### **Project Problem and Hypothesis**

* What's the project about? What problem are you solving?

Working in the world of television, the main attribute to determine the success of an episode is often a rating or reach. What if we took the innate creative features of a program that are closer to a viewer’s knowledge and more salient. “Which of my favorite **characters** are in this episode?” “I laugh whenever Bart gets in trouble at **Springfield Elementary School**.” Homer’s **lines** are iconic.” We can take these attributes to classify the kinds of episodes that come out of the Simpsons. With the episode clusters, how well are they voted for on IMDB? How are they viewed?

**Overall**: Using a comprehensive dataset of 27 seasons of the Simpsons, what groups of episodes using creative attributes like characters, locations, and lines, can we find and which groups perform the best to inform creative content creators?

* Where does this seem to reside as a machine learning problem? Are you predicting some continuous number, or predicting a binary value?

I want to find clusters of episodes based on their in-show attributes. Using clustering, I can input several dimensions that will then label each episode with a specific group. Next, I can look at the success of each cluster based on IMDB votes.

* What kind of impact do you think it could have?

The Simpsons is a very popular program that’s been watched for years since its release in 1987 and continues to perform strongly today. Some of its suc What attributes are required to replicate the success of the Simpsons for a new animated scripted program on broadcast?

* What do you think will have the most impact in predicting the value you are interested in solving for?

### **Datasets**

* Description of data set available, at the field level. (see table)

|  |  |  |
| --- | --- | --- |
| **Field Name** | **Type** | **Description** |
| Simpsons\_script\_lines.csv | | |
| Speaking\_line | string | Untouched line of a character |
| Normalized\_text | string | Line without special characters |
| simpsons\_characters.csv | | |
| Character | string | Name of a character |
| character\_id | integer | Character unique identifier |
| Simpsons\_locations.csv | | |
| Location | string | Name of a location |
| location\_id | integer | Location unique identifier |
| Simspons\_episodes.csv | | |
| Views | Continuous | Amount of views an episode received |
| Season | Continuous | Episode season |
| Number\_in\_season | Continuous | Episode number |
| Original\_air\_date | Continuous | Air Date |
| Us\_viewers\_in\_millions | Continuous | Number of viewers |
| Imdb\_ratings | Continuous | IMDB score |
| imdb\_votes | Continuous | Number of scores submitted for an episode |

### **Domain knowledge**

* What experience do you already have around this area?

I have knowledge within the linear and digital television ratings space. I am familiar with how to measure the success of content from an entire network down to an episode’s telecast. Since ratings are a function of viewers and views, I can use this information to determine how each cluster performs. However, I am not involved in the creative process, so I do not have knowledge around what attributes of a program (i.e. characters, locations, lines) are desired by its audience or cause more repeat viewers.

* What other research efforts exist?
  + Use a quick Google search to see what approaches others have made, or talk with your colleagues if it is work related about previous attempts at similar problems.
  + This could even just be something like "the marketing team put together a forecast in excel that doesn't do well."
  + Include a benchmark, how other models have performed, even if you are unsure what the metric means.

### **Project Concerns**

It is possible that my model can find that the episodes I am looking at are not too different, or there are no meaningful clusters.

### **Outcomes**

I expect to see clusters each with similar information. I believe the episodes with the most imdb views will likely be seasonal episodes. These episodes may also have fewer characters, since they often focus on the main cast. Since The Treehouse of Horror episodes are more topical and often based on real-world horror stories, more unusual locations may be present. I expect older episodes with a greater amount of characters to have the highest amounts of views. Creating interesting segments of Simpsons episodes that are similar and adequately different from other groups of episodes will be a success. I expect to see differing imdb scores for each cluster. To help add features to my dataset that can better show the success of an episode, I can match Nielsen data with each of the episode numbers to see a better representation of the success of an episode using each of the features.