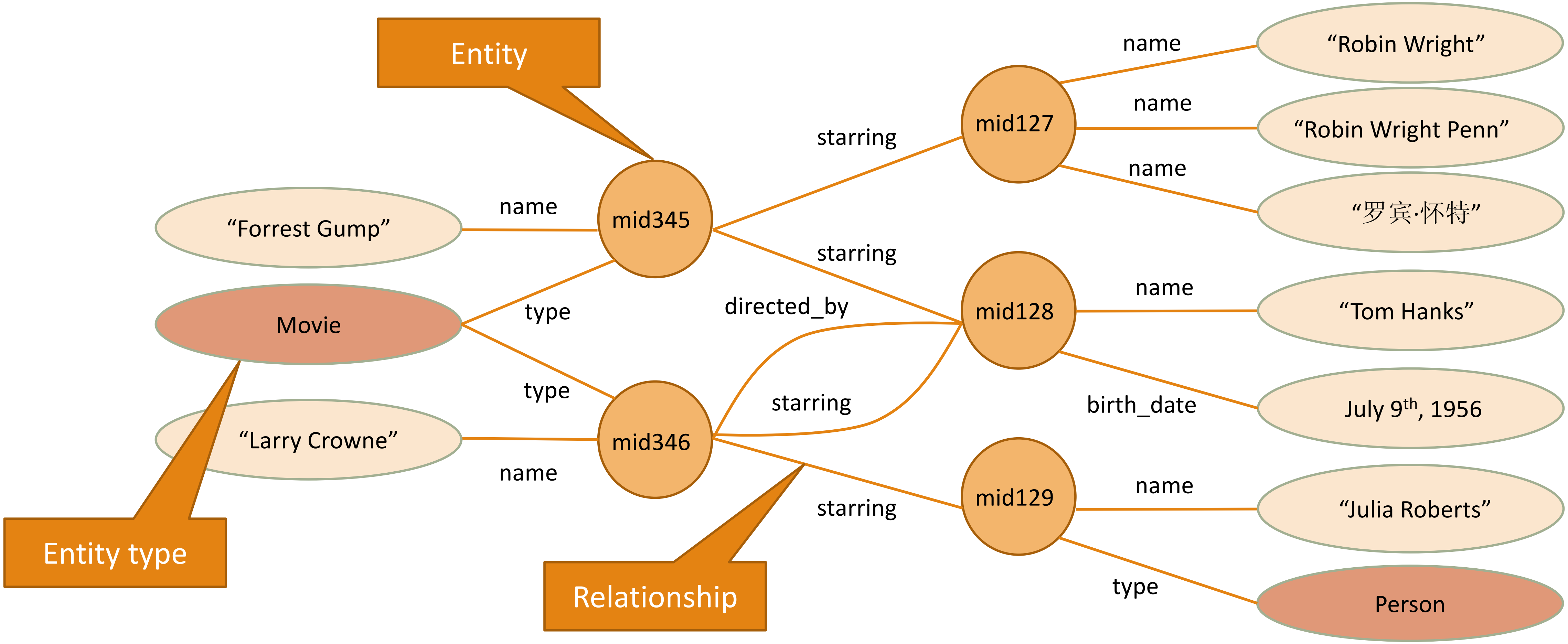
Challenges and Innovations in Building a Product Knowledge Graph

XIN LUNA DONG, AMAZON

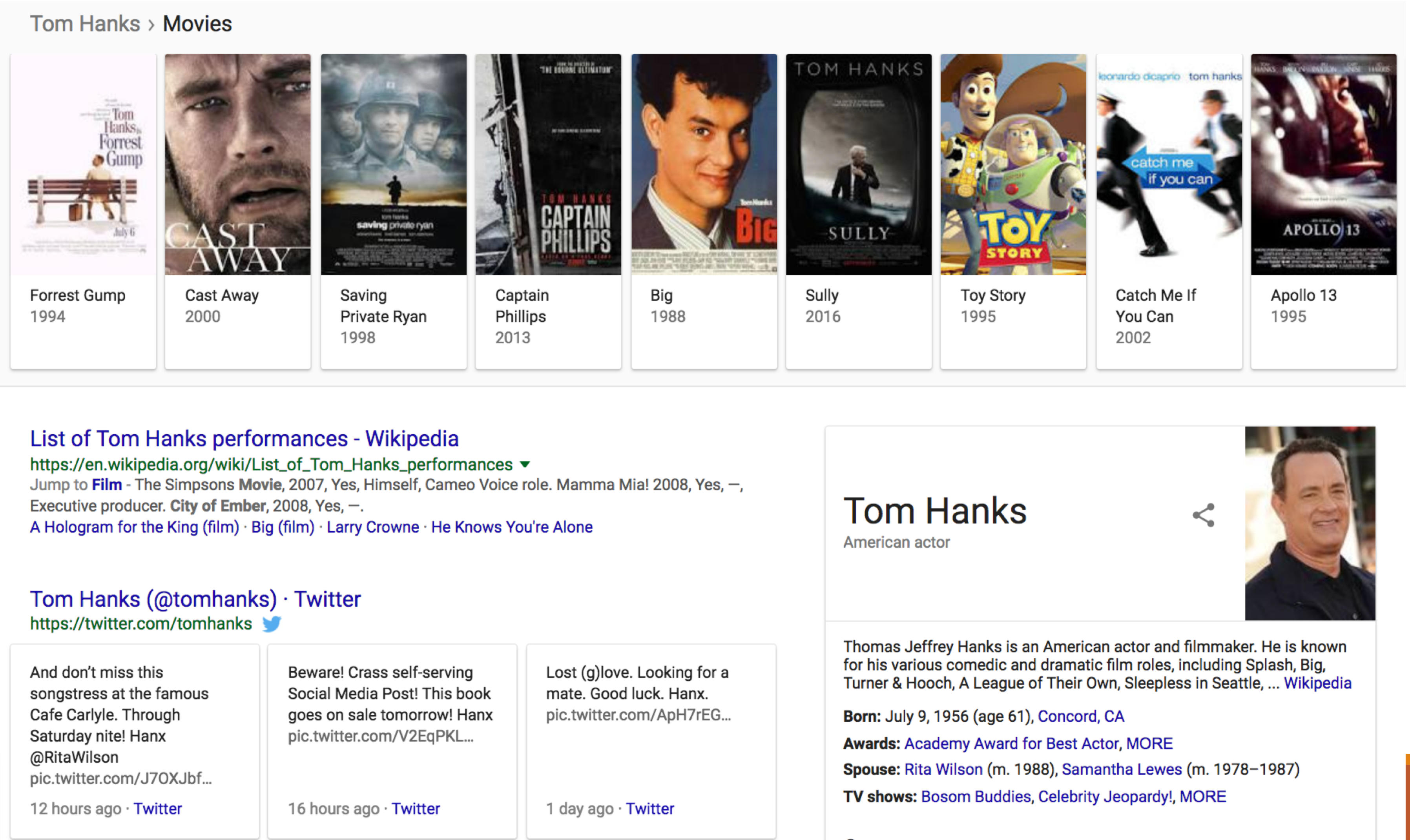
JUNE, 2017

Product Graph vs. Knowledge Graph

Knowledge Graph Example for 2 Movies



Knowledge Graph in Search



Knowledge Graph in Personal Assistant

Alexa, play the music
by Michael Jackson

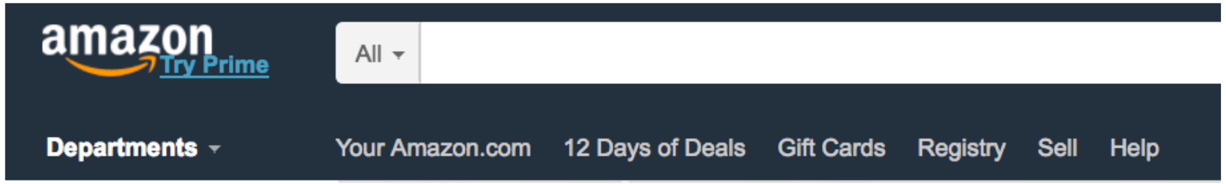
List of officially released compilations and

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

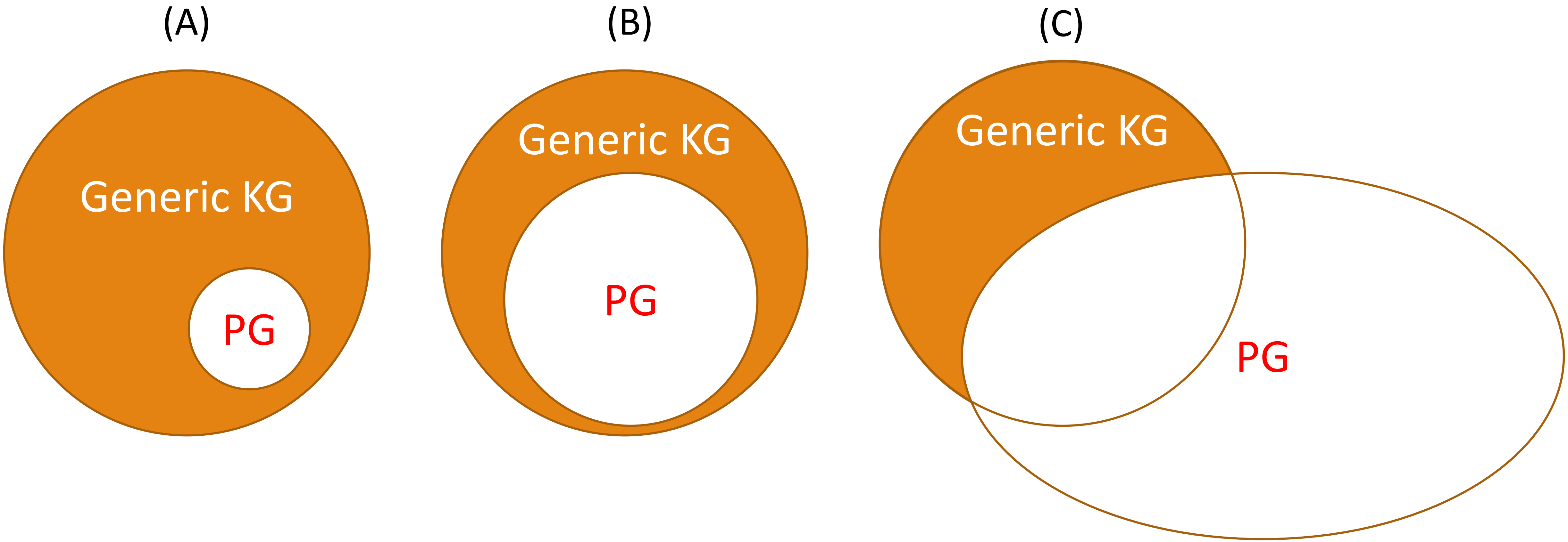
- *Portrait of Michael Jackson / Portrait of Jackson 5* (1973)
- *Os Grandes Sucessos, Vol. 2* (1980)
- *Motown Superstar Series, Vol. 7* (1980)
- *Superstar* (1980)
- *Michael Jackson & The Jackson 5* (1983)
- *Ain't No Sunshine* (1984)
- *The Great Love Songs of Michael Jackson* (1984)
- *Ben / Got 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

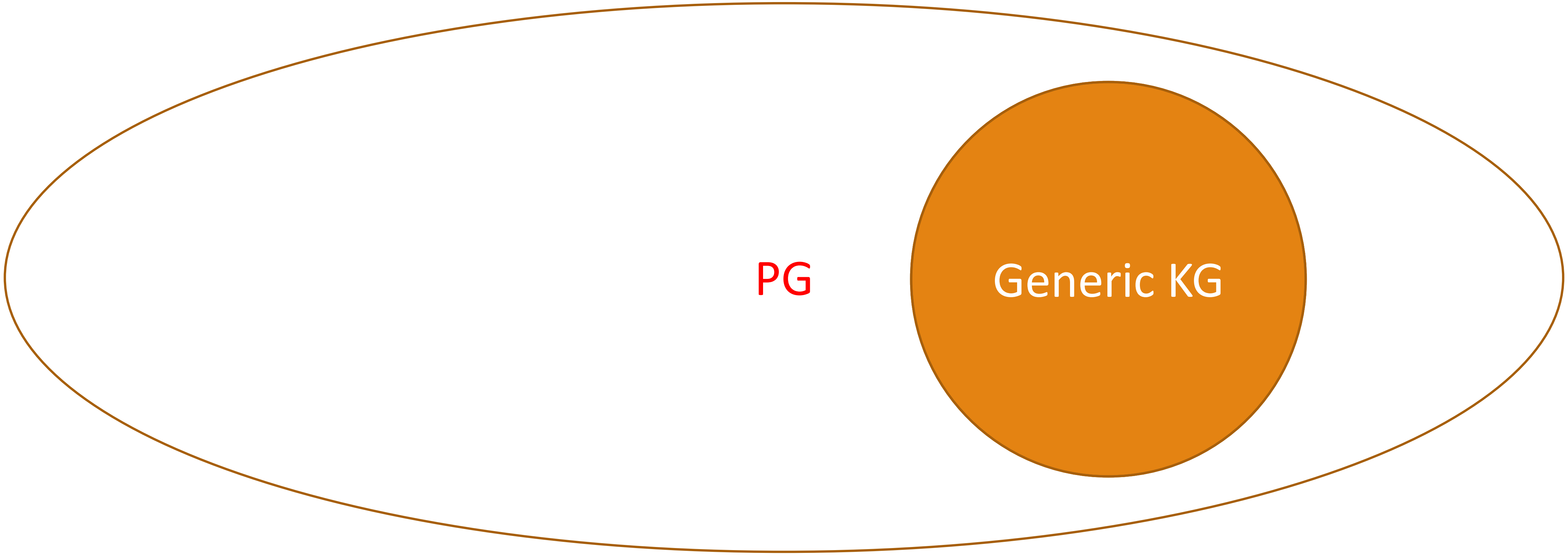


Product Graph vs. Knowledge Graph

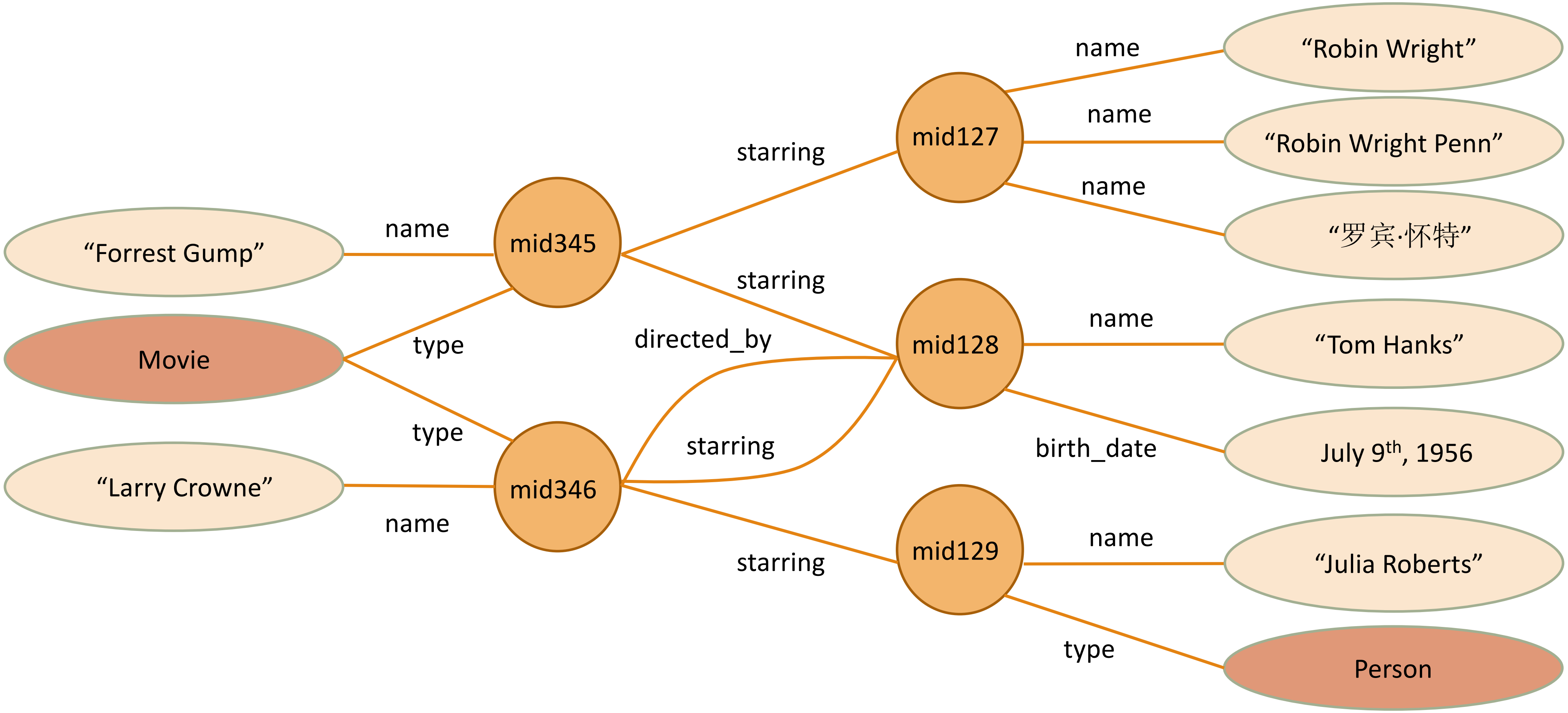


Product Graph vs. Knowledge Graph

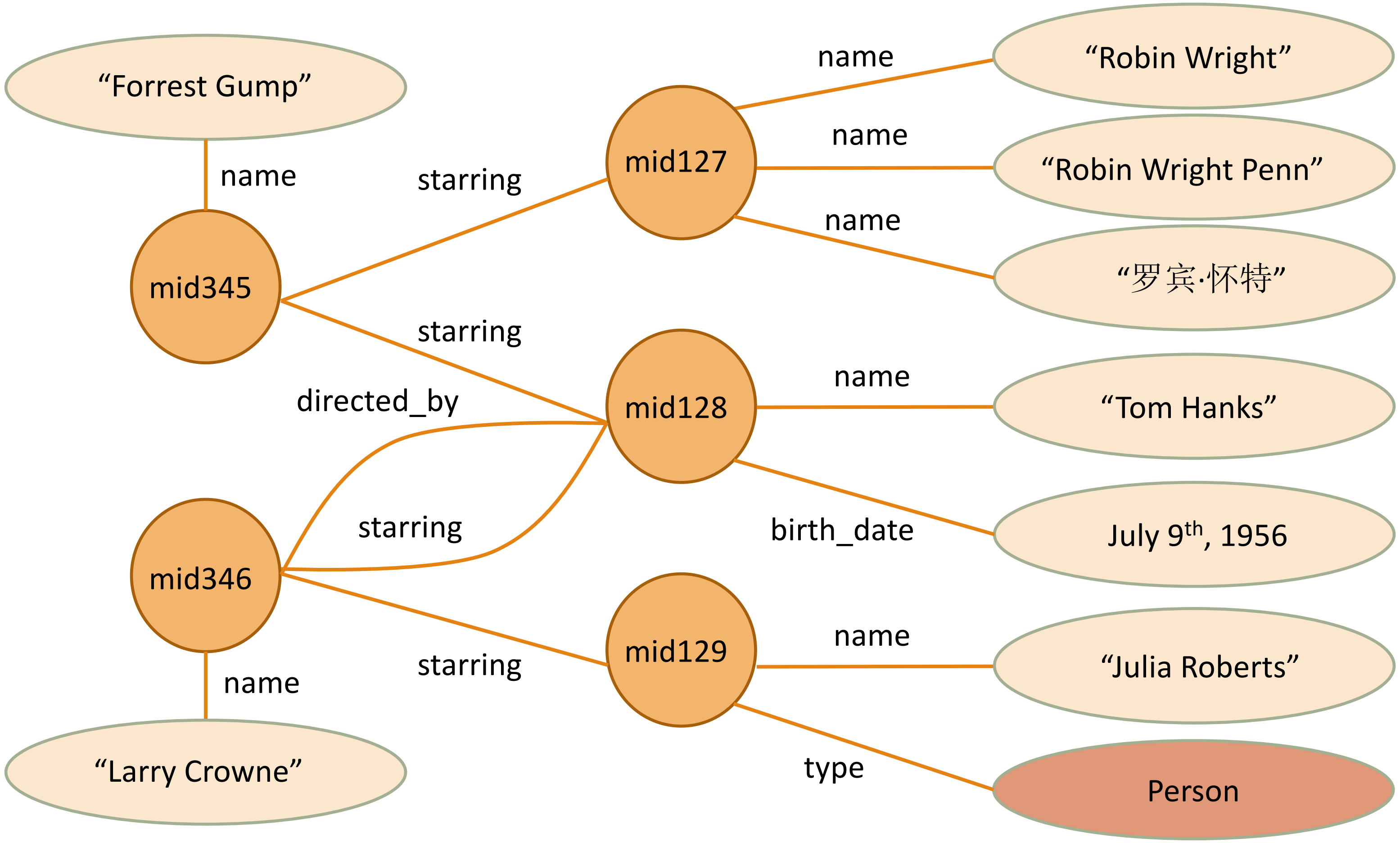
(D)



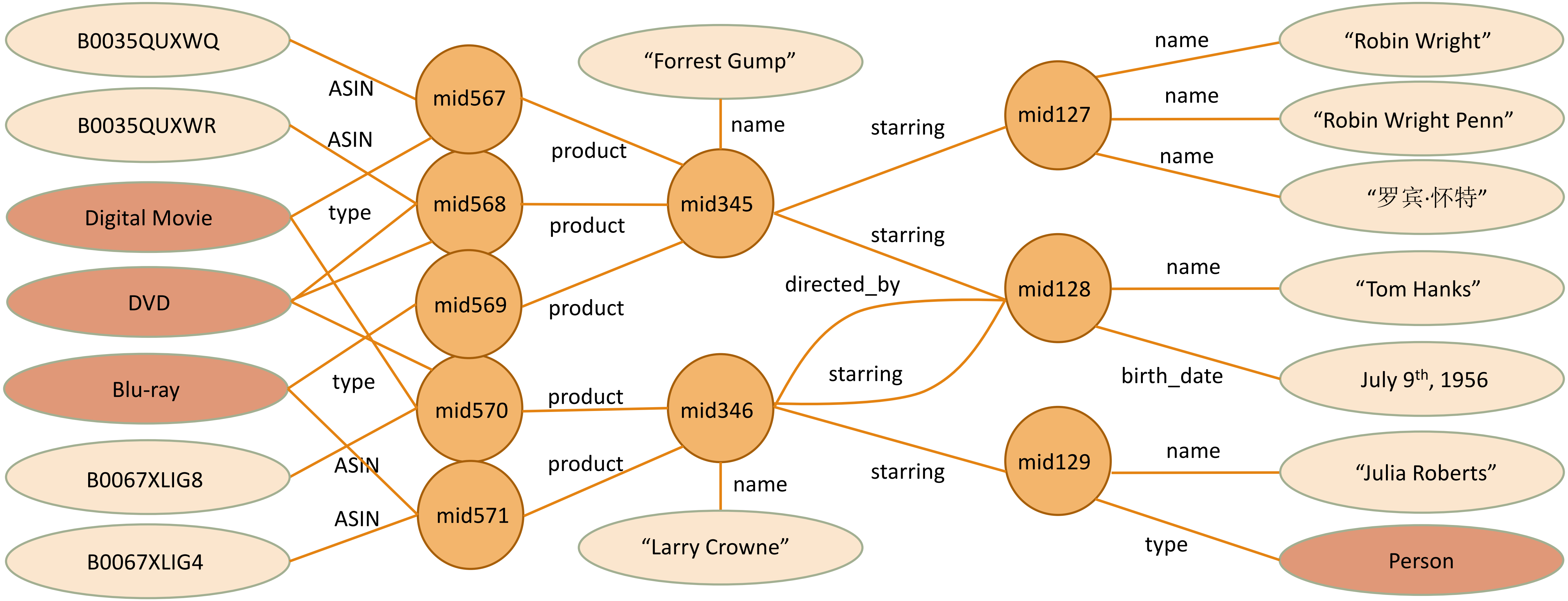
Knowledge Graph Example for 2 Movies



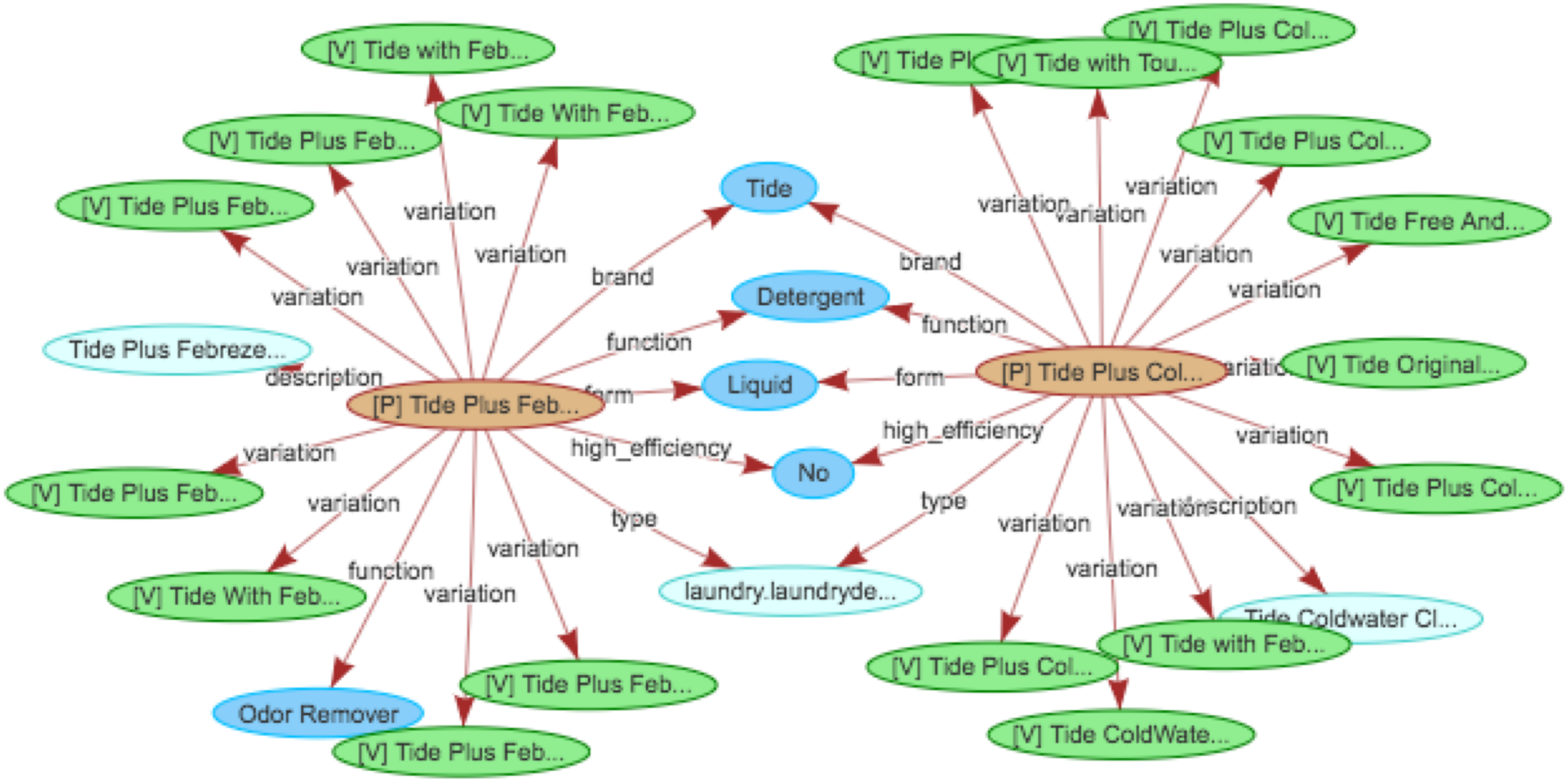
Product Graph vs. Knowledge Graph



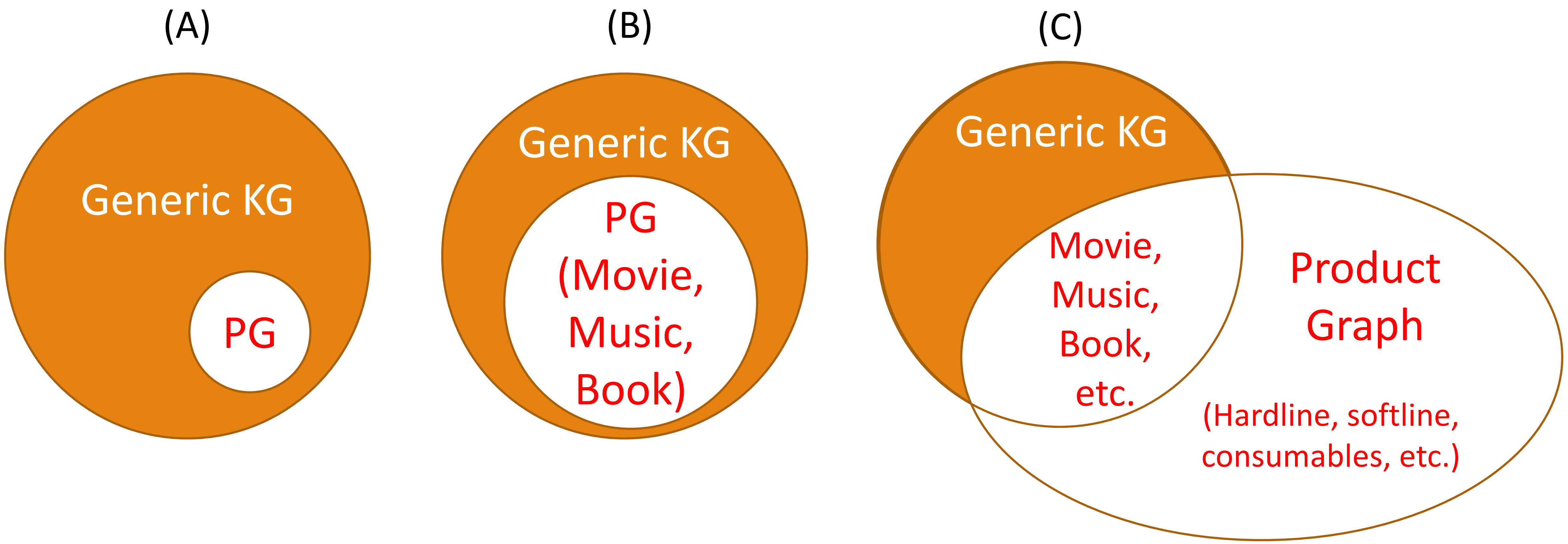
Product Graph vs. Knowledge Graph

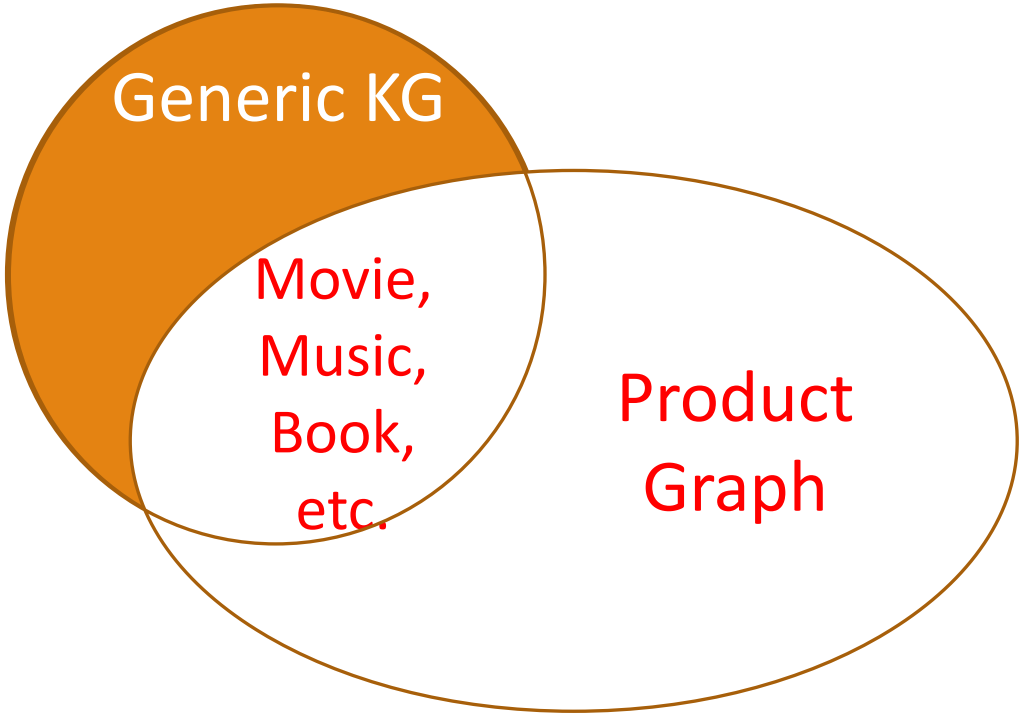


Another Example of Product Graph



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

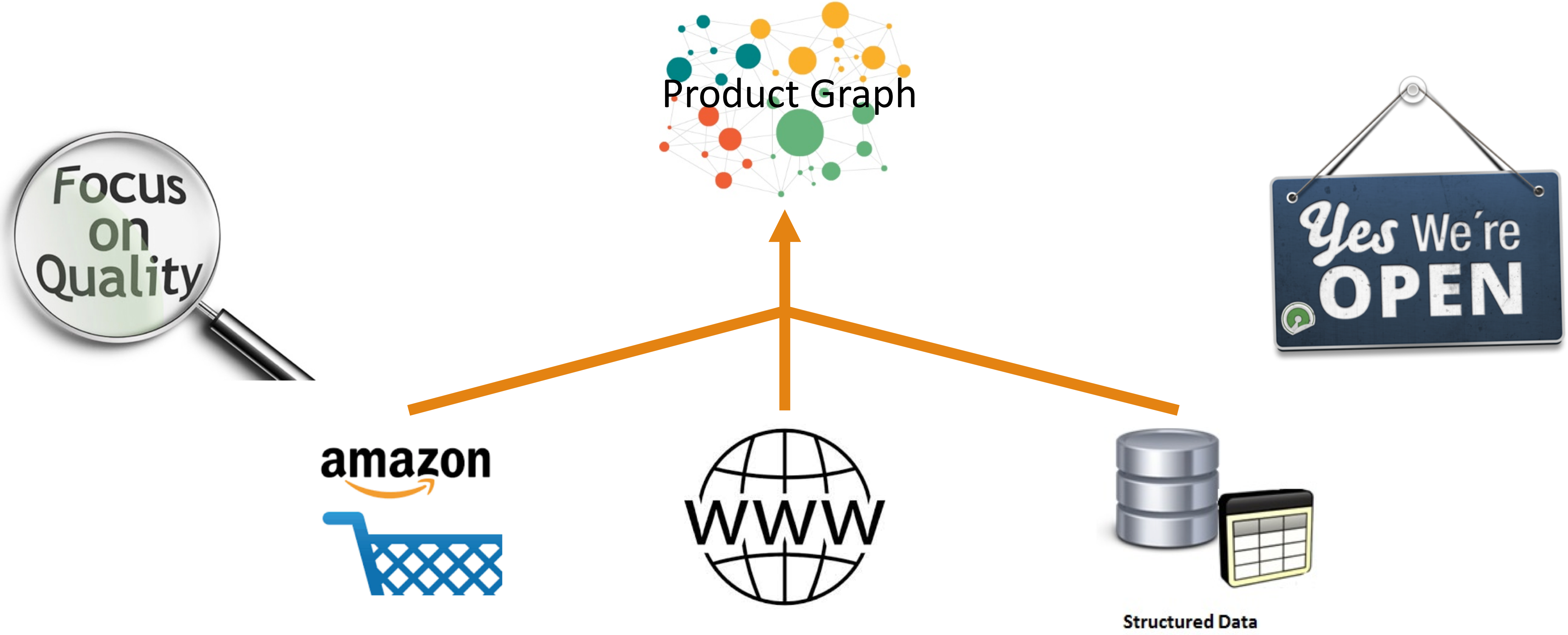
- Large number of product categories
  - A lot of work to manually define ontology
  - Hard to catch the trend of new product categories and properties

Challenges in Building Product Graph IV

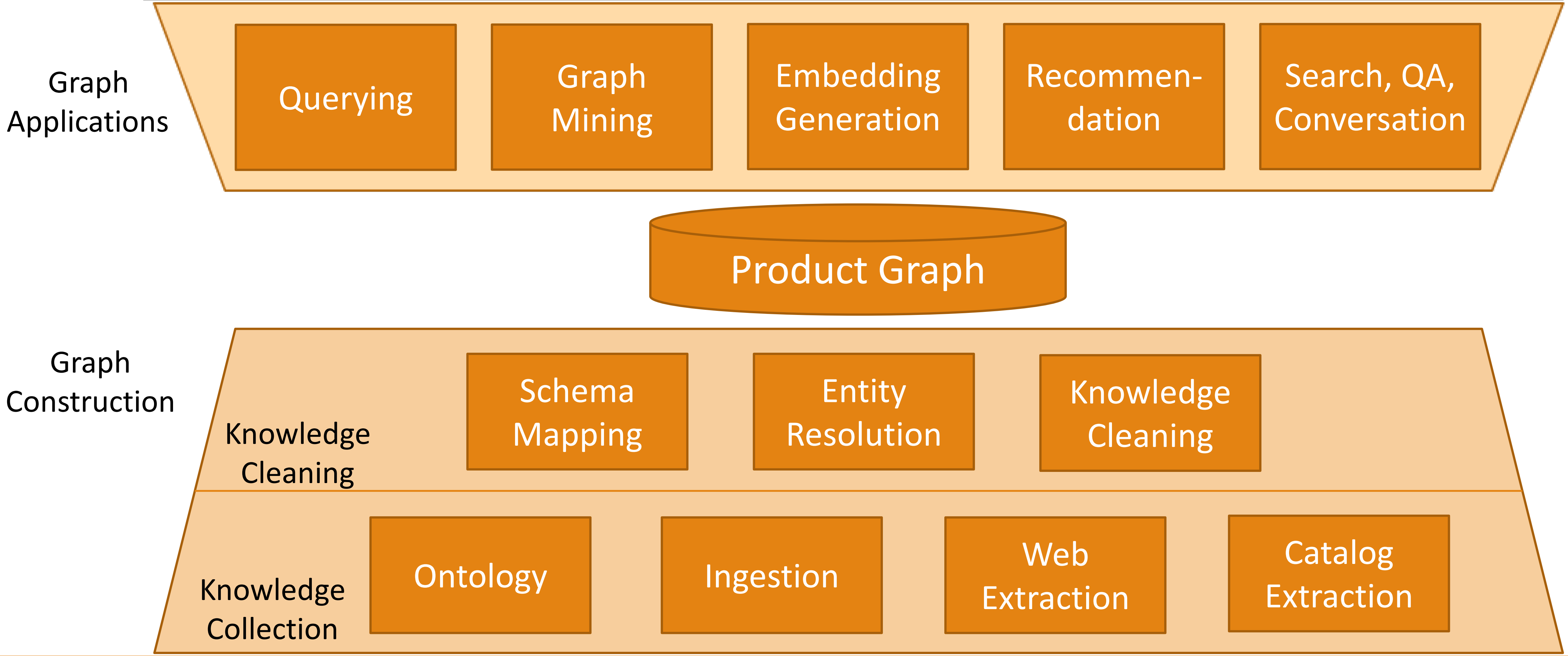
- Many entities are not named entities
  - Named Entity Recognition does not apply
  - New challenges for extraction, linking, and search

How to Build a Product Graph?

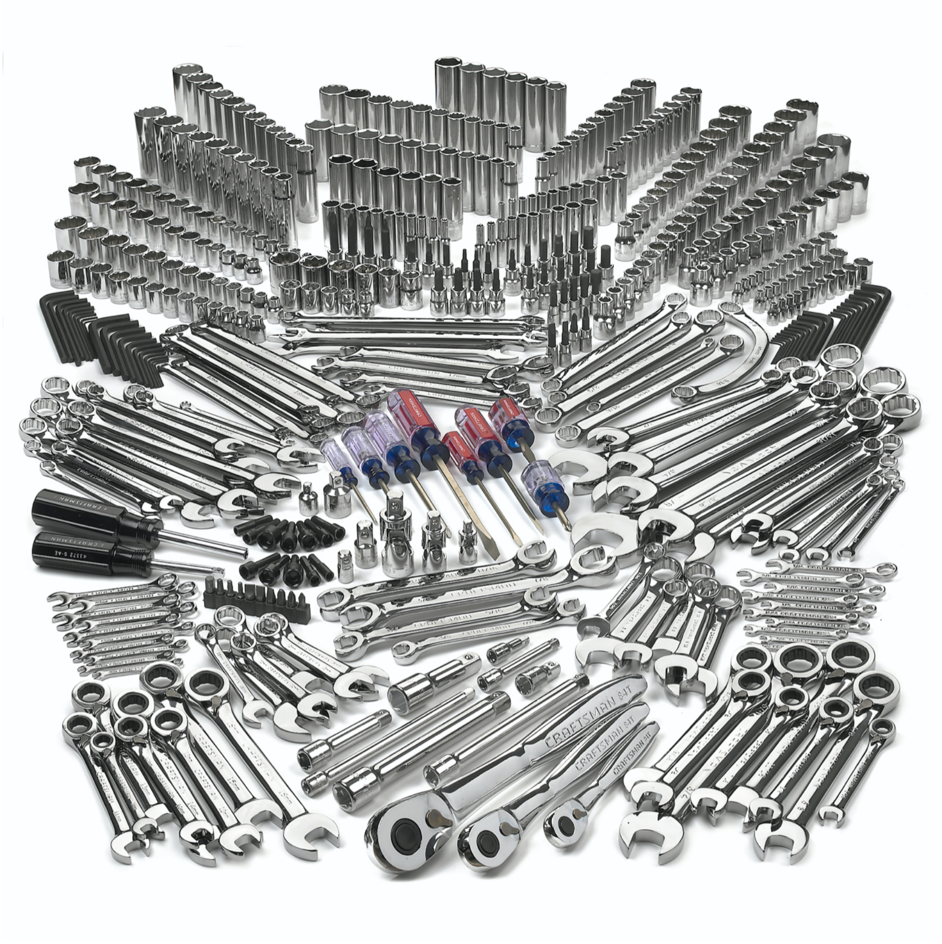
Where is Knowledge from?



Architecture



Which ML Model Works Best?

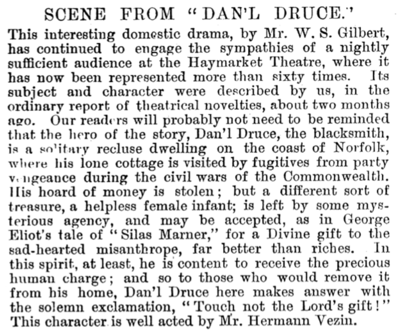


Which ML Model Works Best?

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

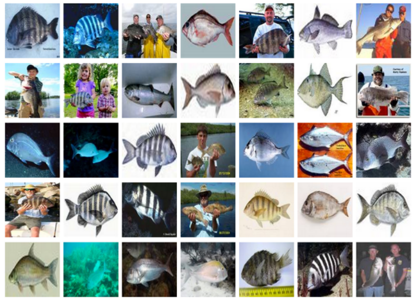
Tree-based models

??



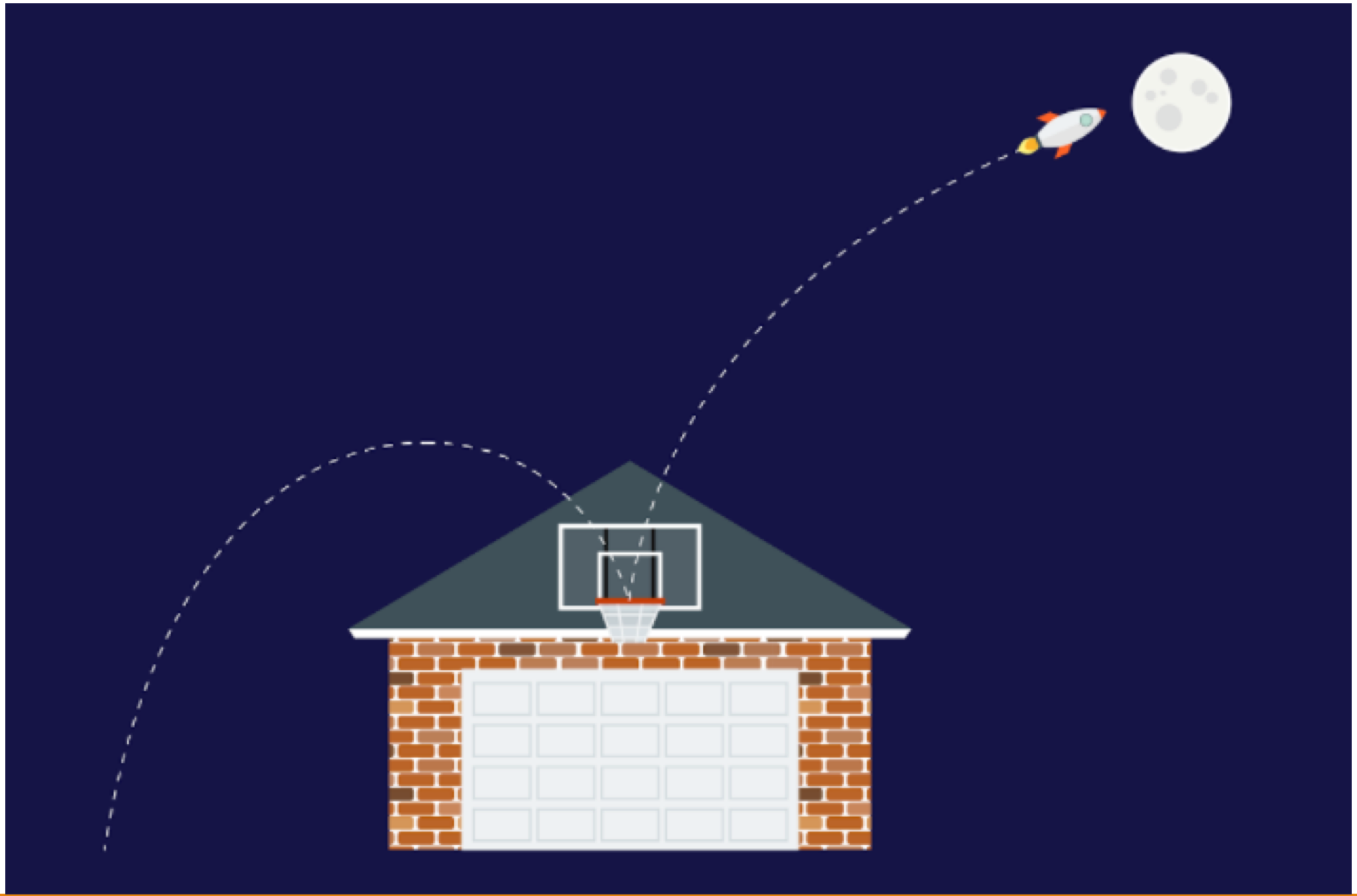
SCENE FROM "DAN'L DRUCE."

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 Vezin.



Neural network

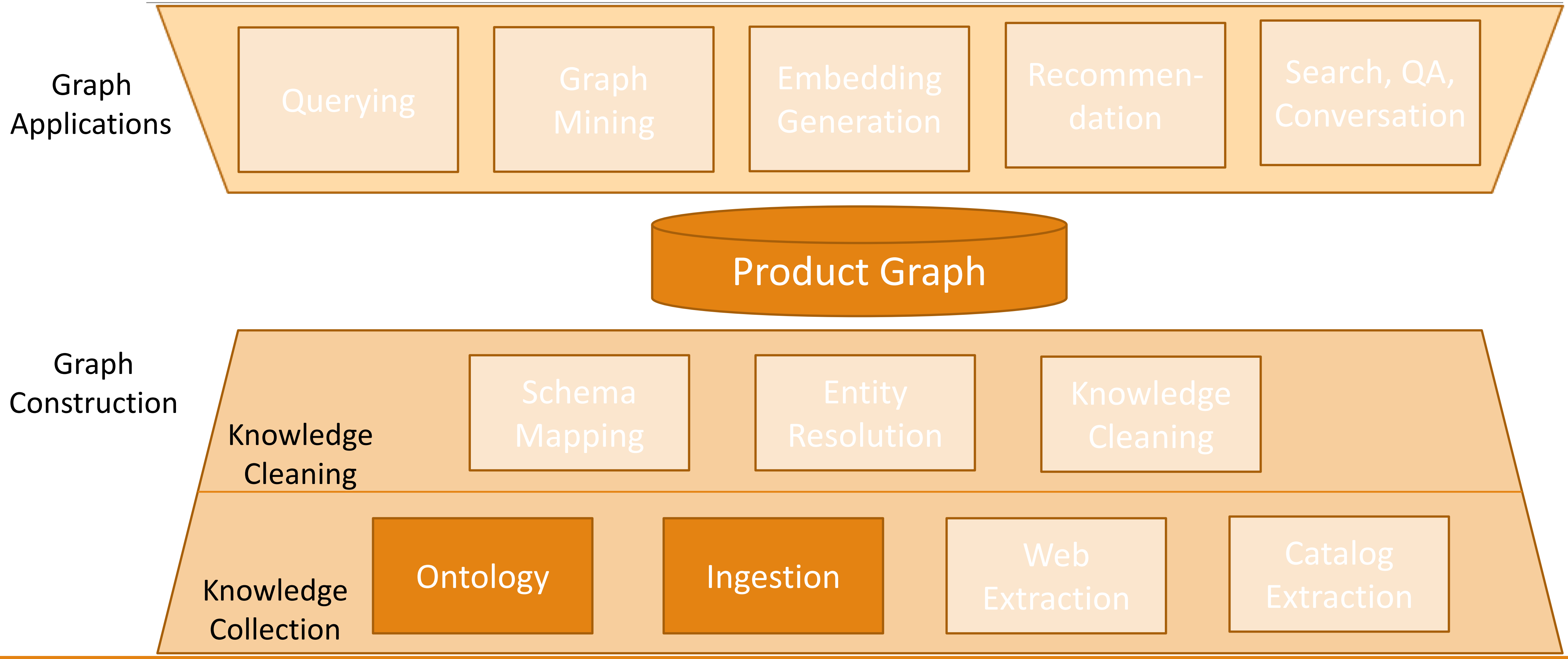
Research Philosophy



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

Roofshots: Deliver incrementally and make production impacts

I. Integrating Knowledge from Structured Sources

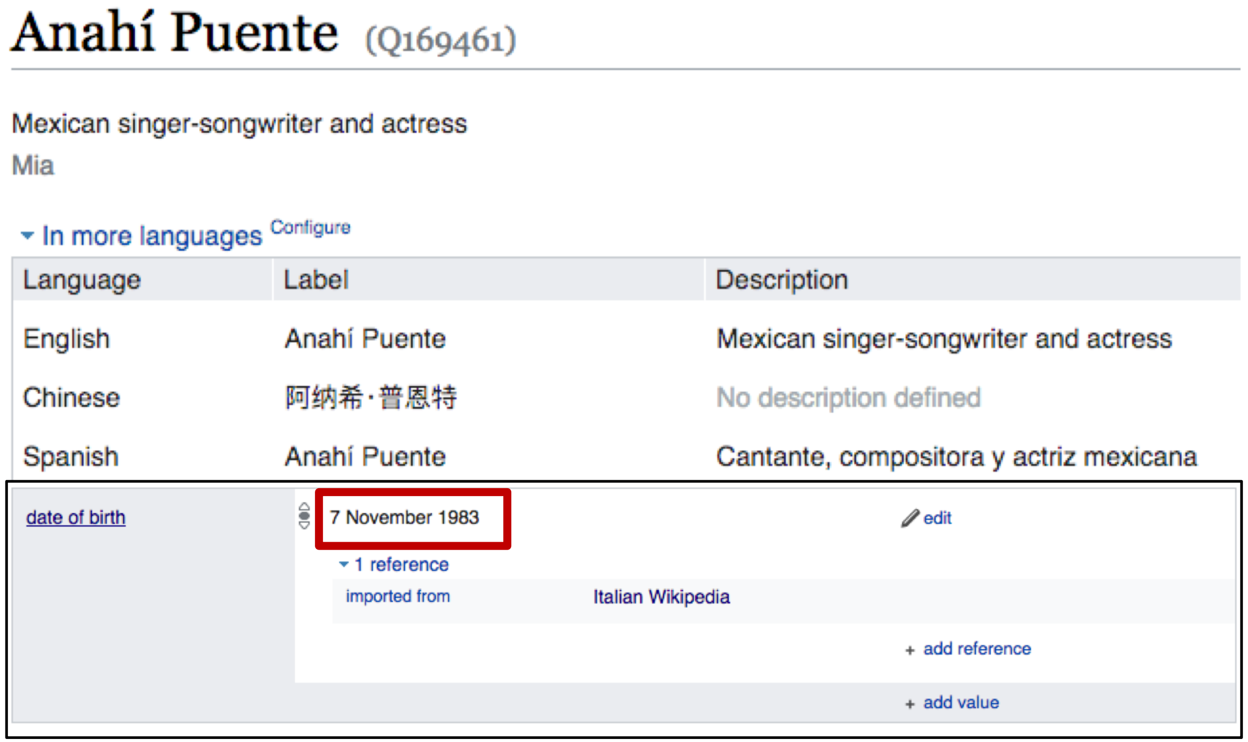


Challenges: Linkage & Quality

IMDB

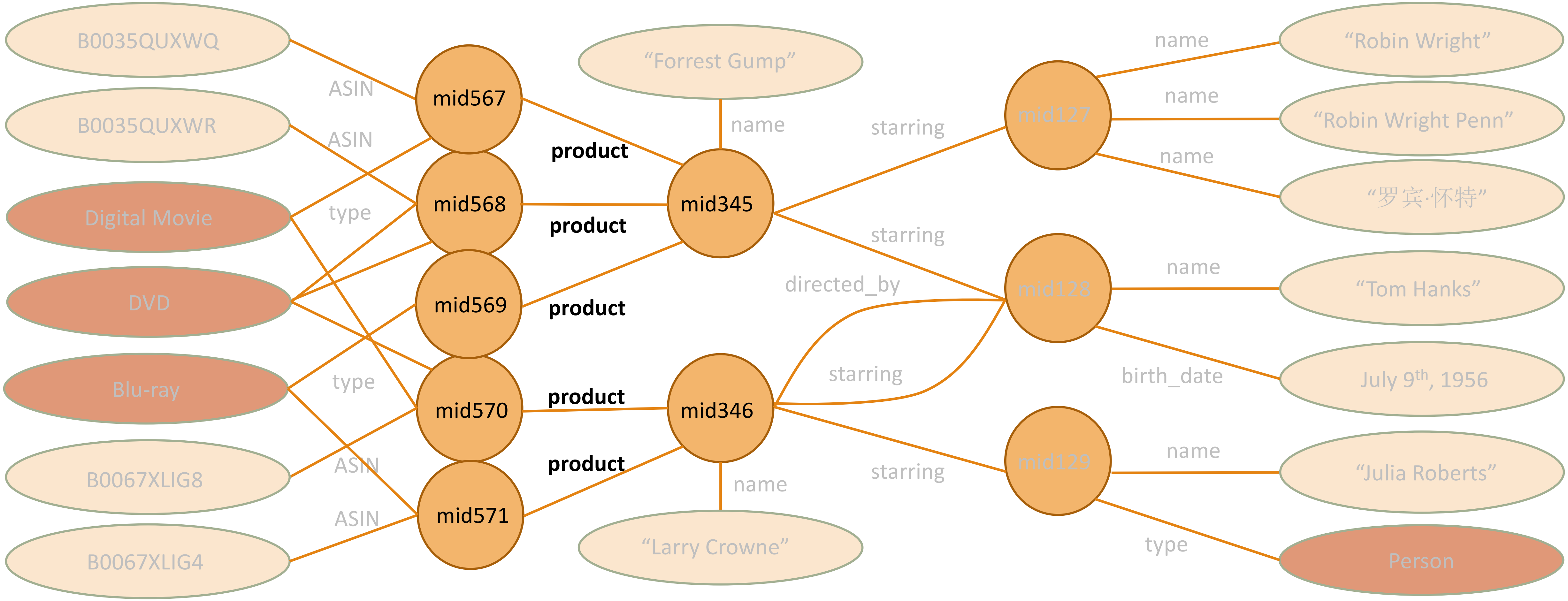
Same entity?

WikiData

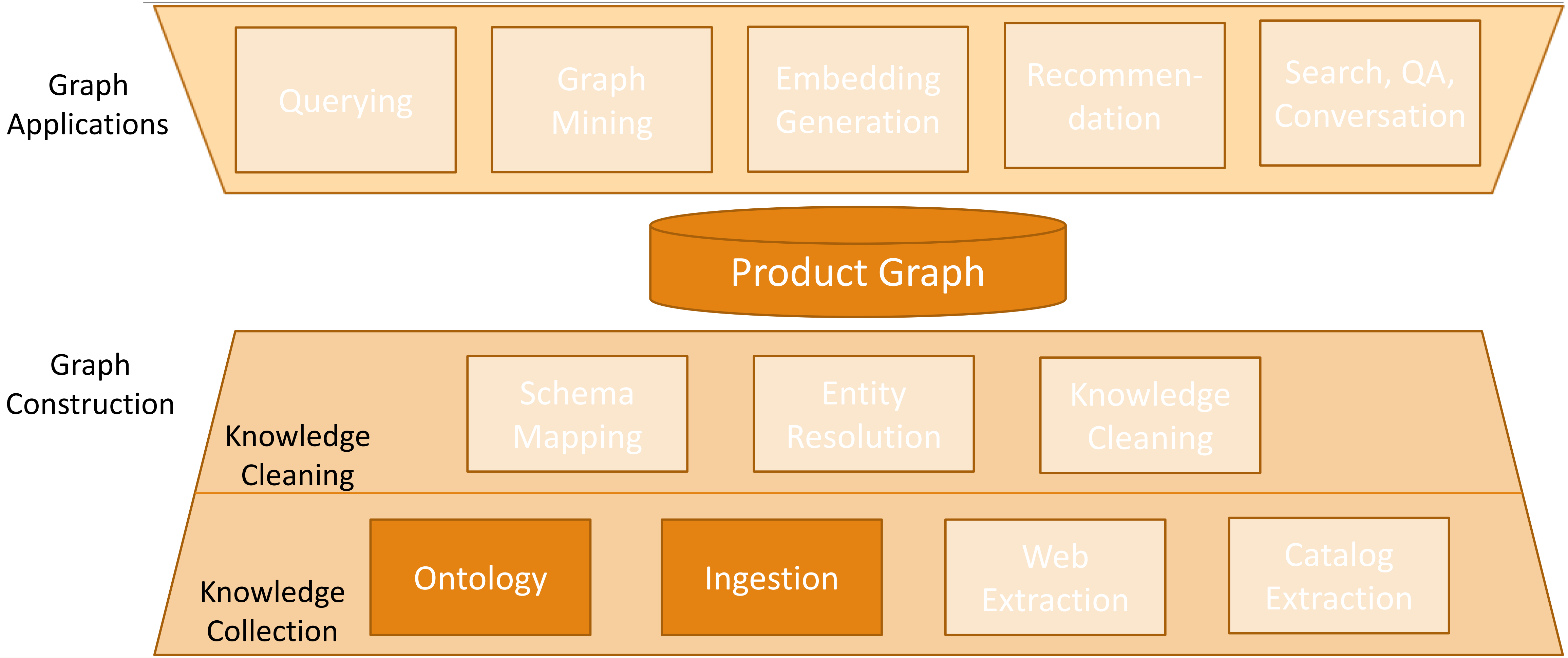


Which BirthDate is correct?

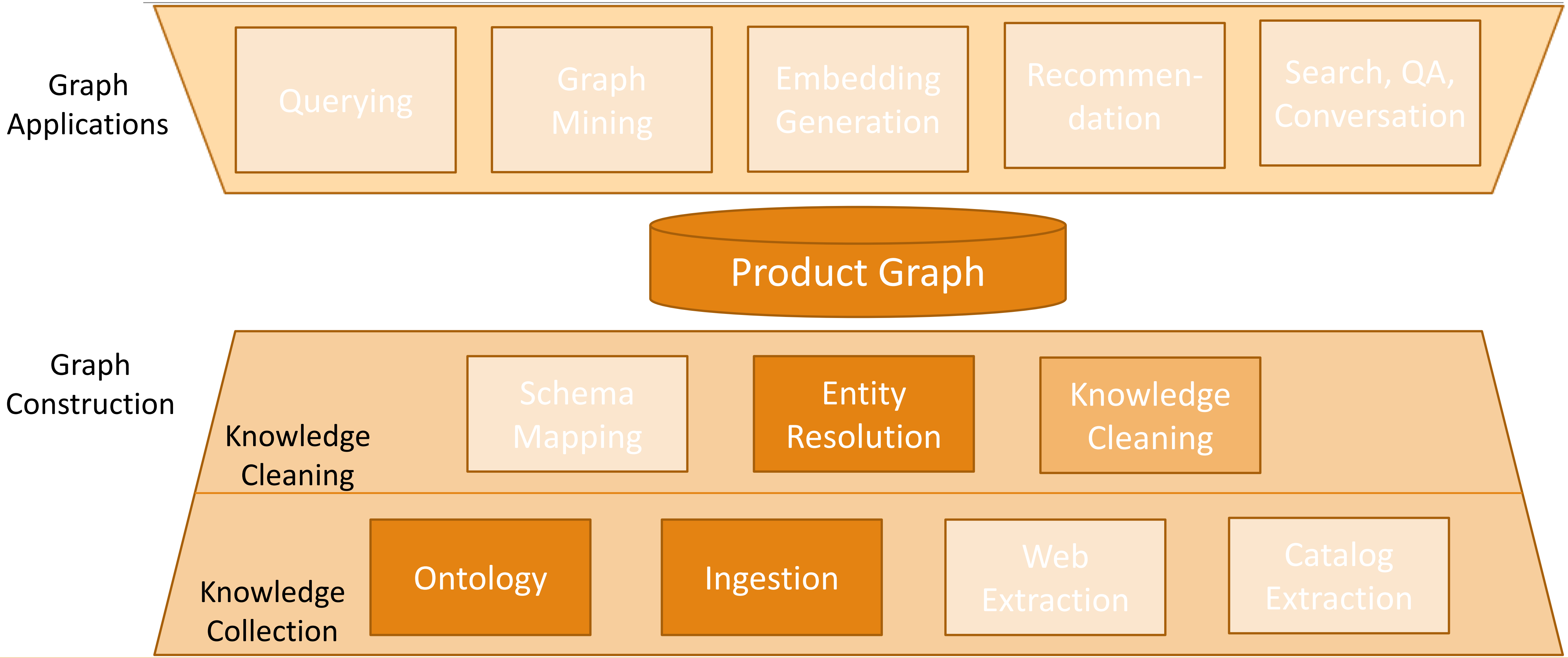
Challenges: Linkage



I. Integrating Knowledge from Structured Sources

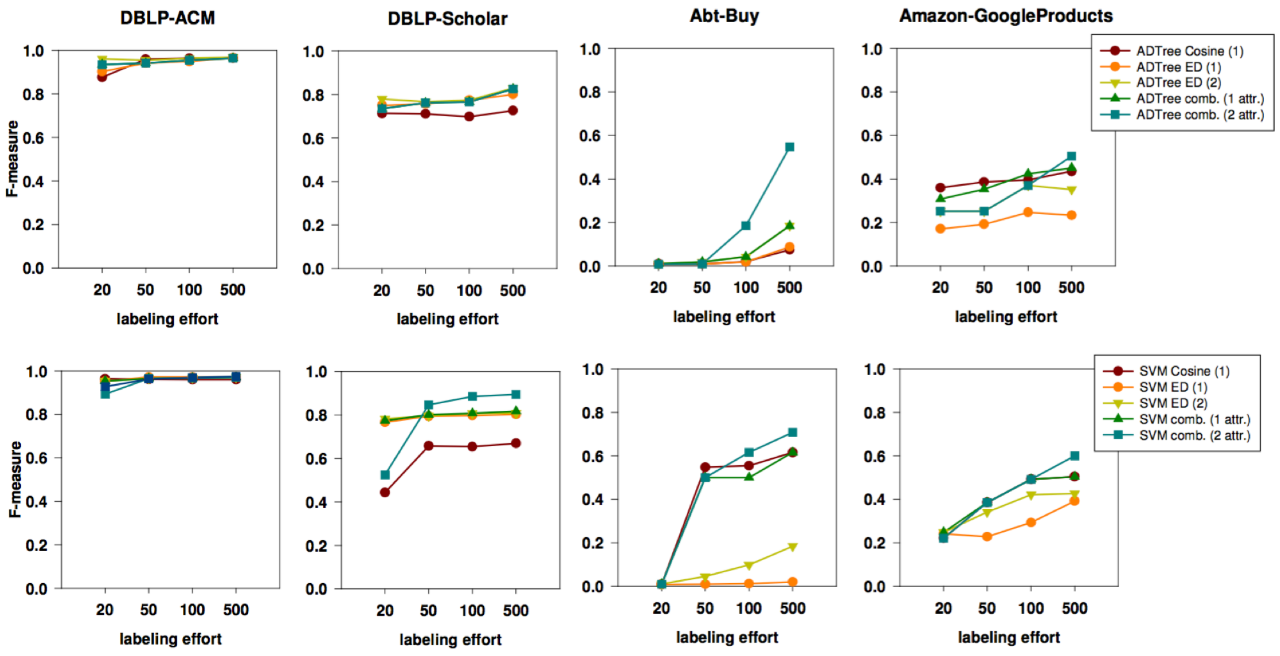


I. Integrating Knowledge from Structured Sources



I. Integrating Knowledge from Structured Sources—Entity Resolution

❑Published results



[Köpcke et al, VLDB 2010]

Dataset	Accuracy (%)			Cost (# Questions)
	P	R	F_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

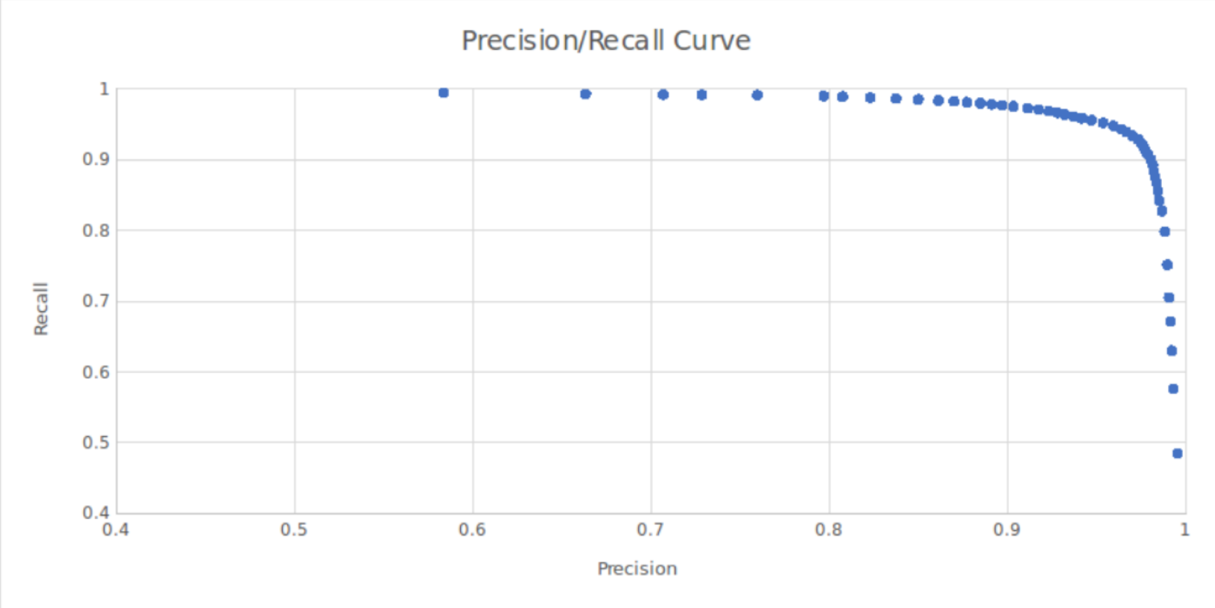
- Our method:
  - Model: Random forest
  - Features: Attribute similarity—various string similarity, number similarity

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

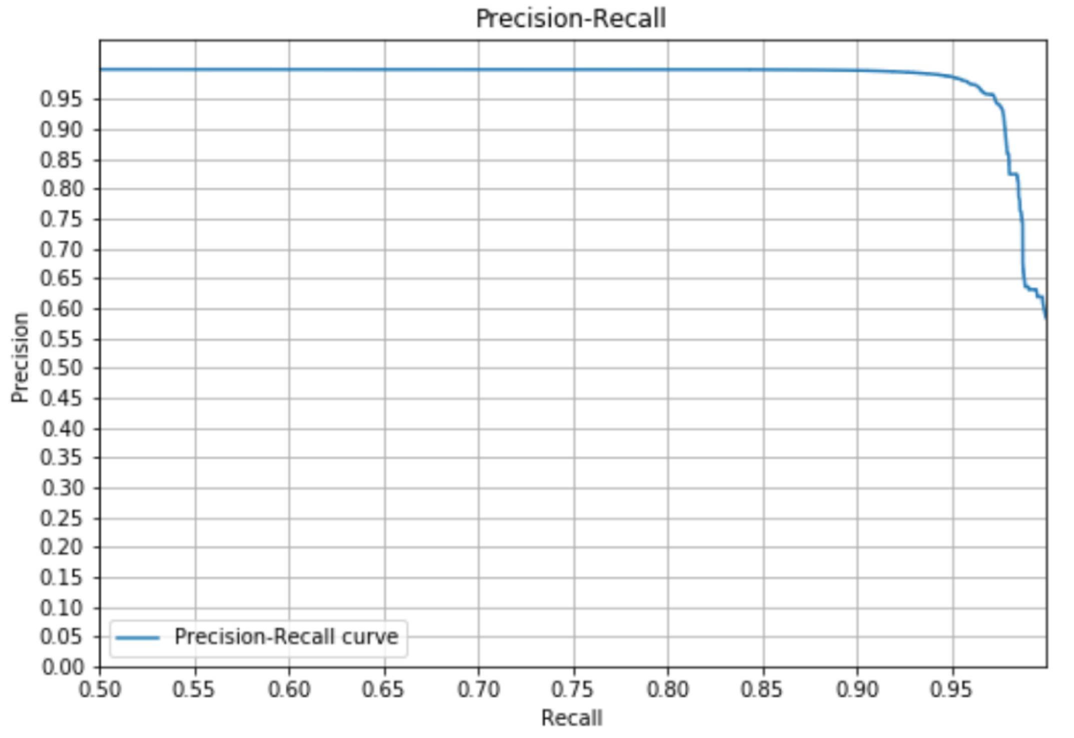
- Our method: Random forest on attribute-wise similarity
- Results between Amazon Movies and IMDb:

  200K labels

  ~150 features

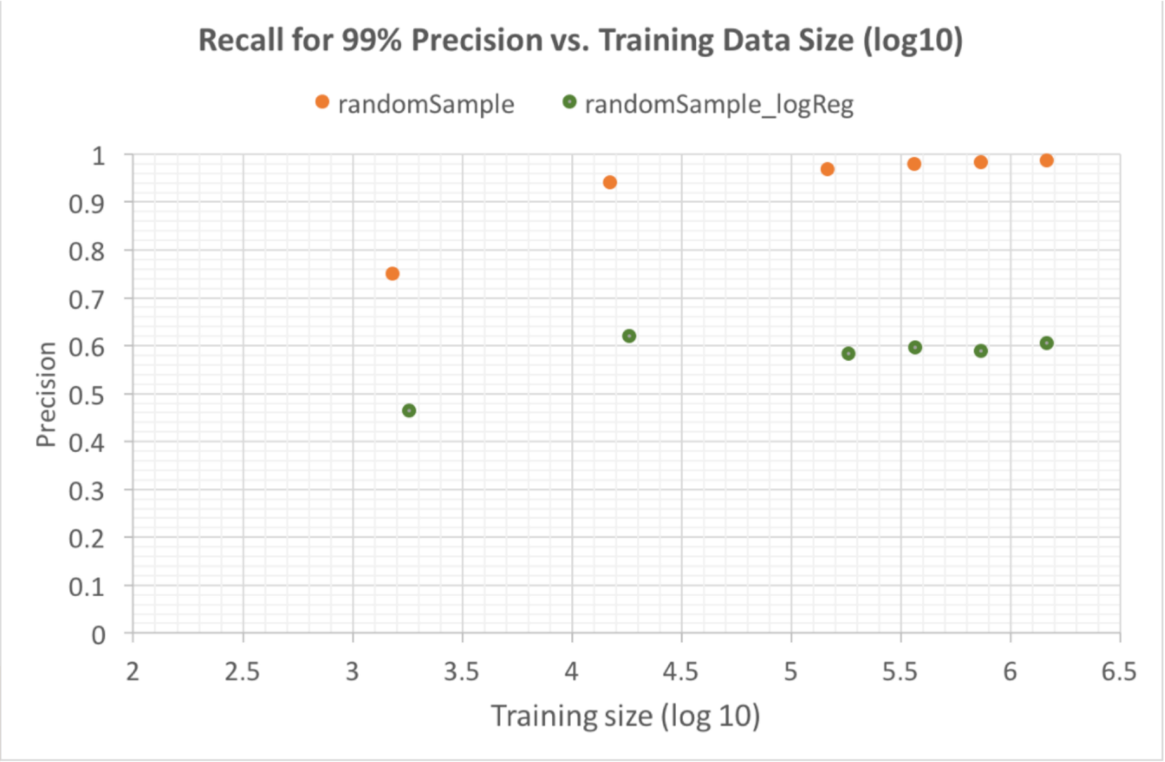
  AUPRC=0.9923

  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

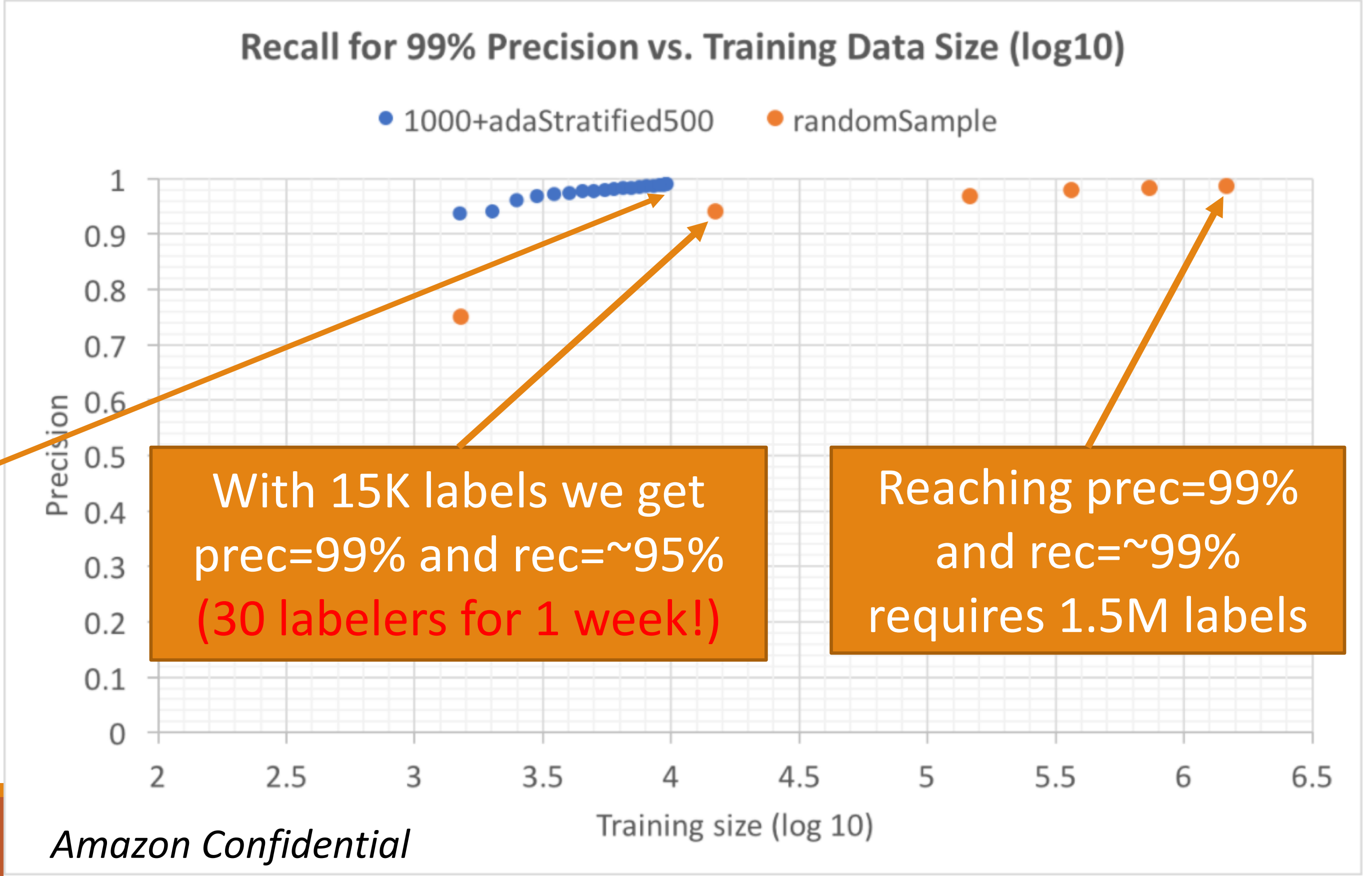


I. Integrating Knowledge—Entity Resolution Which ML Model Works Best?

- 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

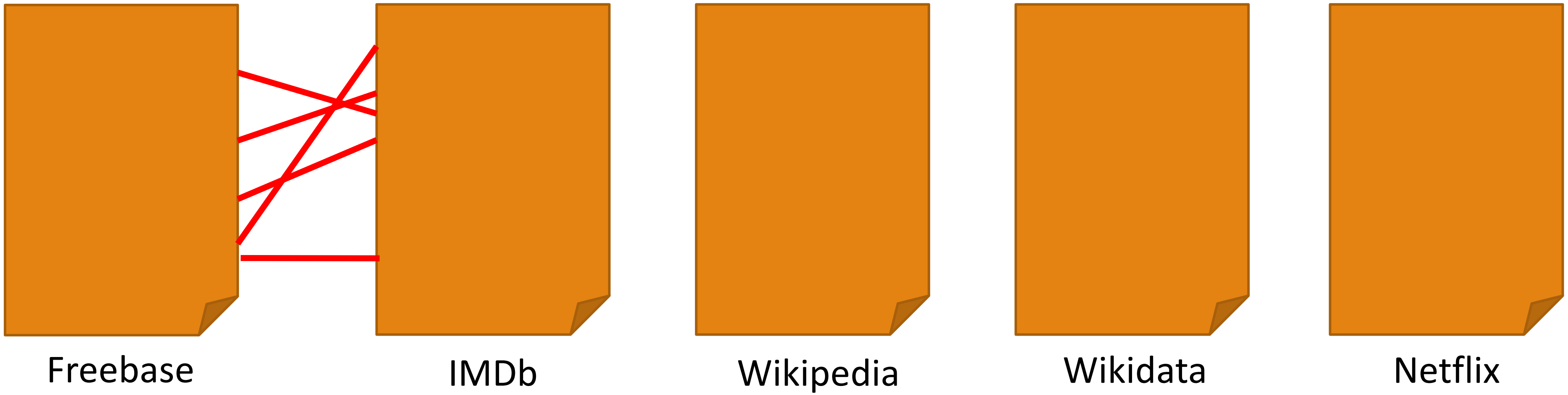
- Moonshot: Apply active learning to minimize #labels



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

I. Integrating Knowledge from Structured Sources—Entity Resolution

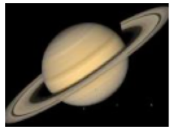
- Moonshots: Seamless incremental graph linkage with high precision and recall

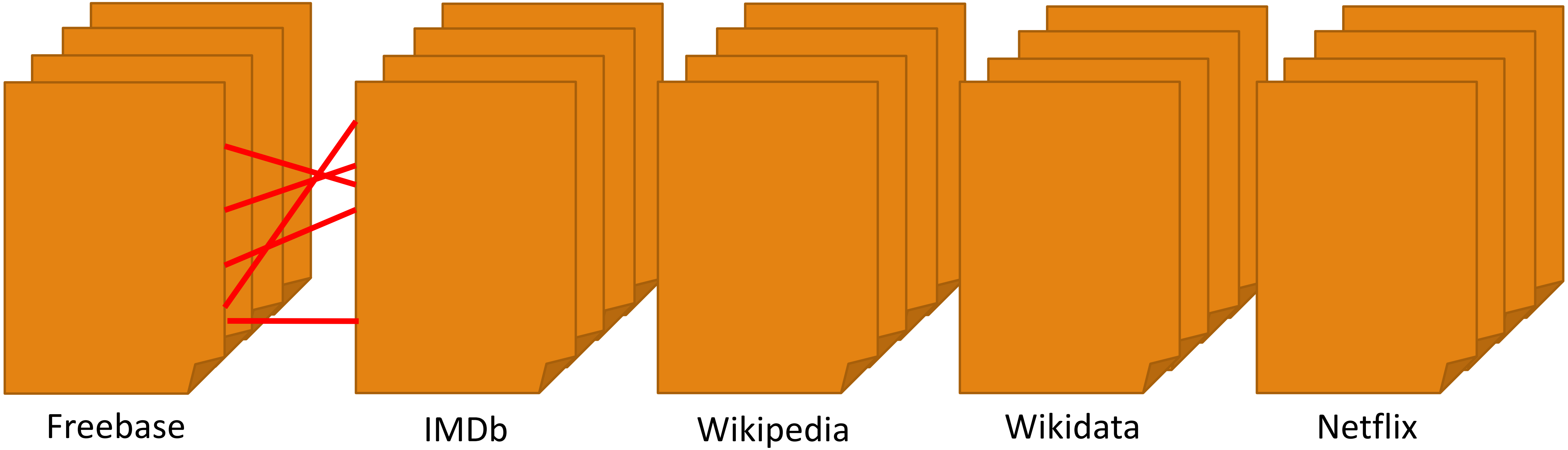


  - Different sources have different characteristics, but share commonalities from the same domain
  - *How to leverage models for existing sources on new sources?*

I. Integrating Knowledge from Structured Sources—Entity Resolution

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

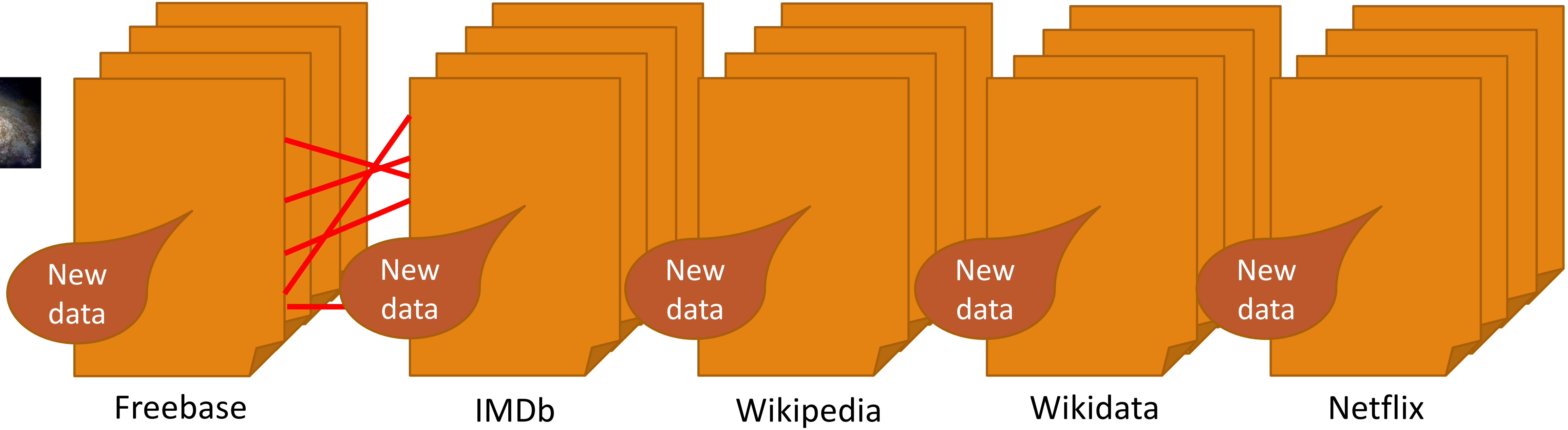




- A complex space can contain tens to thousands of different types and linkage on different types of entities can affect each other
- *How to avoid manual scheduling for linkage?*

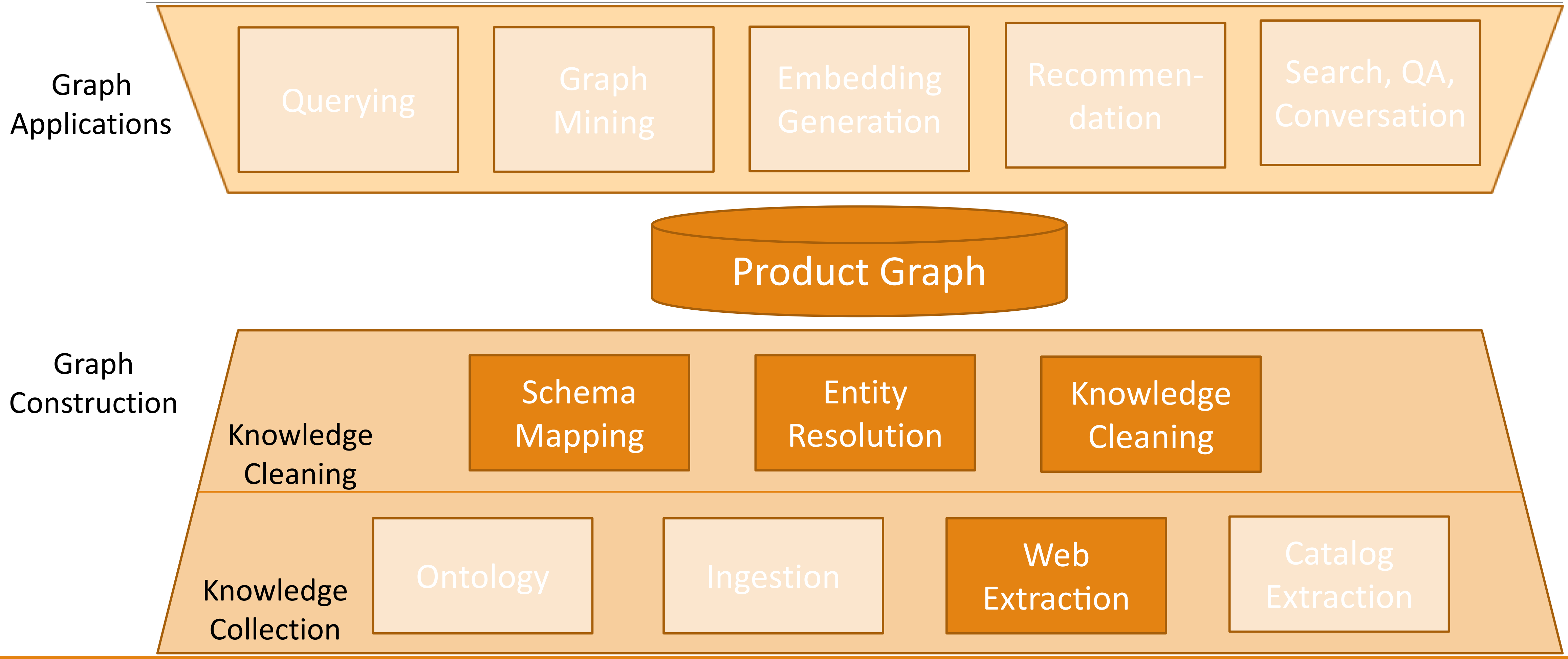
I. Integrating Knowle f
Structured Sources—

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

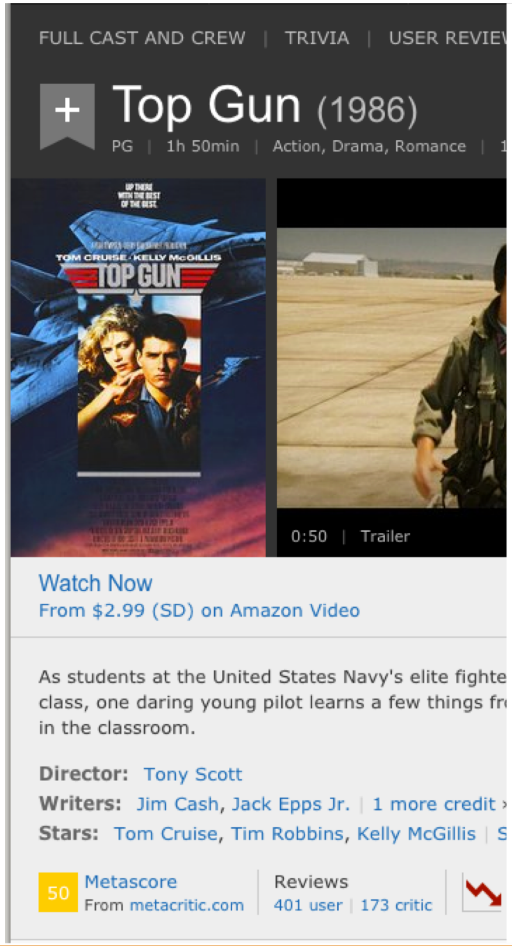


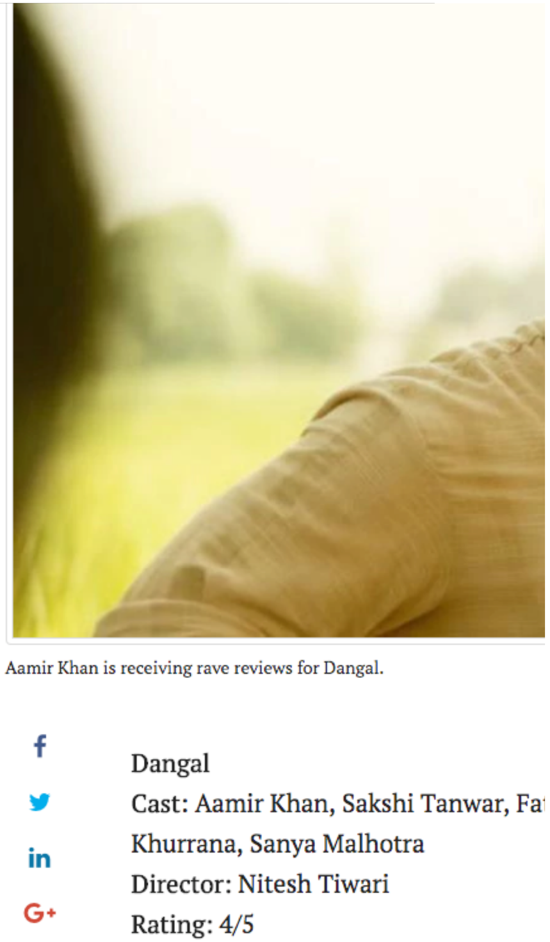
  - New data are arriving over time, requiring incremental linkage and model evolution
  - *How to perform incremental linkage and evolve the model?*

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



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

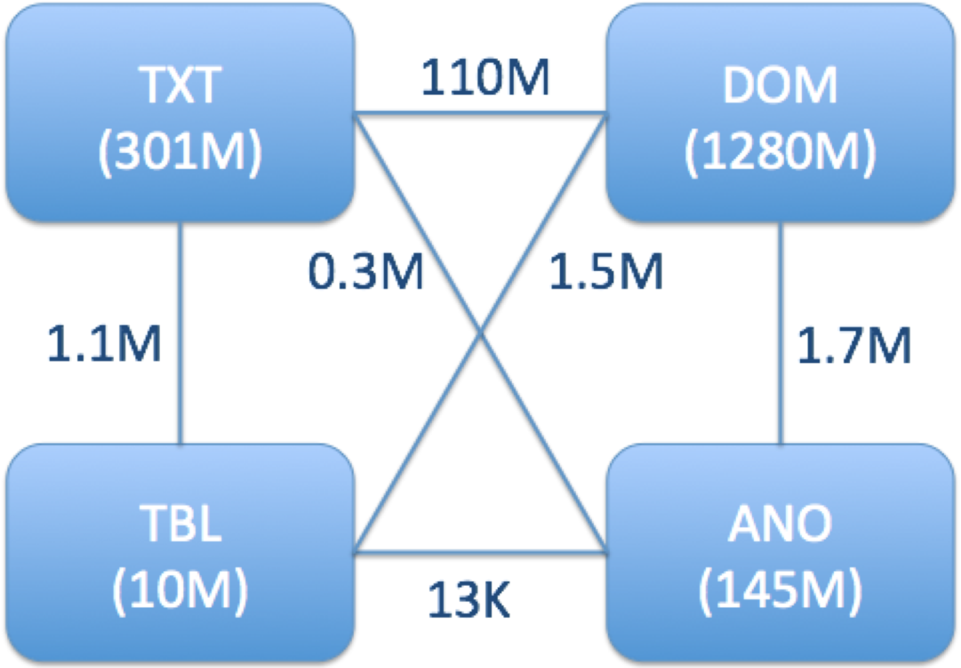




II. Extracting Knowledge from Semi-Structured Data on the Web

- 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]

Relation Annotation → Training (on noisy labels)

Corpus Text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from …
Google was founded by Larry Page …

Training Data

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

[Adapted example from Luke Zettlemoyer]

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

Relation Annotation → Training (on noisy labels)

Corpus Text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from …
Google was founded by Larry Page …

Training Data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

[Adapted example from Luke Zettlemoyer]

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

Relation Annotation → Training (on noisy labels)

Corpus Text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from …
Google was founded by Larry Page …

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Training Data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

[Adapted example from Luke Zettlemoyer]

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

Relation Annotation → Training (on noisy labels)

Corpus Text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from …
Google was founded by Larry Page …

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Training Data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

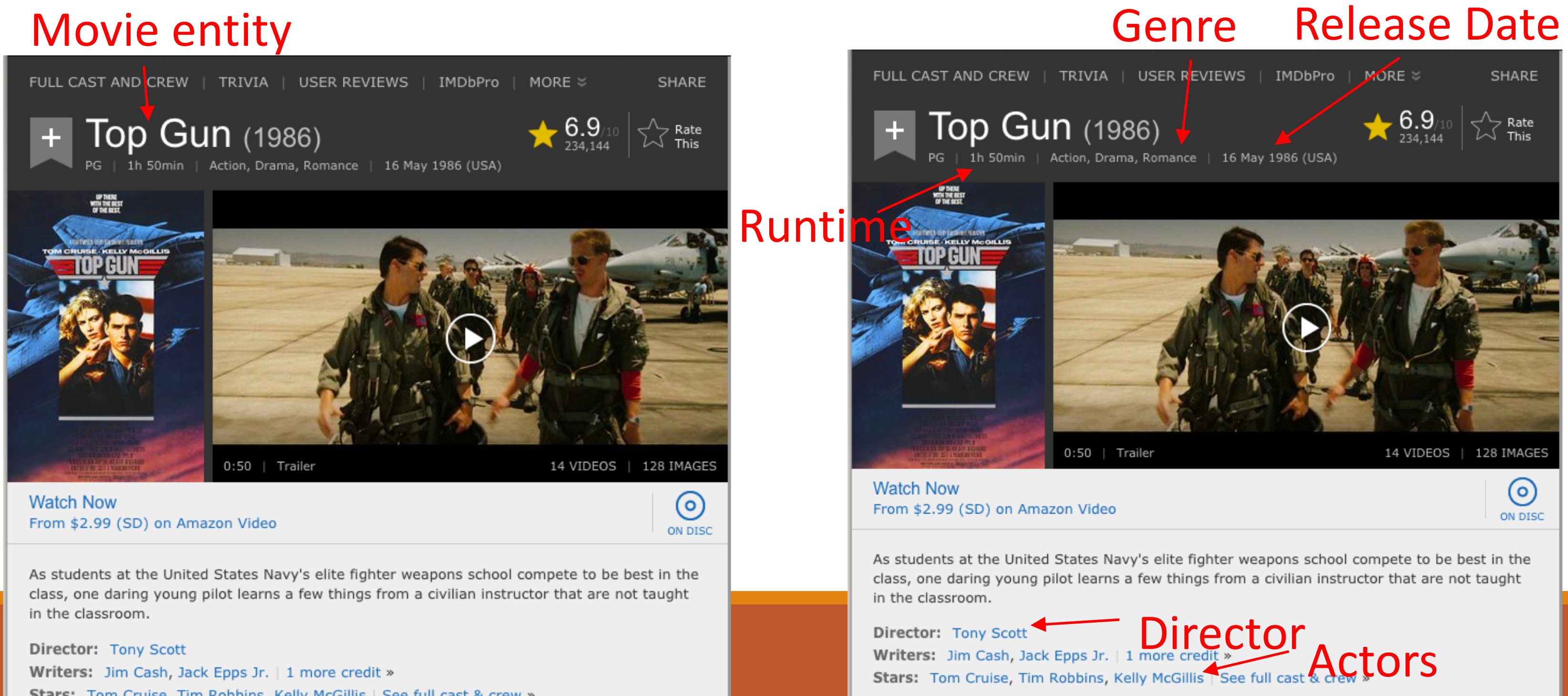
(Bill Gates, Harvard)
Label: CollegeAttended
Feature: X attended Y

[Adapted example from Luke Zettlemoyer]

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

Entity Identification → Relation Annotation → Training (on noisy labels)

Automatic Label Generation



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

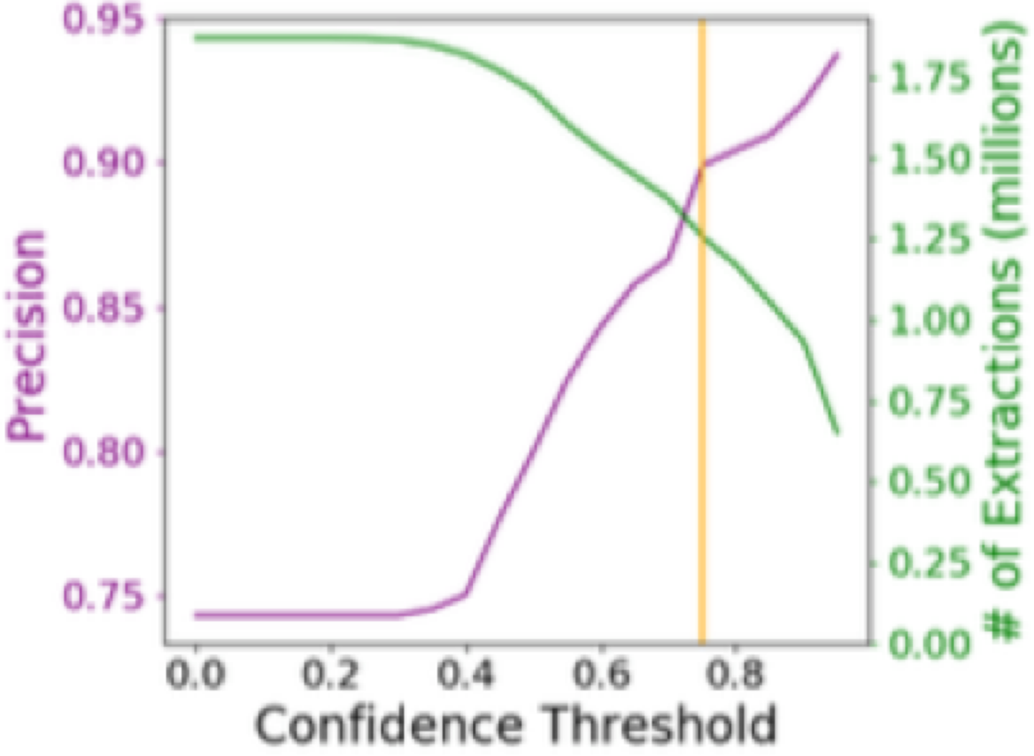
II. Extracting Knowledge from Web—Distantly Supervised DOM Extraction [VLDB'18]

- 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

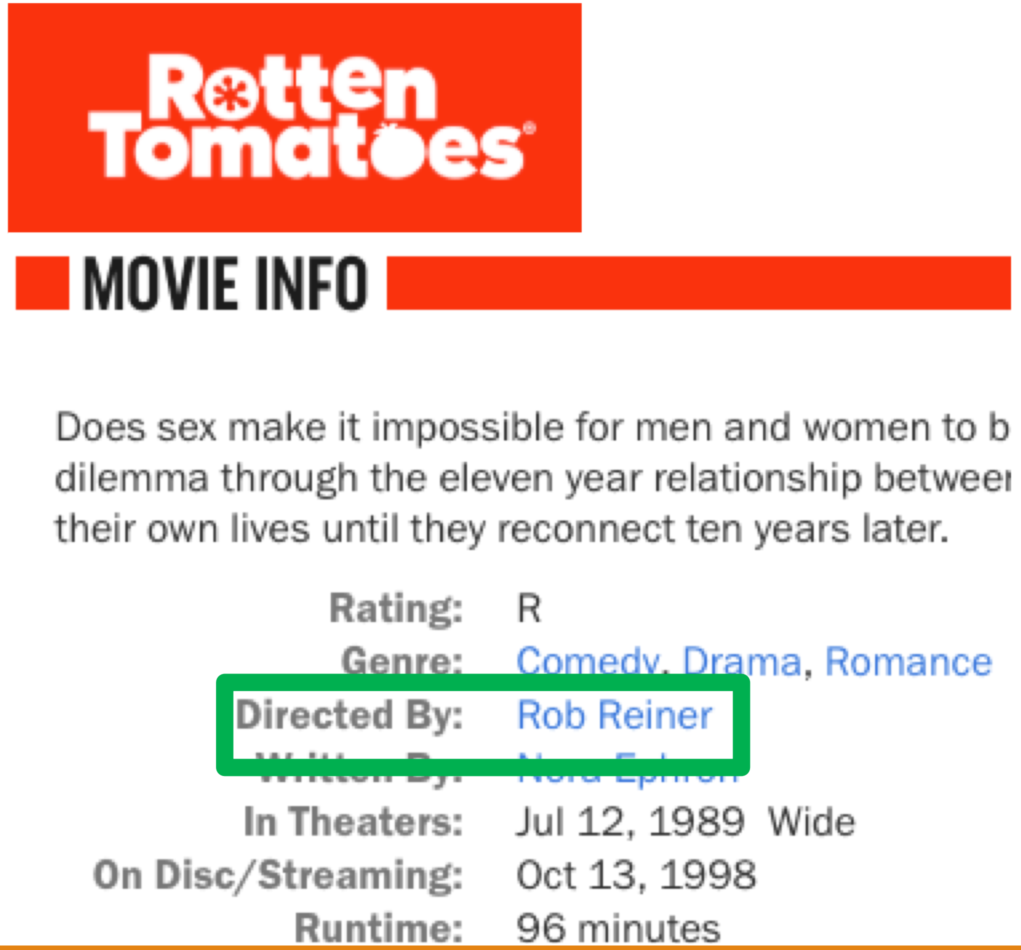


II. Extracting Knowledge from the Web—OpenIE DOM Extraction

- 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")



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”)

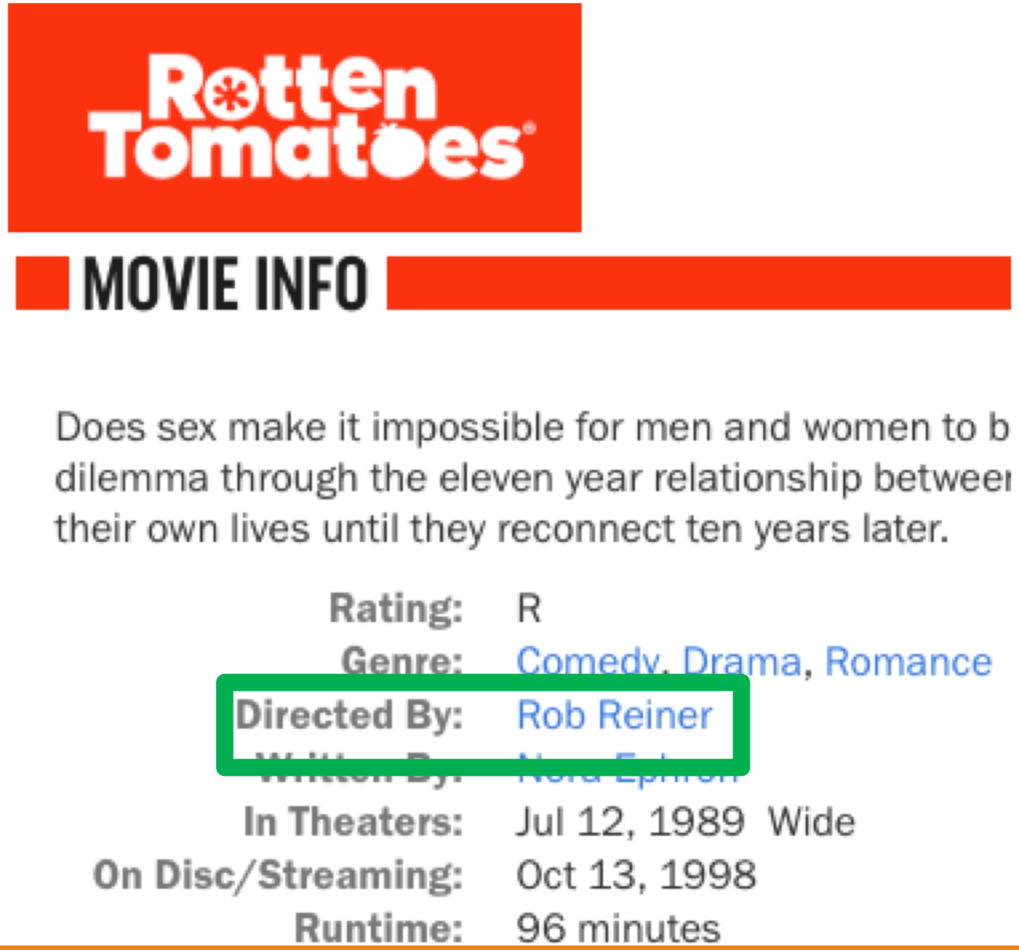
(“When Harry Met Sally”, “Directed By:”, “Rob Reiner”)

(“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”)



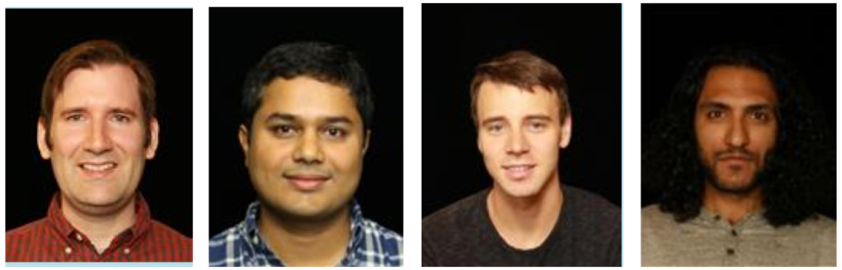
II. Extracting Knowledge from the Web—OpenIE DOM Extraction Preliminary Rslts

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

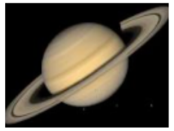
II. Distantly Supervised DOM Extraction Which ML Model Works Best?

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

II. Extracting Knowledge from Web—Distantly Supervised DOM Extraction



Nearly-automatic interactive extraction on any new vertical

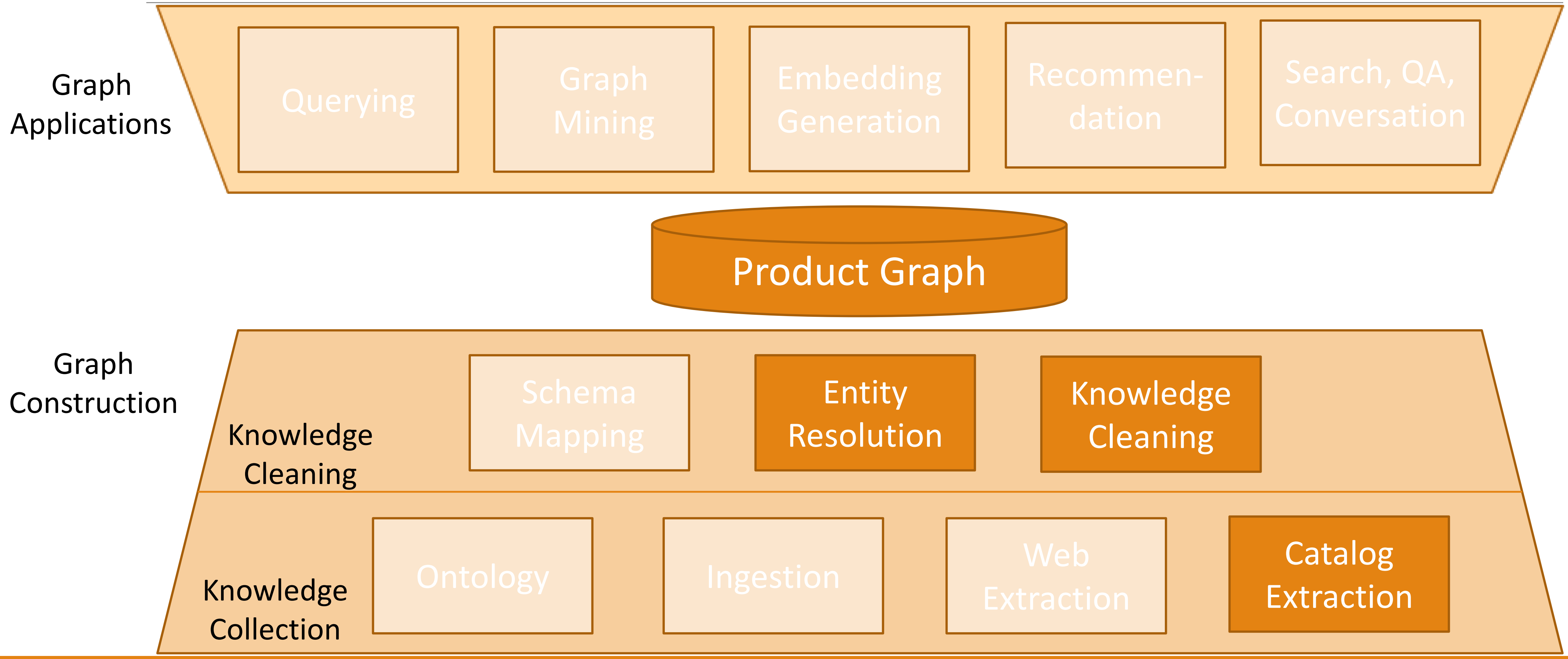


OpenIE
DOM extraction

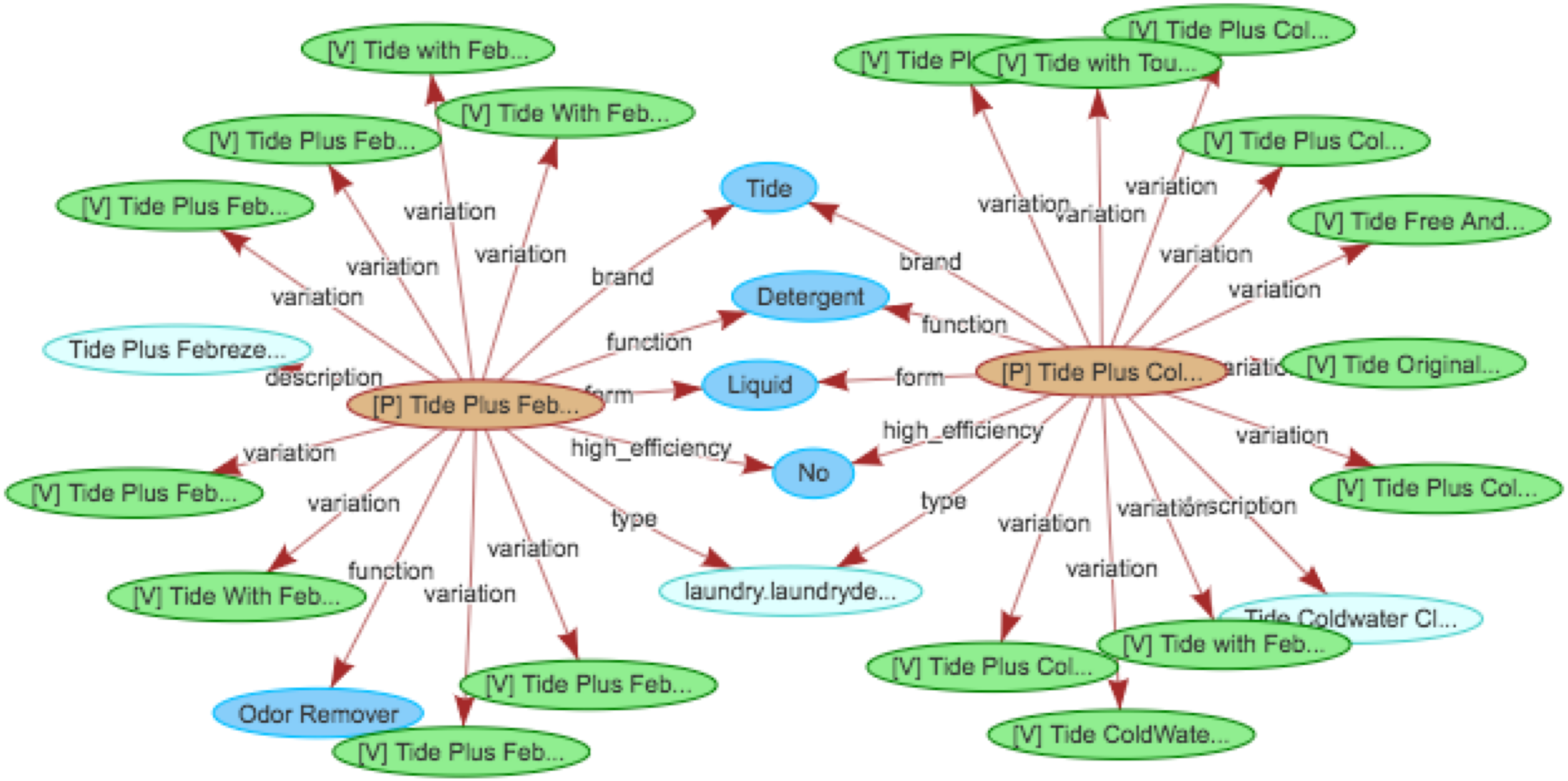
Distantly supervised web extraction

Annotation-based knowledge extraction

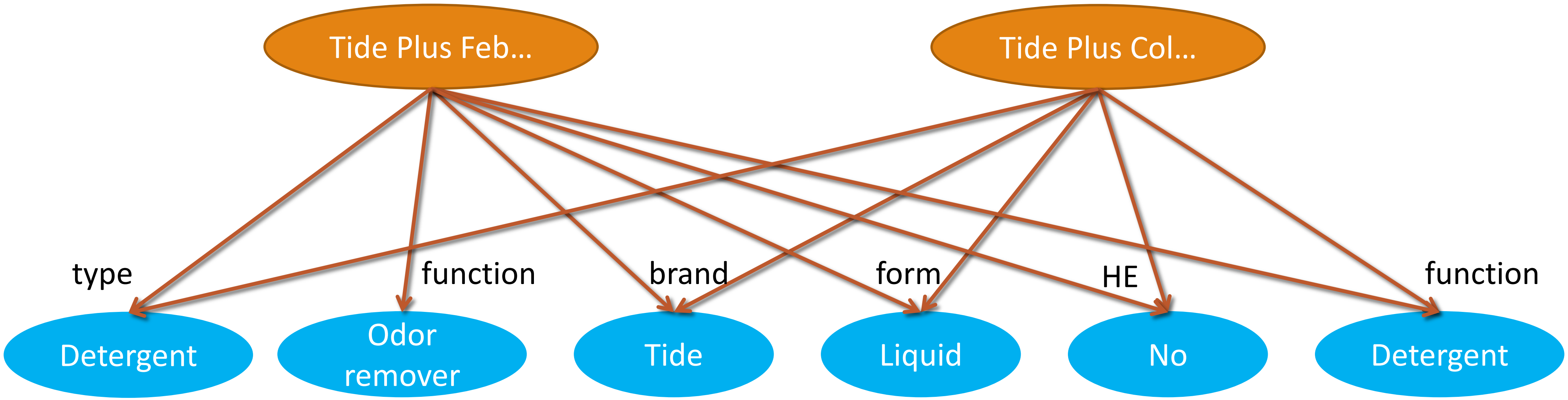
III. Building a Broad & Shallow Graph from Product Profiles in Amazon Catalog



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

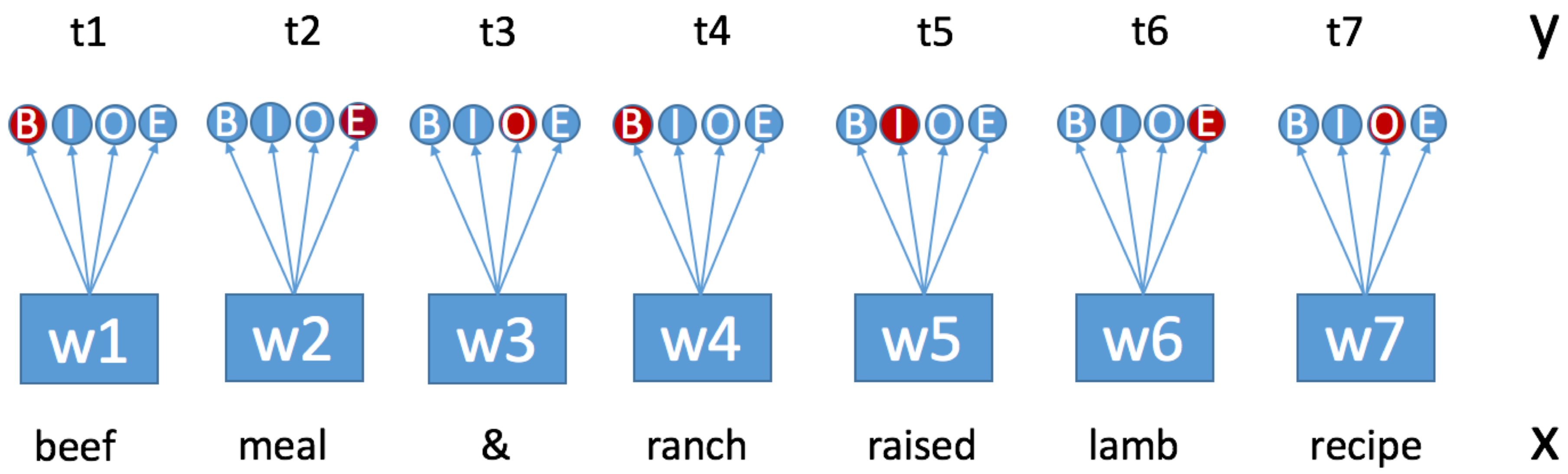
I Inside of entity

O Outside of entity

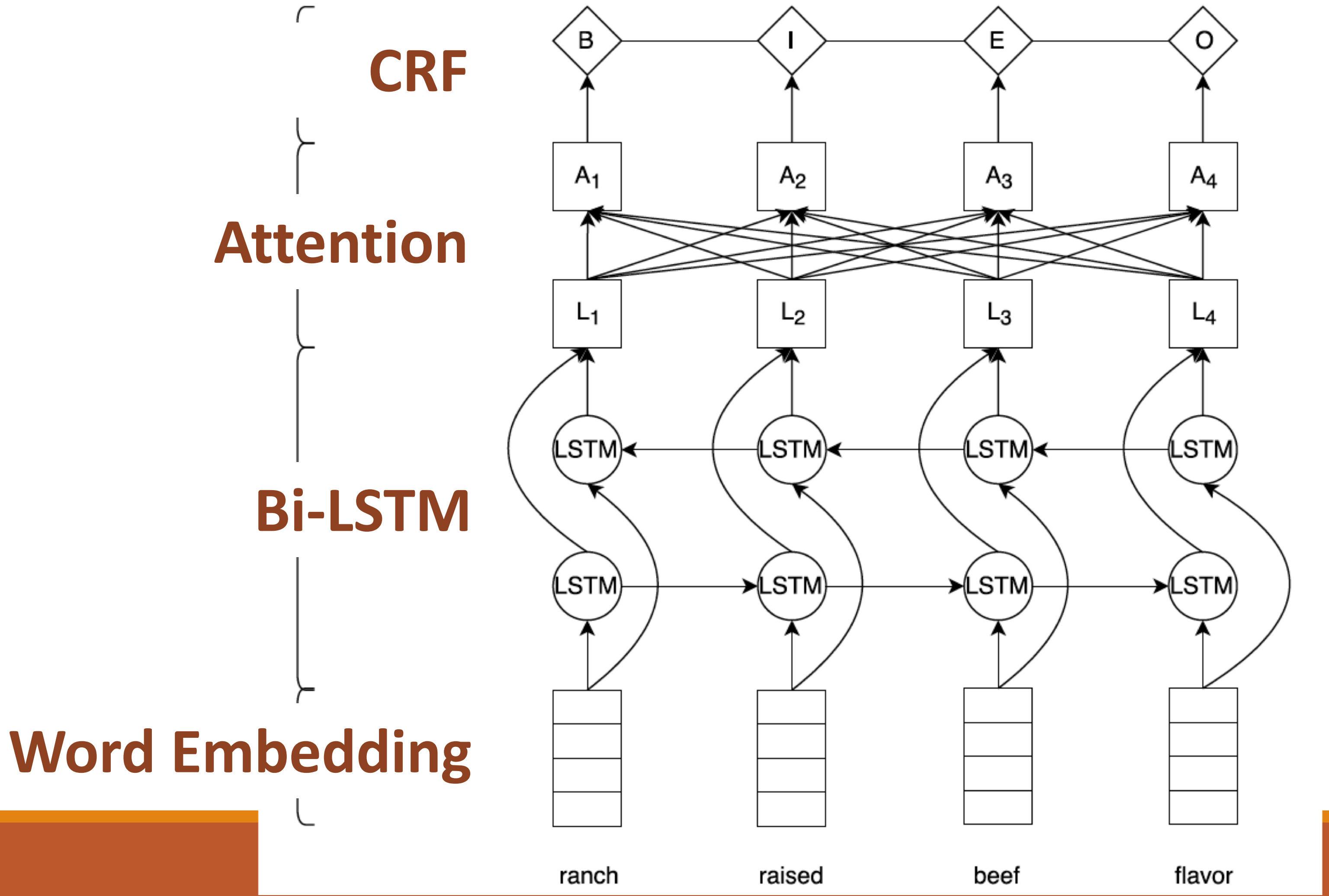
E End of entity

x={w1,w2,…,wn} input sequence

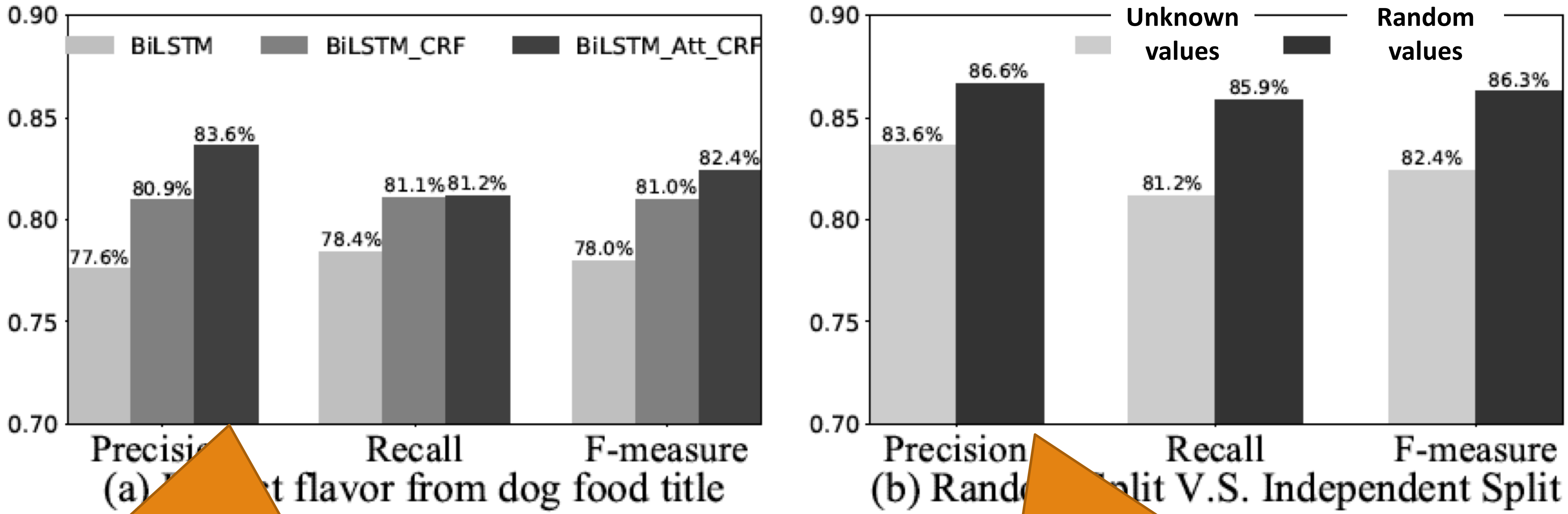
y={t1,t2,…,tn} tagging decision



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



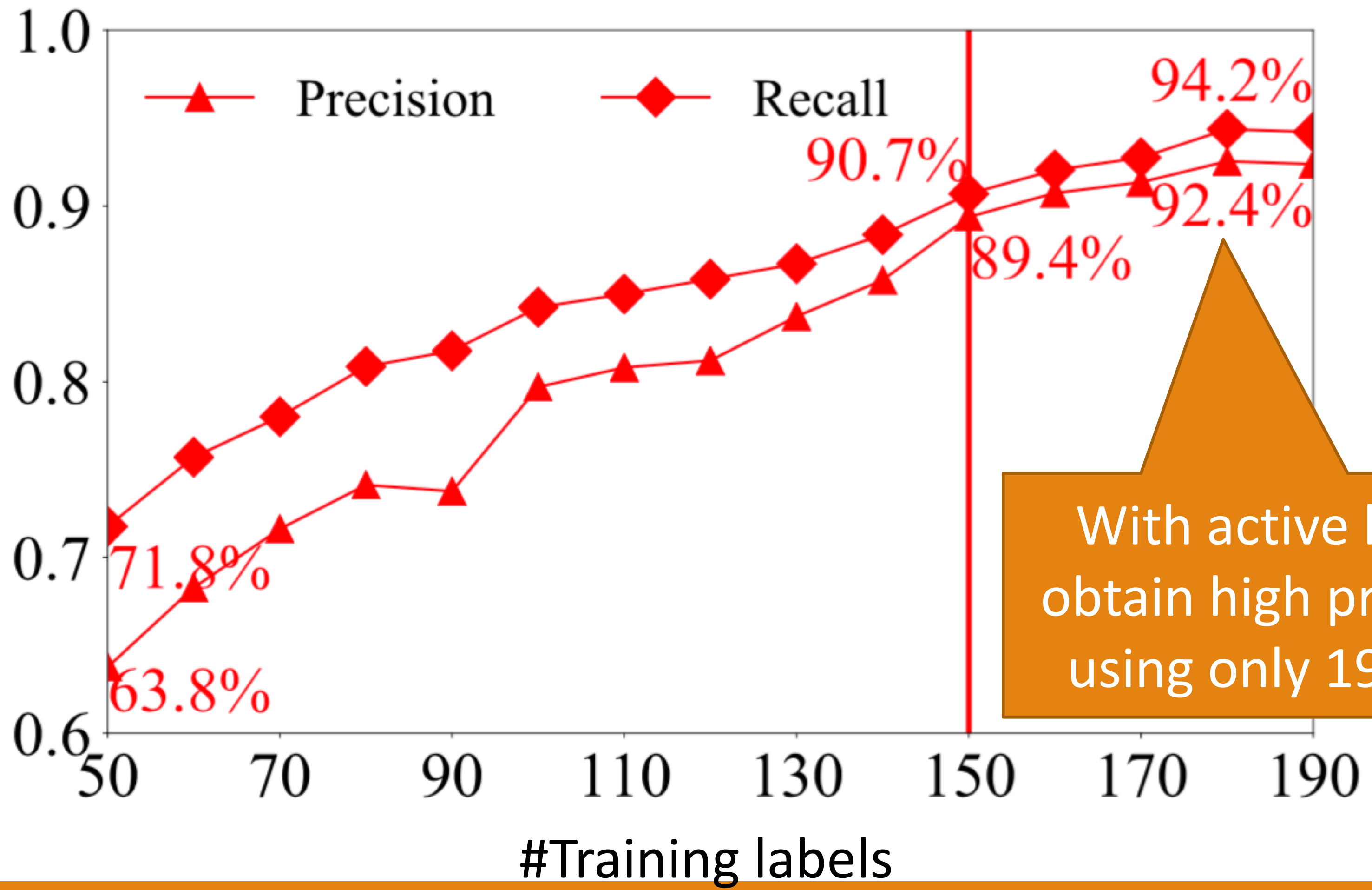
III. Open Attribute Extraction by Named Entity Recognition [KDD'18]



BiLSTM+CRF+Attention obtains best results

Extraction on new values is comparable to already known values

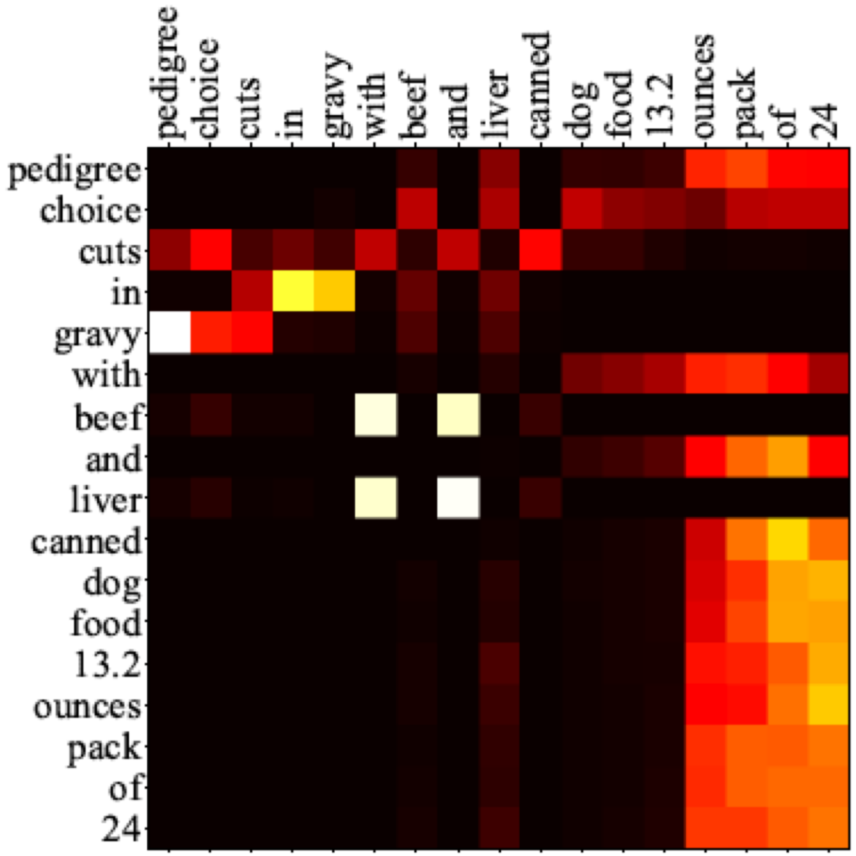
III. Open Attribute Extraction by Named Entity Recognition [KDD'18]



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

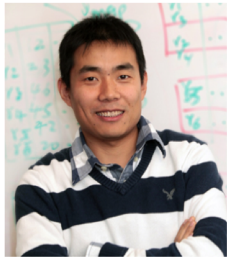
III. Product Profile Extraction by NER —Which ML Model Works Best?

- Recurrent Neural Network, CRF, Attention

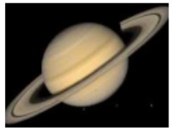


III. Building a Shallow Graph from Product Profiles in Amazon Catalog





Review extraction
& sentiment analysis

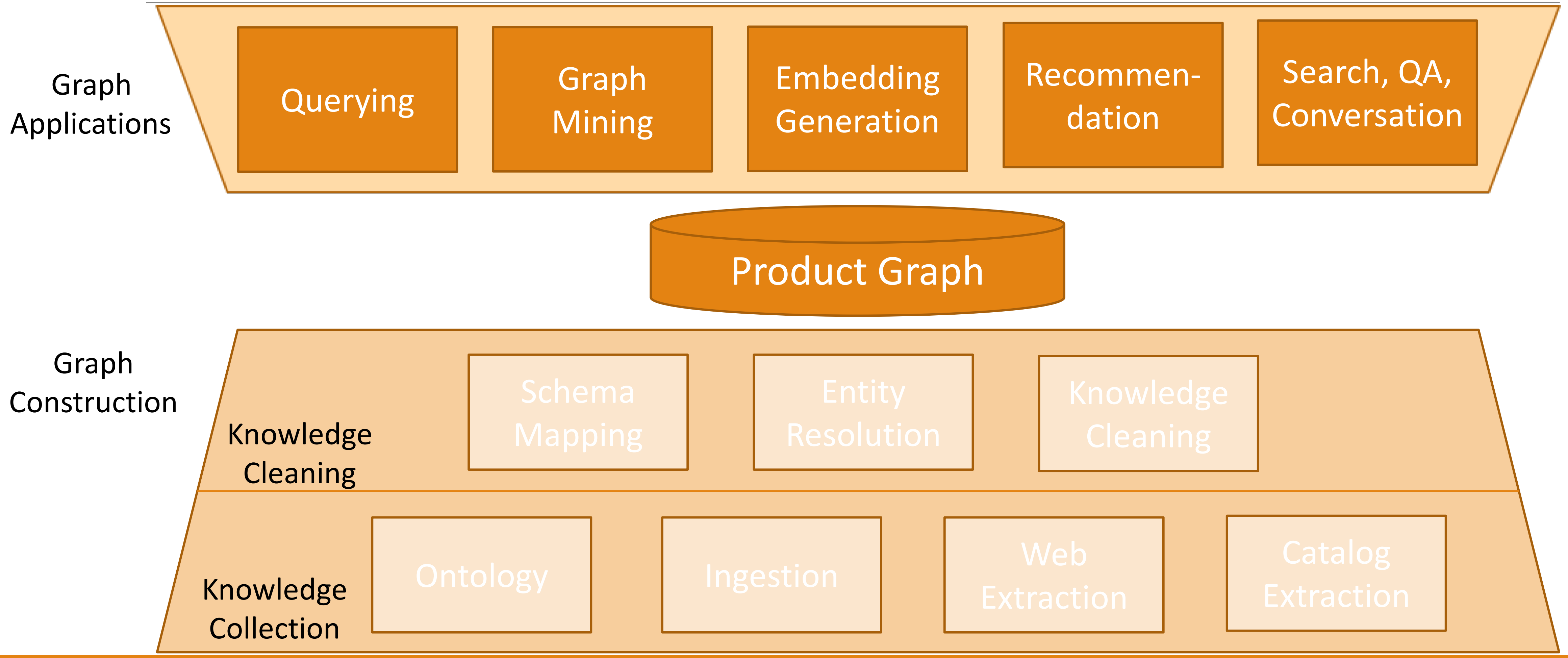


Open aspect
extraction

Automatically
building a
shallow KG

Product profile
extraction

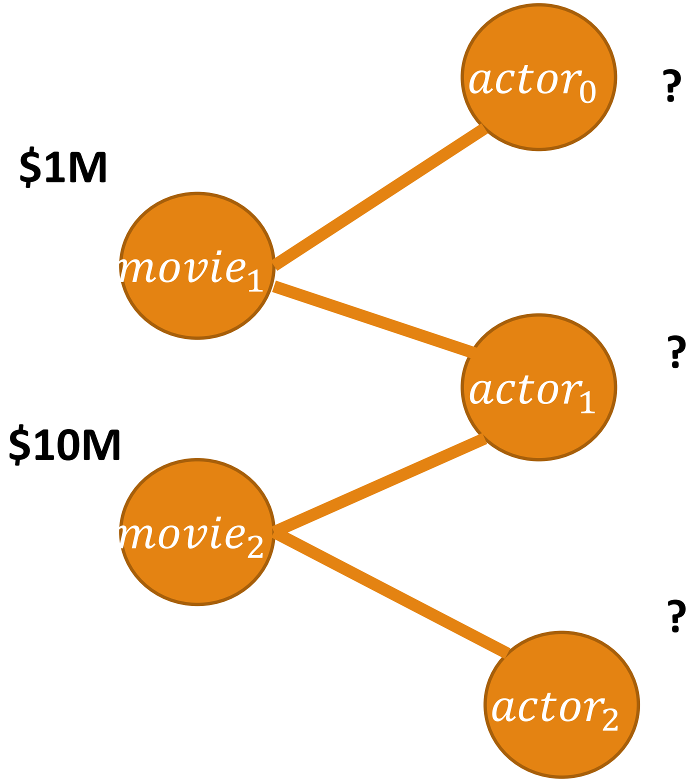
IV. Graph Mining and Embedding



IV. Graph Mining

IV. Finding Entity Importance

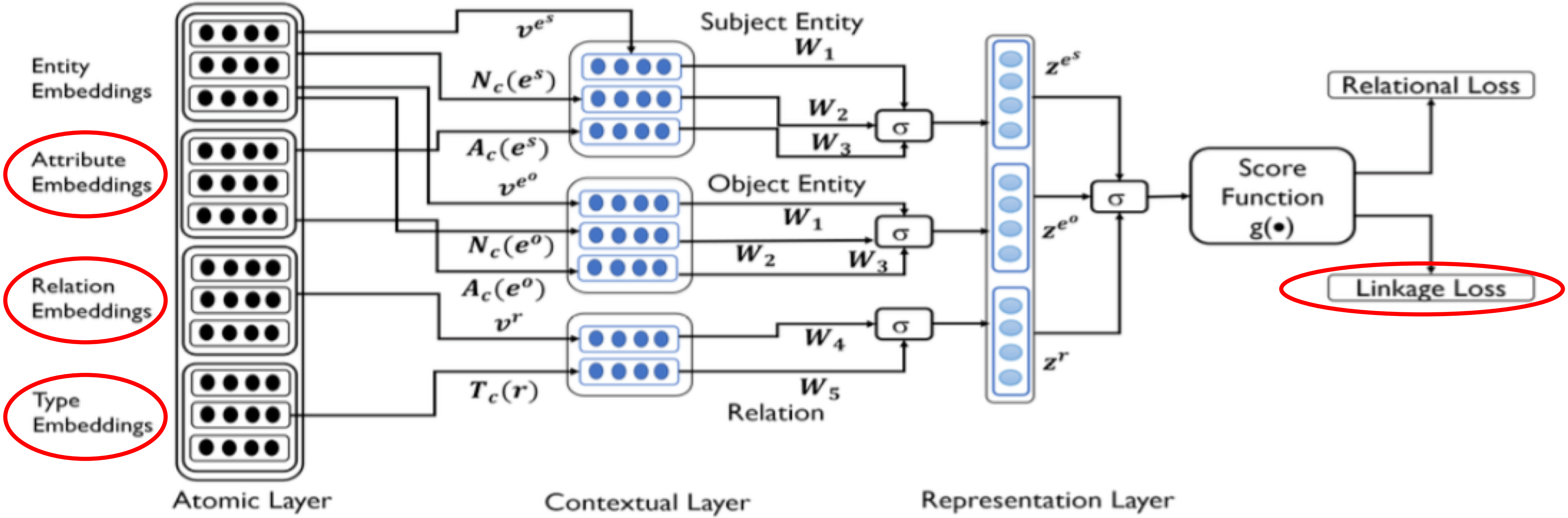
- Input
  - Knowledge Graph
  - Importance scores for some nodes (e.g., PageView for Wikipedia)
- Output:
  Importance scores for all nodes
- Method: PageRank w. Restart



IV. Finding Entity Importance

Actors	Directors	Countries	Companies
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]



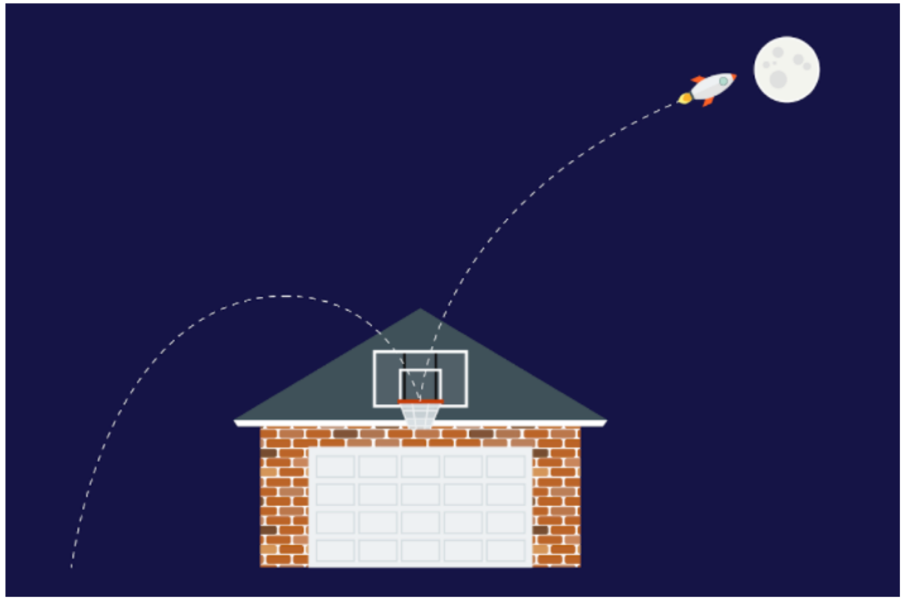
IV. Graph Embedding for




	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!