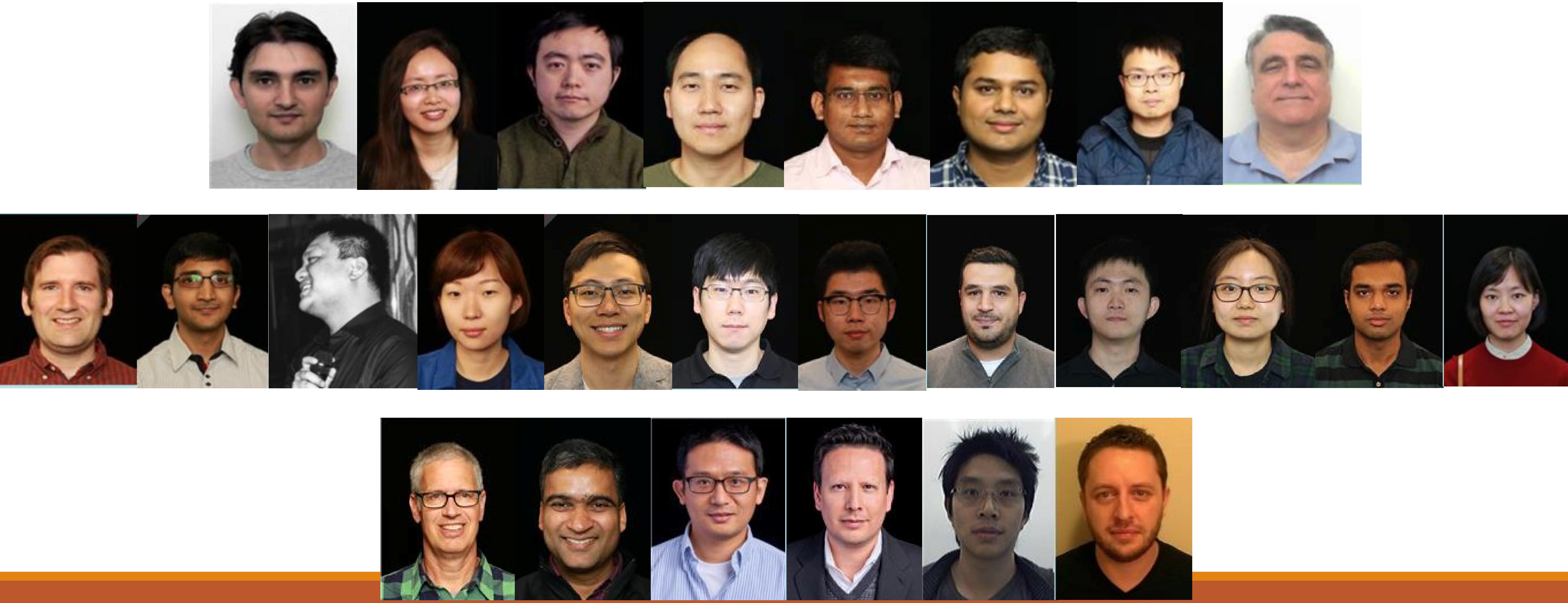
Building a Broad Knowledge Graph for Products

XIN LUNA DONG, AMAZON

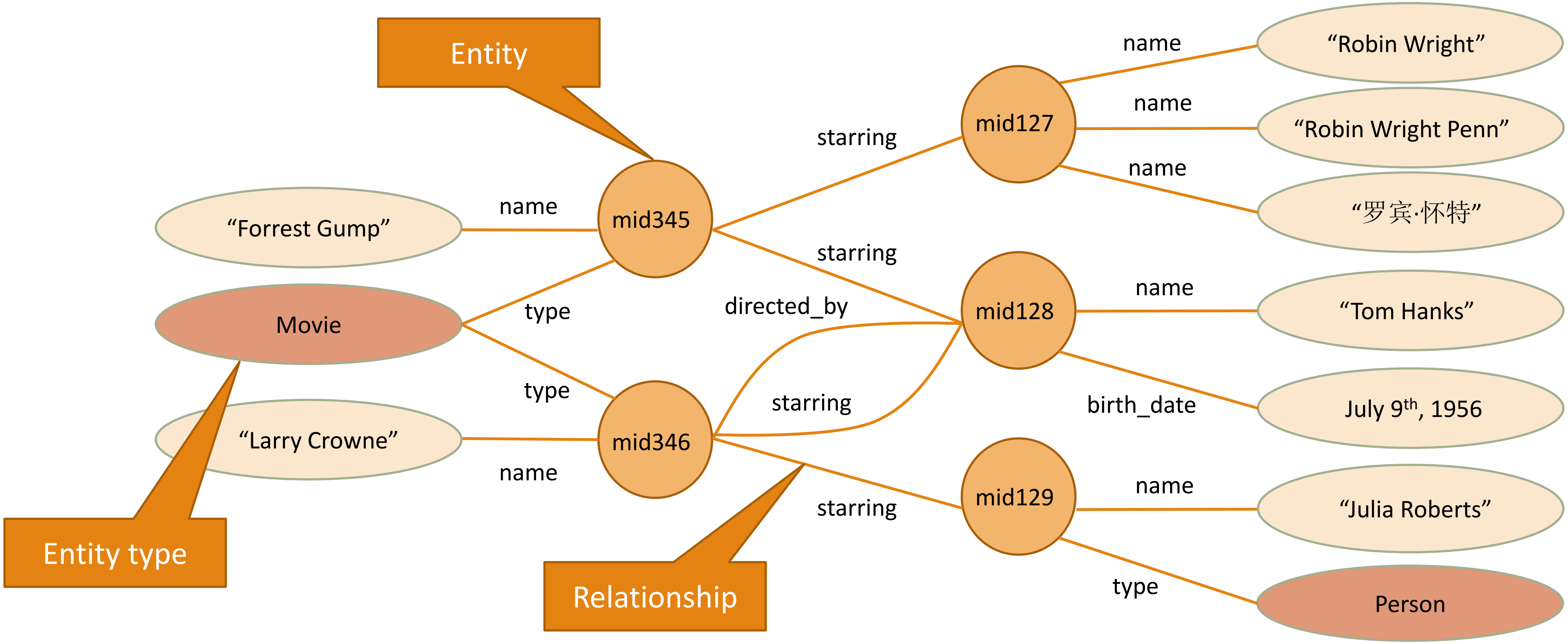
ICDE, 4/2019

Acknowledgement

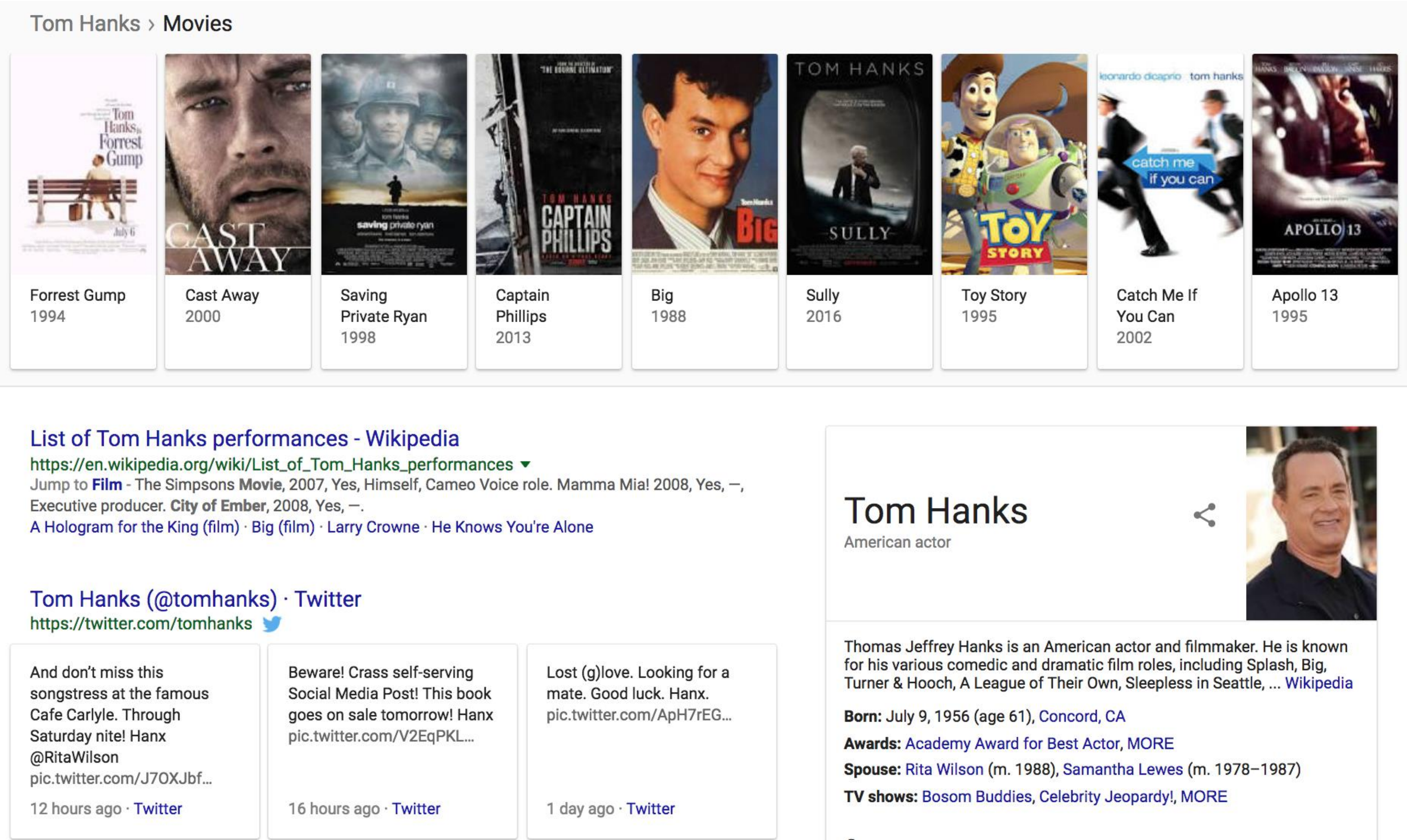


Product Graph vs. Knowledge Graph

Knowledge Graph Example for 2 Movies

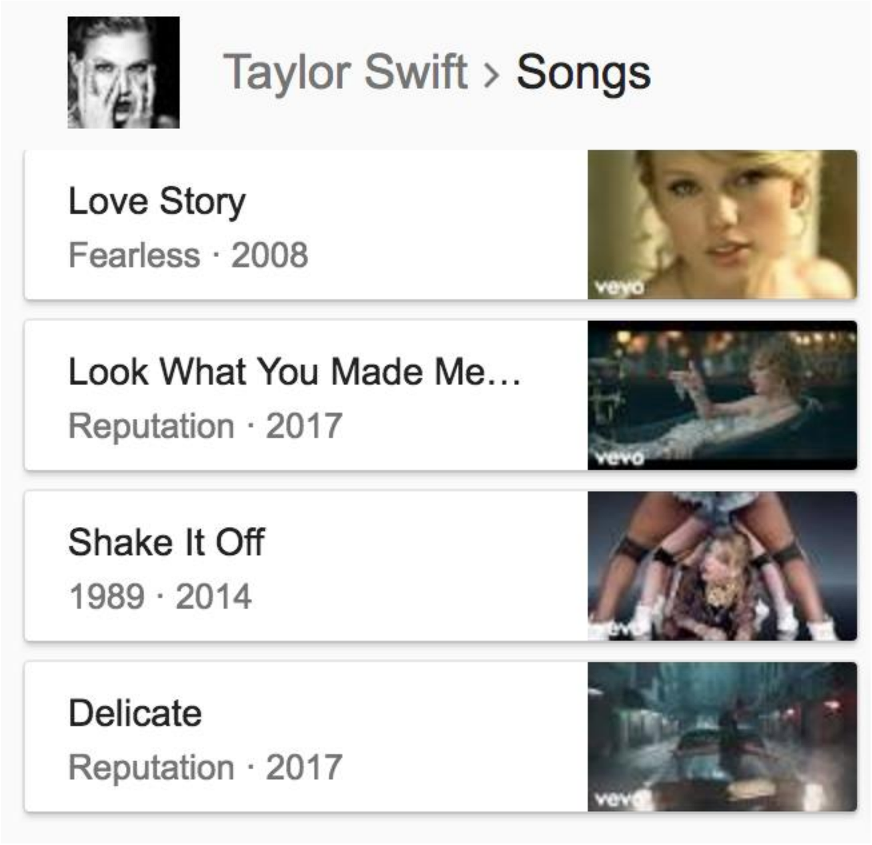


Knowledge Graph in Search



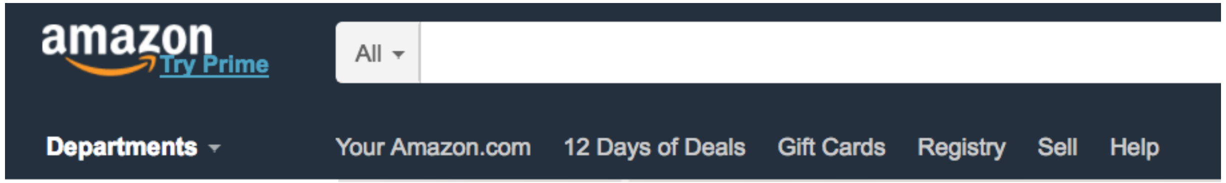
Knowledge Graph in Personal Assistant

Alexa, play Taylor Swift in the past year

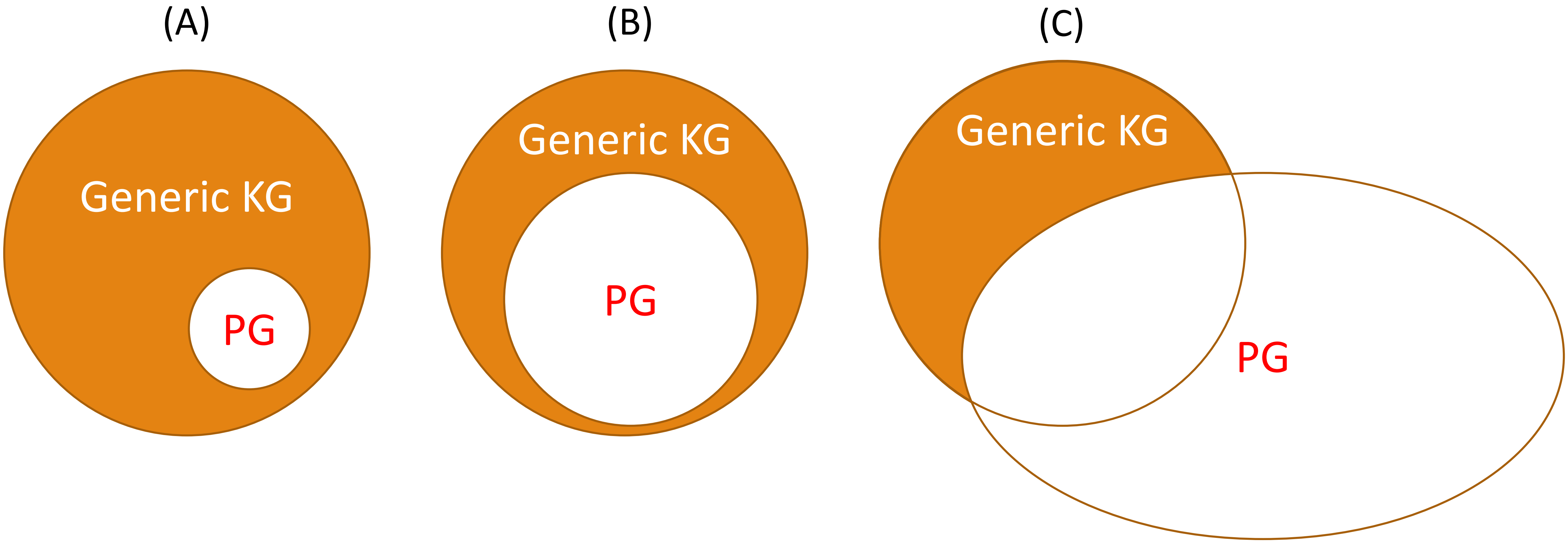


Product Graph

- Mission: To answer any question about products and related knowledge in the world

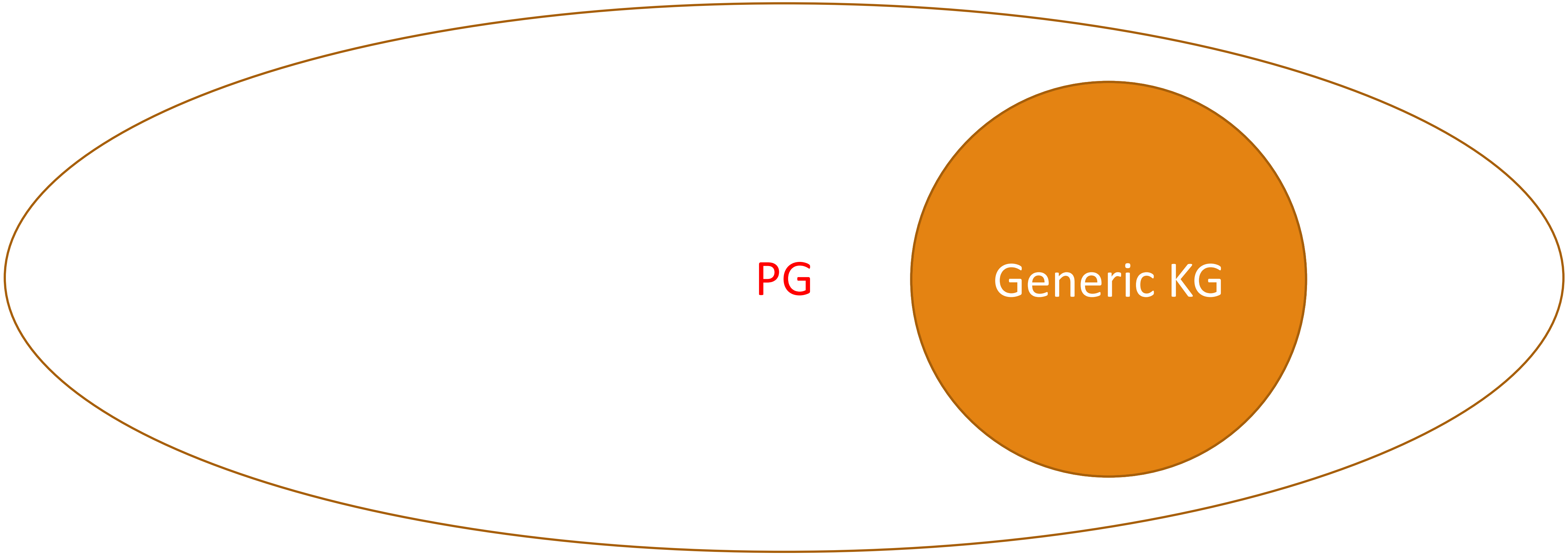


Product Graph vs. Knowledge Graph

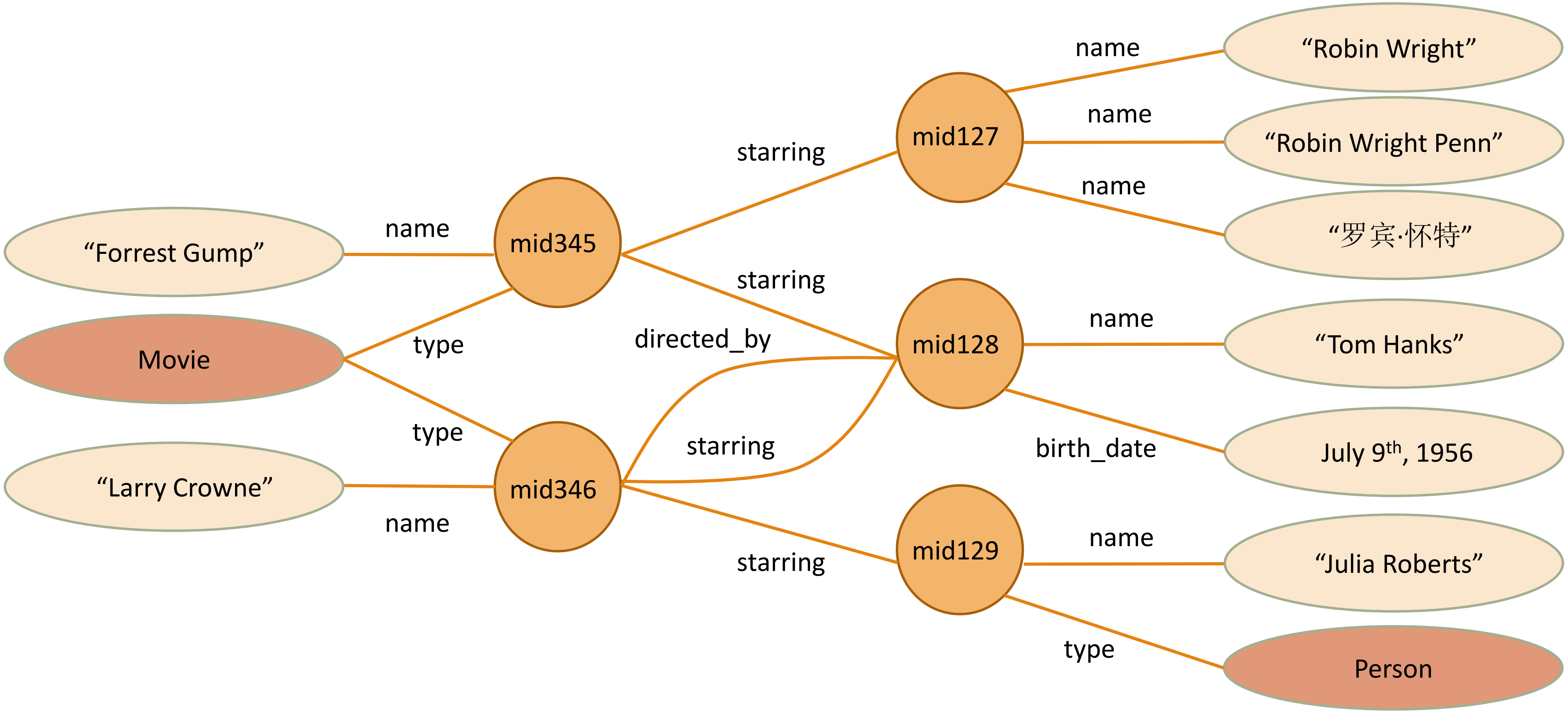


Product Graph vs. Knowledge Graph

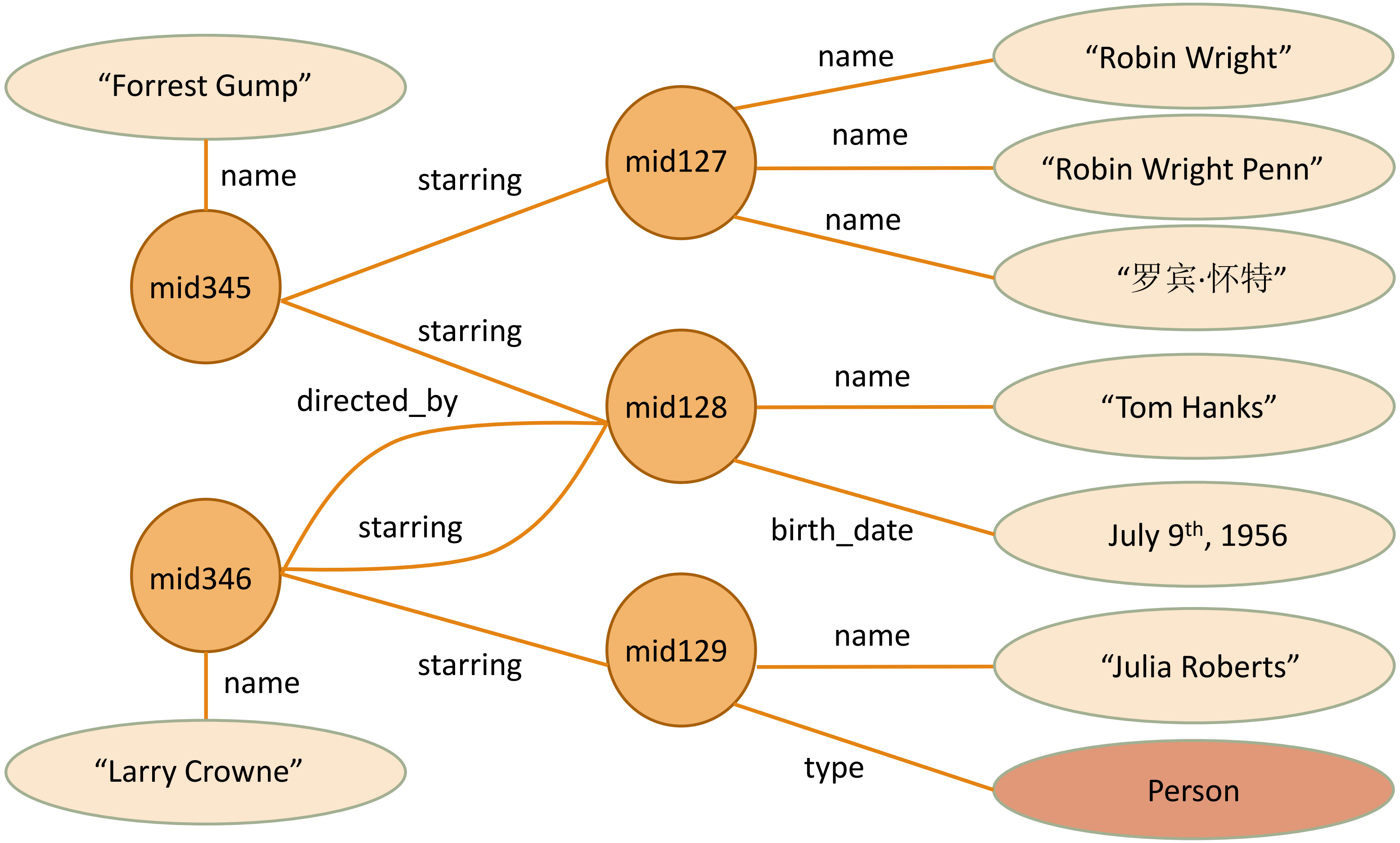
(D)



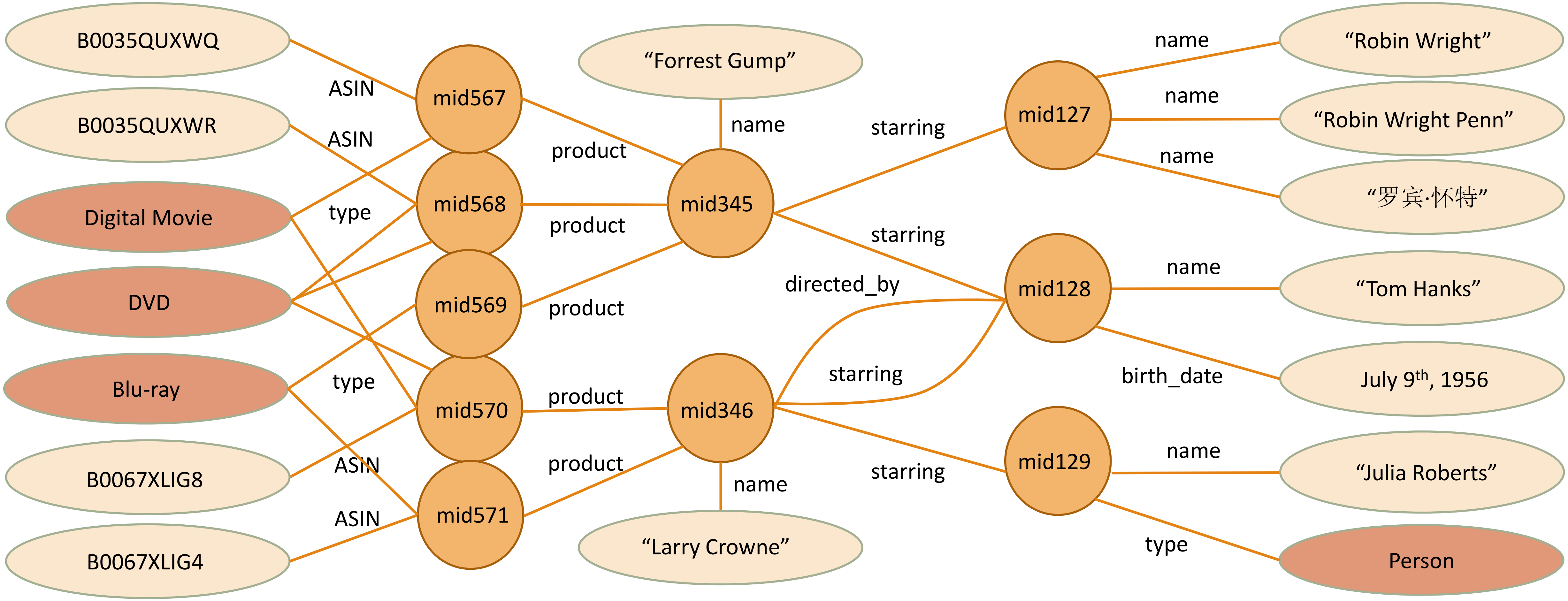
Knowledge Graph Example for 2 Movies



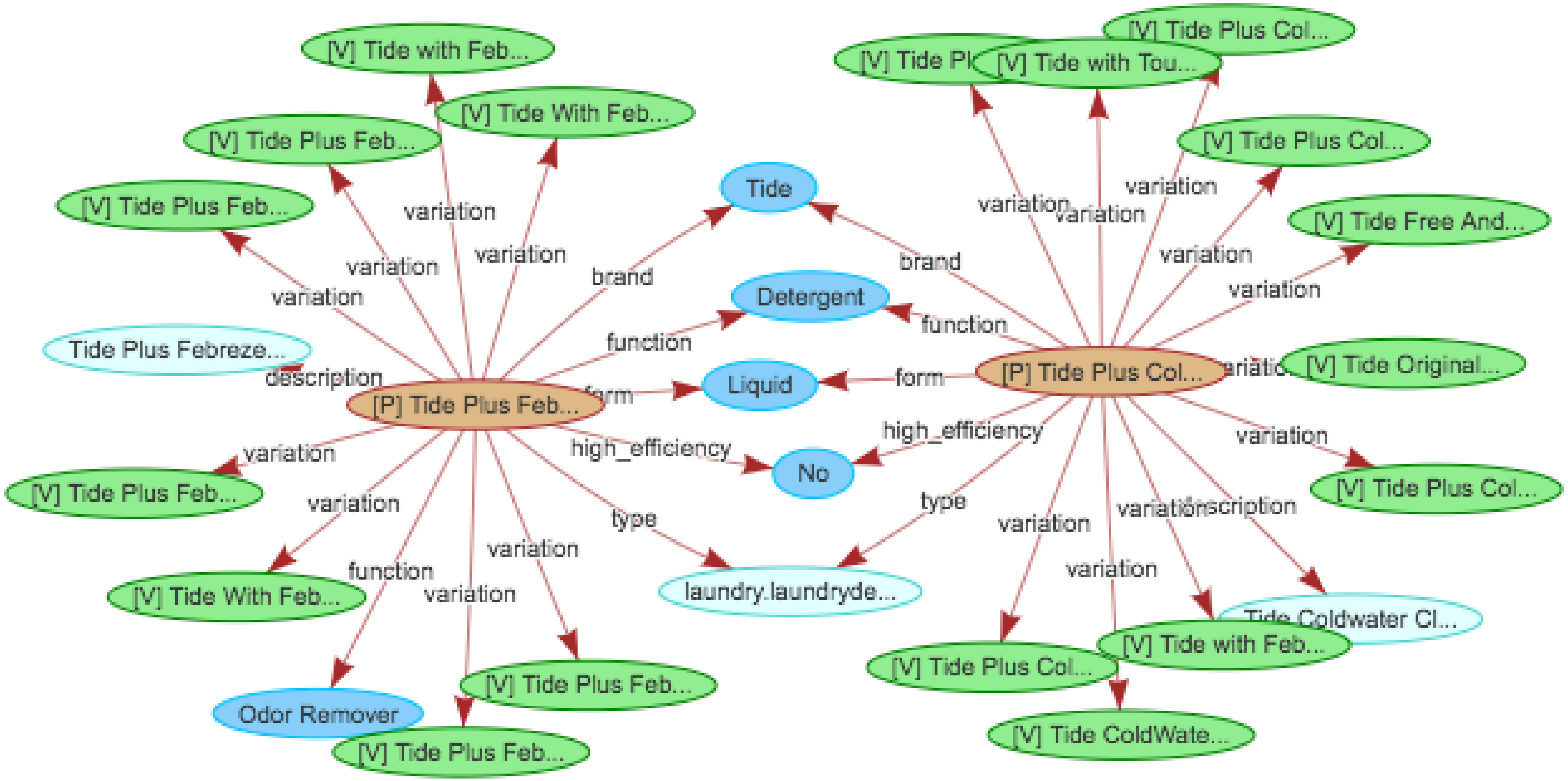
Product Graph vs. Knowledge Graph



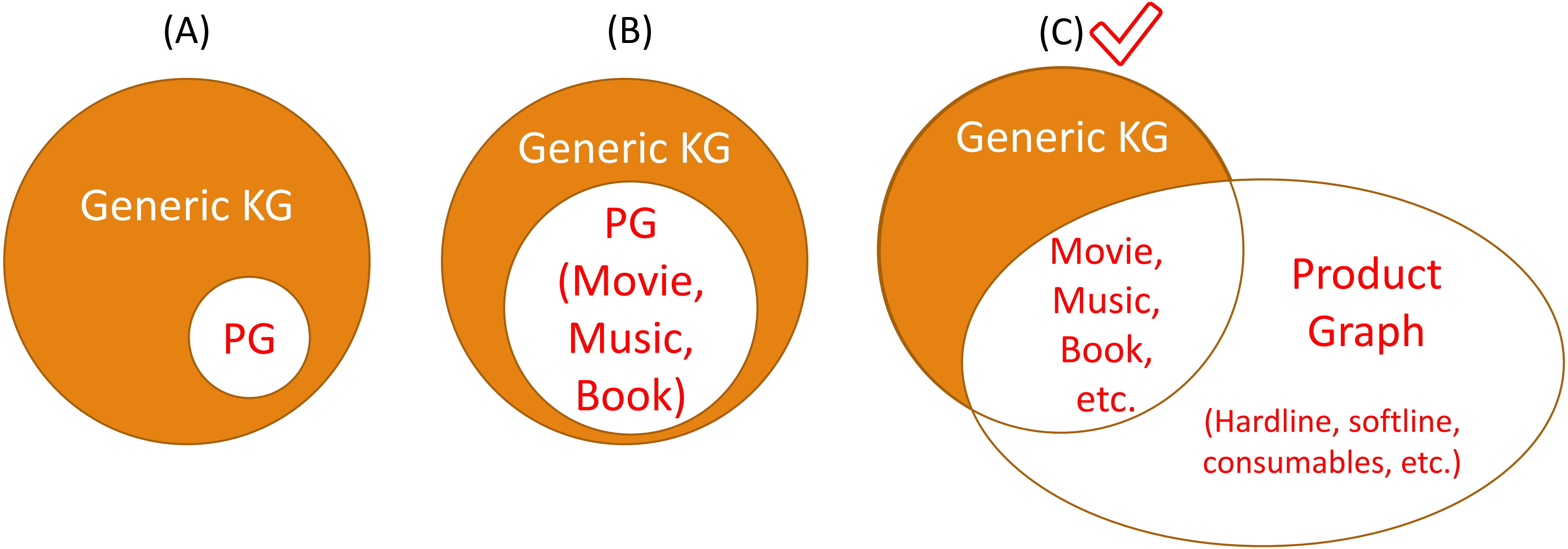
Product Graph vs. Knowledge Graph

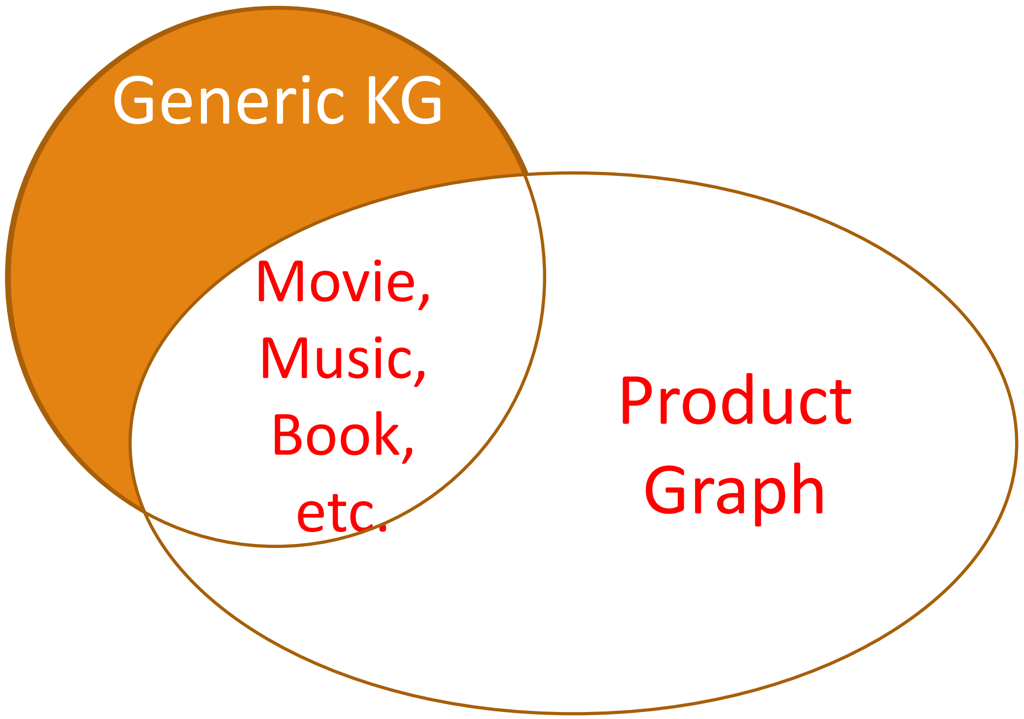


Another Example of Product Graph



Knowledge Graph vs. Product Graph





But, Is The Problem Harder?

Challenges in Building Product Graph I

- No major sources to curate product knowledge from
  - Wikipedia does not help too much
  - A lot of structured data buried in text descriptions in Catalog
  - Retailers gaming with the system so noisy data

Challenges in Building Product Graph II

- Large number of new products everyday
  - Curation is impossible
  - Freshness is a big challenge

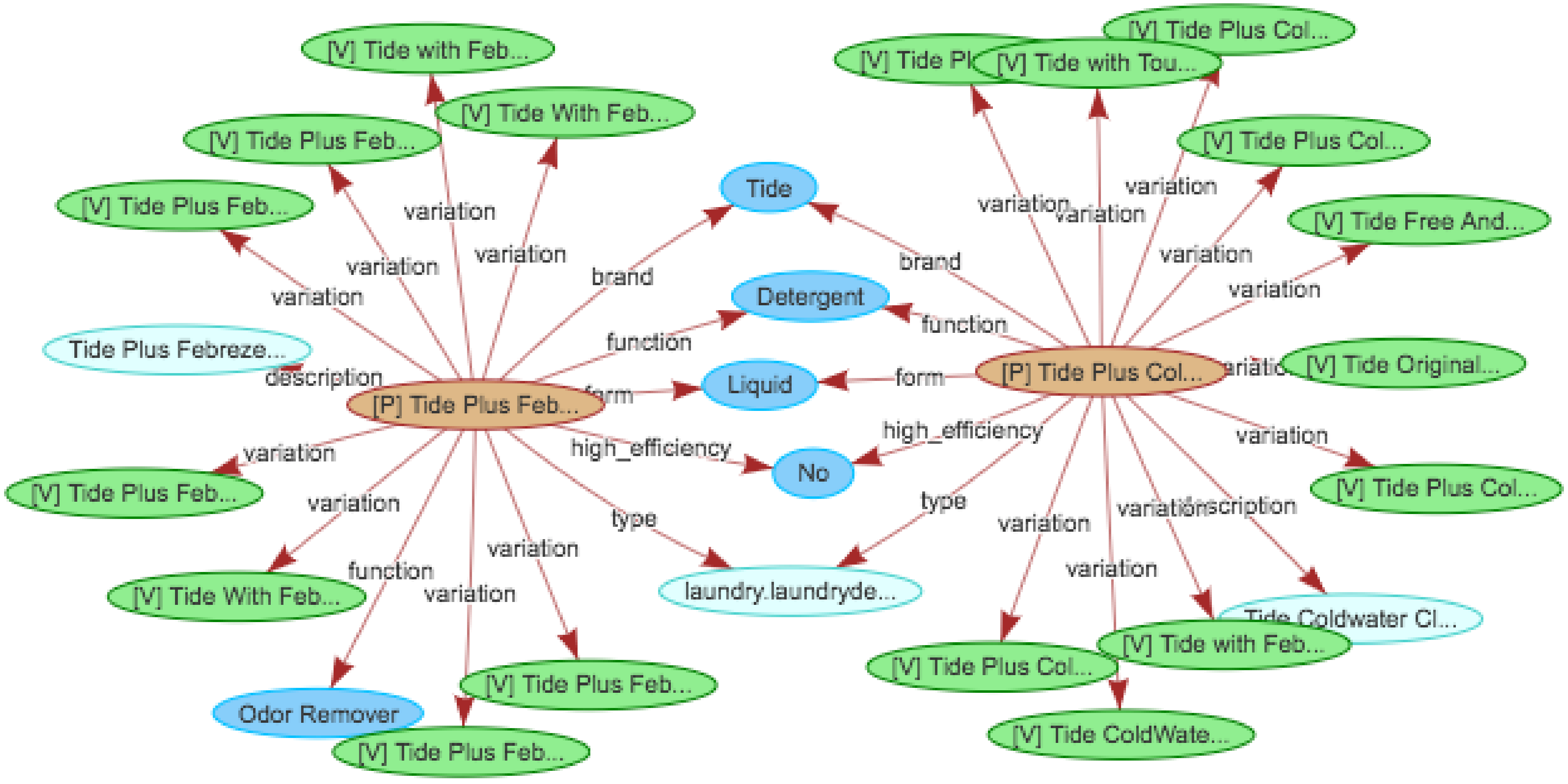
Challenges in Building Product Graph III

- Large number of product categories
  - A lot of work to manually define ontology
  - Hard to catch the trend of new product categories and properties

Challenges in Building Product Graph IV

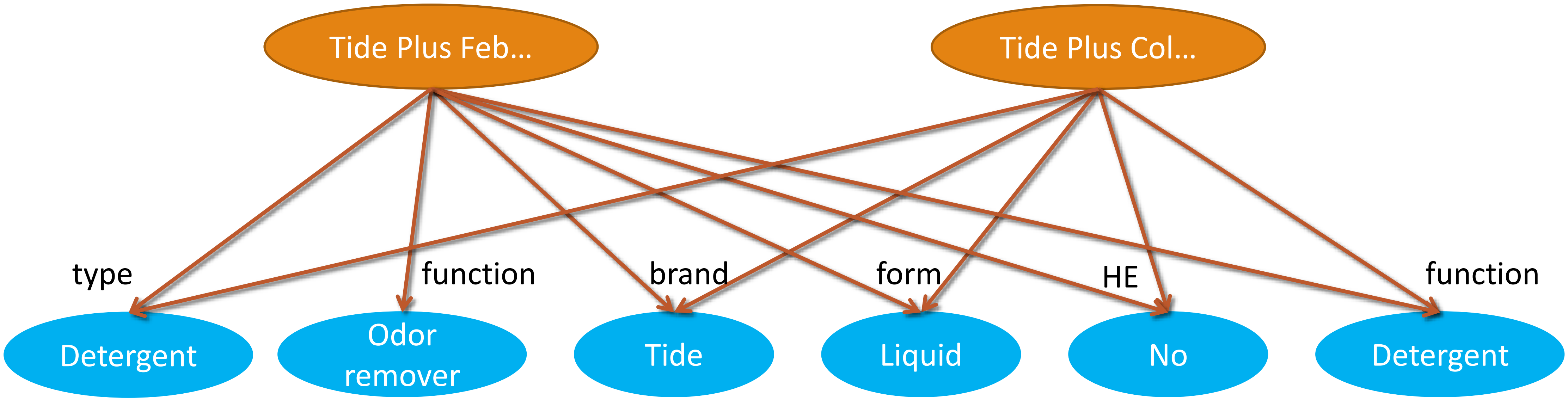
- Many entities are not named entities
  - Named Entity Recognition does not apply
  - New challenges for extraction, linking, and search

A 100-Year Project



Our Solution:
Building a Broad Graph

A Broad & Shallow Version of the Same Graph



Design Principles for Broad Graph

- Start simple
  - Bi-partite graph
  - Core types and relationships
- Grow and clean the graph in a pay-as-you-go fashion
  - Ontology: user log analysis, web extraction
  - Data: product profile extraction, web extraction
  - Cleaning

Input and Output of Broad Graph

INPUT

1. A given vertical
2. Amazon Eco System
3. Semi-structured Web

Hard work and fun

OUTPUT

1. Ontology
2. Broad Graph
3. Comprehensive quality metrics

subject to: $hard\_work \leq fun$

Stages in Building a Broad Graph

Stage 1 Ingestion

amazon

Coverage: 4X

Stage 2a OpenTag text extraction

[Zheng et al., KDD'18]

Size: +7%, predicates: +10%, ExtAccu: 99%

Stage 2b Ceres web extraction

[Lockard et al., VLDB'18]

WWW

Accu: +6%

Stage 3 Cleaning

Stages in Building a Broad Graph

Stage 1 Ingestion

amazon

Coverage: 4X

Stage 2a OpenTag text extraction

[Zheng et al., KDD'18]

Size: +7%, predicates: +10%, ExtAccu: 99%

Stage 2b Ceres web extraction

[Lockard et al., VLDB'18]

Accu: +6%

Stage 3 Cleaning

Stage 2a. OpenTag Extraction from Product Profiles [KDD'18]

name	form	scent
Tide Detergent with Febreze Freshness		
Gain Apple Mango Tango Liquid Laundry Detergent		
Gain Joyful Expression Powder Detergent		
Tide PODS Original Scent HE Turbo Laundry Detergent Pacs 81-load Tub		
Tide PODS Free & Gentle HE Turbo Laundry Detergent Pacs 35-load Bag		

Stage 2a. OpenTag Extraction from Product Profiles [KDD'18]

B Beginning of attribute value

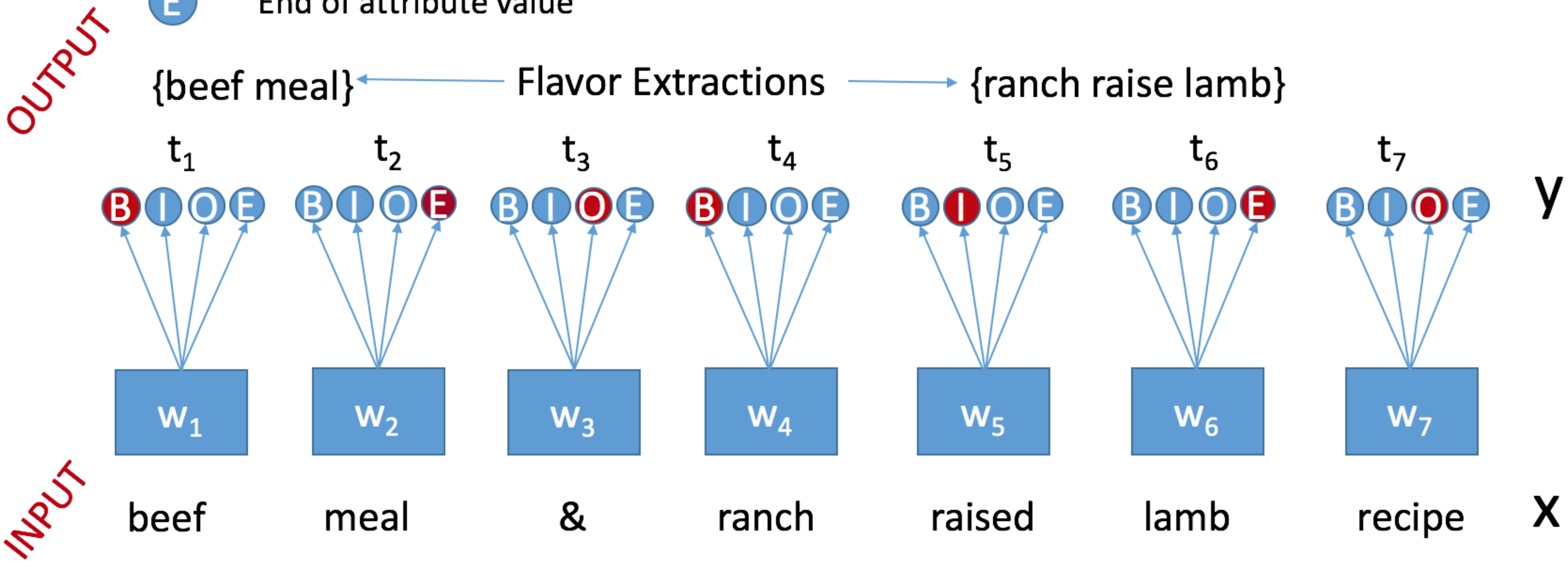
I Inside of attribute value

O Outside of attribute value

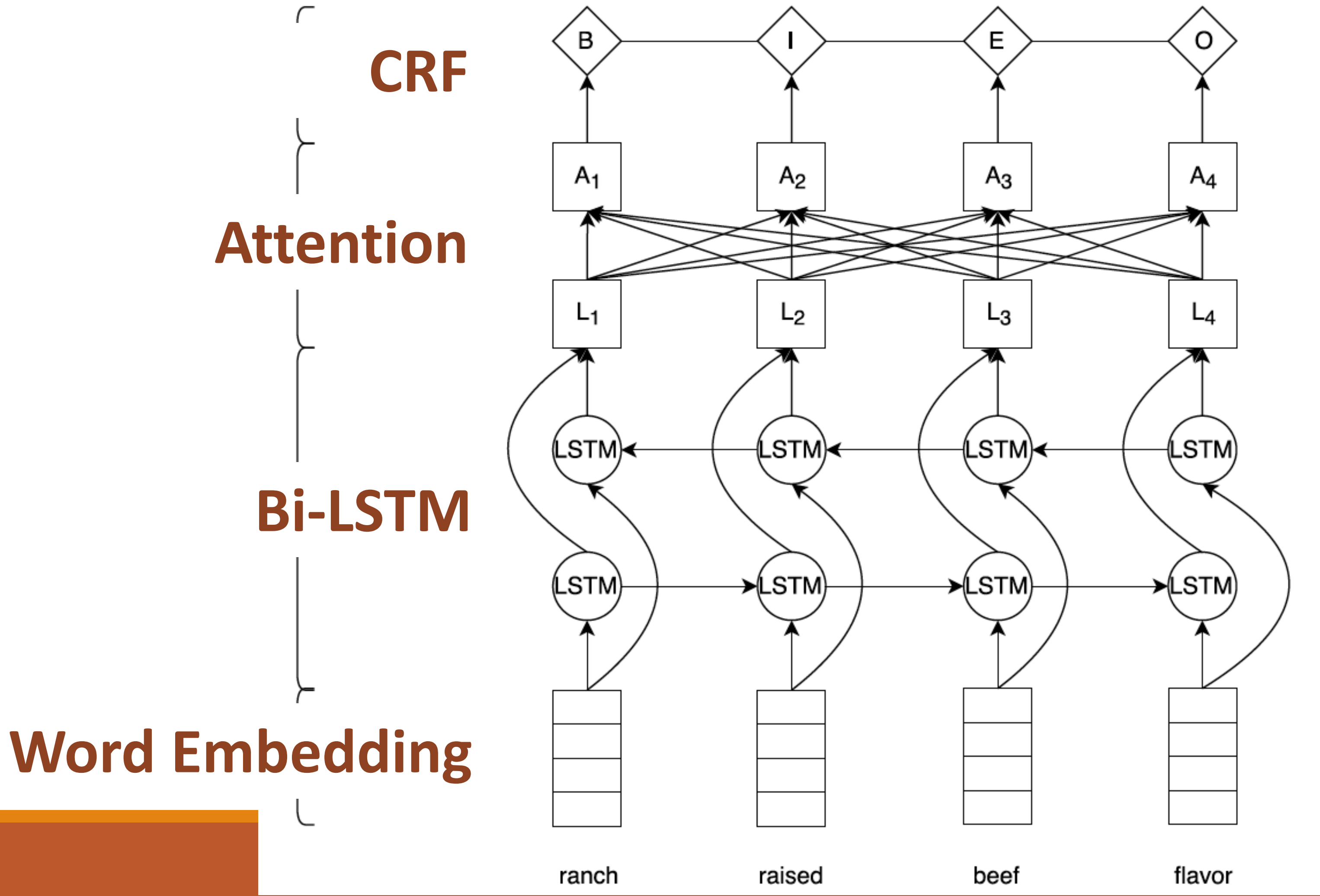
E End of attribute value

$x=\{w_1,w_2,\ldots,w_n\}$ input sequence

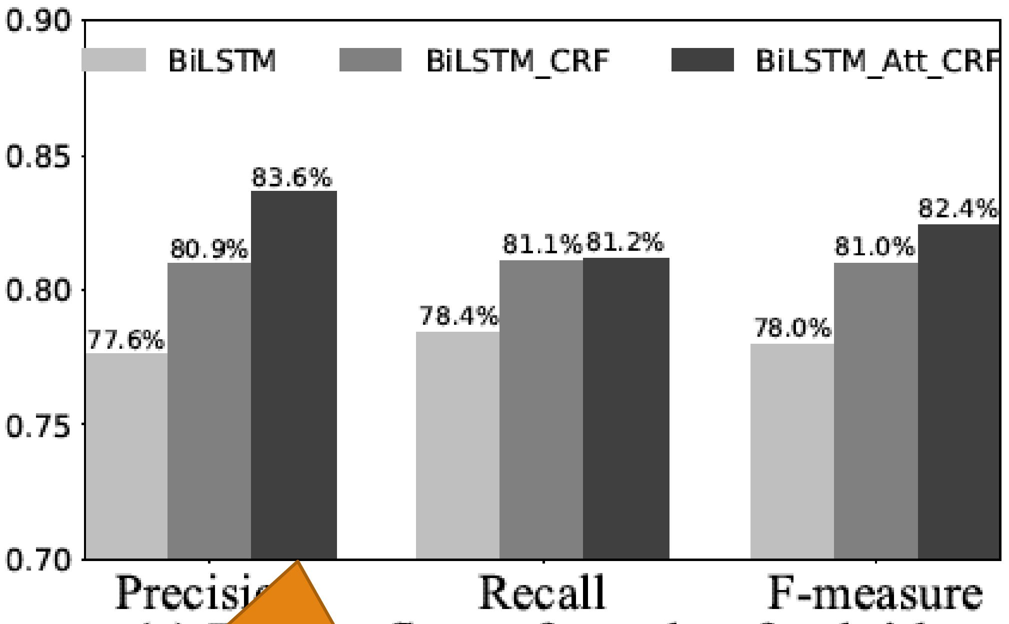
$y=\{t_1,t_2,\ldots,t_n\}$ tagging decision



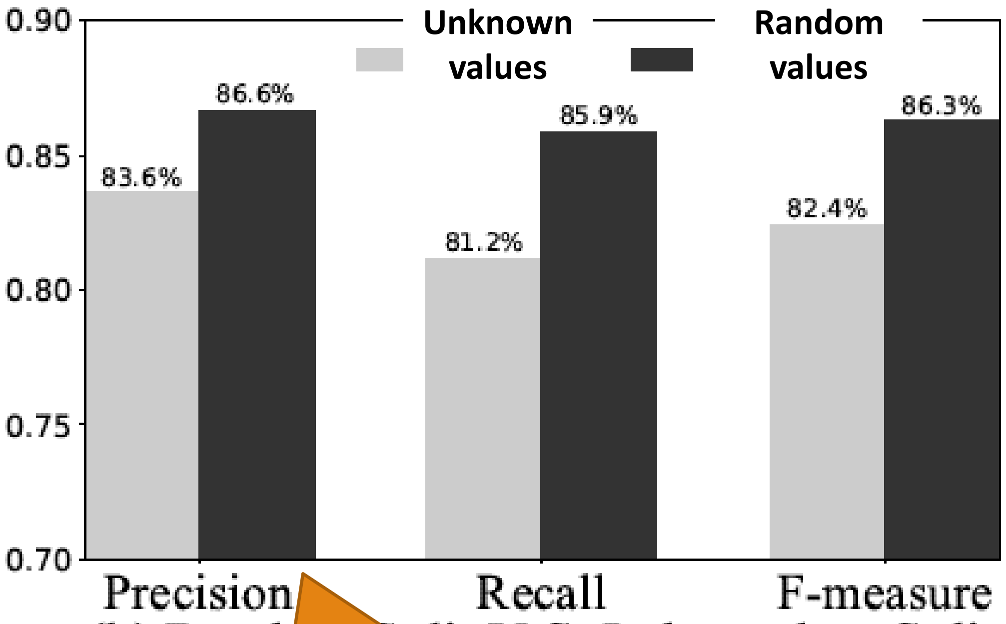
Stage 2a. OpenTag Extraction from Product Profiles [KDD'18]



Stage 2a. OpenTag Extraction from Product Profiles [KDD'18]



(a) t flavor from dog food title

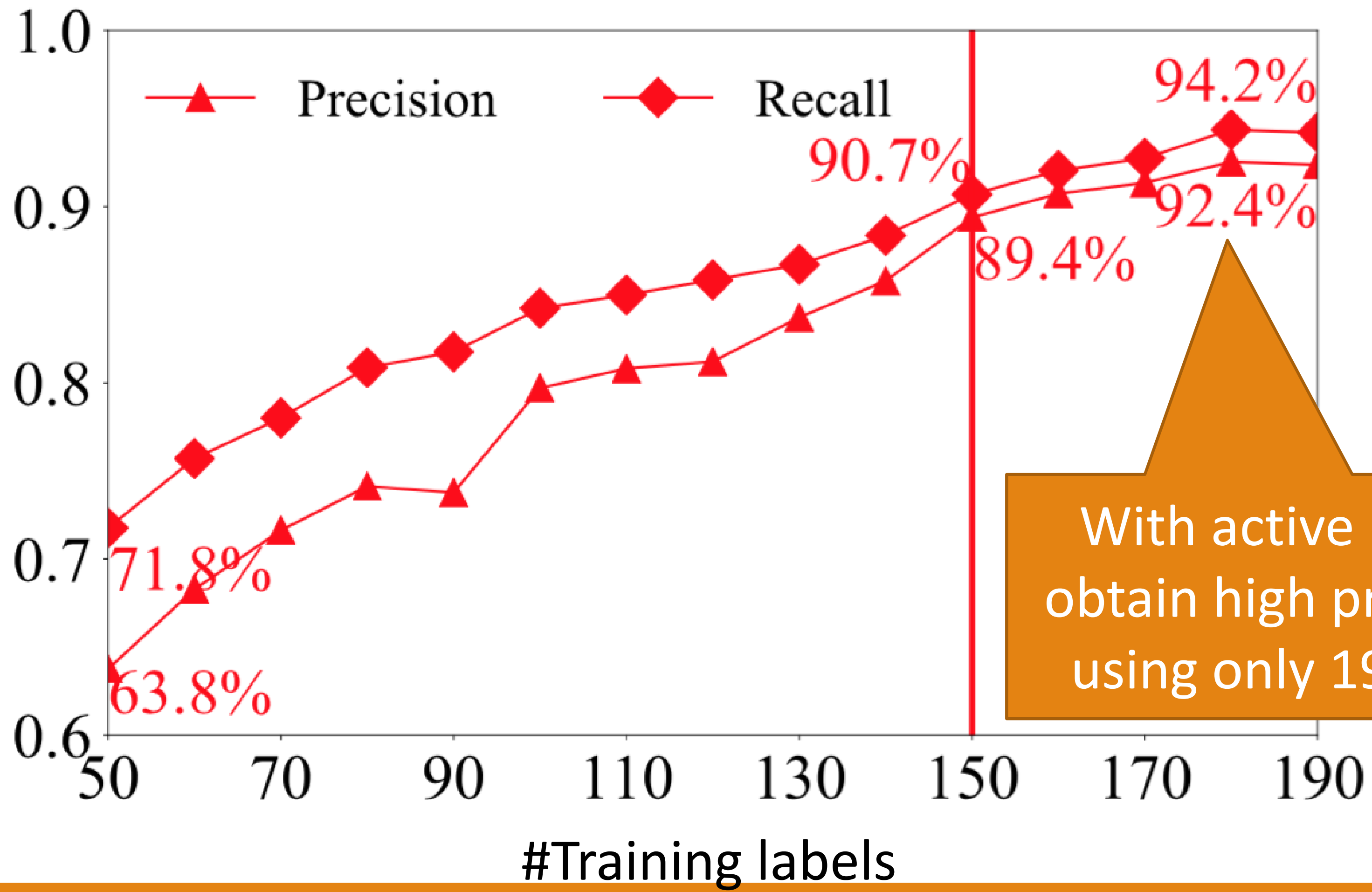


(b) Rand lit V.S. Independent Split

BiLSTM+CRF+Attention obtains best results

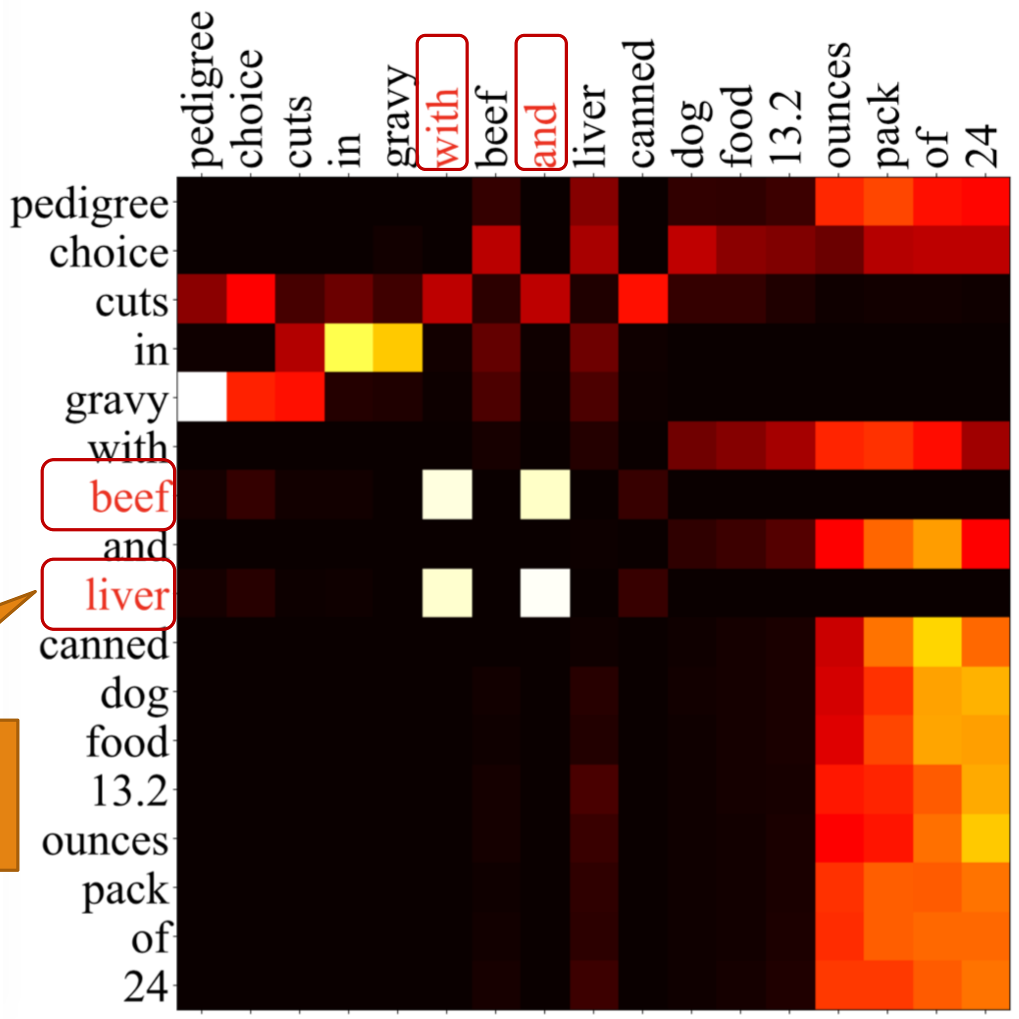
Extraction on new values is comparable to already known values

Stage 2a. OpenTag Extraction from Product Profiles [KDD'18]



With active learning, we can obtain high precision and recall using only 190 training labels

Stage 2a. OpenTag Extraction from Product Profiles [KDD'18]



Interpretability via attention

Stages in Building a Broad Graph

Stage 1 Ingestion

amazon

Coverage: 4X

Stage 2a OpenTag text extraction

[Zheng et al., KDD'18]

Size: +7%, predicates: +10%, ExtAccu: 99%

Stage 2b Ceres web extraction

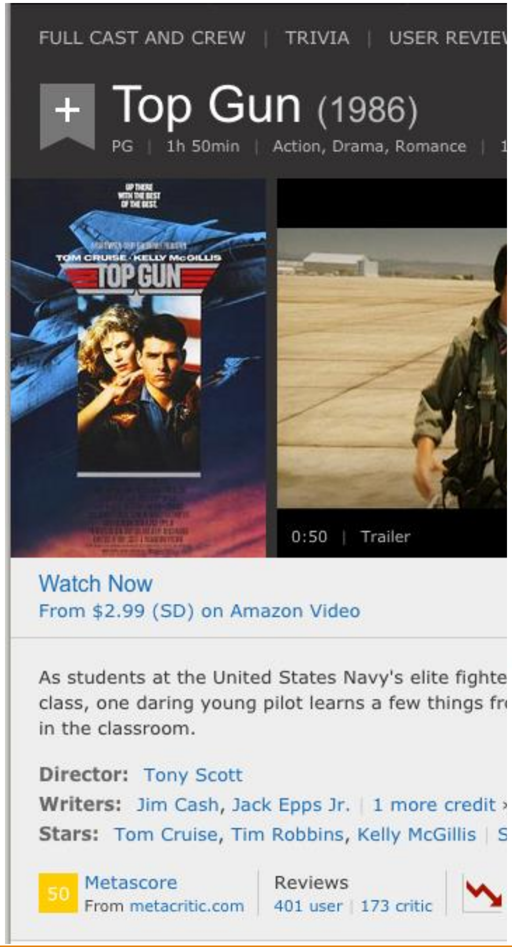
[Lockard et al., VLDB'18]

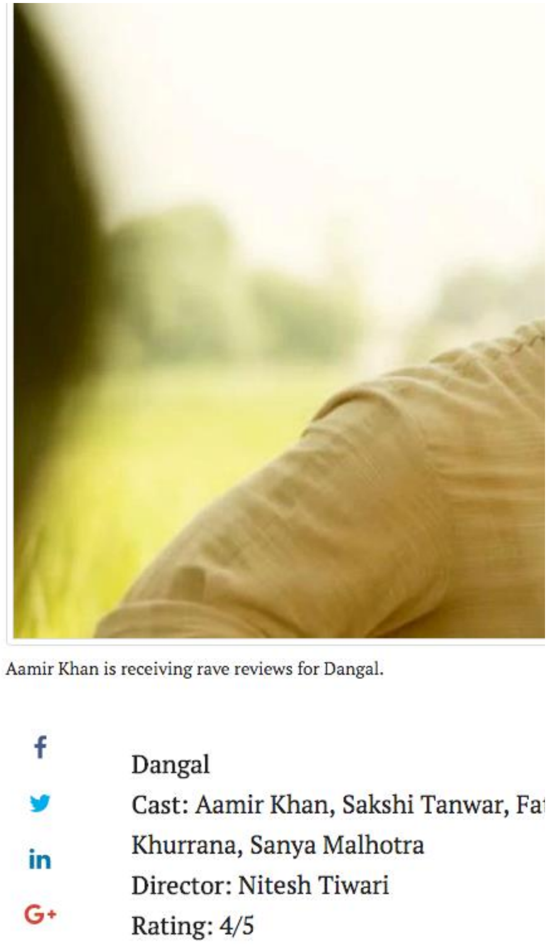
WWW

Accu: +6%

Stage 3 Cleaning

Stage 2b. Extracting Knowledge from Semi-Structured Data on the Web

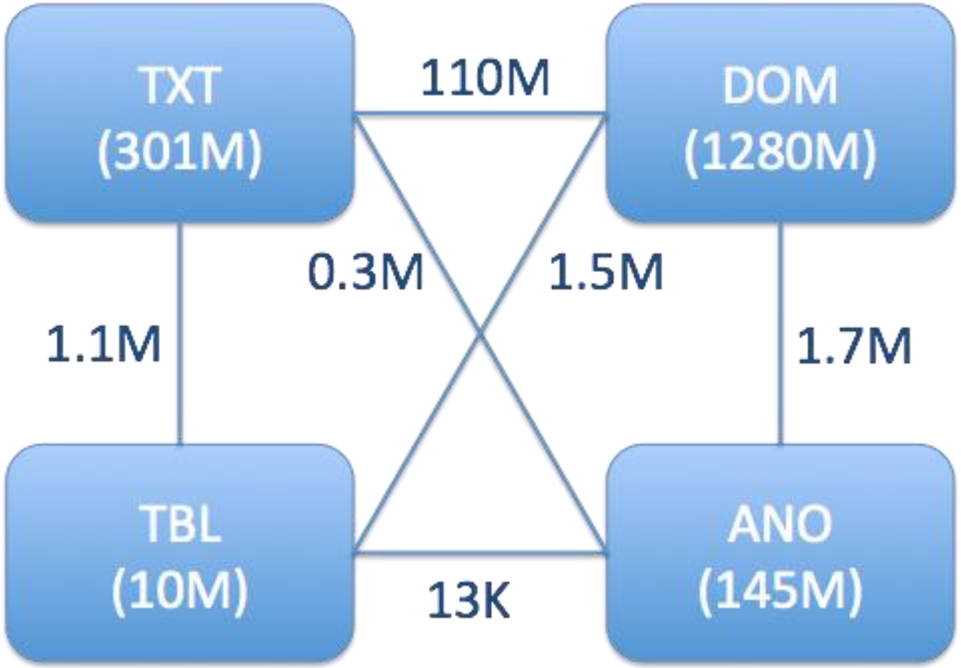




Stage 2b. Extracting Knowledge from Semi-Structured Data on the Web

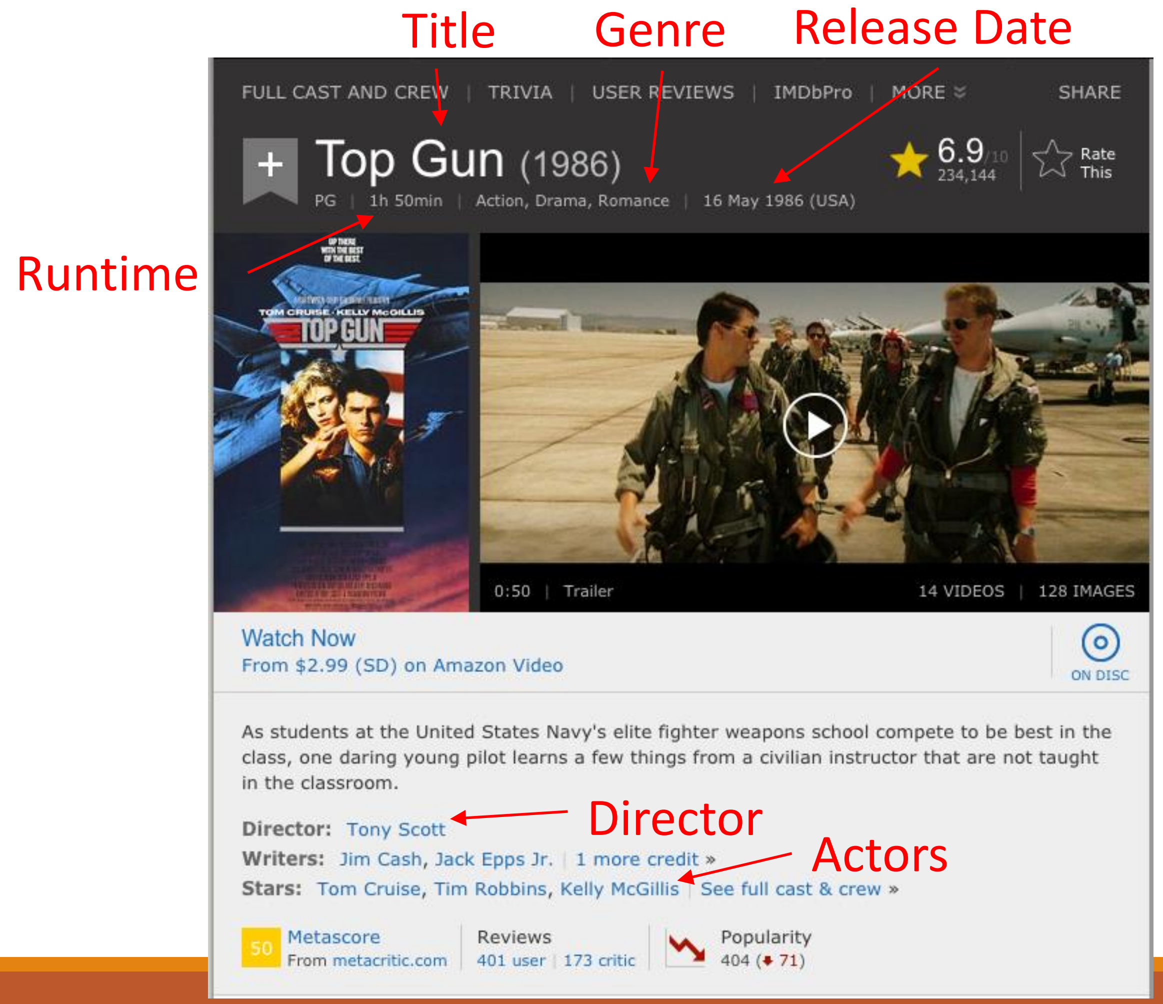
- Knowledge Vault @ Google showed big potential from DOM-tree extraction [Dong et al., KDD'14][Dong et al., VLDB'14]

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



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

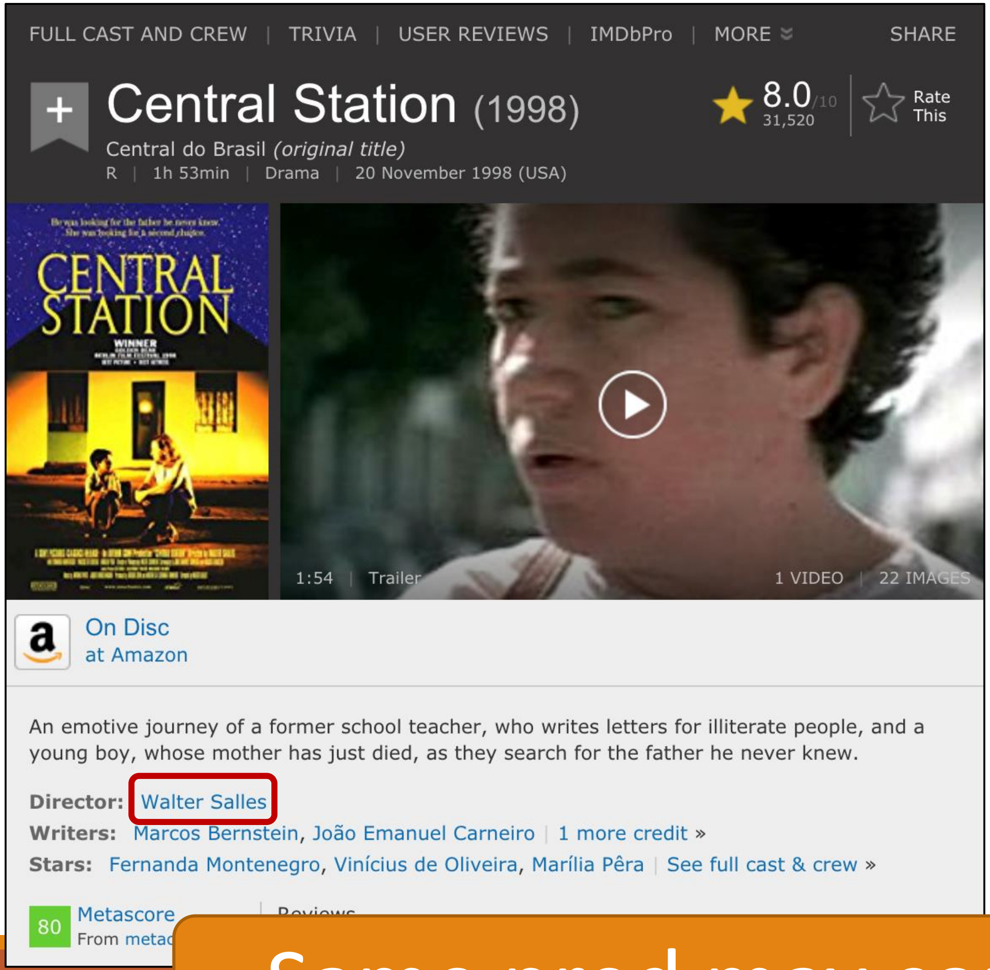
Stage 2b. Extracting Knowledge from Semi-Structured Data on the Web

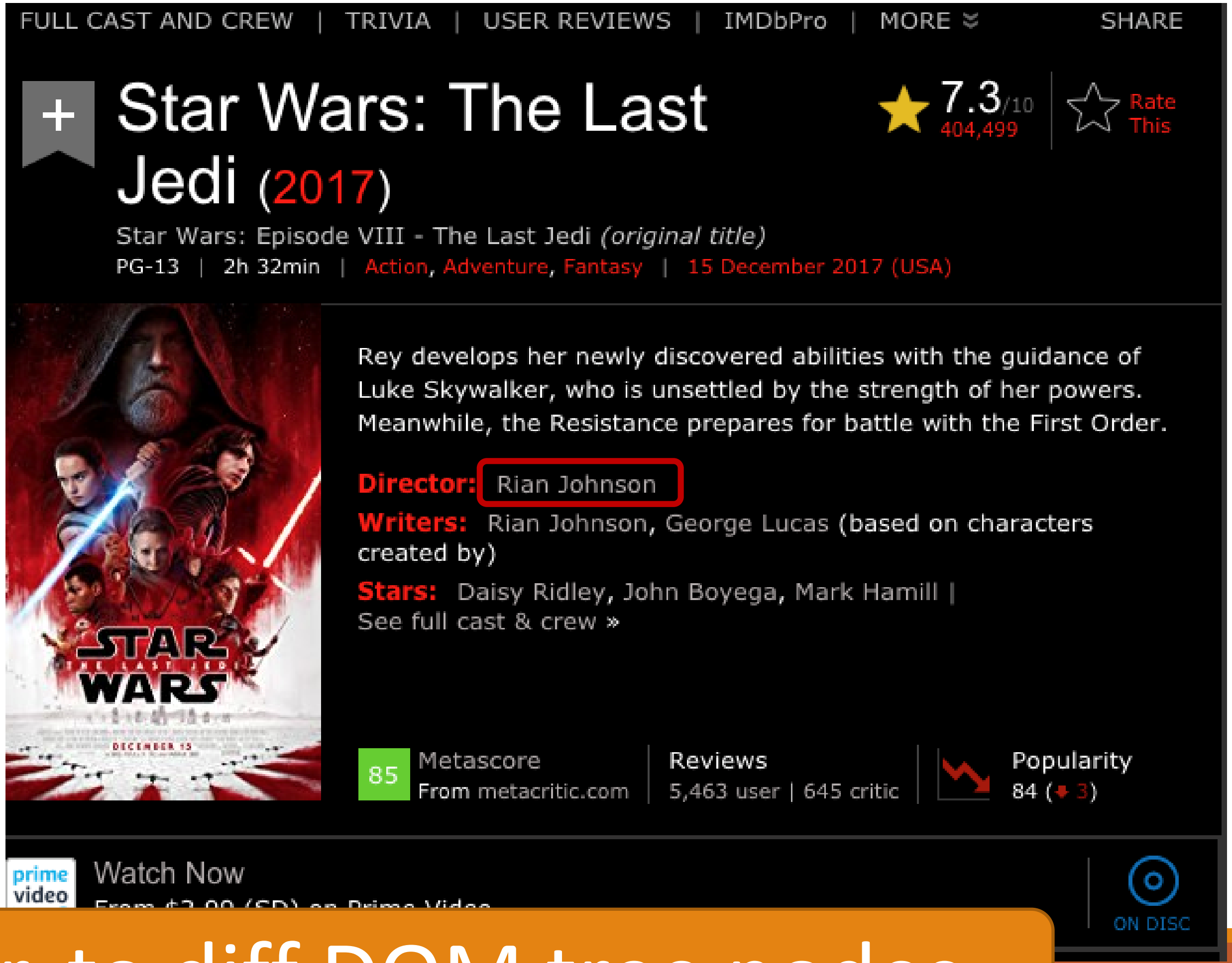


Extracted relationships

- (Top Gun, type.object.name, “Top Gun”)
- (Top Gun, film.film.genre, Action)
- (Top Gun, film.film.directed_by, Tony Scott)
- (Top Gun, film.film.starring, Tom Cruise)
- (Top Gun, film.film.runtime, “1h 50min”)
- (Top Gun, film.film.release_Date_s, “16 May 1986”)

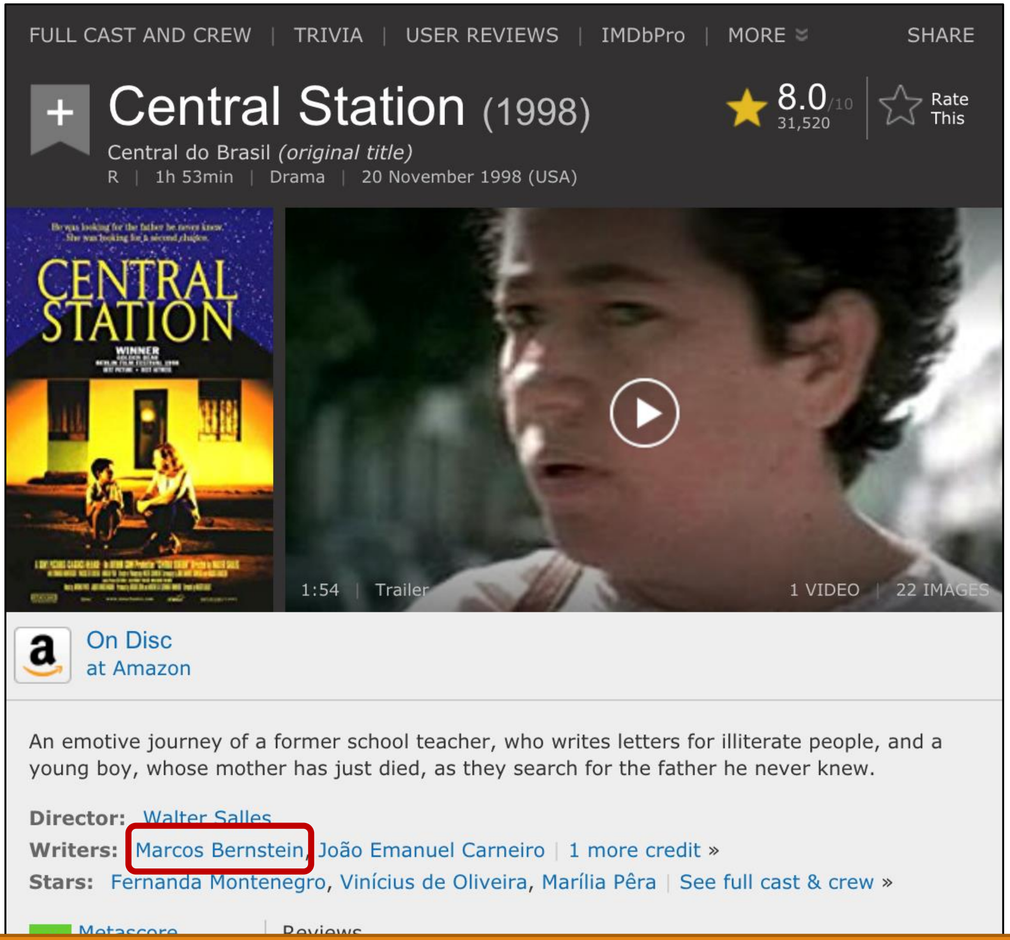
Stage 2b. Extracting Knowledge from Semi-Structured Data on the Web

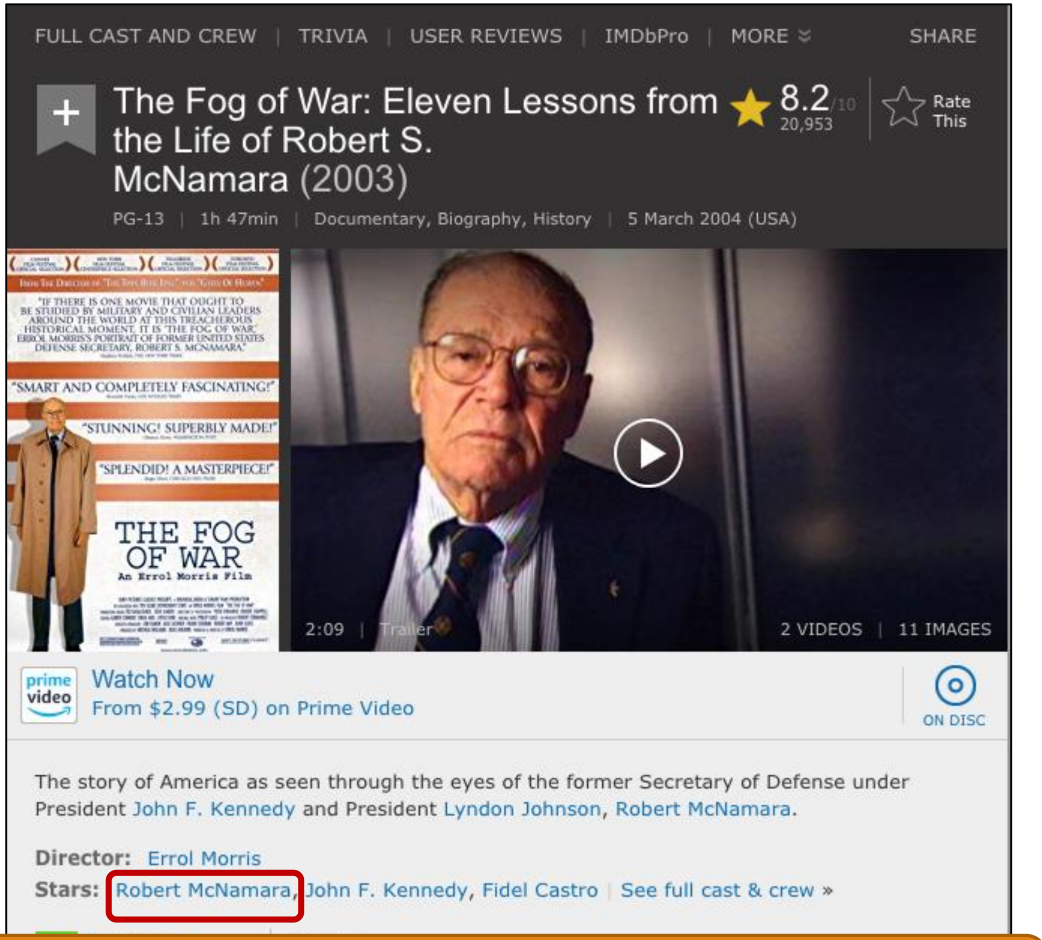




Same pred may corr. to diff DOM tree nodes

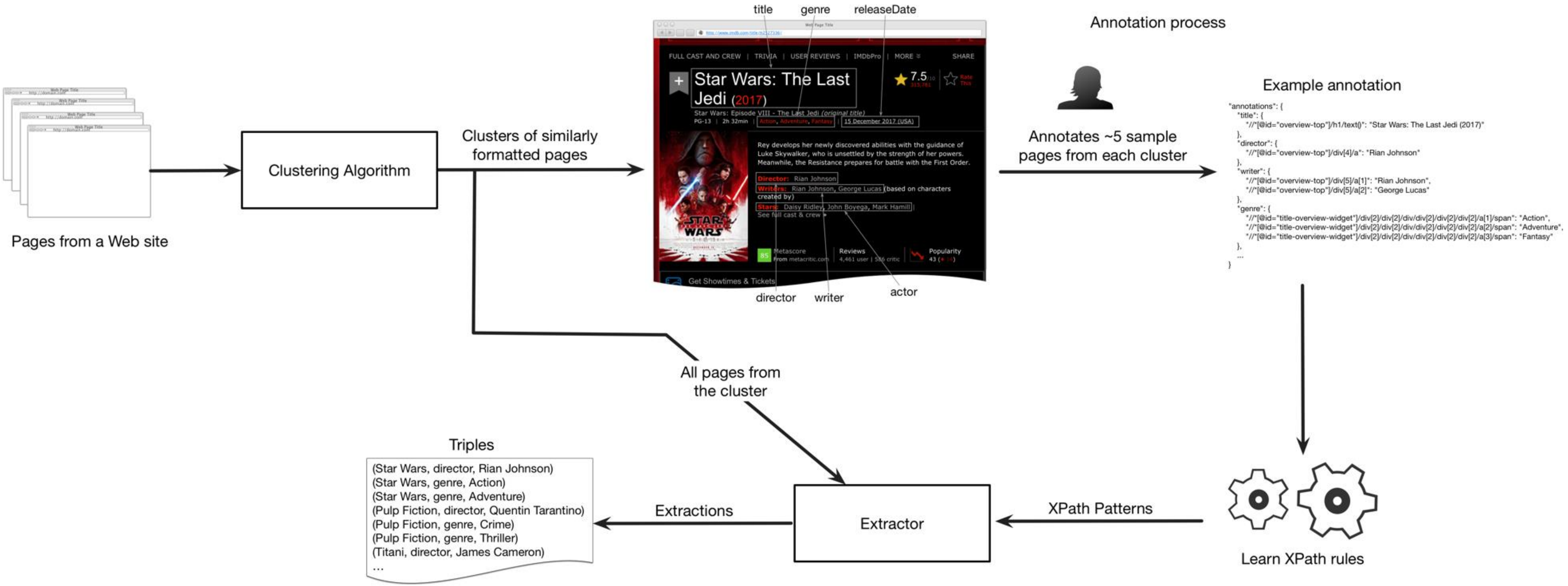
Stage 2b. Extracting Knowledge from Semi-Structured Data on the Web





Same DOM tree node may correspond to diff preds

Stage 2b. Extracting Knowledge from Semi-Structured Data on the Web



Stage 2b. Linking Entities Between Sources

- Extraction experiments on http://swde.codeplex.com/
  - Near perfect precision and recall
  - Annotations on 5 pages per site

Vertical	Predicate	Vertex++ P	Vertex++ R	Vertex++ F1
Movie	Title	1.00	1.00	1.00
	Director	0.99	0.99	0.99
	Genre	0.88	0.87	0.87
	MPAA Rating	1.00	1.00	1.00
	Average	0.97	0.97	0.97
NBAPlayer	Name	0.99	0.99	0.99
	Team	1.00	1.00	1.00
	Weight	1.00	1.00	1.00
	Height	1.00	1.00	1.00
	Average	1.00	1.00	1.00

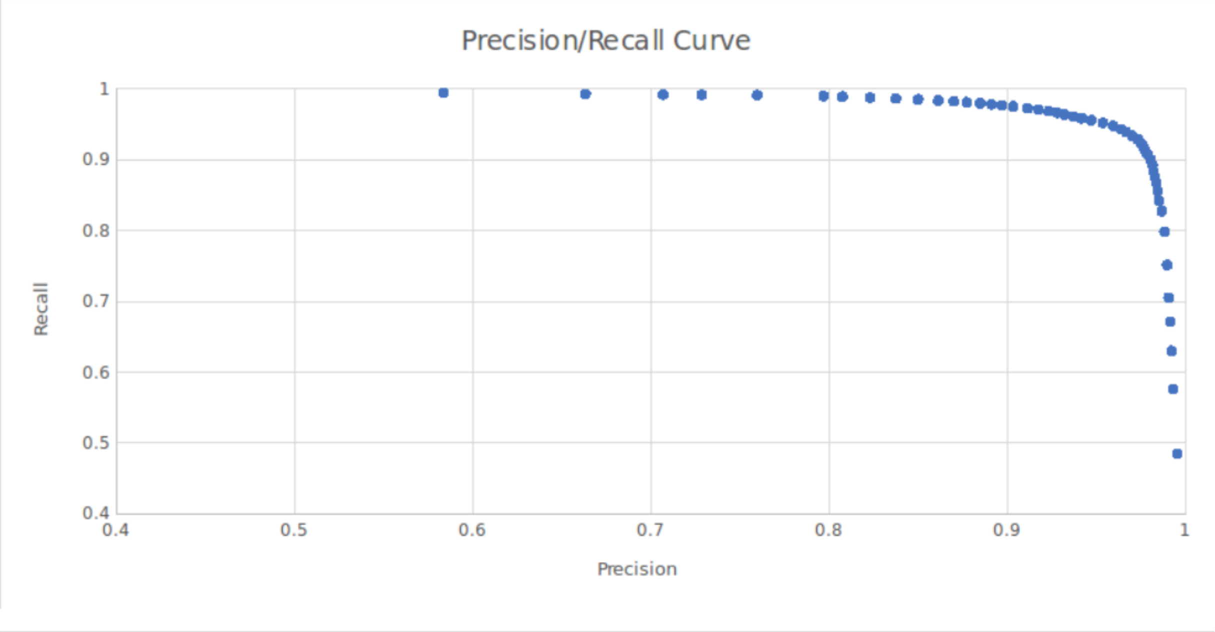
Vertical	Predicate	Vertex++ P	Vertex++ R	Vertex++ F1
University	Name	1.00	1.00	1.00
	Type	1.00	1.00	1.00
	Phone	0.97	0.92	0.94
	Website	1.00	1.00	1.00
	Average	0.99	0.98	0.99
Book	Title	0.99	0.99	0.99
	Author	0.97	0.96	0.96
	Publisher	0.85	0.85	0.85
	Publication Date	0.90	0.90	0.90
	ISBN-13	0.94	0.94	0.94
	Average	0.93	0.93	0.93

Stage 2b. Linking Entities Between Sources

- Random forest on attribute-wise similarity
- Results between Freebase and IMDb movies

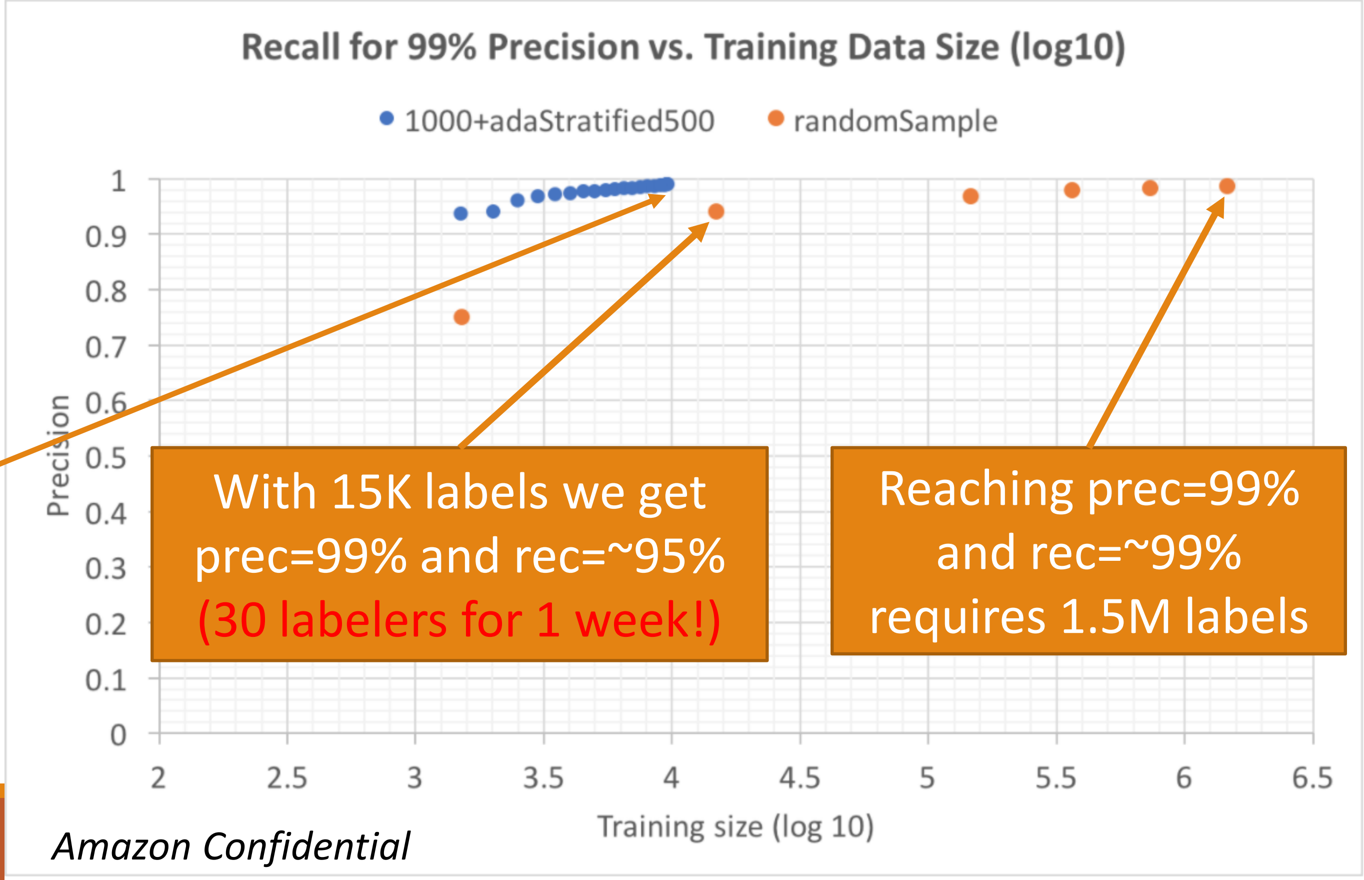
	Precision	Recall
Movie	99.0%	98.7%
People	99.3%	99.6%

1.5M labels



Stage 2b. Linking Entities Between Sources

- Apply active learning to minimize #labels



For 99% precision and recall, active learning reduces #labels by **2** orders of magnitude

With 15K labels we get prec=99% and rec=~95% (30 labelers for 1 week!)

Reaching prec=99% and rec=~99% requires 1.5M labels

Amazon Confidential

Stages in Building a Broad Graph

Stage 1 Ingestion

Coverage: 4X

Stage 2a OpenTag text extraction

[Zheng et al., KDD'18]

Size: +7%, predicates: +10%, ExtAccu: 99%

Stage 2b Ceres web extraction

[Lockard et al., VLDB'18]

Accu: +6%

Stage 3 Cleaning

Stage 3. Supervised Knowledge Cleaning

- Value noise
  - Invalid brands: “1 lb”
  - Lengthy descriptive brands: “The company X was founded in…”
  - Unnormalized values: “Reese’s”, “reese”, “REESE’s”
- Inconsistency
  - E.g., isSugerFree && sugarPerServing > 0

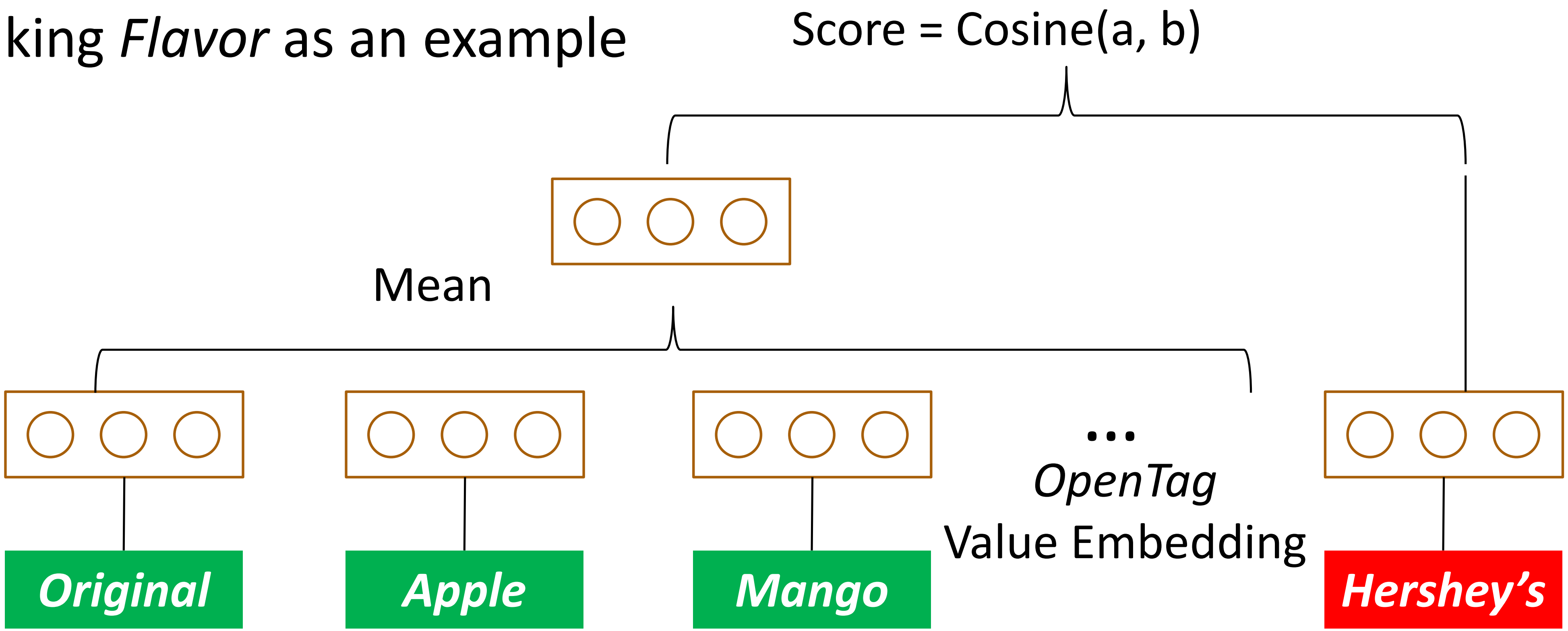
Stage 3. Supervised Knowledge Cleaning I

- A discriminative model to tell: Brand or Not a Brand?
- Features: frequency count, string lengths, special word hit, etc.
- Precision = 99.1%

Stage 3. Supervised Knowledge Cleaning II

- Embedding-based filtering

Taking *Flavor* as an example



Stage 3. Supervised Knowledge Cleaning III

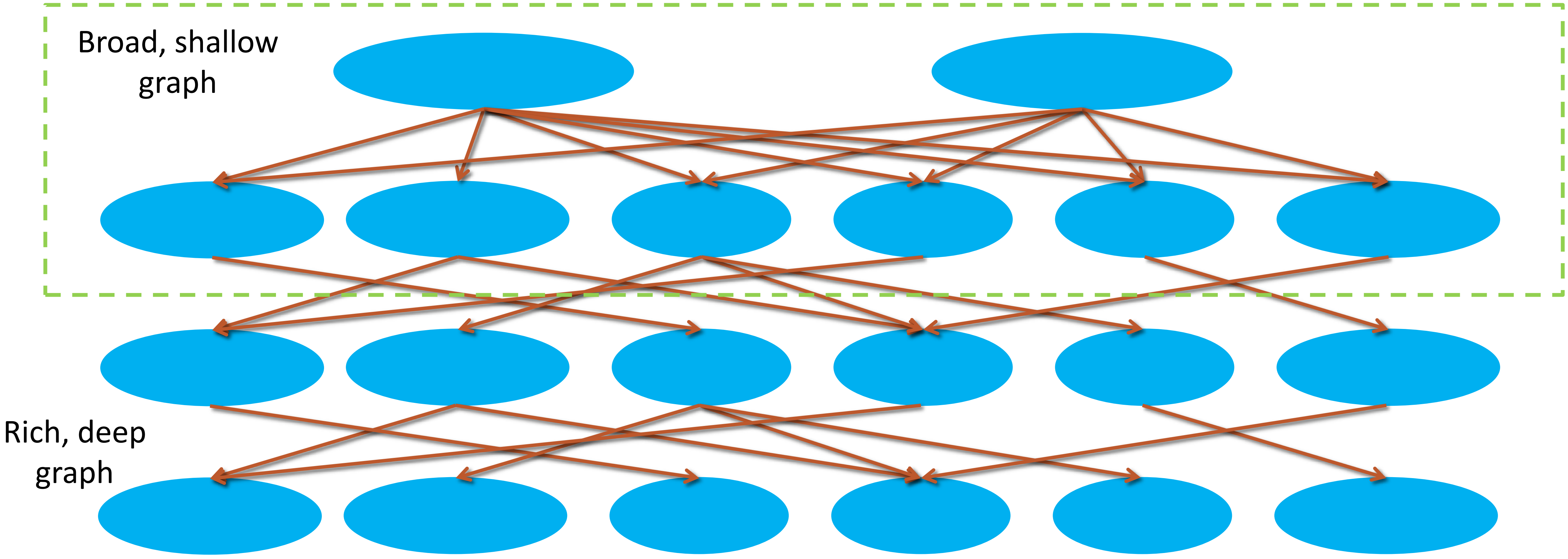
- Relation embedding filtering
  - Learn embedding from IMDb data and identify WikiData errrors

Subject	Relation	Target	Reason
The Moises Padilla Story	writtenBy	César Ámigo Aguilar	Linkage error
Bajrangi Bhaijaan	writtenBy	Yo Yo Honey Singh	Wrong relationship
Piste noire	writtenBy	Jalil Naciri	Wrong relationship
Enter the Ninja	musicComposedBy	Michael Lewis	Linkage error
The Secret Life of Words	musicComposedBy	Hal Hartley	Cannot confirm
...	...	...	...

Broad Graph Summary

- ❑Quick development: ~15 person X month
- ❑Easy ontology design and extension: 2W design
- ❑Higher coverage: +8%~+68%
- ❑Web extraction: +7% new triples, +10% new predicates, with 99% extraction accuracy
- ❑Higher accuracy: +2%~+22%
- ❑Live on Alexa Shopping, being deployed in Amazon Search and Browse

From Broad Graph to Rich Graph



Are We There Yet?

Success Criteria

- Success criteria for Product Graph
  - #Papers?
  - Team size?
  - #Production applications?
  - $$?
- **Becoming part of people's daily lives**

Knowledge Graph Construction Requires Data Integration & Cleaning (DI & DC)?

Hierarchy of thousands of types

One vertical,
A few sources

Thousands-to-millions of sources

Effective search, mining and analysis

Big challenge: Limited training labels for large-scale, rich data

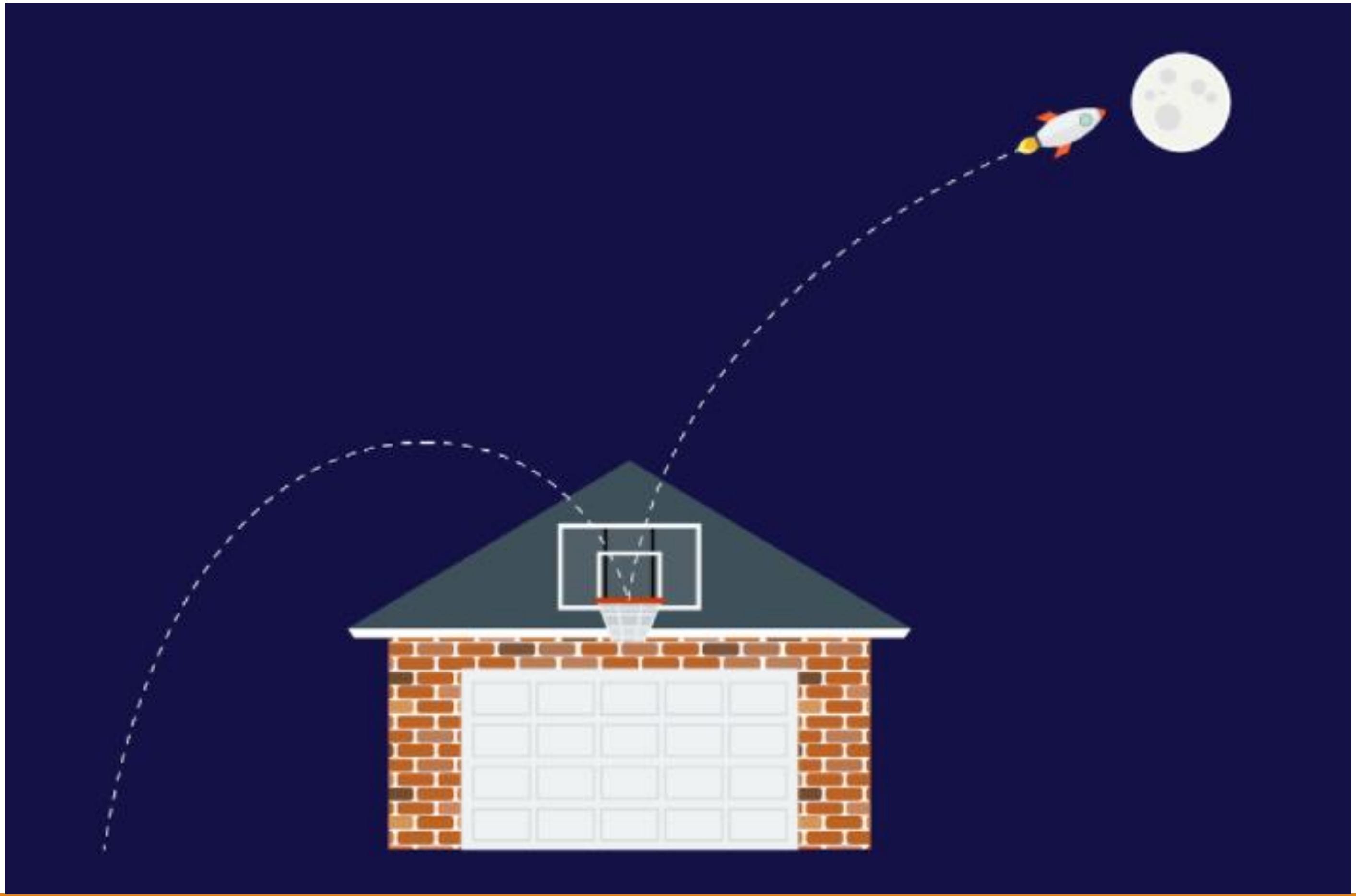
How to Get DI & DC to the Next Level of Success?

- Challenges: Limited training labels for large-scale, rich data
- Solution: Unsupervised learning ✗

How to Get DI & DC to the Next Level of Success?

- Challenges: Limited training labels for large-scale, rich data
- Solution: Learning with limited labels ✓
  - Active learning
  - Weak learning (e.g., distance supervision, data programming)
  - Semi-supervised learning (e.g., graph-based learning)
  - Transfer learning
  - One/few-shot learning

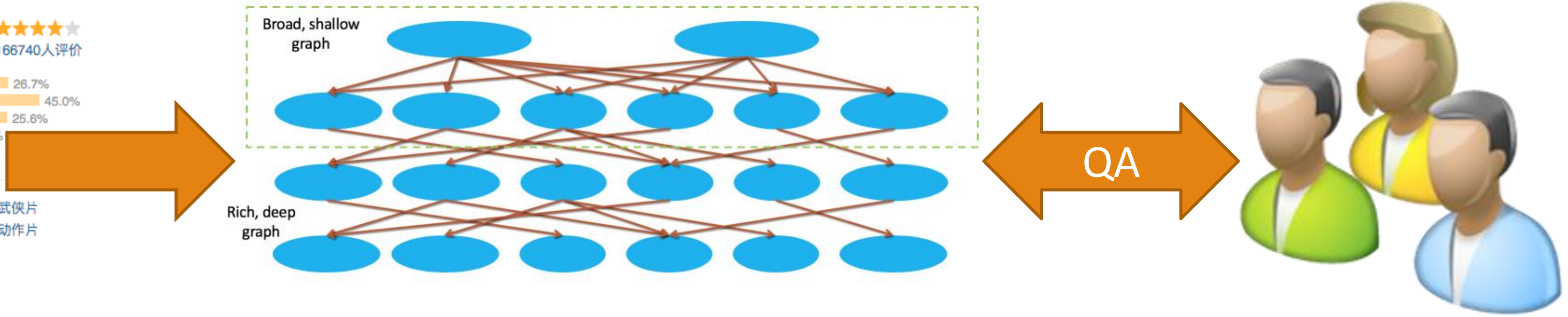
Research Philosophy



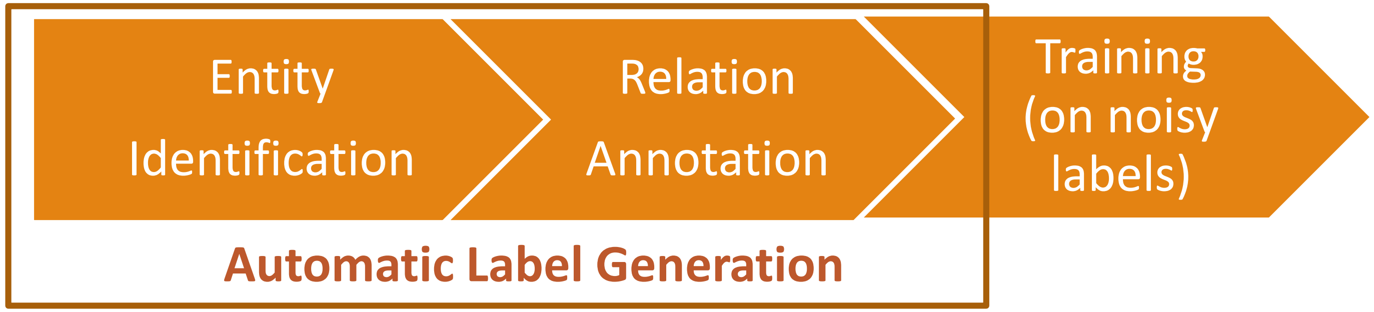
Moonshots: Strive to apply and invent the state-of-the-art

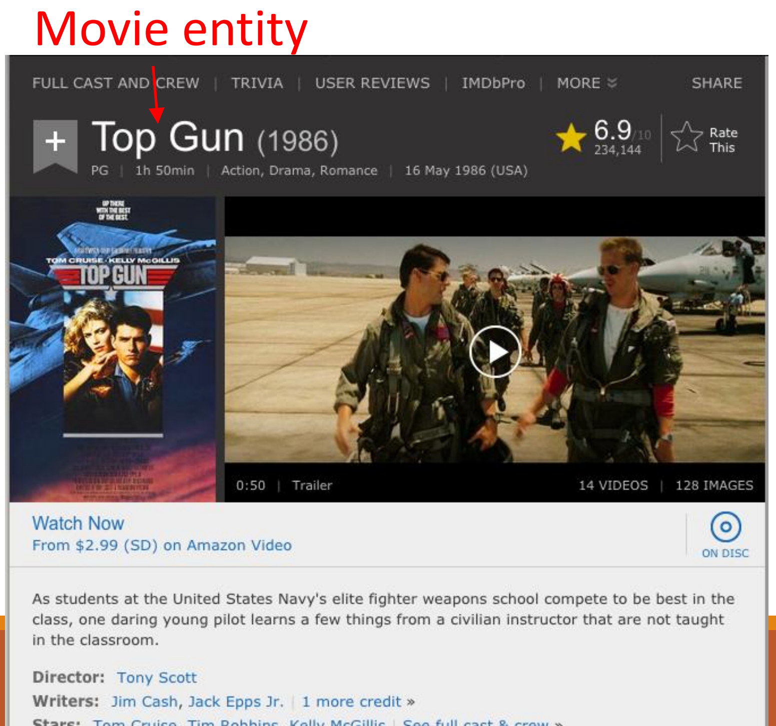
Roofshots: Deliver incrementally and make production impacts

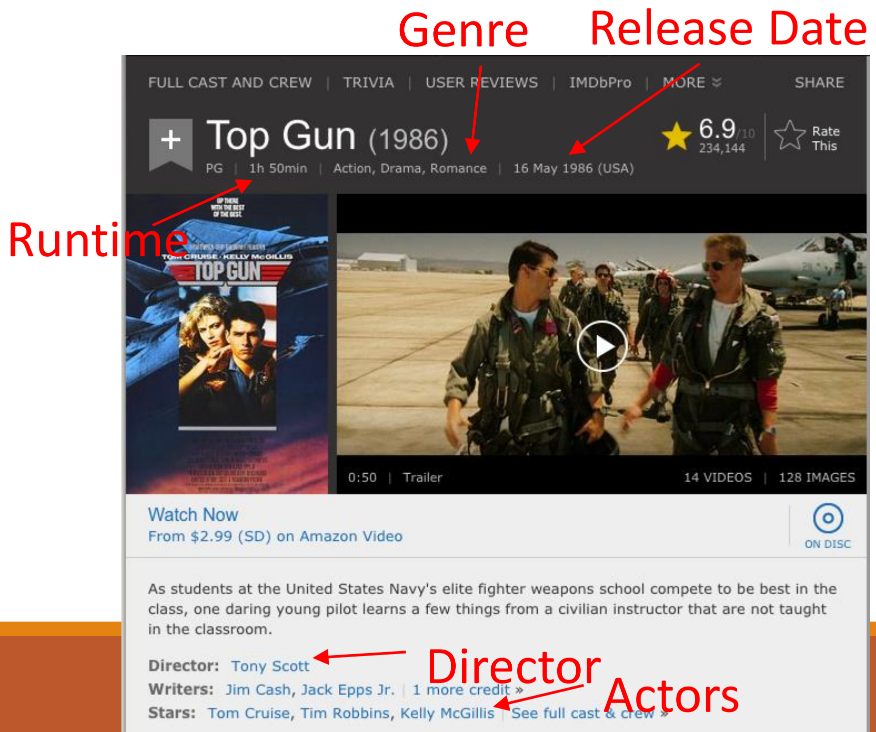
Moonshot: Open Knowledge Extraction and QA from Semi-Structured Web



Ceres: Automatic ClosedIE from Semi-structured Web [VLDB'18]







Extracted triples

- (Top Gun, type.object.name, "Top Gun")
- (Top Gun, film.film.genre, Action)
- (Top Gun, film.film.directed_by, Tony Scott)
- (Top Gun, film.film.starring, Tom Cruise)
- (Top Gun, film.film.runtime, "1h 50min")
- (Top Gun, film.film.release_Date_s, "16 May 1986")

Ceres: Automatic ClosedIE from Semi-structured Web [VLDB'18]

Weak learning

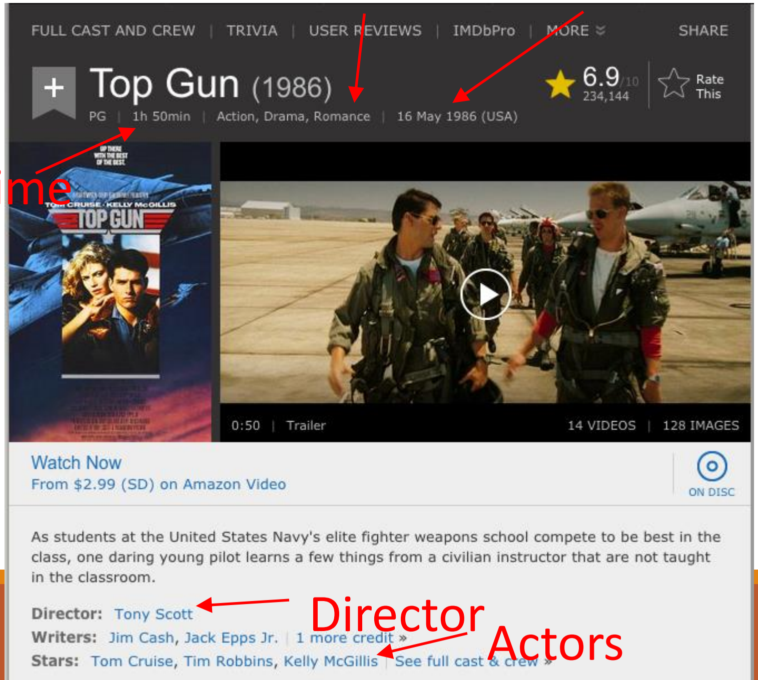
Training (on noisy labels)

Movie entity

Genre

Release Date

Runtime



Director

Actors

Extracted triples

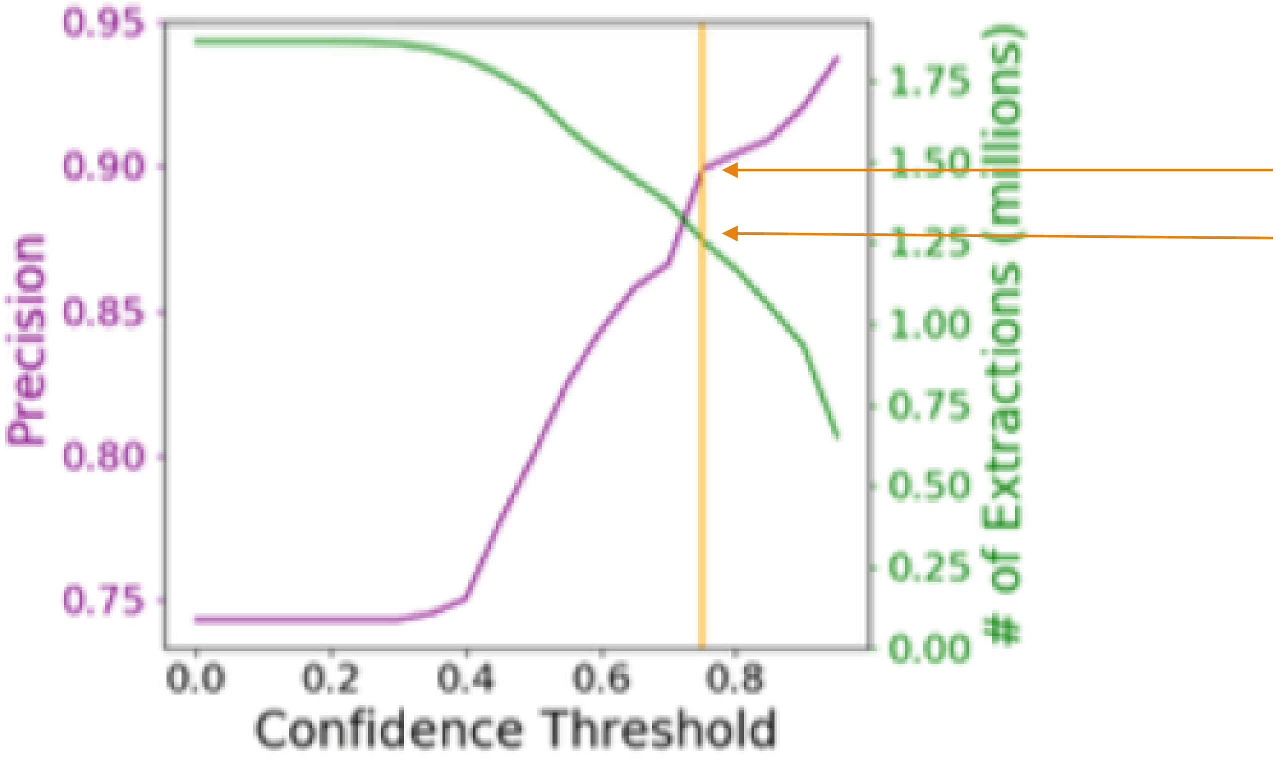
- (Top Gun, type.object.name, "Top Gun")
- (Top Gun, film.film.genre, Action)
- (Top Gun, film.film.directed_by, Tony Scott)
- (Top Gun, film.film.starring, Tom Cruise)
- (Top Gun, film.film.runtime, "1h 50min")
- (Top Gun, film.film.release_Date_s, "16 May 1986")

Ceres: Automatic ClosedIE from Semi-structured Web [VLDB'18]

#Websites / #Webpages	**33 / 434K**
Language	English and 6 other languages
Domains	Animated films, Documentary films, Financial performance, etc.
# Annotated pages	70K (16%)
Annotated : Extracted #entities	1 : **2.6**
Annotated : Extracted #triples	1 : **3.0**
# Extractions	1.25 M
Precision	90%

Ceres: Automatic ClosedIE from Semi-structured Web [VLDB'18]

- Extraction on 33 movie websites with 7 languages

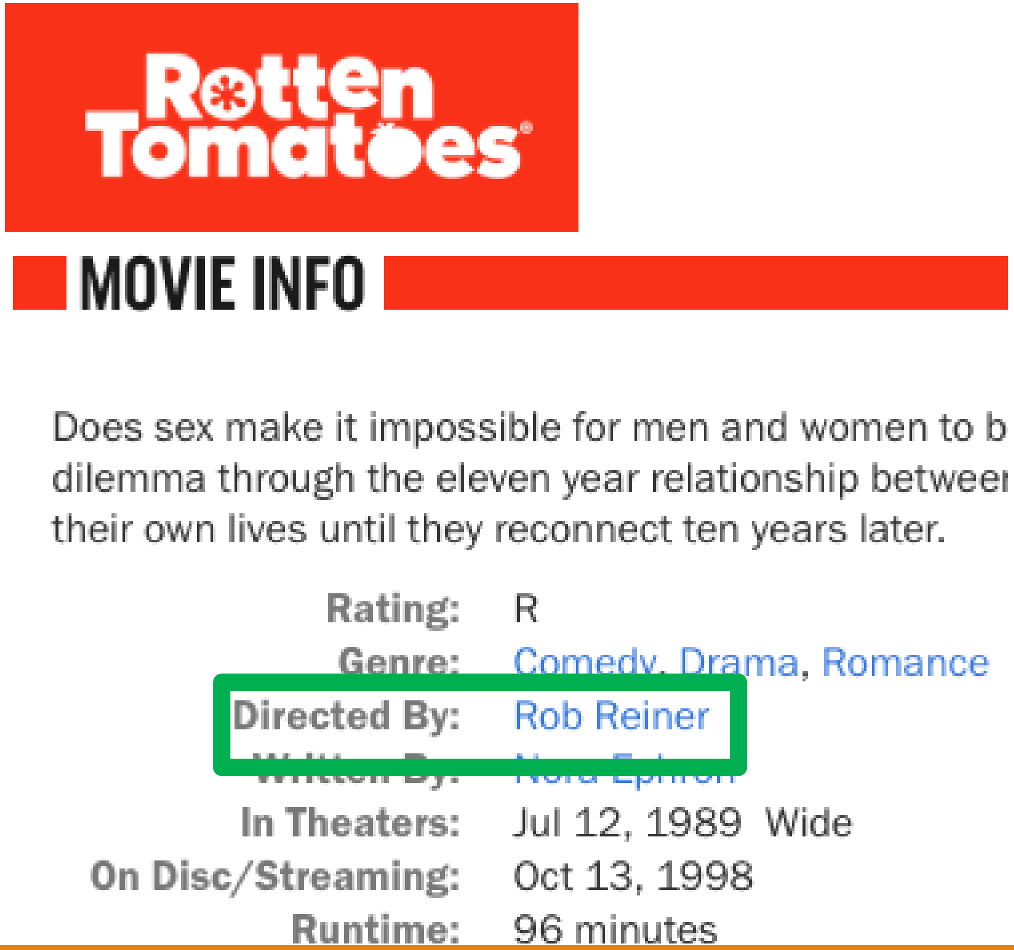


OpenCeres: OpenIE to Identify New Predicates [NAACL'19]

- ClosedIE: Only extracting facts corresponding to ontology
  - ("When Harry Met Sally...", film.film.directed_by, "Rob Reiner")
- OpenIE: Extract all relations expressed on the webpage
  - ("When Harry Met Sally...", "Director", "Rob Reiner")

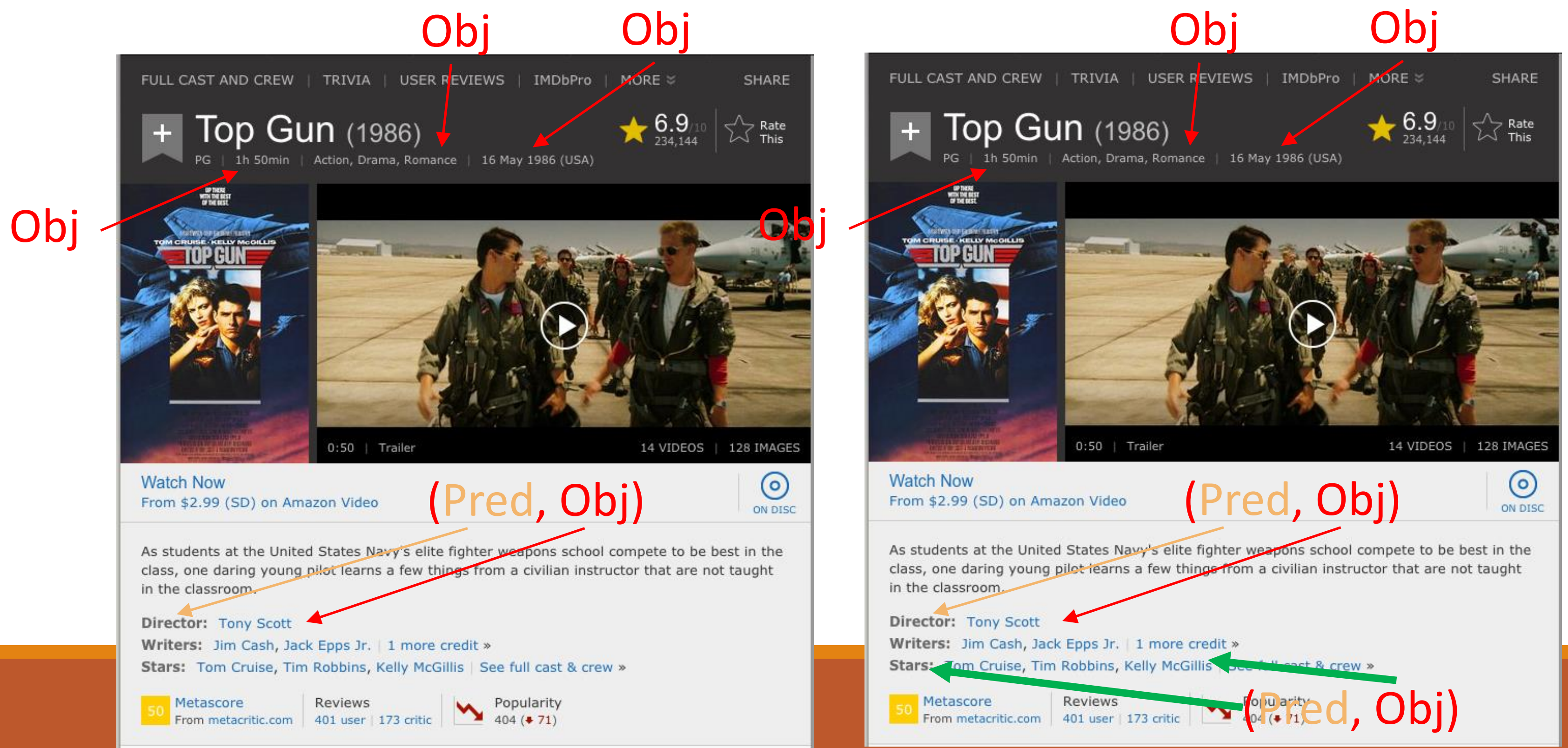
OpenCeres: OpenIE to Identify New Predicates [NAACL'19]

- ClosedIE: Normalize predicates by ontology
  - ("When Harry Met Sally…", film.film.directed_by, "Rob Reiner")
- OpenIE: Predicates are unnormalized strings
  - ("When Harry Met Sally…", "Directed By", "Rob Reiner")



OpenCeres: OpenIE to Identify New Predicates [NAACL’19]

Auto (Pred, Obj) Annotation → Label Propagation → Training



Extracted triples
- (“Top Gun”, “Director”, “Tony Scott”)
- (“Top Gun”, “Writers”, “Jim Cash”)
- (“Top Gun”, “Writers”, “Jack Epps Jr.”)
- (“Top Gun”, “Stars”, “Tom Cruise”)
- (“Top Gun”, “Stars”, “Tim Robbins”)

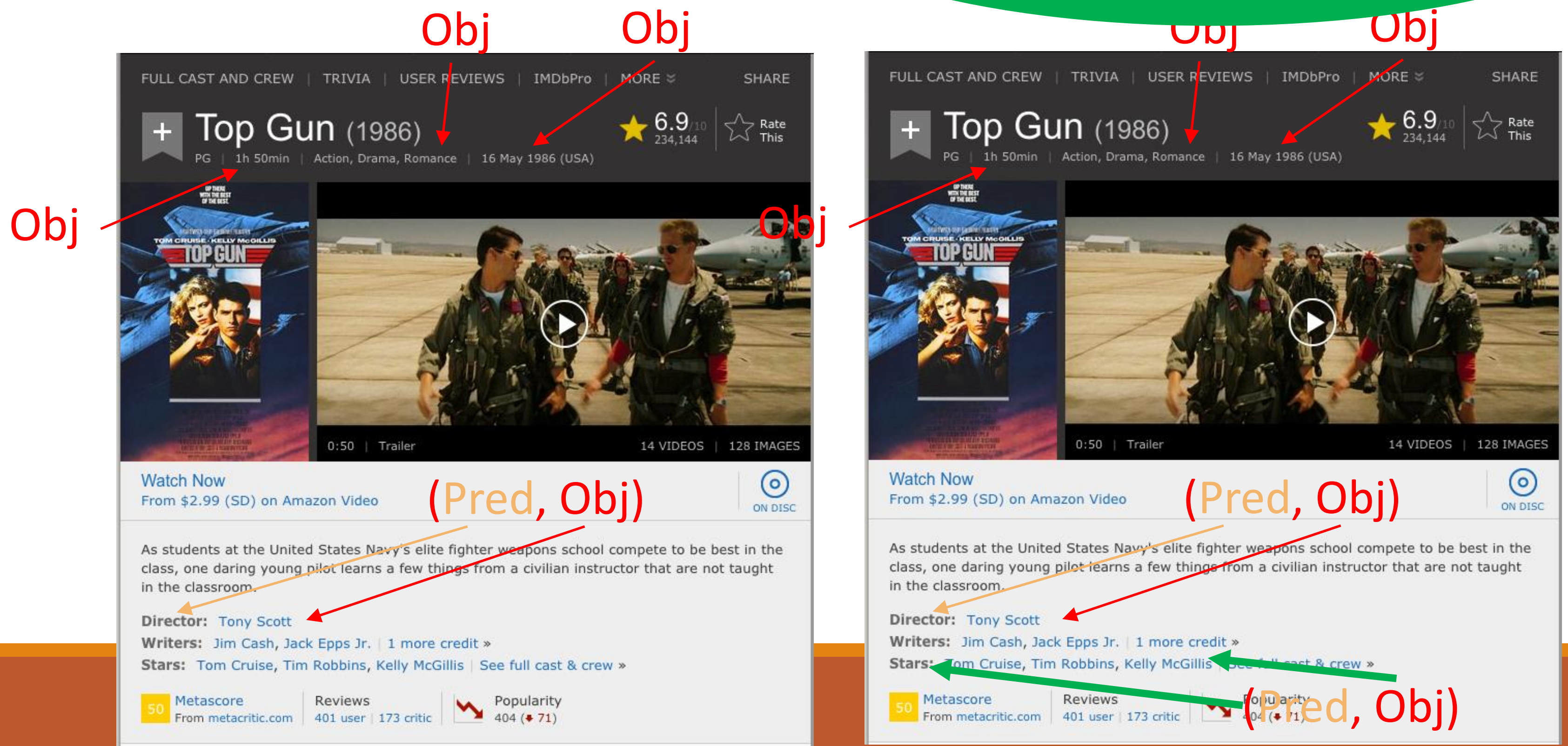
OpenCeres: OpenIE to Identify New Predicates [NAACL'19]

Auto (P

Ann

g

Semi-supervised learning



Extracted triples

- ("Top Gun", "Director", "Tony Scott")
- ("Top Gun", "Writers", "Jim Cash")
- ("Top Gun", "Writers", "Jack Epps Jr.")
- ("Top Gun", "Stars", "Tom Cruise")
- ("Top Gun", "Stars", "Tim Robbins")

OpenCeres: OpenIE to Identify New Predicates [NAACL'19]

Movie

- Seed: Director, Writer, Producer, Actor, Release Date, Genre, Alternate Title
- New: Country, Filmed In, Language, MPAA Rating, Set In, Reviewed by, Studio, Metascore, Box Office, Distributor, Tagline, Budget, Sound Mix

NBA Player

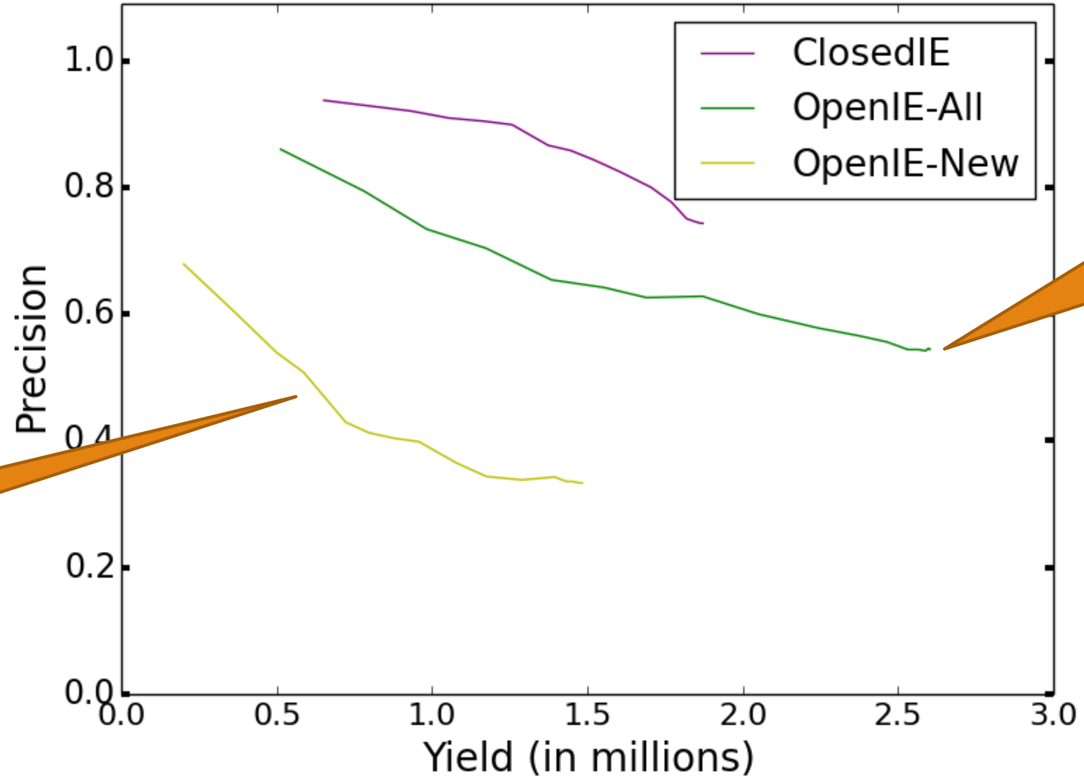
- Seed: Height, Weight, Team
- New: Birth Date, Birth Place, Salary, Age, Experience, Position, College, Year Drafted

University

- Seed: Phone Number, Web address, Type (public/private)
- New: Calendar System, Enrollment, Highest Degree, Local Area, Student Services, President

OpenCeres: OpenIE to Identify New Predicates [NAACL'19]

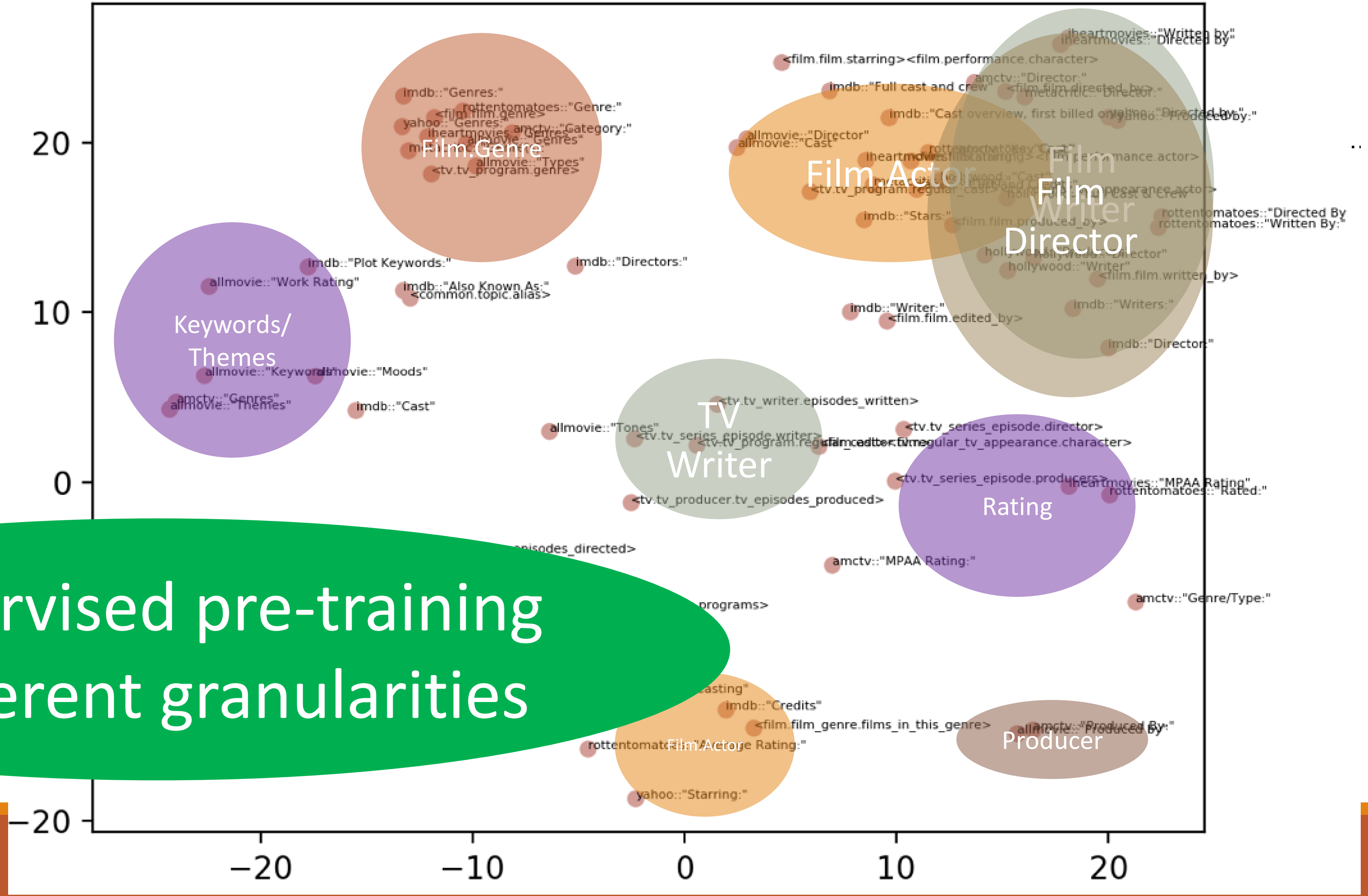
- Extraction on long-tail movie websites



OpenIE added significant amount of knowledge

Still need prec improvement on new relations

OpenKI: Universal Schema to Cluster OpenIE Predicates [NAACL'19]



Unsupervised pre-training for different granularities

QA Preliminary Results: Question→Predicate

- Easy questions we get right
  - “How many calories are there in X” -> caloriesPerServing
  - “Can I eat X while on a low carbohydrates diet?” -> carbsPerServing
  - “Is X nonGMO?” -> isGMOFree
- Hard questions we get right
  - “Is X OK for people with celiac disease” -> isGlutenFree
  - “Was X grown without pesticides” -> IsOrganic
  - “How long can I store X” -> shelfLife
  - “How many ounces of X do I get” -> Weight
  - “Is X a Halal food” -> IsKosher

QA Preliminary Results: Questions→Question

is it vegetarian ?
is it vegetarian friendly
is this vegetarian friendly
is it vegetarian healthy
is vegetarian considered raw
is this contain msg
does it contain gelatine
does it contain gelatine ?
does it contain meat ?

are the ingredients all vegetarian ?
are the products any vegetarian
are any contain vegetarian ?
does it contain msg ?
does it contain gelatine ?
does it contain meat ?

were the tomatoes grown without pesticides ?
were the seed been grown irradiated ?
have the seed been chemically irradiated ?
have the seed been pesticide tested ?
have the honey been irradiated tested ?
does the seed contain sulfites ?
Is the fennel usda irradiated ?
does your fennel seeds irradiated
does these fennel seeds irradiated
are the seeds organic ?
are the seeds organic
are the tomatoes organic ?

QA Preliminary Results: Predicate→Question phrase

isGlutenFree	contains_wheat, gluten_free, gluten, fermented_soy, soy
packageSize	how_many_ounces, bulk_package, 1_lb, 10_oz, ounces, lb, bulk, oz, package
flavor	taste_like, taste, flavored, flavors
kosher	kosher_certified, kosher, halal, certified_kosher
fatPerServing	saturated_fat, saturated, fat
isVegetarian	suitable_for_vegetarians, meat, vegetarians, vegetarian

Unsupervised pre-training

StarFinder: Finding Entity Importance [Submission]

Top 15 Among 1MM Artists

Score	Name
8262	Johann Sebastian Bach
7039	Wolfgang Amadeus Mozart
5028	Ludwig van Beethoven
3591	Frank Sinatra
3118	Elvis Presley
3089	Herbert von Karajan
3052	Frederic Chopin
3001	John Williams
2944	Antonio Vivaldi
2746	Manfred Eicher
2638	Norman Granz
2616	Grateful Dead
2449	Ella Fitzgerald

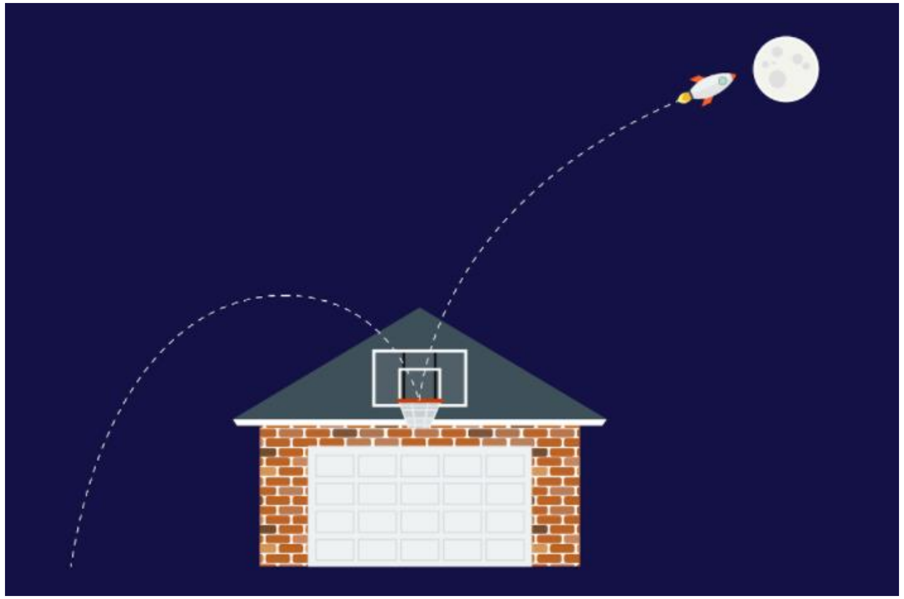
Exploring some other nodes

Rank	**Score**	**Name**
228	726	hans zimmer
267	668	eminem
282	650	michael jackson
369	559	beatles
632	412	cher
678	398	taylor swift
218		

Semi-supervised learning with GCN

Take Aways

- We aim at building an authoritative knowledge graph for all products in the world
- We shoot for roofshot and moonshot goals to realize our mission
- The next-generation of KG could be a combination of rich graph and broad graph
- Learning with limited labels is the rescue to get DI & DC to the next level of success



Thank You!

QUESTIONS?