



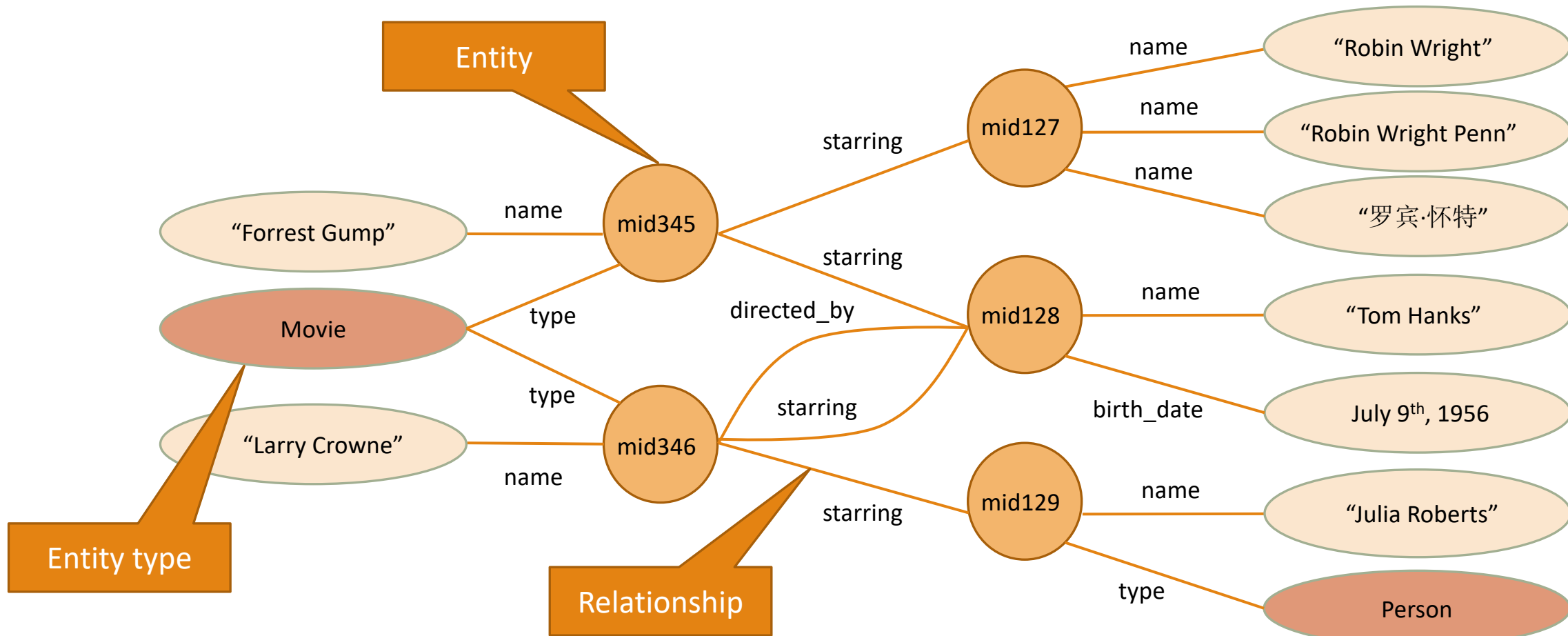
Challenges and Innovations in Building a Product Knowledge Graph

XIN LUNA DONG, AMAZON

JUNE, 2017


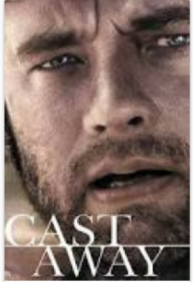







Product Graph vs. Knowledge Graph

Knowledge Graph Example for 2 Movies



Knowledge Graph in Search

Tom Hanks > Movies

| | | | | | | | | |
|--|---|--|--|---|---|---|---|---|
|  <p>Forrest Gump 1994</p> |  <p>Cast Away 2000</p> |  <p>Saving Private Ryan 1998</p> |  <p>Captain Phillips 2013</p> |  <p>Big 1988</p> |  <p>Sully 2016</p> |  <p>Toy Story 1995</p> |  <p>Catch Me If You Can 2002</p> |  <p>Apollo 13 1995</p> |
|--|---|--|--|---|---|---|---|---|

List of Tom Hanks performances - Wikipedia

https://en.wikipedia.org/wiki/List_of_Tom_Hanks_performances

Jump to **Film** - The Simpsons Movie, 2007, Yes, Himself, Cameo Voice role. Mamma Mia! 2008, Yes, —, Executive producer. City of Ember, 2008, Yes, —.

[A Hologram for the King \(film\)](#) · [Big \(film\)](#) · [Larry Crowne](#) · [He Knows You're Alone](#)

Tom Hanks (@tomhanks) · Twitter

<https://twitter.com/tomhanks>

And don't miss this songstress at the famous Cafe Carlyle. Through Saturday nite! Hanx @RitaWilson
<pic.twitter.com/J70XJbf...>

12 hours ago · [Twitter](#)

Beware! Crass self-serving Social Media Post! This book goes on sale tomorrow! Hanx
<pic.twitter.com/V2EqPKL...>

16 hours ago · [Twitter](#)

Lost (g)love. Looking for a mate. Good luck. Hanx.
<pic.twitter.com/ApH7rEG...>

1 day ago · [Twitter](#)

Tom Hanks

American actor



Thomas Jeffrey Hanks is an American actor and filmmaker. He is known for his various comedic and dramatic film roles, including Splash, Big, Turner & Hooch, A League of Their Own, Sleepless in Seattle, ... [Wikipedia](#)

Born: July 9, 1956 (age 61), [Concord, CA](#)

Awards: [Academy Award for Best Actor](#), [MORE](#)

Spouse: [Rita Wilson](#) (m. 1988), [Samantha Lewes](#) (m. 1978–1987)

TV shows: [Bosom Buddies](#), [Celebrity Jeopardy!](#), [MORE](#)



Knowledge Graph in Personal Assistant

Alexa, play the music
by Michael Jackson



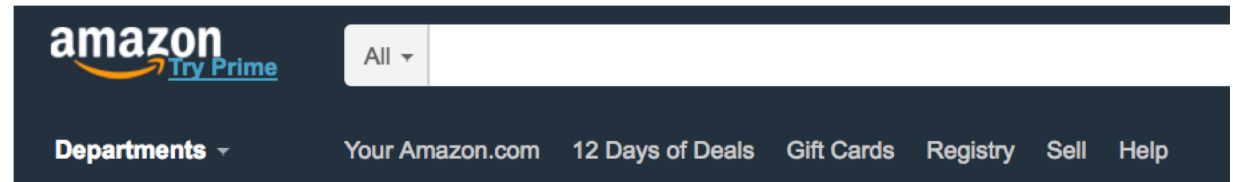
List of officially released compilations and

[92][93][94][95]

- *Portrait of Michael Jackson / Portrait of Jackson 5* (1973)
- *Os Grandes Sucessos, Vol. 2* (1980)
- *Motown Superstar Series, Vol. 7* (1980)
- *Superstar* (1980)
- *Michael Jackson & The Jackson 5* (1983)
- *Ain't No Sunshine* (1984)
- *The Great Love Songs of Michael Jackson* (1984)
- *Ben / Got to Be There* (1986)
- *Looking Back to Yesterday* (1986)
- *The Original Soul of Michael Jackson* (1987)
- *Rockin' Robin* (1993)
- *Dangerous – The Remix Collection* (1993)
- *Michael Jackson Story* (1996)
- *Master Series* (1997)
- *Ghosts – Deluxe Collector Box Set* (1997)
- *Got to Be There / Forever, Michael* (1999)
- *Bad / Thriller* (2000)
- *Forever, Michael / Music & Me / Ben* (2000)
- *Classic – The Universal Masters Collection* (2001)

Product Graph

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



Customers who bought this item also bought

<

Cars 3 Playland with 20 Balls Playset
★★★★☆ 3
\$28.55 ✓prime

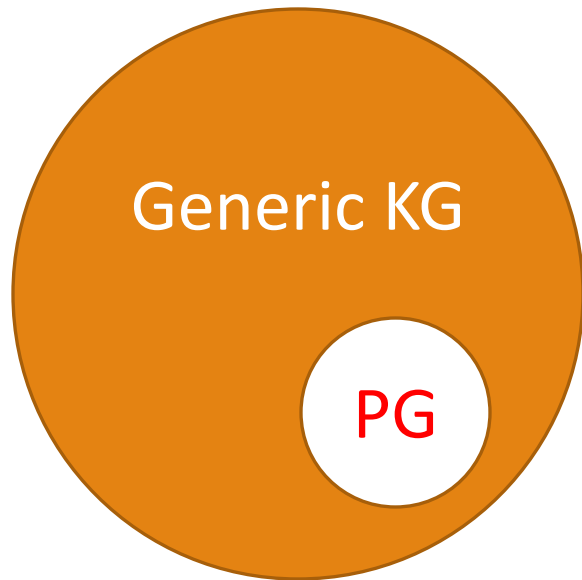
Step2 Push Around Sport Buggy
★★★★☆ 18
\$49.99 ✓prime

GOOD NIGHT, LIGHTNING
› RH Disney
★★★★☆ 299
Board book
\$7.70 ✓prime

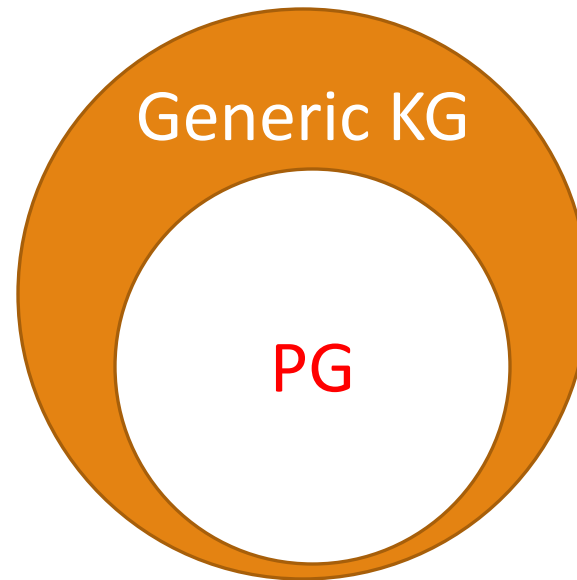


Product Graph vs. Knowledge Graph

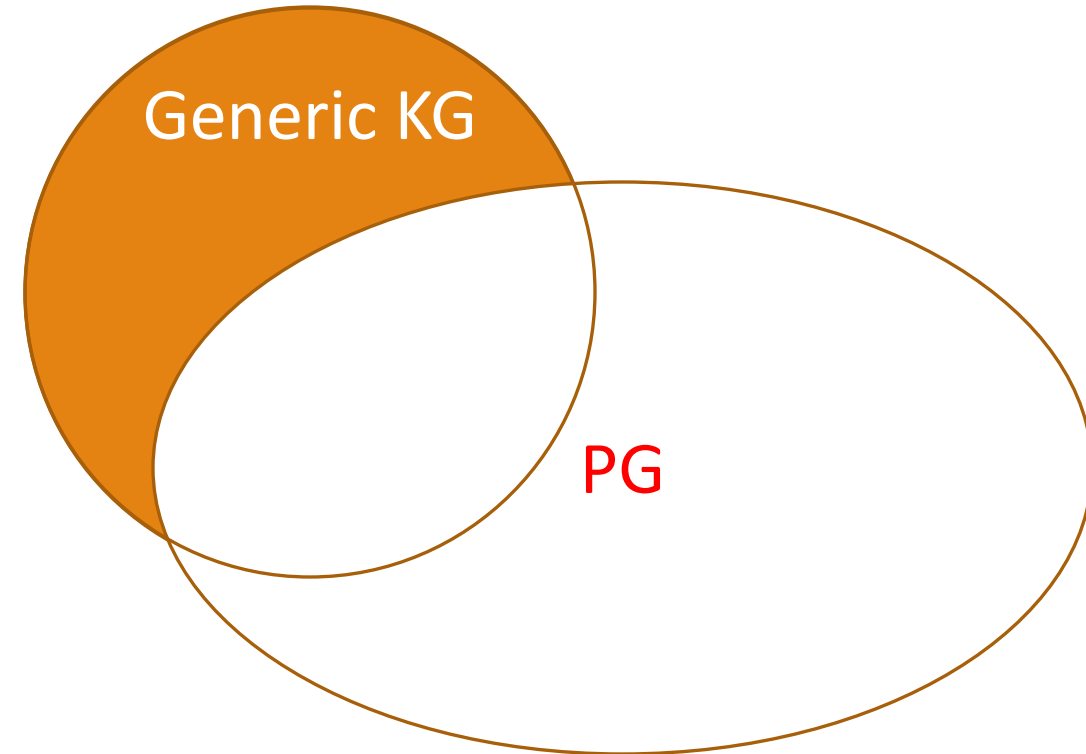
(A)



(B)

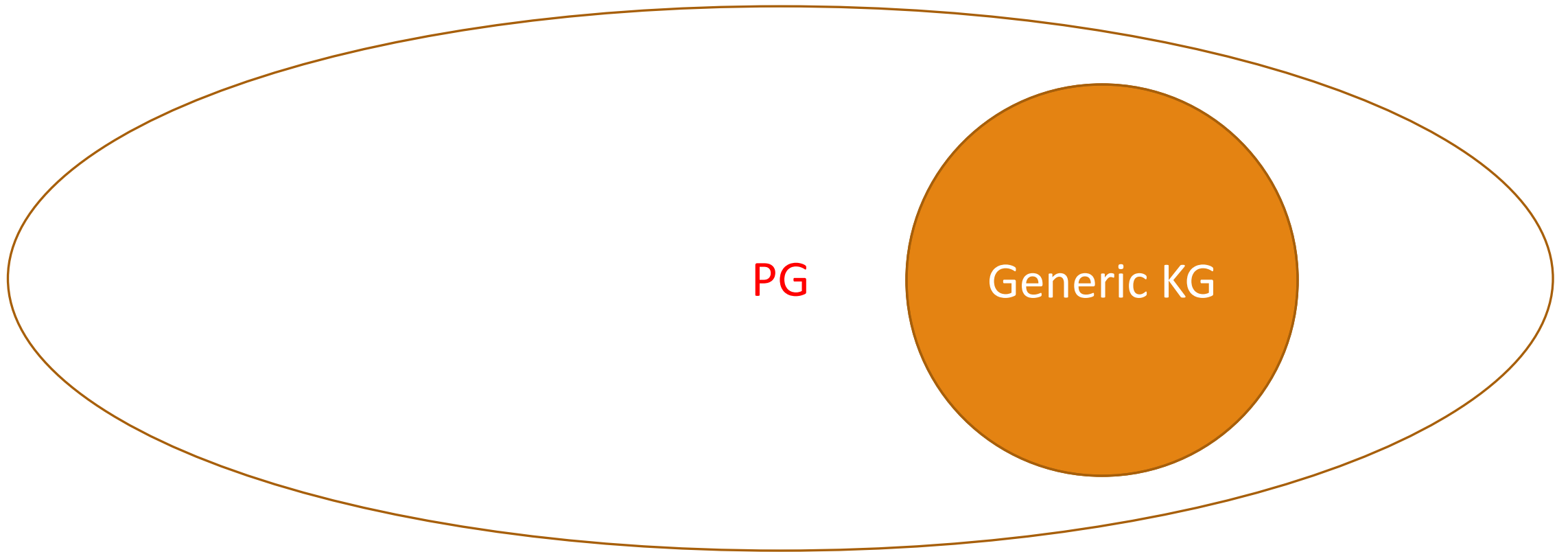


(C)

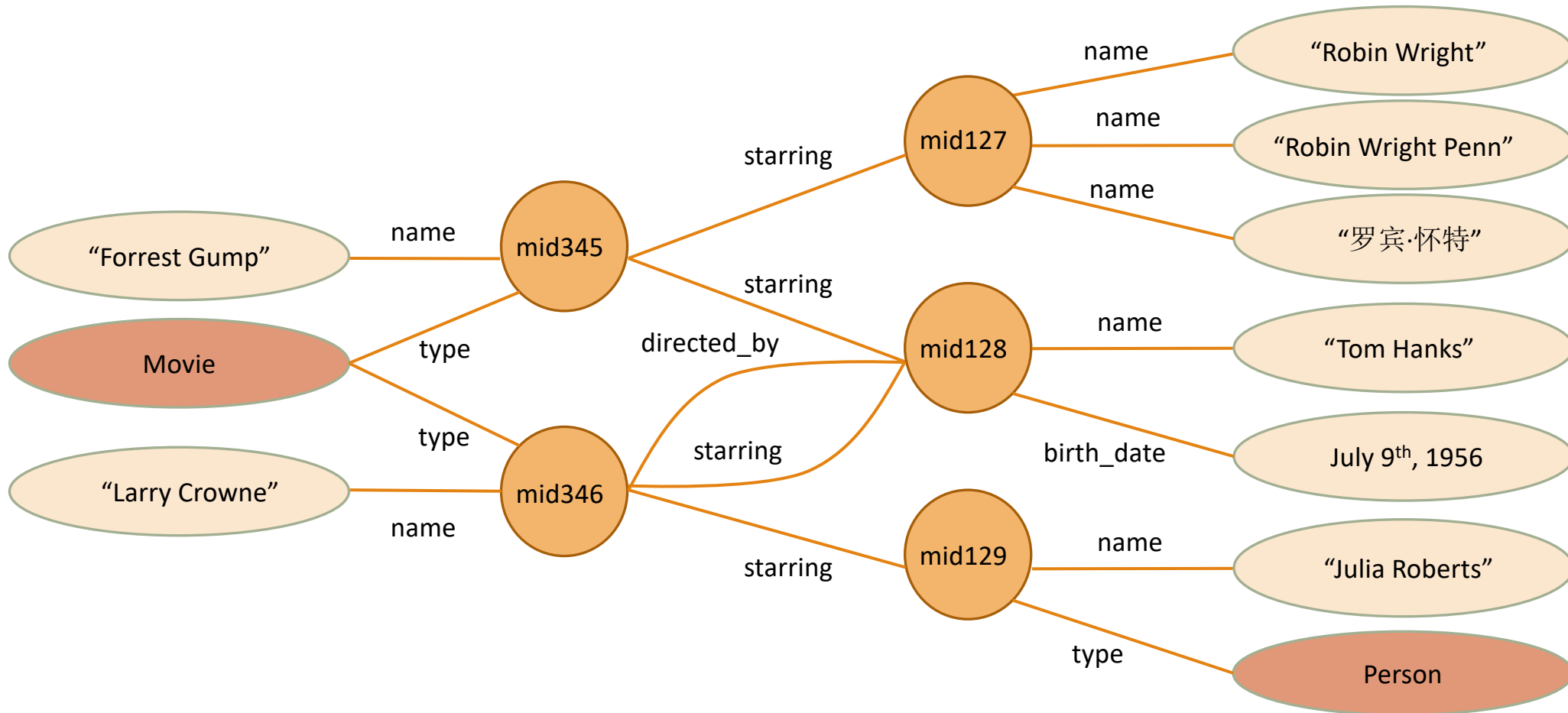


Product Graph vs. Knowledge Graph

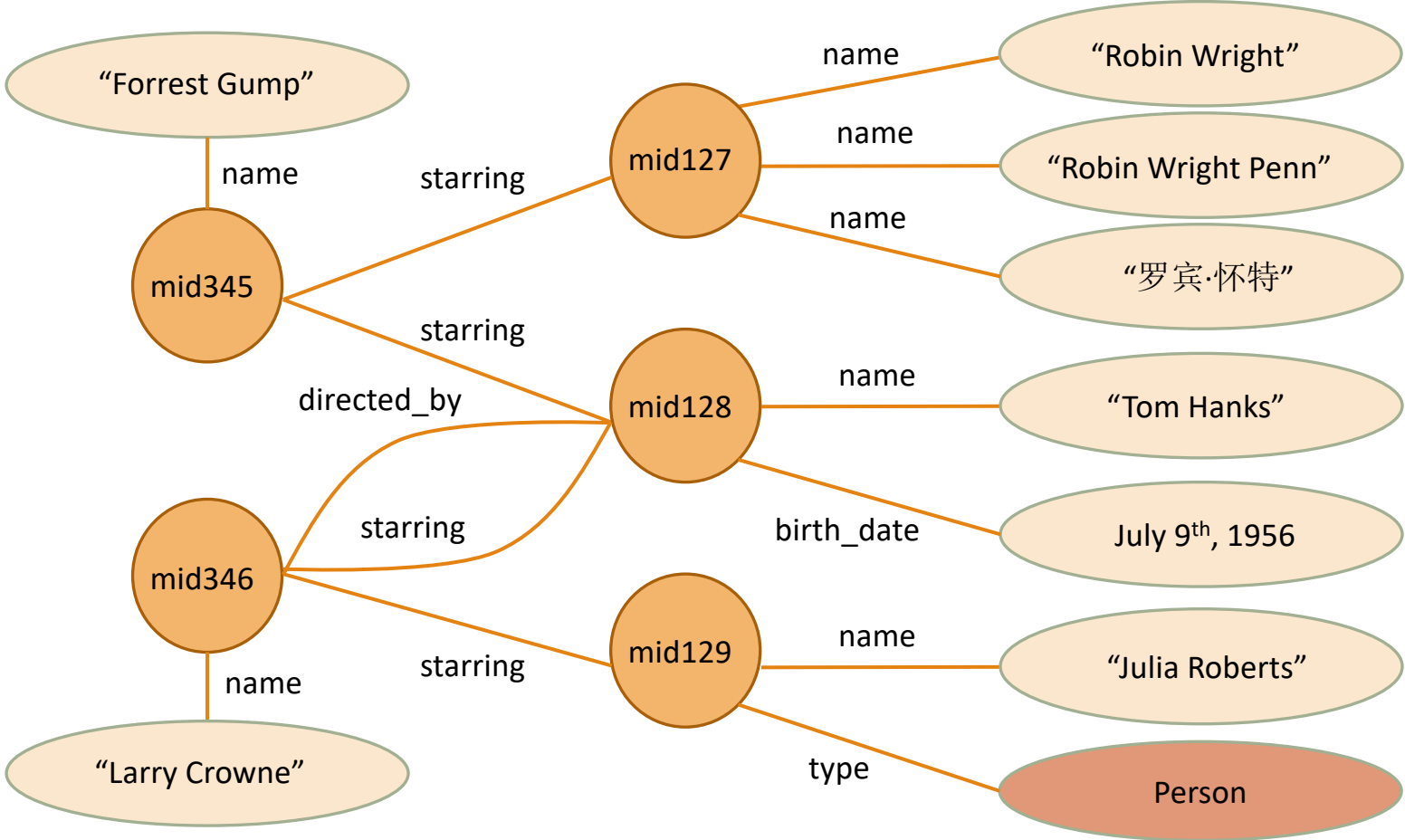
(D)



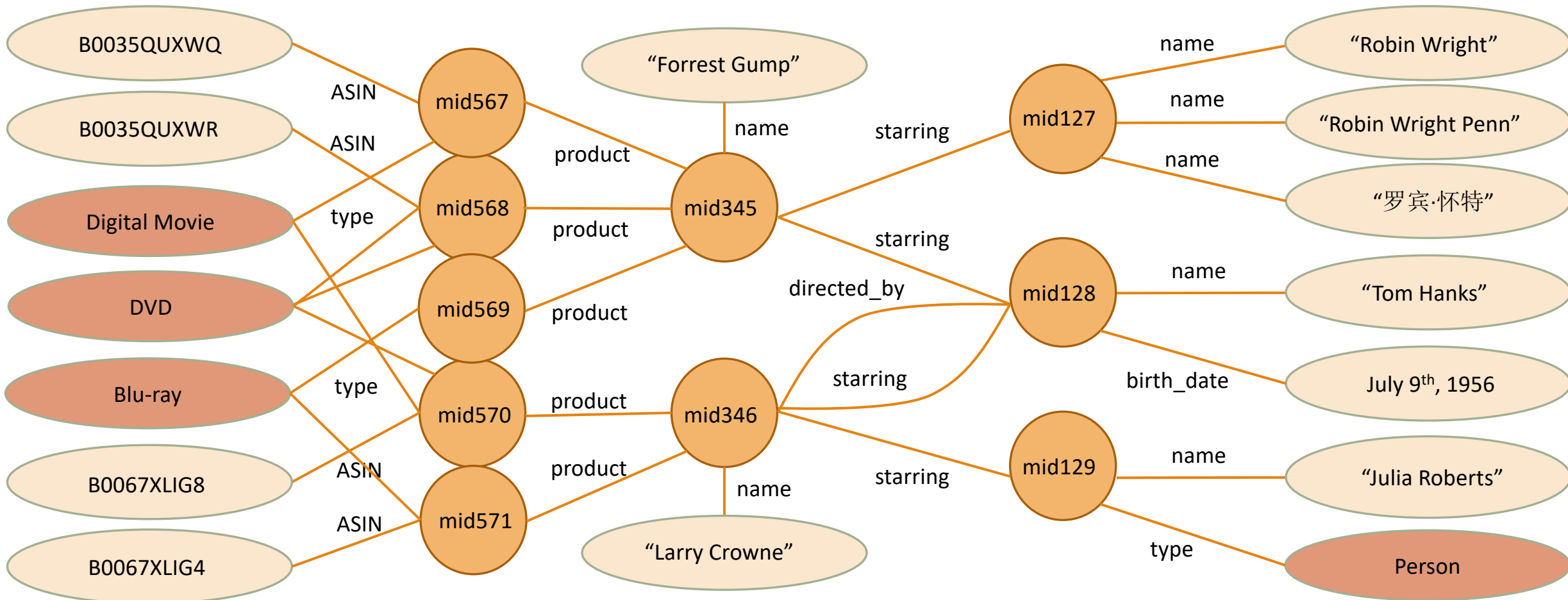
Knowledge Graph Example for 2 Movies



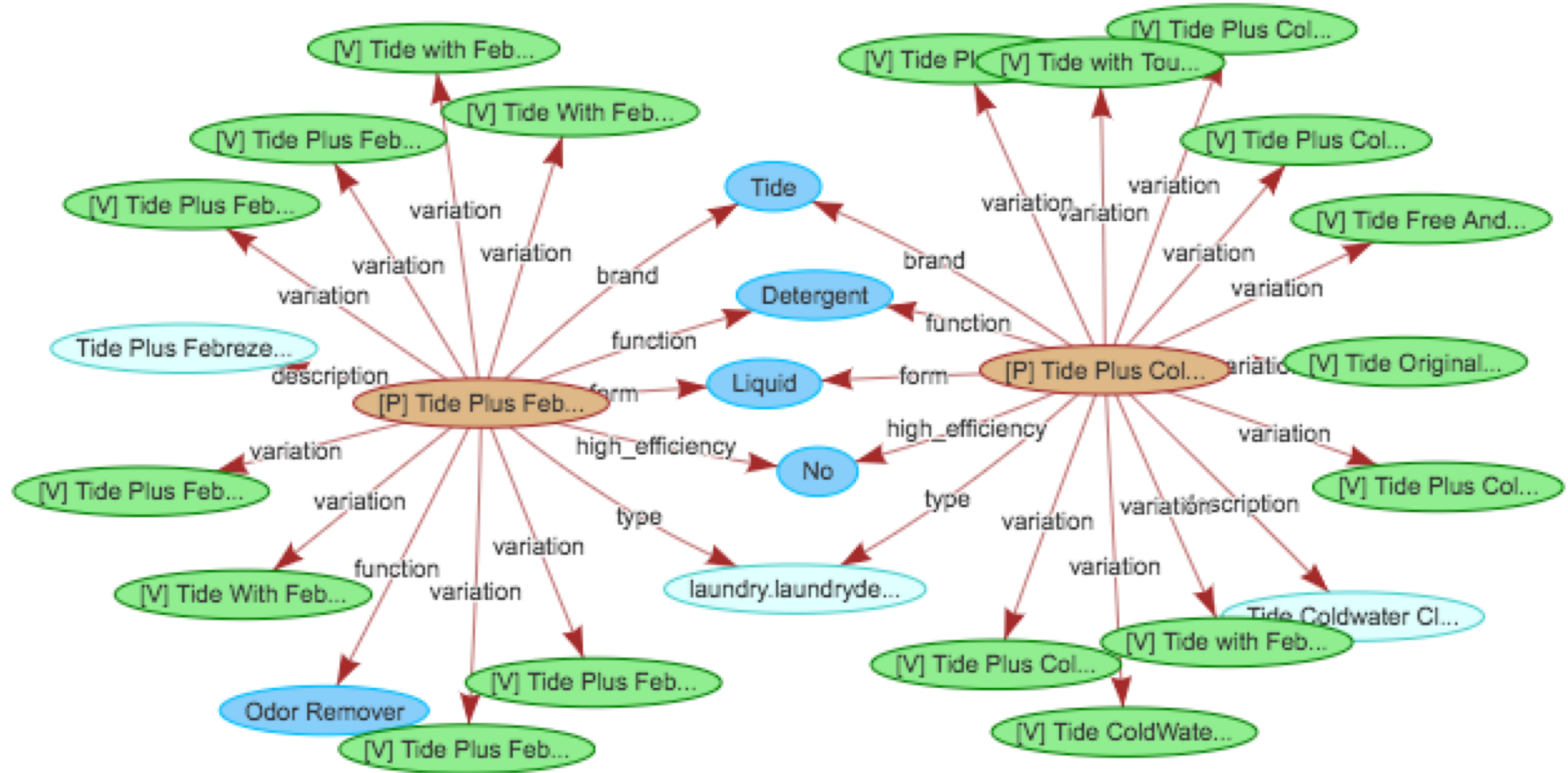
Product Graph vs. Knowledge Graph



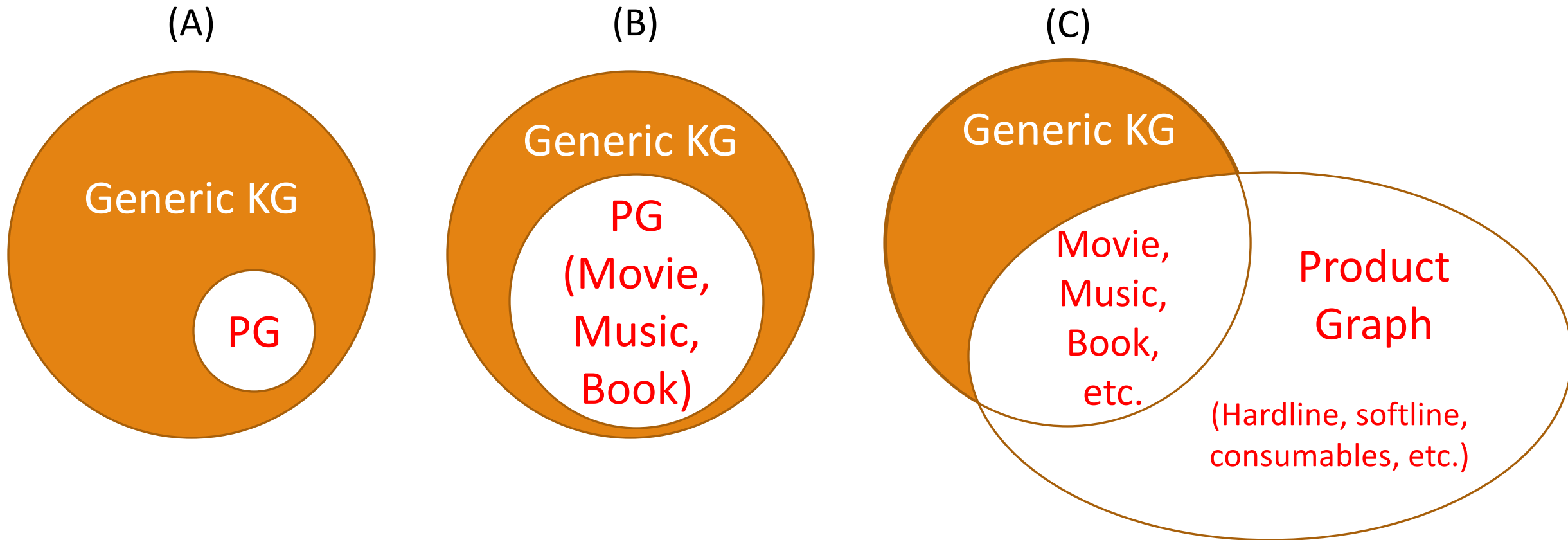
Product Graph vs. Knowledge Graph

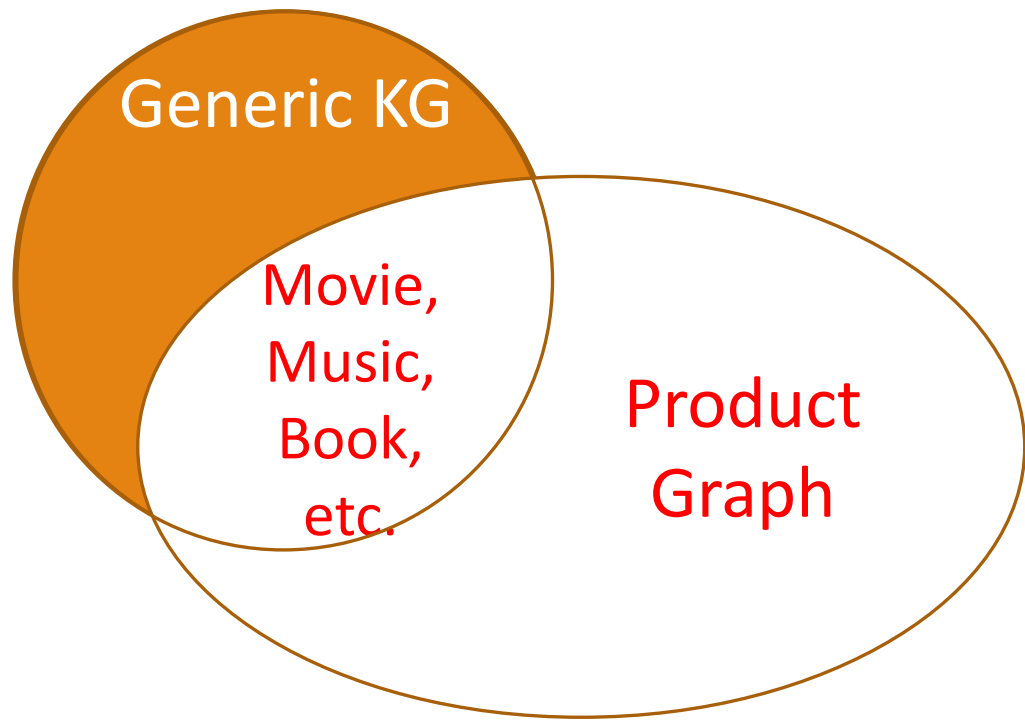


Another Example of Product Graph



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

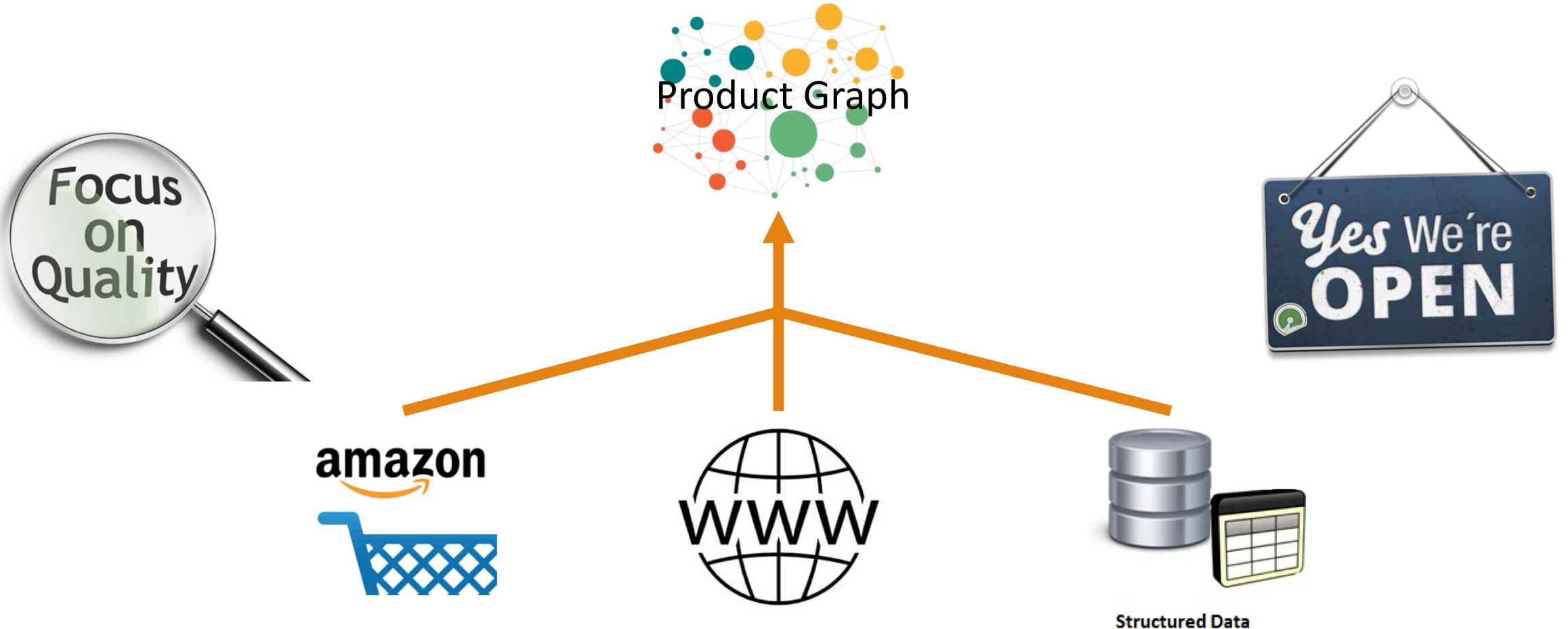
- ❑ Large number of product categories
 - ❑ A lot of work to manually define ontology
 - ❑ Hard to catch the trend of new product categories and properties

Challenges in Building Product Graph IV

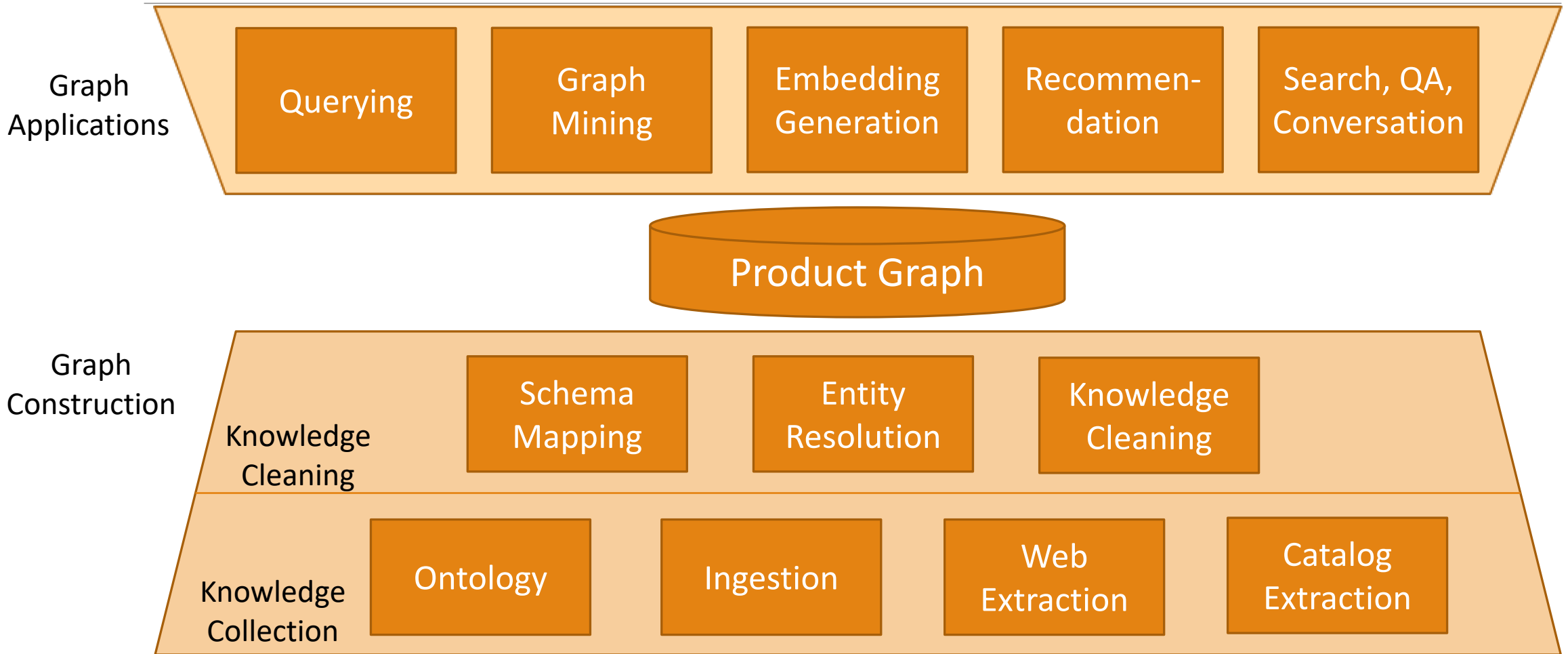
- ❑ Many entities are not named entities
 - ❑ Named Entity Recognition does not apply
 - ❑ New challenges for extraction, linking, and search

How to Build a Product Graph?

Where is Knowledge from?



Architecture



Which ML Model Works Best?



Which ML Model Works Best?

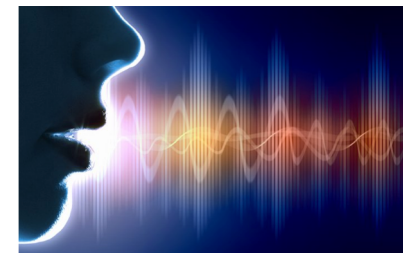
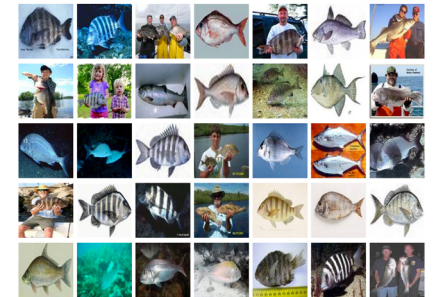
| ID | NAME | CLASS | MARK | SEX |
|----|-------------|-------|------|--------|
| 1 | John Deo | Four | 75 | female |
| 2 | Max Ruin | Three | 85 | male |
| 3 | Arnold | Three | 55 | male |
| 4 | Krish Star | Four | 60 | female |
| 5 | John Mike | Four | 60 | female |
| 6 | Alex John | Four | 55 | male |
| 7 | My John Rob | Fifth | 78 | male |
| 8 | Asruid | Five | 85 | male |
| 9 | Tes Qry | Six | 78 | male |
| 10 | Big John | Four | 55 | female |

Tree-based models



SCENE FROM "DAN'L DRUCE."

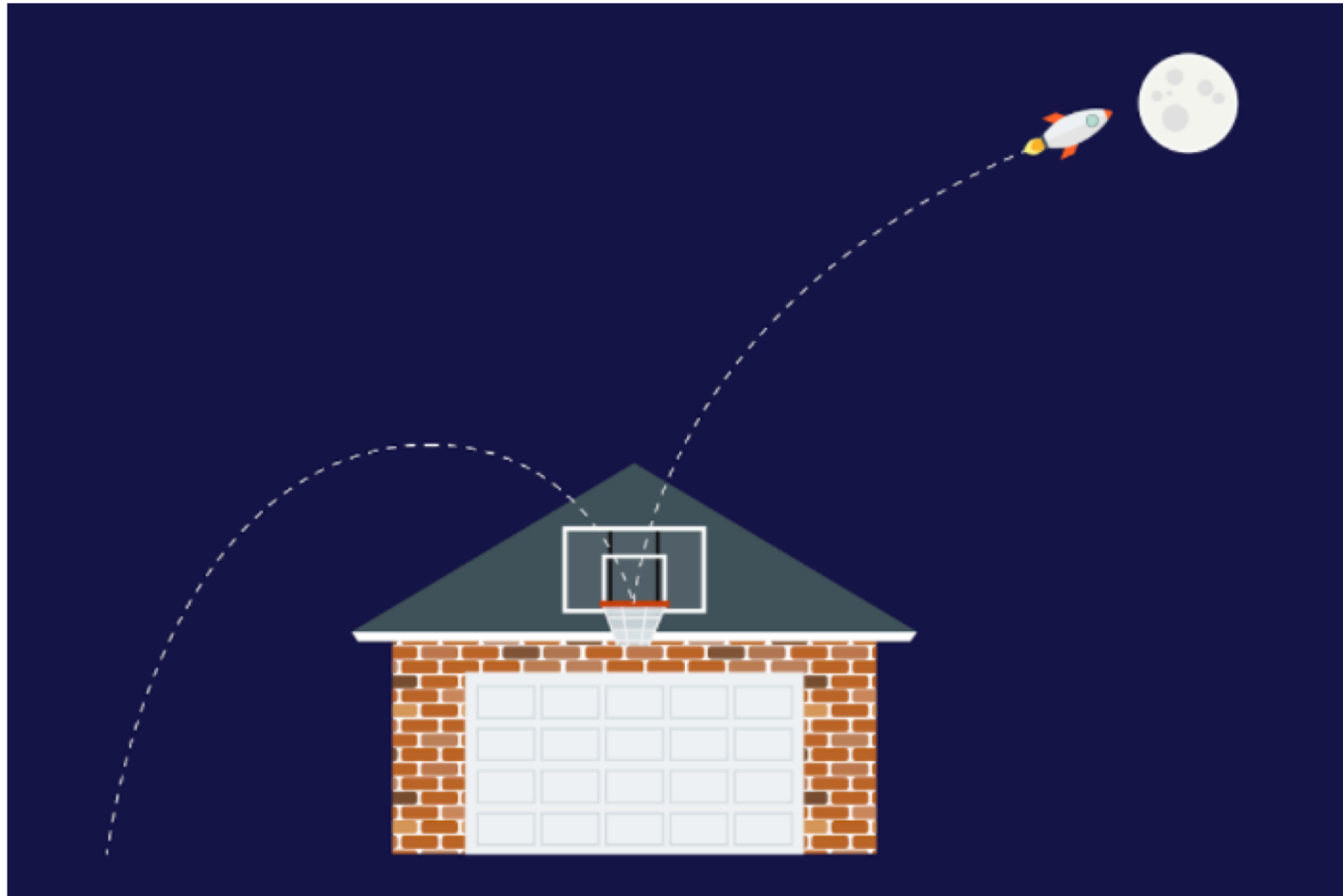
This interesting domestic drama, by Mr. W. S. Gilbert, has continued to engage the sympathies of a nightly sufficient audience at the Haymarket Theatre, where it has now been represented more than sixty times. Its subject and character were described by us, in the ordinary report of theatrical novelties, about two months ago. Our readers will probably not need to be reminded that the hero of the story, Dan'l Druce, the blacksmith, is a solitary recluse dwelling on the coast of Norfolk, where his lone cottage is visited by fugitives from party vengeance during the civil wars of the Commonwealth. His hoard of money is stolen; but a different sort of treasure, a helpless female infant, is left by some mysterious agency, and may be accepted, as in George Eliot's tale of "Silas Marner," for a Divine gift to the sad-hearted misanthrope, far better than riches. In this spirit, at least, he is content to receive the precious human charge; and so to those who would remove it from his home, Dan'l Druce here makes answer with the solemn exclamation, "Touch not the Lord's gift!" This character is well acted by Mr. Hermann Veizin.



Neural network

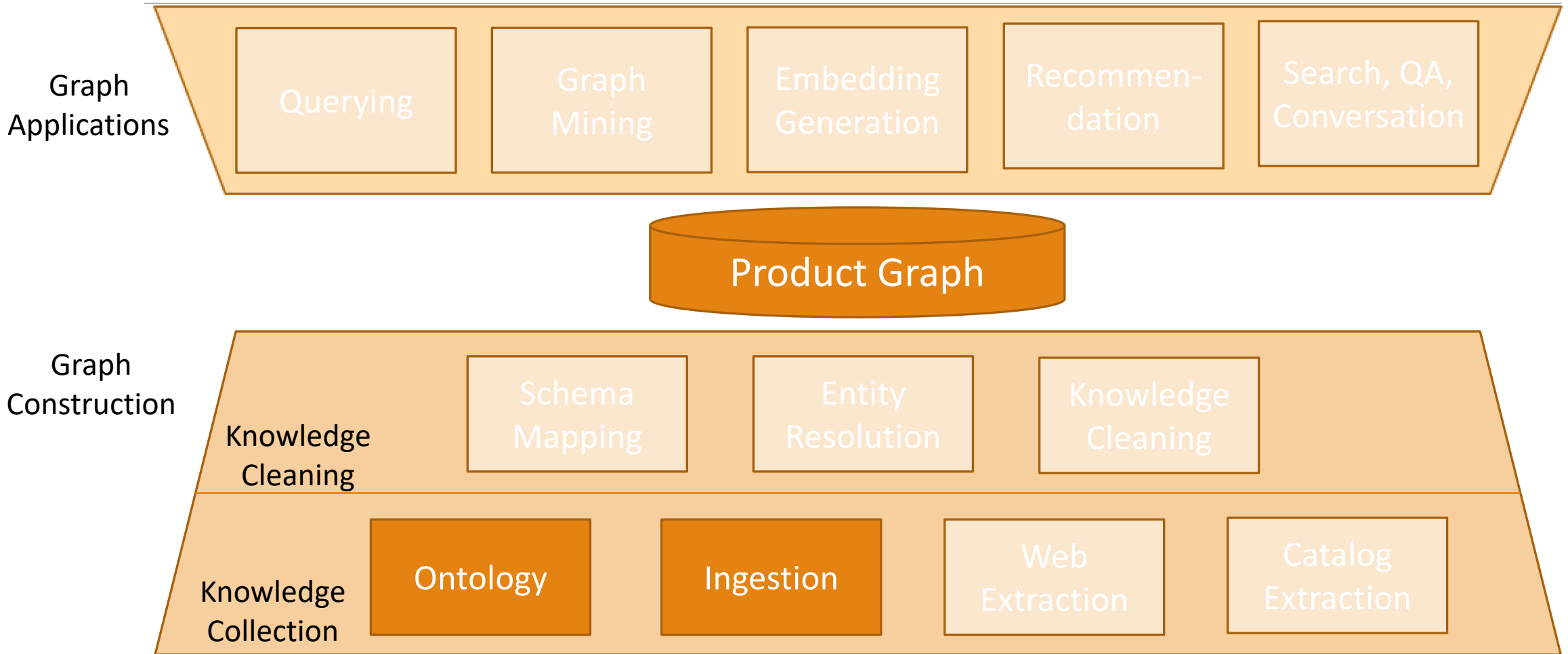
Research Philosophy

Roofshots: Deliver incrementally and make production impacts



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

I. Integrating Knowledge from Structured Sources



Challenges: Linkage & Quality

IMDB



Anahí SEE RANK
Actress | Music Department | Soundtrack

Anahi was born in Mexico. She's had roles in Tu y Yo, in which she played a 17 year old girl while she was 13, and Vivo Por Elena, in which she played Talita, a naive and innocent teenager. Anahi lives with her mother and sister name Marychelo. She hopes to become a fashion designer one day, and is currently pursuing a career in singing.
[See full bio >](#)

Born: May 14, 1982 in Mexico City, Distrito Federal, Mexico

[More at IMDbPro >](#)
[Contact Info: View manager](#)

Same entity?

WikiData

Anahí Puente (Q169461)

Mexican singer-songwriter and actress
Mia

[In more languages](#) [Configure](#)

| Language | Label | Description |
|----------|--------------|---|
| English | Anahí Puente | Mexican singer-songwriter and actress |
| Chinese | 阿纳希·普恩特 | No description defined |
| Spanish | Anahí Puente | Cantante, compositora y actriz mexicana |

date of birth 7 November 1983 [edit](#)

[1 reference](#)

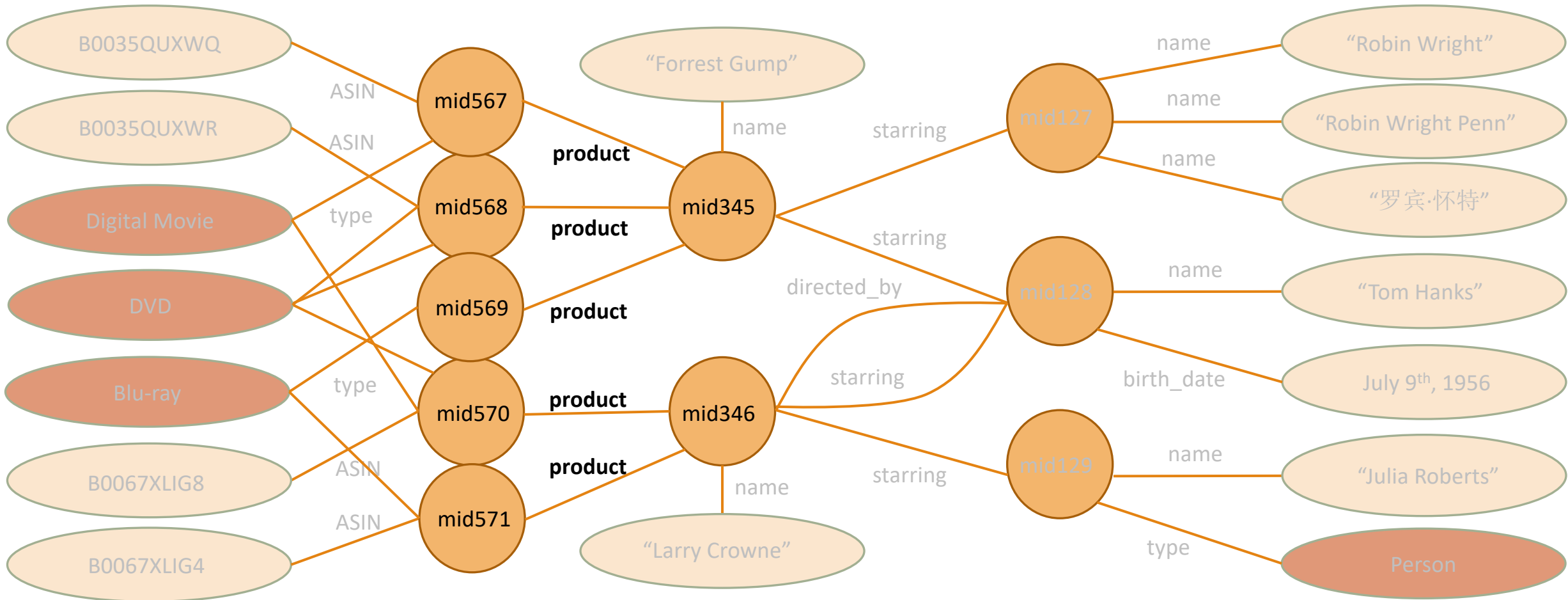
imported from [Italian Wikipedia](#)

[+ add reference](#)

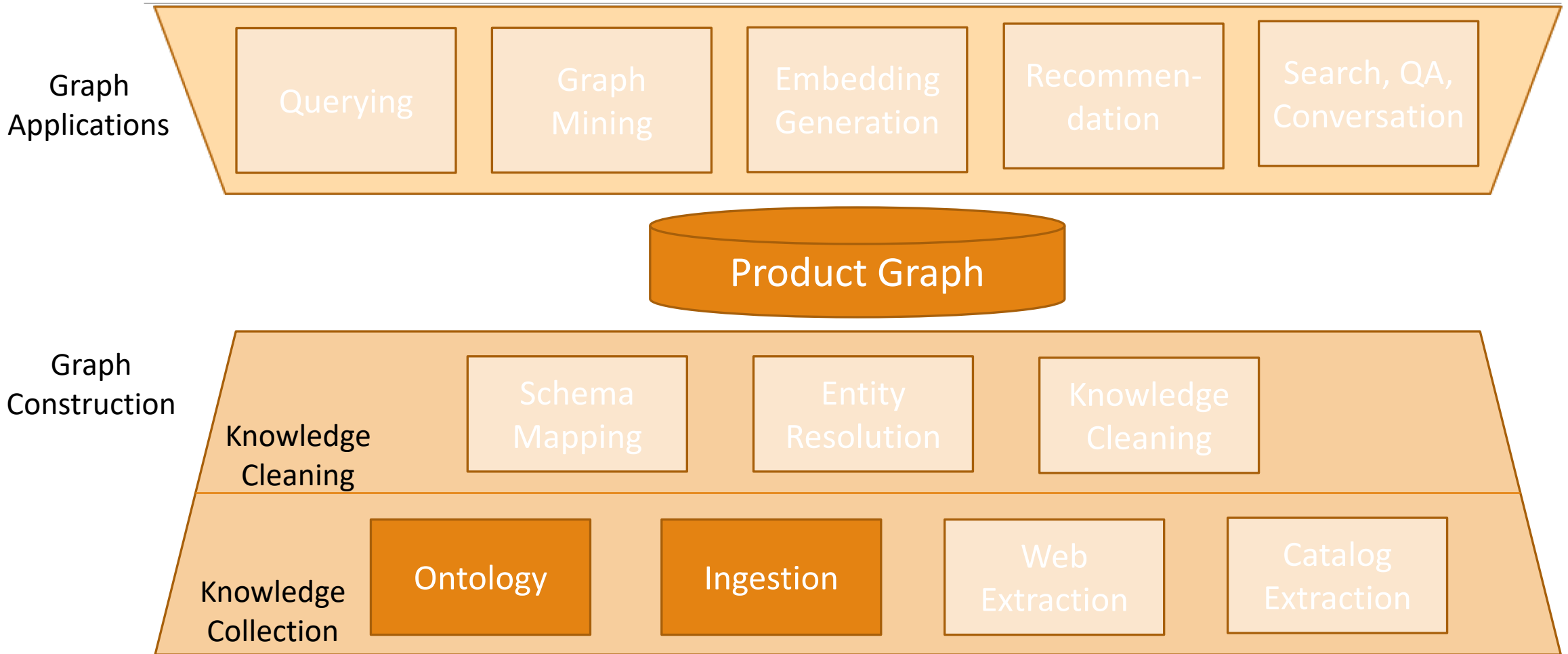
[+ add value](#)

Which BirthDate is correct?

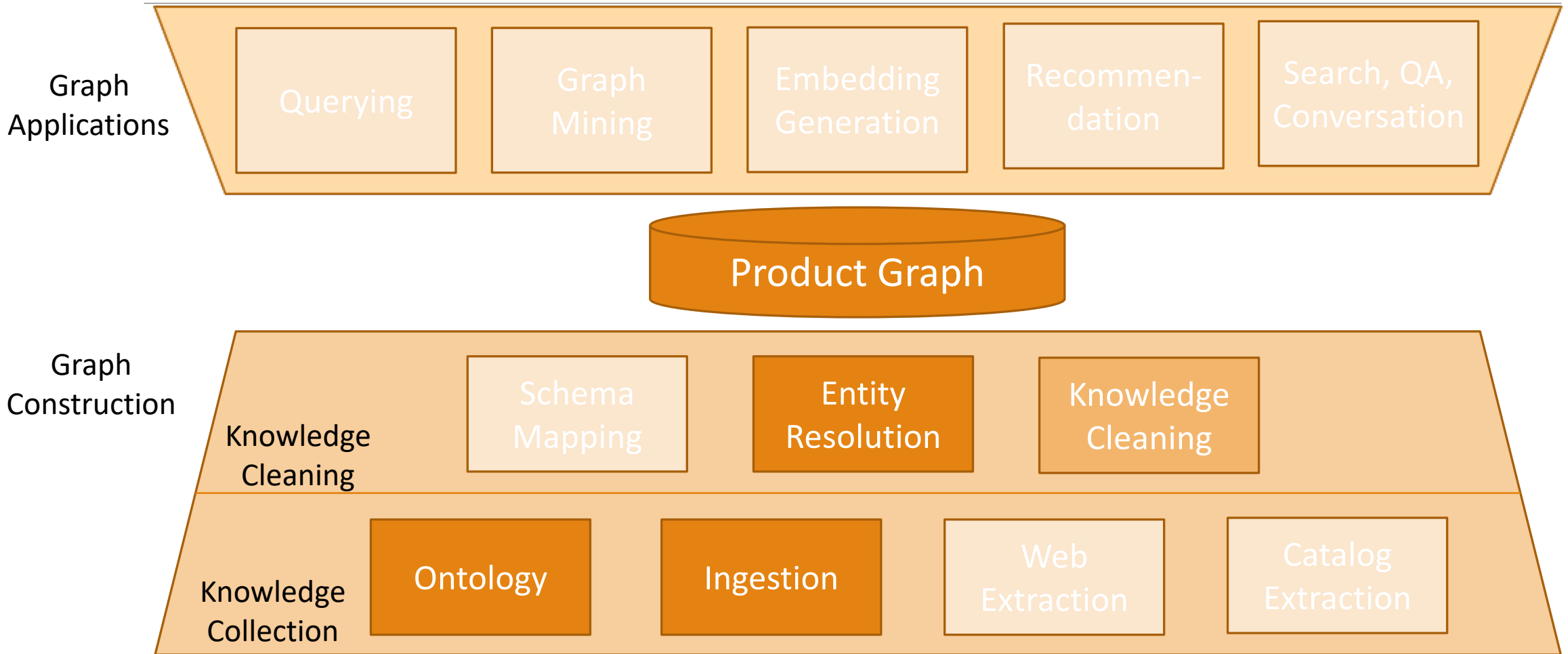
Challenges: Linkage



I. Integrating Knowledge from Structured Sources

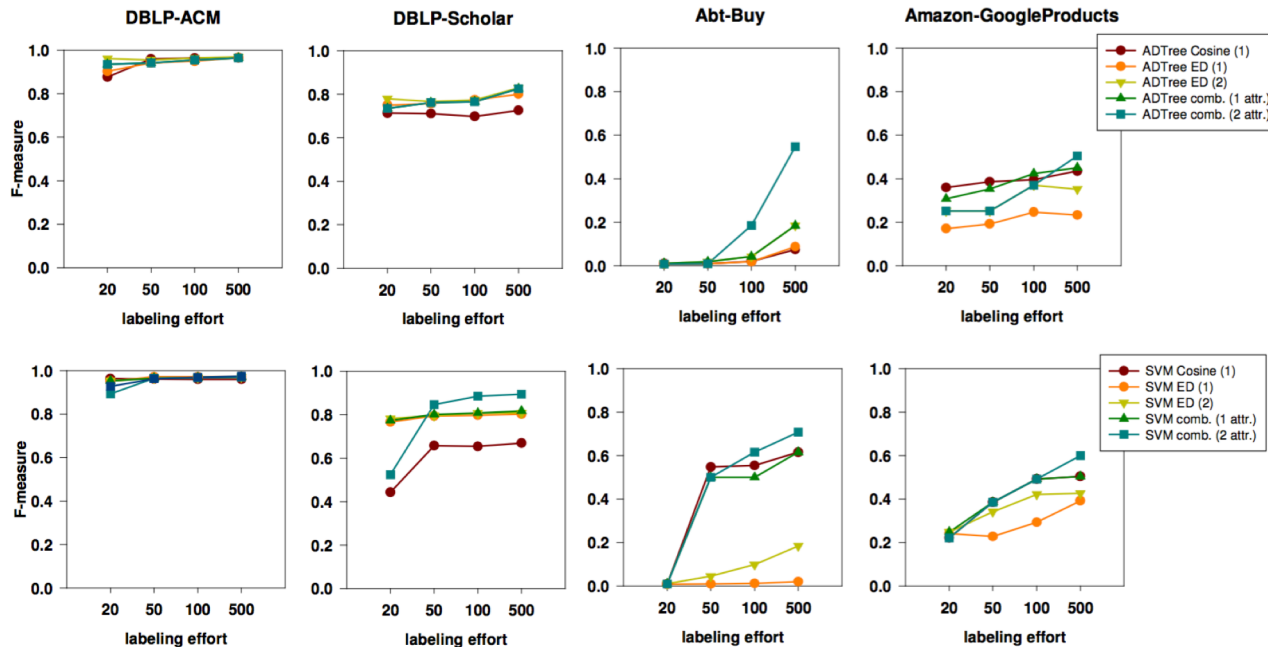


I. Integrating Knowledge from Structured Sources



I. Integrating Knowledge from Structured Sources—Entity Resolution

Published results



[Köpcke et al, VLDB 2010]

| Dataset | Accuracy (%) | | | Cost (# Questions) |
|-----------|--------------|----------|-----------------------|-----------------------|
| | <i>P</i> | <i>R</i> | <i>F</i> ₁ | |
| Products | 90.9 | 74.5 | 81.9 | \$57.6 (960) |
| Songs | 96.0 | 99.3 | 97.6 | \$54.0 (900) |
| Citations | 92.0 | 98.5 | 95.2 | \$65.5 (1087) |

[Das et al, Sigmod 2017]

I. Integrating Knowledge from Structured Sources—Entity Resolution

- Our method:

- Model: Random forest

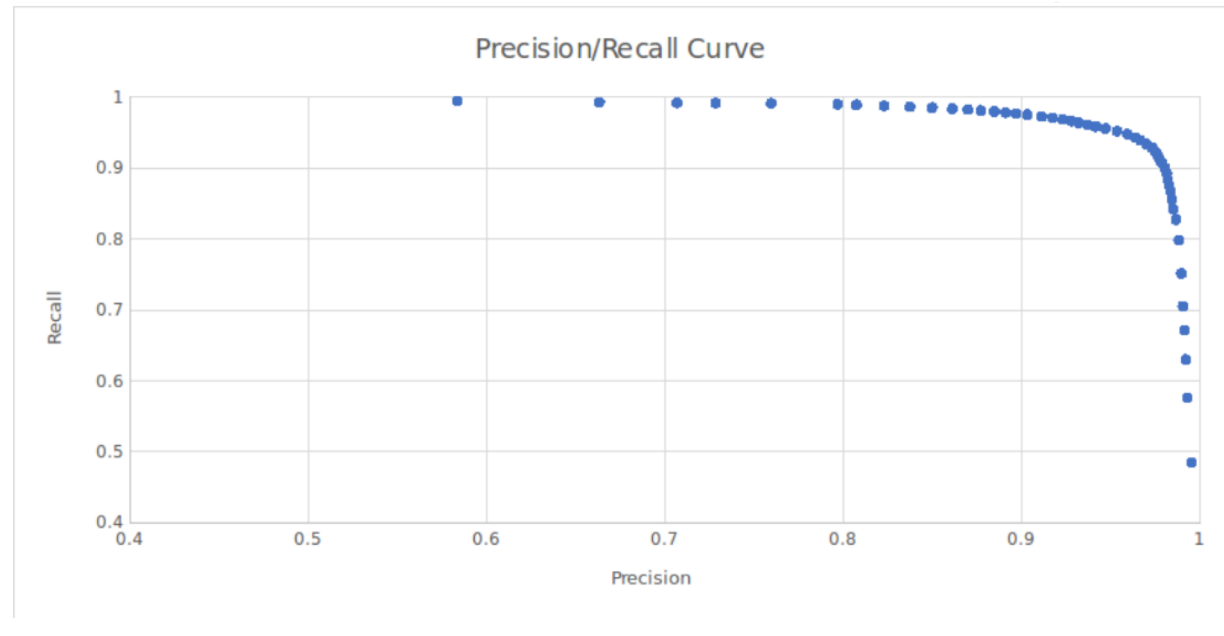
- Features: Attribute similarity—various string similarity, number similarity

I. Integrating Knowledge from Structured Sources—Entity Resolution

- Our method: Random forest on attribute-wise similarity
- Results between Freebase and IMDb: AUPRC=0.9856 (1.5K labels)

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

1.5M labels



I. Integrating Knowledge from Structured Sources—Entity Resolution

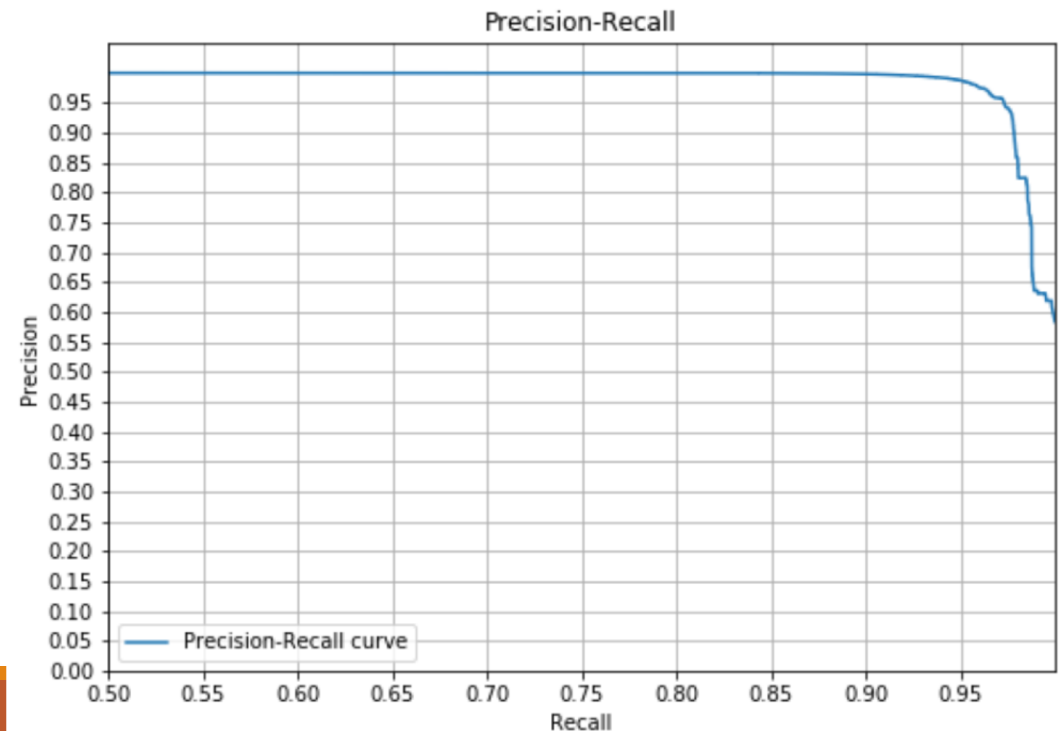
- Our method: Random forest on attribute-wise similarity
- Results between Amazon Movies and IMDb:

200K labels

~150 features

AUPRC=0.9923

Prec=0.982, Rec=0.951

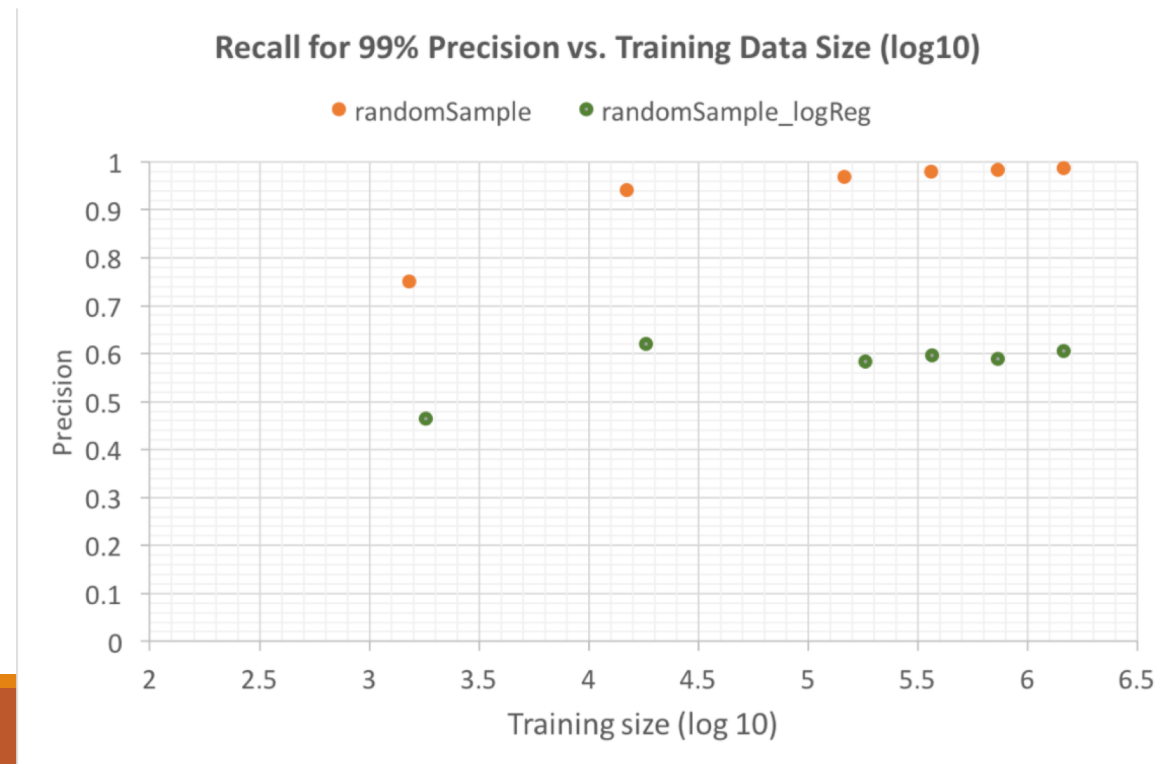


I. Integrating Knowledge—Entity Resolution

Which ML Model Works Best?

❑ Logistic regression: Prec=0.99, Rec=0.6

❑ Random forest: Prec=0.99, Rec=0.99



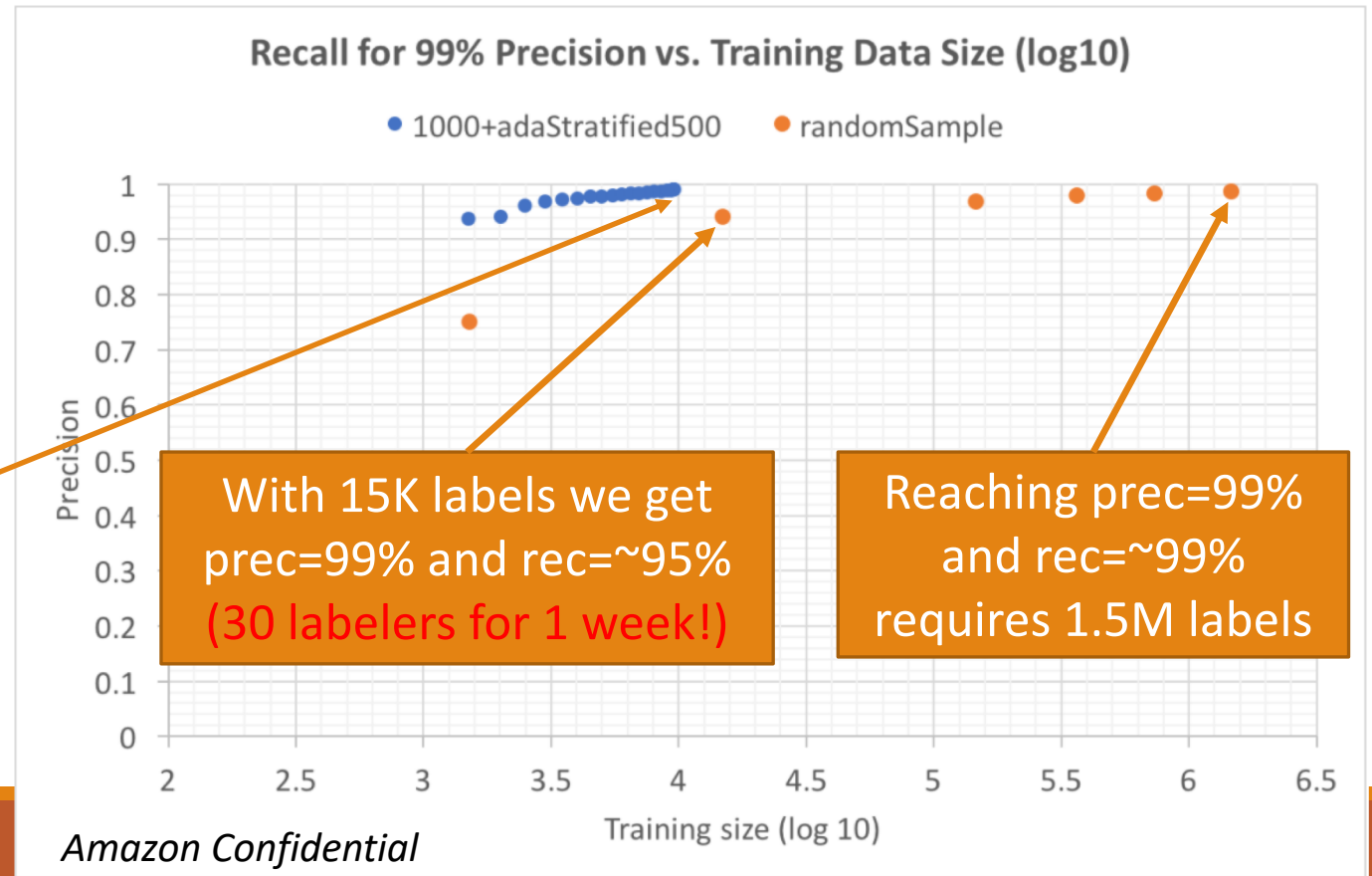
I. Integrating Knowledge—Entity Resolution

Which ML Model Works Best?

- ❑ Logistic regression: Prec=0.99, Rec=0.6
- ❑ Random forest: Prec=0.99, Rec=0.99
- ❑ XGBoost: Marginally better, but sensitive to hyper-parameters
- ❑ Neural network: Similar performance

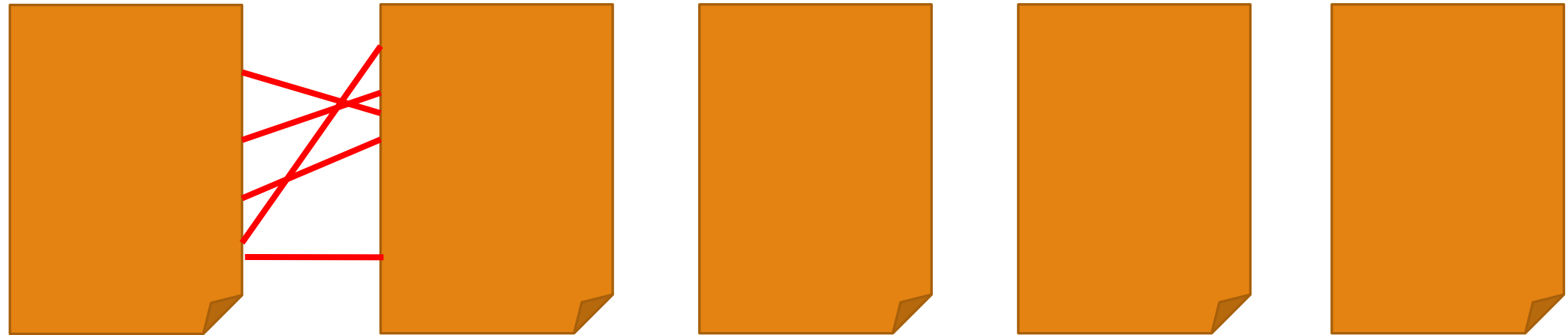
I. Integrating Knowledge from Structured Sources—Entity Resolution

□ Moonshot: Apply active learning to minimize #labels



I. Integrating Knowledge from Structured Sources—Entity Resolution

- Moonshots: Seamless incremental graph linkage with high precision and recall



Freebase

IMDb

Wikipedia

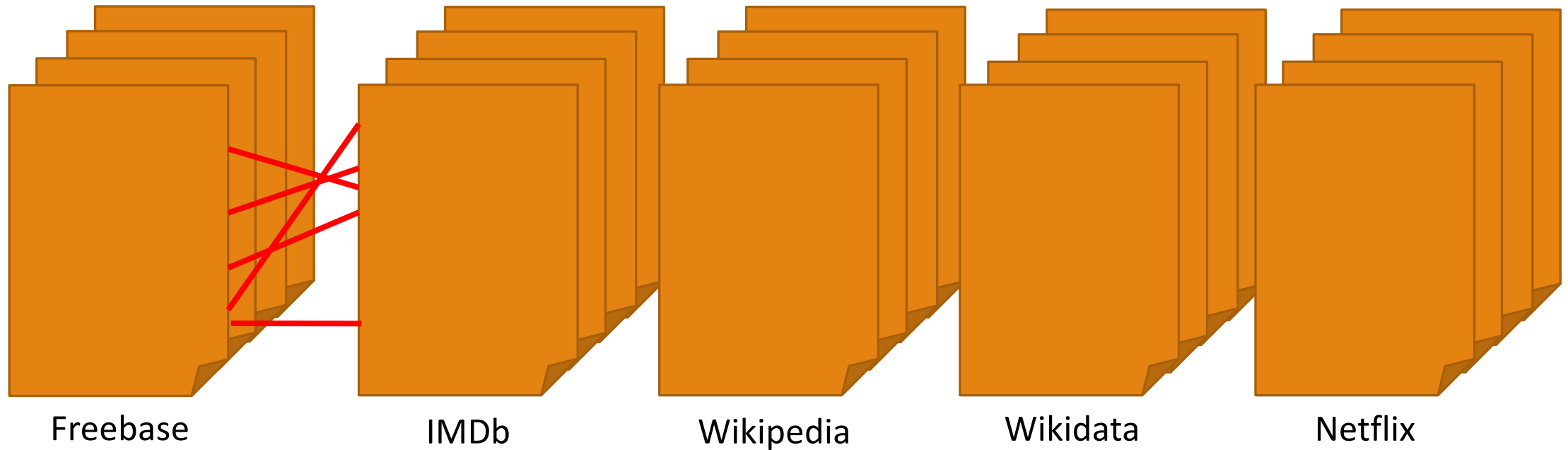
Wikidata

Netflix

- Different sources have different characteristics, but share commonalities from the same domain
- *How to leverage models for existing sources on new sources?*

I. Integrating Knowledge from Structured Sources—Entity Resolution

- Moonshots: Seamless incremental graph linkage with high precision and recall

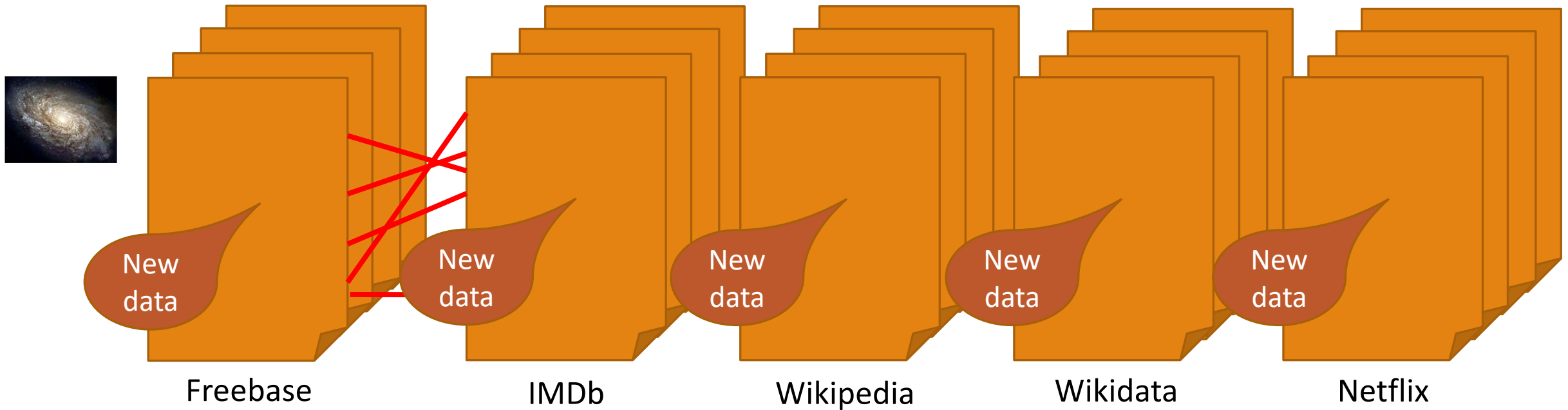


- A complex space can contain tens to thousands of different types and linkage on different types of entities can affect each other
- *How to avoid manual scheduling for linkage?*

I. Integrating Knowledge from Structured Sources

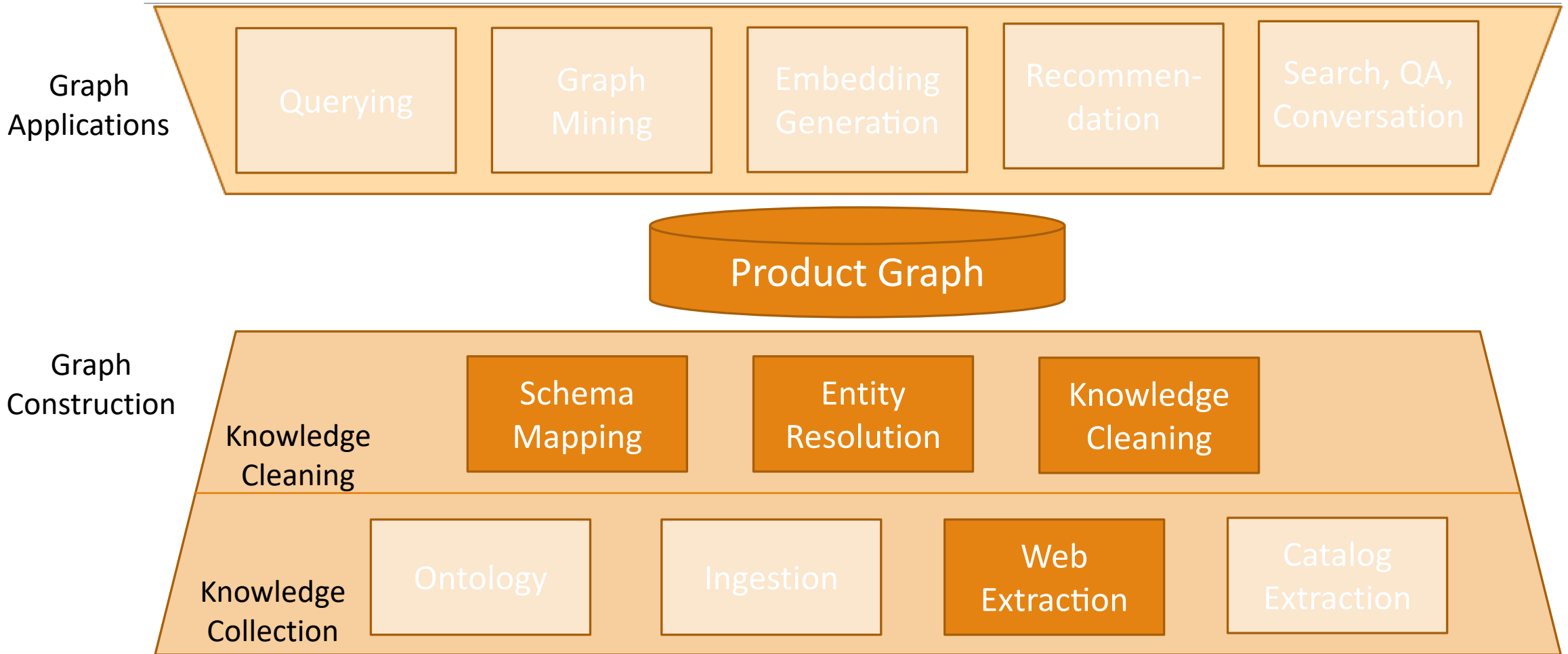


- Moonshots: Seamless incremental graph linkage with high precision and recall



- New data are arriving over time, requiring incremental linkage and model evolution
- *How to perform incremental linkage and evolve the model?*

II. Extracting Knowledge from Semi-Structured Data on the Web



II. Extracting Knowledge from Semi-Structured Data on the Web

FULL CAST AND CREW | TRIVIA | USER REVIEWS

+ Top Gun (1986)

PG | 1h 50min | Action, Drama, Romance



0:50 | Trailer

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

As students at the United States Navy's elite fighter class, one daring young pilot learns a few things from his instructor in the classroom.

Director: Tony Scott
Writers: Jim Cash, Jack Epps Jr. | 1 more credit
Stars: Tom Cruise, Tim Robbins, Kelly McGillis | 5 more credits

Metascore 50
From metacritic.com

Reviews 401 user | 173 critic



Aamir Khan is receiving rave reviews for Dangal.



Dangal
Cast: Aamir Khan, Sakshi Tanwar, Fatima Khurana, Sanya Malhotra
Director: Nitesh Tiwari
Rating: 4/5

卧虎藏龙 臥虎藏龍 (2000)



导演: 李安
编剧: 王蕙玲 / 詹姆斯·夏慕斯 / 蔡国荣
主演: 周润发 / 杨紫琼 / 章子怡 / 张震 / 郑佩佩 / 更多...

类型: 剧情 / 动作 / 爱情 / 武侠 / 古装
制片国家/地区: 台湾 / 香港 / 美国 / 中国大陆
语言: 汉语普通话
上映日期: 2000-10-13(中国大陆) / 2000-05-16(戛纳电影节) / 2000-07-07(台湾) / 2000-07-13(香港) / 2001-01-12(美国)
片长: 120 分钟
又名: Crouching Tiger, Hidden Dragon
IMDb链接: tt0190332

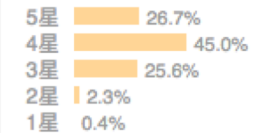
想看 看过 评价: ☆☆☆☆☆

写短评 写影评 + 提问题 分享到

推荐

豆瓣评分

7.9  166740人评价



好于 92% 武侠片
好于 90% 动作片

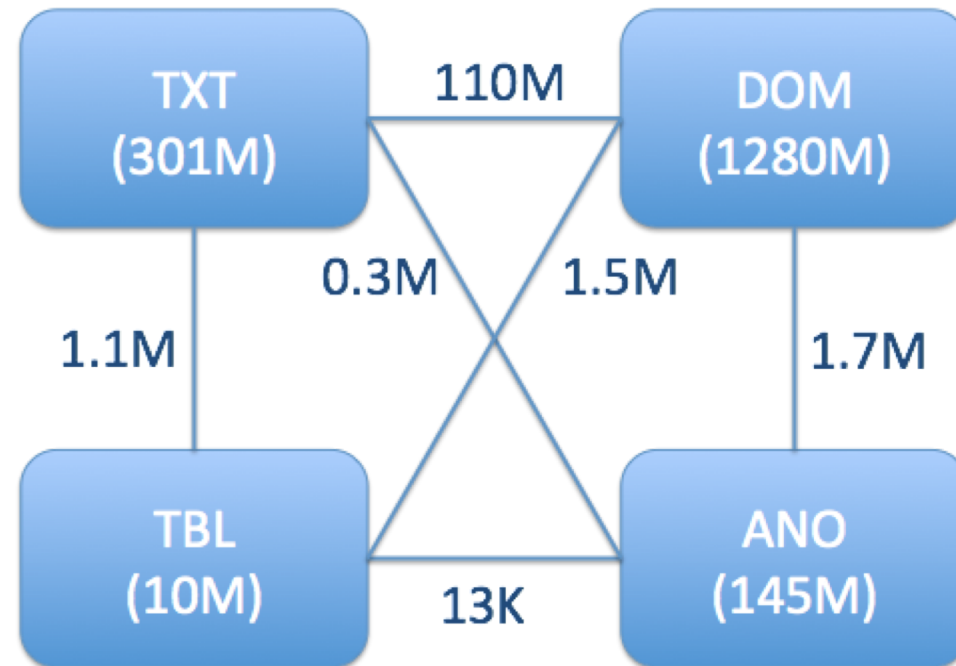
卧虎藏龙的剧情简介 · · · · ·

一代大侠李慕白(周润发饰)有退出江湖之意,托付红颜知己俞秀莲(杨紫琼饰)将青冥剑转交给贝勒爷(郎雄饰)收藏,不料当夜遭玉娇龙(章子怡)窃取。俞秀莲暗中查访也大约知道是玉府小姐玉蛟龙所为,她想办法迫使玉蛟龙归还宝剑,免伤和气。但李慕白发现了害死师傅的碧眼狐狸(郑佩佩饰)的踪迹,她隐匿于玉府并收玉蛟龙为弟子。而玉蛟龙欲以青冥剑来斩断阻碍罗小虎(张震饰)的枷锁,他们私定终身。关系变得错综复杂,俞秀莲和李慕白爱惜玉蛟龙人才难得,苦心引导,但玉蛟龙却使性任气不听劝阻..... ©豆瓣

II. Extracting Knowledge from Semi-Structured Data on the Web

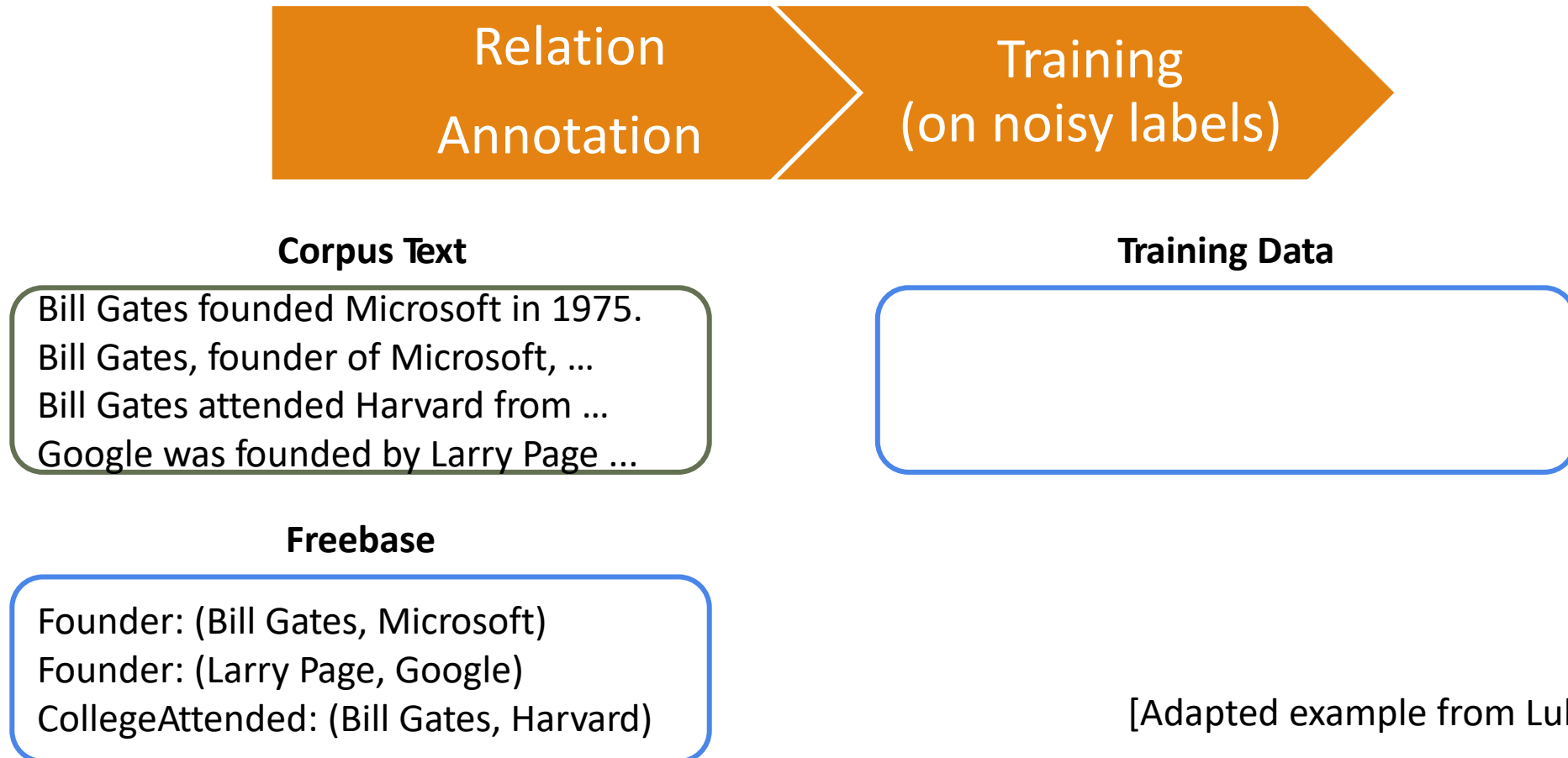
- Knowledge Vault @ Google showed big potential from DOM-tree extraction [Dong et al., KDD'14][Dong et al., VLDB'14]

| Accu | Accu (conf \geq .7) |
|------|-----------------------|
| 0.36 | 0.52 |



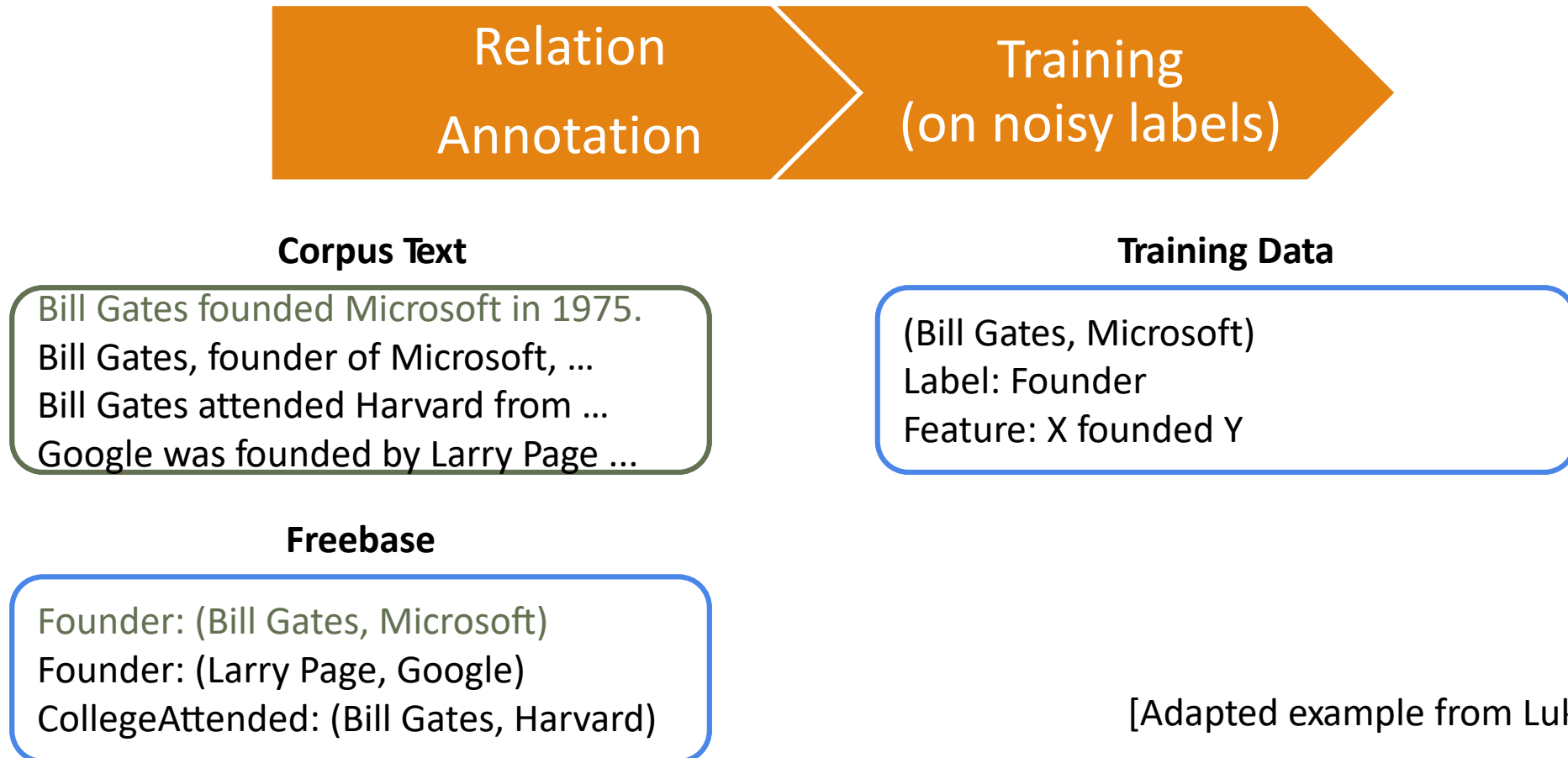
| Accu | Accu (conf \geq .7) |
|------|-----------------------|
| 0.43 | 0.63 |
| 0.09 | 0.62 |

II. Extracting Knowledge from Web— Distantly Supervised DOM Extraction [VLDB'18]



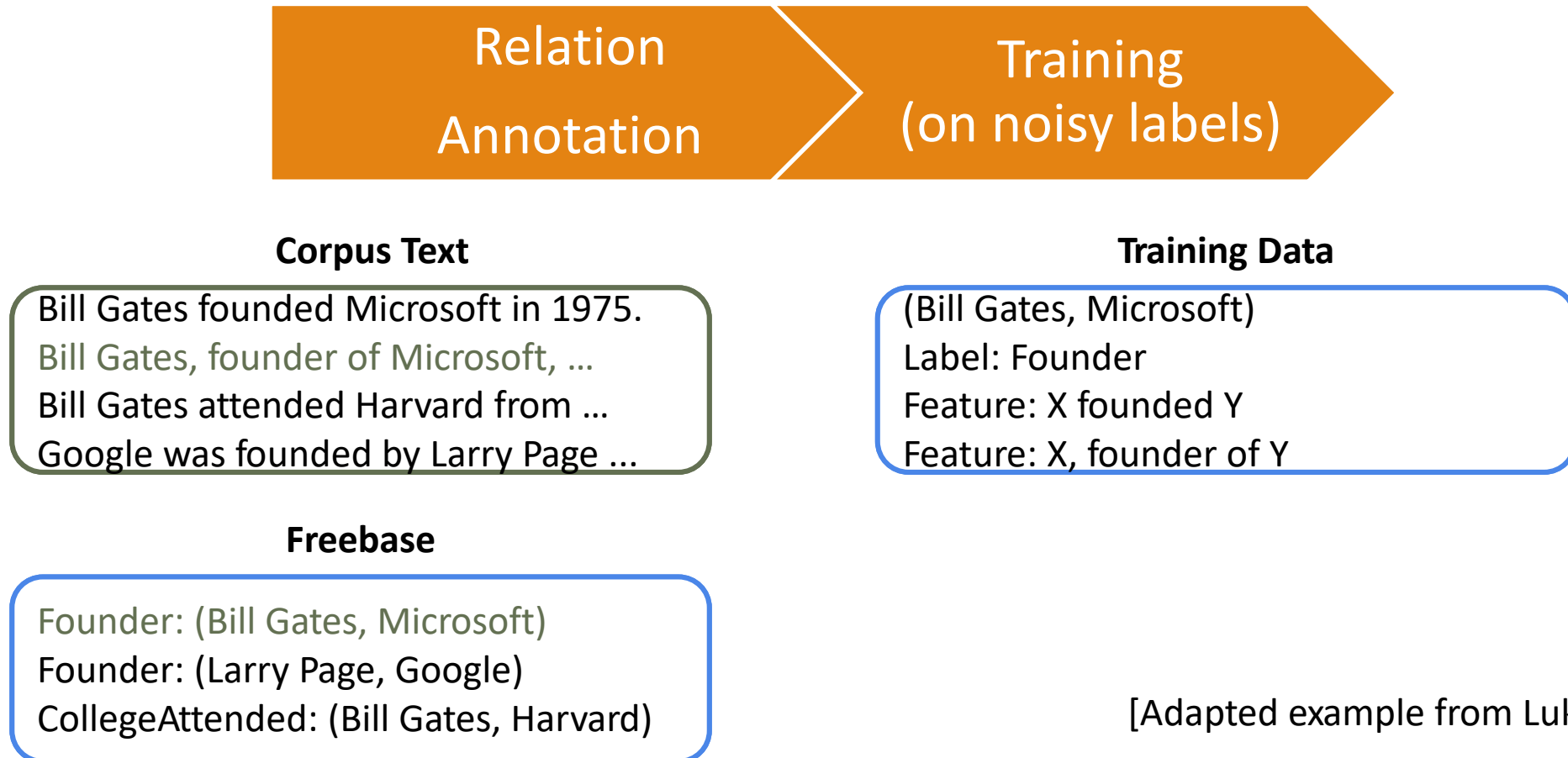
[Adapted example from Luke Zettlemoyer]

II. Extracting Knowledge from Web— Distantly Supervised DOM Extraction [VLDB'18]



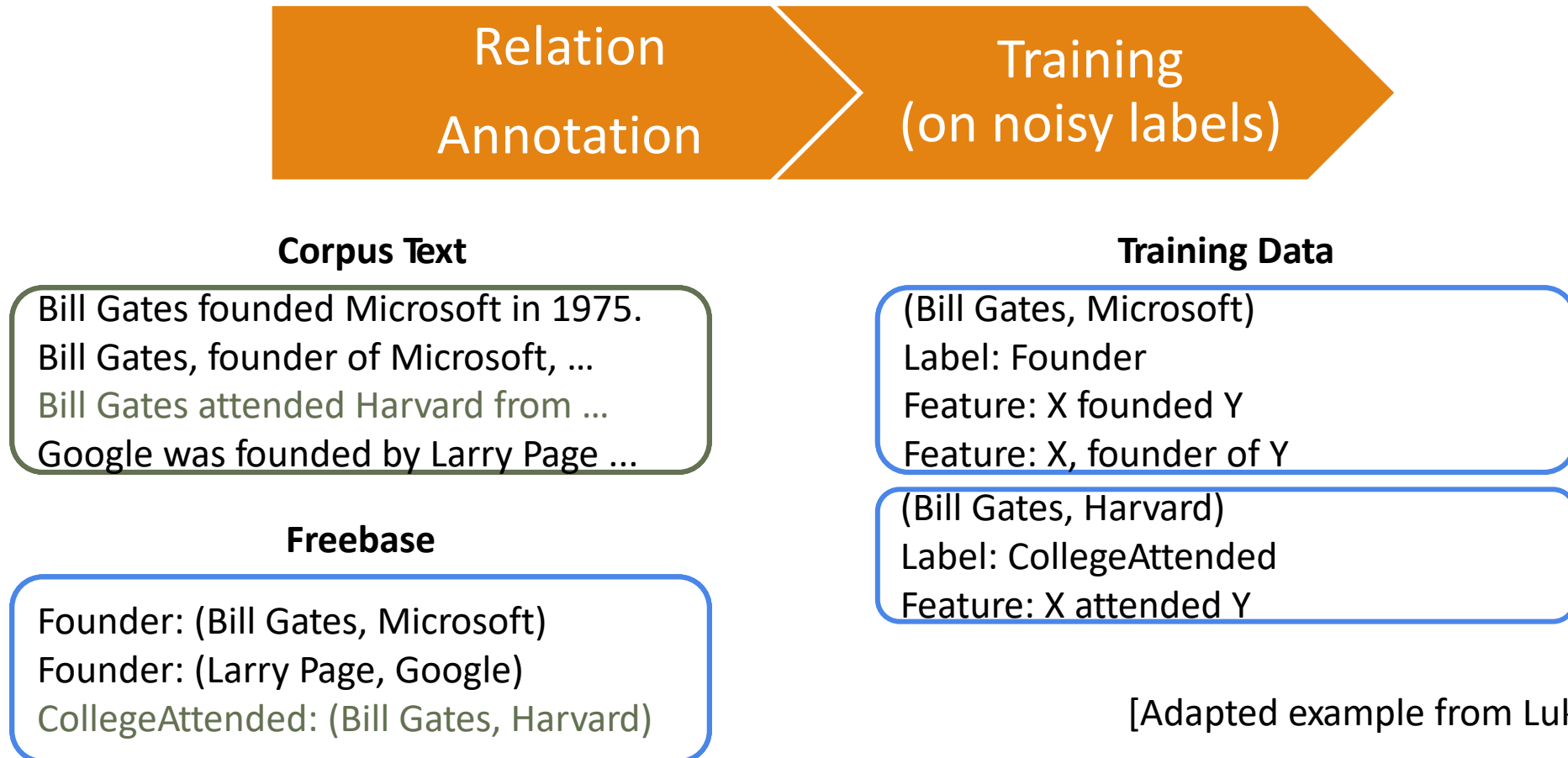
[Adapted example from Luke Zettlemoyer]

II. Extracting Knowledge from Web— Distantly Supervised DOM Extraction [VLDB'18]



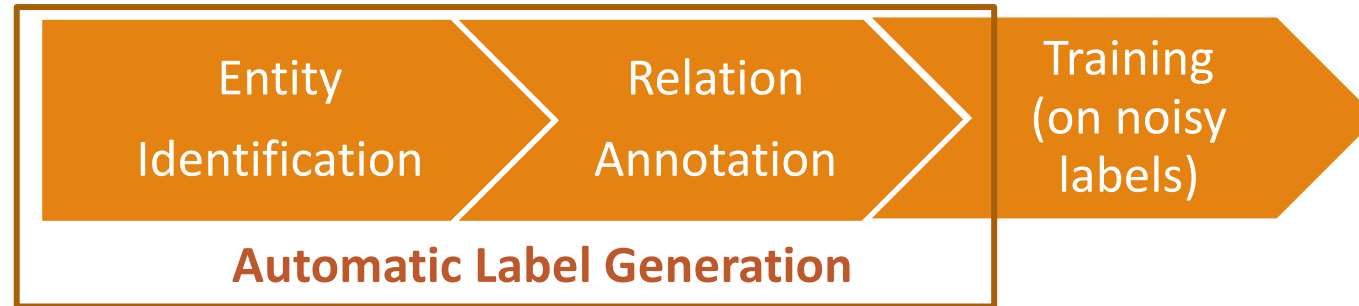
[Adapted example from Luke Zettlemoyer]

II. Extracting Knowledge from Web— Distantly Supervised DOM Extraction [VLDB'18]



[Adapted example from Luke Zettlemoyer]

II. Extracting Knowledge from Web— Distantly Supervised DOM Extraction [VLDB'18]



Movie entity

Genre Release Date

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

+ **Top Gun (1986)** ★ 6.9 ^{1.0}_{234,144} ☆ Rate This

PG | 1h 50min | Action, Drama, Romance | 16 May 1986 (USA)

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

As students at the United States Navy's elite fighter weapons school compete to be best in the class, one daring young pilot learns a few things from a civilian instructor that are not taught in the classroom.

Director: [Tony Scott](#)

Writers: [Jim Cash](#), [Jack Epps Jr.](#) | [1 more credit](#) >

Stars: [Tom Cruise](#), [Tim Robbins](#), [Kelly McGillis](#) | [See full cast & crew](#) >

Runtime

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

+ **Top Gun (1986)** ★ 6.9 ^{1.0}_{234,144} ☆ Rate This

PG | 1h 50min | Action, Drama, Romance | 16 May 1986 (USA)

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

As students at the United States Navy's elite fighter weapons school compete to be best in the class, one daring young pilot learns a few things from a civilian instructor that are not taught in the classroom.

Director: [Tony Scott](#)

Writers: [Jim Cash](#), [Jack Epps Jr.](#) | [1 more credit](#) >

Stars: [Tom Cruise](#), [Tim Robbins](#), [Kelly McGillis](#) | [See full cast & crew](#) >

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")

II. Extracting Knowledge from Web— Distantly Supervised DOM Extraction [VLDB'18]

□ Extraction experiments on <http://swde.codeplex.com/> (2011)

| Vertical | Predicate | Wrapper induction | | | Distant-super | | |
|-----------|-------------|-------------------|------|------|---------------|-------------|-------------|
| | | P | R | F1 | P | R | F1 |
| Movie | Title | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Director | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| | Genre | 0.88 | 0.87 | 0.87 | 0.93 | 0.97 | 0.95 |
| | MPAA Rating | 1.00 | 1.00 | 1.00 | NA | NA | NA |
| | Average | 0.97 | 0.97 | 0.97 | 0.97 | 0.99 | 0.98 |
| NBAPlayer | Name | 0.99 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 |
| | Team | 1.00 | 1.00 | 1.00 | 0.91 | 1.00 | 0.95 |
| | Weight | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Height | 1.00 | 1.00 | 1.00 | 1.00 | 0.90 | 0.95 |
| | Average | 1.00 | 1.00 | 1.00 | 0.98 | 0.98 | 0.98 |

| Vertical | Predicate | Wrapper induction | | | Distant-super | | |
|------------|------------------|-------------------|------|------|---------------|-------------|-------------|
| | | P | R | F1 | P | R | F1 |
| University | Name | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Type | 1.00 | 1.00 | 1.00 | 0.72 | 0.80 | 0.76 |
| | Phone | 0.97 | 0.92 | 0.94 | 0.85 | 0.95 | 0.90 |
| | Website | 1.00 | 1.00 | 1.00 | 0.90 | 1.00 | 0.95 |
| | Average | 0.99 | 0.98 | 0.99 | 0.87 | 0.94 | 0.90 |
| Book | Title | 0.99 | 0.99 | 0.99 | 1.00 | 0.90 | 0.95 |
| | Author | 0.97 | 0.96 | 0.96 | 0.72 | 0.88 | 0.79 |
| | Publisher | 0.85 | 0.85 | 0.85 | 0.97 | 0.77 | 0.86 |
| | Publication Date | 0.90 | 0.90 | 0.90 | 1.00 | 0.40 | 0.57 |
| | ISBN-13 | 0.94 | 0.94 | 0.94 | 0.99 | 0.19 | 0.32 |
| | Average | 0.93 | 0.93 | 0.93 | 0.94 | 0.63 | 0.70 |

Very high precision

Competent w. Wrapper induction
with manual annotation

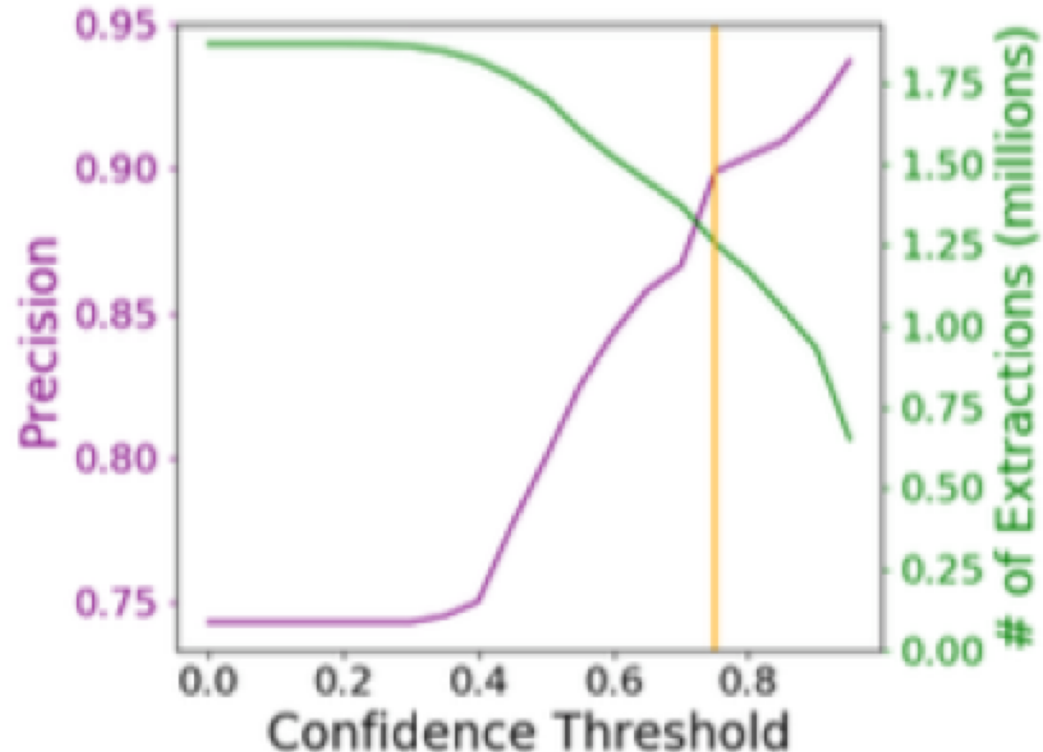
II. Extracting Knowledge from Web— Distantly Supervised DOM Extraction [VLDB'18]

□ Extraction on long-tail movie websites

| | |
|--|--|
| #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% |

II. Extracting Knowledge from Web— Distantly Supervised DOM Extraction

- Extraction on long-tail movie websites



II. Extracting Knowledge from the Web— OpenIE DOM Extraction

- ❑ 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”)



II. Extracting Knowledge from the Web— OpenIE DOM Extraction

- ❑ 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”)



MOVIE INFO

Does sex make it impossible for men and women to b dilemma through the eleven year relationship between their own lives until they reconnect ten years later.

Rating: R
Genre: [Comedy](#), [Drama](#), [Romance](#)
Directed By: [Rob Reiner](#)
Written By: [Nora Ephron](#)
In Theaters: Jul 12, 1989 Wide
On Disc/Streaming: Oct 13, 1998
Runtime: 96 minutes

II. Extracting Knowledge from the Web— OpenIE DOM Extraction

("When Harry Met Sally", "Rating:", "R")

("When Harry Met Sally", "Genre:", "Comedy")

("When Harry Met Sally", "Genre:", "Drama")

("When Harry Met Sally", "Genre:", "Romance")

("When Harry Met Sally", "Directed By:", "Rob Reiner")

("When Harry Met Sally", "Written By:", "Nora Ephron")

("When Harry Met Sally", "In Theaters", "Jul 12, 1989
Wide")

("When Harry Met Sally", "On Disc/Streaming", "Oct 13,
1998")

("When Harry Met Sally", "Runtime", "96 minutes")



MOVIE INFO

Does sex make it impossible for men and women to b dilemma through the eleven year relationship between their own lives until they reconnect ten years later.

Rating: R
Genre: [Comedy](#), [Drama](#), [Romance](#)
Directed By: [Rob Reiner](#)
Written By: [Nora Ephron](#)
In Theaters: Jul 12, 1989 Wide
On Disc/Streaming: Oct 13, 1998
Runtime: 96 minutes

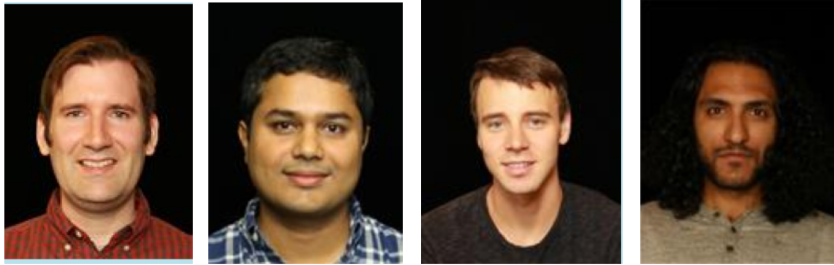
II. Extracting Knowledge from the Web— OpenIE DOM Extraction Preliminary Rslts

| Site | # New Preds | Pred Precision | Pred Recall | Triple Precision | Triple Recall |
|----------|-------------|----------------|-------------|------------------|---------------|
| Slam | 4 | 1.0 | 0.5 | <i>0.95</i> | <i>0.5</i> |
| Wiki | 7 | 1.0 | ~1.0 | <i>0.9</i> | <i>0.9</i> |
| ESPN | 9 | 1.0 | 1.0 | <i>0.7</i> | <i>0.7</i> |
| Fanhouse | 6 | 1.0 | 1.0 | <i>1.0</i> | <i>1.0</i> |
| SI | 5 | 0.88 | 1.0 | <i>0.8</i> | <i>1.0</i> |
| USAToday | 0 | 0.33 | 0.2 | <i>0.2</i> | <i>0.2</i> |
| Yahoo | 3 | 1.0 | 1.0 | <i>1.0</i> | <i>1.0</i> |

II. Distantly Supervised DOM Extraction Which ML Model Works Best?

- ❑ Logistic regression: Best results (20K features on one website)
- ❑ Random forest: lower precision and recall

II. Extracting Knowledge from Web— Distantly Supervised DOM Extraction



Annotation-based
knowledge extraction



Distantly supervised
web extraction

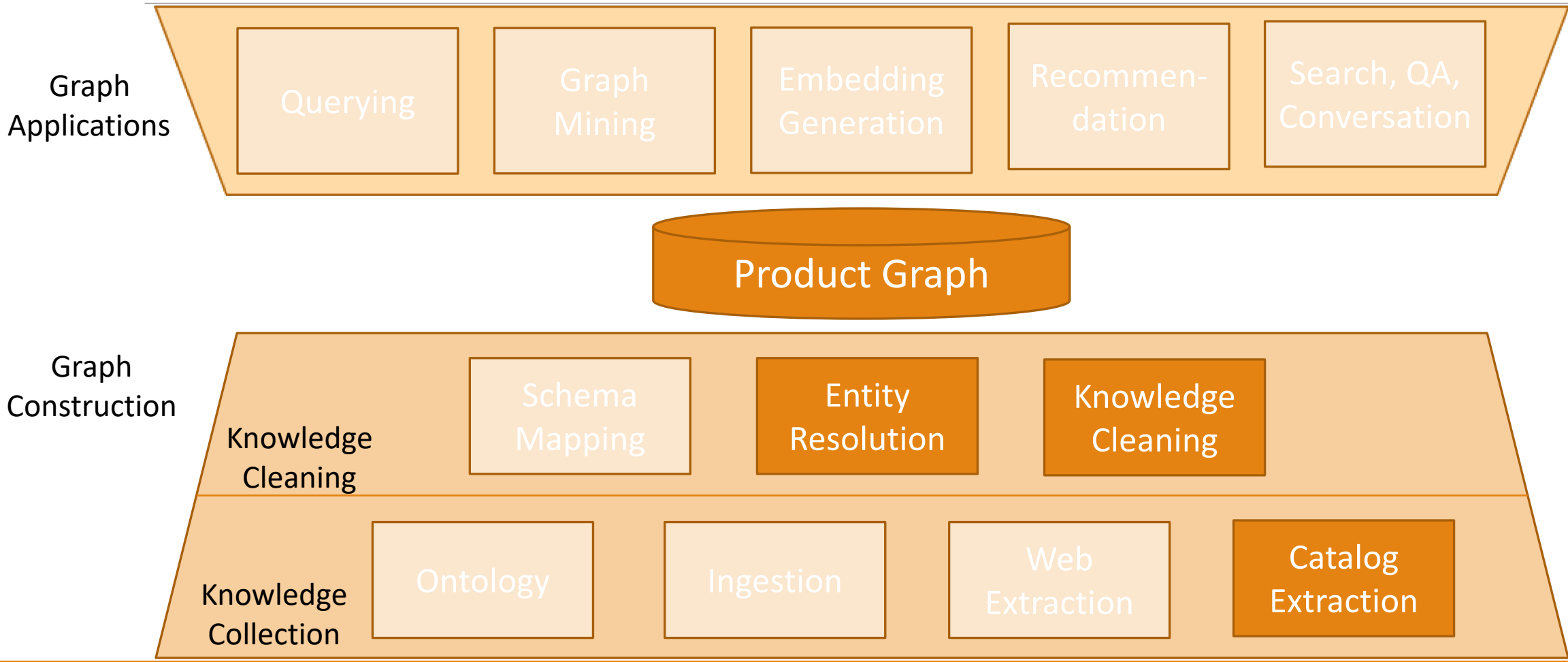


OpenIE
DOM extraction

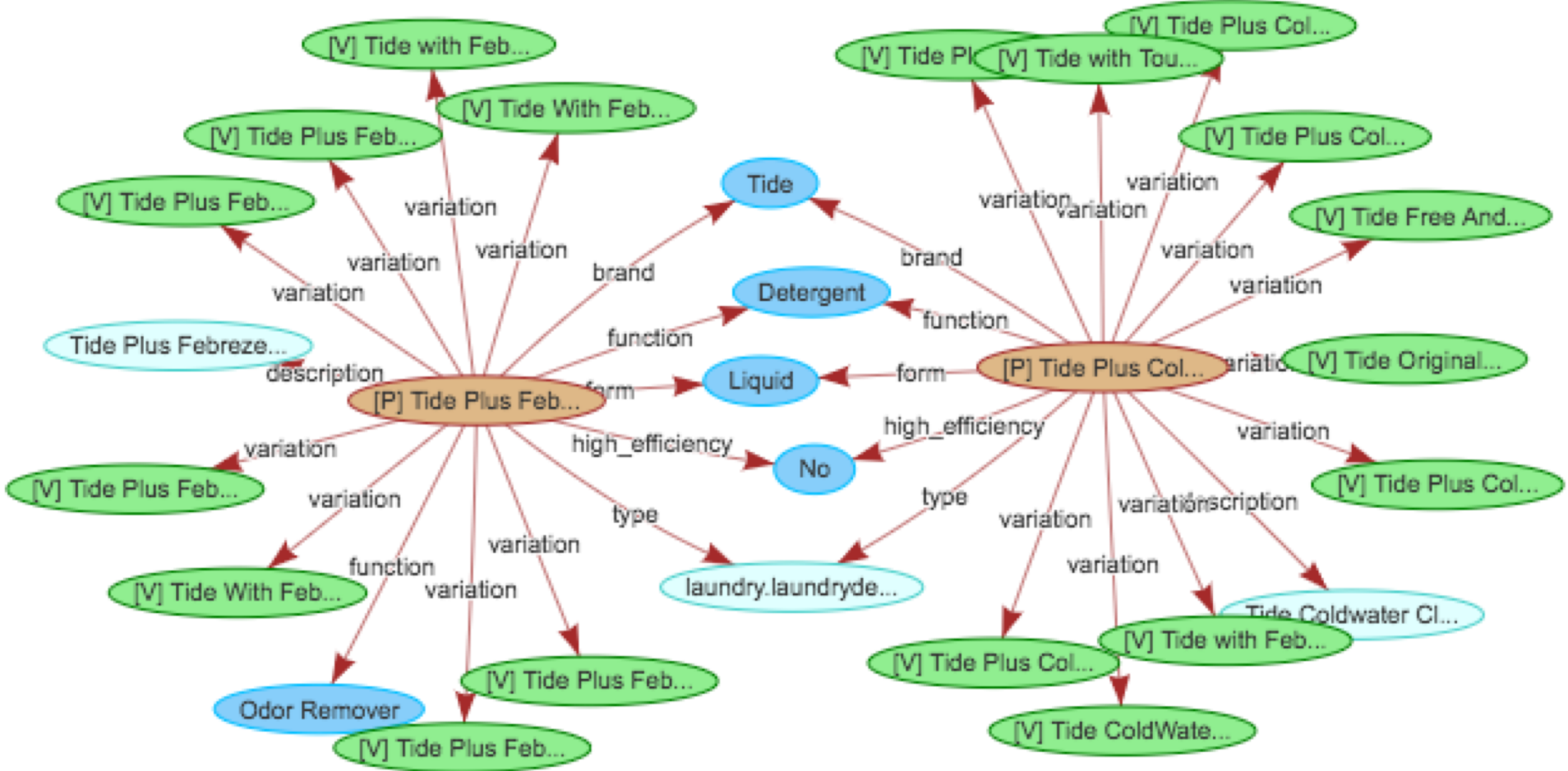


Nearly-automatic
interactive extraction
on any new vertical

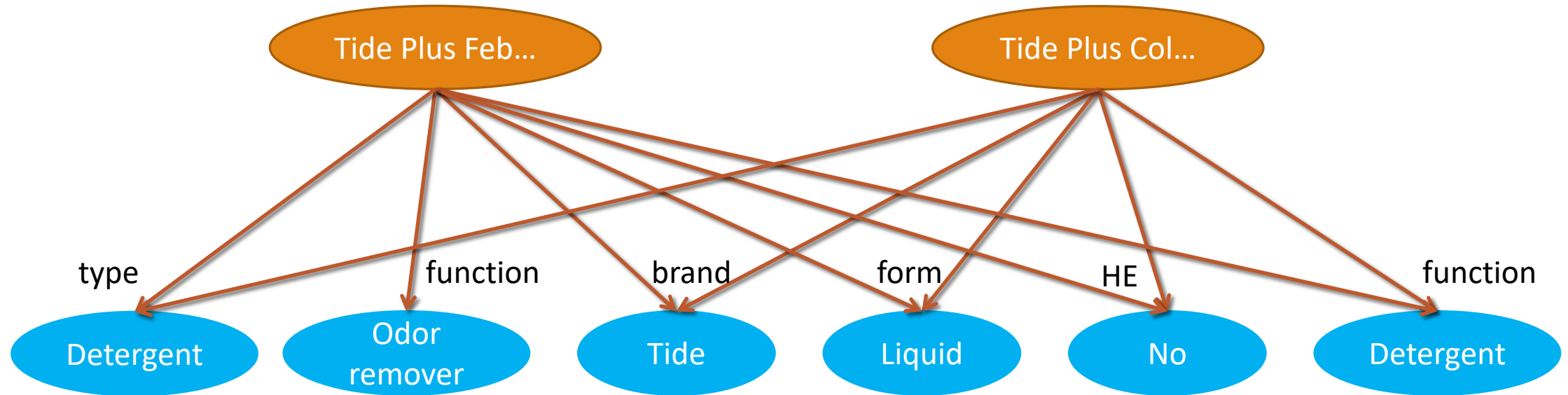
III. Building a Broad & Shallow Graph from Product Profiles in Amazon Catalog



Another Example of Product Graph



III. Building a Broad & Shallow Graph from Product Profiles in Amazon Catalog



III. Building a Broad & Shallow Graph from Product Profiles in Amazon Catalog

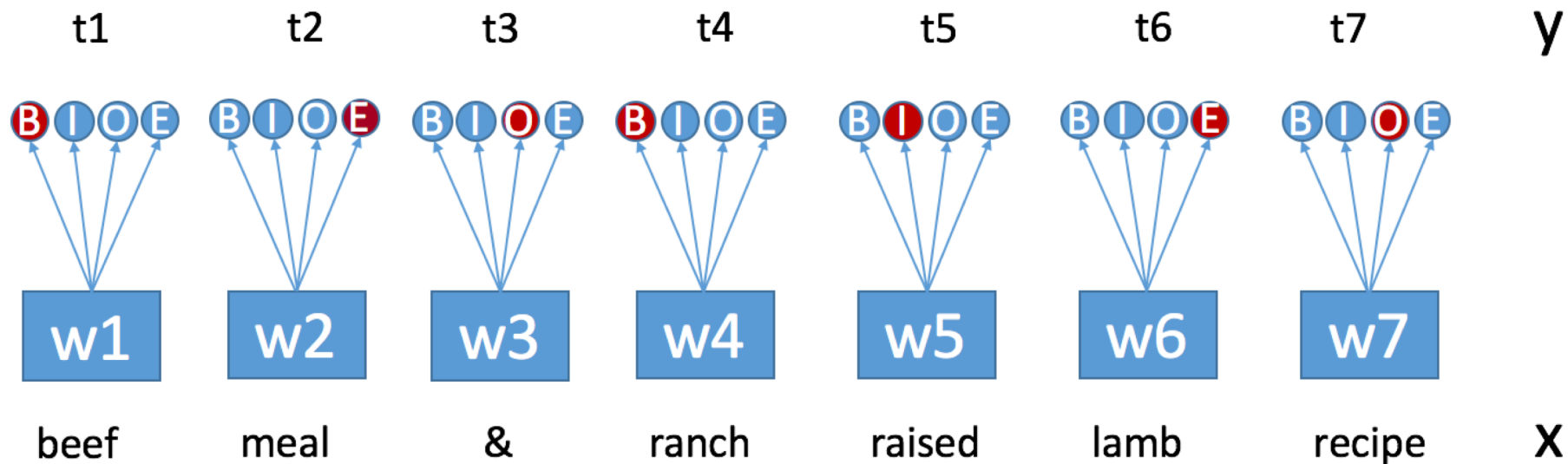
| name | form | scent |
|--|------|-------|
| Tide Detergent with Febreze Freshness | | |
| Gain Apple Mango Tango Liquid Laundry Detergent | | |
| Gain Joyful Expression Powder Detergent | | |
| Tide PODS Original Scent HE Turbo Laundry Detergent Pacs 81-load Tub | | |
| Tide PODS Free & Gentle HE Turbo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

III. Open Attribute Extraction by Named Entity Recognition [KDD'18]

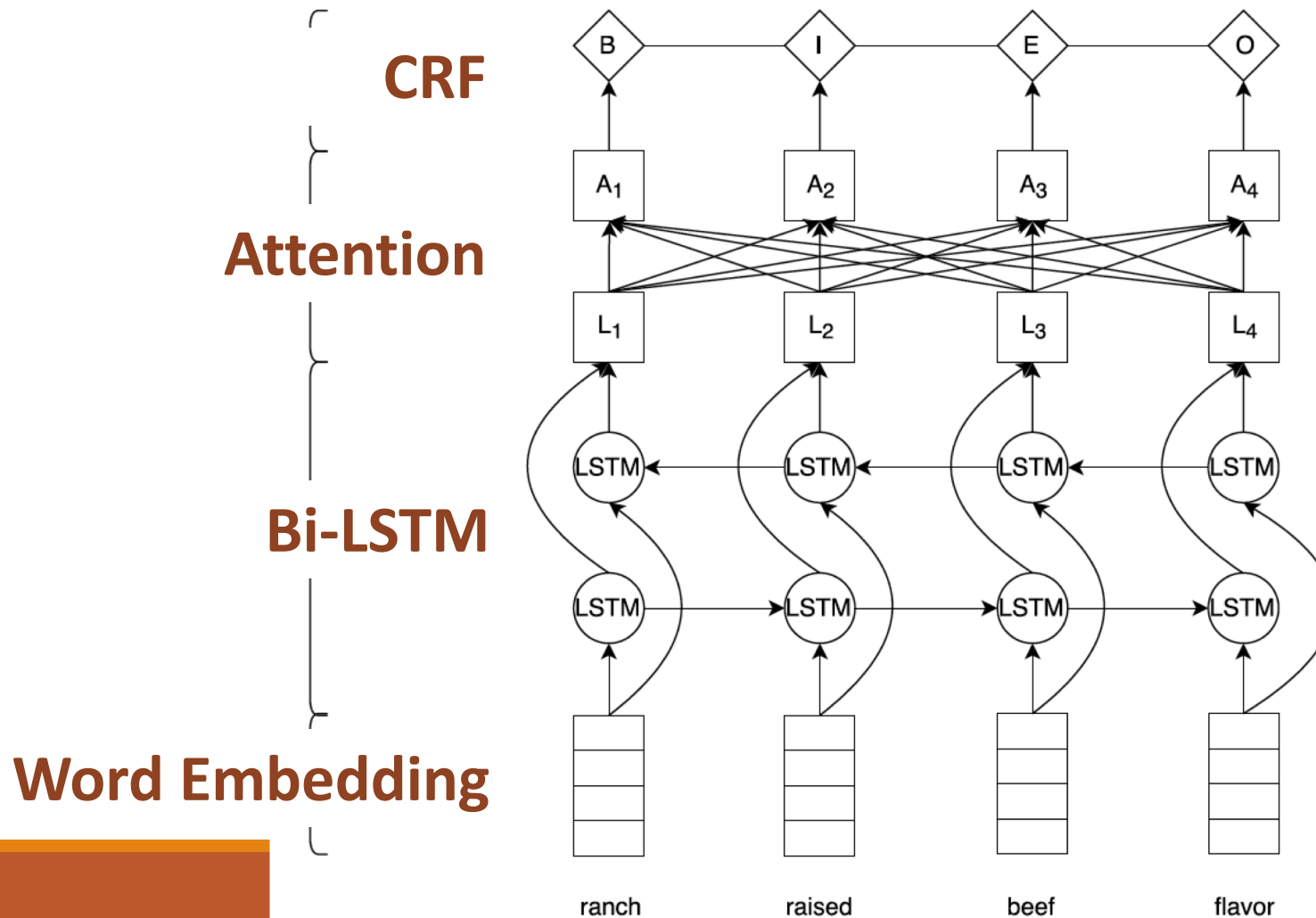
- B** Beginning of entity
- I** Inside of entity
- O** Outside of entity
- E** End of entity

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

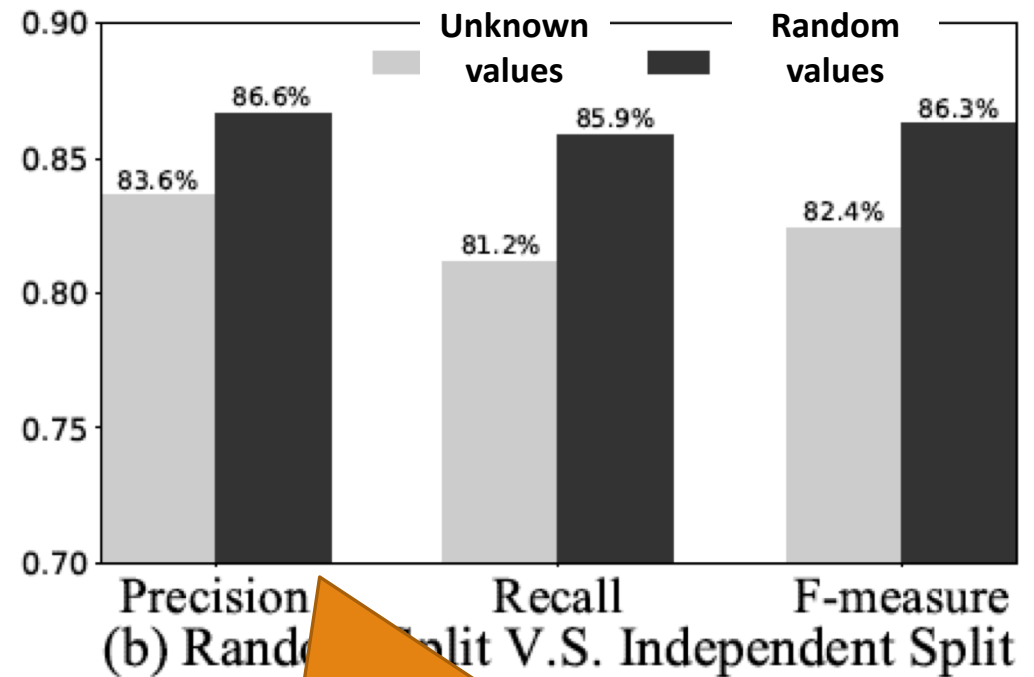
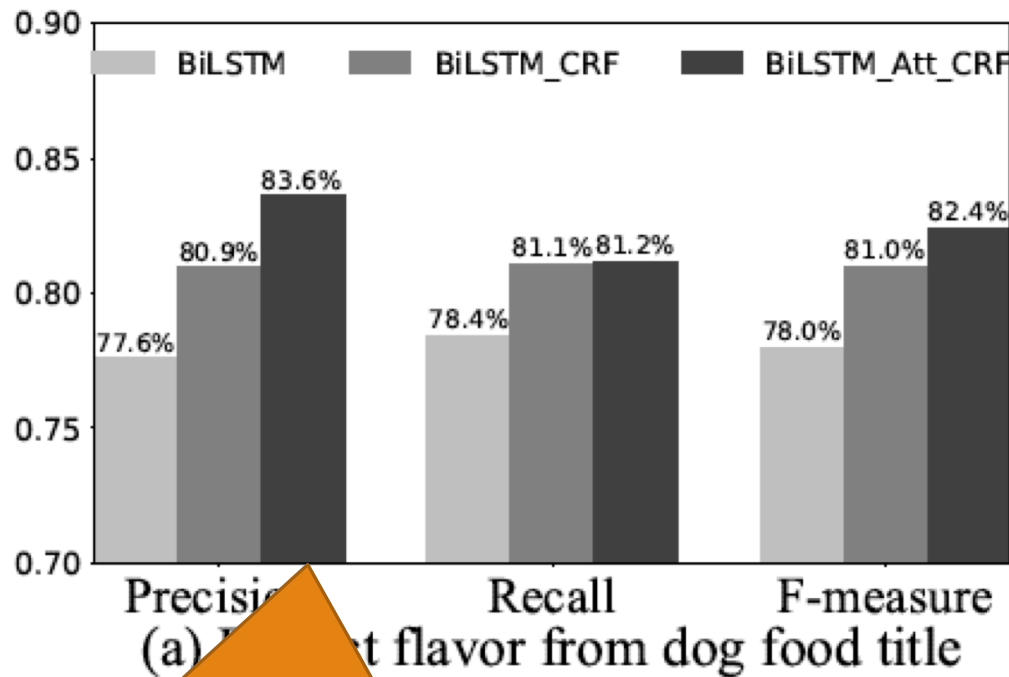
$y = \{t_1, t_2, \dots, t_n\}$ tagging decision



III. Open Attribute Extraction by Named Entity Recognition [KDD'18]



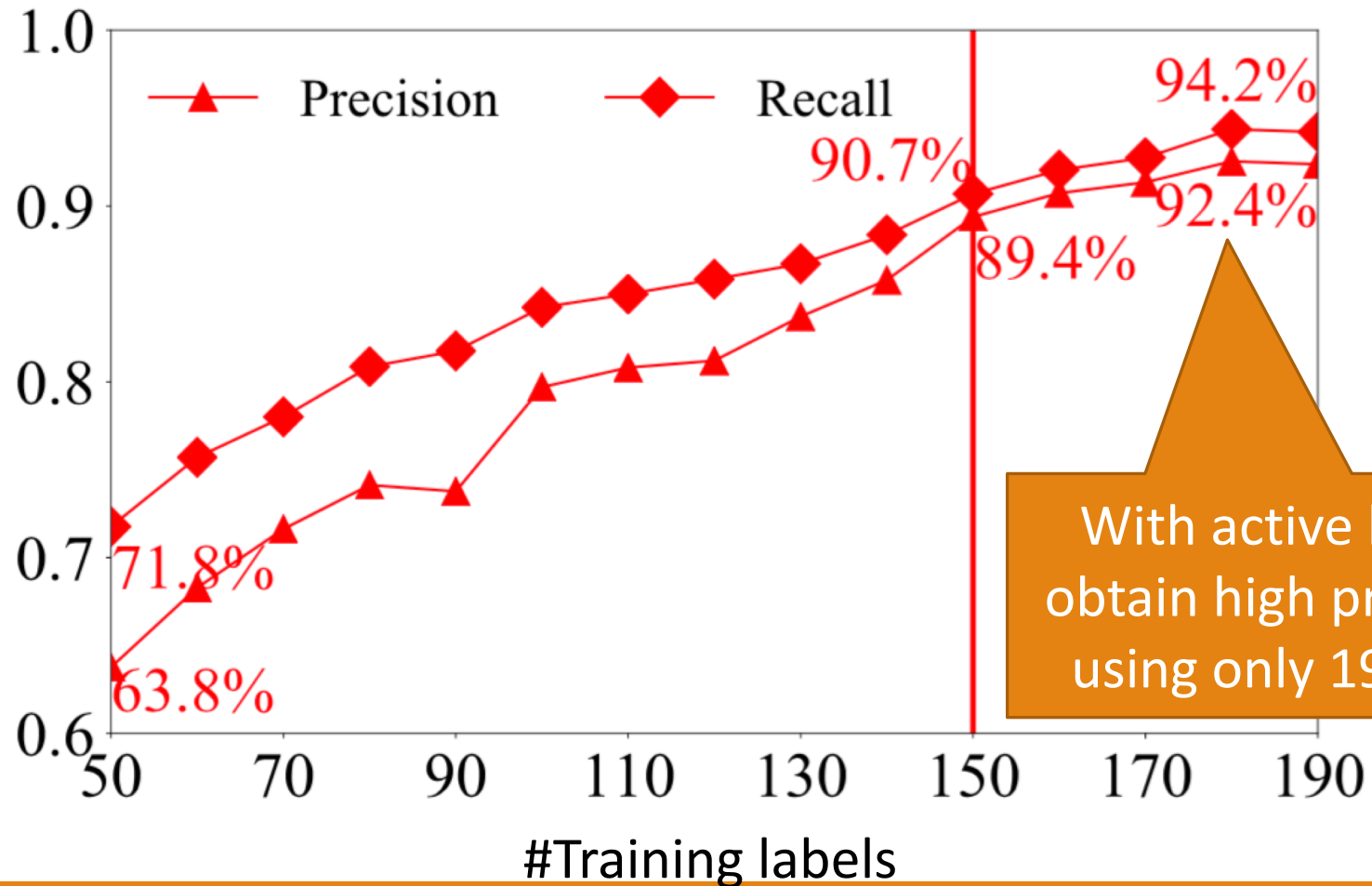
III. Open Attribute Extraction by Named Entity Recognition [KDD'18]



BiLSTM+CRF+Attention obtains best results

Extraction on new values is comparable to already known values

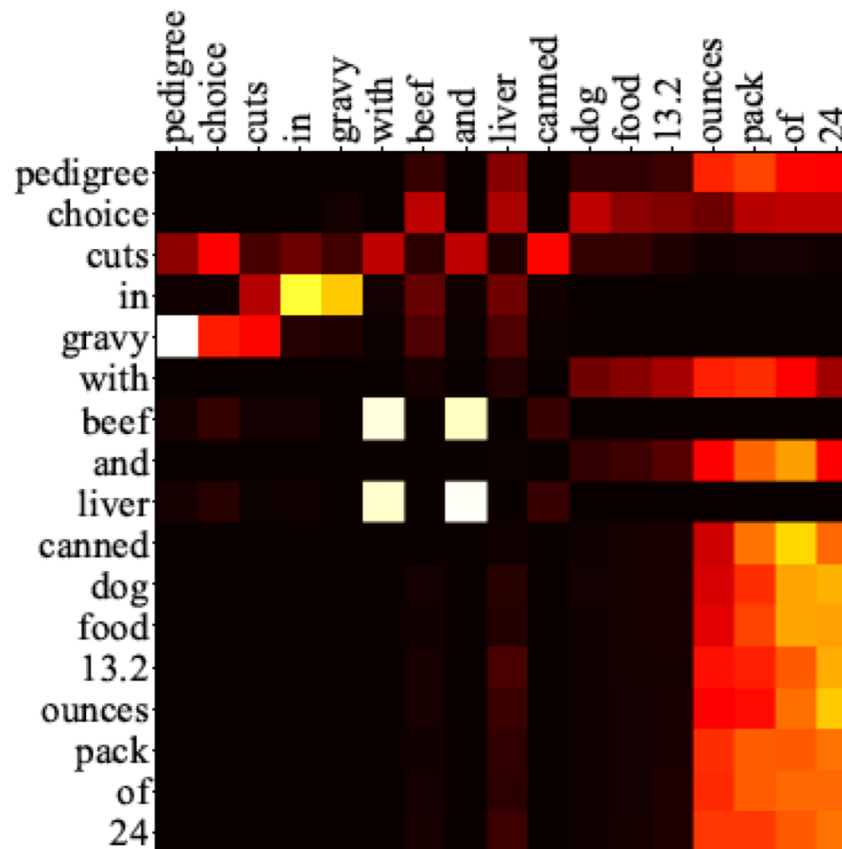
III. Open Attribute Extraction by Named Entity Recognition [KDD'18]



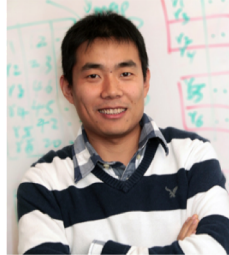
With active learning, we can obtain high precision and recall using only 190 training labels

III. Product Profile Extraction by NER —Which ML Model Works Best?

□ Recurrent Neural Network, CRF, Attention



III. Building a Shallow Graph from Product Profiles in Amazon Catalog



Product profile
extraction



Automatically
building a
shallow KG

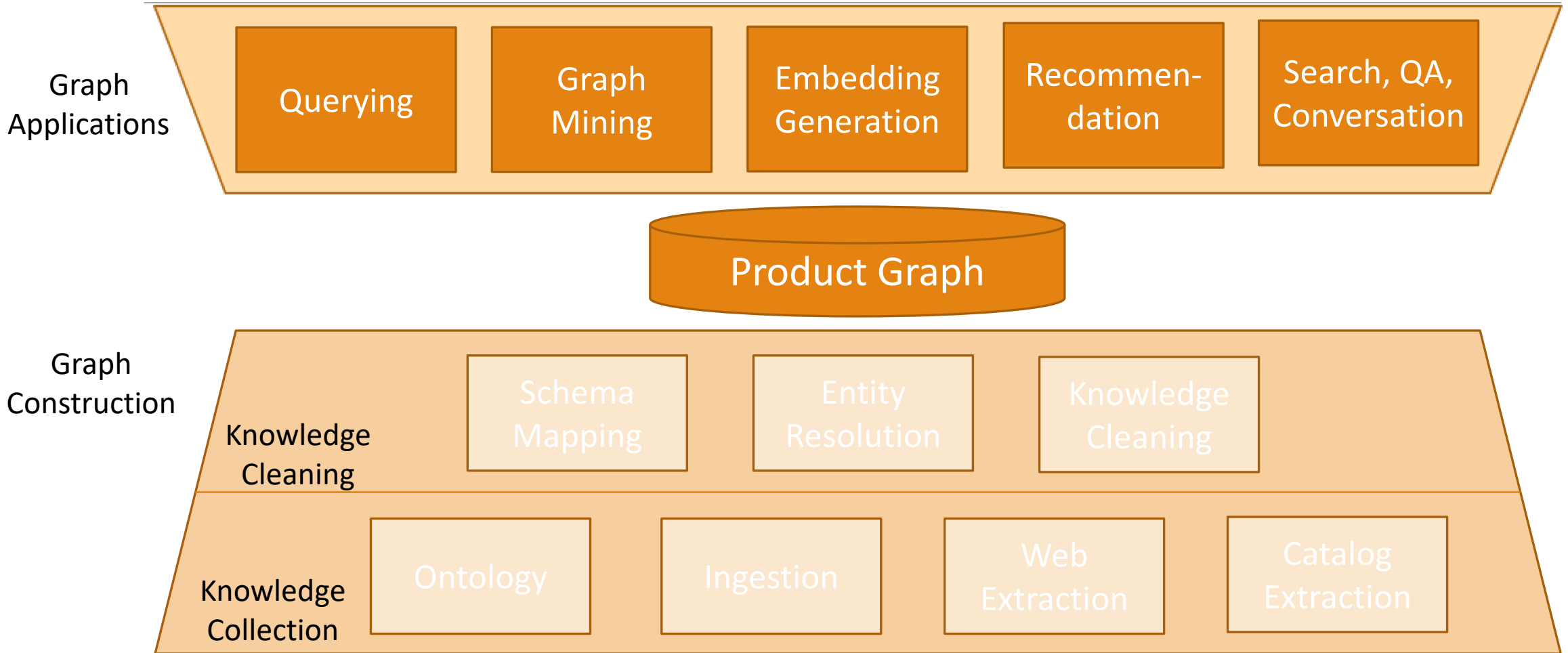


Open aspect
extraction



Review extraction
& sentiment analysis

IV. Graph Mining and Embedding



IV. Graph Mining

Which char more important?

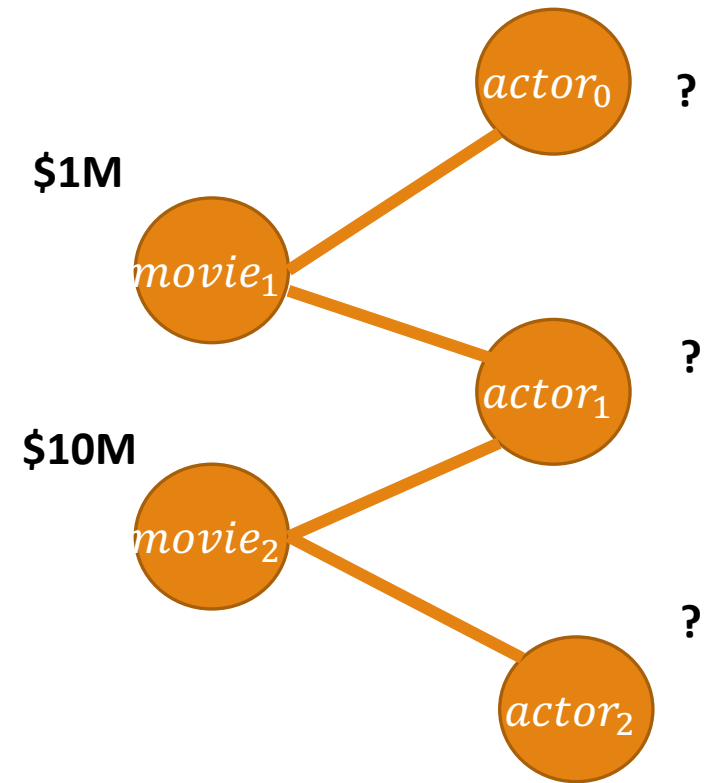


Trend of Darth Vader lamps?

Why people who bought this lamp also bought this chair?

IV. Finding Entity Importance

- Input
 - Knowledge Graph
 - Importance scores for some nodes (e.g., PageView for Wikipedia)
- Output: Importance scores for all nodes
- Method: PageRank w. Restart

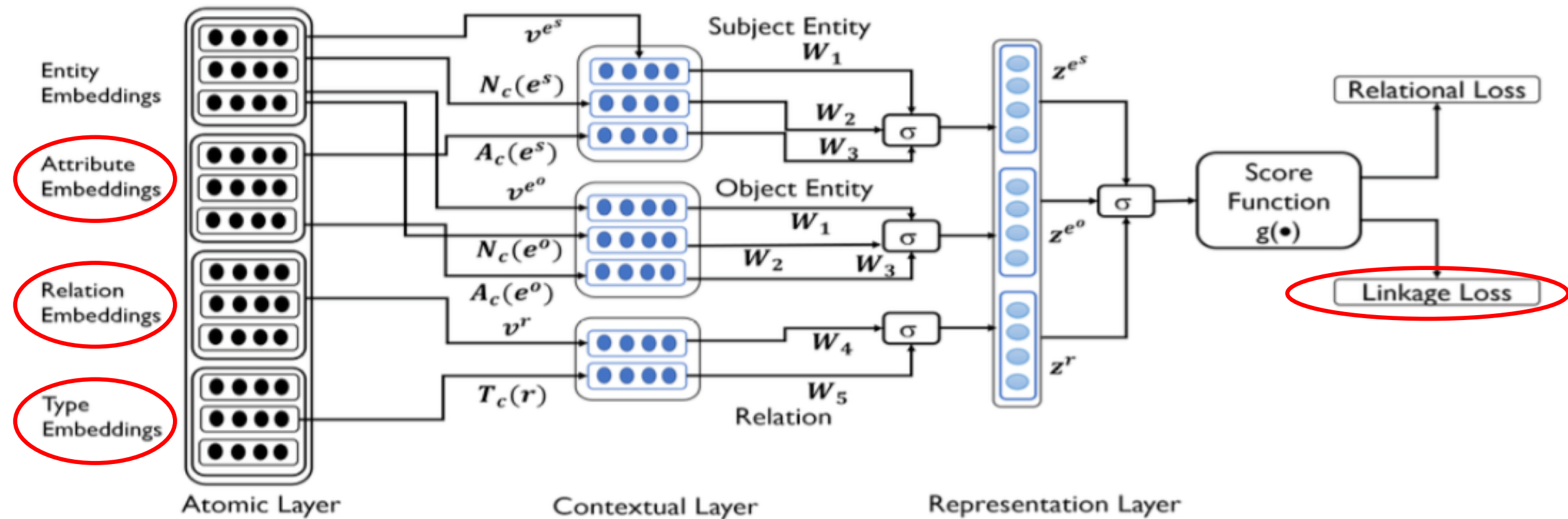


IV. Finding Entity Importance

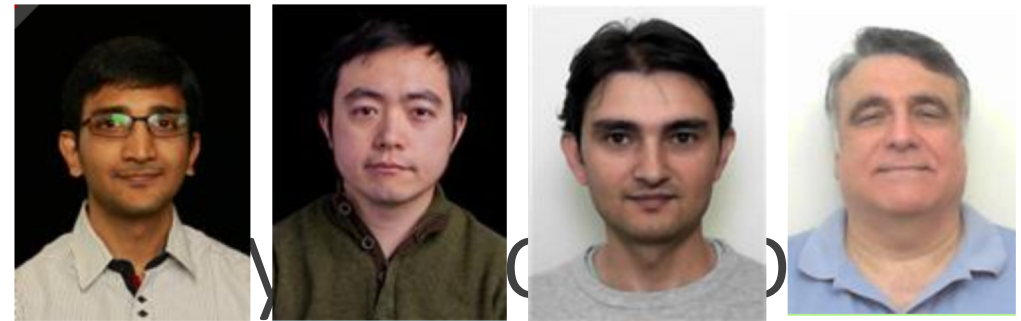


| <u>Actors</u> | <u>Directors</u> | <u>Countries</u> | <u>Companies</u> |
|---------------------------------|----------------------------------|---------------------------------|--------------------------------------|
| 133, "Samuel L. Jackson", actor | 48, "Woody Allen", Director | 7585, "U.S.A.", country | 667, "Warner Bros.", company |
| 123, "Robert De Niro", actor | 43, "Steven Spielberg", Director | 1120, "United Kingdom", country | 630, "Paramount Pictures", company |
| 109, "Morgan Freeman", actor | 38, "Ridley Scott", Director | 601, "Germany", country | 547, "Universal Pictures", company |
| 99, "Owen Wilson", actor | 36, "Steven Soderbergh" | 571, "France", country | 431, "20 Century Fox", company |
| 95, "Susan Sarandon", actor | 36, "Renny Harlin", Director | 448, "Canada", country | 386, "Columbia Pictures", company |
| 93, "Brad Pitt", actor | 35, "Spike Lee", Director | 215, "Australia", country | 317, "New Line Cinema", company |
| 85, "Steve Buscemi", actor | 35, "Martin Scorsese", Director | 137, "Spain", country | 227, "Walt Disney Pictures", company |

IV. Graph Embedding [ACL'18]



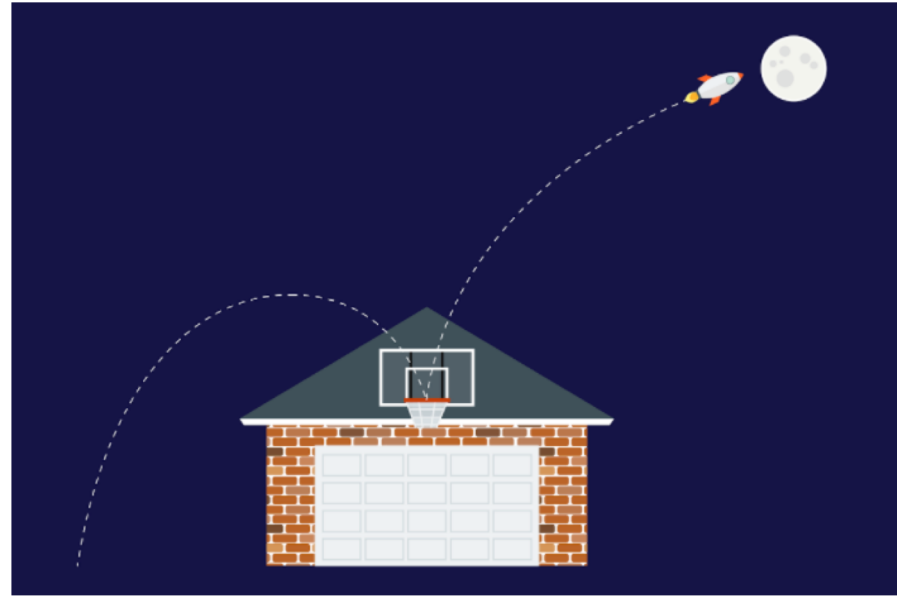
IV. Graph Embedding for



| | Link prediction | | Entity resolution |
|-----------------|-----------------|--------------|--|
| | IMDB-MRR | FB-MRR | AUPRC |
| RASCAL | 0.592 | 0.147 | 0.327 |
| DistMult | 0.691 | 0.556 | 0.292 |
| ComplEx | 0.752 | 0.629 | 0.359 |
| STransE | 0.421 | 0.397 | 0.231 |
| GAKE | 0.114 | 0.093 | 0.457 |
| Ours | 0.733 | 0.677 | 0.553 (unsuper) / 0.691 (super) |

Take Aways

- ❑ We aim at building an authoritative knowledge graph for all products in the world
- ❑ The next-generation of KG could be a combination of **rich** graph and **broad** graph
- ❑ We shoot for roofshot and moonshot goals to realize our mission



Thank You!