Yelp

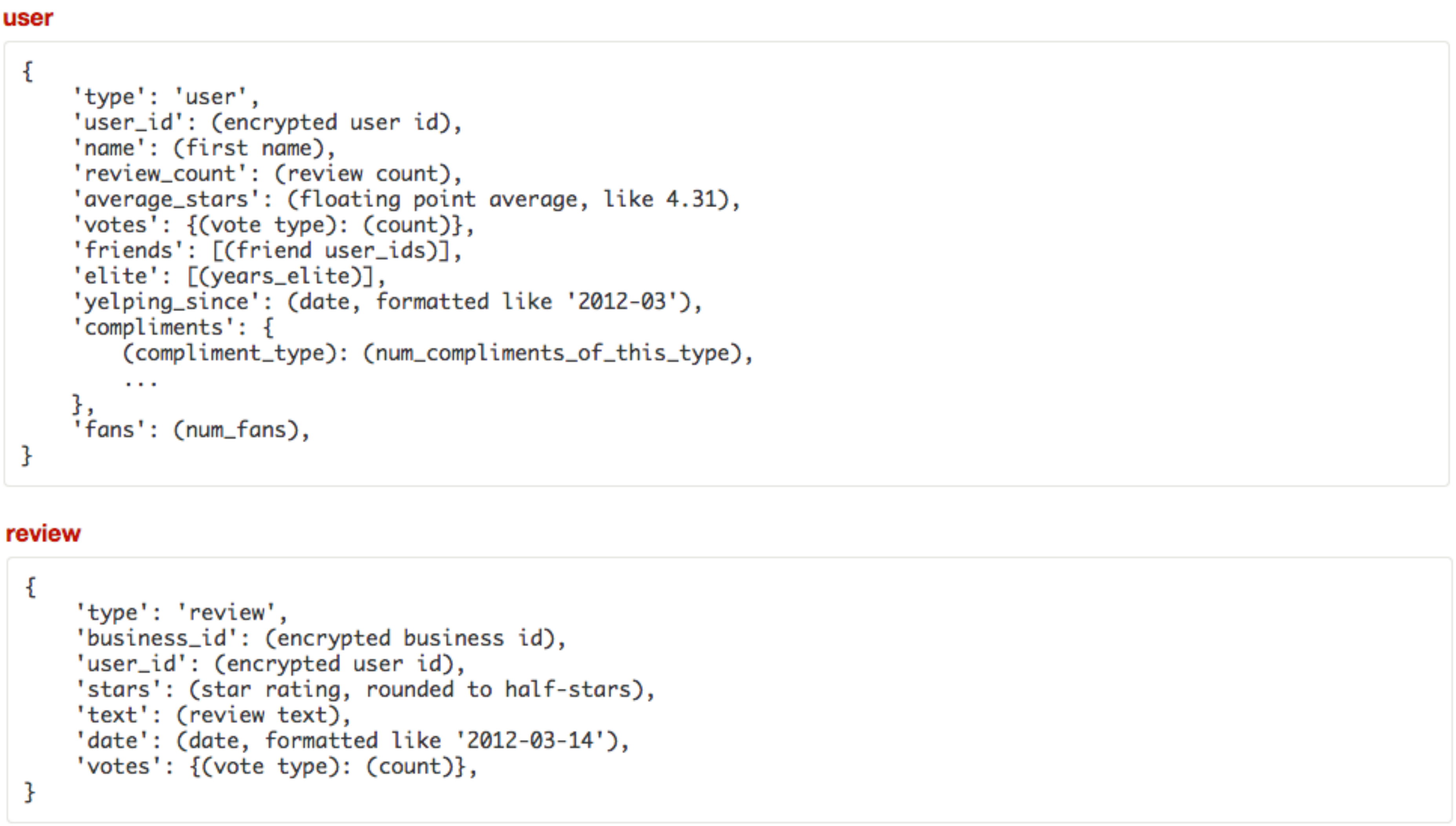
Mitsue Iwata, Kyle Magida, Scarlett Swerdlow

Data

- Yelp is releasing a sample of its data set as part of the Yelp Dataset Challenge
- 1.6M reviews and 500K tips by 366K users for 61K businesses in ten cities
- Only 1.5 GB

Go big

- Cosine similarity analysis of every pair of users or businesses, possibly in select cities
- 366K (users) choose 2 = 66B pairs
- 61K (businesses) choose 2 = 1.8B pairs

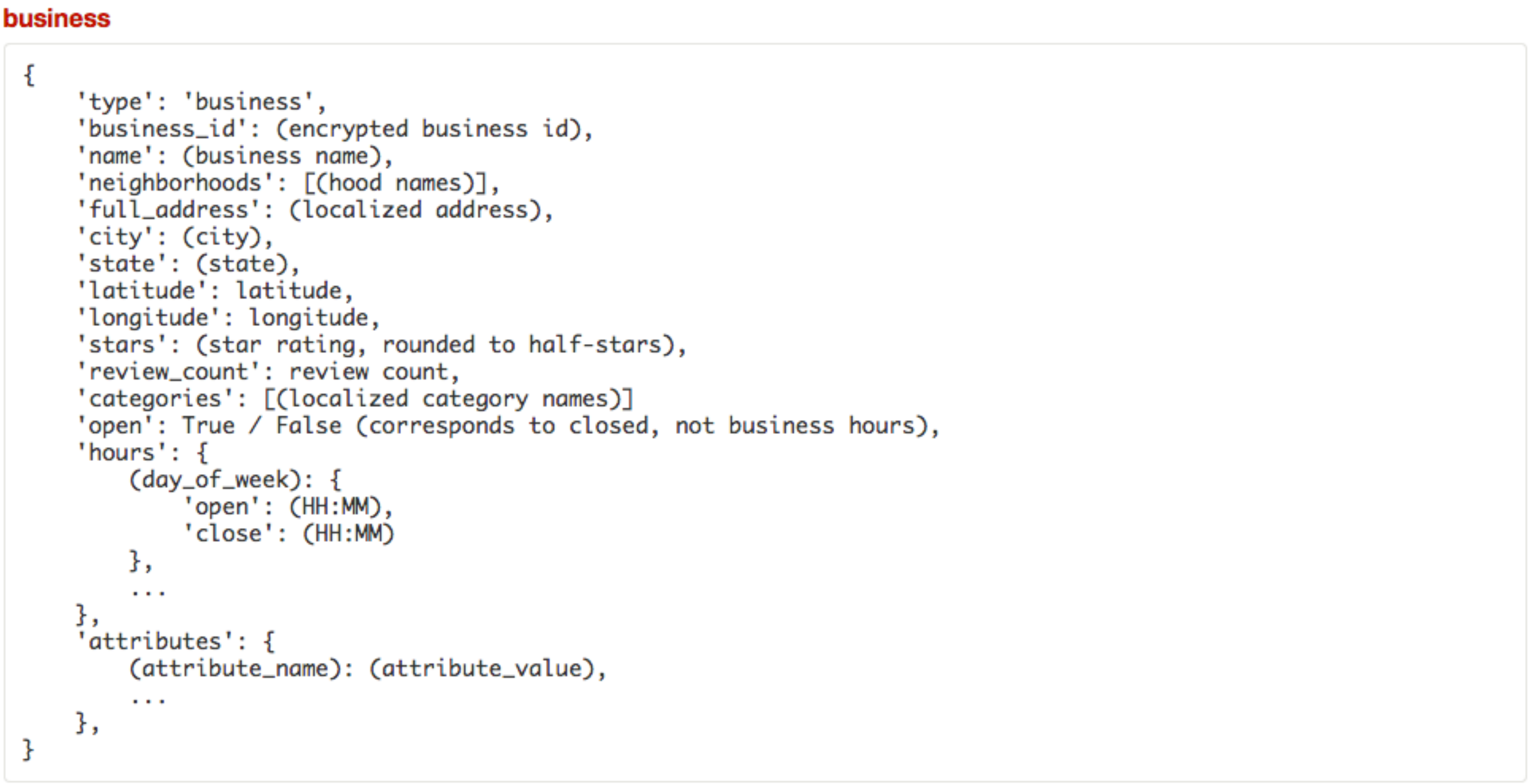


user

```
{
    'type': 'user',
    'user_id': (encrypted user id),
    'name': (first name),
    'review_count': (review count),
    'average_stars': (floating point average, like 4.31),
    'votes': {(vote type): (count)},
    'friends': [(friend user_ids)],
    'elite': [(years_elite)],
    'yelping_since': (date, formatted like '2012-03'),
    'compliments': {
        (compliment_type): (num_compliments_of_this_type),
        ...
    },
    'fans': (num_fans),
}
```

review

```
{
    'type': 'review',
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'stars': (star rating, rounded to half-stars),
    'text': (review text),
    'date': (date, formatted like '2012-03-14'),
    'votes': {(vote type): (count)},
}
```



business

```
{
    'type': 'business',
    'business_id': (encrypted business id),
    'name': (business name),
    'neighborhoods': [(hood names)],
    'full_address': (localized address),
    'city': (city),
    'state': (state),
    'latitude': latitude,
    'longitude': longitude,
    'stars': (star rating, rounded to half-stars),
    'review_count': review count,
    'categories': [(localized category names)]
    'open': True / False (corresponds to closed, not business hours),
    'hours': {
        (day_of_week): {
            'open': (HH:MM),
            'close': (HH:MM)
        },
        ...
    },
    'attributes': {
        (attribute_name): (attribute_value),
        ...
    },
}
```

Hypotheses

Do Yelp users who are similar based on cosine similarity:

- Share social networks?
- Go to the same types of places?
- Give similar ratings to businesses?
- Write similar reviews?

Timeline

- Implement cosine similarity analysis to identify similar users (week 5)
- Do exploratory data analysis on sample of data (week 7)
    - Each person takes 1-2 hypotheses
- Decide on hypothesis to test (week 7)
- Report on progress (week 7)

Timeline continued

- Refine algorithm for entire data set (week 8)
- Make mistakes and break program (weeks 9 and 10)
- Implement analysis on all data (week 10)
- Report findings (week 10)
- Submit findings to Yelp Dataset Challenge and win $5000 (week 11)