

OPENCERES: WHEN OPEN INFORMATION EXTRACTION MEETS THE SEMI-STRUCTURED WEB

Colin Lockard, Prashant Shiralkar, Xin Luna Dong

W PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING



amazon

6 QUESTIONS YOU MIGHT HAVE

- What are semi-structured websites?
- Why do we want to extract from them?
- Why OpenIE?
- Is this problem hard?
- Well then how are you going to solve it?
- Does a dataset even exist to evaluate these extractions anyway?

6 QUESTIONS YOU MIGHT HAVE

- What are semi-structured websites?
 - Exactly what they sound like
- Why do we want to extract from them?
 - They have lots of information
- Why OpenIE?
 - They have LOTS of information
- Is this problem hard?
 - Yes.
- Well then how are you going to solve it?
 - By automatically creating training data
- Does a dataset even exist to evaluate these extractions anyway?
 - It does now.

Enjoy unlimited streaming on Prime Video.

Includes thousands of titles. Start your 30-day free trial.

[Start your 30-day free trial](#)[FULL CAST AND CREW](#)[TRIVIA](#)[USER REVIEWS](#)[IMDbPro](#)[MORE](#)[SHARE](#)

+ When Harry Met Sally... (1989)

R

1h 36min

Comedy, Drama, Romance

21 July 1989 (USA)

 7.6 / 10
160,124

Rate This



Harry and Sally have known each other for years, and are very good friends, but they fear sex would ruin the friendship.

Director: Rob Reiner**Writer:** Nora Ephron**Stars:** Billy Crystal, Meg Ryan, Carrie Fisher |[See full cast & crew »](#)

76 Metascore

From metacritic.com

Reviews

283 user | 122 critic

 Popularity
1,178 (▲ 657)[Watch Now](#)

From \$2.99 (SD) on Amazon Video

**Nominated for 1 Oscar.** Another 4 wins & 17 nominations. [See more awards »](#)[Photos](#)

Entertainment included with Prime

[Start your free trial](#)[ad feedback](#)

Scary Good: IMDb's Guide to Horror



Can't get enough of movies and TV shows that scare up a good fright? Check out [Scary Good](#), IMDb's Horror Entertainment Guide.

IMDb

Find Movies, TV shows, Celebrities and more...

All

IMDbPro | Help

Movies, TV & Showtimes

Celebs, Events & Photos

News & Community

Watchlist

Sign in with Facebook

Other Sign in options

Enjoy unlimited streaming on Prime Video.
Includes thousands of titles. Start your 30-day free trial.

Start your free trial

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

+ When Harry Met Sally... (1989)

R | 1h 36min | Comedy, Drama, Romance | 21 July 1989 (USA)



Harry and Sally have known each other for years, good friends, but they fear sex would ruin the fr...

Director: Rob Reiner
Writer: Nora Ephron
Stars: Billy Crystal, Meg Ryan, Carrie Fisher | See full cast & crew »

76 Metascore From metacritic.com

Reviews 283 user | 122 critic

Watch Now From \$2.99 (SD) on Amazon Video

Nominated for 1 Oscar. Another 4 wins & 17 nominations. See more awards »

Photos

Semi-structured:

- Rich layout features
- Each page provides info about an entity
- Data represented as key-value pairs and lists
- Text fields consisting of just entity/attribute name

There are lots of semi-structured websites

A collage of semi-structured website snippets:

- Top Left:** A movie poster for "Jab Harry Met Sejal". Below it is a screenshot of the NMDb (Now Playing Movie Database) website showing details for the movie.
- Middle Left:** A screenshot of the AllMusic website showing album highlights for "The Beatles - Abbey Road".
- Bottom Left:** A screenshot of the hifiengine website showing a product page for a digital integrated amplifier.
- Right Side:** A sidebar from a football statistics page showing the 2017 season stats for Aaron Rodgers.

hifiengine

News • Library • Database • Gallery • Requests • FAQ • VE

User Login
[Login/Register](#)

Manual Library / A and D

A And D DA-U7000

Digital Integrated Amplifier (1988)

★★★★★ add a review

Latest Comments

- Re: Sx-950
- Re: Sw-156
- Re: M-5030
- Re: Picture Of My Hitachi...
- Re: Onkyo Model No....
- Re: Sg-9500
- Re: A-50
- Re: Control 5

Who's Online

There are currently 153 Users And 596 Guests Online.

We Have 505,556 Registered Users

Specifications

Power output: 120 watts per channel into 6Ω (stereo)

Frequency response: 3Hz to 100kHz

Total harmonic distortion: 0.03%

Signal to noise ratio: 85dB (MM), 100dB (line)

Digital inputs: coaxial, optical

Dimensions: 425 x 177 x 152mm

+ When Harry Met Sally... (1989)

R | 1h 36min | Comedy, Drama, Romance | 21 July 1989 (USA)



A NEW FRIEND TO ROB REINER

RATED R

Can two friends sleep together and still love each other in the morning?

Harry and Sally have known each other for years, and are very good friends, but they fear sex would ruin the friendship.

Director: Rob Reiner**Writer:** Nora Ephron**Stars:** Billy Crystal, Meg Ryan, Carrie Fisher |[See full cast & crew »](#)

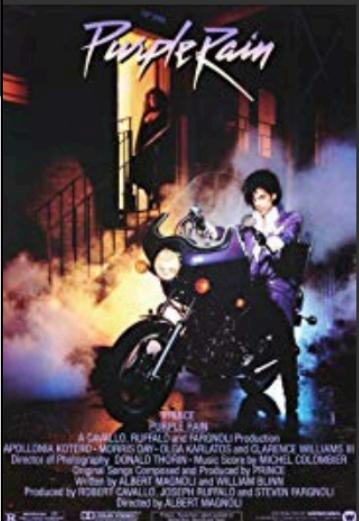
★ 7.6 /10
160,124

Rate This



Purple Rain (1984)

R | 1h 51min | Drama, Music, Musical | 27 July 1984 (USA)



0:17 | Trailer

★ 6.5 /10
17,575

Rate This

2 VIDEOS | 68 IMAGES



Semi-structured websites
have lots of pages with very
similar template structure



Watch Now
From \$3.99 on Prime Video

A young musician, tormented by an abusive situation at home, must contend with a rival singer, a burgeoning romance, and his own dissatisfied band, as his star begins to rise.

Director: Albert Magnoli**Writers:** Albert Magnoli, William Blinn**Stars:** Prince, Apollonia Kotero, Morris Day | [See full cast & crew »](#)

And they have a lot of information

Knowledge Vault
@ Google found
4x more facts
from semi-
structured than
unstructured text
[Dong et al.,
KDD'14][Dong et al.,
VLDB'14]

Produced by

Nora Ephron ... associate producer
Steve Nicolaides ... co-producer
Rob Reiner ... producer
Andrew Scheinman ... producer
Jeffrey Stott ... co-producer

Cinematography by

Barry Sonnenfeld ... director of photography

Film Editing by

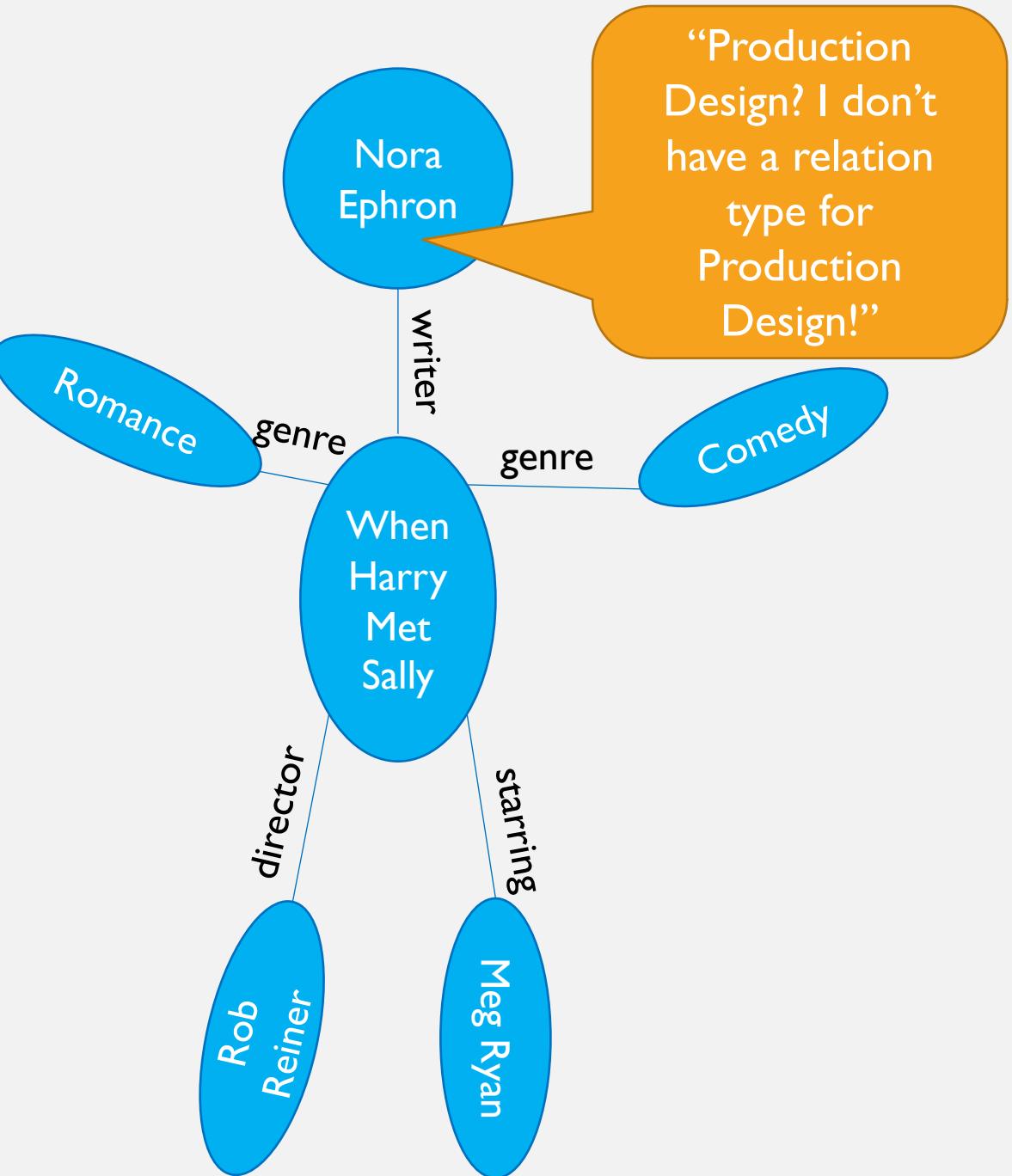
Robert Leighton

Casting By

Janet Hirshenson
Jane Jenkins

Production Design by

Jane Musky



Produced by

Nora E.
Steve
Rob Re
Andrew
Jeffrey Scott

... co-producer

Cinematography by

Barry Sonnenfeld

... director of photography

Film Editing by

Robert Leighton

Casting By

Janet Hirshenson
Jane Jenkins

Production Design by

Jane Musky

They might have TOO MUCH information

TRADITIONAL EXTRACTION

The screenshot shows the IMDb page for the movie 'When Harry Met Sally... (1989)'. At the top, there's a navigation bar with links for 'FULL CAST AND CREW', 'TRIVIA', 'USER REVIEWS', 'IMDbPro', and 'MORE'. Below the title, there's a rating of 7.6/10 from 160,124 users. A yellow box highlights the title 'When Harry Met Sally... (1989)'. To the right of the title is a 'Rate This' button. Below the title, it says 'R | 1h 36min | Comedy, Drama, Romance | 21 July 1989 (USA)'. On the left, there's a movie poster showing a man and a woman standing on a bridge over a city skyline. The plot summary below the poster states: 'Harry and Sally have known each other for years, and are very good friends, but they fear sex would ruin the friendship.' The director is listed as 'Rob Reiner', the writer as 'Nora Ephron', and the stars as 'Billy Crystal, Meg Ryan, Carrie Fisher'. There's also a link to 'See full cast & crew'. At the bottom, there's a Metascore of 76 from Metacritic.com, reviews from 283 users and 122 critics, and a popularity metric of 1,178 with a green upward arrow indicating growth of 657. An 'ON DISC' button is also present. A yellow box at the bottom says 'Nominated for 1 Oscar. Another 4 wins & 17 nominations. See more awards »'.

TOPIC ENTITY

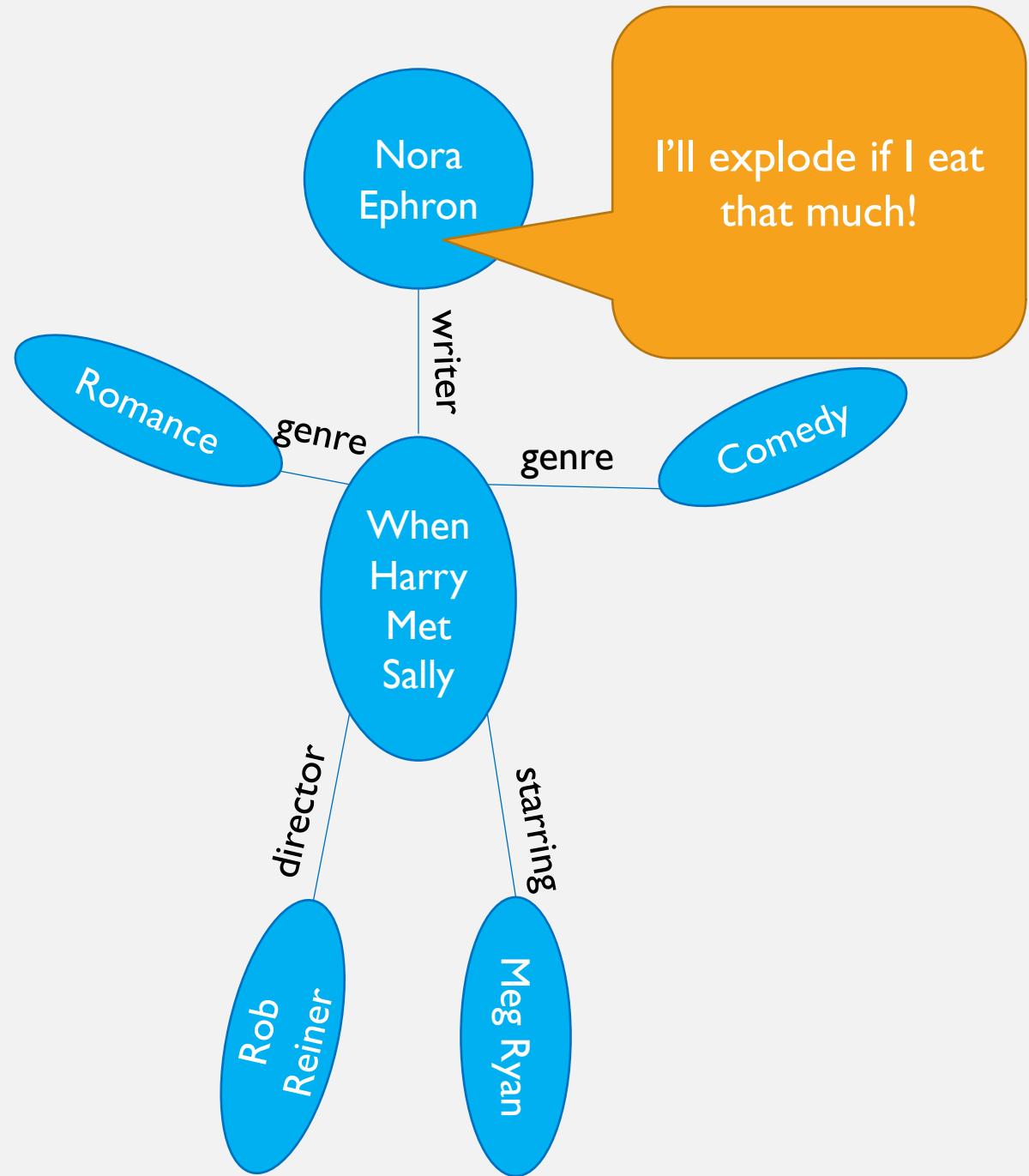
“When Harry Met Sally”

Relation types defined
in ontology

(film.written_by, “Nora Ephron”)

(film.actor, “Meg Ryan”)

On 10 semi-structured movie websites, the IMDb ontology covered only 7% of relations.



OPEN INFORMATION EXTRACTION¹

- Instead of relying on ontology for relation type, just extract relation string from page
- Widely explored in unstructured text
- Not explored on semi-structured websites

¹ Michele Banko, Michael Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. “Open Information Extraction from the Web.” *IJCAI* (2008)

PROBLEM DEFINITION

- Input:
 - Pages from a semi-structured website
 - Seed KB (and ontology)
- Output:
 - A set of triples (s, r, o) , where
 - s is a string corresponding to the subject (page topic entity),
 - o is a string corresponding to the object
 - r is a string indicating the relation/predicate

OPEN INFORMATION EXTRACTION

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

+ When Harry Met Sally... (1989)

7.6 /10 160,124 Rate This

R | 1h 36min | Comedy, Drama, Romance | 21 July 1989 (USA)

Harry and Sally have known each other for years, and are very good friends, but they fear sex would ruin the friendship.

Director: Rob Reiner

Writer: Nora Ephron

Stars: Billy Crystal, Meg Ryan, Carrie Fisher | See full cast & crew »

76 Metascore From metacritic.com

Reviews 283 user | 122 critic

Popularity 1,178 (▲ 657)

ON DISC

Watch Now From \$2.99 (SD) on Amazon Video

Nominated for 1 Oscar. Another 4 wins & 17 nominations. See more awards »

TOPIC ENTITY

“When Harry Met Sally”

(“Writer:”, “Nora Ephron”)

(“Stars:”, “Meg Ryan”)

OpenIE triples can be used for:

- KB Completion
 - See our paper “[OpenKI: Integrating Open Information Extraction and Knowledge Bases with Relation Inference](#)” by Dongxu Zhang et al!
- Question Answering
- Better KG Embeddings
- Ontology Discovery
- Fact Checking

To make use of all facts on these sites, we need Open Information Extraction

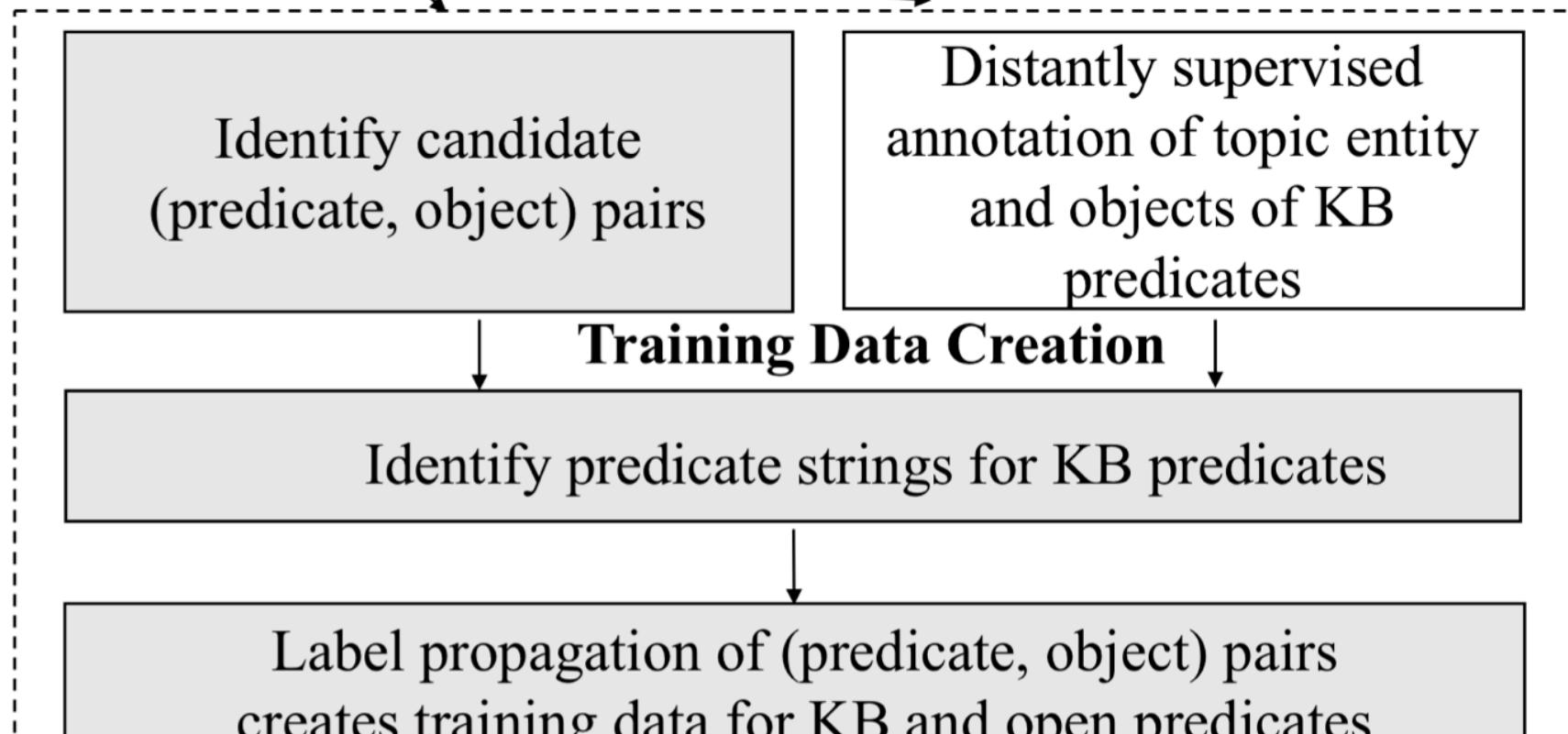
WHAT ARE THE CHALLENGES?

WHY IS THIS HARD?

- Too much variety in expression for rules
- Need to consider **pairs** of text fields
- No labeled data
 - Lots of websites, would need lots of annotations
- Traditional DOM (HTML) features too limited to find new relation types

Semi-structured website W

Knowledge base (KB)



Supervised learning

DOM node
classifier

AUTOMATIC SEED DATA CREATION

Tape (2001)

Add to My Favorites

Genres - Drama | Sub-Genres - Psychological Drama, Reunion Films

Cast



Ethan Hawke

Vince



Robert Sean Leonard

Johnny



Uma Thurman

Amy

Crew

Richard Linklater

Director

Maryse Alberti

Cinematographer

Step 1: Identify candidate pairs

- Use site-wide statistics
 - Relation string must be more common than object string
- Relation and object should be “nearby” in visual or DOM distance
- Reduces # pairs from quadratic to quasilinear.

AUTOMATIC SEED DATA CREATION

Tape (2001)

Add to My Favorites

(“Tape”, film.genre, “Drama”)

Genres - Drama

Sub-Genres - Psychological Drama, Reunion Films

Cast



Ethan Hawke



Robert Sean Leonard

Johnny



Uma Thurman

(“Tape”, film.actor, “Uma Thurman”)

(“Tape”, film.director, “Richard Linklater”)

Crew

Richard Linklater

Director

Maryse Alberti

Cinematographer

Caroline Kaplan

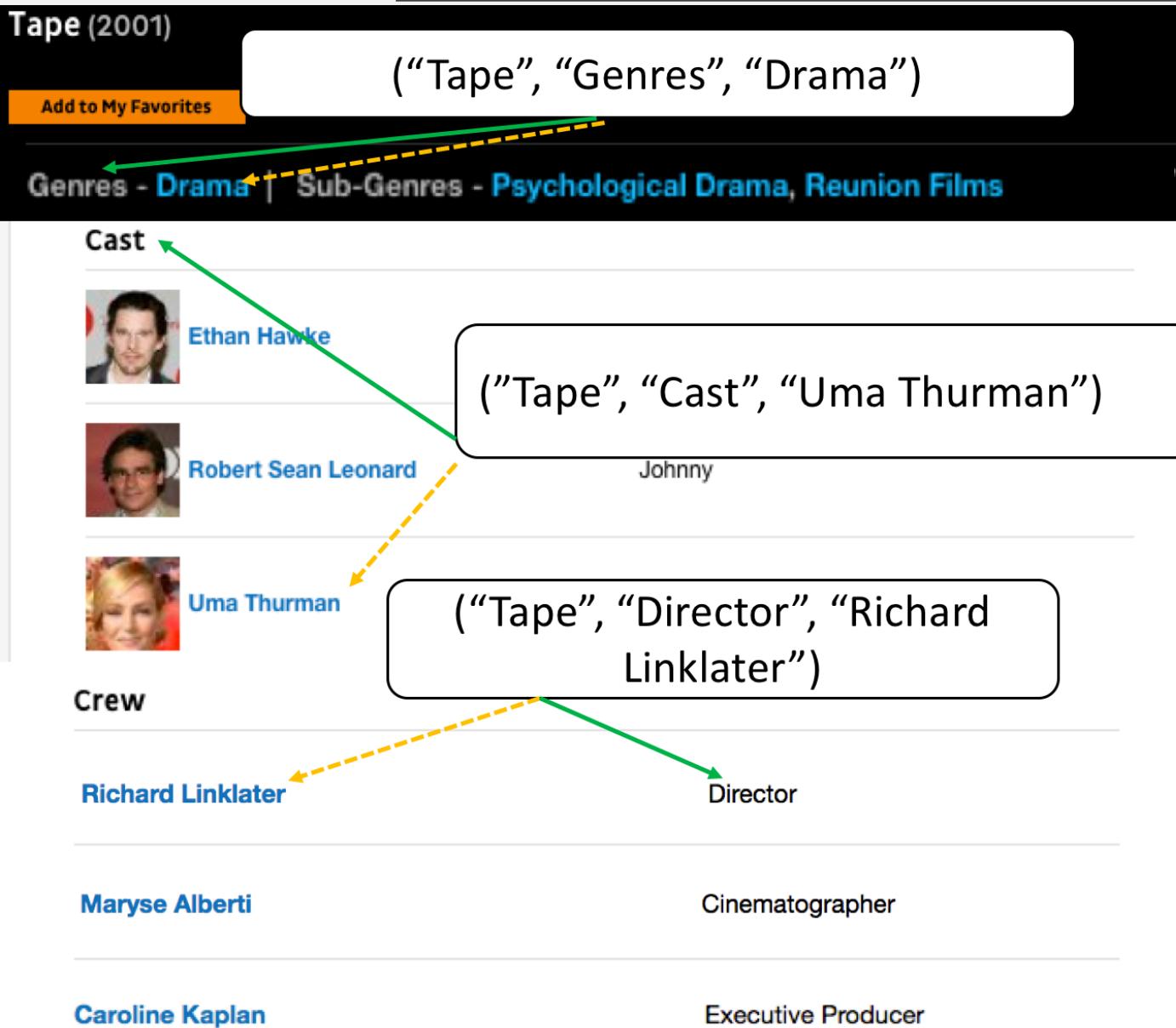
Executive Producer

Step 2:

- Distantly supervised annotation of facts in existing knowledge base²

² Lockard, Colin, Xin Dong, Prashant Shiralkar and Arash Einolghozati. “CERES: Distantly Supervised Relation Extraction from the Semi-Structured Web.” PVLDB 11 (2018): 1084-1096.

AUTOMATIC SEED DATA CREATION



Step 3:

- Use lexicon or embedding similarity to find relation string
- Annotation now defines a *pair* of text fields

AUTOMATIC SEED DATA CREATION

Tape (2001)

Add to My Favorites

(“Tape”, “Genres”, “Drama”)

Genres - Drama Sub-Genres - Psychological Drama, Reunion Films

Cast

Ethan Hawke

(“Tape”, “Cast”, “Uma Thurman”)

Robert Sean Leonard Johnny

Uma Thurman

(“Tape”, “Director”, “Richard Linklater”)

Crew

Richard Linklater Director

Maryse Alberti Cinematographer

Caroline Kaplan Executive Producer

```
graph TD; Drama[Drama] -- "Green Arrow" --> Ethan[Ethan Hawke]; Drama -- "Yellow Dashed Arrow" --> Uma[Uma Thurman]; Director[Director] -- "Yellow Dashed Arrow" --> Richard[Richard Linklater]
```

Step 4:

- Identify pairs of text fields similar to our seed pairs

AUTOMATIC TRAINING DATA CREATION

- Use **visual features** to understand similarity
 - Font
 - Font size
 - Bold/Italic
 - Color
 - Horizontal and vertical location
 - Horizontal/Vertical distance between relation and object
- Graph-based label propagation

SIMILARITY GRAPH CONSTRUCTION

Tape (2001)

Add to My Favorites

Genres - Drama | Sub-Genres - Psychological Drama, Reunion Films

Cast



Ethan Hawke

Vince



Robert Sean Leonard

Johnny



Uma Thurman

Amy

Crew

Richard Linklater

Director

Maryse Alberti

Cinematographer

Candidate pairs become nodes in graph.

Edges weighted by similarity metric based on visual and layout features.

SIMILARITY GRAPH CONSTRUCTION

Tape (2001)

Add to My Favorites

Genres - Drama | Sub-Genres - Psychological Drama, Reunion Films

Cast



Ethan Hawke

Vince



Robert Sean Leonard

Johnny



Uma Thurman

Amy

Crew

Richard Linklater

Director

Maryse Alberti

Cinematographer

Similar:
Relations have same font and size and are vertically aligned

SIMILARITY GRAPH CONSTRUCTION

Tape (2001)

Add to My Favorites

Genres - Drama | Sub-Genres - Psychological Drama, Reunion Films

Cast



Ethan Hawke

Vince



Robert Sean Leonard

Johnny



Uma Thurman

Amy

Crew

Richard Linklater

Director

Maryse Alberti

Cinematographer

Similar:
Objects have same font and size

SIMILARITY GRAPH CONSTRUCTION

Tape (2001)

Add to My Favorites

Genres - Drama | Sub-Genres - Psychological Drama, Reunion Films

Cast



Ethan Hawke



Robert Sean Leonard



Uma Thurman

Vince

Johnny

Amy

Crew

Richard Linklater

Director

Maryse Alberti

Cinematographer

Different:
Horizontal, vertical
difference between relation
and object

SIMILARITY GRAPH CONSTRUCTION

Tape (2001)

Add to My Favorites

Genres - Drama | Sub-Genres - Psychological Drama, Reunion Films

Cast



Ethan Hawke

Vince



Robert Sean Leonard

Johnny



Uma Thurman

Amy

Crew

Richard Linklater

Director

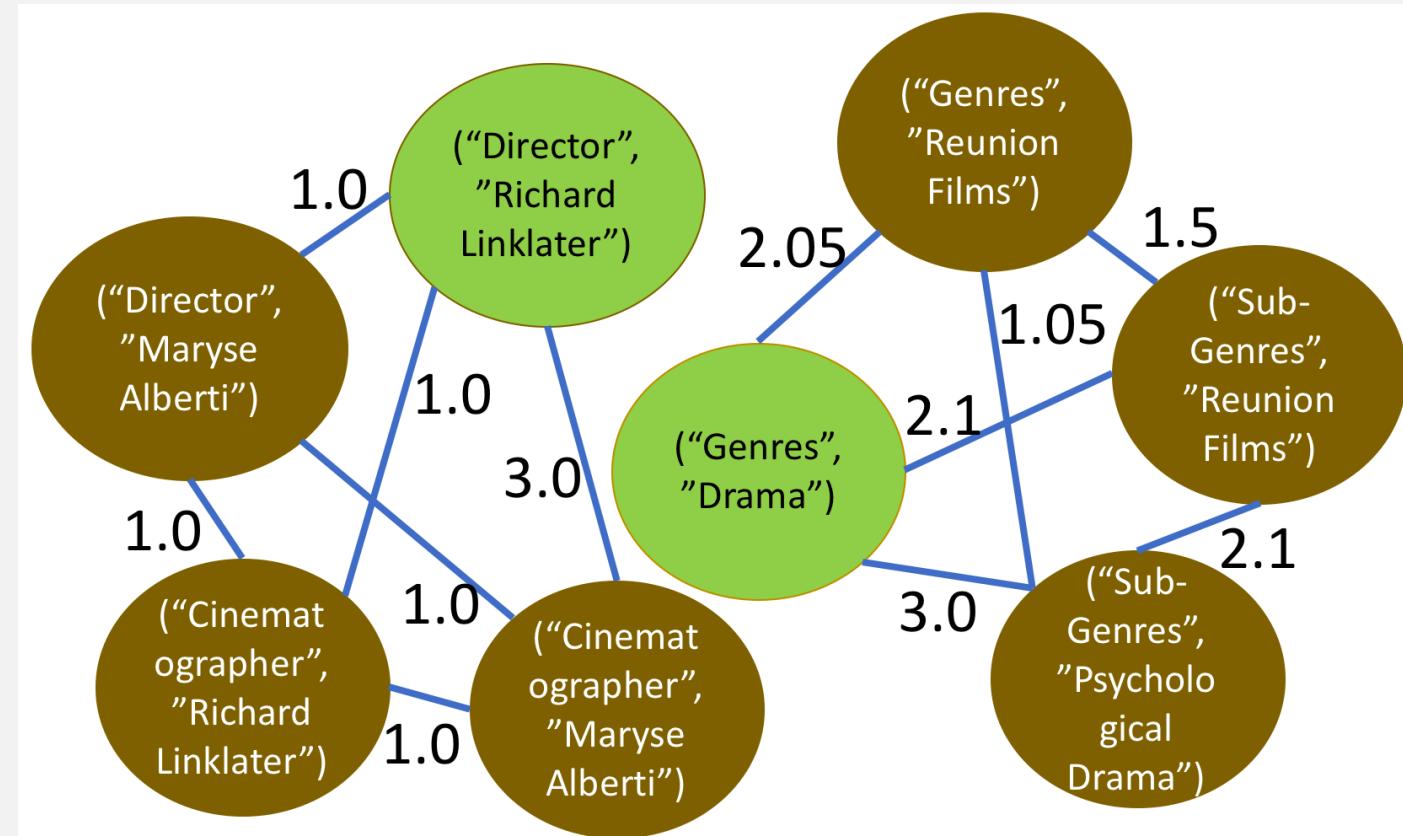
Maryse Alberti

Cinematographer

These pairs will have a higher similarity score

TRAINING DATA CREATION BY LABEL PROPAGATION

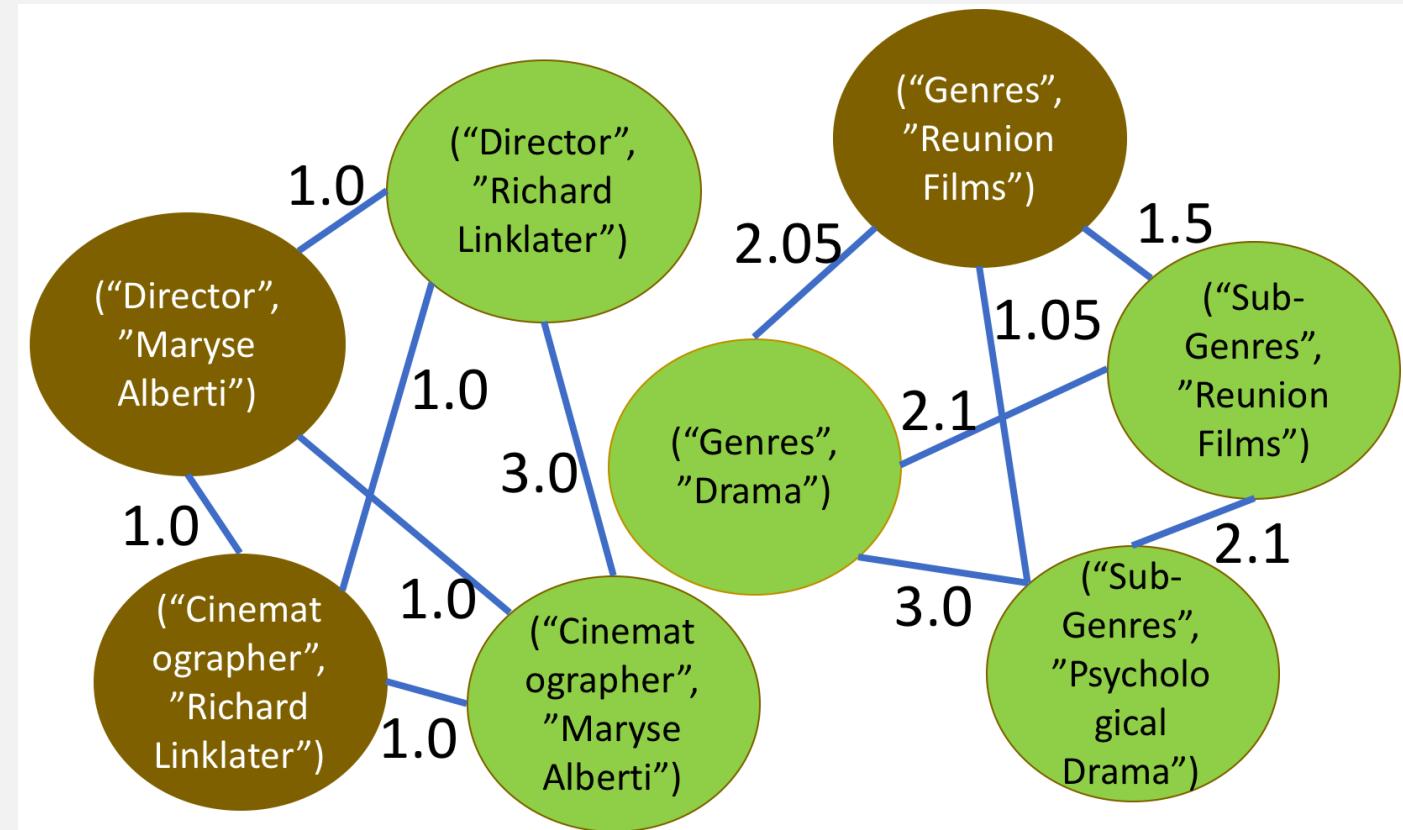
- Multi-Rank Walk algorithm¹
- Uses Personalized PageRank to score similarity to seed nodes



¹ Lin, Frank and William W. Cohen. “Semi-Supervised Classification of Network Data Using Very Few Labels.” *2010 International Conference on Advances in Social Networks Analysis and Mining* (2010): 192-199.

TRAINING DATA CREATION BY LABEL PROPAGATION

- Multi-Rank Walk algorithm¹
- Uses Personalized PageRank to score similarity to seed nodes



¹ Lin, Frank and William W. Cohen. "Semi-Supervised Classification of Network Data Using Very Few Labels." *2010 International Conference on Advances in Social Networks Analysis and Mining* (2010): 192-199.

LEARN CLASSIFIER

- We now have training data for **new relation types for some pages of a website**
- Learn a logistic regression model and apply it to other pages of the site

DATASET AND EXPERIMENTS

EXPANDED SWDE LABEL SET

- Existing SWDE³ dataset labels 3-4 predicates on a set of 80 websites
 - All English language
- We created OpenIE labels for 21 of these sites
- 5 – 272 relation types per website
 - ~800,000 labels
- Released at homes.cs.washington.edu/~lockardc/expanded_swde.html

³ Hao, Qiang, Rui Cai, Yanwei Pang and Lei Zhang. “From one tree to a forest: a unified solution for structured web data extraction.” *SIGIR* (2011).

EXPERIMENTS

System	Movie		NBA		University	
	P	R	P	R	P	R
WEIR (Bronzi et al., 2013)	0.23	0.17	0.08	0.17	0.13	0.18
Colon Baseline	0.63	0.21	0.51	0.33	0.46	0.31
OpenCeres	0.77	0.68	0.74	0.48	0.65	0.29

Fairly high precision

EXAMPLES OF DISCOVERED RELATION TYPES

Movie

Seed: Director, Writer, Producer, Actor, Release

New: Country, Filmed In, Language, MPAA Rating, Set In, Reviewed By, Studio, Metascore, Box Office, Distributor, Tagline, Budget, Sound Mix

4x increase in
relation types in
Movie/NBA

NBA Player

Seed: Height, Weight, Team

New: Birth Date, Birth Place, Salary, Age, Experience

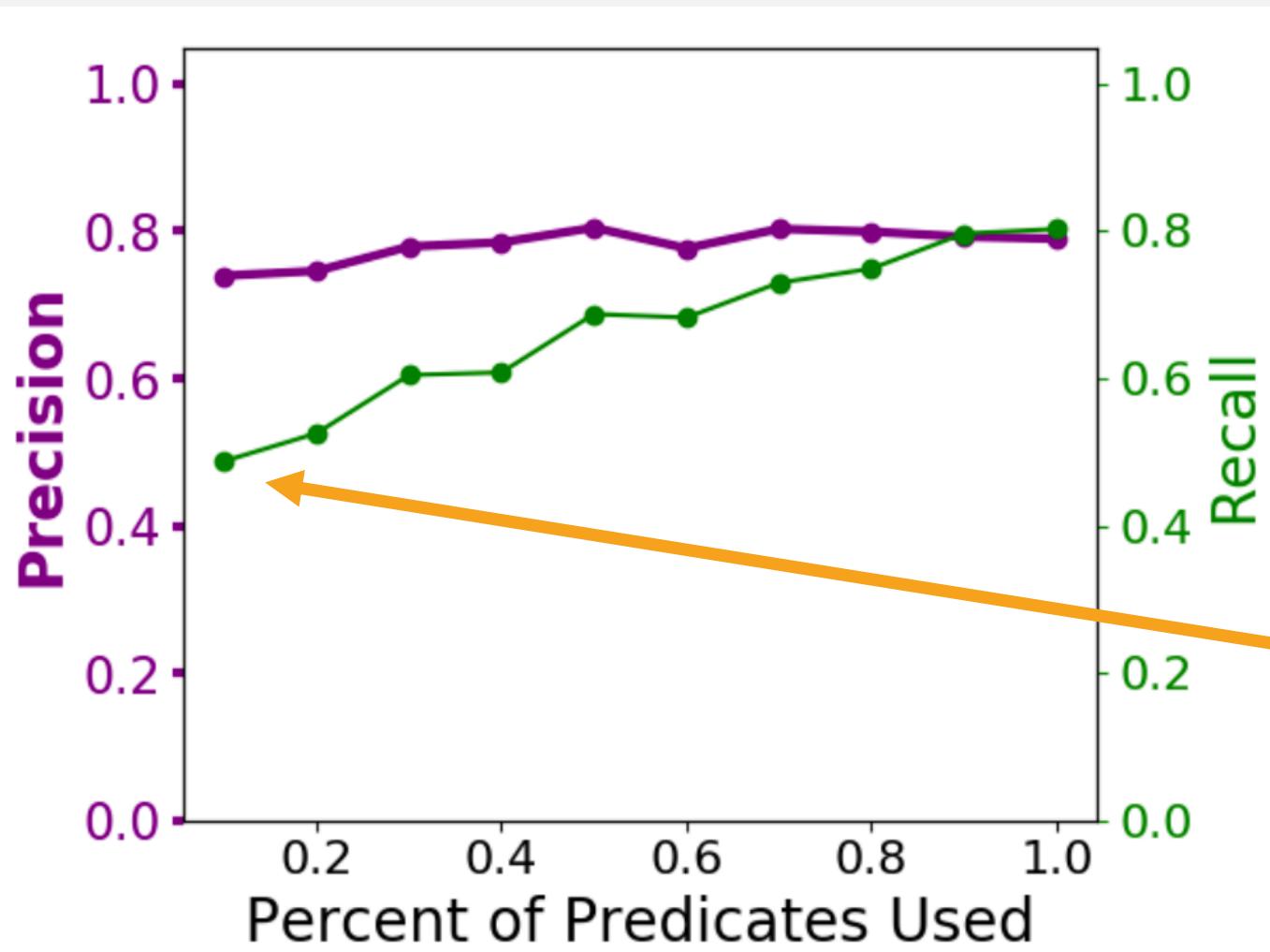
23x increase in
relation types in
University!

University

Seed: Phone Number, Web address, Type (public/private)

New: Calendar System, Enrollment, Highest Degree, Local Area, Student Services, President

VARYING RELATION TYPES IN SEED



Able to reach 50% recall when only 10% of predicates are in seed set!

FALSE NEGATIVES

Santa Sangre details

Overall Rating 

Overview Details Cast & C

Movie Details:

Director: [Alejandro Jodorowsky](#)

Produced By: Produzioni Intersound

Year: 1989

Run Time: 123 minutes

Country: Italy, Mexico

Language: English, Spanish

Alternate Titles: Holy Blood

Key Cast: [Axel Jodorowsky](#), [Blanca Guerra](#), [Sabrina Dennison](#), [Adan Jodorowsky](#), [Guy Stockwell](#), [Thelma Tixou](#), [Faviola Tapia](#), [Jesus Juarez](#), [Full Credits](#)

Hard to find this section
since format is different
from rest of page

Similar Movies

[Psycho \(1960\)](#)

[The Shining \(1980\)](#)

[The Fly \(1986\)](#)

[Shivers \(1975\)](#)

[El Topo \(1971\)](#)

[Danzon \(1991\)](#)

[Suspiria \(1977\)](#)

[Providence \(1977\)](#)

[Eyes Without a Face \(1960\)](#)

[Tenebre \(1982\)](#)

FALSE POSITIVES

Noble Lee Lester

Best known as a **Supporting Actor** based on a credit in that role in 1 film, with \$15,432,314 worldwide aggregate box office (rank #42,574)

Best-Known Acting Roles: Media Jackel ([The Bonfire of the Vanities](#))

Most productive collaborators: [Tom Hanks](#), [Brian De Palma](#), [Bruce Willis](#), [Michael Christofer](#), [Melanie Griffith](#)

("Supporting Actor", "Tom Hanks")

"Supporting Actor" has similar format as true relation types

CONCLUSIONS

- Semi-structured OpenIE can yield millions of facts for QA, KB Completion, and more.
- OpenCeres can discover lots of new relation types.



Use our dataset to build/evaluate even better systems!

homes.cs.washington.edu/~lockardc/expanded_swde.html