The challenge:

<https://www.yelp.com/dataset_challenge>

To get the data:

<https://www.yelp.com/dataset_challenge/dataset>

Downloaded data (.tgz).

Used Winzip to unzip. Files in JSON format. Will try to get info from smallest file (checkin)

Tried to use pandas.read\_json but error.

There is more material in <https://github.com/Yelp/dataset-examples>

Going to try to use added file: json\_to\_csv\_converter.py

It actually requires simplejson so I had to find it and install it via pip

simplejson-3.8.2-cp27-cp27m-win\_amd64.whl

It feels a little bit like cheating because I used provided code to convert it to csv … . But maybe ok

**The Challenge Dataset:**

* **2.2M** reviews and **591K** tips by **552K** users for **77K** businesses
* **566K** business attributes, e.g., hours, parking availability, ambience.
* Social network of **552K** users for a total of **3.5M** social edges.
* Aggregated check-ins over time for each of the **77K** businesses
* **200,000** pictures from the included businesses

## Notes on the Dataset

Each file is composed of a single object type, one json-object per-line.  
Take a look at some examples to get you started: <https://github.com/Yelp/dataset-examples>.

### 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),

...

},

}

### 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)},

}

### 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),

}

### check-in

{

'type': 'checkin',

'business\_id': (encrypted business id),

'checkin\_info': {

'0-0': (number of checkins from 00:00 to 01:00 on all Sundays),

'1-0': (number of checkins from 01:00 to 02:00 on all Sundays),

...

'14-4': (number of checkins from 14:00 to 15:00 on all Thursdays),

...

'23-6': (number of checkins from 23:00 to 00:00 on all Saturdays)

}, # if there was no checkin for a hour-day block it will not be in the dict

}

### tip

{

'type': 'tip',

'text': (tip text),

'business\_id': (encrypted business id),

'user\_id': (encrypted user id),

'date': (date, formatted like '2012-03-14'),

'likes': (count),

}

### photos (from the photos auxiliary file)

This file is formatted as a JSON list of objects.

[

{

"photo\_id": (encrypted photo id),

"business\_id" : (encrypted business id),

"caption" : (the photo caption, if any),

"label" : (the category the photo belongs to, if any)

},

{...}

]

## The Challenge

Not only would we like to give you our data, we’d also like to announce the seventh round of the **Yelp Dataset Challenge**. We challenge you to use this data in an innovative way and break ground in research. Here are some examples of topics we find interesting, but remember these are only to get you thinking and we welcome novel approaches!

**Cultural Trends:** By adding a diverse set of cities, we want participants to compare and contrast what makes a particular city different. For example, are people in international cities less concerned about driving in to a business, indicated by their lack of mention about parking? What cuisines are Yelpers raving about in these different countries? Do Americans tend to eat out late compared to the Germans and English? In which countries are Yelpers sticklers for service quality? In international cities such as Montreal, are French speakers reviewing places differently than English speakers?

I guess, generically: what do customers care about?

**Location Mining and Urban Planning:** How much of a business' success is really just location, location, location? Do you see reviewers' behavior change when they travel?

Very interesting question!

**Seasonal Trends:** What about seasonal effects: Are HVAC contractors being reviewed just at onset of winter, and manicure salons at onset of summer? Are there more reviews for sports bars on major game days and if so, could you predict that?

I assume businesses are categorized by type?

**Infer Categories:** Do you see any non-intuitive correlations between business categories e.g., how many karaoke bars also offer Korean food, and vice versa? What businesses deserve their own subcategory (i.e., Szechuan or Hunan versus just "Chinese restaurants"), and can you learn this from the review text?

Very interesting

**Natural Language Processing (NLP):** How well can you guess a review's rating from its text alone? What are the most common positive and negative words used in our reviews? Are Yelpers a sarcastic bunch? And what kinds of correlations do you see between tips and reviews: could you extract tips from reviews?

Very interesting

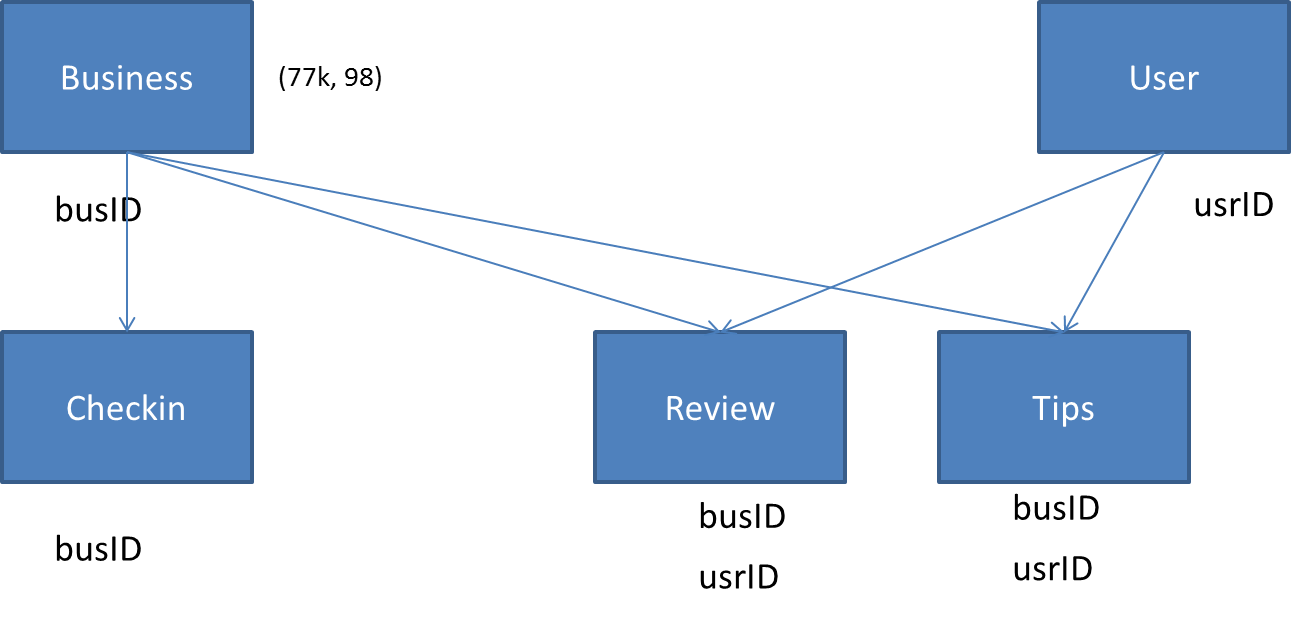
**Changepoints and Events:** Can you detect when things change suddenly (i.e. a business coming under new management)? Can you see when a city starts going nuts over cronuts?

**Social Graph Mining:** Can you figure out who the trend setters are and who found the best waffle joint before waffles were cool? How much influence does my social circle have on my business choices and my ratings?

**Data Structure**

The data consists of 5 sets (Figure 1).

Figure 1



**6/20/2016**

There are sample codes. Little clue about what they do. Need to take a deeper look: naïve bayes, markov chains, and something else.

**6/21/2016**

# One question could be:

# Ratings change for restaurants

# Bias US/EU in features: are there features that they care more/less about?

# Should translate into a supervised problem where target = nationality or continent