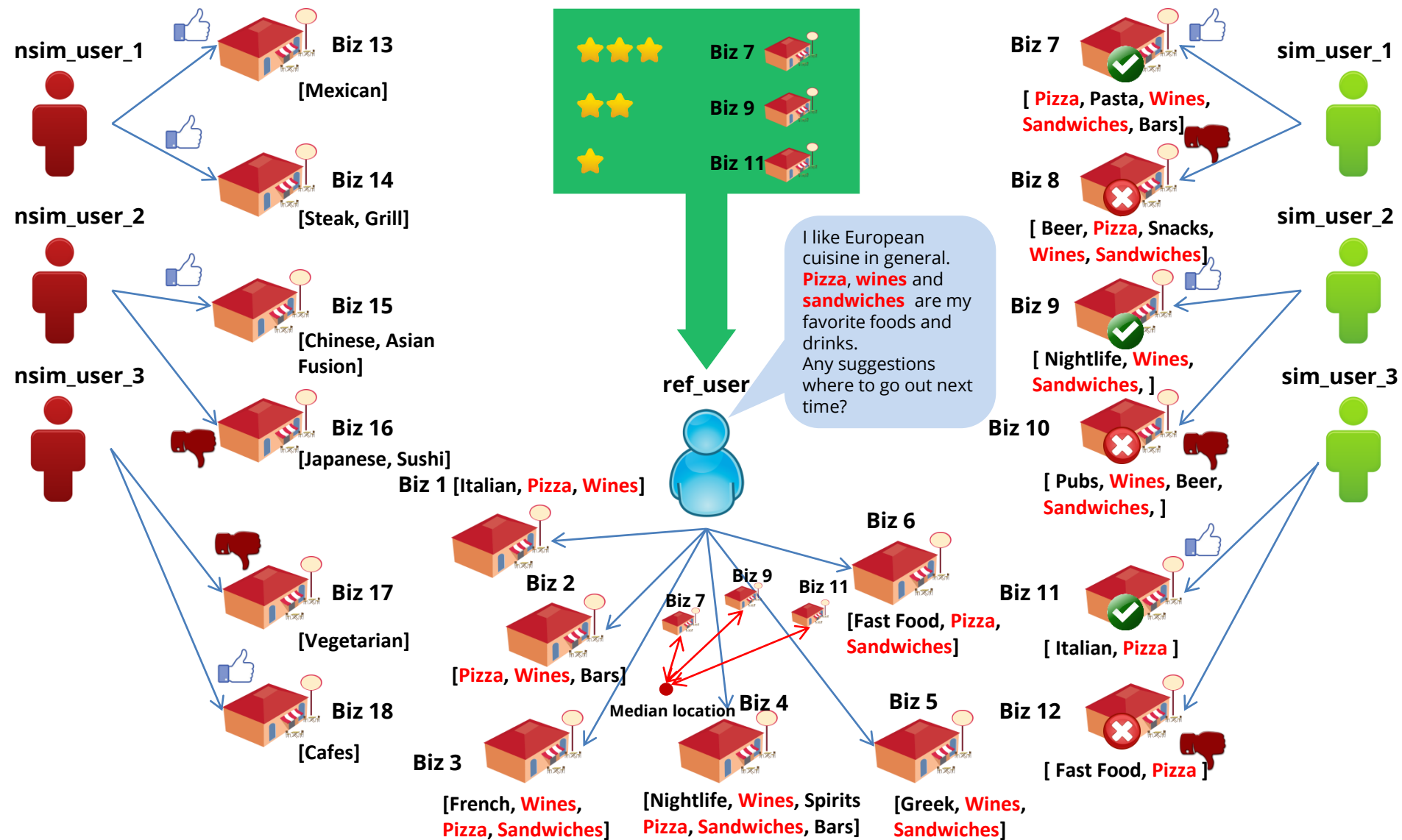
500 users SN graph



Reviews:

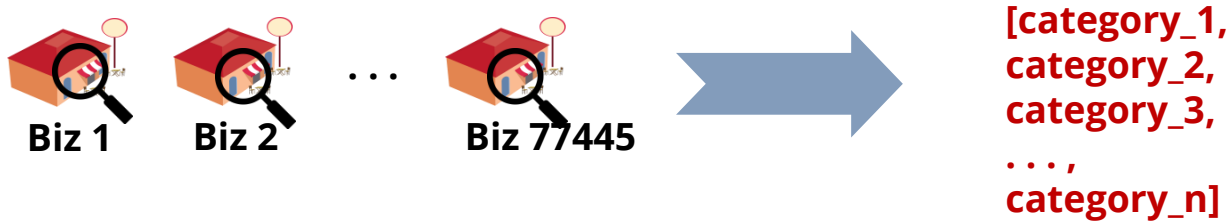
- 2 225 213 reviews
- 552 339 unique users
- 77 079 unique businesses
- 3.75 avg. stars







Learning users tastes and preferences



User preference vector

user_1

2 reviews

10 stars max cumulative rating

★★★★★
★★★★★

category_1

★★★★★
★★★★★

5/10

category_2

★★★★★

3/10

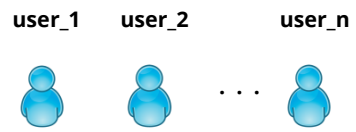
...

category_n

★★★★★

2/10

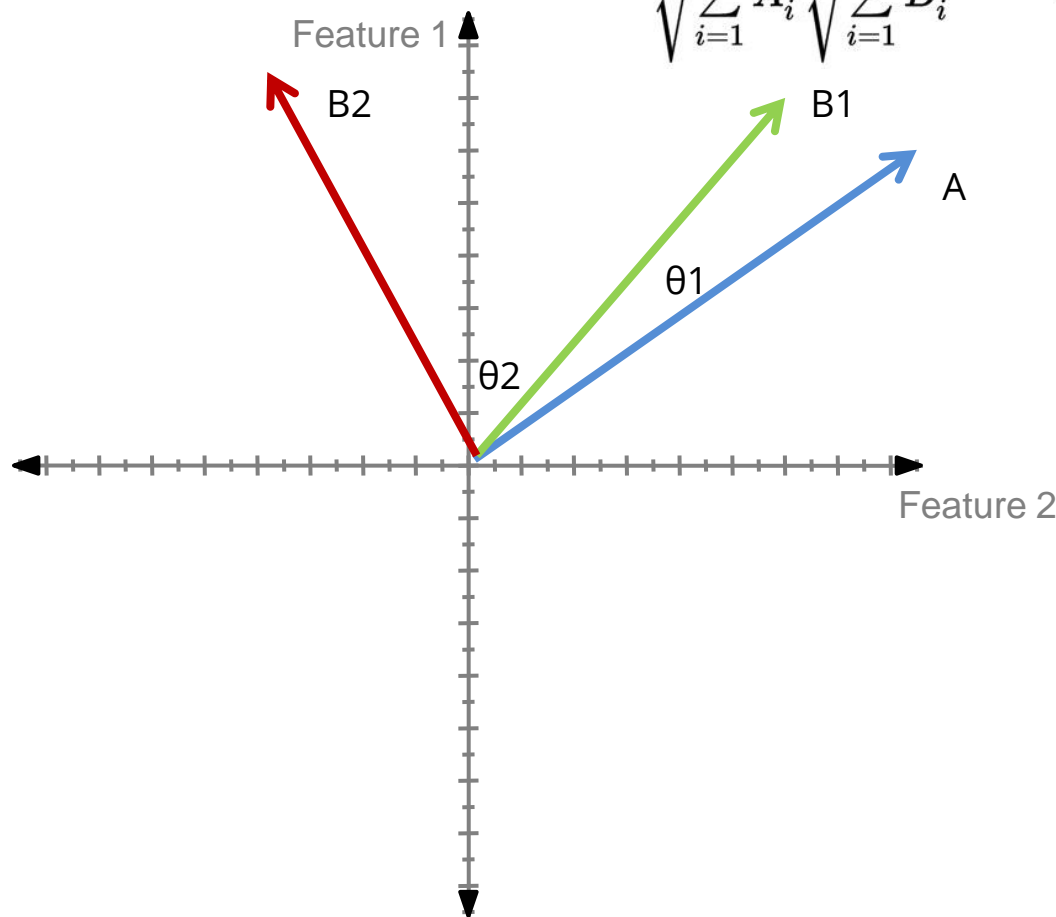
	Irish	Juice Bars & Smoothies	Korean	Cheese Shops	Arts & Entertainment	American	Peruvian	Shanghai
-0itF0VWVB3k2AdfUReGA	0	0	0.012346	0	0	0.012336	0	0
-0p7UrssdF5FvDKl5y6RA	0	0	0	0	0	0.072727	0	0
-0wUMy3vgInUD4S6KJInnw	0	0	0	0	0	0.044444	0	0
-19j4ExVKylIU8JcEV55Q	0	0	0	0	0	0.030769	0	0
-24OurJb1F03IWgo0EbMEv	0	0	0	0	0	0.063281	0	0
-26J4wzYrjU7IIQAKr1EWg	0	0.092261458	0	0	0	0.09157	0	0
-2EuoueswhqEERWezJY8gv	0	0	0	0	0.02	0.013556	0	0
-2PFmYGBW6rQhZtRh6PIPA	0.007712	0	0.012847	0	0.01025641	0.007712	0	0
-2RVxn8RRN3j9bOiqGkjug	0	0	0	0	0	0.1	0	0
-2gxrMXH8_F1Ep_ppWYTk	0	0	0	0	0.090909091	0.090909	0	0
-2iNr0vd-jCCDGcXW6lDdg	0.08	0	0	0	0	0.08	0	0
-2r-OO0dtCQq3BLYdHXlg	0	0	0	0	0	0.019501	0	0
-3AYE63sussmBAylWvDnOv	0	0	0	0	0	0.038023	0	0
-3HcO3CnxzhxdEW8Soplig	0	0	0.055926	0	0	0.044444	0	0
-3HkgwyxVPf49MXM4dT8T	0.028776	0	0	0	0.02877551	0.028777	0	0



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

\mathbf{A}, \mathbf{B} - 2 users pref. vectors

A_i, B_i - components of the corresponding pref. vectors



ref_user ref_user`s pref. vector



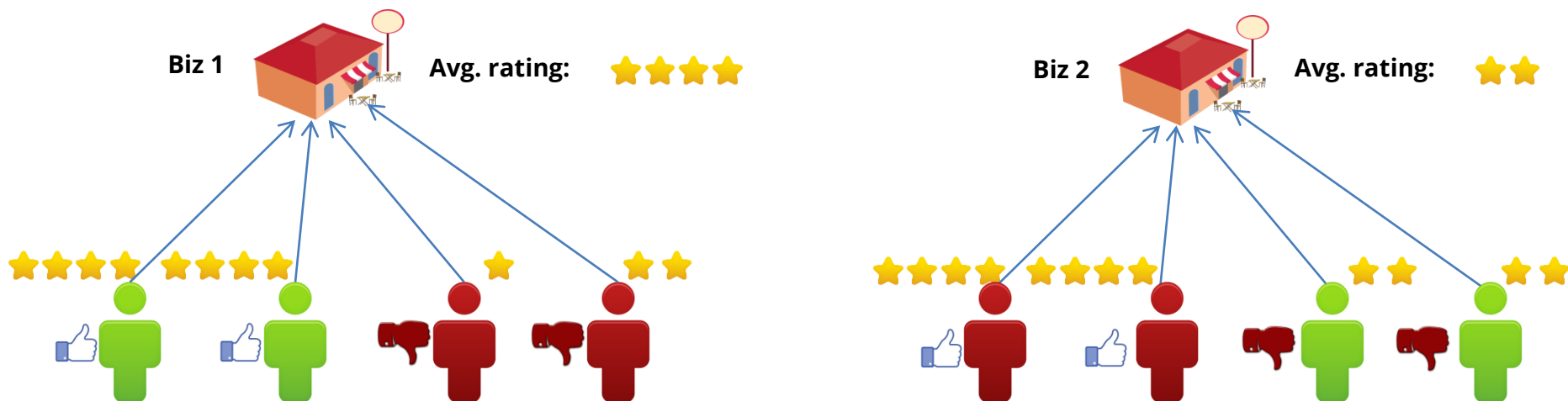
sim_user sim_user`s pref. vector



nsim_user nsim_user`s pref. vector



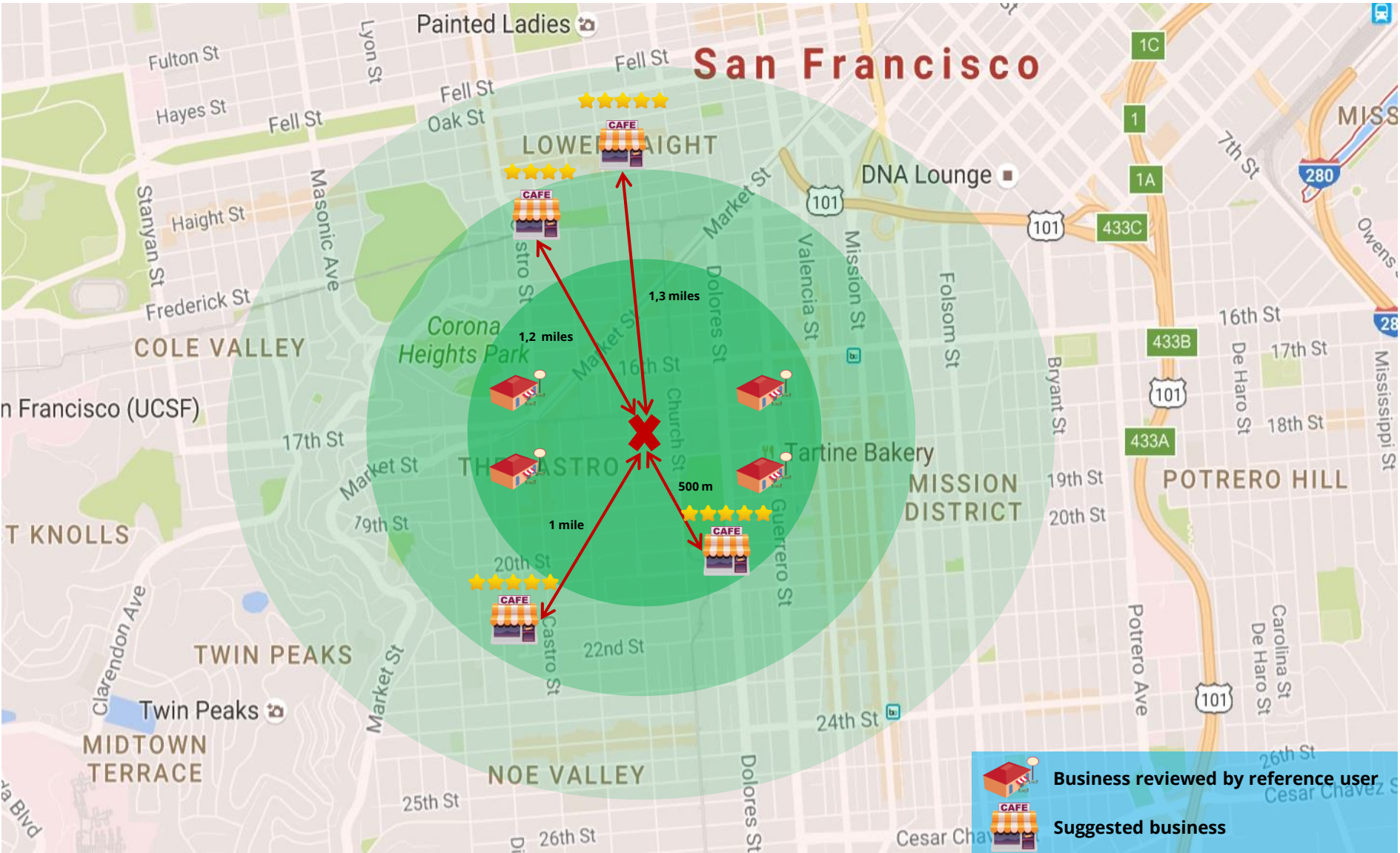
$0 \leq \text{similarity} \leq 1$

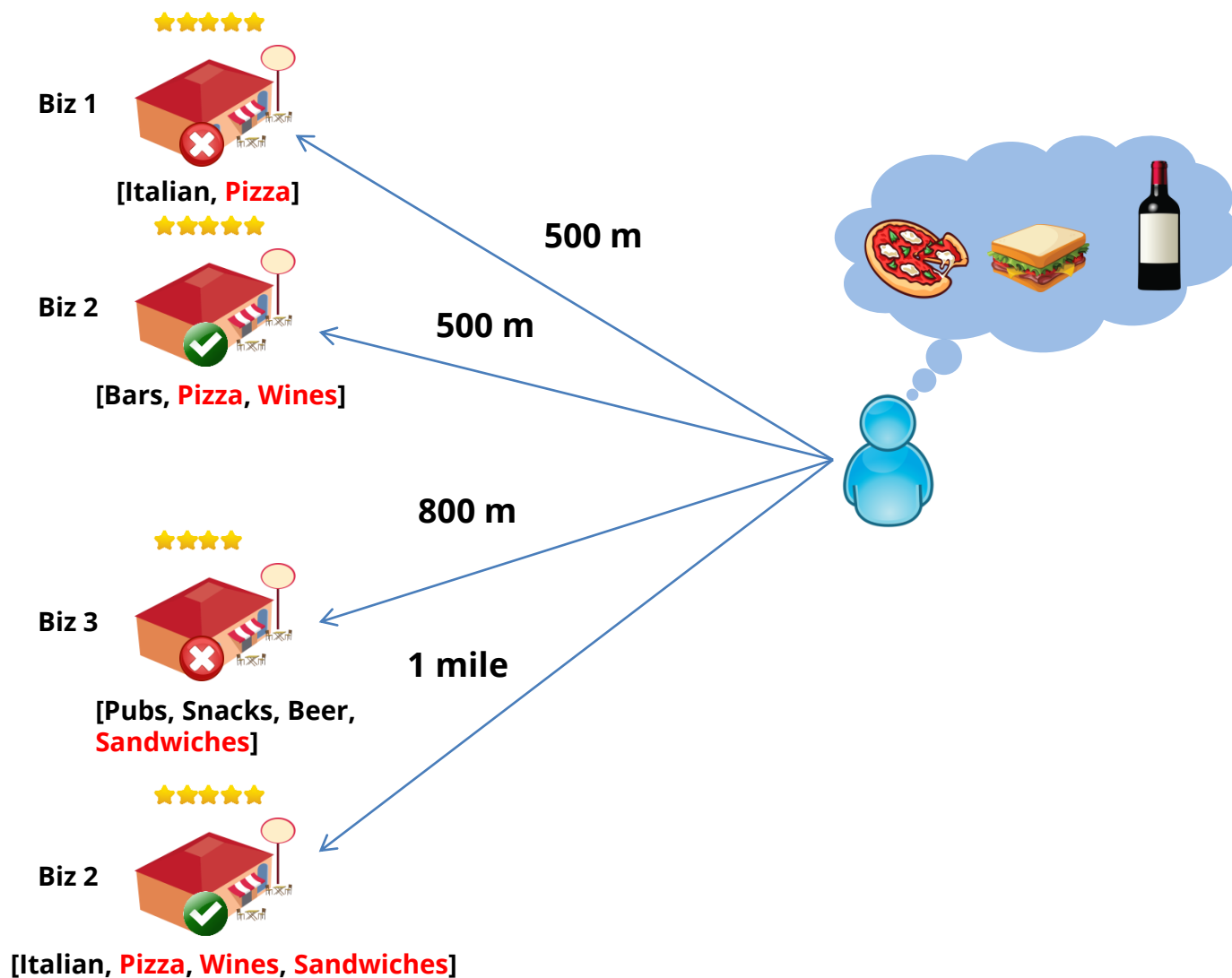


$$\frac{\sum_{i=1}^n rating_i \times similarity_i^2}{\sum_{i=1}^n similarity_i^2}$$

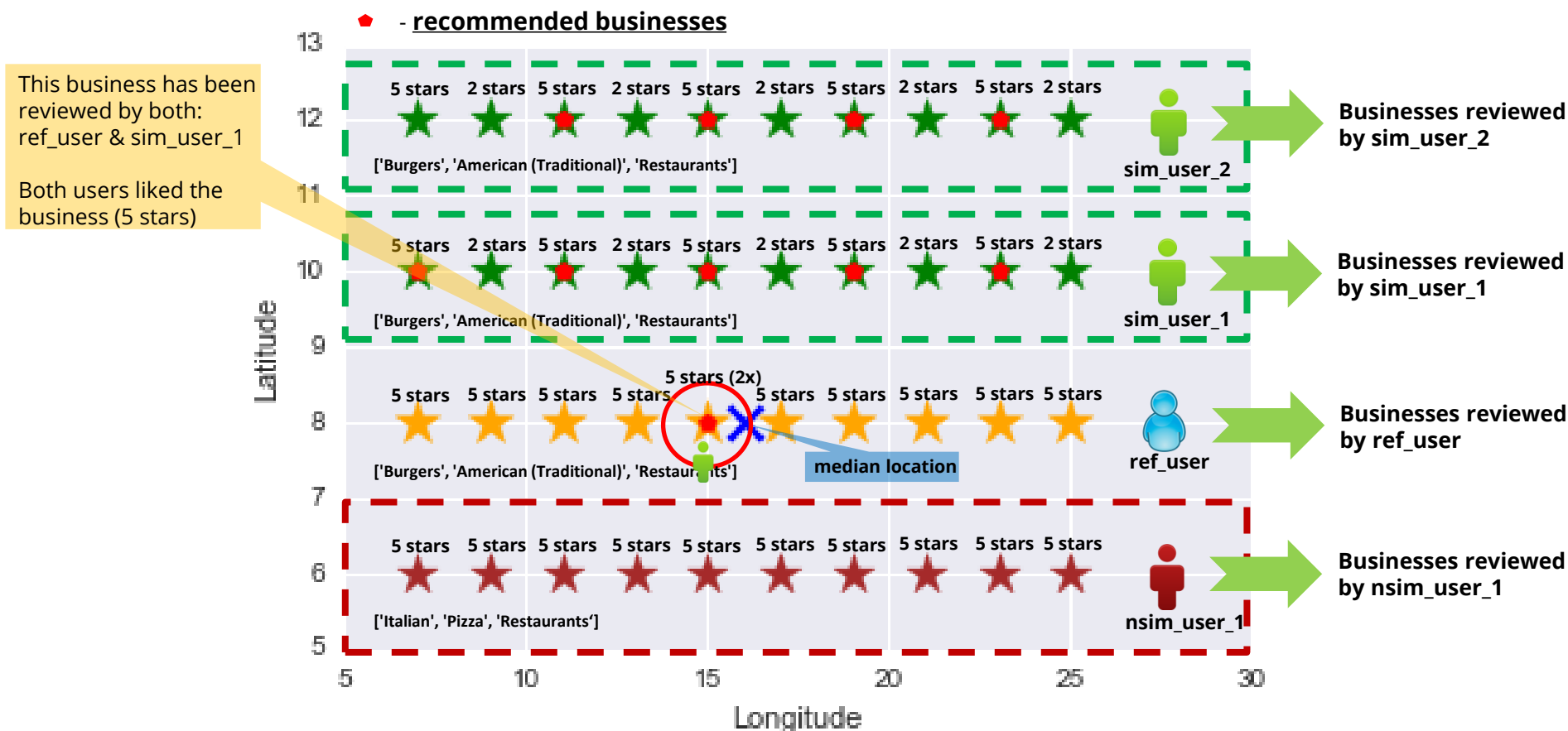


Business combination metric





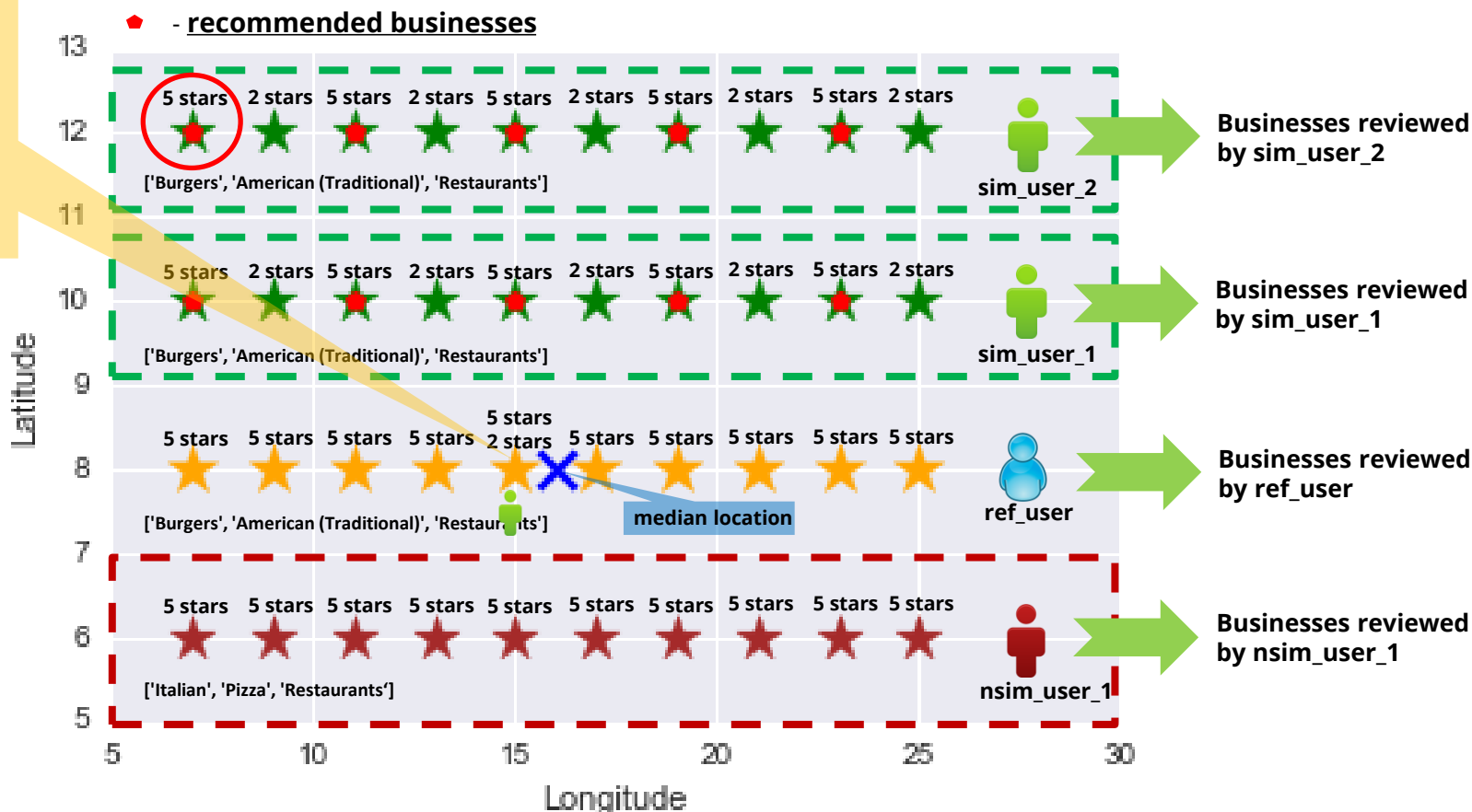
- Test case 1:**
- 1 reference user - 40 businesses
 - 2 similar users (30 same category)
 - 1 not similar user - sim_users: 5 stars reviews count >> 11
 - sim_users: 2 stars reviews count >> 10
 - n_sim_user: 5 stars reviews count >> 10



- Test case 2:**
- 1 reference user - 40 businesses
 - 2 similar users (30 same category)
 - 1 not similar user - sim_users: 5 stars reviews count >> 10
 - sim_users: 2 stars reviews count >> 11
 - n_sim_user: 5 stars reviews count >> 10

This business has been reviewed by both: ref_user & sim_user_1

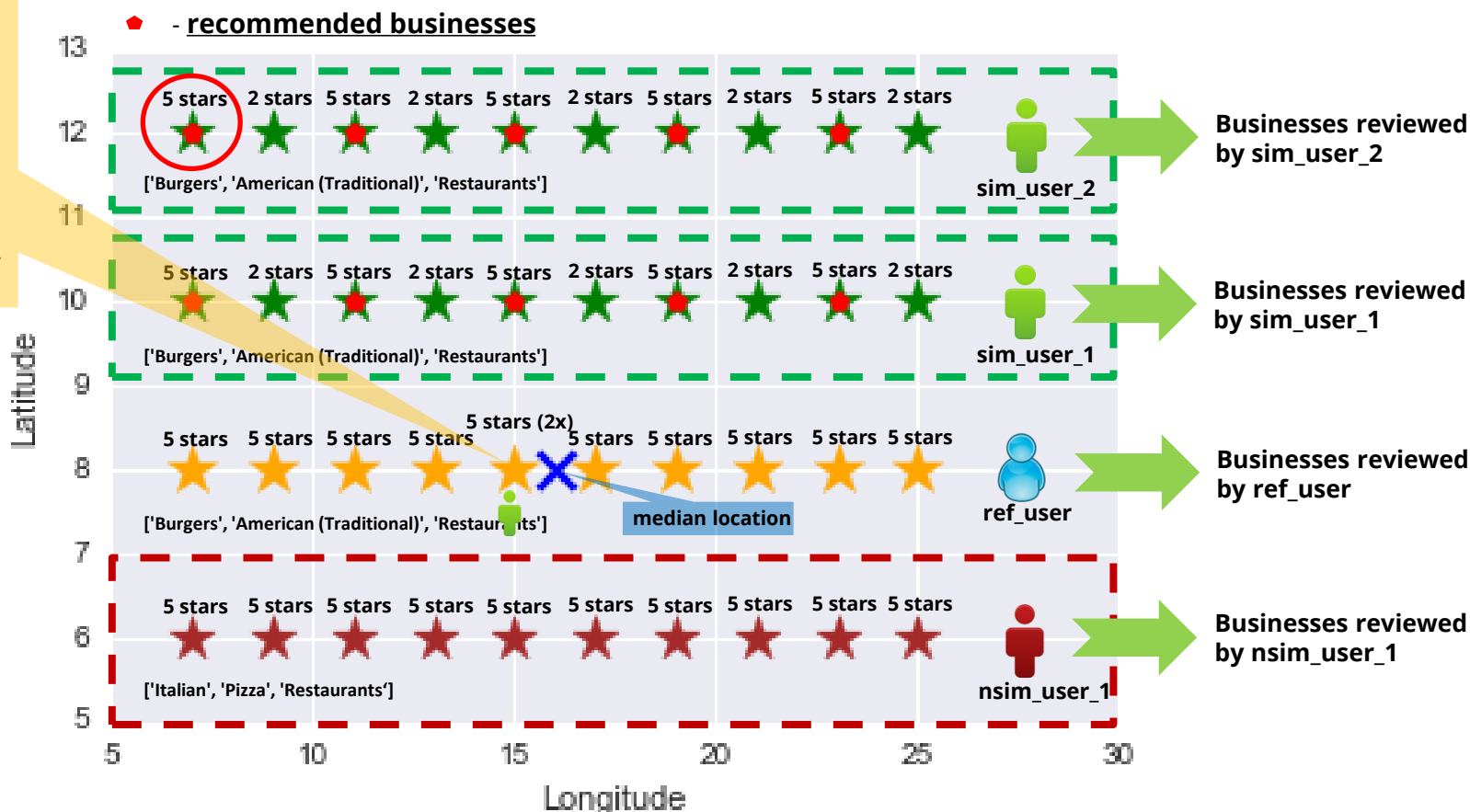
However, this time sim_user_1 did not like the business (2 stars)



This business has been reviewed by both: `ref_user` & `sim_user_1`

Both users liked the business (5 stars)

However, this business belongs only to ['Burgers'] category





Running model on real data



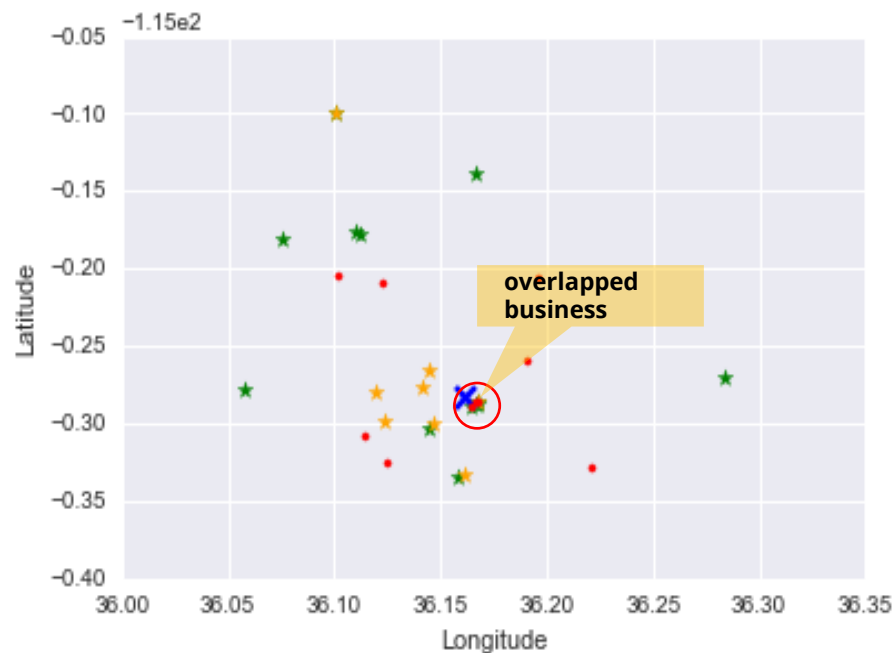
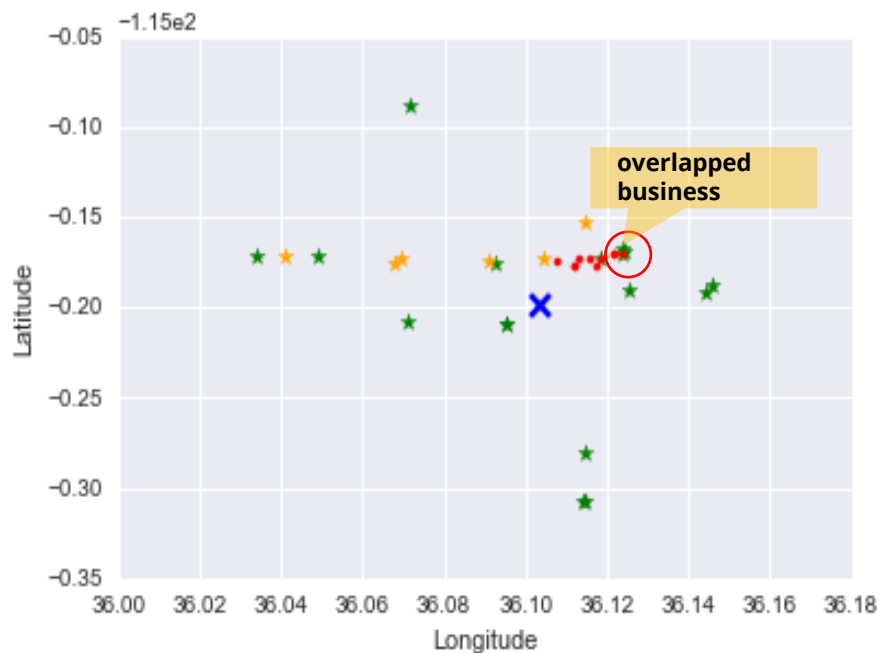
Tipped business



Reviewed business



Recommended business



$$RMSE_{100\ users} = 1,48$$



Thank you for attention!

Contacts:

i.abdrashitov.ds@gmail.com

radojkovic.n@gmail.com

Github: https://github.com/iabdrashitov/Recommendation_engine_yelp