* **Introduction/Motivation/Problem Definition (15%)**
  + What is it that you are trying to solve/achieve? Who cares and why does it matter?
  + Identify, define, and motivate the problem that you are addressing.
  + How (precisely) will a machine learning solution address the problem?

We are trying to predict the annual percentage growth in value of urban zones (defined by postal codes), based on observed growth rates and the public perception of the zone according to the businesses located in it.

Real state owners would be benefited from this prediction since they would be able to make more informed decisions when planning to sell their properties or invest in a certain urban zone. Also, business owners and entrepreneurs could plan to open a new business in a zone that is predicted to experiment a significant amount of growth in value.

On the other hand, city Government could deploy measures to prevent a decrease in value of zones with predicted high probability of experiencing a negative percentage growth in value.

Having historic information for each urban zone, we will be dealing with a supervised learning. Using machine learning we will be able to develop a model that predicts the expected value growth for each zone in the following year. By fitting multiple machine learning algorithms, we will be able to select the model that delivers a low prediction error.

* **Data Understanding and Preparation (15%)**
  + Identify and describe the data (and data sources) that will support machine learning to address the problem.
  + Include various aspects of the data such as its size (GB/TB/etc), type(s), format, etc.
  + Specify how these data are integrated to produce the format required for machine learning.

Our data comes from 2 sources:

1. **Yelp’s challenge dataset**: the dataset was provided to us as a TAR file of 1.8 GB, with 6 files as json-objects. Two of the six files were selected for our project:

* **Business dataset:** information from 144,072 business with 16 attributes
* **Reviews dataset:** information from 4,153,150 reviews with 10 attributes

1. **Zillow research data:** provided as CVS files. The selected file for our project was:

* **Median List Price Per Sq Ft:** Median of list prices divided by the square footage of a home for all postal codes in the US (10,651 postal codes). The file’s size was 5.7 MB.

Using Python libraries we were able to manipulate the data to select values corresponding to the Greater Pittsburgh area and to aggregate it by postal code and year. Our final data frame contains information from 88 postal codes, which multiplied by the years where each had reviews from Yelp users, result in 634 data points.

The data frame is composed of 17 predictive attributes:

* **Postal\_Code:** used as the joining attribute between the Yelp and the Zillow datasets. Represent the urban areas for which the growth prediction is intended.
* **Year:** used as the joining attribute between the Yelp and the Zillow datasets. Years where a postal code had reviews from Yelp’s users and median list price per square foot information from Zillow.
* **Useful:** average number of useful ratings per review of Yelp’s users for a business in the postal code.
* **Funny:** average number of funny ratings per review of Yelp’s users for a business in the postal code.
* **Cool:** average number of cool ratings per review of Yelp’s users for a business in the postal code.
* **Price:** average price range for businesses in the postal code.
* **Stars\_1:** percentage of reviews with 1 star rating.
* **Stars\_2:** percentage of reviews with 2 stars rating.
* **Stars\_3:** percentage of reviews with 3 stars rating.
* **Stars\_4:** percentage of reviews with 4 stars rating.
* **Stars\_5:** percentage of reviews with 5 stars rating.
* **Stars:** average star rating for the business in the postal code
* **Prev\_stars:** average star rating for the business in the postal code in the previous year
* **Reviews:** percentage of total reviews of the year corresponding to the postal code.
* **MedPriceGrowth\_sqft(t):** growth inmedian list price per square foot in the current year compared to the previous year
* **MedPriceGrowth\_sqft(tminus1):** growth inmedian list price per square foot in the previous year compared to one year before

And the value to be predicted:

* **MedPriceGrowth\_sqft(tplus1):** growth inmedian list price per square foot in the next year compared to the current year.
* **Methodology (30%)**  
  This is where you give a detailed description of your primary contributions. It is especially important that this part be clear and well written so that we can fully understand what you did.
  + How did you approach the problem? What challenges did you face? In what (unique) ways did you handle those challenges?
  + Specify the type of model(s) built and/or information/knowledge extracted.
  + Discuss choices for machine learning algorithm: what are other alternatives, and what are their pros and cons (in the context of the problem and as compared to your proposed solution)?
  + Discuss why and how this model should "solve" the problem (i.e., improve along some dimension of interest).

The main challenge was the construction of the data frame…

We performed descriptive statistics to understand the composition of each original dataset and how we could aggregate and combine information to obtain the final data frame

We originated new attributes from existing ones.

We constructed a simple baseline model to evaluate the performance of our machine learning algorithms. The model consists in calculating the mean percentage value growth and predict all postal codes would have that growth throughout the years.

We performed descriptive analysis in the resulting data frame to better understanding the behavior of our independent and dependent variables.

Our choices for machine learnings algorithms will be:

* Regression and Lasso regression
* Random Forest
* Generalized Additive Models

After the prediction is made we will use Density Estimation to additionally classify each postal code as:

* Negative growth zone
* Low growth zone
* Medium growth zone
* High growth zone
* **Evaluation and Results (30%)**  
  We are interested in seeing a clear and conclusive set of experiments which successfully evaluate the problem you set out to solve. Make sure to interpret the results and talk about what we can conclude and learn from your approach.
  + How do you evaluate your machine learning solution to the specific question(s) you have addressed?
  + What do these evaluation methods tell you about your solution?

It is not so important how well your method performs but rather, (a) how thorough and careful your evaluation is, and (b) how interesting and clever your results and findings are.

Descriptive analysis

Baseline Model

**References:**

* "Yelp Dataset Challenge." Yelp. N.p., n.d. Web. 05 Apr. 2017.
* "Data." Zillow Research. N.p., n.d. Web. 05 Apr. 2017.
* Kotzias Dimitros, Misha Denil, Nando De Freitas, and Padhraic Smyth. *From Group to Individual Labels Using Deep Features.*
* Maurya Abhinav, Murray Kenton, Lui Yandong, Dyer Chris, Cohen William W., Neill Daniel B. *Semantic Scan: Detecting Subtle, Spatially Localized Events in Text Streams.*
* Ahmari Ramin, Karimi Soraya, Xu Kenneth, Mistele Matt. *Gentrification: Causation and Identification.*
* Royall, Emily Binet. *Towards an Epidemiology of Gentrification: Modeling Urban Change as a Probabilistic Process Using K-means Clustering and Markov Models.* Thesis. Massachusetts Institute of Technology, 2016. N.p.: n.p., n.d. Print.