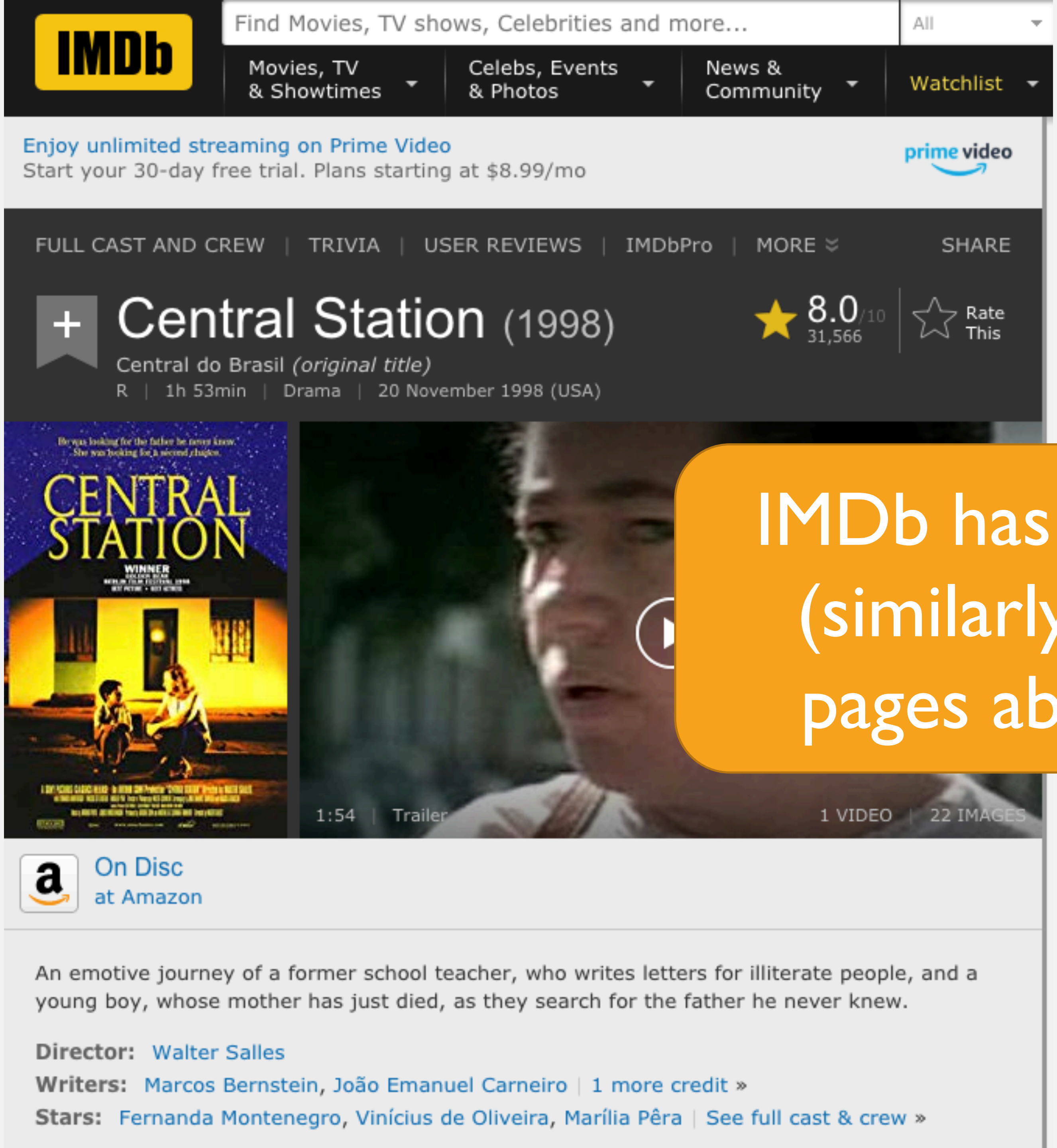
CERES: DISTANTLY SUPERVISED RELATION EXTRACTION FROM THE SEMI-STRUCTURED WEB

Colin Lockard, Xin Luna Dong, Arash Einolghozati, Prashant Shiralkar



IMDb has thousands of (similarly formatted) pages about movies.

NMDb

MOVIES TV SHOWS ACTORS CREW EVENTS

Nigerian films

Twisted (Short film)

D" Trailer

Not on IMDb!

Year of production: 2014

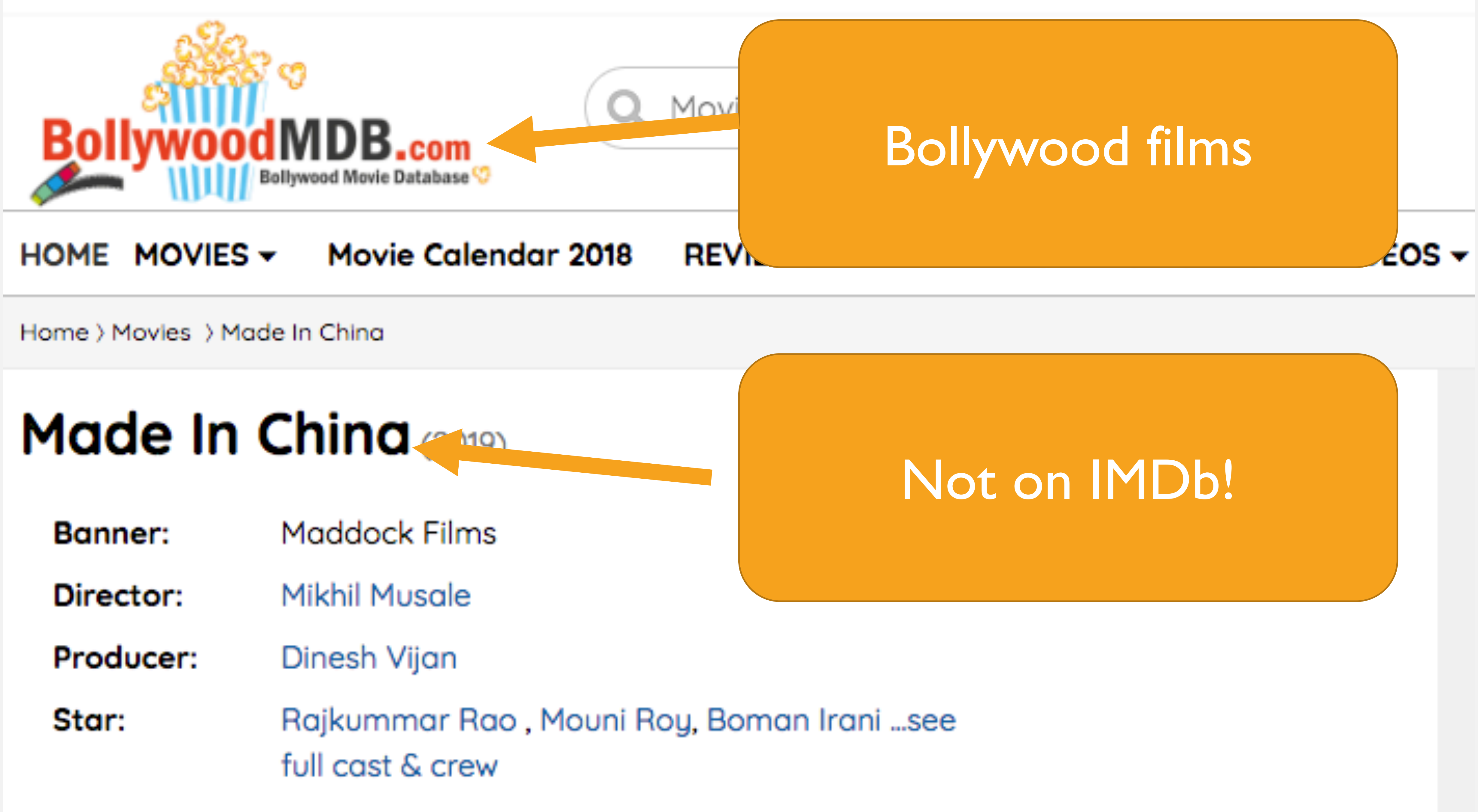
Running Time: 2:12 mins

Written by: Daniel Ademinokan

Produced by: Daniel Ademinokan

Directed by: Daniel Ademinokan

Starring: Stella Damasus Rob Byrnes, Matt Meinsen and David Ademinokan


BollywoodMDB.com
Bollywood Movie Database
HOME
MOVIES
Movie Calendar 2018
Home › Movies › Made In China
Made In China
Banner: Maddock Films
Director: Mikhil Musale
Producer: Dinesh Vijan
Star: Rajkummar Rao , Mouni Roy, Boman Irani ...see full cast & crew
Bollywood films
Not on IMDb!

ONF NFB

SEARCH

FILMS

DOCUMENTARY

ANIMATION

INTERACTIVE

EDUCATION

SIGN IN

Abegweit

Serge Morin

1998 | 1 h 11 min

AVAILABLE

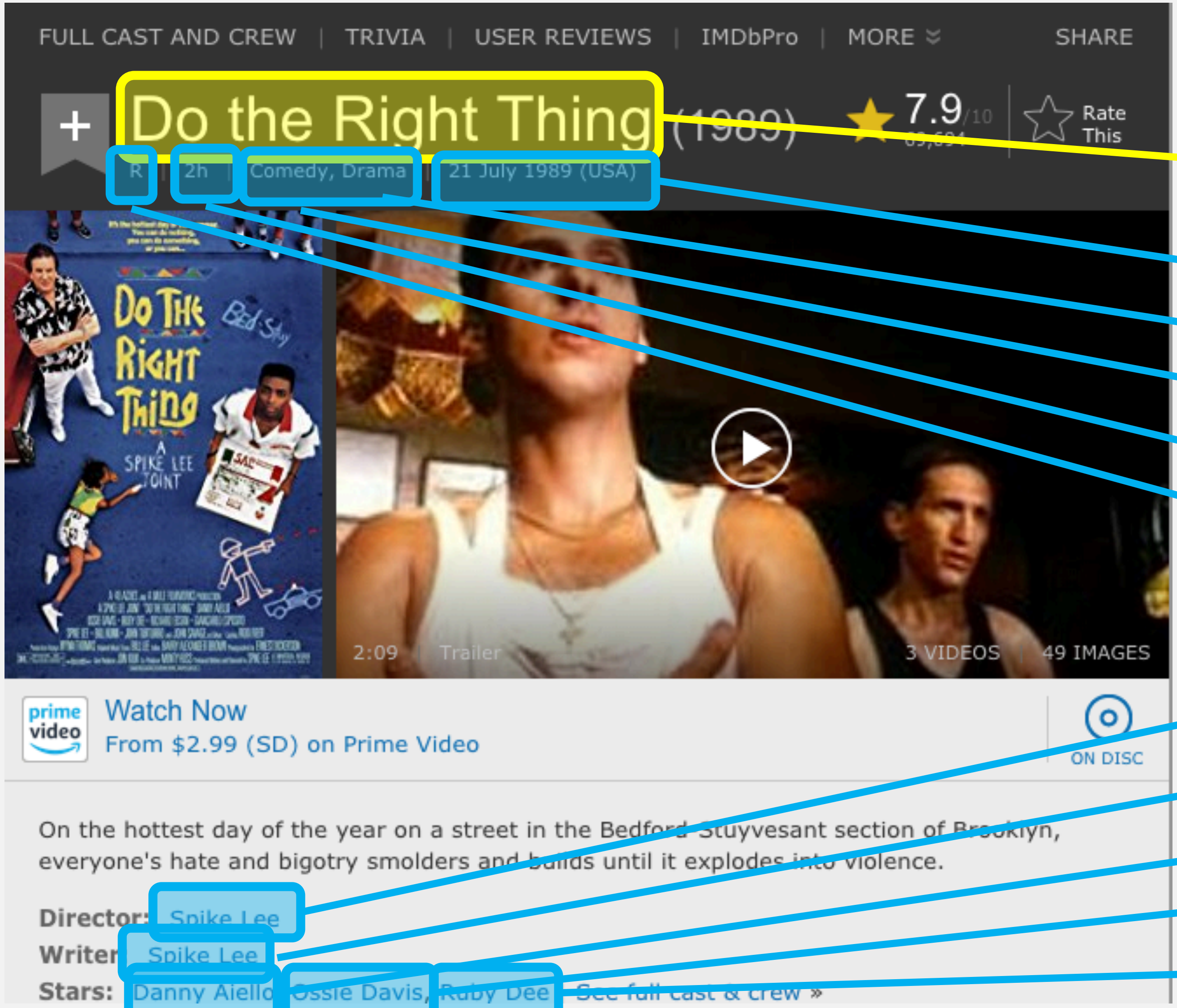
Canadian films

Not on IMDb!

▼ CREDITS

DIRECTOR Serge Morin	**SCRIPT** Serge Morin	**PRODUCER** Pierre Bernier Diane Poitras	**CAMERA** Marc Paulin
SOUND Georges Hannan	**EDITING** Fernand Bélanger	**RE-RECORDING** Serge Boivin Jean Paul Vialard	**SOUND EDITING** Fernand Bélanger Claude Langlois
NARRATOR Alex Madsen	**MUSIC** Richard Gibson		**PARTICIPATION** Francine Blais Peter Briden Ralph Murray Guy Cormier Jim Feltham Kim Gallant

SEMI-STRUCTURED WEB EXTRACTION



TOPIC ENTITY

"Do the Right Thing"

"Do the Right Thing", film.release, "21 July 1989"

"Do the Right Thing", film.genre, "Comedy"

"Do the Right Thing", film.genre, "Drama"

"Do the Right Thing", film.runtime, "2h"

"Do the Right Thing", film.rating, "R"

"Do the Right Thing", film.director, "Spike Lee"

"Do the Right Thing", film.writer, "Spike Lee"

"Do the Right Thing", film.actor, "Danny Aiello"

"Do the Right Thing", film.actor, "Ossie Davis"

"Do the Right Thing", film.actor, "Ruby Dee"

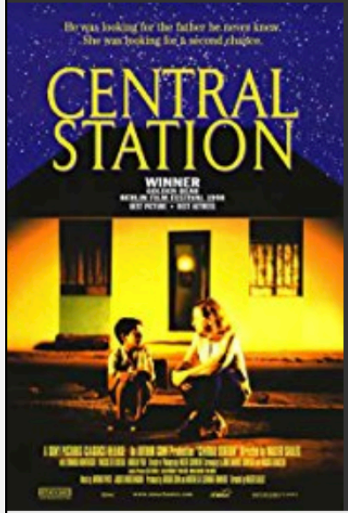
FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE SHARE

Central Station (1998)

8.0/10 31,520 Rate This

Central do Brasil *(original title)*

R | 1h 53min | Drama | 20 November 1998 (USA)



On Disc
at Amazon

An emotive journey of a former school teacher, who writes letters for illiterate people, and a young boy, whose mother has just died, as they search for the father he never knew.

Director: Walter Salles

Writers: Marcos Bernstein, João Emanuel Carneiro | 1 more credit »

Stars: Fernanda Montenegro, Vinícius de Oliveira, Marília Pêra | See full cast & crew »

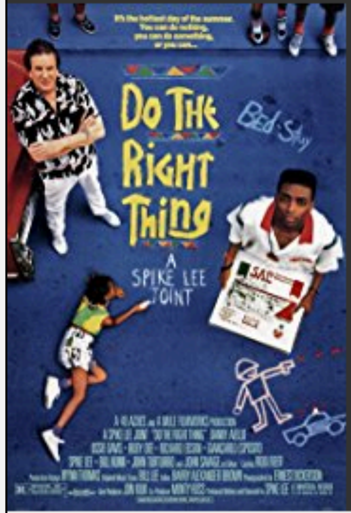
80 Metascore From metacritic.com | Reviews 261 user | 73 critic

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE SHARE

Do the Right Thing (1989)

7.9/10 69,694 Rate This

R | 2h | Comedy, Drama | 21 July 1989 (USA)



prime video Watch Now
From $2.99 (SD) on Prime Video
ON DISC

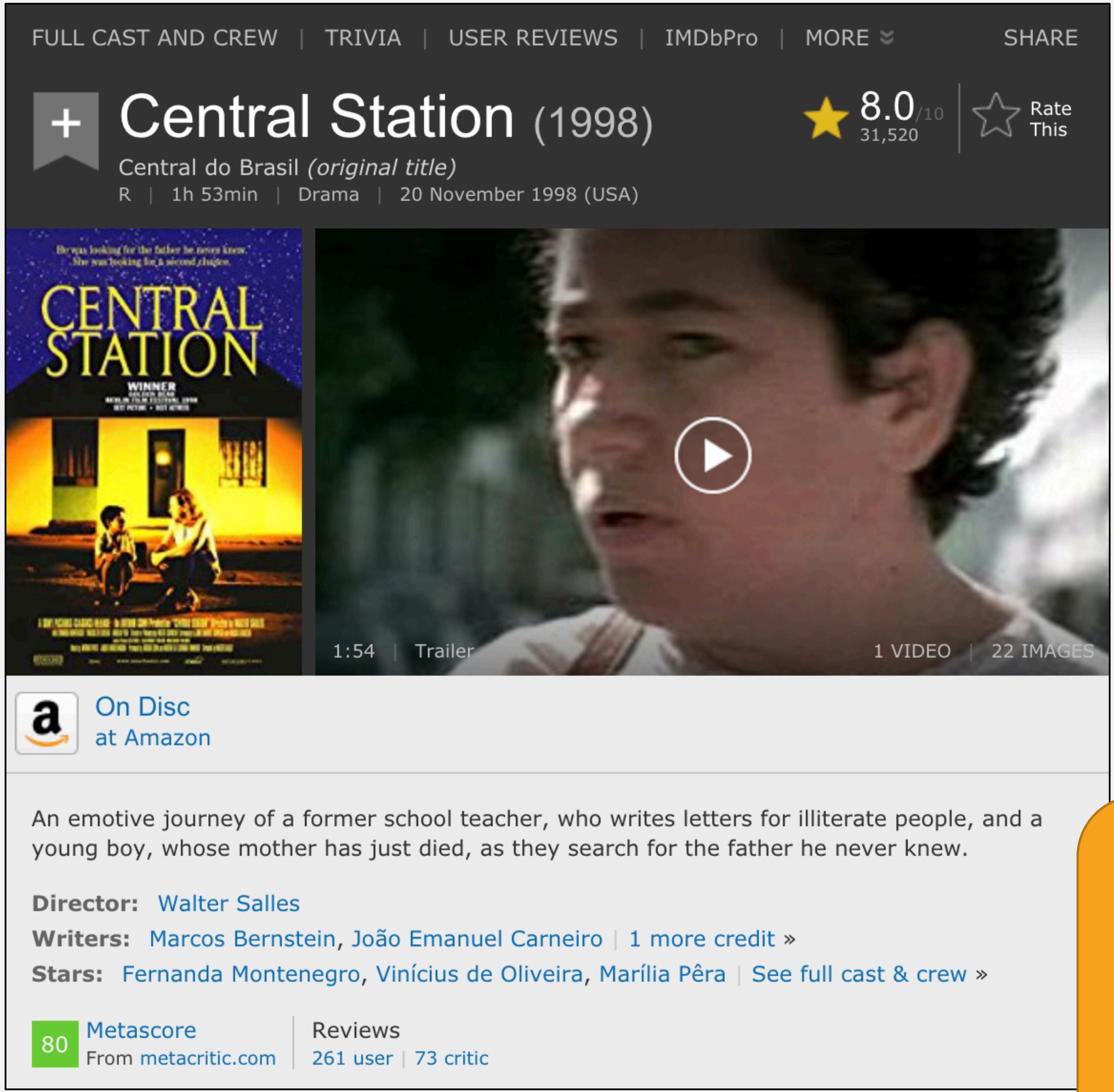
On the hottest day of the year on a street in the Bedford-Stuyvesant section of Brooklyn, everyone's hate and bigotry smolders and builds until it explodes into violence.

Director: Spike Lee

Writer: Spike Lee

Stars: Danny Aiello, Ossie Davis, Ruby Dee | See full cast & crew »

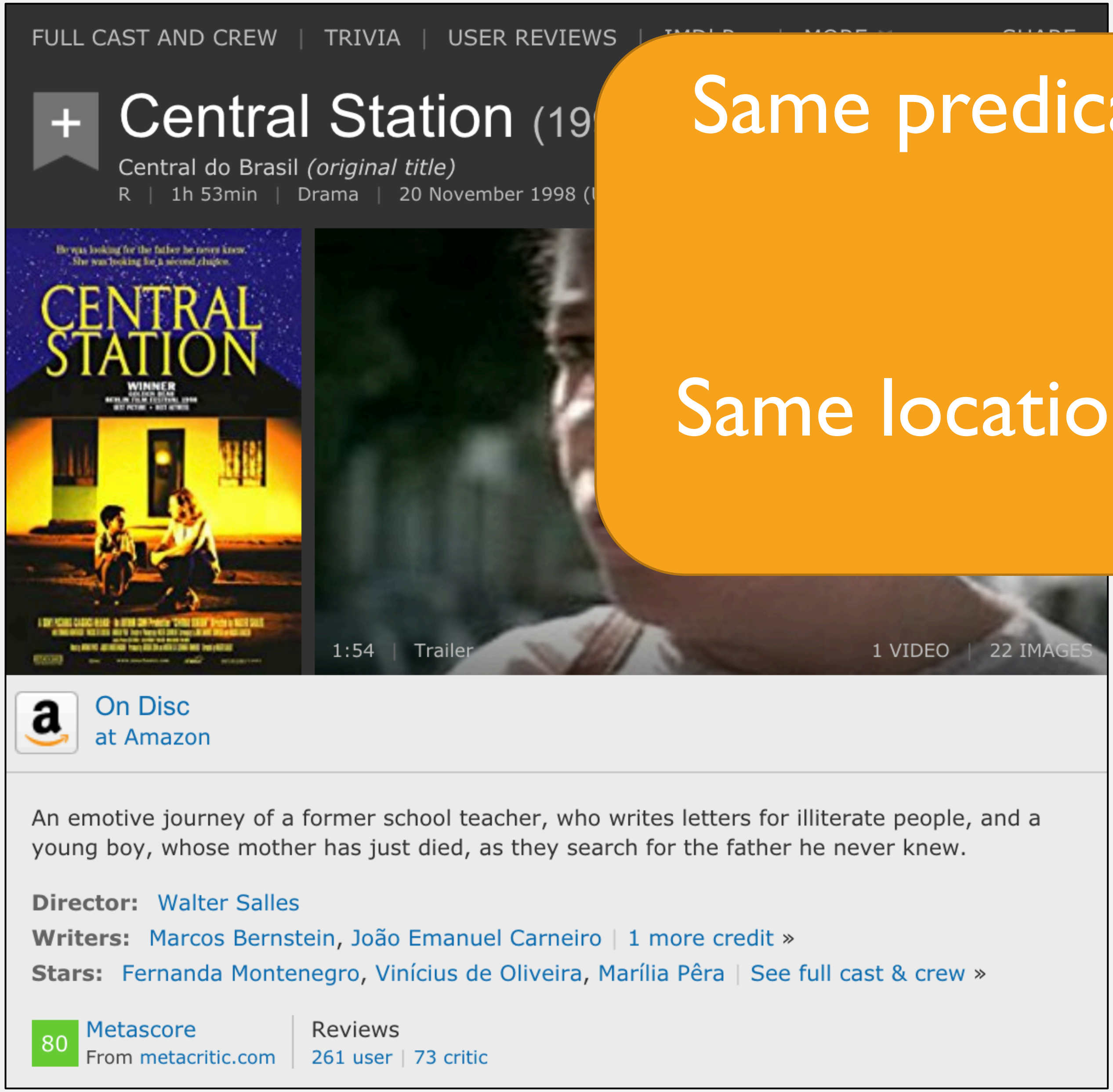
91 Metascore From metacritic.com | Reviews 436 user | 110 critic | Popularity 2,391 (↑ 446)

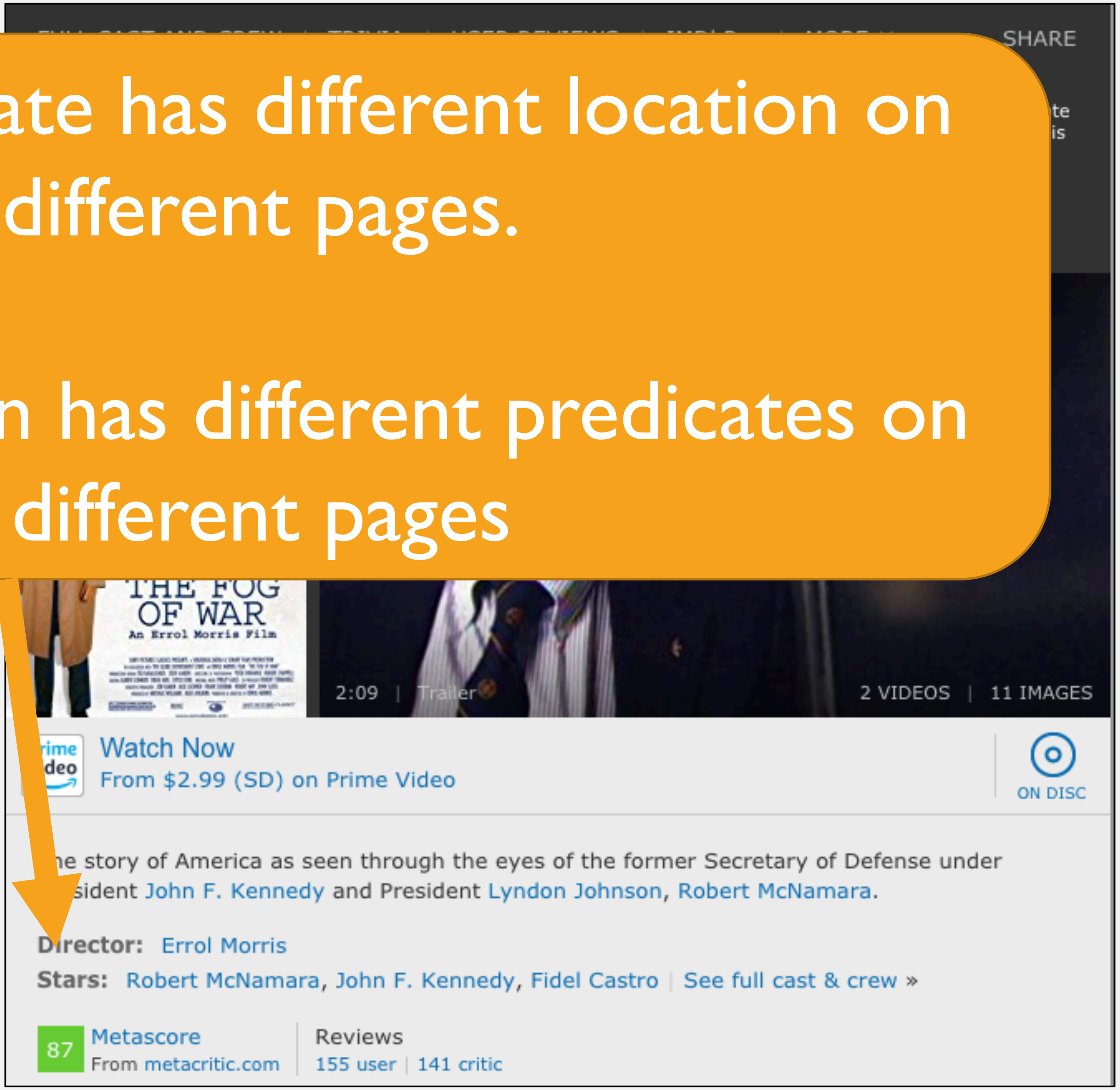


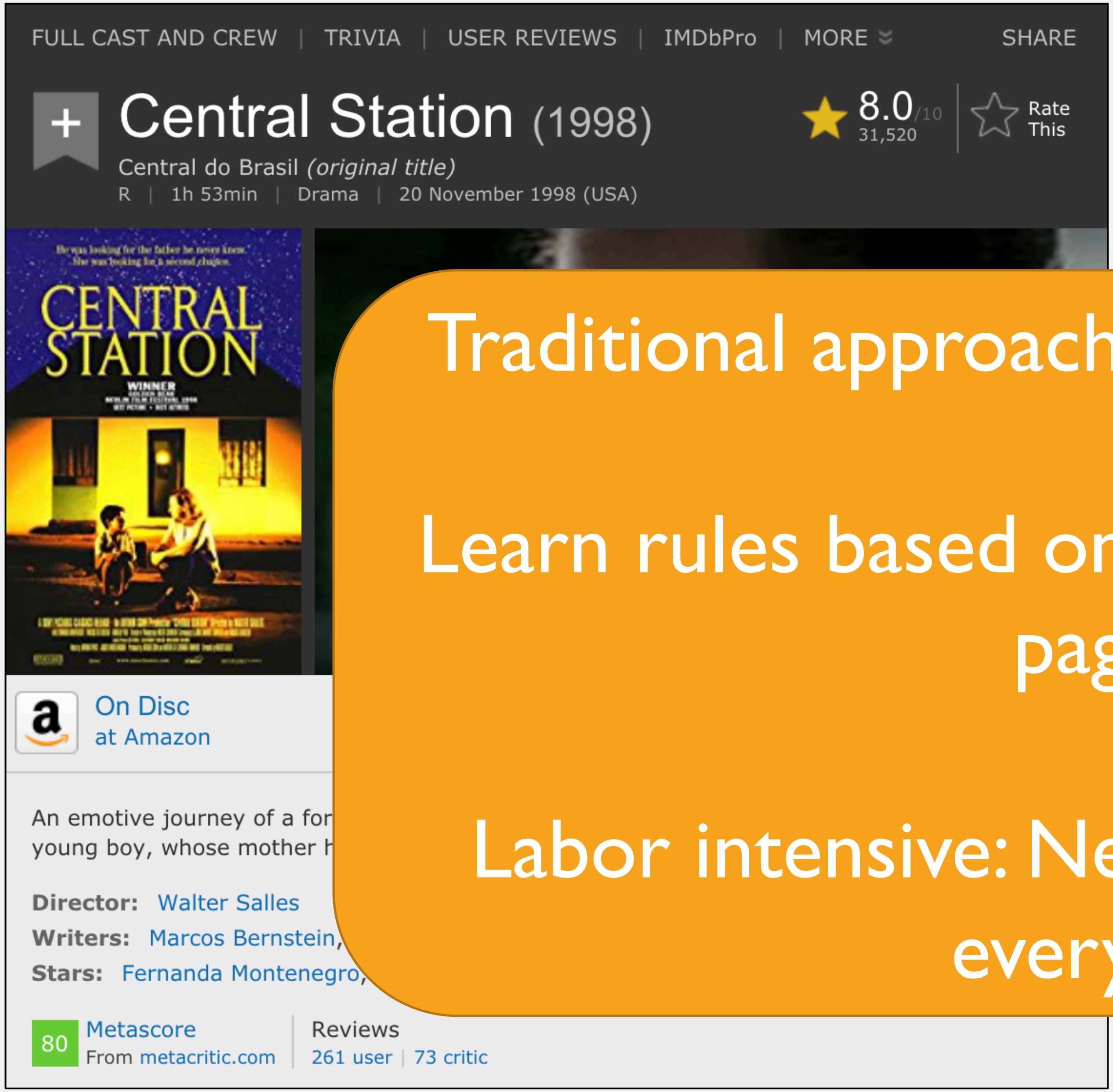
Page formats can differ slightly

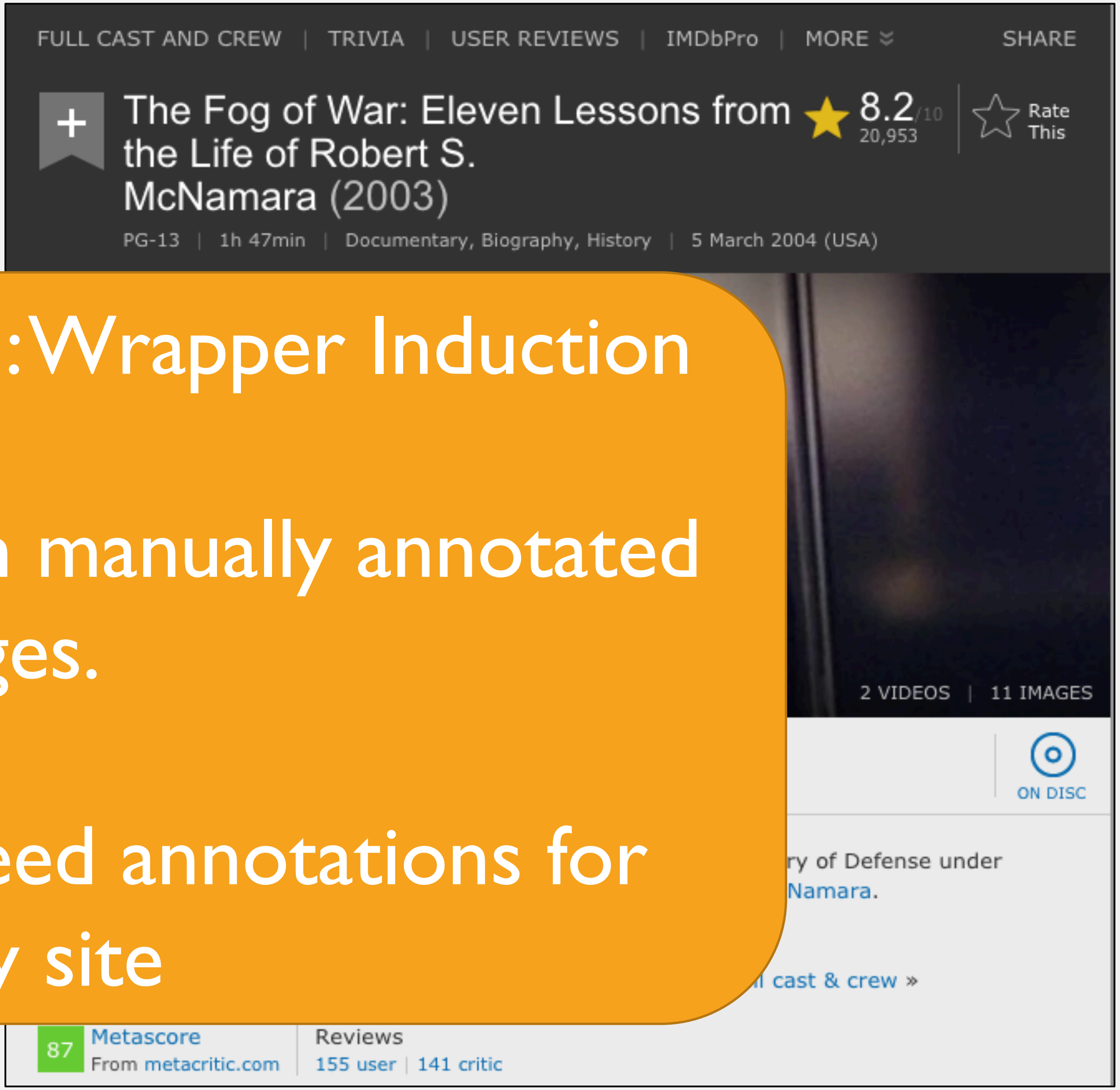
Same predicate has different location on different pages.

Same location has different predicates on different pages









Traditional approach: Wrapper Induction

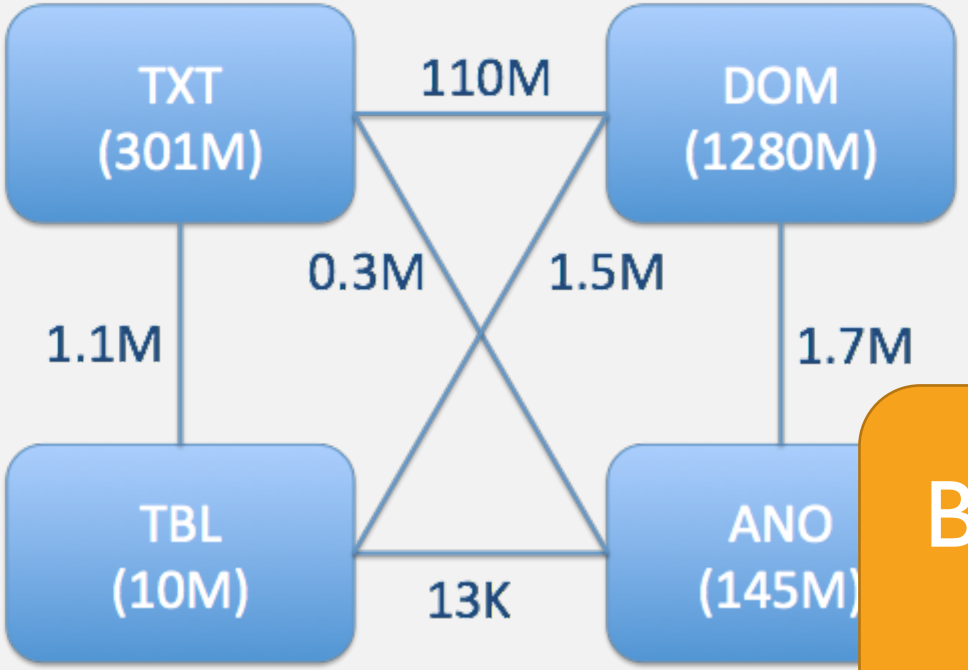
Learn rules based on manually annotated pages.

Labor intensive: Need annotations for every site

BIG PROMISE FROM SEMI-STRUCTURED DATA

- Knowledge Vault @ Google showed big potential from distantly supervised DOM-tree extraction [Dong et al., KDD'14][Dong et al., VLDB'14]

Accu	Accu (conf $\geq$ .7)
0.36	0.52



Accu	Accu (conf $\geq$ .7)
0.43	0.63
0.09	0.62

Can't find new entities

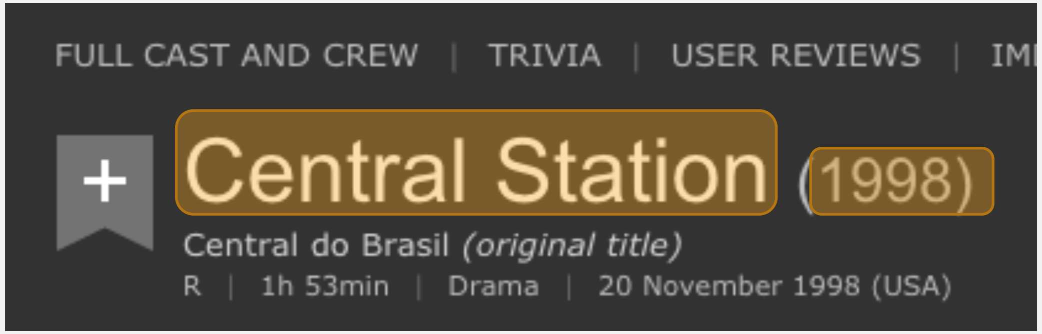
But accuracy is still low

Can we automatically and accurately extract from semi-structured pages?

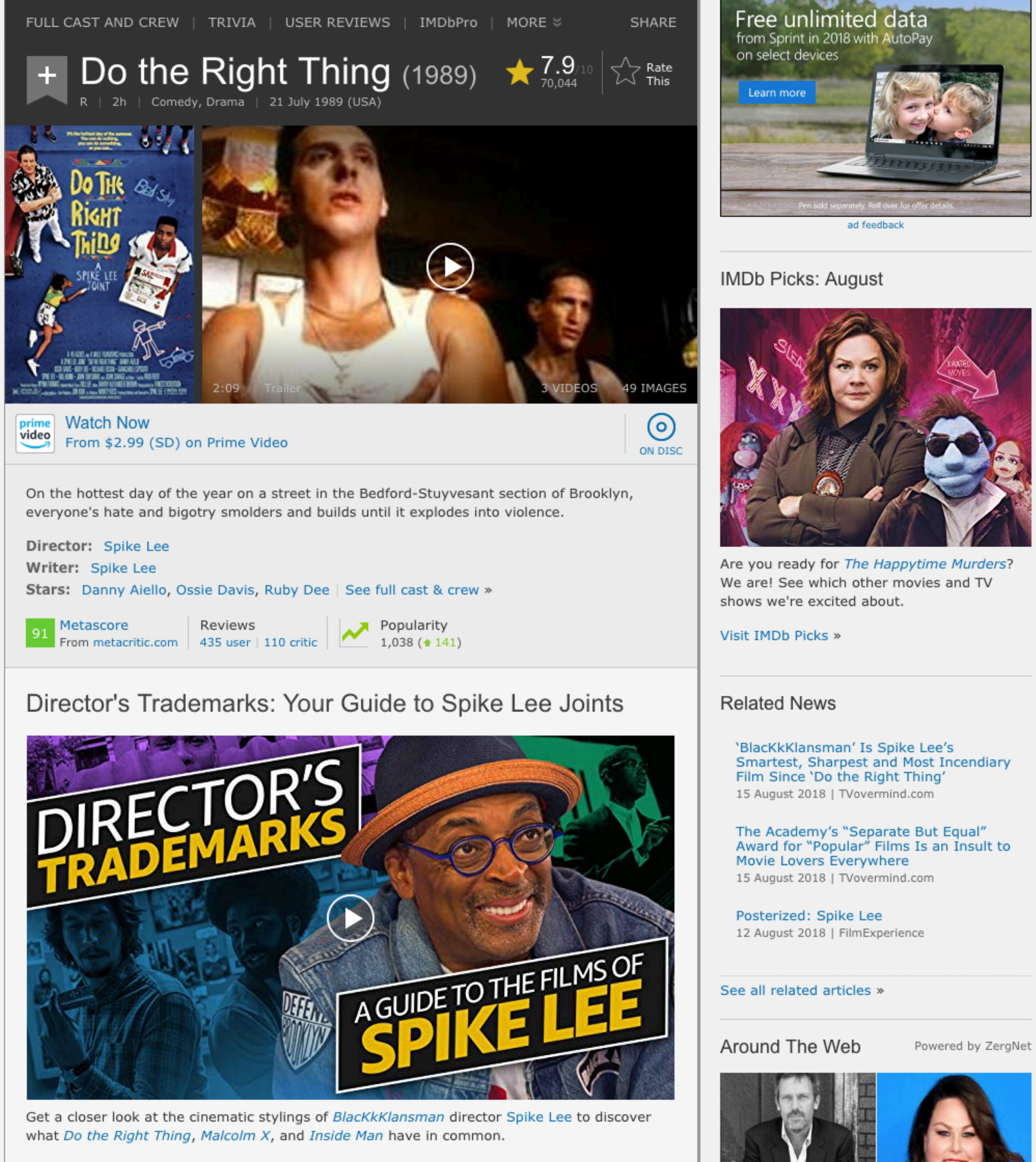
PROBLEM DEFINITION

- Input:
  - Pages from a semi-structured website
  - Seed KB (and ontology)
- Output:
  - Newly extracted triples corresponding to the given ontology
  - Subject and object are strings
    - We do not address entity linkage or knowledge fusion

DISTANT SUPERVISION



- Assume joint mentions of subjects and objects indicate relation
- Use KB to automatically annotate data
- Use this as training data to learn classifier



Problem:

Webpages may mention thousands of entities, creating millions of pairs:

1. Computationally complex.
2. Spurious matches very likely

+ Do the Right Thing (1989) 7.9/10 70,044 Rate This

R | 2h | Comedy, Drama | 21 July 1989 (USA)

Automatic annotation is hard: Same person is writer and director

Director: Spike Lee

Writer: Spike Lee

Stars: Danny Aiello, Ossie Davis, Ruby Dee | See full cast & crew »

On the hottest day of the year on a street in the Bedford-Stuyvesant section of Brooklyn, everyone's hate and bigotry smolders and builds until it explodes into violence.

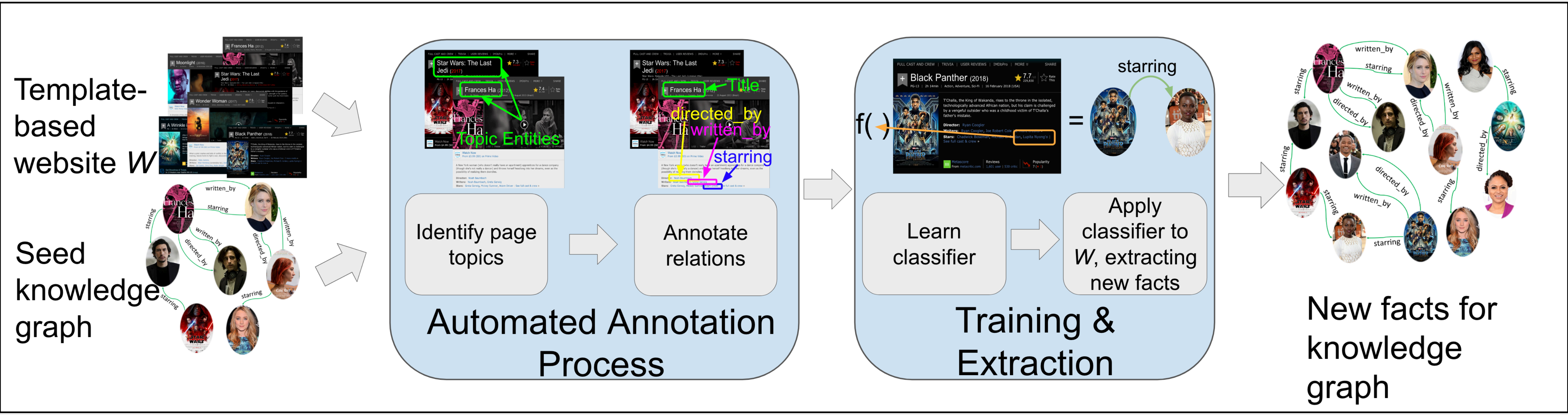
Director: Spike Lee

Writer: Spike Lee

Stars: Danny Aiello, Ossie Davis, Ruby Dee | See full cast & crew »


Do the Right Thing (1989)
R | 2h | Comedy, Drama | 21 July 1989 (USA)
7.9/10
70,044
Rate This
More Like This
Learn more
Crooklyn (1994)
PG-13 Comedy | Drama
6.9/10
Spike Lee's vibrant semi-autobiographical portrait of a school teacher, her stubborn jazz musician husband and their five kids living in Brooklyn in 1973.
Add to Watchlist
Next »
Director: Spike Lee
Stars: Alfre Woodard, Delroy Lindo, ...
Prev 6 Next 6
On the h
everyone
Director
Writer:
Stars: Danny Aiello, Ossie Davis, Ruby Dee | See full cast & crew »
Same person also mentioned in recommendation

CERES OVERVIEW

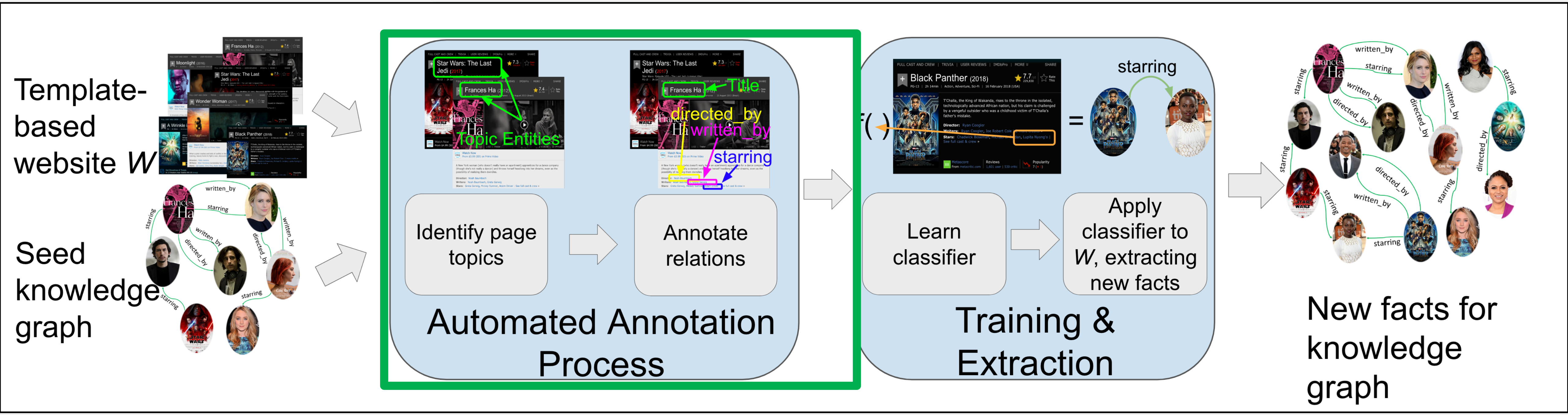


Contribution 1:
CERES Annotator

Contribution 2:
CERES Extractor

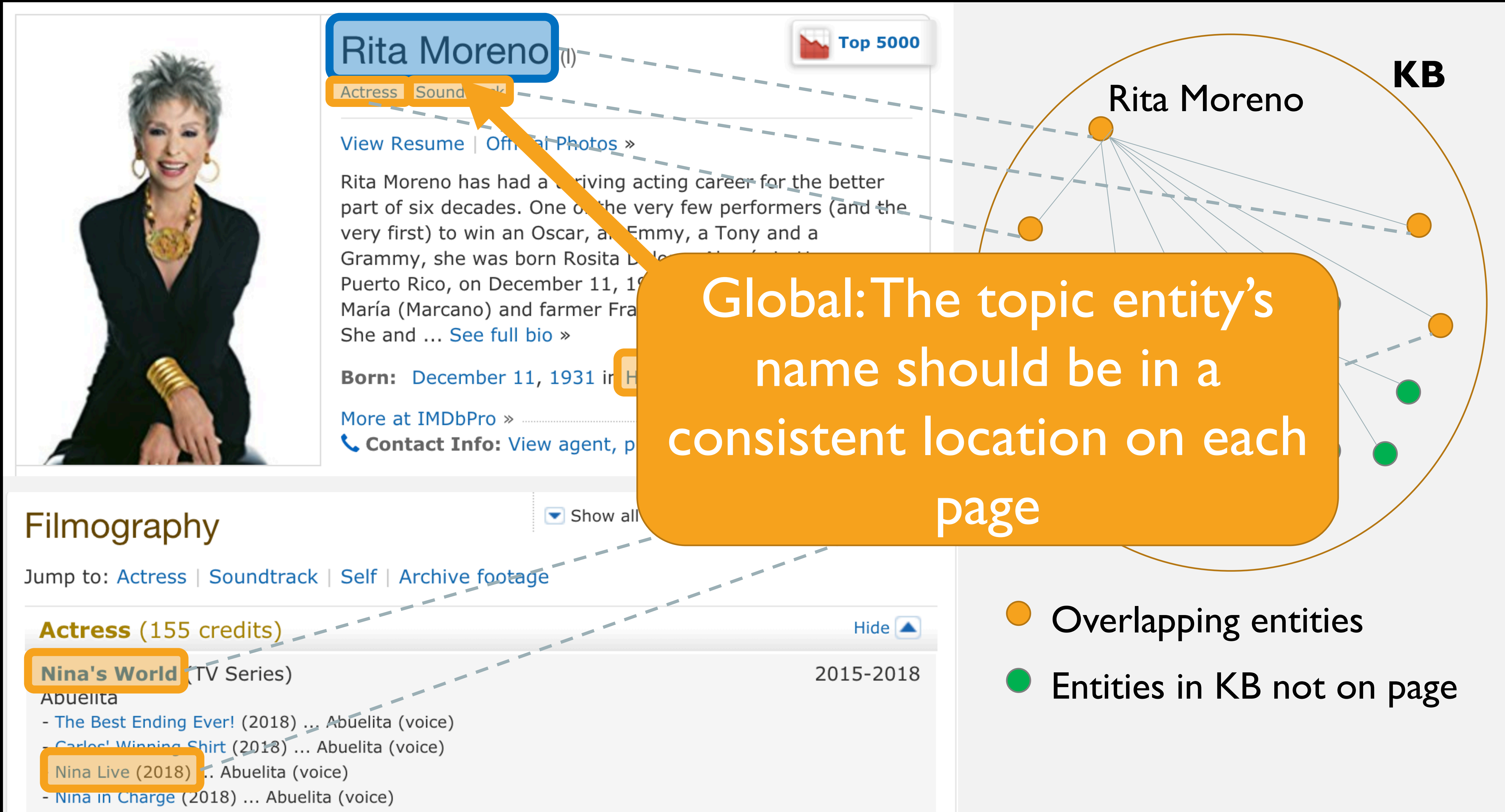
	No manual annotations	Finds new entities	Accurate
Supervised Wrapper Induction	✗	✓	✓
Knowledge Vault	✓	✗	✗
CERES	✓	✓	✓

AUTOMATED ANNOTATION PROCESS



Use both local (on a single page) and global (site-wide patterns) information

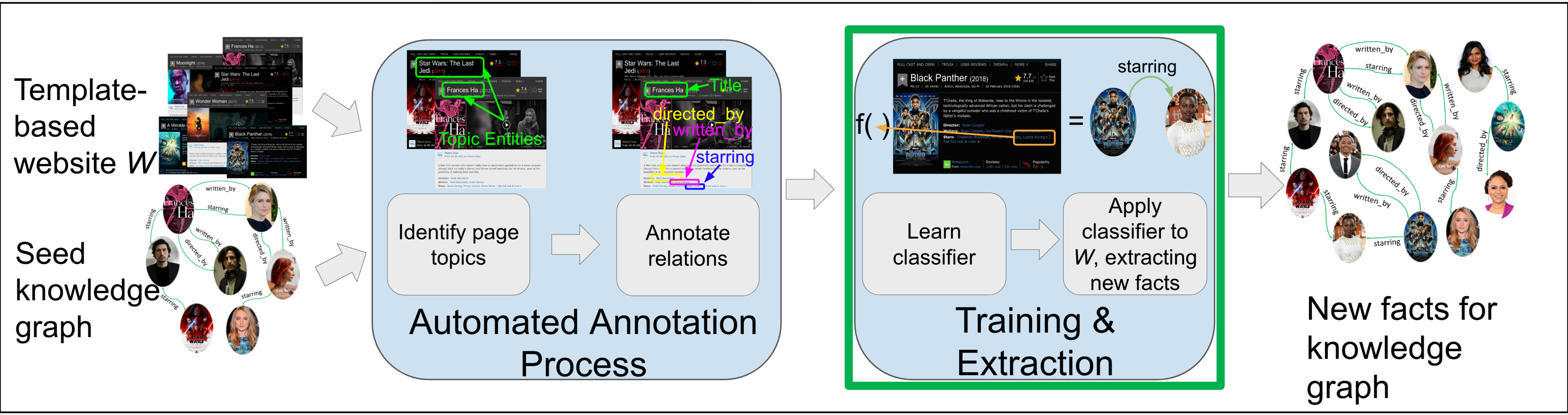
Rita Moreno (I)
Actress
Soundtrack
Top 5000
View Resume | Official Photos »
Rita Moreno has had a thriving acting career for the better part of six decades. One of the very few performers (and the very first) to win an Oscar, an Emmy, a Tony and a Grammy, she was born Rosita Dolores Alverío in Humacao, Puerto Rico, on December 11, 1931, to seamstress Rosa María (Marcano) and farmer Francisco José "Paco" Alverío. She and ... See full bio »
Born: December 11, 1931 in Humacao, Puerto Rico
More at IMDbPro »
Filmography
Jump to: Actress | Soundtrack | Self
Actress (155 credits)
Nina's World (TV Series)
- The Best Ending Ever! (2018) ... Abuelita (voice)
Nina Live (2018) ... Abuelita (voice)
- Nina in Charge (2018) ... Abuelita (voice)
KB
Rita Moreno
Local: The topic entity should be associated with a large number of entities on the page
rlapping entities
ties in KB not on page


Rita Moreno (I)
Top 5000
Actress
Sound
View Resume | Off al Photos »
Rita Moreno has had a riving acting career for the better part of six decades. One o the very few performers (and the very first) to win an Oscar, a Emmy, a Tony and a Grammy, she was born Rosita D
Puerto Rico, on December 11, 19
María (Marcano) and farmer Fra
She and ... See full bio »
Born: December 11, 1931 in H
More at IMDbPro »
Contact Info: View agent, p
Filmography
Show all
Jump to: Actress | Soundtrack | Self | Archive footage
Actress (155 credits)
Hide
Nina's World (TV Series) 2015-2018
Abuelita
- The Best Ending Ever! (2018) ... Abuelita (voice)
- Carlos' Winning Shirt (2018) ... Abuelita (voice)
- Nina Live (2018) ... Abuelita (voice)
- Nina in Charge (2018) ... Abuelita (voice)
KB
Rita Moreno
Global: The topic entity's name should be in a consistent location on each page
Overlapping entities
Entities in KB not on page

RELATION ANNOTATION

- Annotate known facts found on page
- If ambiguous:
  - Local: Objects of same predicate should be in same section of page
  - Global: Predicates should be in *similar* location on all pages. Cluster all potential mentions of a relation across site, choose most common location.

TRAINING PROCESS



- Probabilistic classifier
- Robust to noise in training data
  - (compared to wrapper induction)

MODEL & FEATURES

- Multi-class logistic regression model
  - Input: Featurized DOM node
  - Output: Relation label (or "None")
- Features based on Vertex (Gulhane et al, ICDE 2011)
  - Tag, ID, Class of ancestors/siblings
  - DOM path to template strings
- Important to limit # of features to prevent overfitting to noise in training data

Experiments show CERES is competitive with state-of-the-art supervised extractors.

BENCHMARK DATASET

- Baselines:
  - CERES-Baseline: Naïve Distant Supervision
  - CERES-Topic: Uses topic identification, but does not resolve ambiguous objects
- SWDE dataset (Hao et al, SIGIR 2011)
- 10 sites in each of 4 domains (Book, Movie, NBA, University)

SWDE

System	Manual Labels	Movie	NBA Player	University	Book
Hao *et al.* [19]	yes	0.79	0.82	[illegible]	[illegible]
XTPath [7]	yes	0.94	**0.98**	[illegible]	[illegible]
BigGrams [26]	yes	0.74	0.90	[illegible]	[illegible]
LODIE-Ideal [15]	no	0.86	0.9	[illegible]	[illegible]
LODIE-LOD [15]	no	0.76	0.87 [a]	0.91	0.78
RR+WADaR [29]	no	0.73	0.80	0.79	0.70
RR+WADaR 2 [30]	no	0.75	0.91	0.79	0.71
WEIR [4]	no	0.93	0.89	0.97	0.91
Vertex++	yes	0.90	0.97	**1.00**	0.94
CERES-Baseline	no	NA [b]	0.78	0.72	0.27
CERES-Topic	no	**0.99** [a]	0.97	0.96	0.72
CERES-Full	no	**0.99** [a]	**0.98**	0.94	0.76

State-of-the-art results in two verticals.
(better than supervised systems!)

SWDE

Competitive with top unsupervised systems in a third

	Manual Labels	Movie	NBA Player	University	Book
	yes	0.79	0.82	0.83	0.86
	yes	0.94	**0.98**	0.98	**0.97**
BigGrams [26]	yes	0.74	0.90	0.79	0.78
LODIE-Ideal [15]	no	0.86	0.9	0.96	0.85
LODIE-LOD [15]	no	0.76	0.87 [a]	0.91 [a]	0.78
RR+WADaR [29]	no	0.73	0.80	0.79	0.70
RR+WADaR 2 [30]	no	0.75	0.91	0.79	0.71
WEIR [4]	no	0.93	0.89	0.97	0.91
Vertex++	yes	0.90	0.97	**1.00**	0.94
CERES-Baseline	no	NA [b]	0.78	0.72	0.27
CERES-Topic	no	**0.99** [a]	0.97	0.96	0.72
CERES-Full	no	**0.99** [a]	**0.98**	0.94	0.76

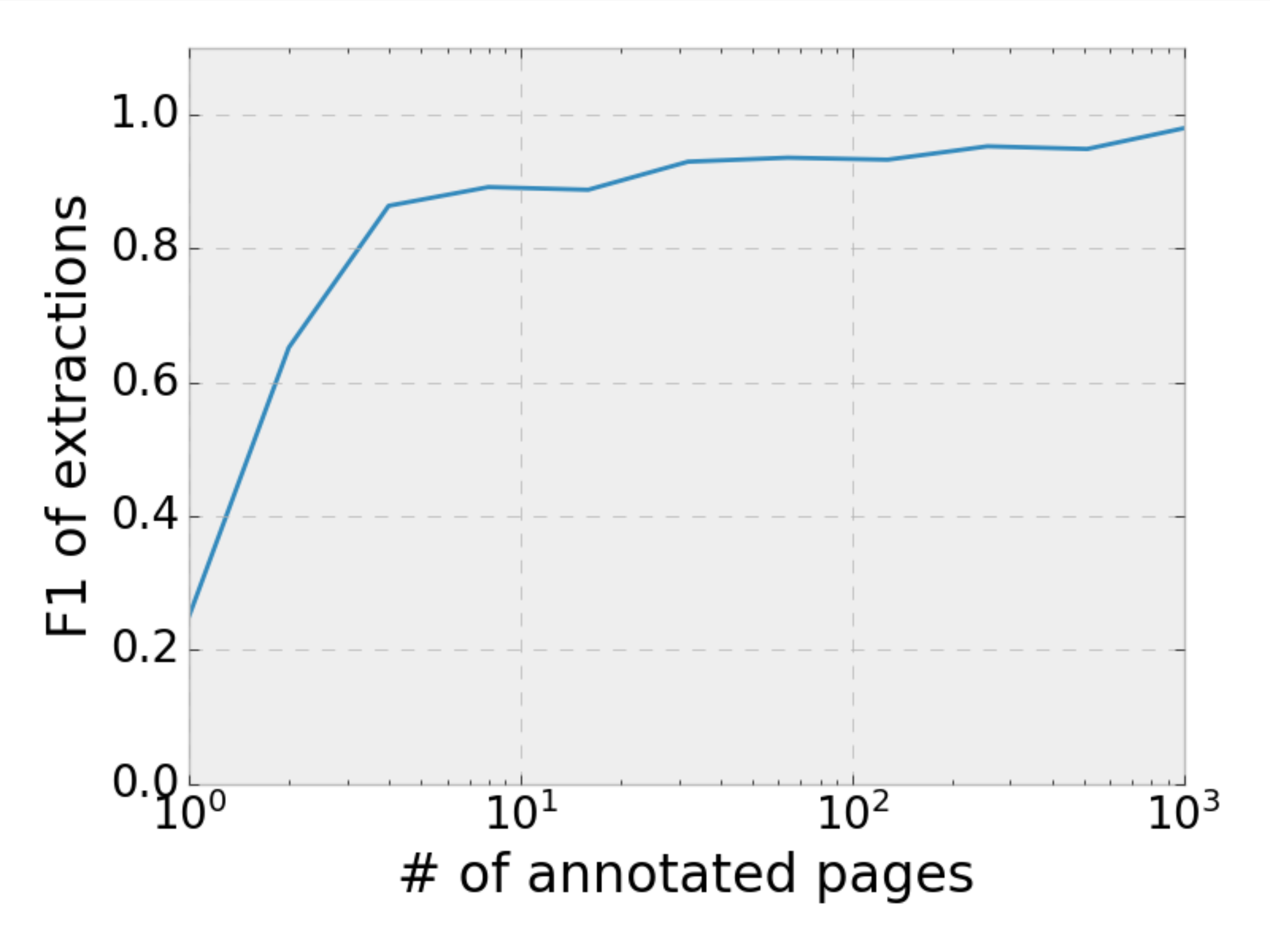
SWDE

System	Manual Labels	Movie	NBA Player	University	Book
Hao *et al.* [19]				0.83	0.86
XTPath [7]				0.98	**0.97**
BigGrams [26]				0.79	0.78
LODIE-Ideal [15]				0.96	0.85
LODIE-LOD [15]	no	0.76	0.87	0.91[a]	0.78
RR+WADaR [29]				0.79	0.70
RR+WADaR 2 [30]				0.79	0.71
WEIR [4]				0.97	0.91
Vertex++				**1.00**	0.94
CERES-Baseline	no	NA[b]	0.78	0.72	0.27
CERES-Topic	no	**0.99**[a]	0.97	0.96	0.72
CERES-Full	no	**0.99**[a]	**0.98**	0.94	0.76

Weak on one vertical. Why?

Only a few overlapping entities between websites and seed KB

ABOUT 10 ANNOTATED PAGES ARE NEEDED TO LEARN

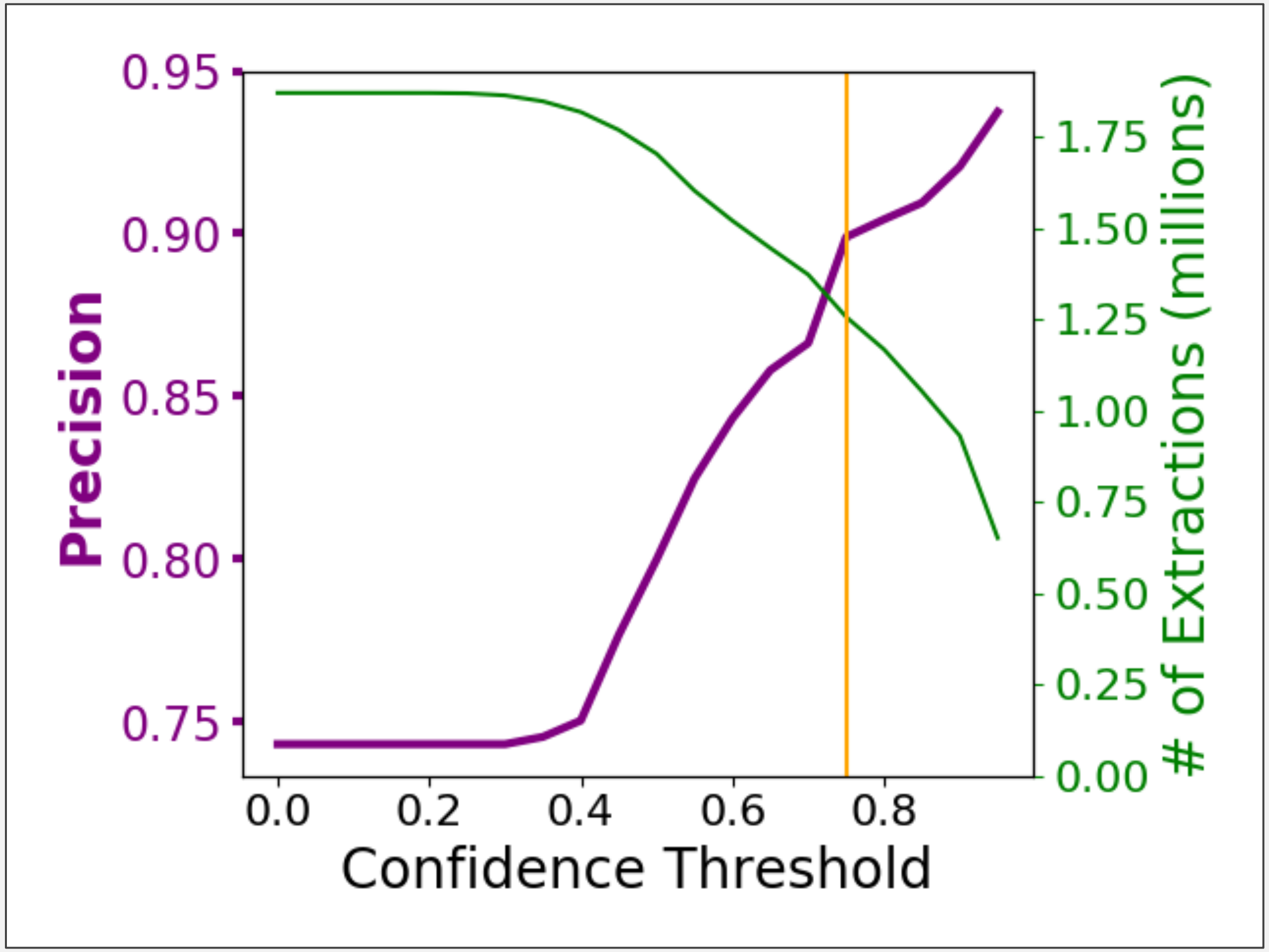


COMMONCRAWL MOVIE WEBSITES

- 33 websites, 400,000 pages (not all semi-structured)
- Mostly long-tail (foreign, documentary, animated)
- 7 languages

COMMONCRAWL MOVIE WEBSITES

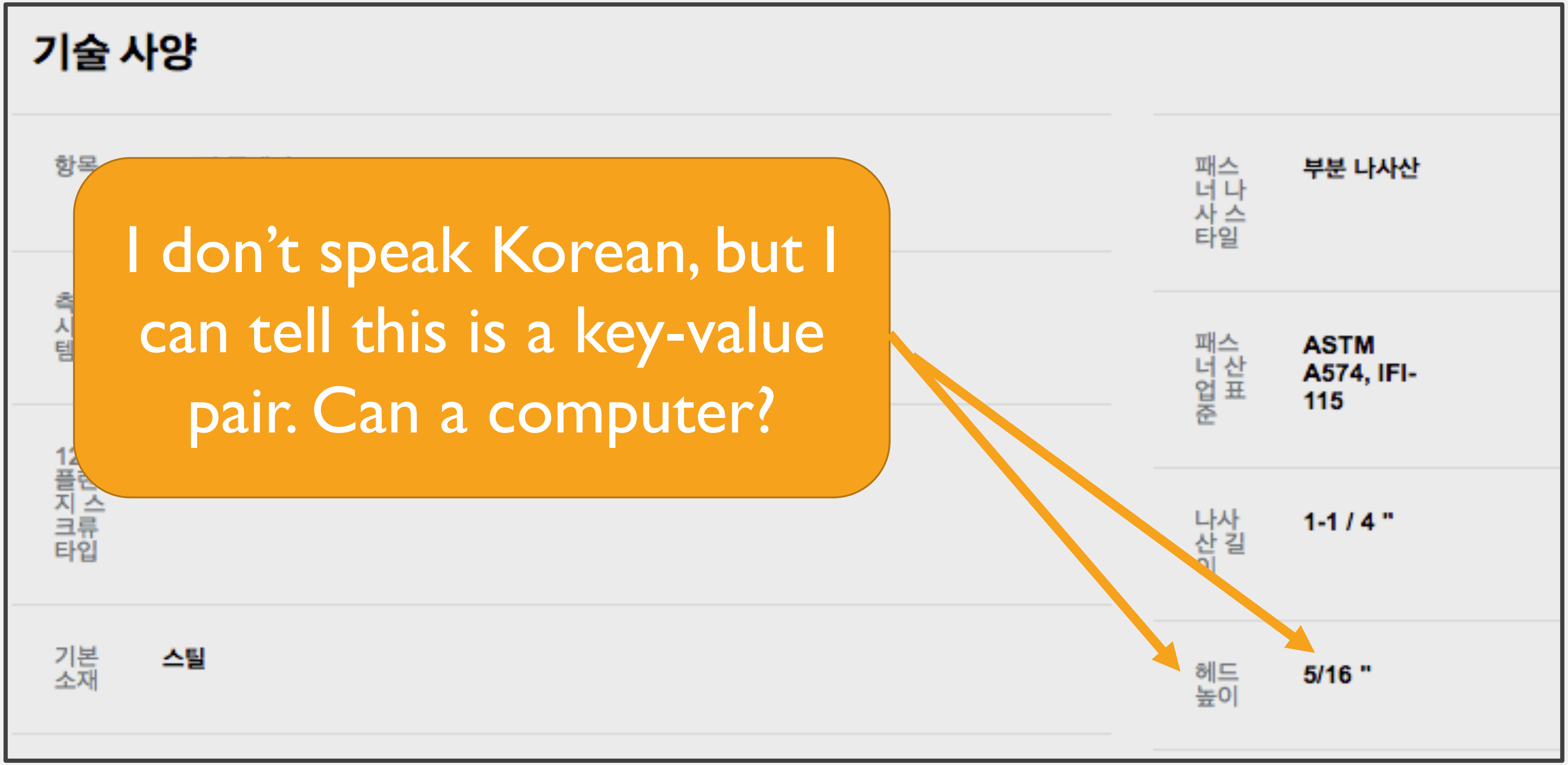
- 90% precision (compared to ~63% with Knowledge Vault)
- 1.25 million extractions
- Extracted 2.6 new entities for every annotated entity



TAKEAWAYS

- Automatic extraction from semi-structured pages is practical
  - If you have existing KB you can:
    - Identify new entities
    - Extract new facts with high-precision
- Probabilistic classifiers are effective at semi-structured extraction
  - Allow for precision/recall tradeoff

FUTURE WORK



I don’t speak Korean, but I can tell this is a key-value pair. Can a computer?