OpenCeres: When Open Information Extraction Meets the Semi-Structured Web

Colin Lockard, Prashant Shiralkar, Xin Luna Dong

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.

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)

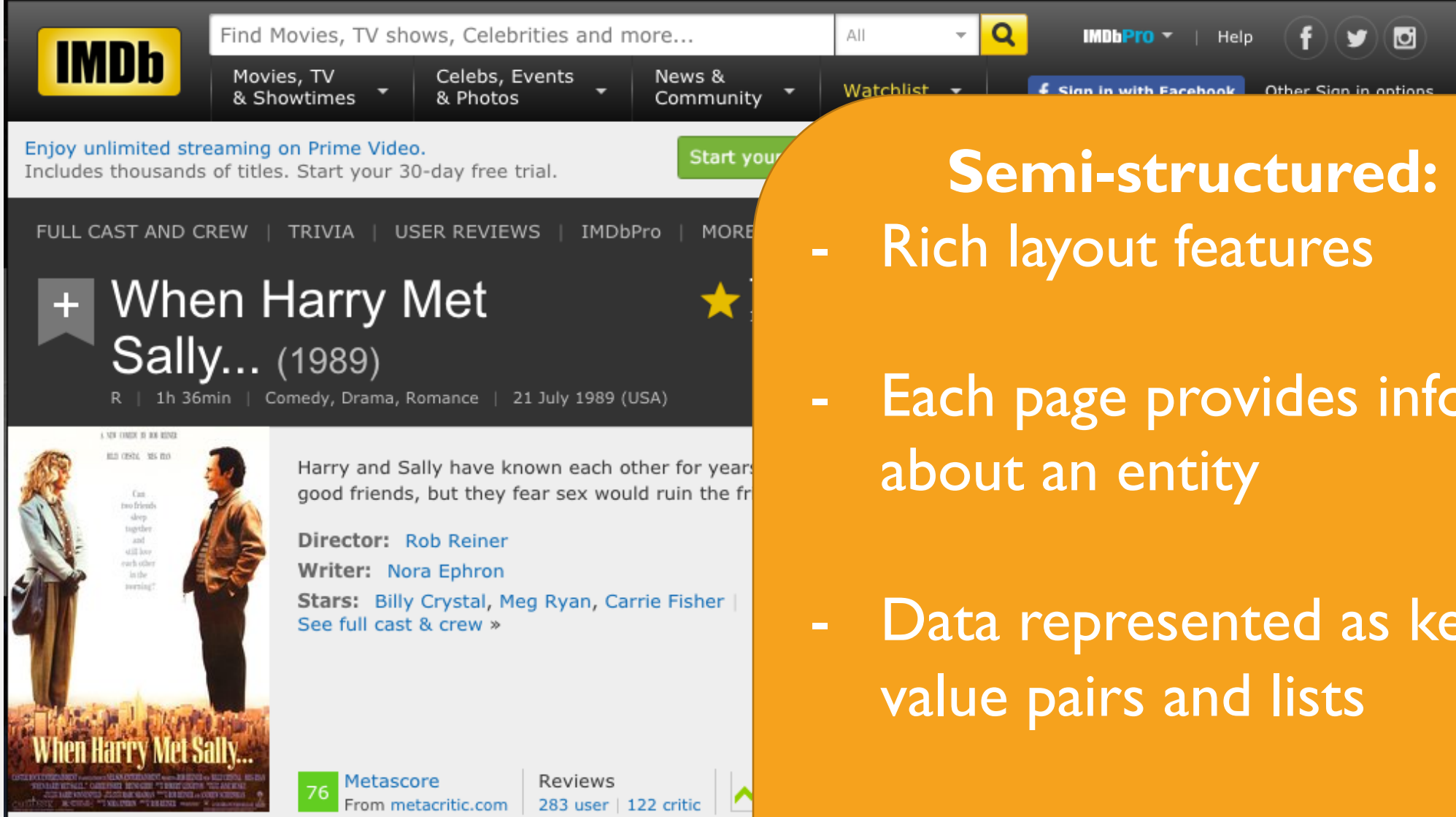
Watch Now
From $2.99 (SD) on Amazon Video
ON DISC

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

Photos

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.



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

FULL CAST AND CREW | TRIVIA | USER REVIEW

When Harry Met S

BollywoodMD

HOME MOVIES

Home › Movies › Jab Harry

Jab Harry M

2 hrs 24 mins | Comedy,

jab Harry met

4th august, 2017

CAST & CREW

NMDb

Daniel Ade

Year of produ

Running Time

Written by: D

Produced by:

Directed by: D

Starring: Stell

ALLMUSIC

New Releases Discover A

The most p

Read Full B

Overvi

Player Profile

Album Highlights

The Beatles
Abbey Road

hifiengine

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

User Login

Login/Register

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*

Manual Library / A and D

A And D DA-U7000

Digital Integrated Amplifier (1988)

add a review

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

Stream or buy on:

amazon

iTunes

2017 Season

RAT	YDS
104.1	1,36

Career

104.1	38,1

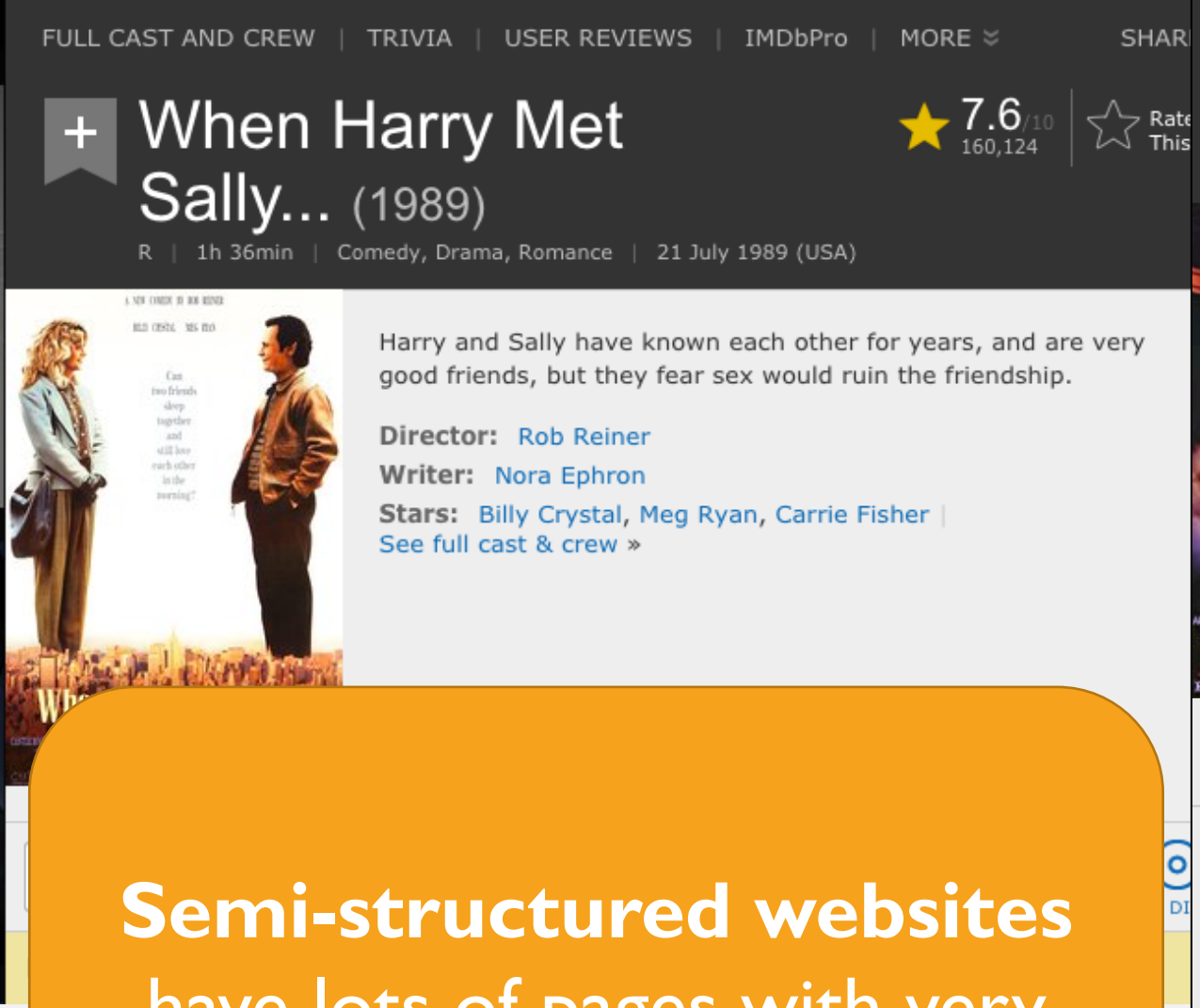
Go to

Aaron Rodgers

elic/Garage
Roll

ck

Harrison
non
Cartney
arr
t
utcliffe



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

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

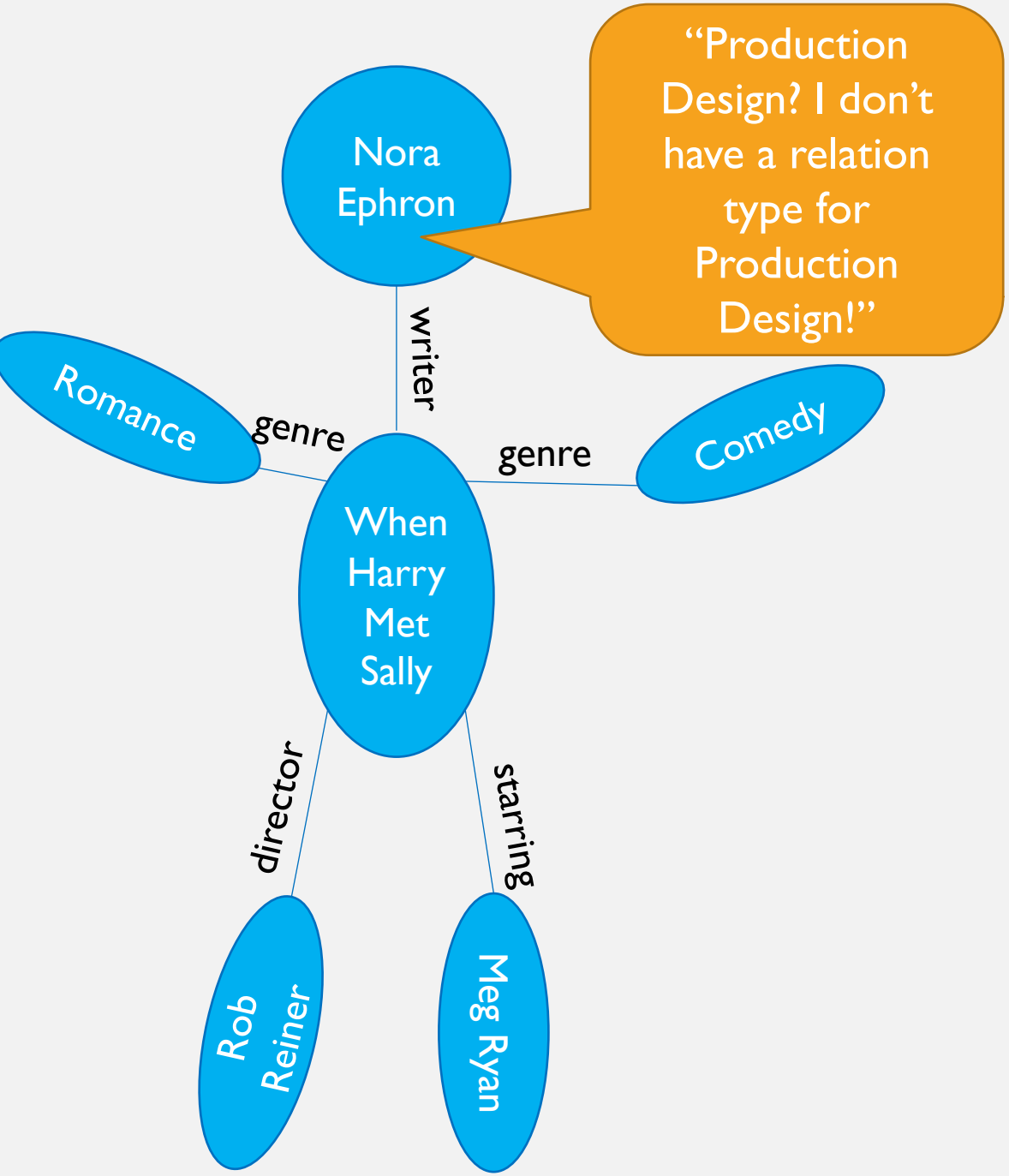
Robert Leighton

Casting By

Janet Hirshenson
Jane Jenkins

Production Design by

Jane Musky



Produced by

Nora E
Steve
Rob Re
Andrew
Jeffrey Stott ... co-producer

They might have TOO MUCH information

Cinematography by

Barry Sonnenfeld ... director of photography

Film Editing by

Robert Leighton

Casting By

Janet Hirshenson
Jane Jenkins

Production Design by

Jane Musky

TRADITIONAL EXTRACTION

TOPIC ENTITY

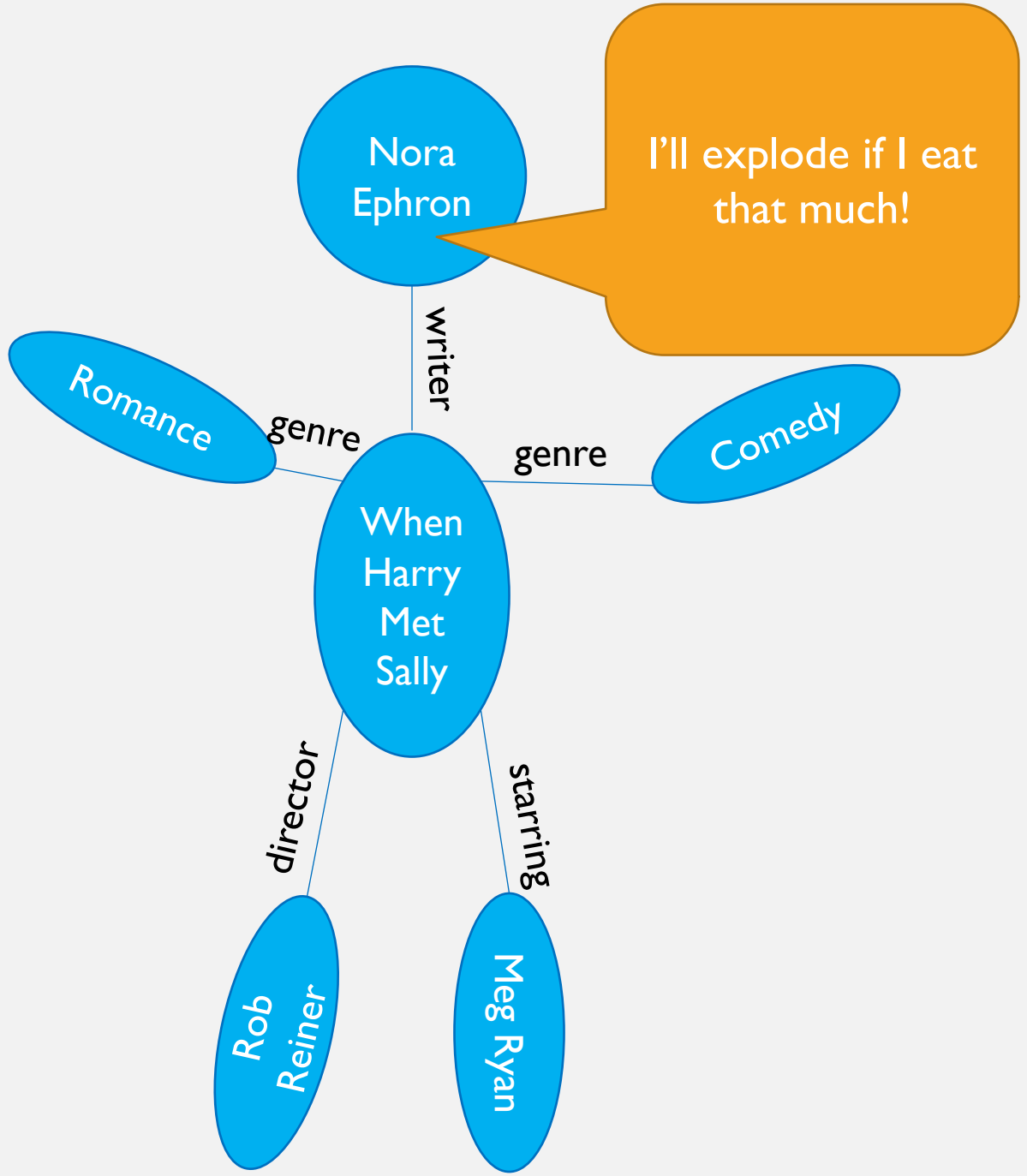
"When Harry Met Sally"

Relation types defined in ontology

(film.written_by, "Nora Ephron")

(film.actor, "Meg Ryan")

On10 semi-structured movie websites, the IMDb ontology covered only **7%** of relations.



OPEN INFORMATION EXTRACTION[1]

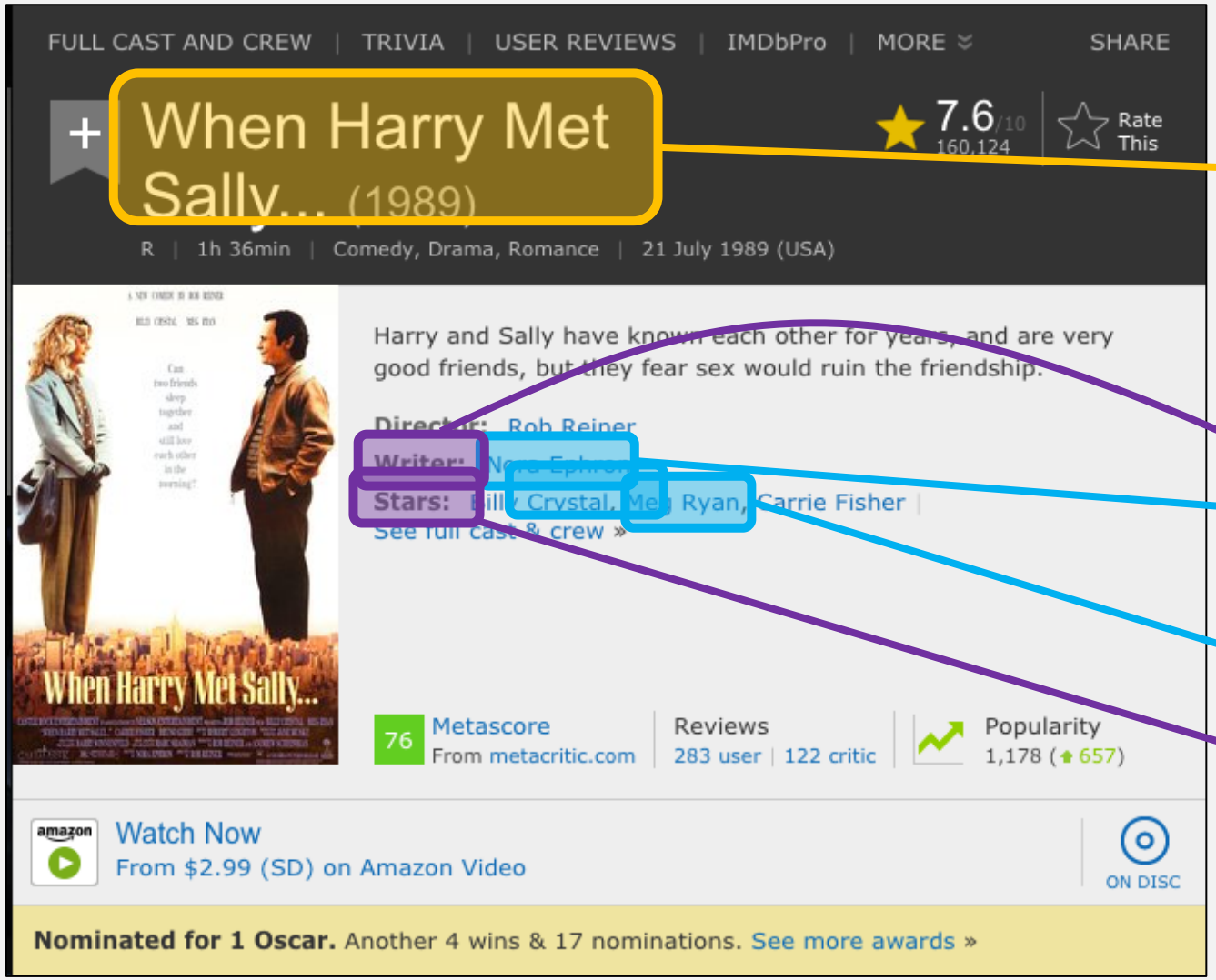
- Instead of relying on ontology for relation type, just extract relation string from page
- Widely explored in unstructured text
- Not explored on semi-structured websites

[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



TOPIC ENTITY

"When Harry Met Sally"

("Writer:", "Nora Ephron")

("Stars:","Meg Ryan")

OpenIE triples can be used for:

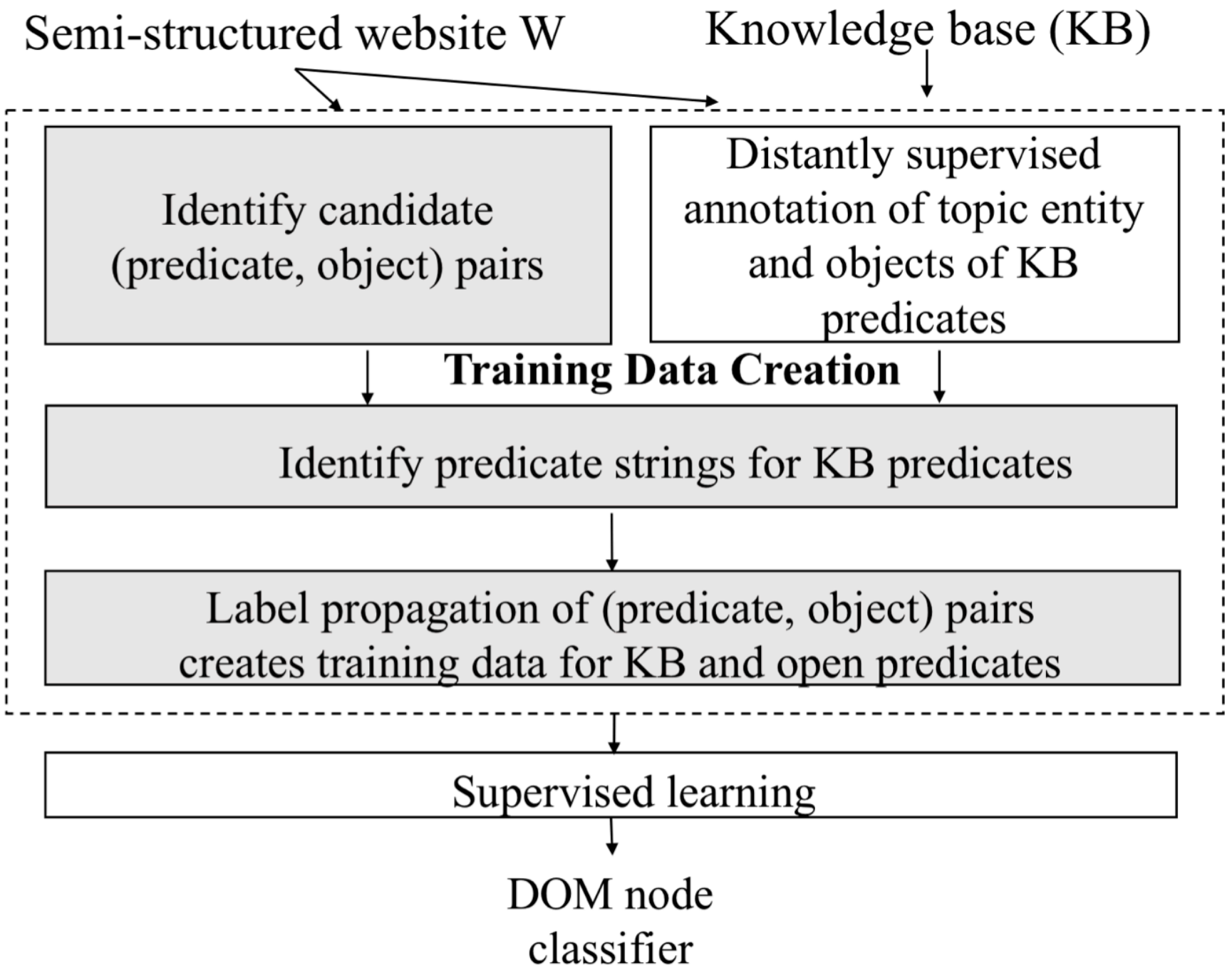
- KB Completion
  - See our paper "OpenKI: Integrating Open Information Extraction and Knowledge Bases with Relation Inference" by Dongxu Zhang et al!
- Question Answering
- Better KG Embeddings
- Ontology Discovery
- Fact Checking

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

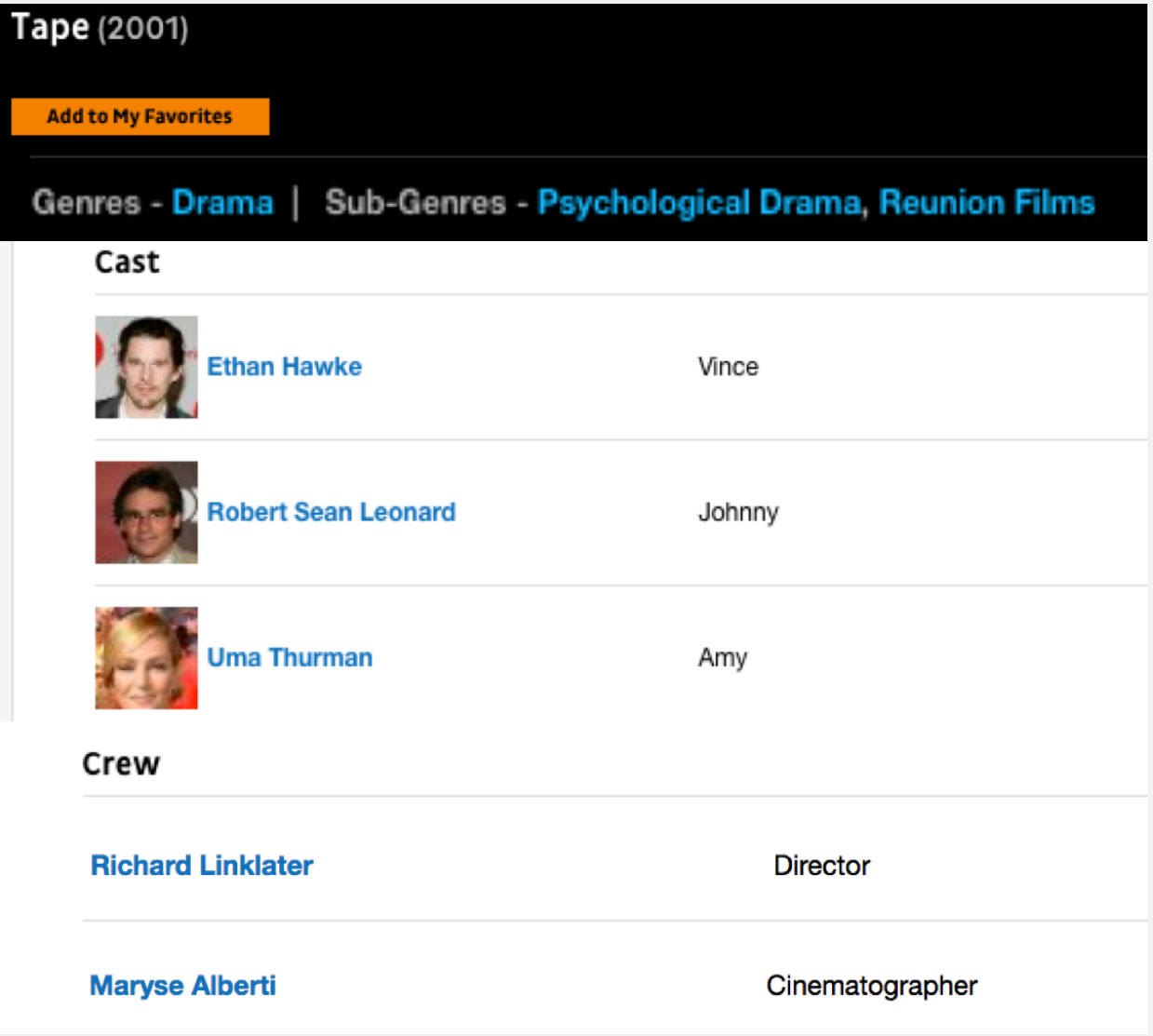
WHAT ARE THE CHALLENGES?

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)
Identify candidate (predicate, object) pairs
Distantly supervised annotation of topic entity and objects of KB predicates
Training Data Creation
Identify predicate strings for KB predicates
Label propagation of (predicate, object) pairs creates training data for KB and open predicates
Supervised learning
DOM node classifier

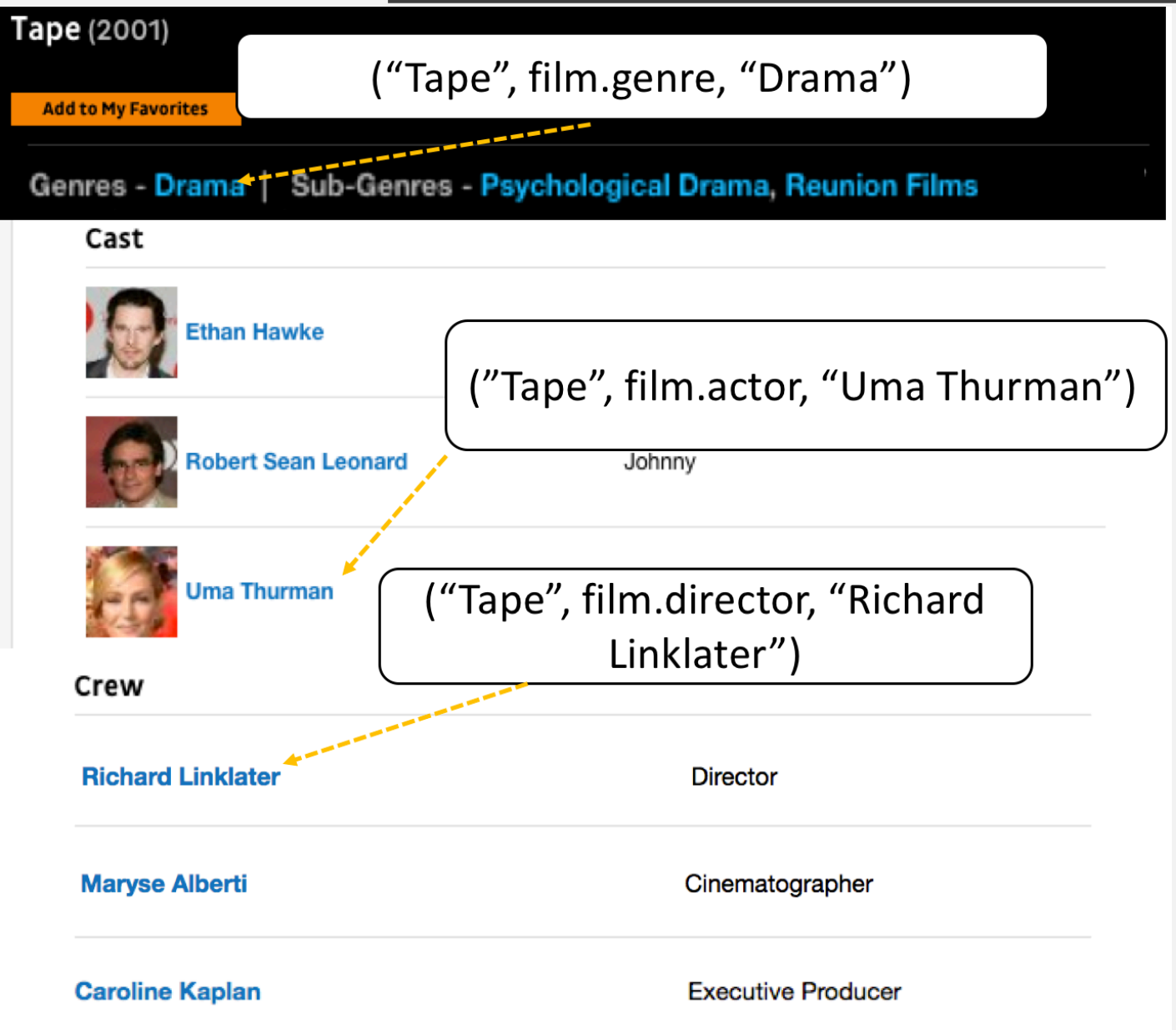
AUTOMATIC SEED DATA CREATION



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

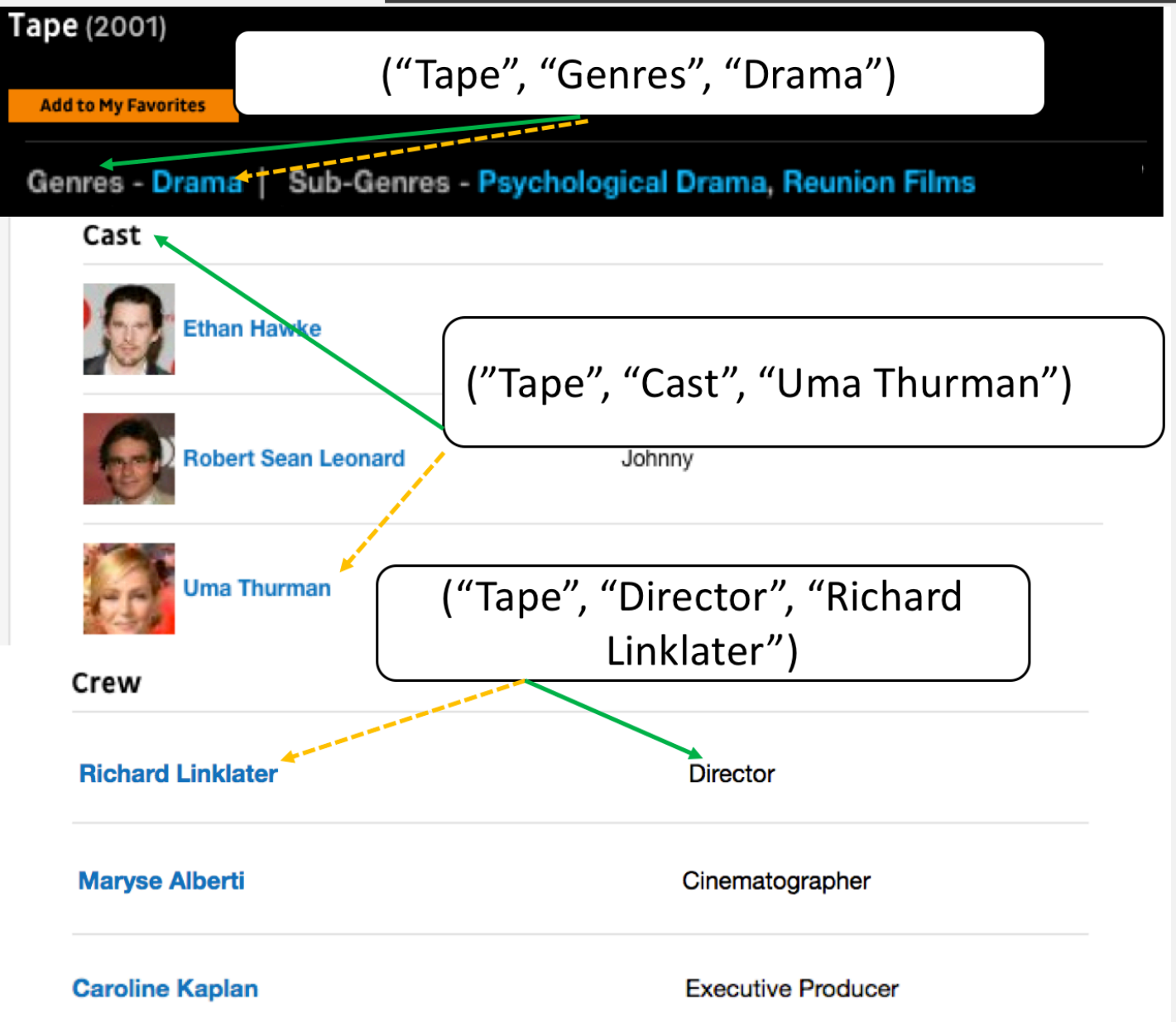


Step 2:

- Distantly supervised annotation of facts in existing knowledge base[2]

[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.*

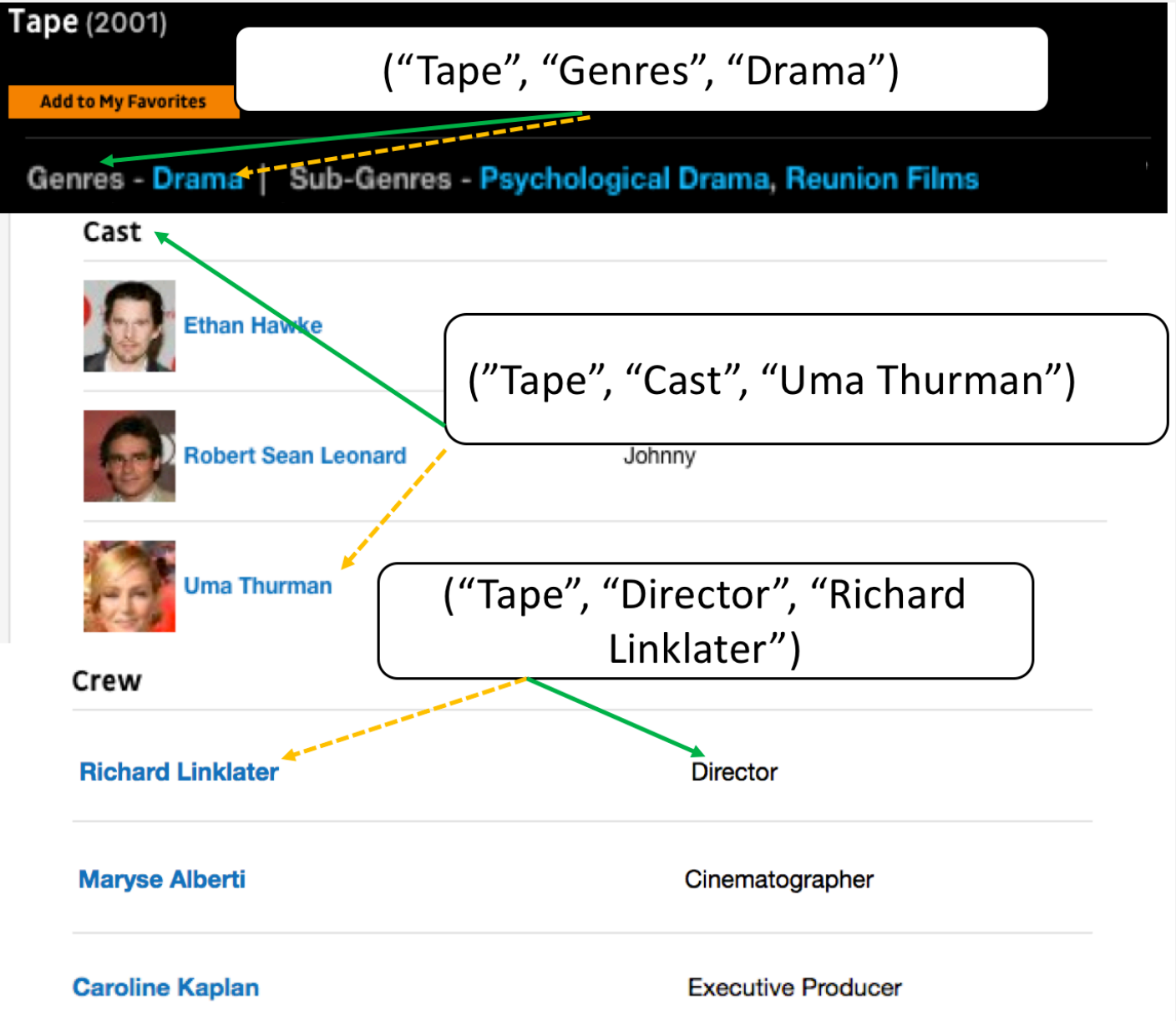
AUTOMATIC SEED DATA CREATION



Step 3:

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

AUTOMATIC SEED DATA CREATION



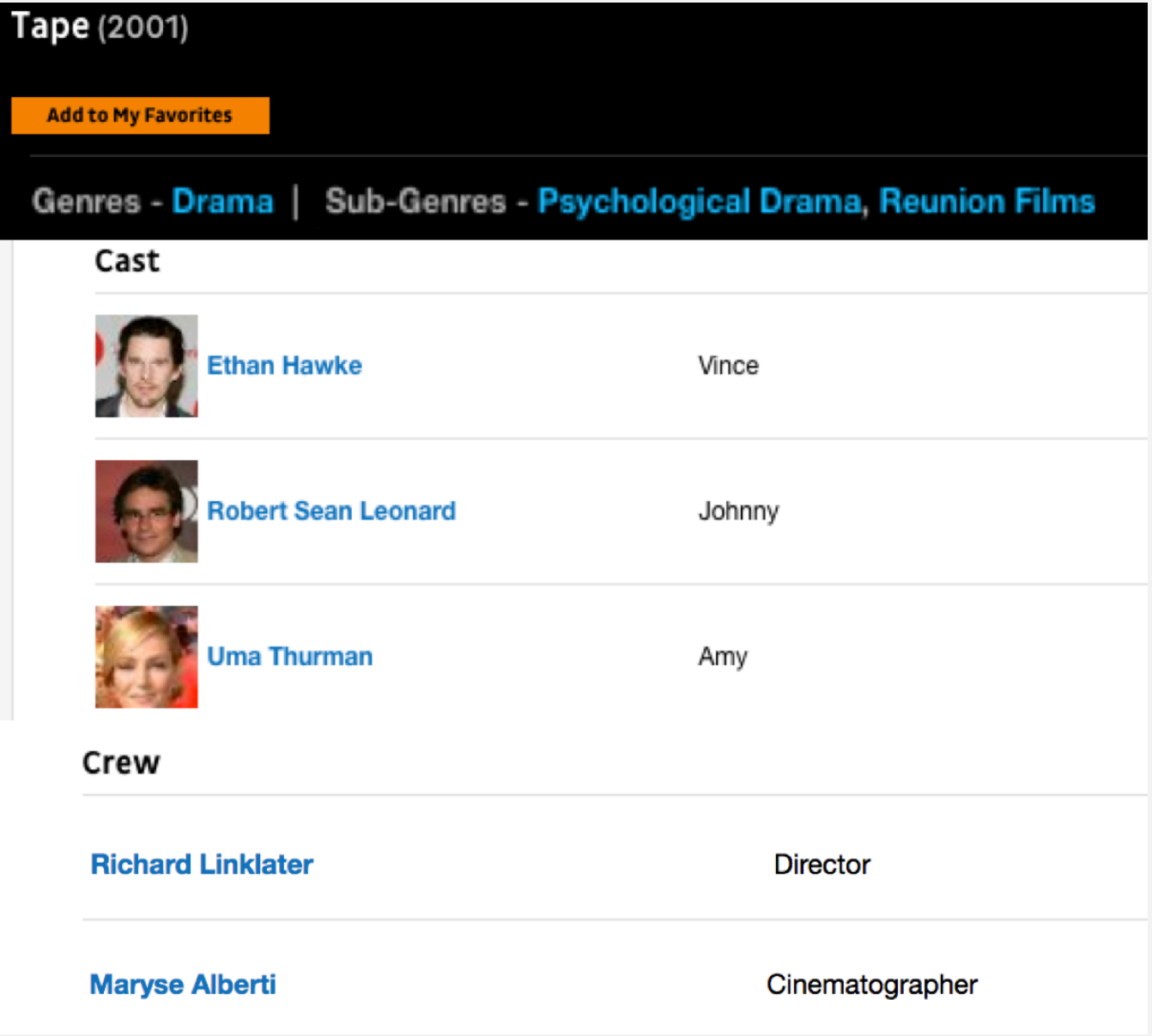
Step 4:

- 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

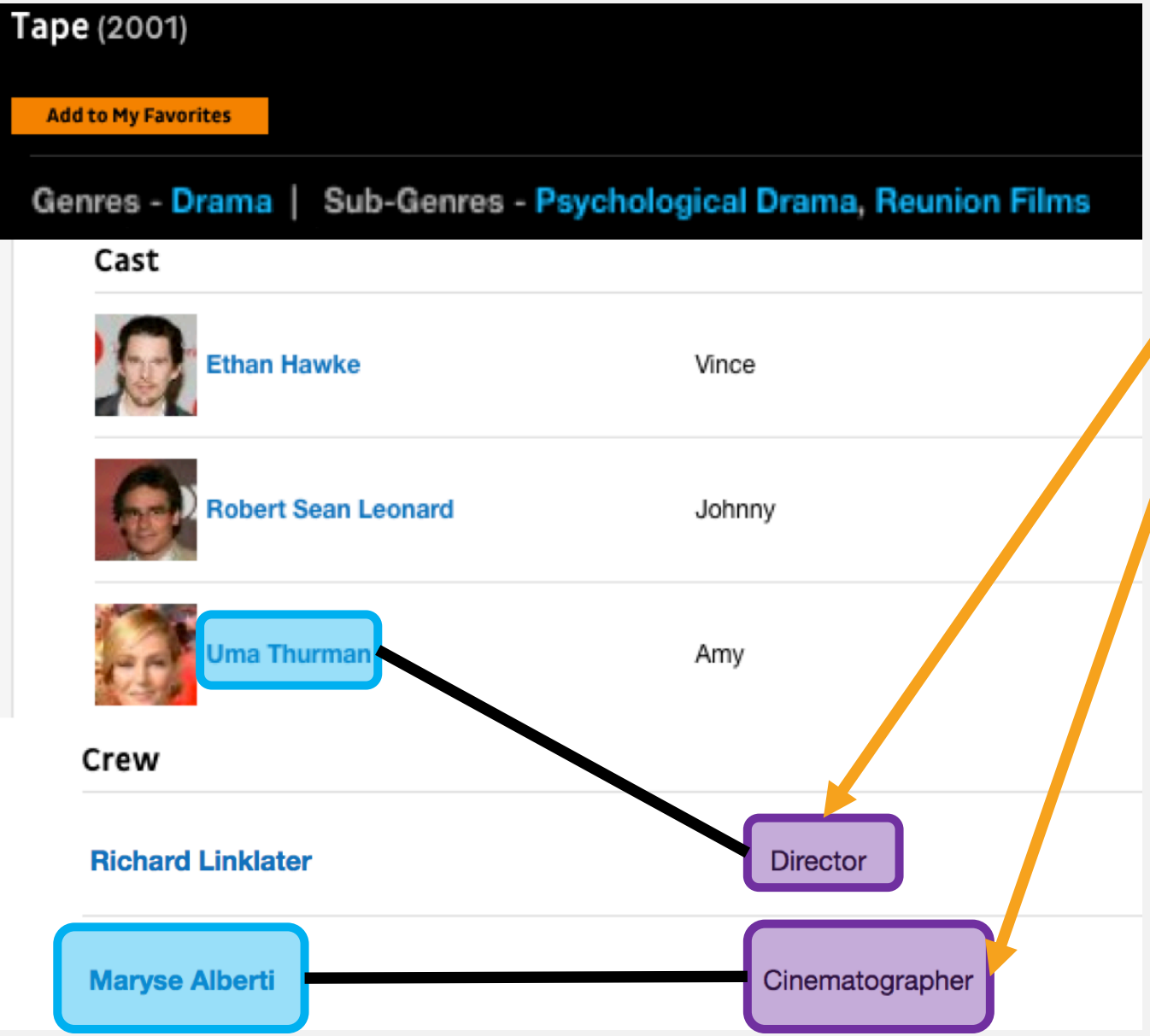
SIMILARITY GRAPH CONSTRUCTION



Candidate pairs become nodes in graph.

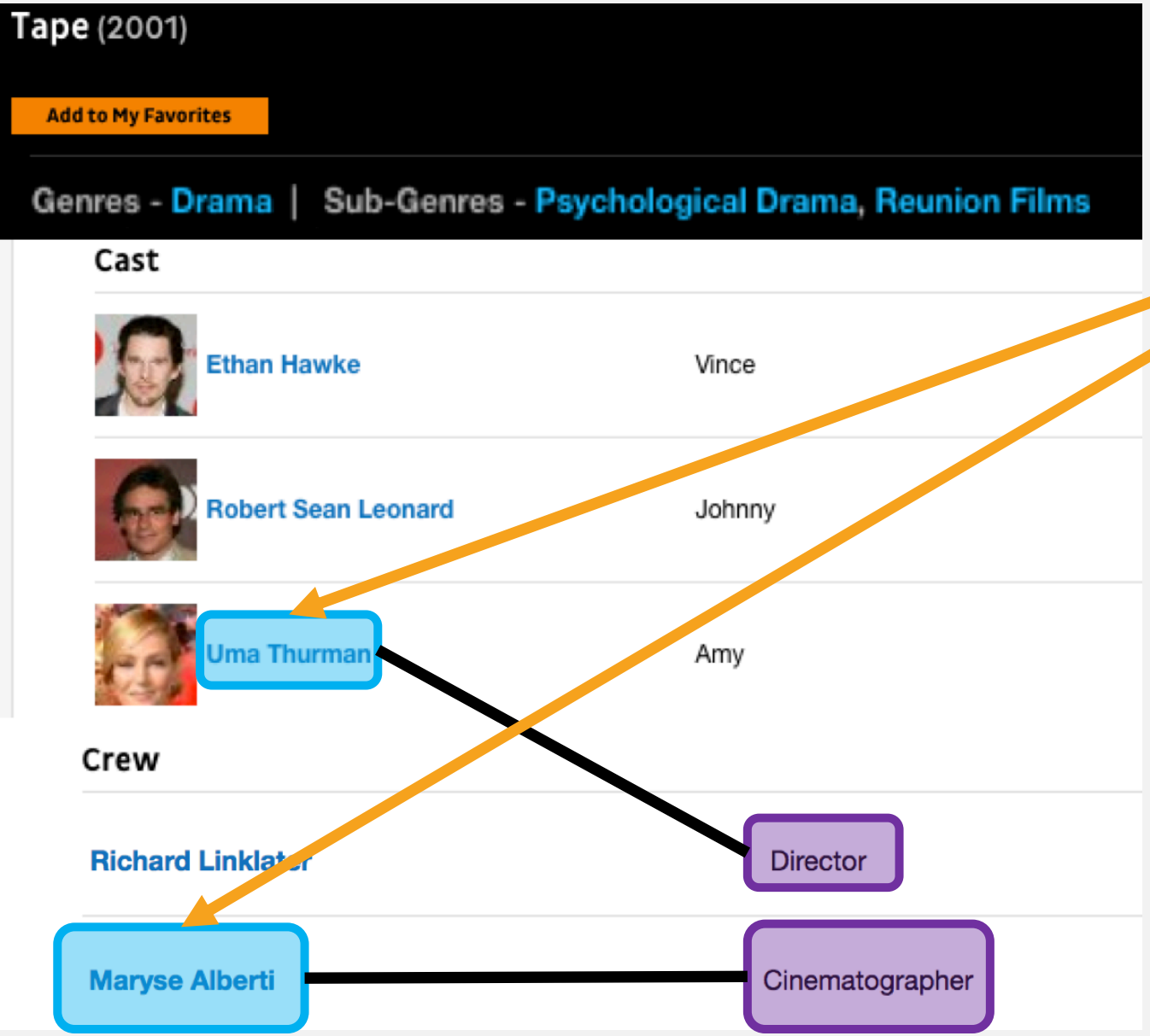
Edges weighted by similarity metric based on visual and layout features.

SIMILARITY GRAPH CONSTRUCTION



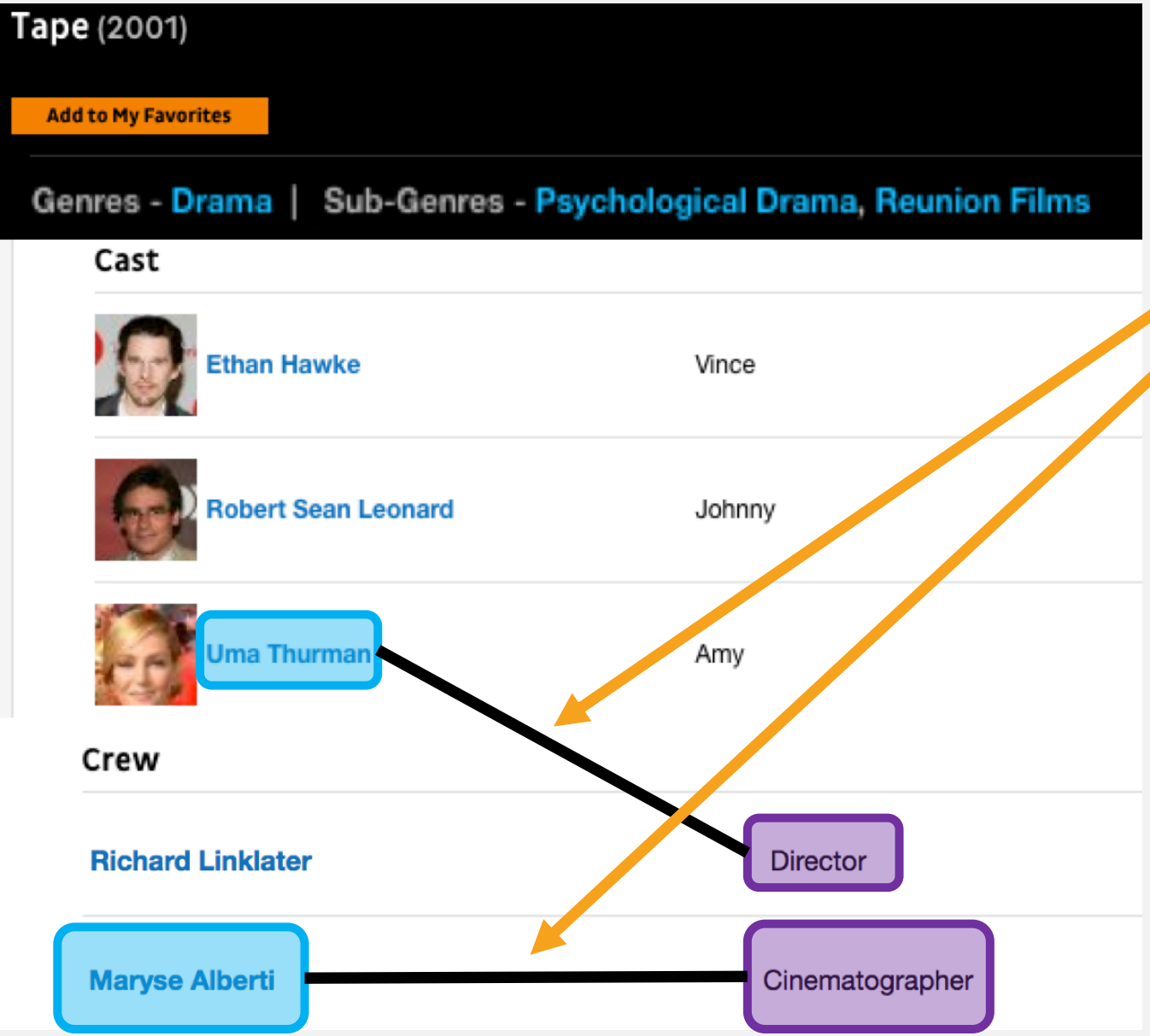
Similar:
Relations have same font and size and are vertically aligned

SIMILARITY GRAPH CONSTRUCTION



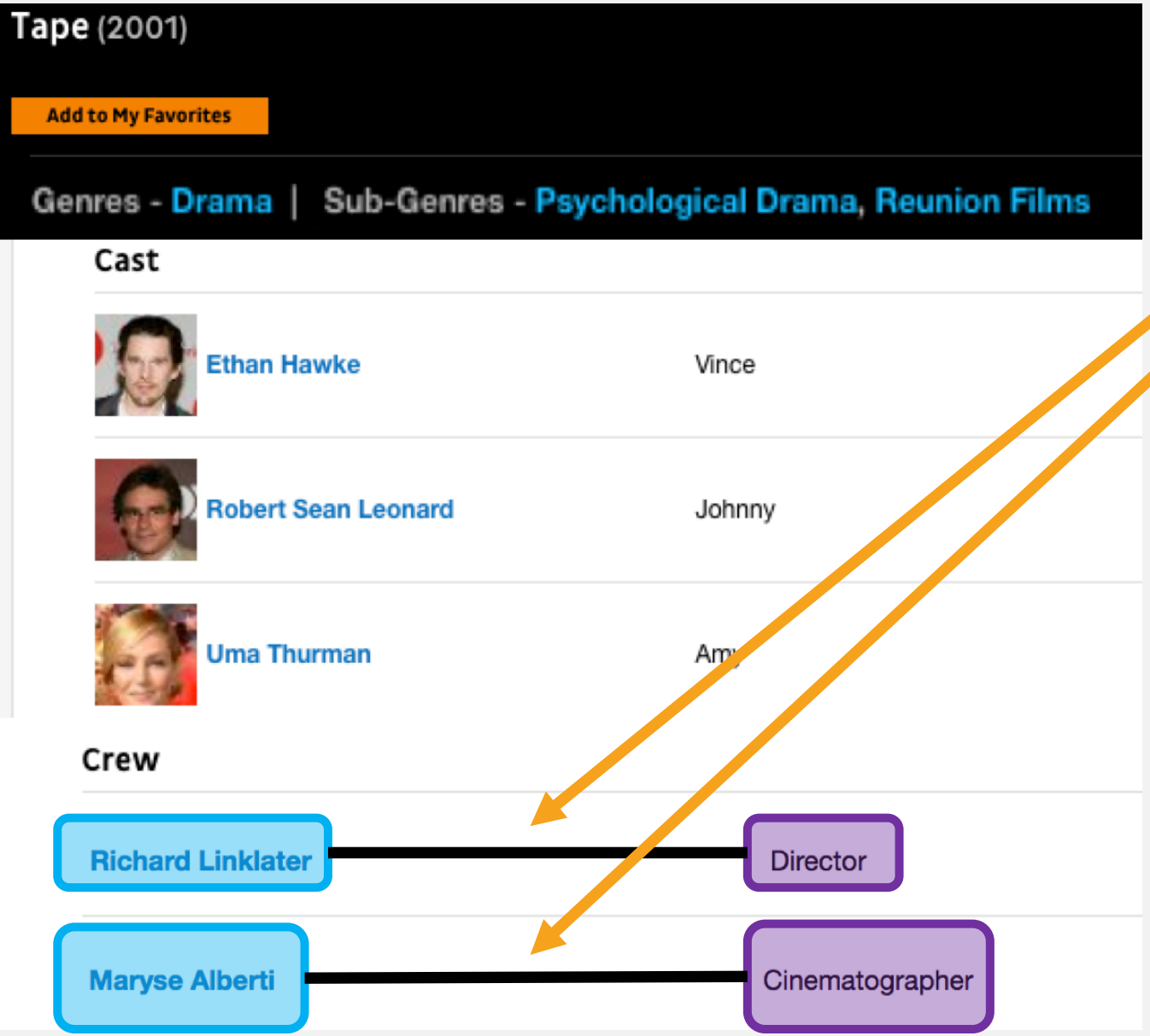
Similar:
Objects have same font and size

SIMILARITY GRAPH CONSTRUCTION



Different:
Horizontal, vertical difference between relation and object

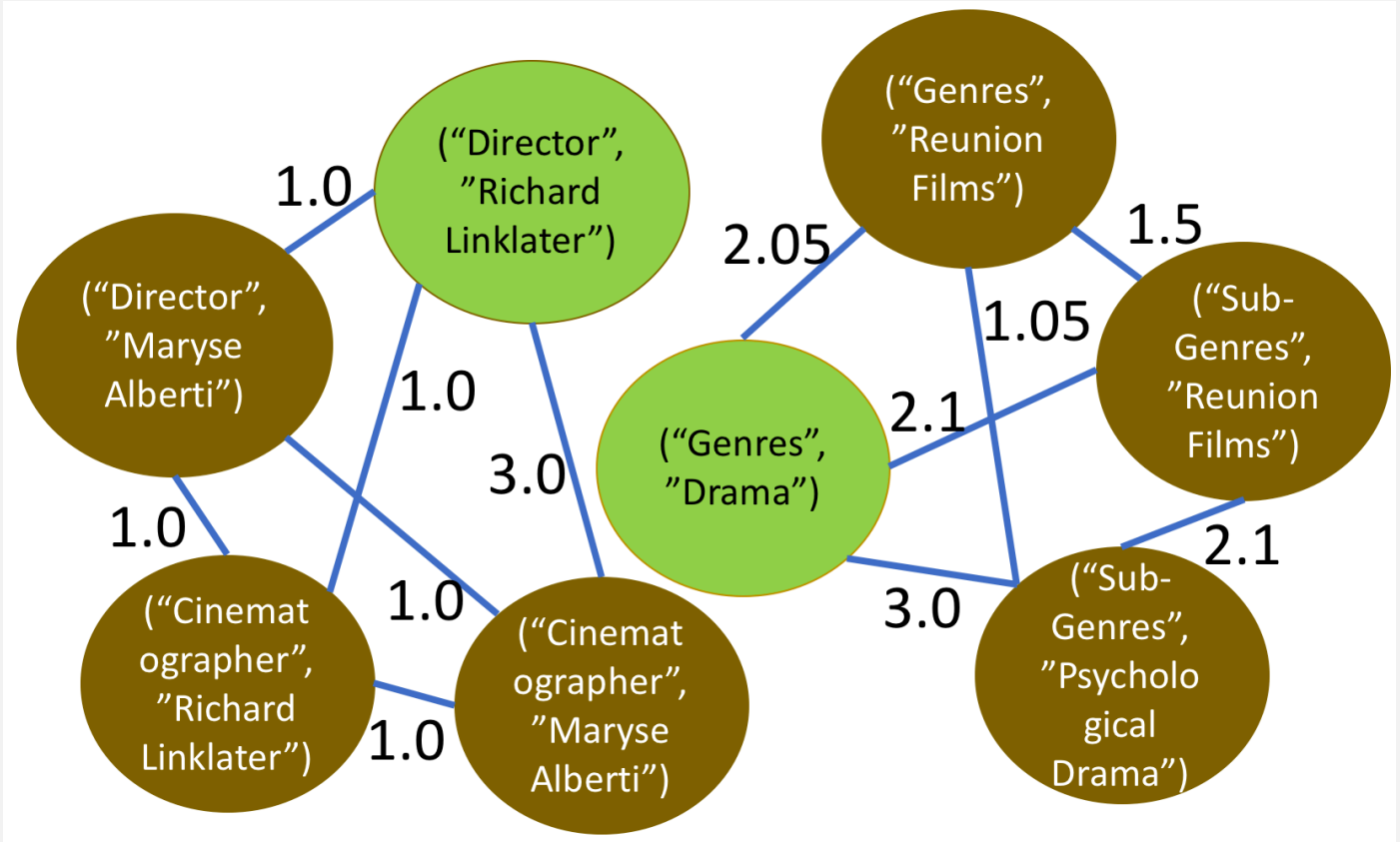
SIMILARITY GRAPH CONSTRUCTION



These pairs will have a higher similarity score

TRAINING DATA CREATION BY LABEL PROPAGATION

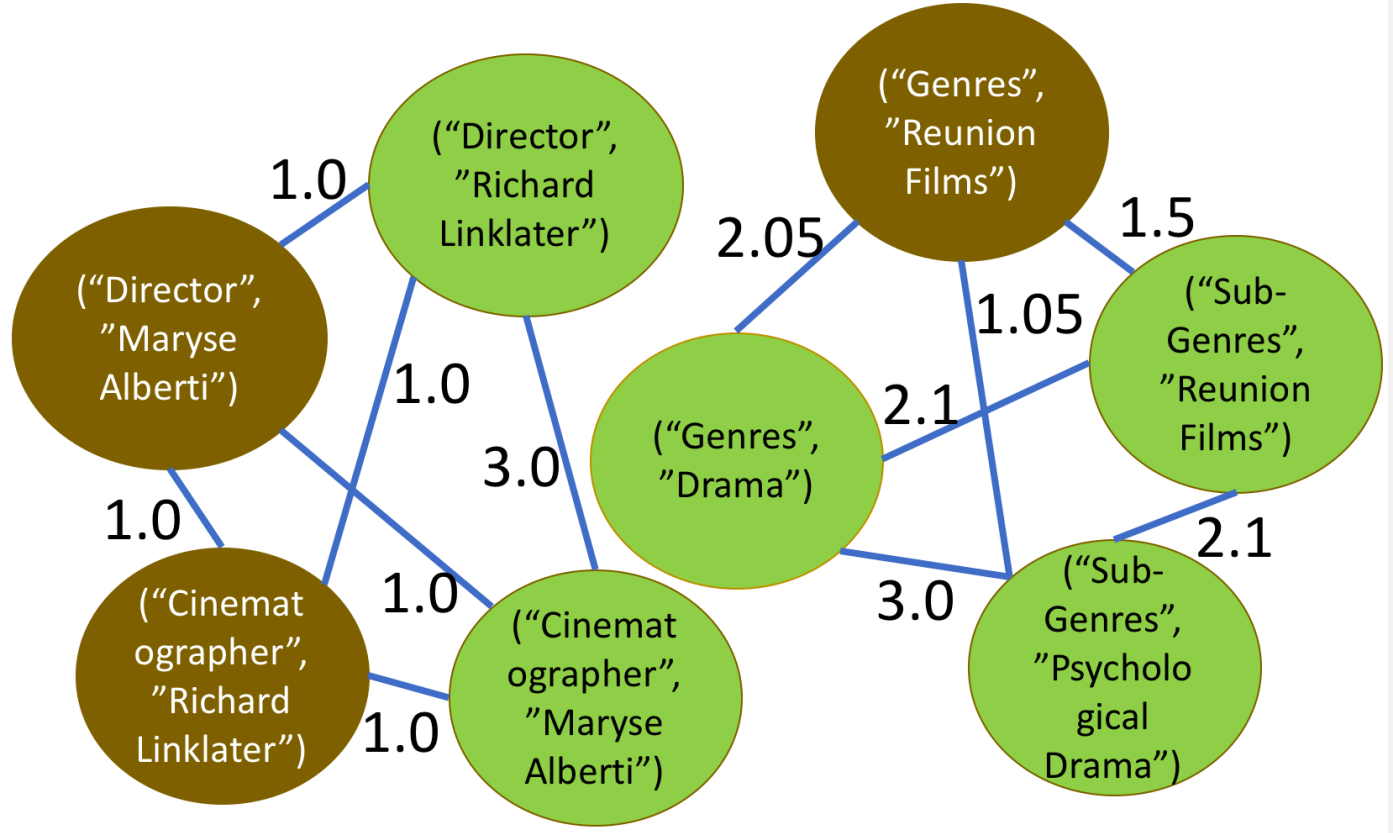
- Multi-Rank Walk algorithm[1]
- Uses Personalized PageRank to score similarity to seed nodes



[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 algorithm[1]
- 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 SWDE[3] 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, 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, Exper

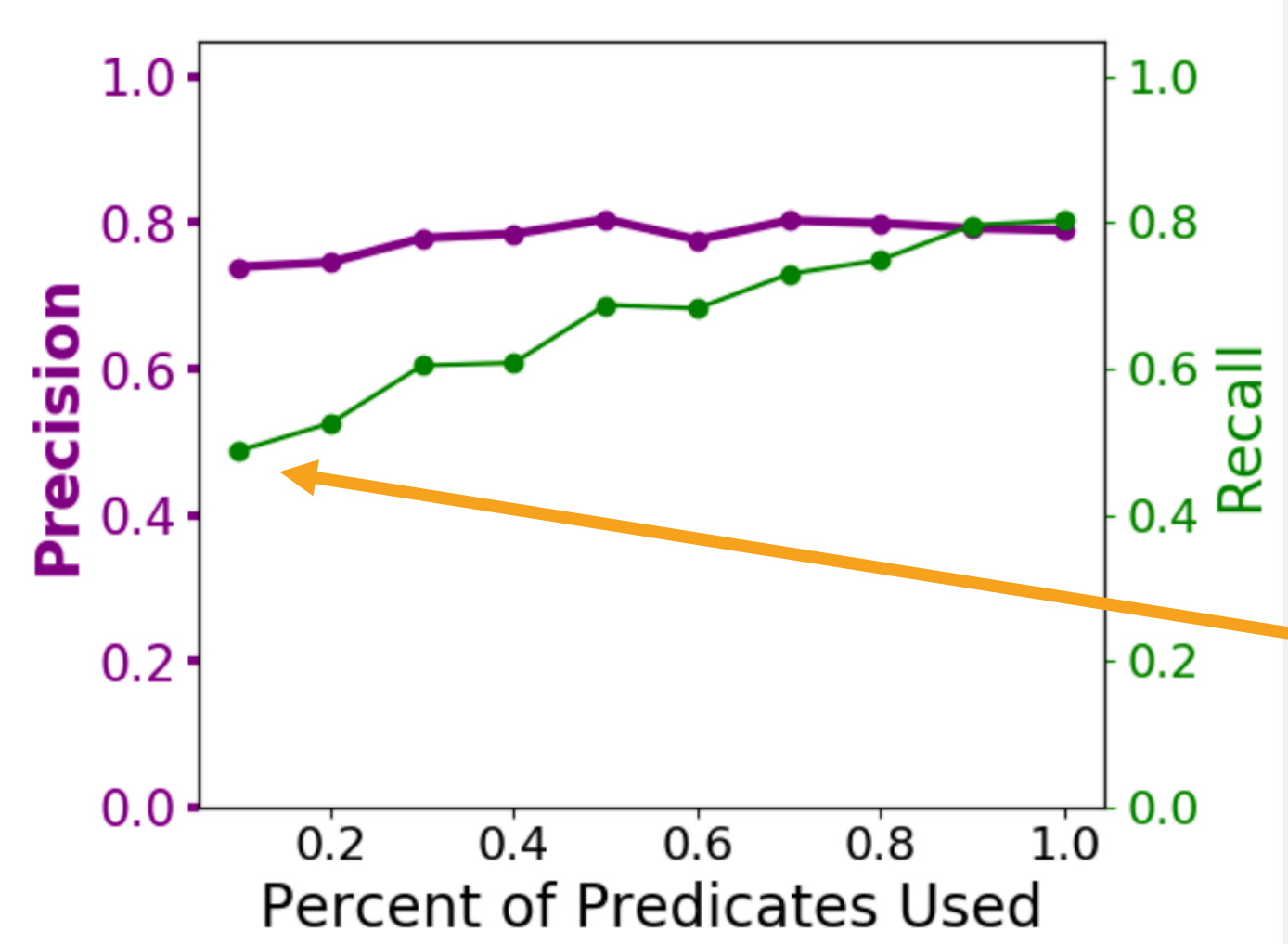
23x increase in relation types in University!

University

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

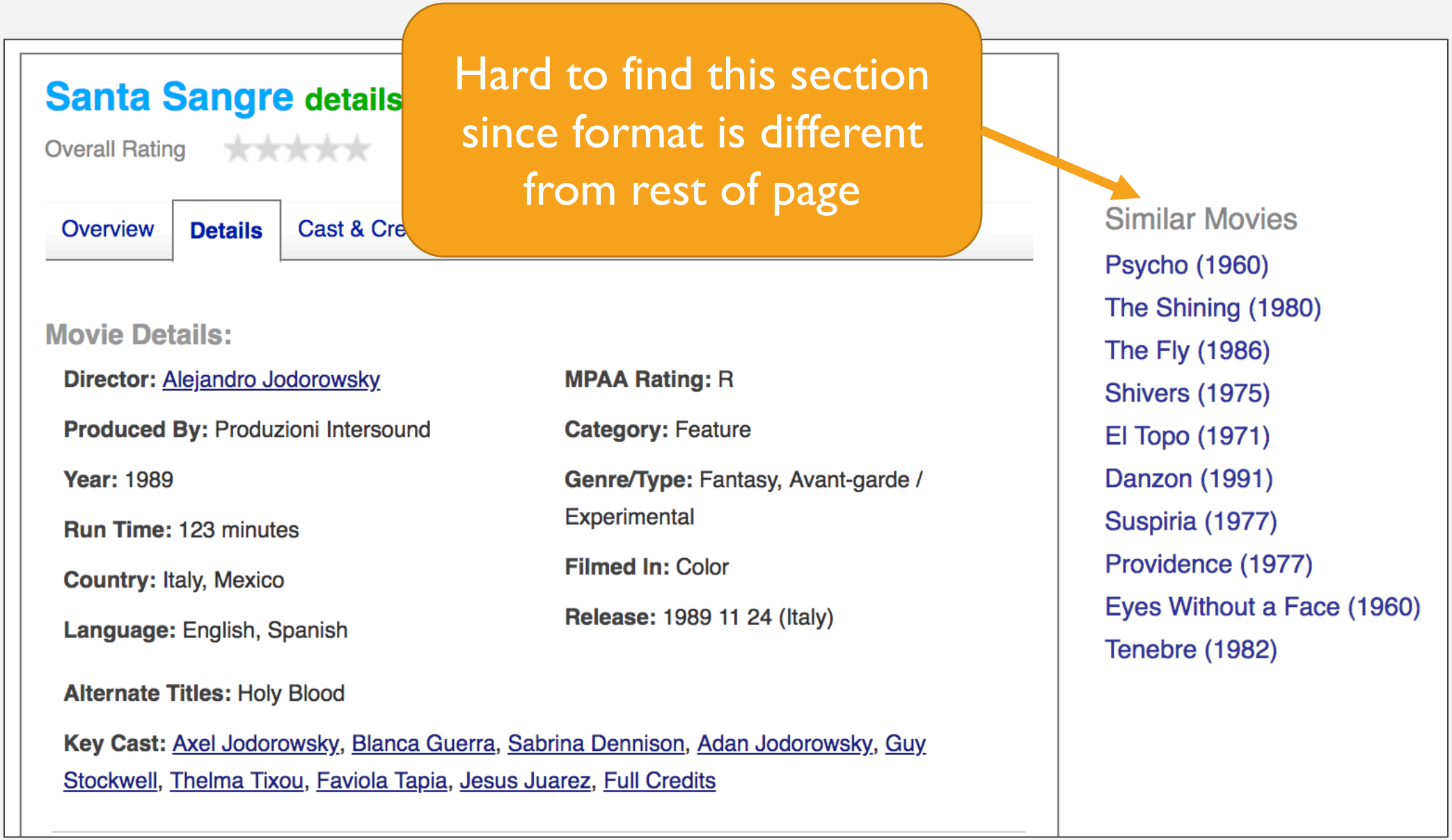
New: Calendar System, Enrollment, Highest Degree, Local Area, Student Services, President

VARYING RELATION TYPES IN SEED



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

FALSE NEGATIVES



FALSE POSITIVES

("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