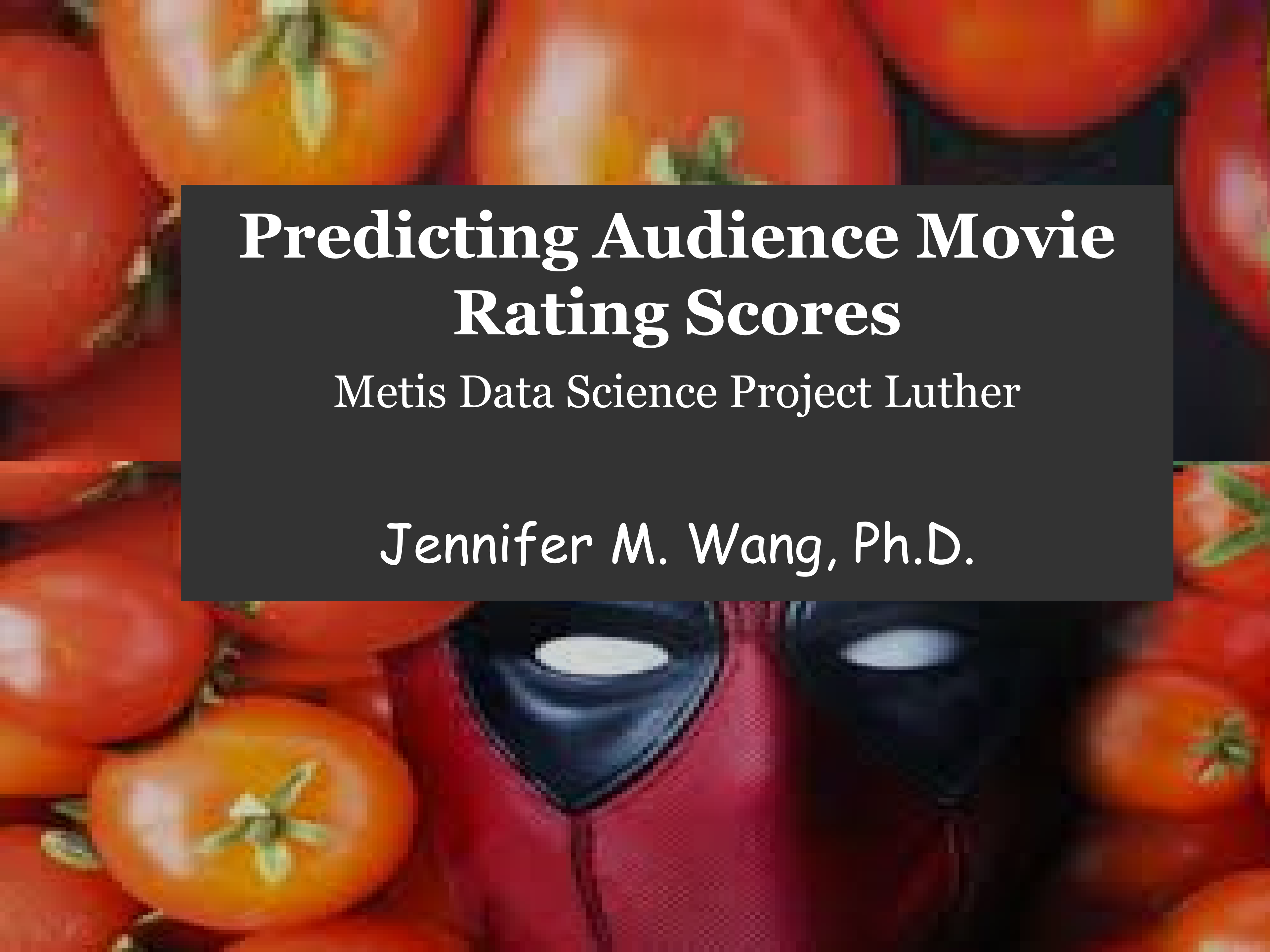


Predicting Audience Movie Rating Scores

Metis Data Science Project Luther

Jennifer M. Wang, Ph.D.

Background Overview

- **Audience rating scores**
  - Audience ratings affect sales
  - Audience ratings serve as important source of information for consumers and movie-goers
  - May be particularly important given the massive amount of movies released today

Overall Project Goal

- **Predict audience's movie rating scores**
  - What variables are associated with audience's moving rating scores?
- Results may prove useful across different industries, including film (directors, writers, producers, investors), advertising, and marketing

Method

- **Sources:**
    - IMDB
    - Rotten Tomatoes
    - Box Office Mojo
- Web-scraping with **BeautifulSoup** in Python
- **n** = **7,117**

Full Model (Initial Model)

- **Predictors/Features:**
  - Movie length time
  - Movie release year
  - Budget (in USD)
  - Gross revenue (in USD)
  - IMDB User Ratings
  - Rotten Tomatoes Critics rating scores
  - Genres (Animation, Action, Comedy, Drama, Crime, Fantasy, Horror, Musical, Biography, Western, Family)
    - Average genre per movie: 3

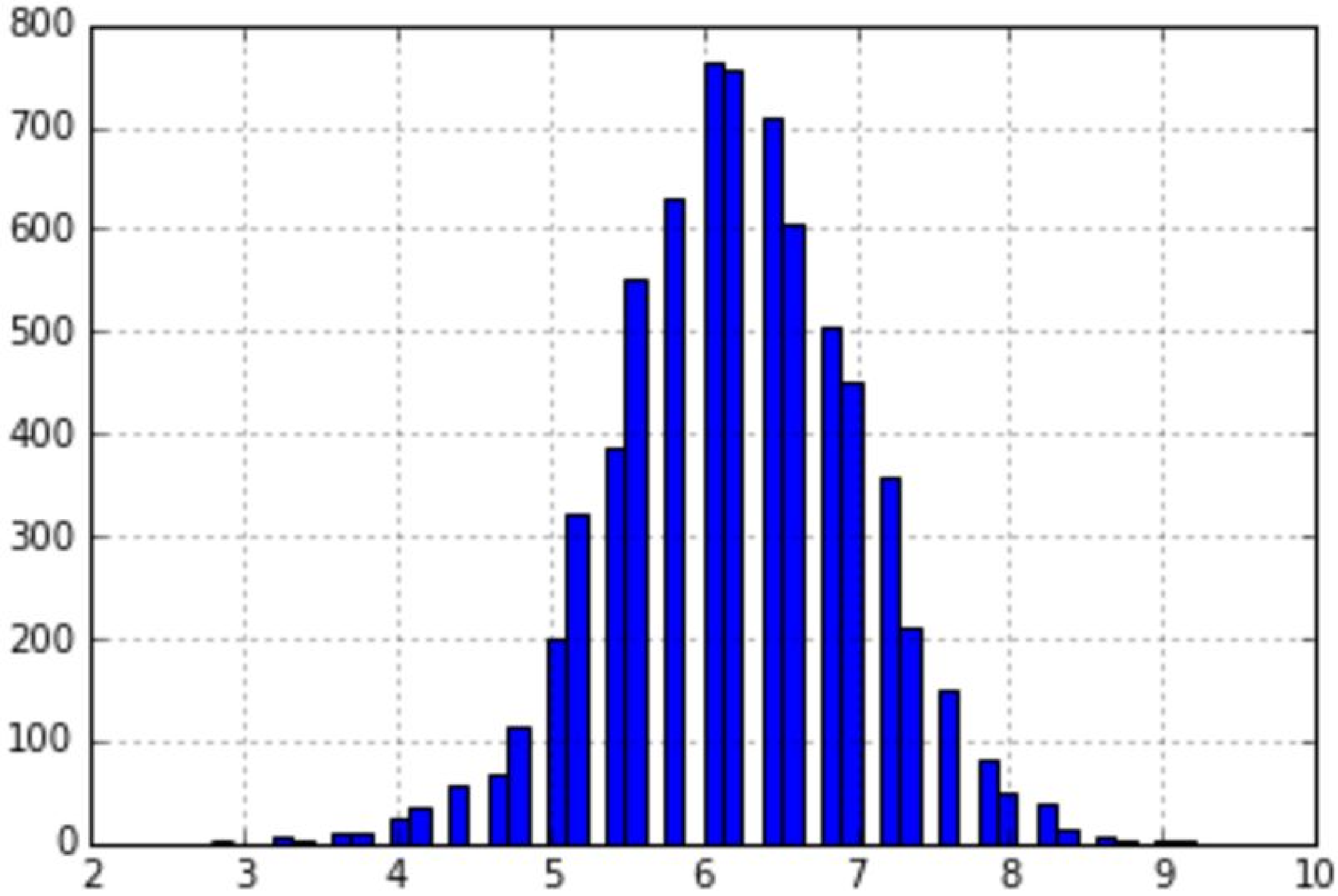
- All non-normally distributed variables were transformed (log, square root) to account for skewness

Target/Dependent Outcome: Audience Rating Scores

- Score values, not %
- May be more reliable than IMDB
- Rotten Tomatoes users may be more relevant and ideal for industries and companies targeting today’s audience

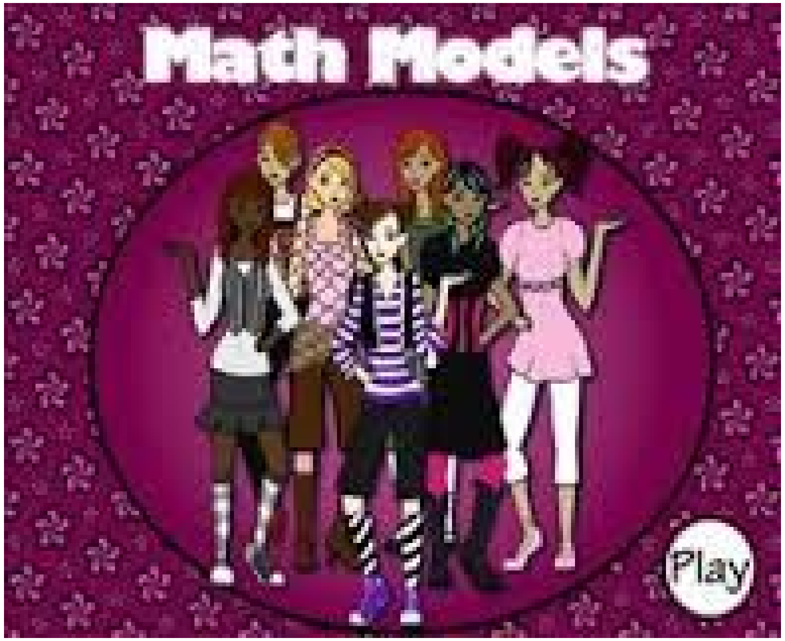
  ↞ **Trust the Wang :)**

Distribution of RT Audience Rating Scores



Models

- Linear Regression
- Ridge
- ElasticNet
- Lasso
- RandomForest

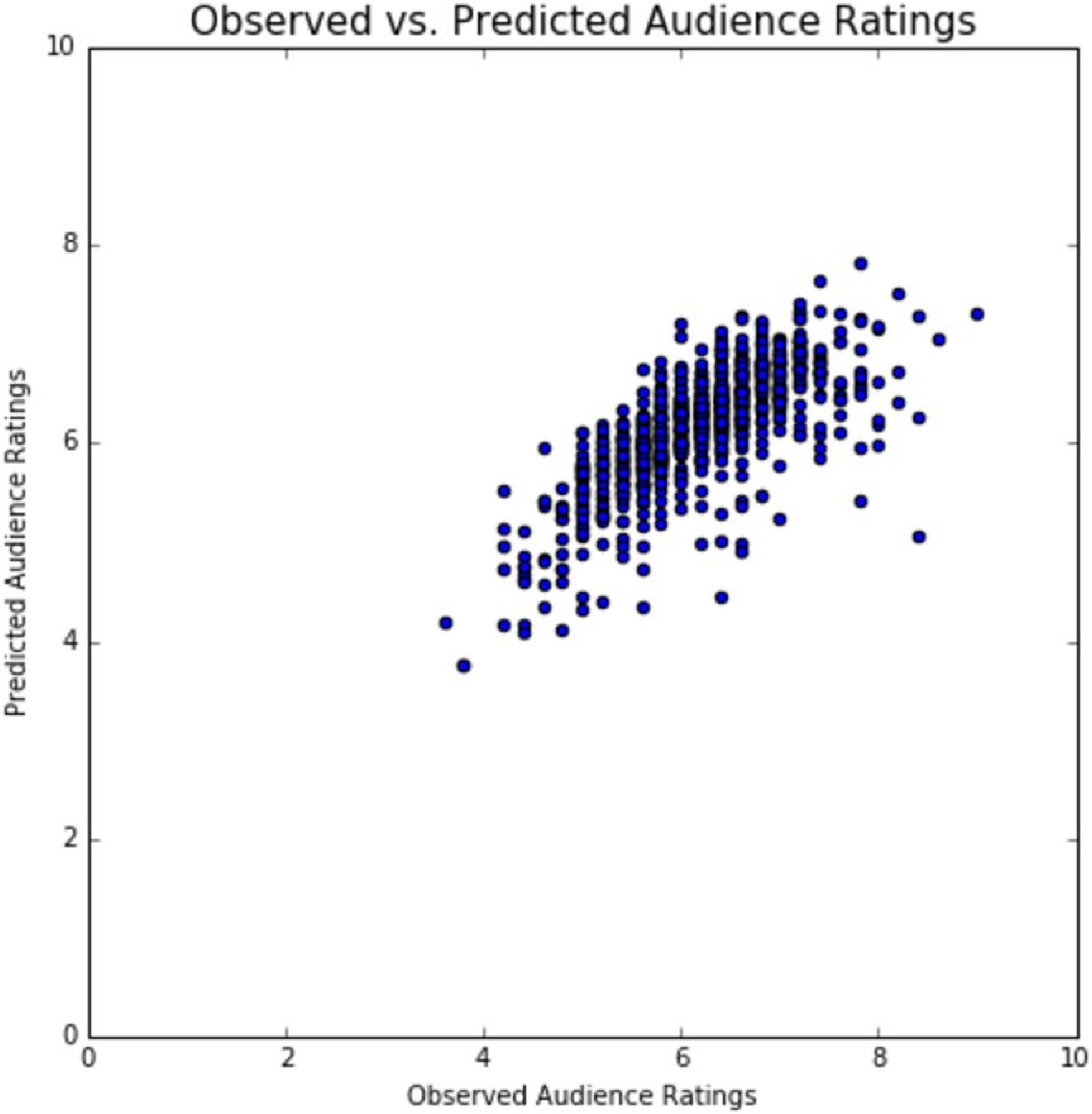


- **Feature engineering:**
  - Retained variables with significant coefficients
  - Large sample size (minimized overfitting)

Results

- **Linear Regression** as the best-fitting model ($R^2 = .59$)
- Predictors/features in final model (all $p < .05$)

+ **Associated with Audience Ratings**	- **Associated with Audience Ratings**
IMDB audience ratings	Critics rating scores
Gross revenue	**Budget**
Year of release	
Length of movie	
Animation, Drama, Musical	Action, Comedy, Crime, Fantasy, **Horror**


Observed vs. Predicted Audience Ratings
Predicted Audience Ratings
Observed Audience Ratings
0
2
4
6
8
10
0
2
4
6
8
10

NOT cheating! :)

- Great differences between I**MDB users** vs. **RT users**

- **Model without IMDB audience ratings:**
  - $R^2 = .39$
  - Coefficients remained significant ($p < .05$)

Implications

- Examine characteristics and behaviors of IMDB users
- Longer movies may be better rated
- Highlight certain genres (Animation, Drama, Musical) and downplay others (Action, Comedy, Crime, Fantasy, **Horror)**
- Understand discrepancy between critics and audience
- Budget might not be everything!

I'm Just Saying...

Future Directions

- Directors
- Actors/actresses
- Production companies/studios
- Awards (e.g., Oscar's)
- Individual characteristics of audience reviewers
- Texts/keywords in reviews! (e.g., sentiment analysis)
- Looking at interacting and moderating variables

"AMAZING!" "DAZZLING!"
"MAJESTIC!" "BRILLIANT!"
"STUNNING!" "CHILLING!"
"AWESOME" "SHOCKING!"
"THRILLING" "STRIKING!"
"BREATHTAKING!" "HOT!"
"WONDERFUL!" "TORRID!"

Thank you! :)

Email me with any questions:
jennifermadisonwang@gmail.com