# Abstract – Analysis of E-commerce Data

DATADIGGER: hcui15, jliu265, yyuan50, yyuan51

## Hypothesis

In this project, we analyze the E-commerce items and hypothesize that there is a connection between being a best seller, which is defined as rank top 50 sellers in the project, and how the item is, which means such as the rating counts, price, average rating, etc. Based on this motivation, we then come up with three hypotheses and three experiments, which are tested by hypothesis testing and techniques of machine learning.

## Data

We collected the data from Amazon. The data is the information of top 100 sellers in each category, and there are 10 categories, such as ‘baby’, ‘electronics’, etc. Therefore, there are 10 times 100 items in total. As for how we clean the data, some of the information have N/A property, so we have to remove these items. For the convenience of analysis, we define the top 50 sellers in each category as best sellers.

## Findings

**Claim #1:** There’s a significant difference between the distribution of ratings for the ‘baby' category compared to the distribution of other categories.
**Support for Claim #1:** We use z-test method to support this claim, since for this testing, alpha = 0.05 and the test stats is about 5.98, which is positive and indicates that the rejection region is on the right, so we have the corresponding value in z-table as 1.960, which is less than 5.98.

**Claim #2:** The length of the product title of the electronic category is irrelavant to being best seller.
**Support for Claim #2:** We use t-test method to support this claim, since the degree of freedom is 90.99 and the test stats is -1.45, which is negative and indicates that the rejection region is on the left, so we have the corresponding value in t-table as -1.987, which is less than -1.45.

**Claim #3:** There is an association between the number of rating count and being a best seller.
**Support for Claim #3:** We use chi-square test method to support this claim, since according to the chi-square table, alpha = 0.05 and the number of the category is 8, which means that $d_f = 7$, we have the critical value is 14.067, and this value is less than the chi-square stats.

**Claim #4:** We get about 2 percent accuracy improvement using K-fold cv parameter searching.
**Support for Claim #4:** To determine the optimum hyper parameters and assess their accuracy, we employ K-fold cross validation. We test different regularization methods and parameters and found out that the best parameter is 0.1 and the $l_1$ regularization has the best performance in our setting.

**Claim #5:** By probing the data, we find that the value range is relatively large and minmax normalization will squeeze the small values to zeros, which may make the prediction much harder.
**Support for Claim #5:** To support this claim, standard normalization and MinMax normalization are two distinct scale approaches that we try out. Both normalization approaches reduce test accuracy when compared to the unnormalized version. While ordinary normalization has little effect, minmax reduces accuracy by 6%.

**Claim #6:** The feature that has the most influence on being a best seller is rating number, while the other features have lower influence.
**Support for Claim #6:** To support the claim, we do K-fold cross validation, once a feature in the dataset is disabled, to discover the optimal hyper parameters under that condition. Then we calculate how the final accuracy will be affected if the model is not trained using a specific feature.