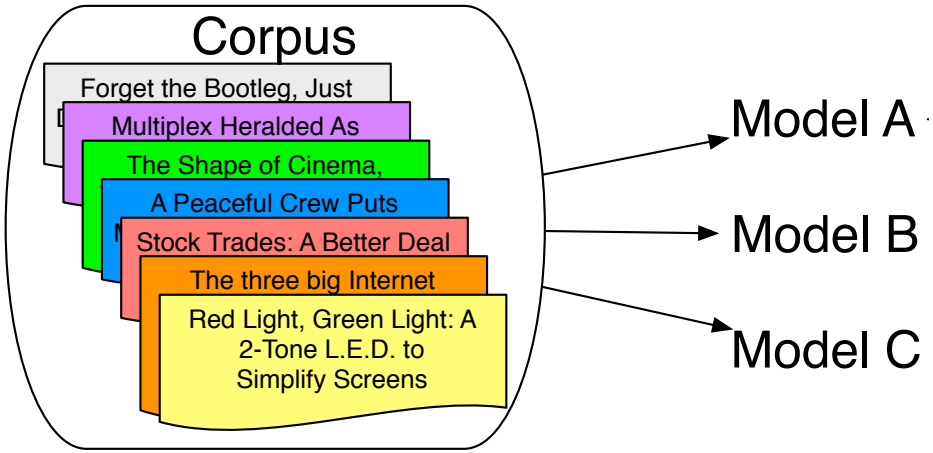
Topic Models

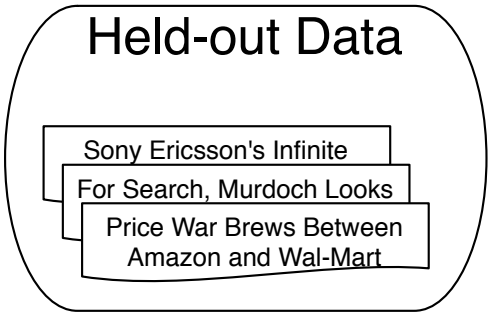
Advanced Machine Learning for NLP

Jordan Boyd-Graber

EVALUATION

Evaluation

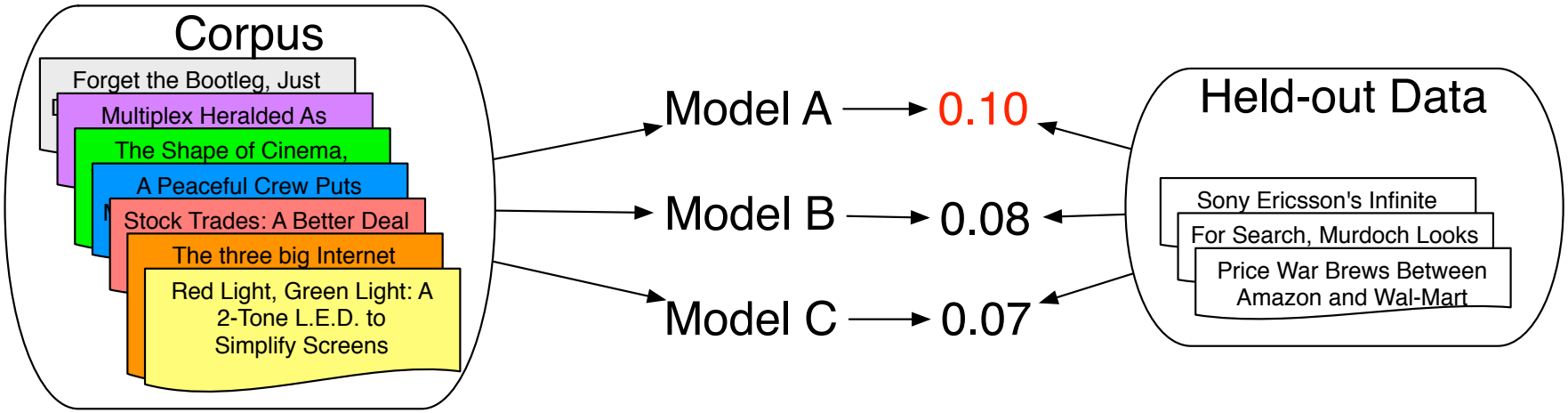




$$P(\boldsymbol{w} \mid \boldsymbol{w}', \boldsymbol{z}', \alpha\boldsymbol{m}, \beta\boldsymbol{u}) =$$
$$\sum_{\boldsymbol{z}} P(\boldsymbol{w}, \boldsymbol{z} \mid \boldsymbol{w}', \boldsymbol{z}', \alpha\boldsymbol{m}, \beta\boldsymbol{u})$$

How you compute it is important too

Evaluation



Measures predictive power, not what the topics are

$$P(\boldsymbol{w} \mid \boldsymbol{w}', \boldsymbol{z}', \alpha\boldsymbol{m}, \beta\boldsymbol{u}) = \sum_{\boldsymbol{z}} P(\boldsymbol{w}, \boldsymbol{z} \mid \boldsymbol{w}', \boldsymbol{z}', \alpha\boldsymbol{m}, \beta\boldsymbol{u})$$

How you compute it is important too

Word Intrusion

TOPIC 1	TOPIC 2	TOPIC 3
computer, technology, system, service, site, phone, internet, machine	sell, sale, store, product, business, advertising, market, consumer	play, film, movie, theater, production, star, director, stage

Word Intrusion

1. Take the highest probability words from a topic

Original Topic

dog, cat, horse, pig, cow

Word Intrusion

1. Take the highest probability words from a topic

Original Topic

dog, cat, horse, pig, cow

2. Take a high-probability word from another topic and add it

Topic with Intruder

dog, cat, apple, horse, pig, cow

Word Intrusion

1. Take the highest probability words from a topic

Original Topic

dog, cat, horse, pig, cow

2. Take a high-probability word from another topic and add it

Topic with Intruder

dog, cat, apple, horse, pig, cow

3. We ask users to find the word that doesn't belong

Hypothesis

If the topics are interpretable, users will consistently choose true intruder

Word Intrusion

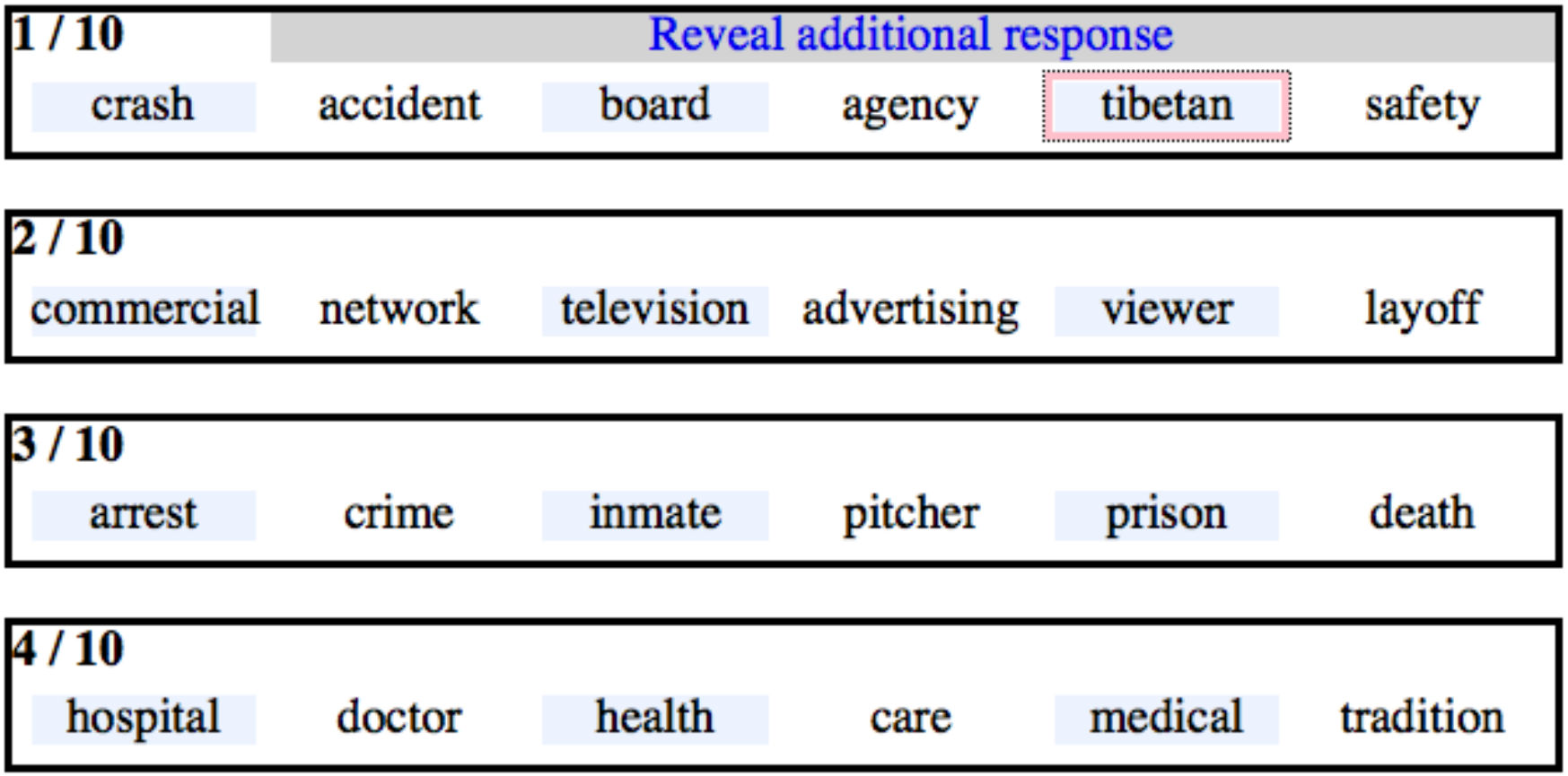
1 / 10					
crash	accident	board	agency	tibetan	safety

2 / 10					
commercial	network	television	advertising	viewer	layoff

3 / 10					
arrest	crime	inmate	pitcher	prison	death

4 / 10					
hospital	doctor	health	care	medical	tradition

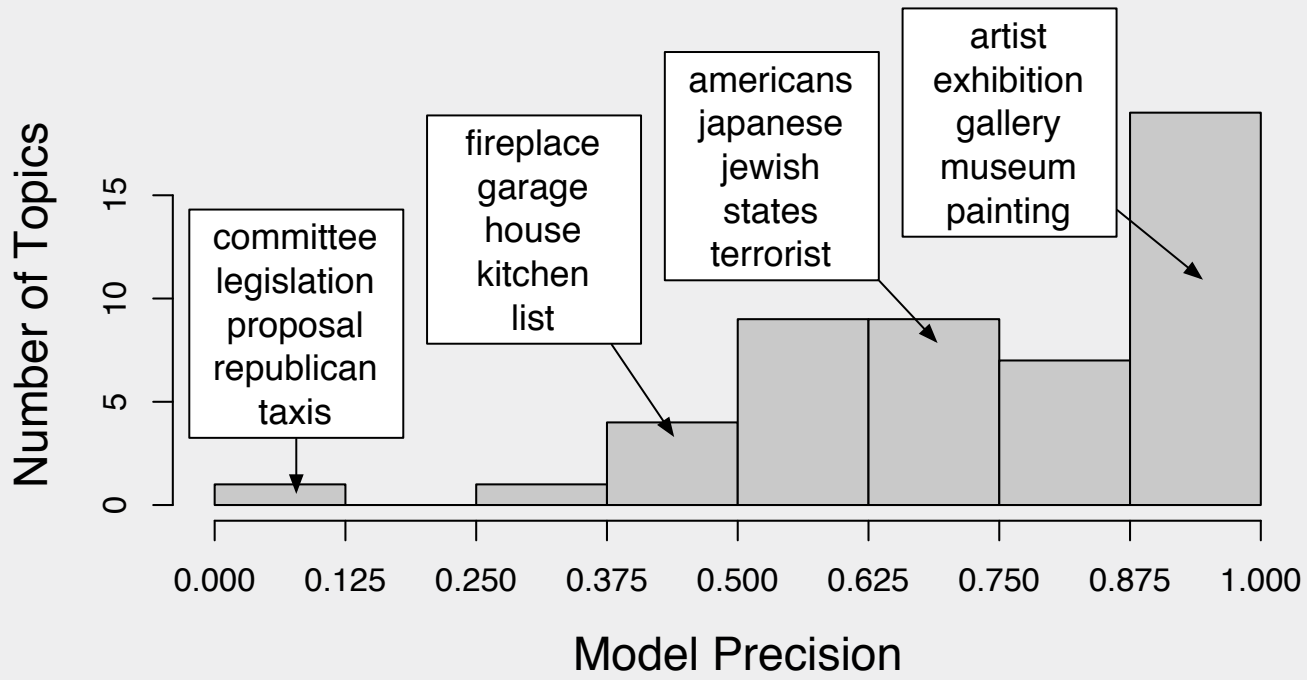
Word Intrusion



- Order of words was shuffled
- Which intruder was selected varied
- Model precision: percentage of users who clicked on intruder

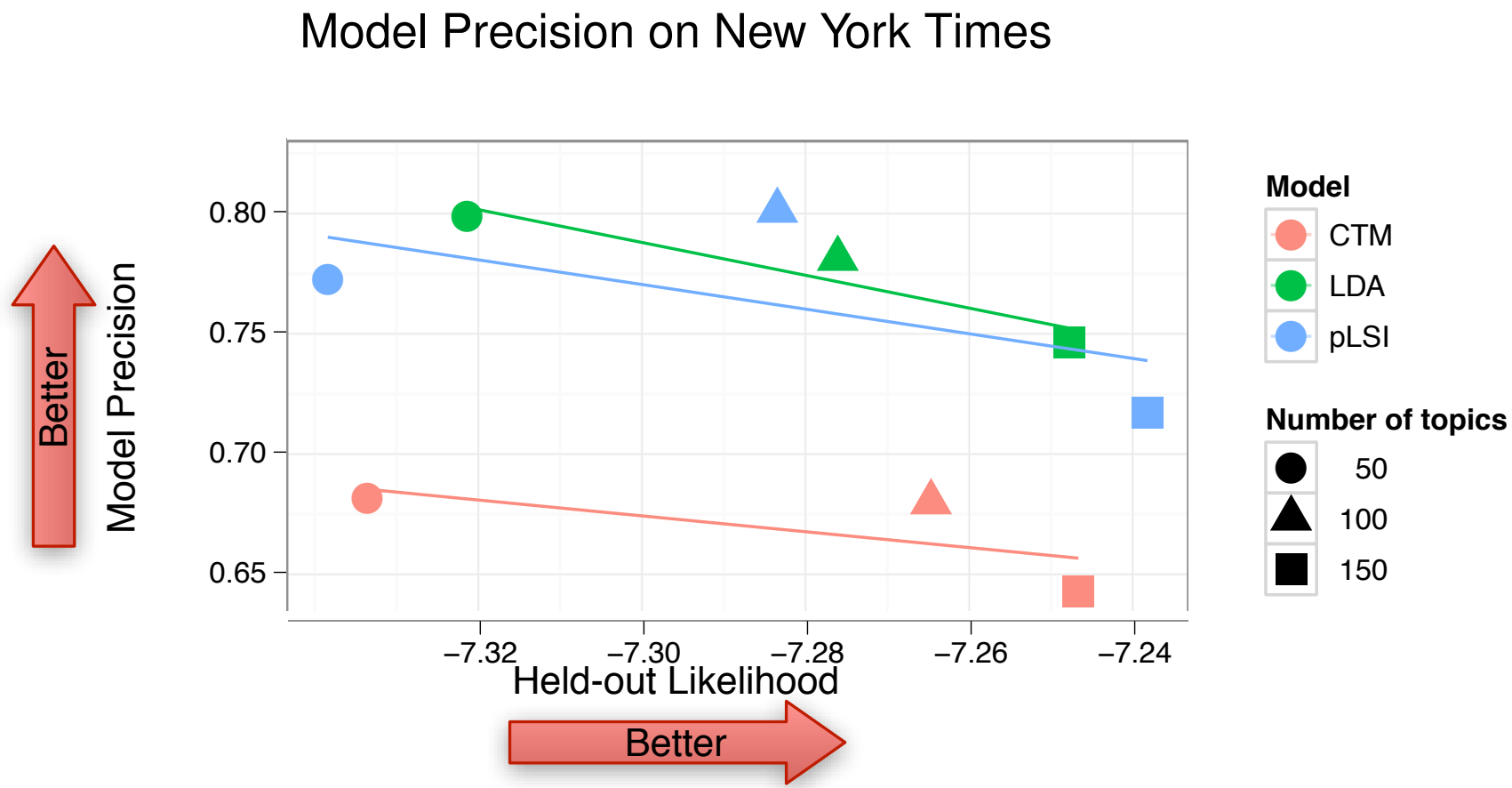
Word Intrusion: Which Topics are Interpretable?

New York Times, 50 LDA Topics



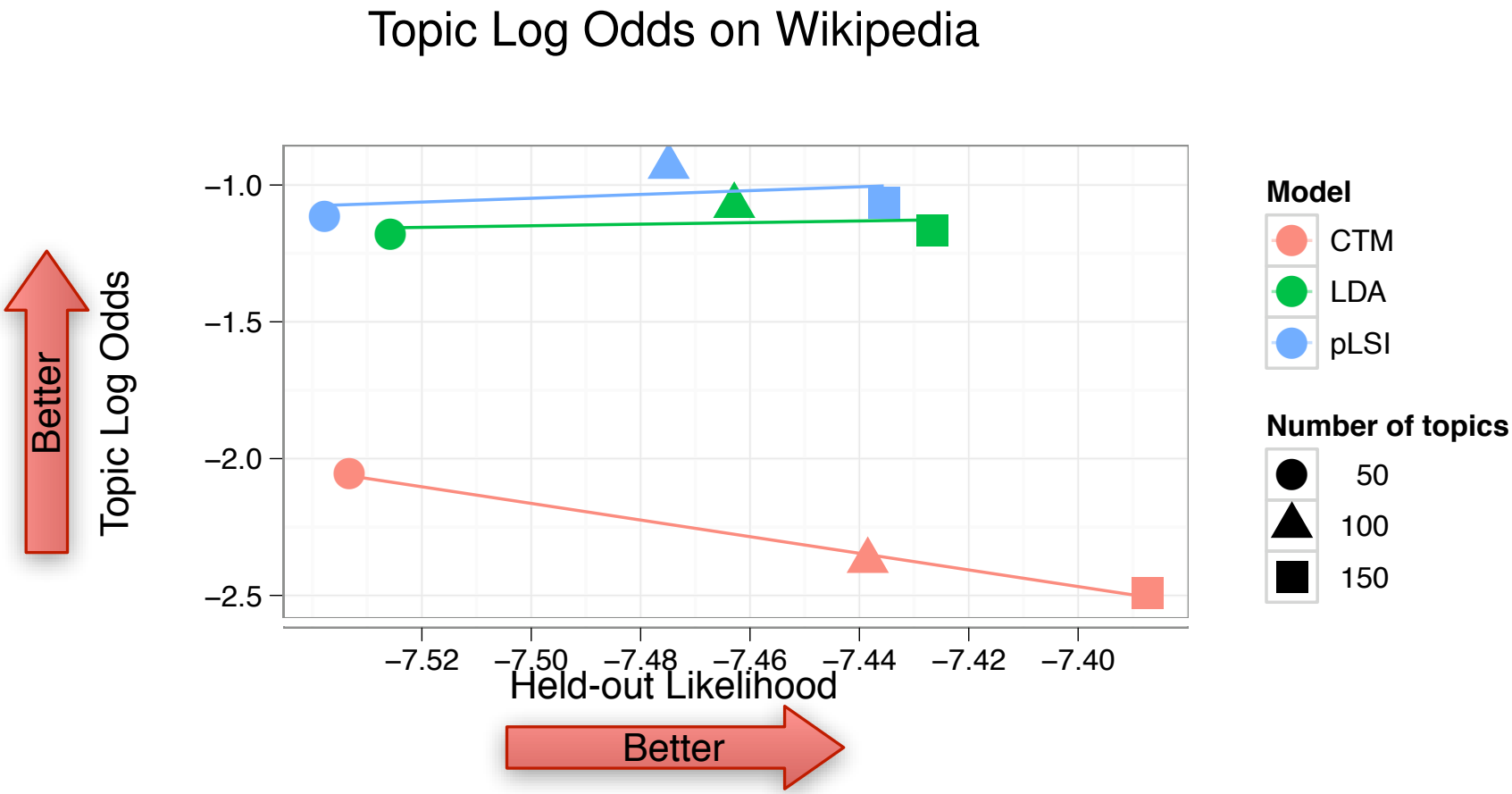
Model Precision: percentage of correct intruders found

Interpretability and Likelihood



within a model, higher likelihood ≠ higher interpretability

Interpretability and Likelihood



across models, higher likelihood ≠ higher interpretability

Downstream Tasks

- Classification
- Machine Translation
- Political Polarization/Framing

Downstream Tasks

- Classification
- Machine Translation
- Political Polarization/Framing