Welcome!
Please take a picture of this slide.

Copy the content on the usb drive to your laptop, and install Anaconda.

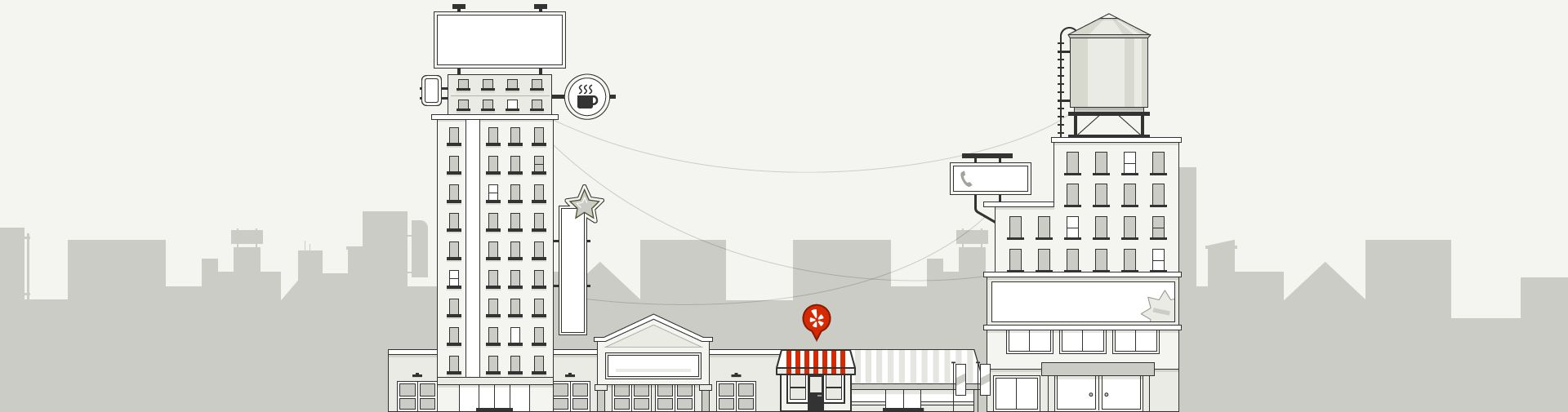
What's in the usb drive:

1. A software platform we'll use: Anaconda. Pick for your OS (Windows, Mac, Linux). Copy & install.
2. A dataset we'll play with: Yelp open data.
3. A github repo we'll cover. https://goo.gl/QjWtyS

WIFI: name is Cuckoo, password is redbirdpounceflag

A Hands-on Dive into Making Sense of Real World Data

Xun Tang
linkedin.com/in/xuntang | @whoisxun | xun@yelp.com



Plan for Today

-1-

Install Jupyter* & Learn how to use it

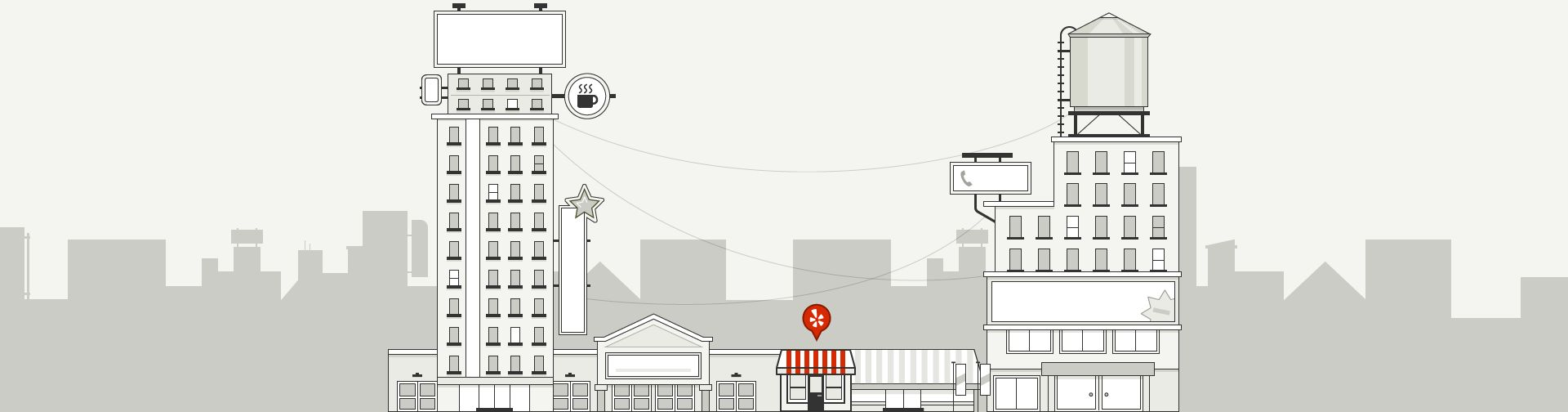
*** We recommend installing via Anaconda distribution**

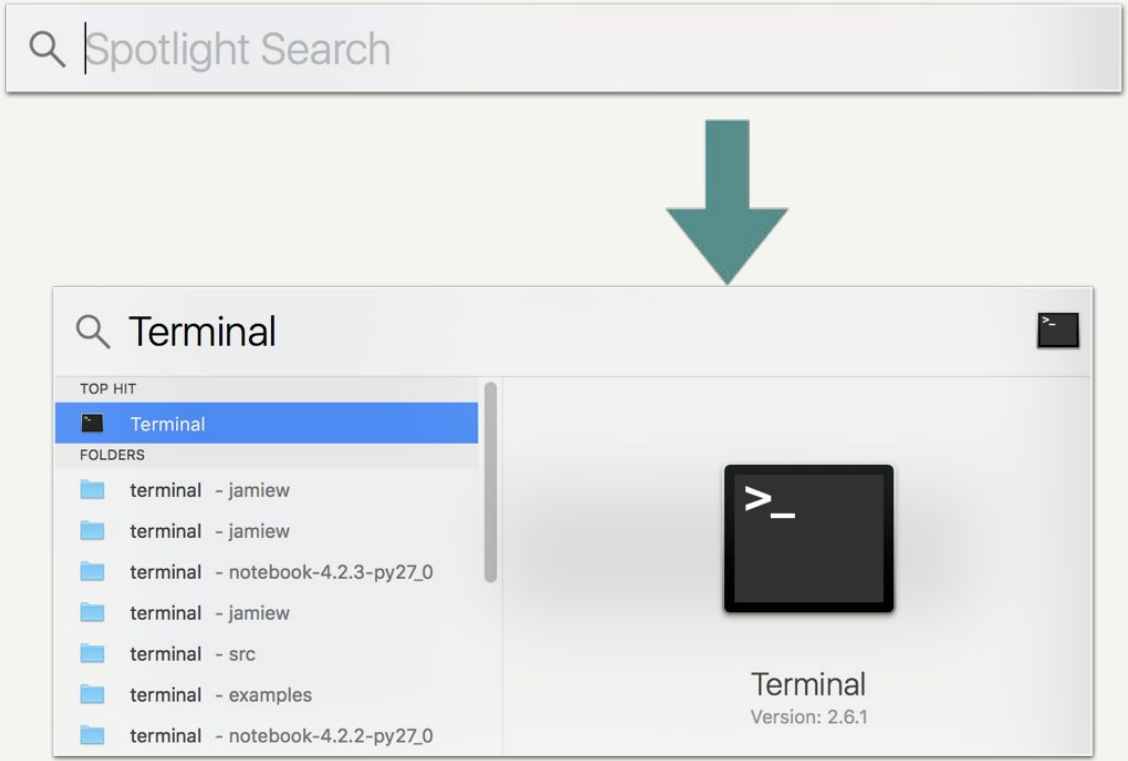
-2-

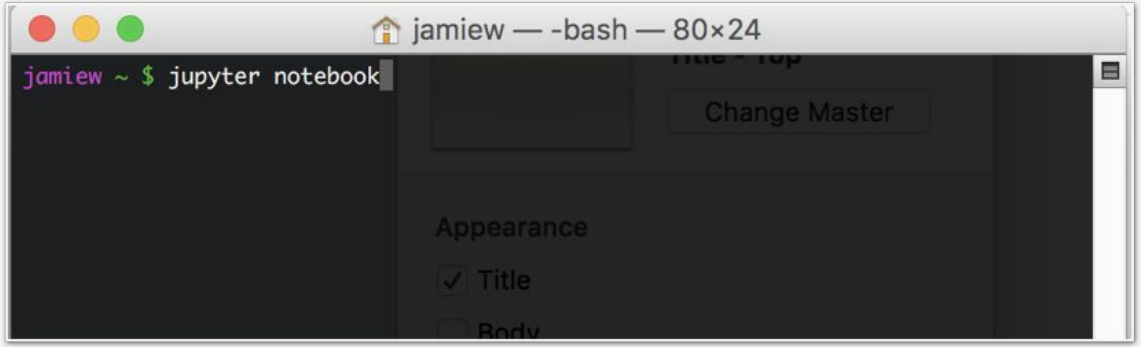
Complete a machine learning case study using the Jupyter Notebook & Yelp open dataset

Github: https://goo.gl/QjWtyS

Let’s Get Started

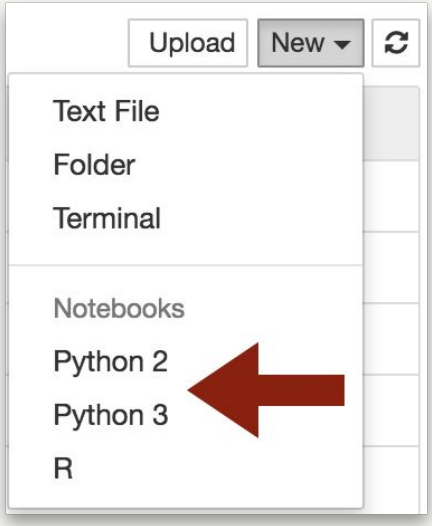



Spotlight Search
Terminal
TOP HIT
Terminal
FOLDERS
terminal - jamiew
terminal - jamiew
terminal - notebook-4.2.3-py27_0
terminal - jamiew
terminal - src
terminal - examples
terminal - notebook-4.2.2-py27_0
Terminal
Version: 2.6.1

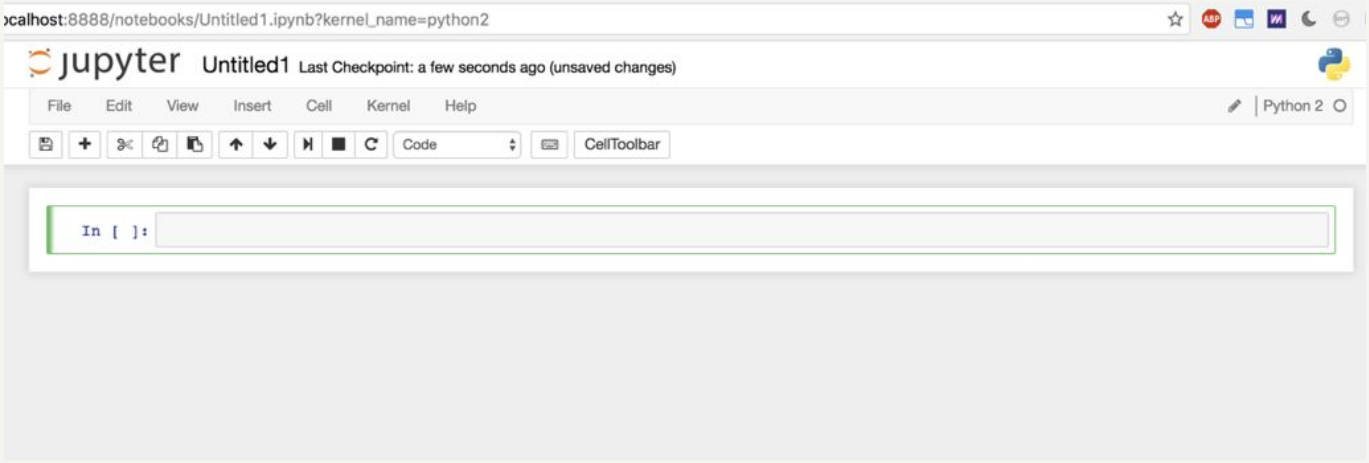

jamiew — -bash — 80×24
jamiew ~ $ jupyter notebook
Change Master
Appearance
Title

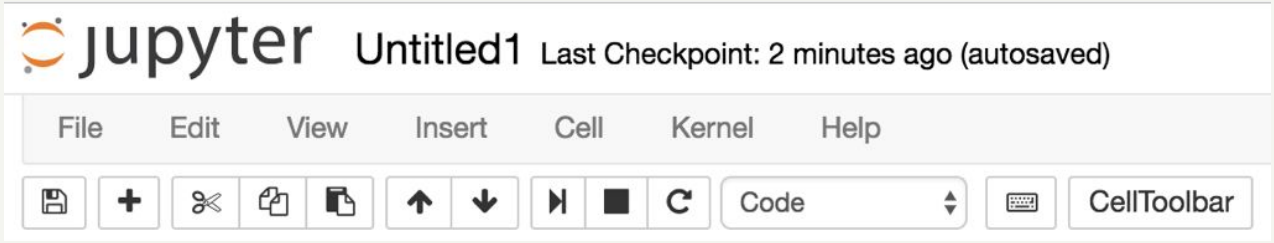

localhost:8888/tree
jupyter
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
2015
anaconda
Applications
Biorepository Legacy Migration
bokeh-notebooks-master
book-exercises
Coursera-HowToUseGitandGitHub
datasciencecoursera
Desktop
development
Documents


localhost:8888/tree
jupyter
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
2015
anaconda
Applications


Upload
New
Text File
Folder
Terminal
Notebooks
Python 2
Python 3
R

http://wikiclipart.com/candy-corn-clipart_1783/


calhost:8888/notebooks/Untitled1.ipynb?kernel_name=python2
jupyter
Untitled1
Last Checkpoint: a few seconds ago (unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Help
Python 2
Code
CellToolbar
In [ ]:


jupyter
Untitled1
Last Checkpoint: 2 minutes ago (autosaved)
File
Edit
View
Insert
Cell
Kernel
Help
Code
CellToolbar

Warm Up

```
#Define a variable
x = 5
print(x)
```

Importing Libraries

import ____ as _____

CONVENTIONS

import matplotlib.pyplot as plt

%matplotlib inline

import pandas as pd

import seaborn as sns

from sklearn import _____

Reading in Data - Pandas

- Reading data in / querying data sources / writing data out

```
import pandas as pd

PATH = '/scratch/xun/docs/yelp_dataset_challenge_round11/'
biz_df = pd.read_csv(PATH + 'business.csv')
user_df = pd.read_csv(PATH + 'user.csv')
review_df = pd.read_csv(PATH + 'review.csv')
```

Did it work?

```
review_df.head()
```

review_df.head(): Print top rows in the data frame.

review_df.describe(): Generate various summary statistics, mean, max, count, etc.

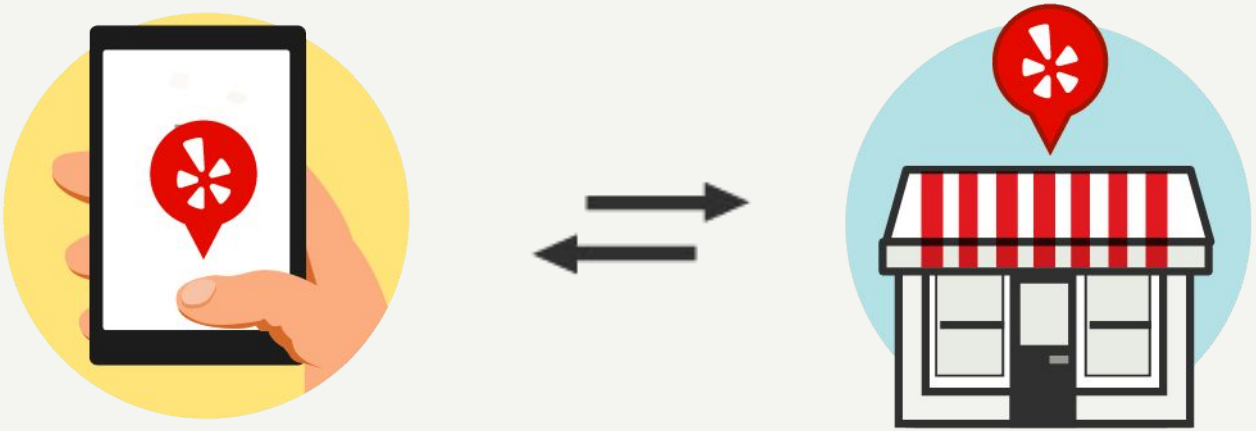
```
review_df.head()
```

	funny	user_id	text	business_id	stars	useful	type	cool	datetime	year
review_id										
NxL8SIC5yqOdnlXCg18IBg	0	KpkOkG6RIf4Ra25Lhhxf1A	If you enjoy service by someone who is as comp...	2aFiy99vNLklCx3T_tGS9A	5	0	review	0	2011-10-10	2011
pXbbIgOXvLuTi_SPs1hQEQ	0	bQ7fQq1otn9hKX-gXRsrgA	After being on the phone with Verizon Wireless...	2aFiy99vNLklCx3T_tGS9A	5	1	review	0	2010-12-29	2010
wslW2Lu4NYylb1jEapAGsw	0	r1NUhdNmL6yU9Bn-Yx6FTw	Great service! Corey is very service oriented....	2aFiy99vNLklCx3T_tGS9A	5	0	review	0	2011-04-29	2011
GP6YEearUWrzPtQYSF1vVg	0	aW3ix1KNZAvoM8q-WghA3Q	Highly recommended. Went in yesterday looking ...	2LfluF3_sX6uwe-IR-P0jQ	5	0	review	1	2014-07-14	2014
25RlYGq2s5qShi-pn3ufVA	0	YOo-Cip8HqvKp_p9nEGphw	I walked in here looking for a specific piece ...	2LfluF3_sX6uwe-IR-P0jQ	4	0	review	0	2014-01-15	2014

```
biz_df.describe()
```

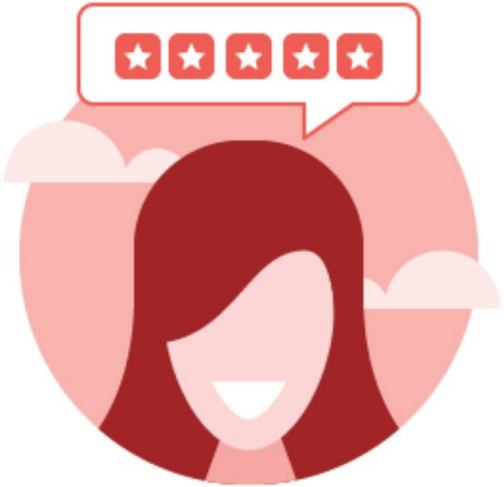
Yelp's Mission

Connecting people with great local businesses.



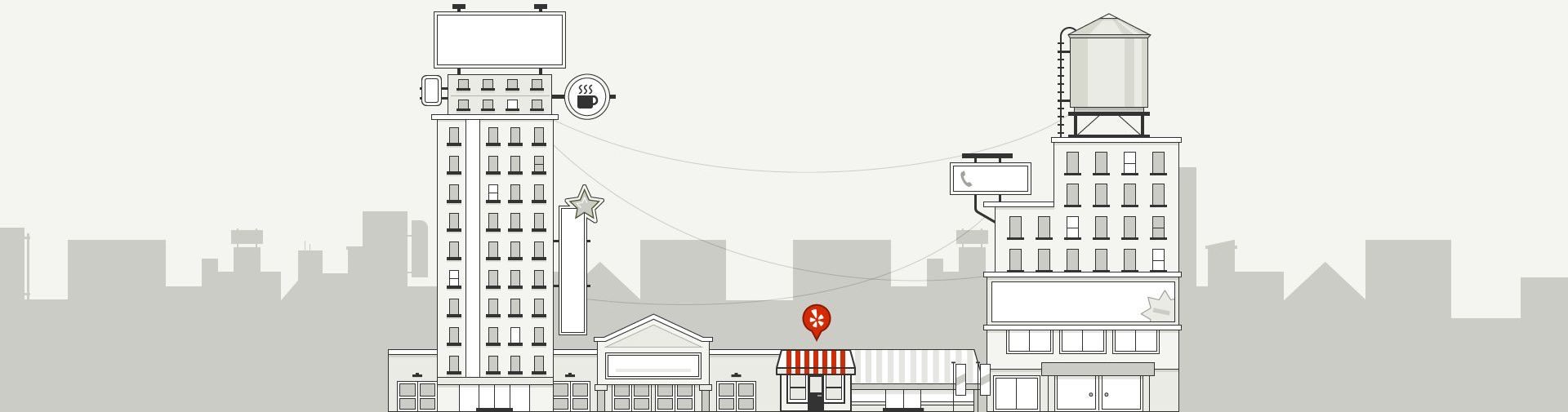
Yelp Open Dataset

yelp.com/dataset



5,200,000 reviews **174,000 businesses** **200,000 pictures**

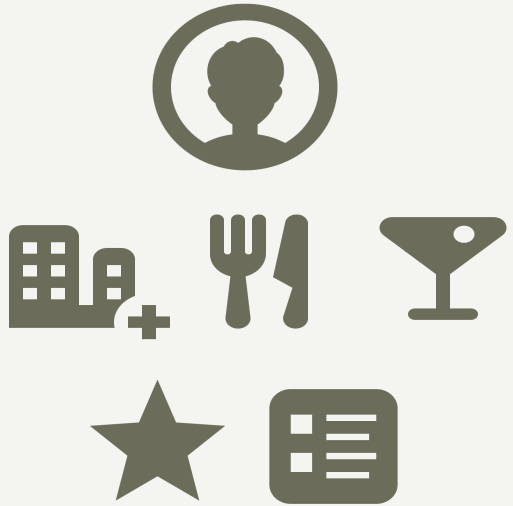
Will Your Next yelp Review Be a ★★★★★ Review?



Given users' past reviews on Yelp

yelp Open Dataset

When the user writes a review for a business she hasn't reviewed before



Will it be a ★★★★★ review? (True / False)

To Solve a Data Science Problem

Step 1: Load the Data

Step 2: Explore and Visualize the Data

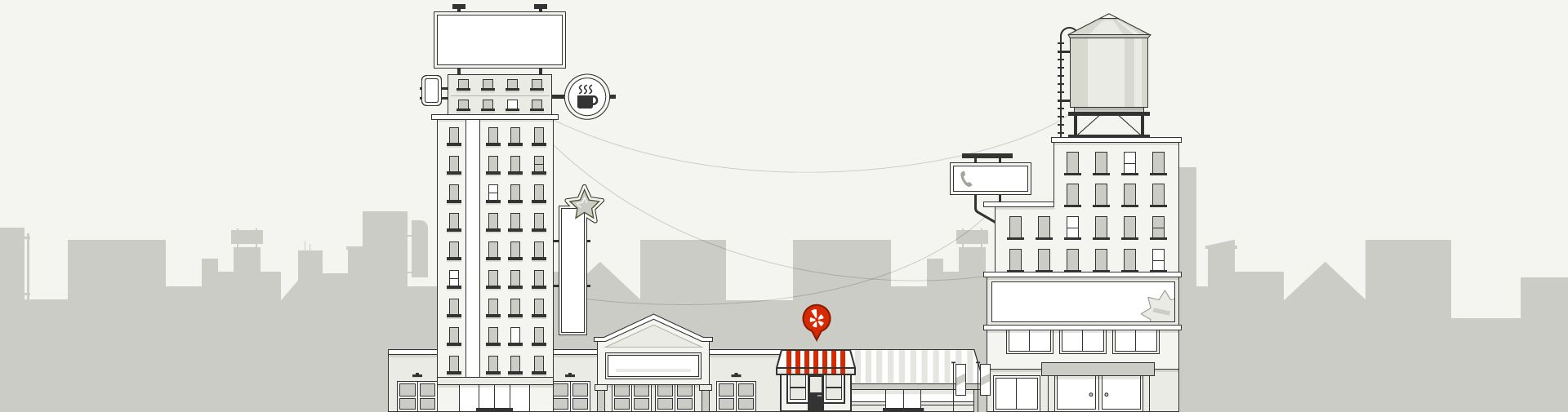
Step 3: Generate the Features

Step 4: Train a Model

Step 5: Evaluate the Model

Step 6 & Beyond: Iterate Through the Process

Step 1
Load the Data



Store Data in Pandas DataFrames

```
import pandas as pd

PATH = '/scratch/xun/docs/yelp_dataset_challenge_round10/'
biz_df = pd.read_csv(PATH + 'business.csv')
user_df = pd.read_csv(PATH + 'user.csv')
review_df = pd.read_csv(PATH + 'review.csv')
```

```
review_df = review_df.set_index('review_id')
user_df = user_df.set_index('user_id')
biz_df = biz_df.set_index('business_id')
```

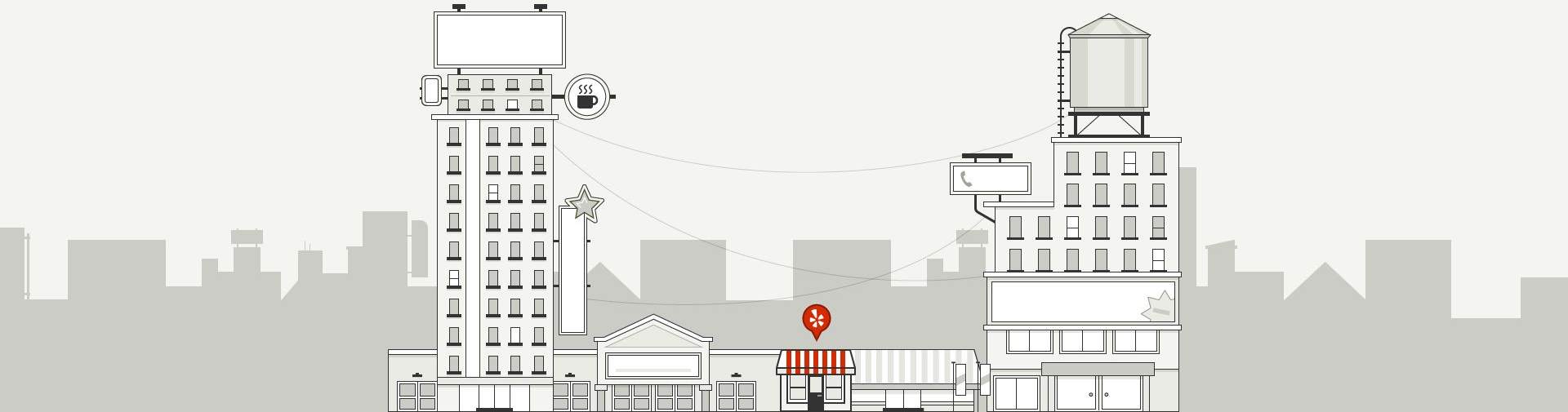
What's in a Review DataFrame?

```
review_df.head()
```

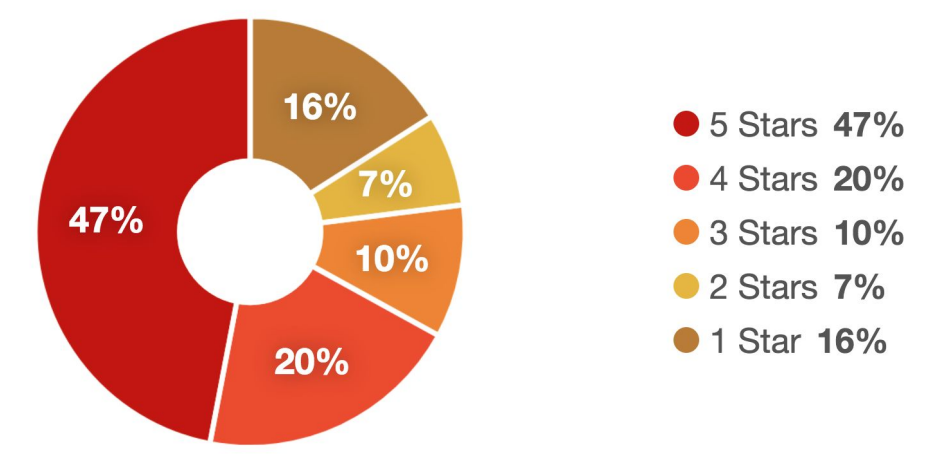
	funny	user_id	text	business_id	stars	useful	type	cool	datetime	year
review_id										
NxL8SIC5yqOdnlXCg18IBg	0	KpkOkG6RIf4Ra25Lhhxf1A	If you enjoy service by someone who is as comp...	2aFiy99vNLklCx3T_tGS9A	5	0	review	0	2011-10-10	2011
pXbbIgOXvLuTi_SPs1hQEQ	0	bQ7fQq1otn9hKX-gXRsrgA	After being on the phone with Verizon Wireless...	2aFiy99vNLklCx3T_tGS9A	5	1	review	0	2010-12-29	2010
wslW2Lu4NYylb1jEapAGsw	0	r1NUhdNmL6yU9Bn-Yx6FTw	Great service! Corey is very service oriented....	2aFiy99vNLklCx3T_tGS9A	5	0	review	0	2011-04-29	2011
GP6YEearUWrzPtQYSF1vVg	0	aW3ix1KNZAvoM8q-WghA3Q	Highly recommended. Went in yesterday looking ...	2LfluF3_sX6uwe-IR-P0jQ	5	0	review	1	2014-07-14	2014
25RlYGq2s5qShi-pn3ufVA	0	YOo-Cip8HqvKp_p9nEGphw	I walked in here looking for a specific piece ...	2LfluF3_sX6uwe-IR-P0jQ	4	0	review	0	2014-01-15	2014

```
biz_df.describe()
```

Step 2
Explore and Visualize the Data



Review Star Rating Distribution Published on Yelp's Factsheet

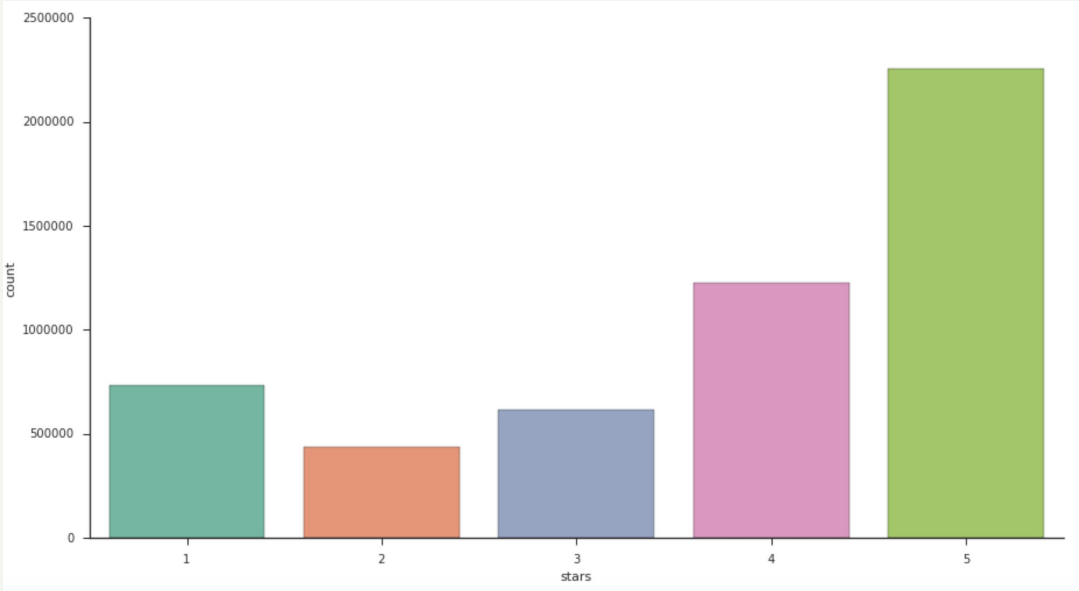


Source: https://www.yelp.com/factsheet

Plot Review Star Rating Distribution from Open Dataset

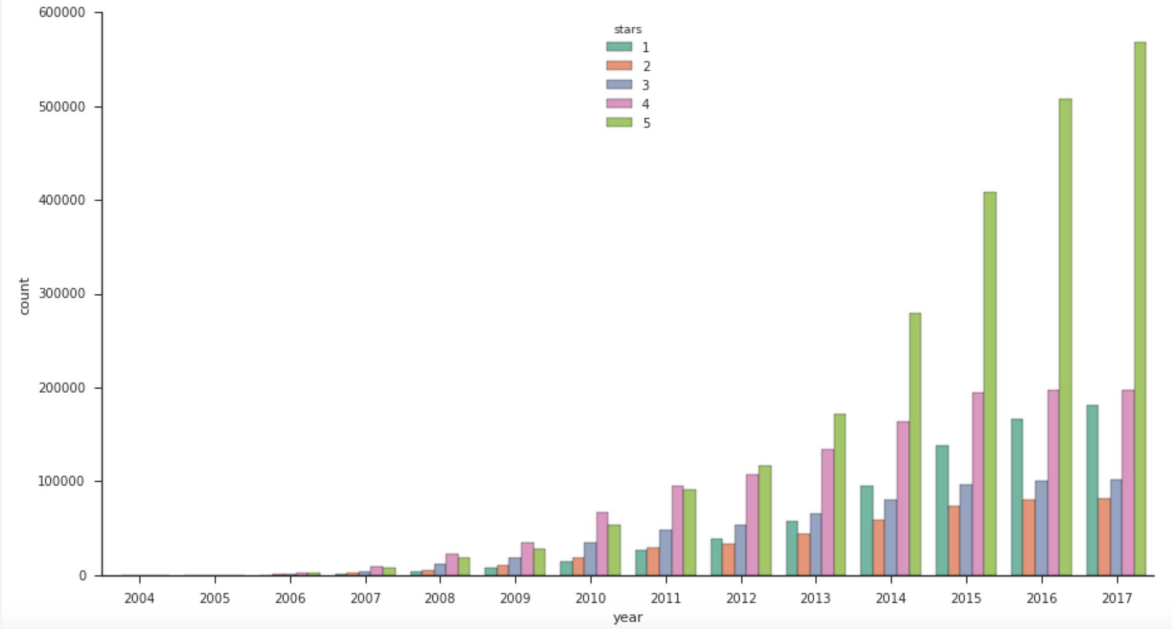
```
import seaborn as sns
%matplotlib inline
```

```
ax = sns.countplot(x='stars', data=review_df)
```

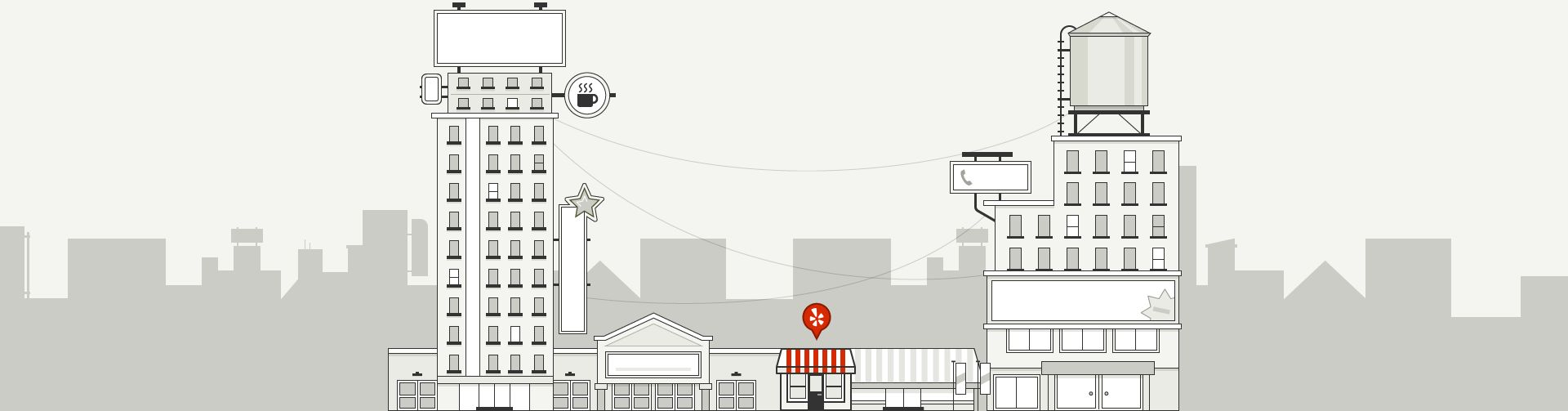


Plot Star Ratings by Year

```
ax = sns.countplot(x='year', data=review_df, hue='stars')
```



Step 3
Generate the Features



For Example..

Convert date string to date delta

e.g. business_age

Convert strings to categorical features

e.g. noise level: {'quiet', 'loud', 'very loud'}.

Drop unused features

e.g. business_name

```
def calculate_date_delta(df, from_column, to_column):
    datetime = pd.to_datetime(df[from_column])
    time_delta = datetime.max() - datetime
    df[to_column] = time_delta.apply(lambda x: x.days)
    df.drop(from_column, axis=1, inplace=True)
```

```
def to_length(df, from_column, to_column):
    df[to_column] = df[from_column].apply(lambda x: len(x))
    df.drop(from_column, axis=1, inplace=True)
```

```
def drop_columns(df, columns):
    for column in columns:
        df.drop(column, axis=1, inplace=True)
```

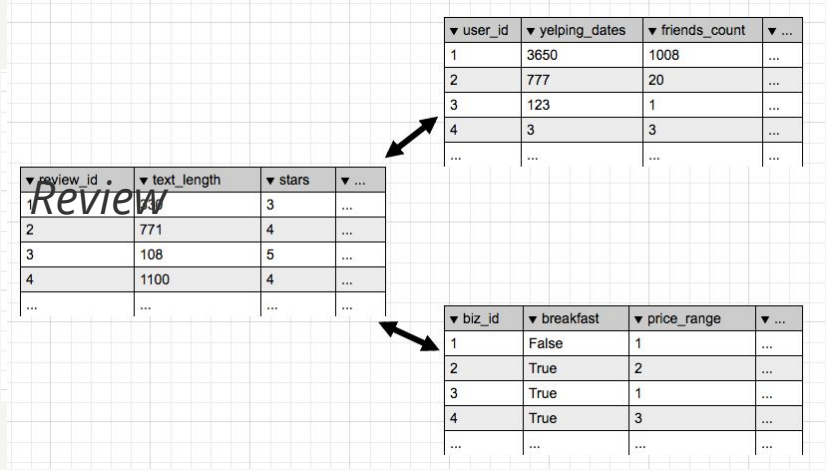
```
def to_boolean(df, columns):
    for column in columns:
        to_column = column+'_bool'
        df[to_column] = df[column].apply(lambda x: bool(x))
        df.drop(column, axis=1, inplace=True)
```

```
FILL_WITH = 0.0
```

```
def to_category(df, columns):
    for column in columns:
        df[column] = df[column].astype('category')
        # add FILL_WITH category for fillna() to work w/o error
        if (FILL_WITH not in df[column].cat.categories):
            df[column] = df[column].cat.add_categories([FILL_WITH])
        #print 'categories for ', column, ' include ', df[column].cat.c
```

```
def category_rename_to_int(df, columns):
    for column in columns:
        df[column].cat.remove_unused_categories()
        size = len(df[column].cat.categories)
        #print 'column ', column, ' has ', size, ' columns, include ', c
        df[column] = df[column].cat.rename_categories(range(1, size+1))
        #print 'becomes ', df[column].cat.categories
```

Join DataFrames to Populate the Features



Review

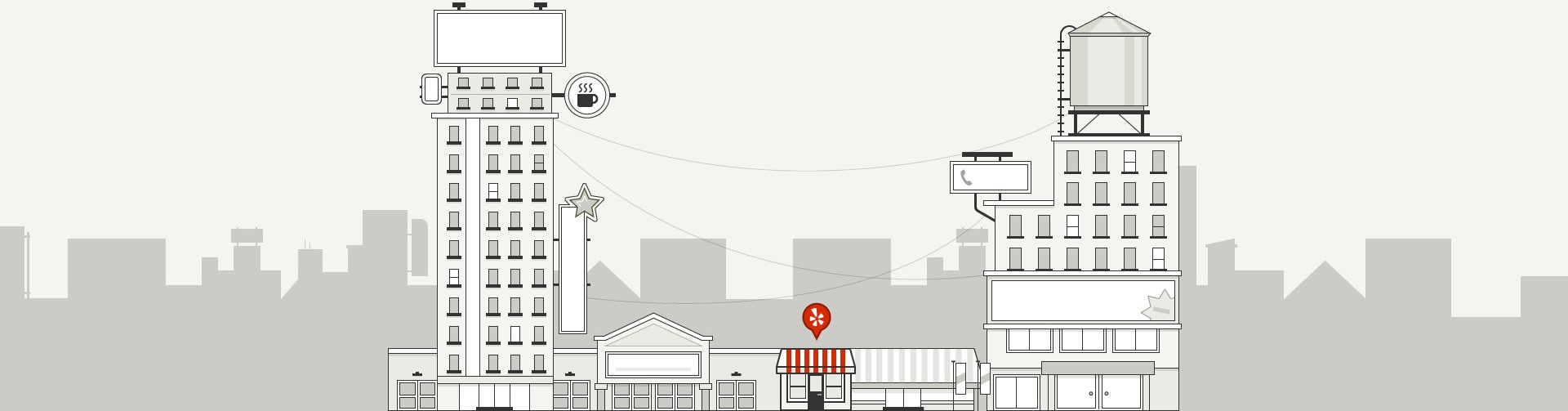
User

Biz

```
# The `user_df` DataFrame is already indexed by the join key (`user_id`). Make sure it's on t
review_join_user = review_df.join(user_df, on='user_id', lsuffix='_review', rsuffix='_user')
```

```
review_join_user_join_biz = review_join_user.join(biz_df, on='business_id', rsuffix='_biz')
```

Step 4
Train a Model



Arrange Data into a Feature Matrix and a Target Array

Feature matrix (X)

All features generated from biz, user, review dataframes

Target array (y)

What we predict: Whether the review is Five-star or not

```
# Target y is whether a review is five-star (True / False)
y = review_join_user_join_biz.stars.apply(lambda x: x == 5)

# Exclude the `stars` columns from the feature matrix, since it is the target
X = review_join_user_join_biz
review_join_user_join_biz.drop('stars', axis=1, inplace=True)
```

Split Training and Testing Set

Training set: used for an machine learning algorithm to train from

Testing set: used to to estimate / evaluate how well the model has been trained

Split them s.t. we don't evaluate on the same dataset we train from

```
from sklearn.cross_validation import train_test_split

# Split the data into a training set and a test set
X_train, X_test, y_train, y_test = train_test_split(X, y)
```

```
training data shape (3552672, 109)
test data shape (1184225, 109)
converted label data shape (3552672,)
```

Model: Logistic Regression (LR)

Estimates the probability of a **binary** response based on the features

Here we estimate the probability of a review being five-star

Normalize the Features

Standardize features by removing the mean and scaling to unit variance

Logistic Regression requires all features normalized

```
from sklearn import preprocessing

scaler = preprocessing.StandardScaler().fit(X_train)

X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Cross Validation

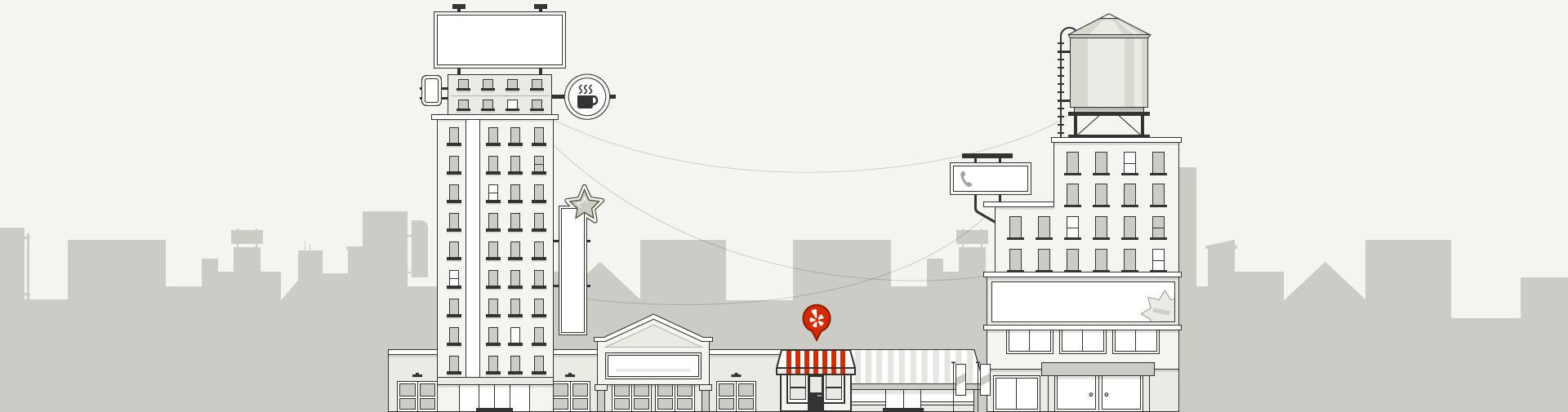
Holding out a portion of the training data for model validation, and do this for `n_folds`

- Ensure that the model does not overfit the training data
- Select optimal model parameters

```
from sklearn.cross_validation import StratifiedKFold

# cross-validation
cv = StratifiedKFold(y_train, n_folds=5, shuffle=True)
```

Step 5
Evaluate the Model



Metrics

Accuracy:
Percentage of labels correctly predicted. The higher the better.

```
from sklearn.cross_validation import cross_val_score
import numpy as np

# Function used to print cross-validation scores
def training_score(est, X, y, cv):
    acc = cross_val_score(est, X, y, cv = cv, scoring='accuracy')
    roc = cross_val_score(est, X, y, cv = cv, scoring='roc_auc')
    print '5-fold Train CV | Accuracy:', round(np.mean(acc), 3),'+/-', \
    round(np.std(acc), 3),'| ROC AUC:', round(np.mean(roc), 3), '+/-', round(np.std(roc), 3)
```

```
# print cross-validation scores
training_score(est=lrc, X=X_train_scaled, y=y_train, cv=cv)
```

```
5-fold Train CV | Accuracy: 0.76 +/- 0.001 | ROC AUC: 0.836 +/- 0.001
```

Evaluation via Confusion Matrix

Does It Work?

Given users' past reviews on Yelp

When the user writes a review for a business she hasn't reviewed before

Will it be a ★★★★★ review? (True / False)

Given users' past reviews on Yelp

```
user1 = user_df[user_df.index == 'kEtR1ZVL3Xr-tEX7lg16dQ']
#print user1.review_count
print user1.average_stars
```

```
user_id
kEtR1ZVL3Xr-tEX7lg16dQ    4.96
Name: average_stars, dtype: float64
```

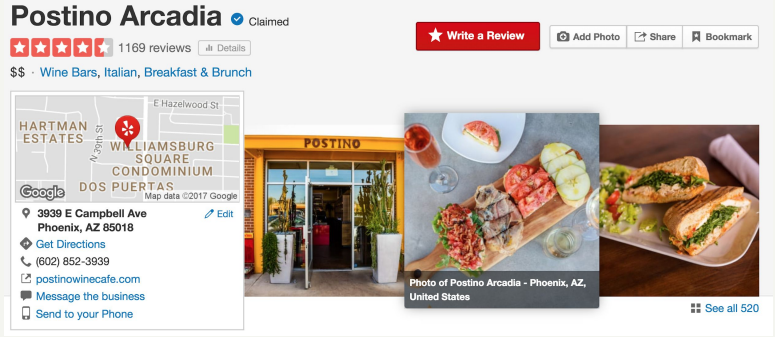
```
user2 = user_df[user_df.index == 'Hj20fg3vyzKnJwnLn_rMqw']
#print user2.review_count
print user2.average_stars
```

```
user_id
Hj20fg3vyzKnJwnLn_rMqw    4.55
Name: average_stars, dtype: float64
```

```
user3 = user_df[user_df.index == 'om5ZiponkpRqUNa3pVPiRg']
#print user2.review_count
print user3.average_stars
```

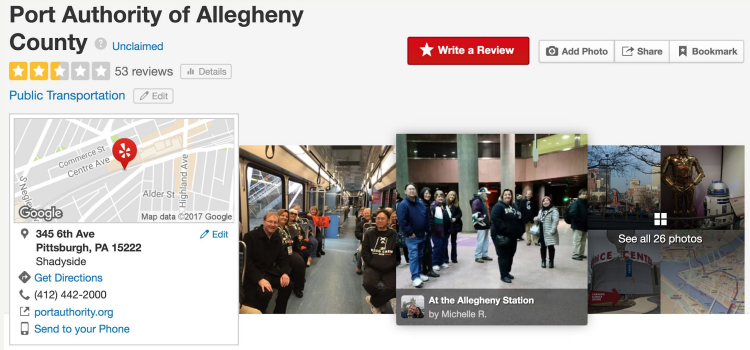
```
user_id
om5ZiponkpRqUNa3pVPiRg    3.94
Name: average_stars, dtype: float64
```

When the user writes a review for a business she hasn't reviewed before



biz1

biz2



Will it be a  review?

Make predictions for user[1,2,3]'s review on biz1

```
predict_given_user_biz(user=user1, biz=biz1, review_df=review_df)
predict_given_user_biz(user=user2, biz=biz1, review_df=review_df)
predict_given_user_biz(user=user3, biz=biz1, review_df=review_df)
```

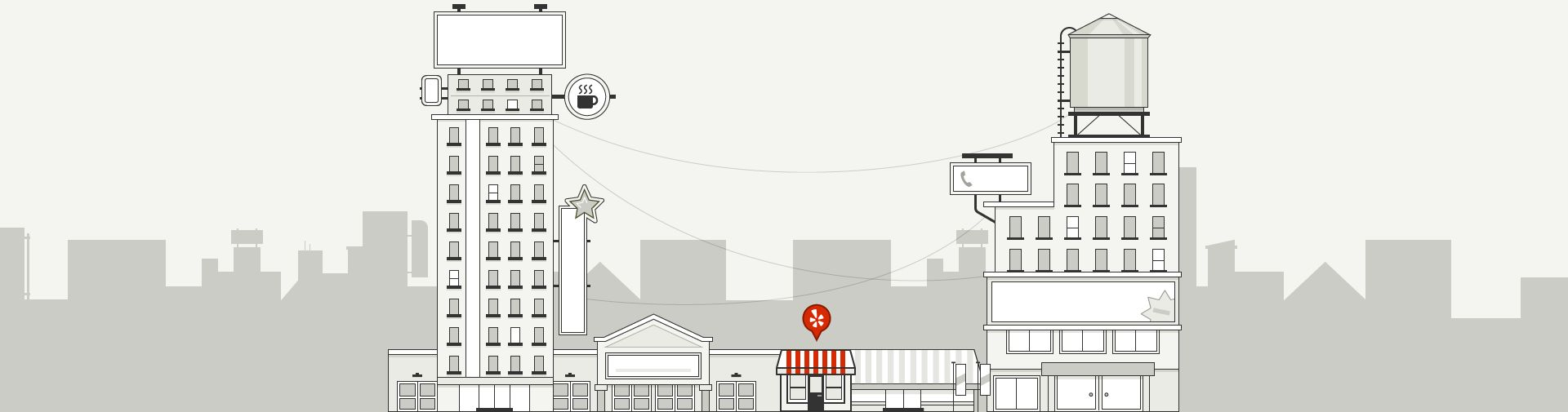
```
True , with probability [False, True] ==  [ 0.07176054  0.92823946]
True , with probability [False, True] ==  [ 0.17819623  0.82180377]
False , with probability [False, True] ==  [ 0.86403705  0.13596295]
```

Make predictions for user[1,2,3]'s review on biz2

```
predict_given_user_biz(user=user1, biz=biz2, review_df=review_df)
predict_given_user_biz(user=user2, biz=biz2, review_df=review_df)
predict_given_user_biz(user=user3, biz=biz2, review_df=review_df)
```

```
True , with probability [False, True] ==  [ 0.29932791  0.70067209]
False , with probability [False, True] ==  [ 0.5450868  0.4549132]
False , with probability [False, True] ==  [ 0.97231228  0.02768772]
```

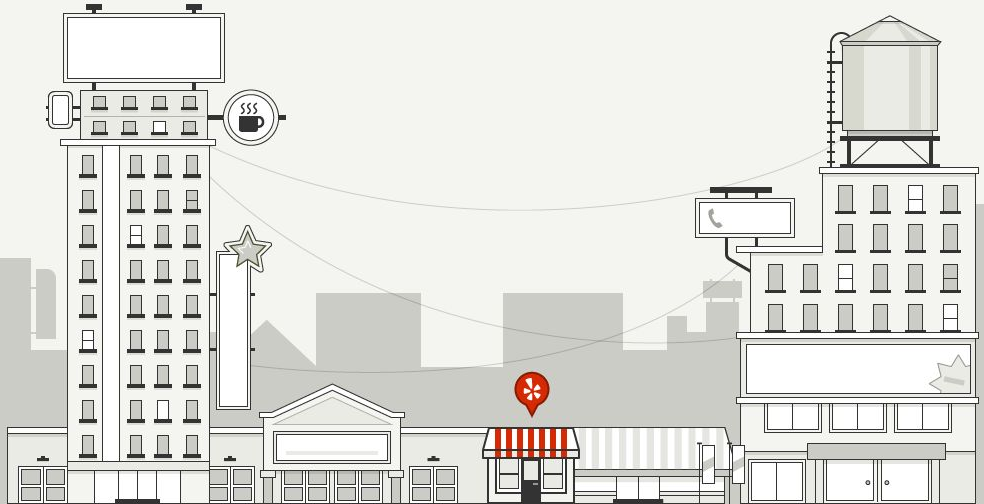
Step 6 & Beyond
Iterate Through the Process



Yelp Dataset Challenge
Round 11

yelp.com/dataset/challenge

January 18, 2018 ~ June 30, 2018



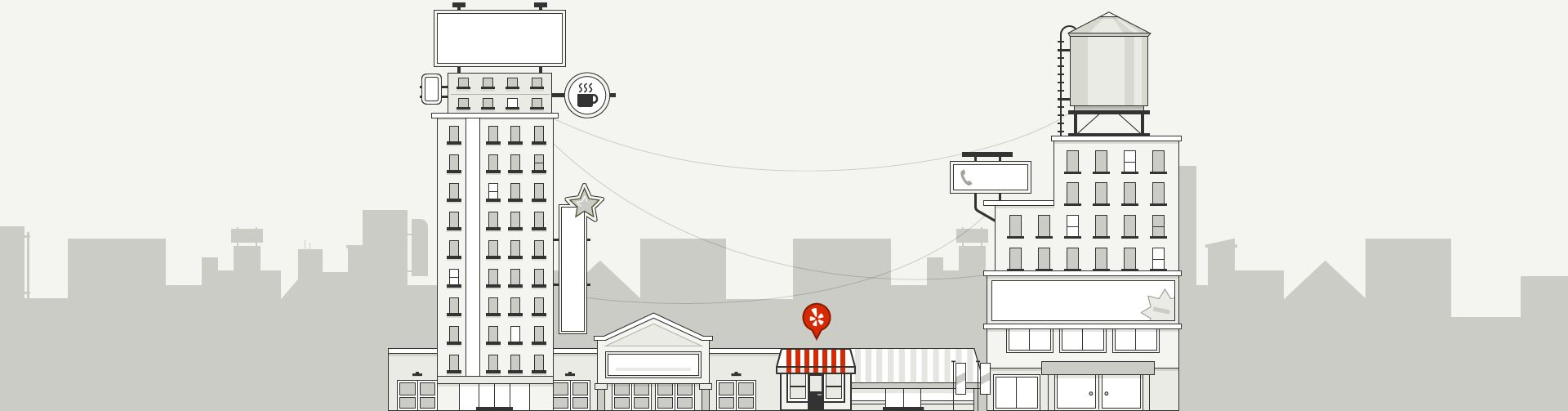
Questions?

Repo: https://goo.gl/QjWtyS

LinkedIn: linkedin.com/in/xuntang

Email: xun@yelp.com

Twitter: @whoisxun




We're Hiring!
www.yelp.com/careers/

fb.com/YelpEngineers

@YelpEngineering

engineeringblog.yelp.com

github.com/yelp