Unearth the Intrinsic values of Movies

Arvind Krishna, Georgia Institute of Technology, Statistics in ISyE
Haesong Choi, Georgia Institute of Technology, Statistics in ISyE
Namjoon Suh, Georgia Institute of Technology, Statistics in ISyE

Motivation and Goal

- Public/critics movie rate helps people to make a choice on the movie they want to watch

- Ratings of the movie keep changing until a large number of critics have rated it.

- Goal : predict the stabilized rating of a movie on a particular reviewing website before it is released

Data

- “Kaggle” contains more than 5000 movies released from the year 1916 to 2016 in 66 countries

- 30 movies/ 6 movie reviewer groups:
  - 2 Groups: Rotten tomatoes
  - 1 Group: Flixter, Metacritic, MRQE, IMDb

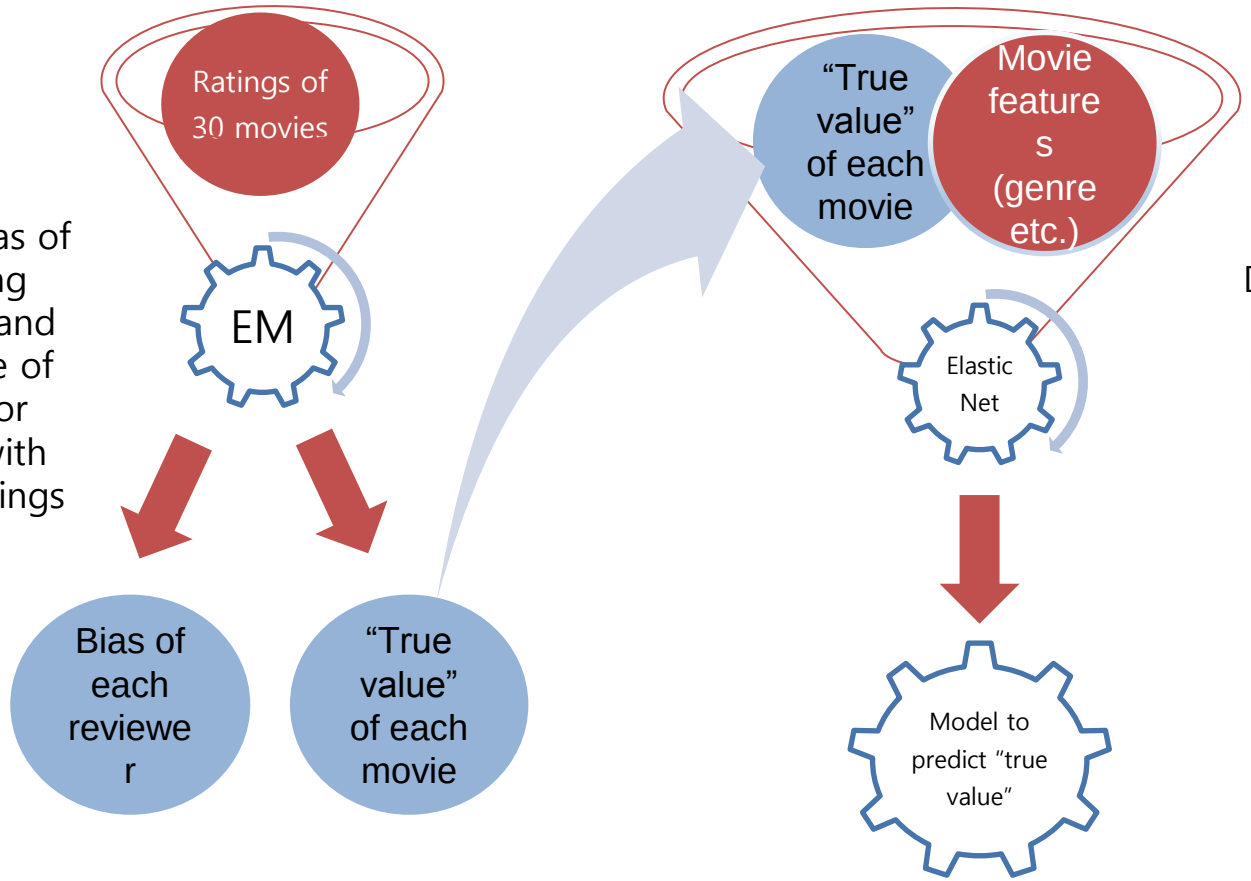
- From IMDB database
  - 22 characteristic variables of movies (shown in next slide)

No.	Data_Name	Data_Description
1	Color	Color movie:1/Black&White movie:0
2	Num_critic	Number of critics who reviewed
3	Duration	Duration of the movie(minutes)
4	Director_Facebook_likes	Number of likes on director's FB page
5	Actor_3_Facebook likes	Number of likes on 3rd actor's FB page
6	Actor_1_Facebook likes	Number of likes on 1st actor's FB page
7	Gross	Gross earning by the movie($)
8	SciFi	1 if the Genre of the movie is SciFi/0 otherwise
9	Drama	1 if the Genre of the movie is Drama/0 otherwise
10	Action	1 if the Genre of the movie is Action/0 otherwise
11	Thriller	1 if the Genre of the movie is Thriller/0 otherwise
12	Num_Voted_Users	Number of users voted on IMDb
13	Cast_Total_Facebook_Likes	Total FB likes for all cast members
14	Face Number_in Poster	Number of the actor who featured in the movie poster
15	Num_User_for_Reviews	Number of users who gave a review
16	USA	1 if the country the movie was produced in is U.S.A /0 otherwise
17	PG_13	1 if there is a need of parental guidance for kids less than 13 years old/ 0 otherwise
18	Budget	Budget of the movie($)
19	Title_Year	Year the movie released in
20	Actor_2_Facebook_Likes	Number of FB likes for actor 2
21	Aspect_Ratio	Aspect ratio the movie was made in
22	Movie_Facebook_Likes	Number of likes on the movie's FB page

Methodology

Bias of movie-rating websites and “true value” of movie

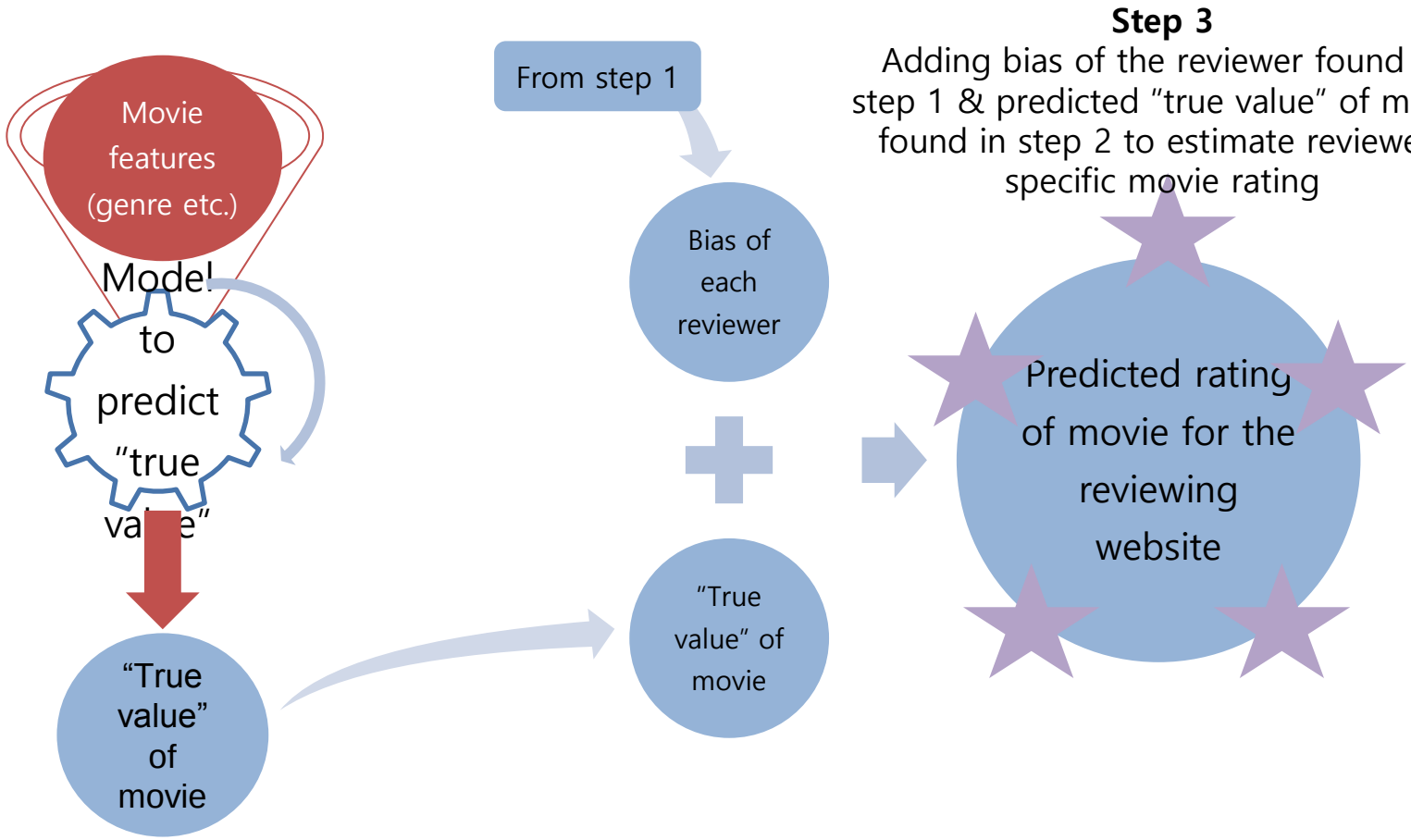
Step 1
Finding bias of reviewing websites and true value of movie for movies with known ratings



Step 2
Developing a model to predict “true value” of movies with unknown ratings

Methodology

Prediction of reviewer-specific movie rating



Step 3

Adding bias of the reviewer found in step 1 & predicted "true value" of movie found in step 2 to estimate reviewer-specific movie rating

EM Algorithm-Finding the True Movie values and the Bias of the Reviewers

- Variables
  - *m:* movies *r:* review scores
  - $x^{(mr)}$: the score that reviewer *r* gave to movie *m*
  - $y^{(mr)}$: "intrinsic" true value($y^{(mr)}$)
  - $Z^{(mr)}$: "bias" of each reviewers($Z^{(mr)}$)

- Assumptions
  - $y^{(mr)} \sim \mathrm{N}\left(\ \mu_m\ , \sigma_m^2\ \right)$
  - $Z^{(mr)} \sim \mathrm{N}\left(\ \mu_r\ , \sigma_r^2\ \right)$
  - $y^{(mr)}$,$Z^{(mr)}$: latent random variables
  - $x^{(mr)} |\ y^{(mr)}, Z^{(mr)}\ \sim\ \mathrm{N}\left(\ y^{(mr)} + \ Z^{(mr)}\ , \sigma^2\right)$
  - $x^{(mr)}$:observed data (σ^2 :2.5),
  - Wilk-Shapiro Test (p-value: 0.5193):

    To check whether $x^{(mr)}$ follows **normal distribution**

Derivation of EM Algorithm

1) E-step

$$\theta = \arg\max_{\theta} \sum_{m=1}^{M} \sum_{r=1}^{R} E_Q[\log P(x^{(mr)}, y^{(mr)}, Z^{(mr)}\ ;\ \theta\ ) \mid x^{(mr)}, \theta^{(old)}]$$

2) M-step

Setting derivatives w. r. t parameters μ_m , σ_m^2 , μ_r , σ_r^2 to 0

$$-\frac{1}{2\sigma_m^2}\sum_{r=1}^{R}(2\,\mu_m - 2\mu_{mr,Y}) = 0 \ \ ===>\ \ \mu_m = \frac{1}{R}\sum_{r=1}^{R}\mu_{mr,Y}$$

$$-\frac{1}{2\sigma_r^2}\sum_{m=1}^{M}(2\,\mu_r - 2\mu_{mr,Z}) = 0 \ \ ===>\ \ \mu_r = \frac{1}{M}\sum_{m=1}^{M}\mu_{mr,Z}$$

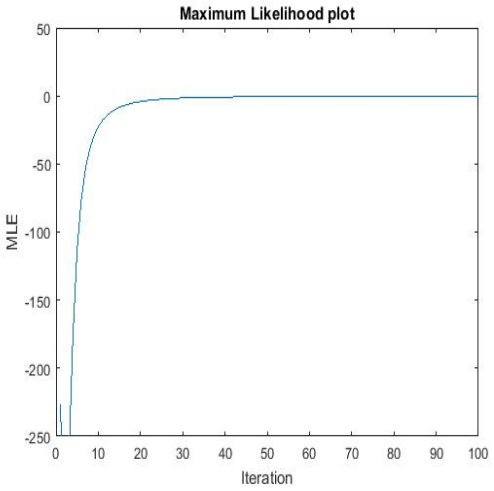
$$\sum_{r=1}^{R}[-\frac{1}{\sigma_p} - \frac{1}{\sigma_m^3}(\Sigma_{mr,YY} + \mu^2{}_{mr,Y} - 2\mu_{mr,\ Y}\mu_m + \mu_m^2)] = 0 \ \ ===>\ \ \sigma_m{}^2 = \sum_{r=1}^{R}(\Sigma_{mr,YY} + \mu^2{}_{mr,Y} - 2\mu_{mr,Y}\mu_m + \mu_m^2)$$

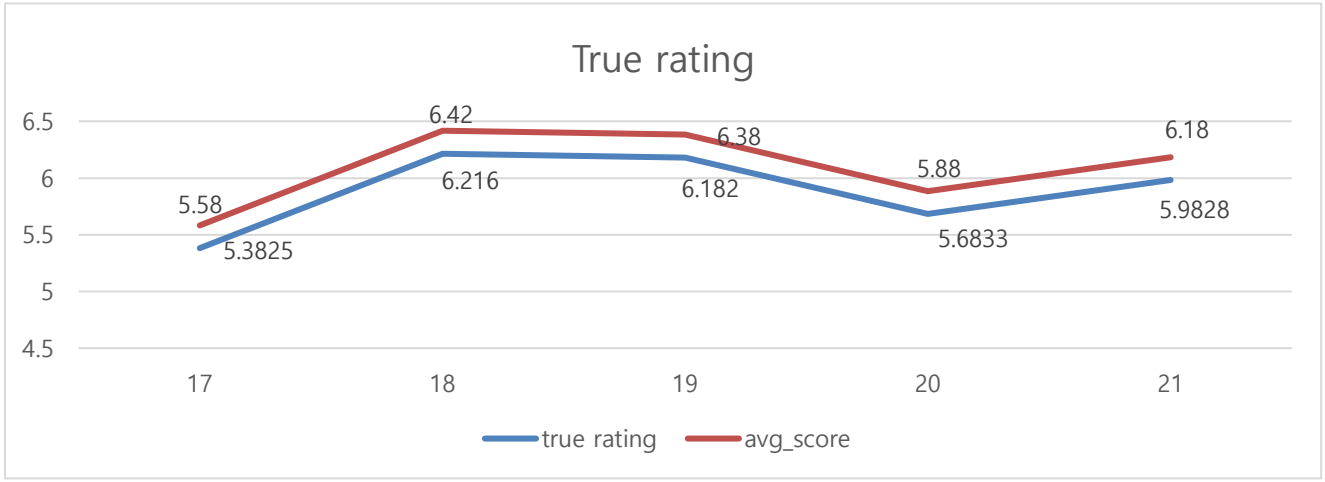
$$\sum_{m=1}^{M}[-\frac{1}{\sigma_r} - \frac{1}{\sigma_r^3}(\Sigma_{mr,ZZ} + \mu^2{}_{mr,Z} - 2\mu_{mr,\ Z}\mu_r + \mu_r^2)] = 0 \ \ ===>\ \ \sigma_r{}^2 = \sum_{m=1}^{M}(\Sigma_{mr,ZZ} + \mu^2{}_{mr,Z} - 2\mu_{mr,Z}\mu_r + \mu_r^2)$$

Result of EM Algorithm

- Whether the expected log-likelihood function converges over iterations.
- After 72th iteration, ML converges
- Gap between true value and averaged score : 0.2= the averaged bias

(∵By LLN, $\mathrm{E}(x^{(mr)}) = \mathrm{E}(y^{(mr)}) + \mathrm{E}(z^{(mr)})$)

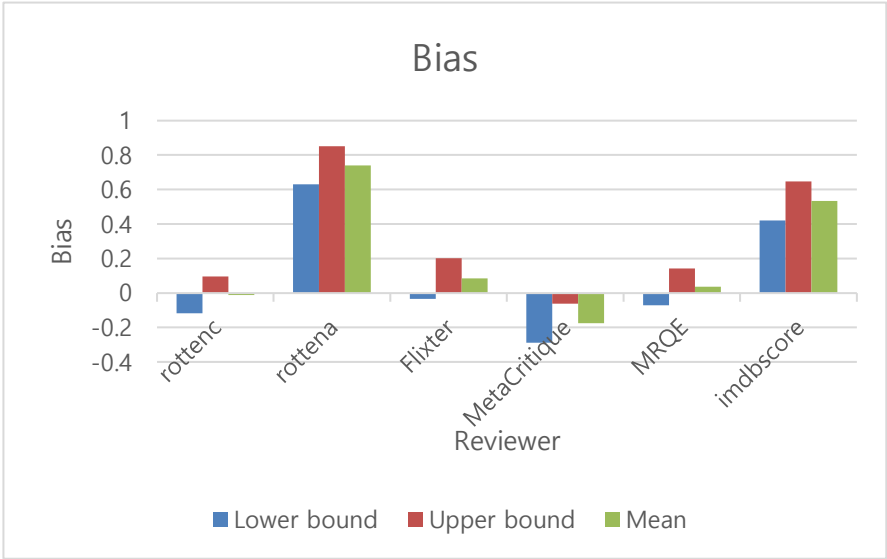




Result of EM Algorithm

- Positive Bias : the ordinary audiences in Rotten tomato, Flixter, Meta-Critics, MRQE, and IMDB
- Negative Bias: critiques from Rotten Tomato and Meta Critics

	RT_C	RT_A	Flixter	Meta Critics	MRQE	IMDb
Bias-mean	-0.0117	0.7403	0.0842	-0.1755	0.0354	0.5338
Bias_High	-0.1186	0.6299	-0.0333	-0.2887	-0.0706	0.4207
Bias_Low	0.0952	0.8508	0.2017	-0.0624	0.1414	0.6469



Elastic-net model- Predicting the Unreleased Movies' True value

$$\hat{\beta} = \underset{\beta}{argmin} \frac{1}{2N} \|y - X\beta\|_2^2 + \lambda/2\|\beta\|_1 + (1-\lambda)/2\|\beta\|_2^2$$

- 25 data : Training dataset/ 10 fold Cross Validation
- Lambda : 0.02327

Figure 7 - Path of the Elastic-Net Regression Coefficients

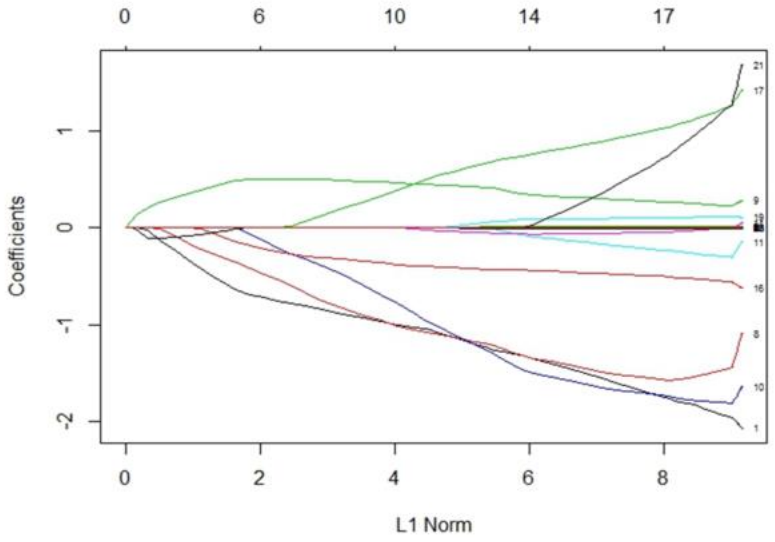
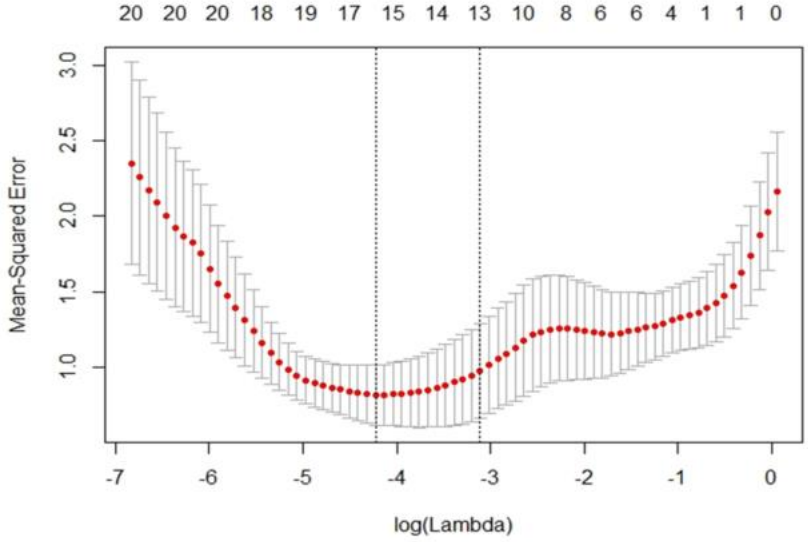


Figure 8 - Cross-Validation Error Curve



Analysis of Elastic-net Model

- The top 5 highest absolute coefficient values

: Action(-1.66), Color(-1.55), Sci-Fi(-1.49), PG_13(0.89), USA(-0.47)

Fithteen Coefficients Selected using Elastic Net Regression

No.	Data_Name	Coefficients
10	Action	-1.66
1	Color	-1.55
8	Sci_Fi	-1.49
17	PG_13	0.89
16	USA	-0.47
21	Aspect_ratio	0.35
9	Drama	0.30
11	Thriller	-0.17
19	Title_tear	0.098
14	Face_number_in_poster	-0.057
3	Duration	0.0157
2	Num_critic_for_reviews	0.00976
4	actor_2_facebook_likes	-3.99E-05
20	actor_1_facebook_likes	-5.21E-09
18	budget	-4.27E-06

Prediction of the Movie Rating

with re

- (Tı
- Bia
- Te
- The

Movie	Rotten tomatoes (A)		Rotten tomatoes (C)		Flixter	
	Actual rating	Our estimation	Actual rating	Our estimation	Actual rating	Our estimation
Whiplash	8.6	8.2	9	8.9	9.4	8.3
Mission: Impossible - Rogue Nation	7.5	7.5	8.2	8.3	8.7	7.6
The Finest Hours	6.1	5.9	7.2	6.7	6.6	6.0
Magic Mike	5.9	5.6	6.8	6.4	5.6	5.7
Oculus	6.5	6.0	6.4	6.7	5.3	6.0

Movie	Meta Critic		MRQE		IMDB	
	Actual rating	Our estimation	Actual rating	Our estimation	Actual rating	Our estimation
Whiplash	8.8	8.0	8.5	8.2	8.5	8.7
Mission: Impossible - Rogue Nation	7.5	7.4	7.4	7.6	7.4	8.0
The Finest Hours	5.8	5.7	6.8	6.0	6.8	6.4
Magic Mike	6.0	5.5	5.7	5.7	5.7	6.1
Oculus	6.1	5.8	6.5	6.0	6.5	6.5

Conclusion and Discussion

- The audience : positively biased/ higher rating than critics do.
  (the high positive bias if IMDB and Rotten Tomatoes (Audience score)
  negative bias of Rotten Tomatoes (Critic score) and Metacritic)

- The "true rating" of the movie is proportional to the average rating of all the reviewing websites (consistent with our modeling assumption)

- We estimate the true value of a movie from the elastic net model.
  (the values of the covariates in the elastic net model + the bias of each reviewing website = the rating of any movie on each of the nation's top 5 reviewing website even before the movie is released)


Q&A