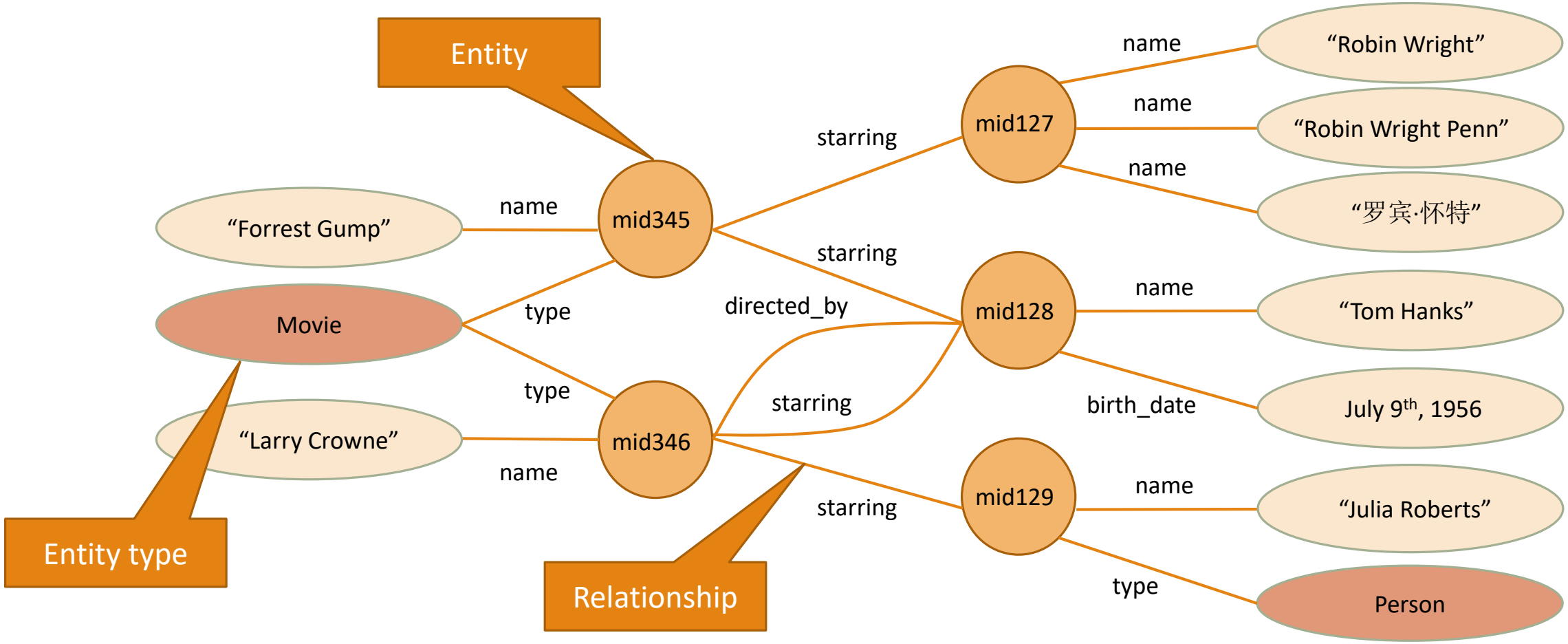
Challenges and Innovations in Building a Product Knowledge Graph

XIN LUNA DONG, AMAZON

FEBRUARY, 2017

Product Graph vs. Knowledge Graph

Knowledge Graph Example for 2 Movies



Knowledge Graph in Search

Tom Hanks › Movies

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

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/J7OXJbf…	Beware! Crass self-serving Social Media Post! This book goes on sale tomorrow! Hanx pic.twitter.com/V2EqPKL…	Lost (g)love. Looking for a mate. Good luck. Hanx. pic.twitter.com/ApH7rEG…
12 hours ago · Twitter	16 hours ago · Twitter	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

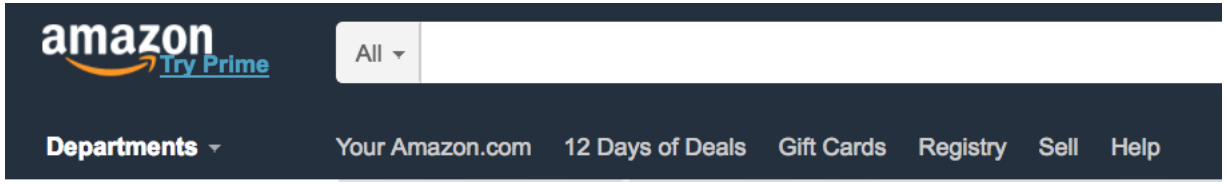
List of officially released compilations and

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

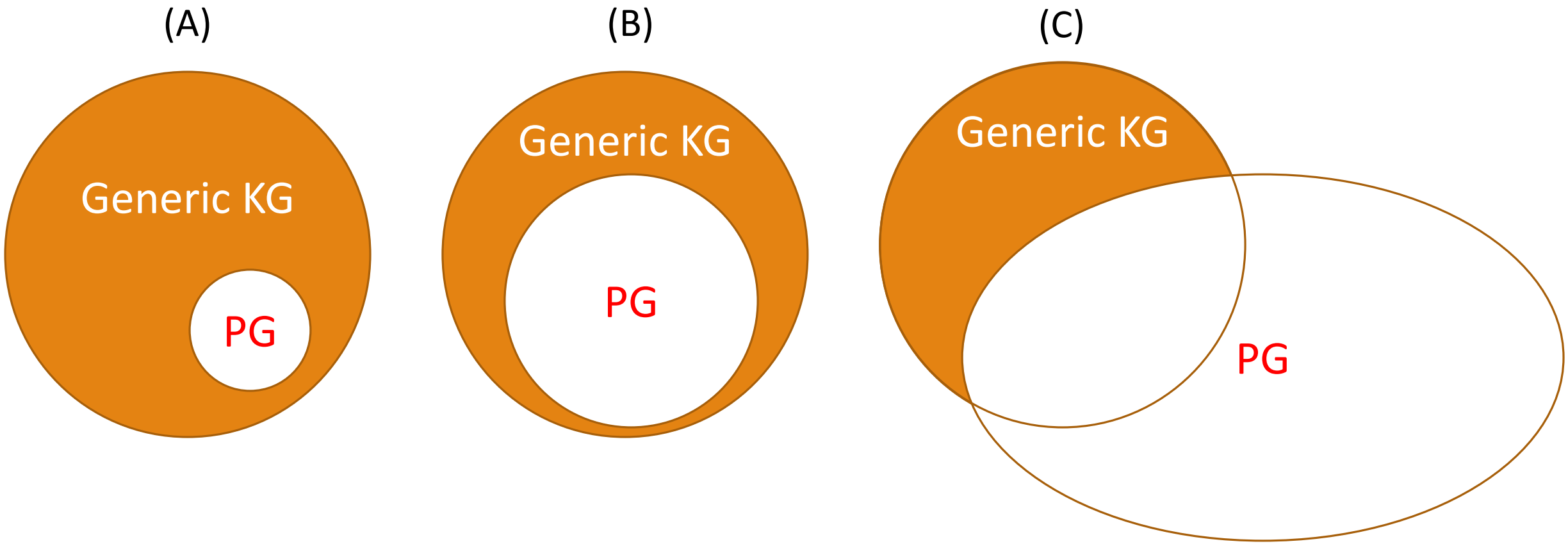
- *Portrait of Michael Jackson / Portrait of Jackson 5* (1973)
- *Os Grandes Sucessos, Vol. 2* (1980)
- *Motown Superstar Series, Vol. 7* (1980)
- *Superstar* (1980)
- *Michael Jackson & The Jackson 5* (1983)
- *Ain't No Sunshine* (1984)
- *The Great Love Songs of Michael Jackson* (1984)
- *Ben / Got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

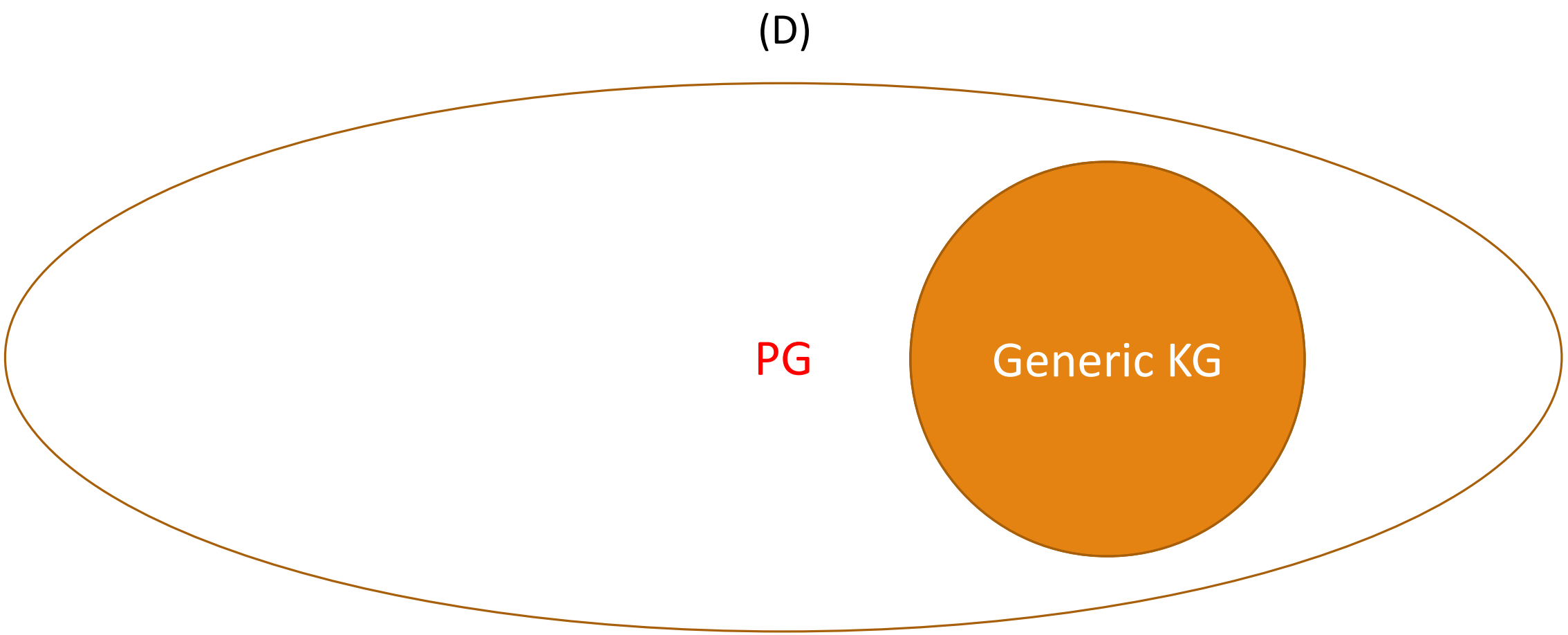
- Mission: To answer any question about products and related knowledge in the world



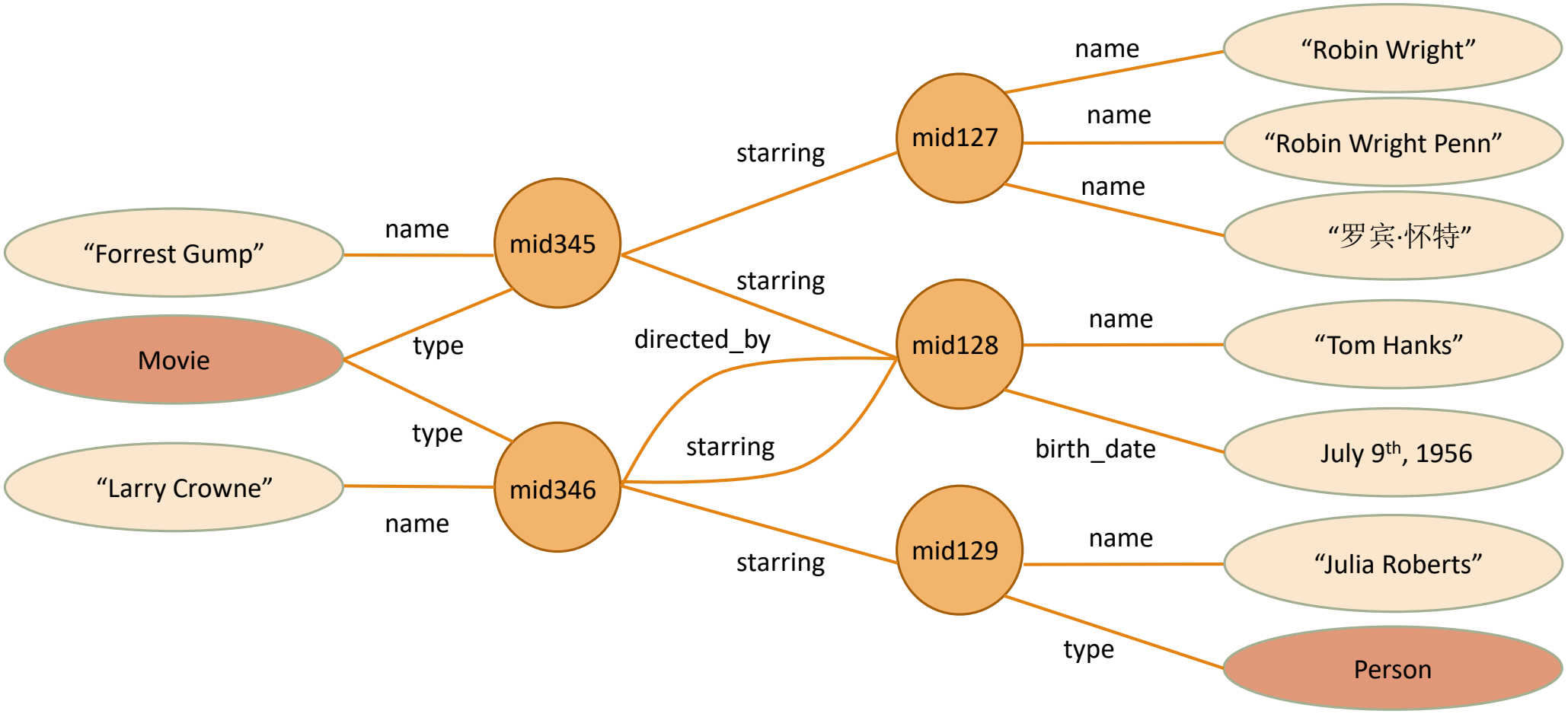
Product Graph vs. Knowledge Graph



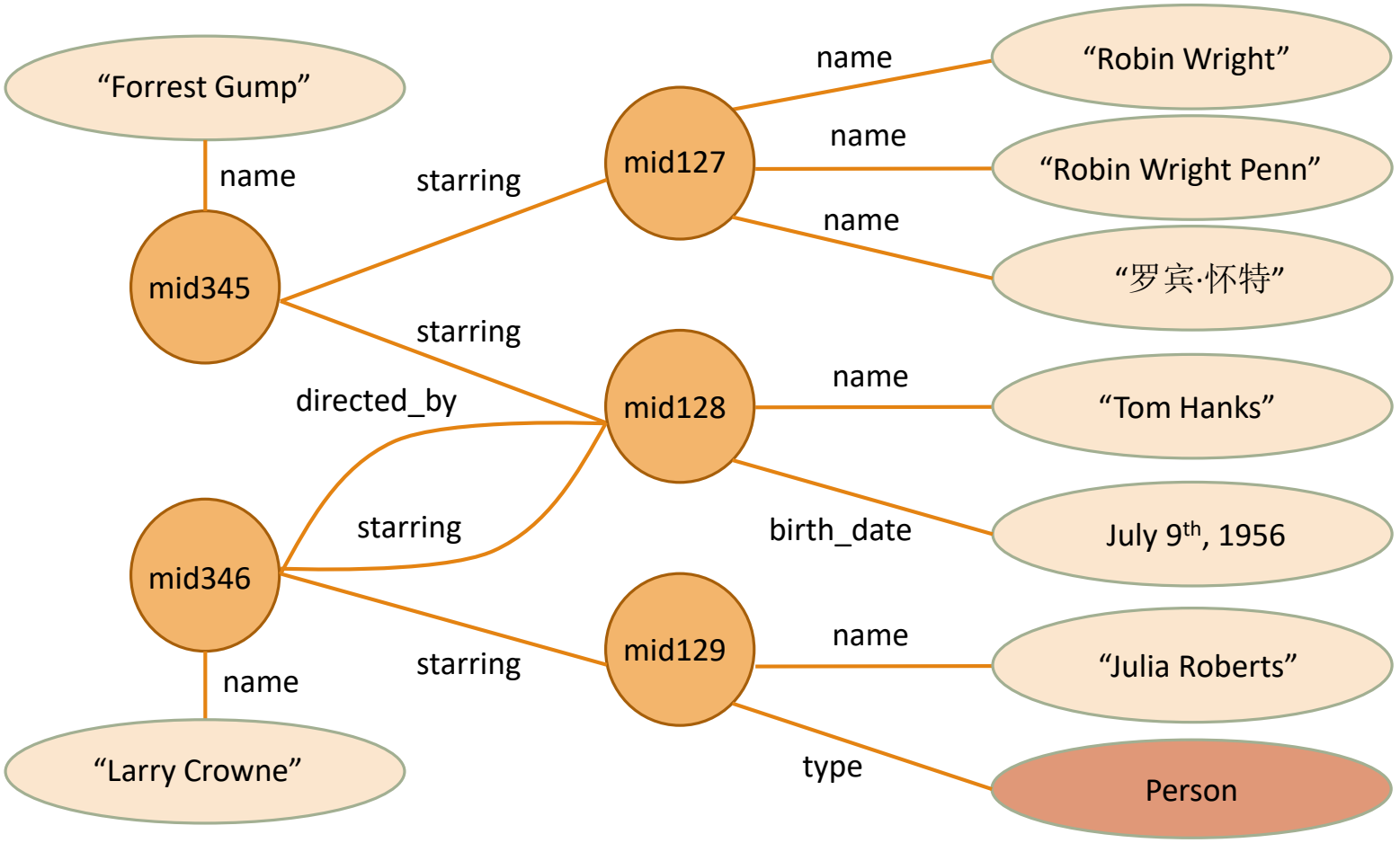
Product Graph vs. Knowledge Graph



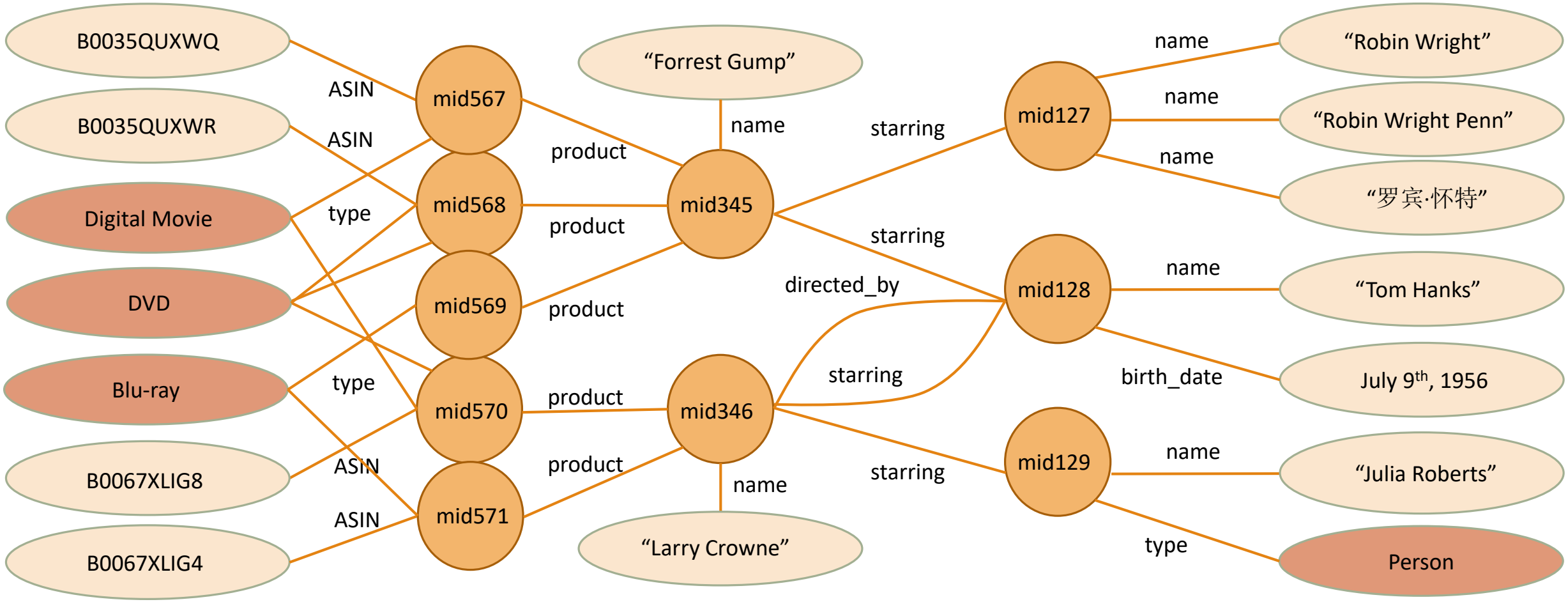
Knowledge Graph Example for 2 Movies



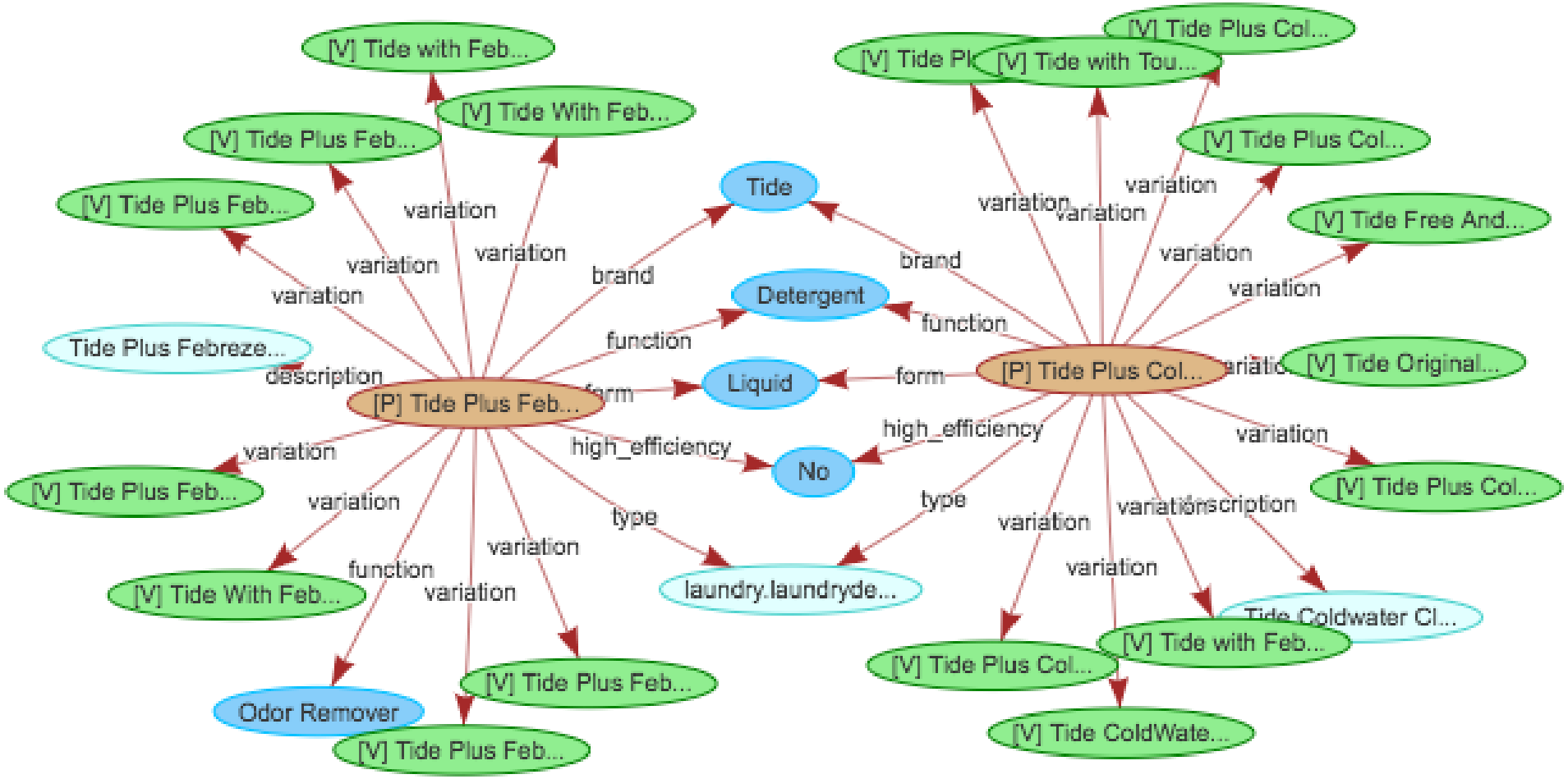
Product Graph vs. Knowledge Graph



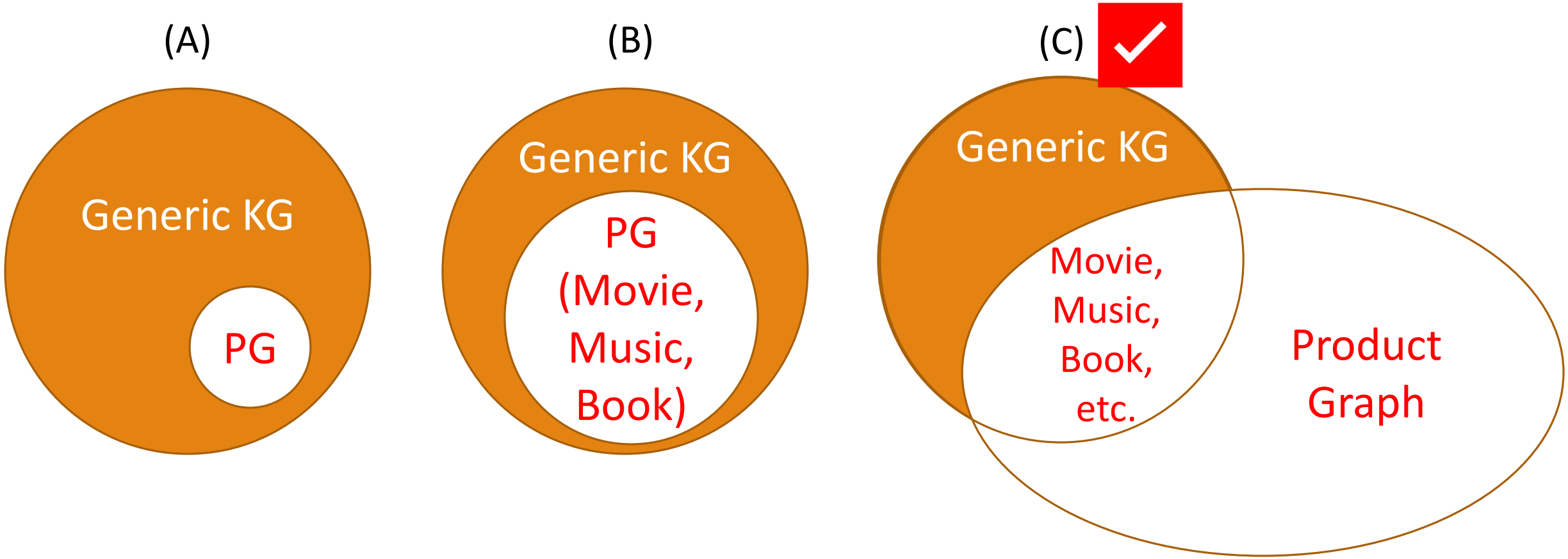
Product Graph vs. Knowledge Graph

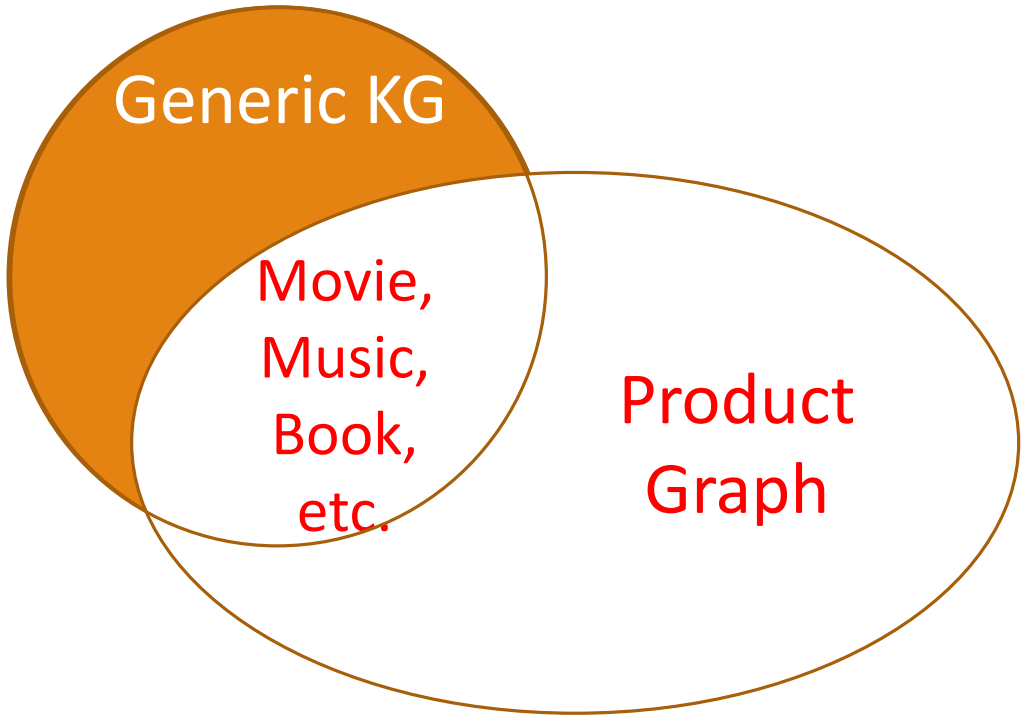


Another Example of Product Graph



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

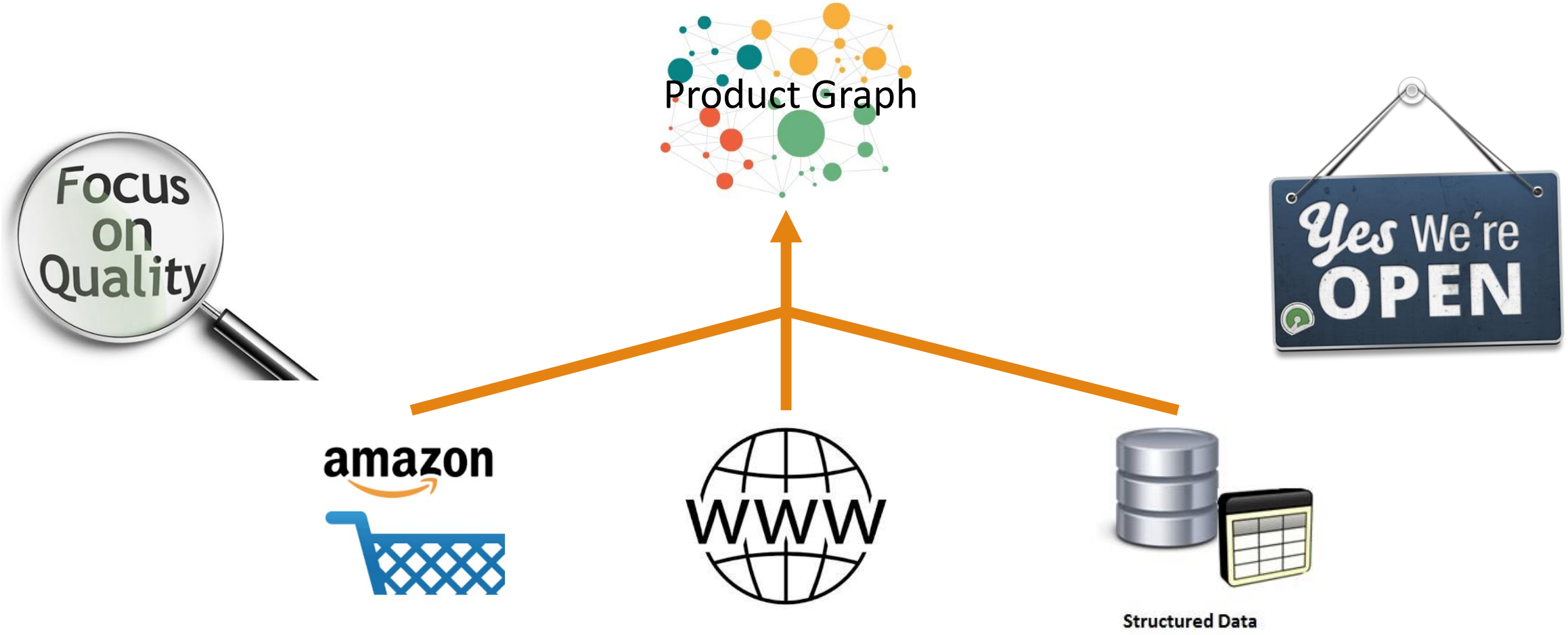
- Large number of product categories
  - A lot of work to manually define ontology
  - Hard to catch the trend of new product categories and properties

Challenges in Building Product Graph IV

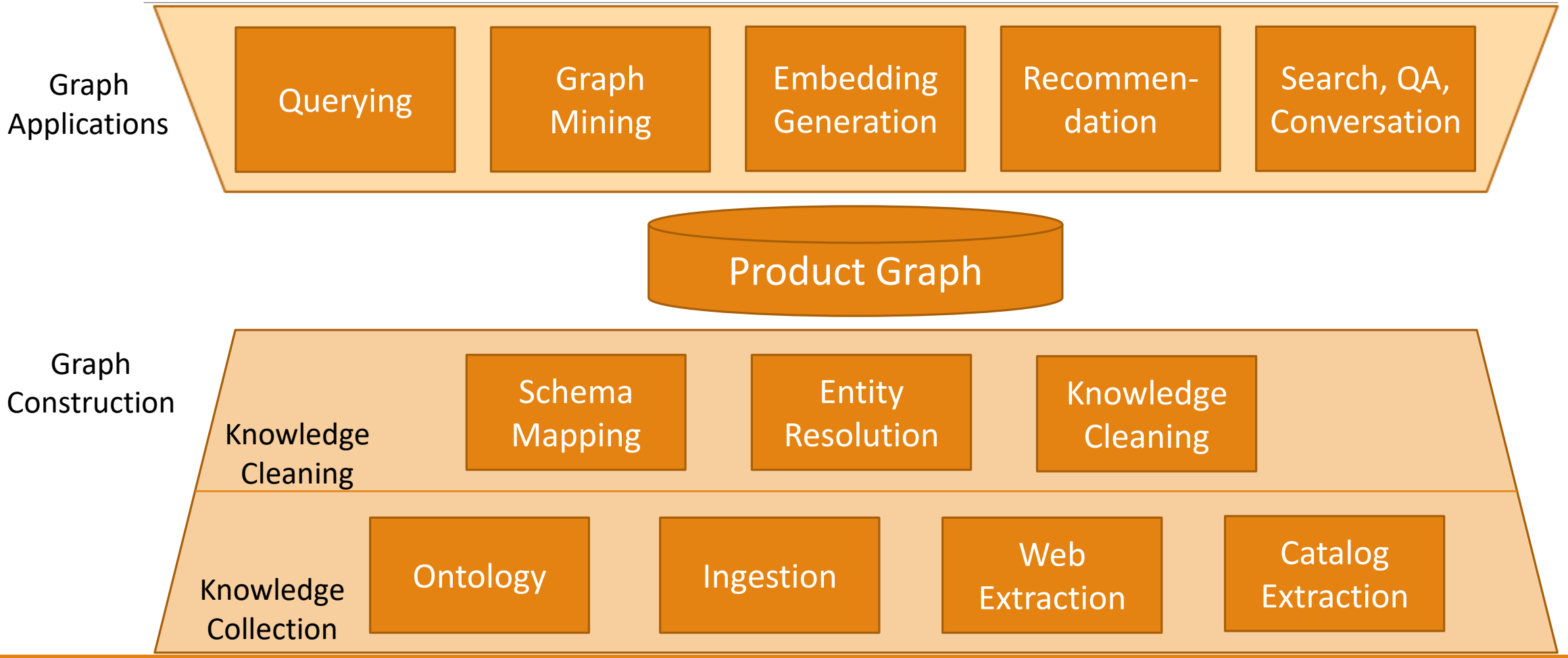
- Many entities are not named entities
  - Named Entity Recognition does not apply
  - New challenges for extraction, linking, and search

How to Build a Product Graph?

Where is Knowledge from?



Architecture

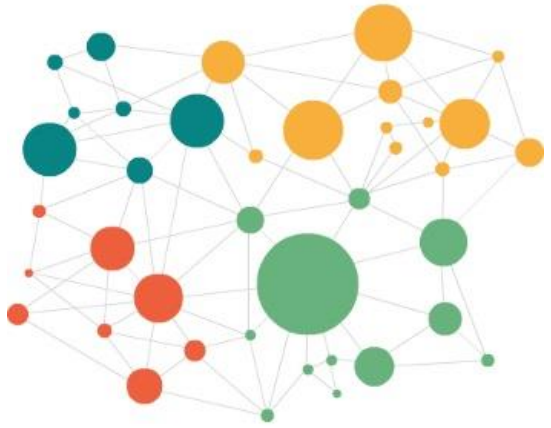


Which ML Model Works Best?

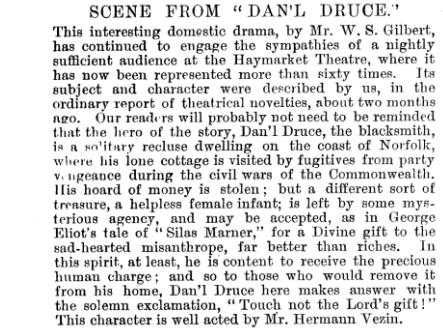
Which ML Model Works Best?

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.

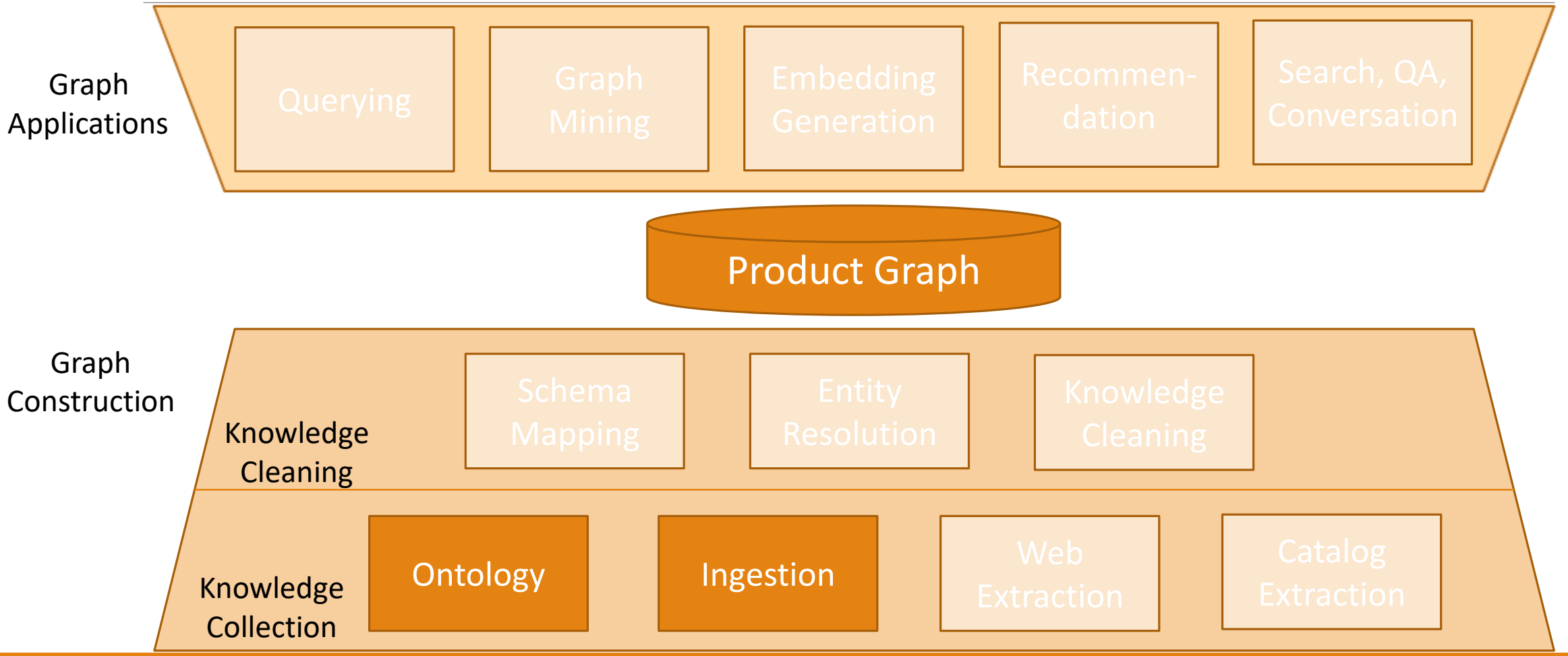
Neural network

Research Philosophy

Roofshots: Deliver incrementally and make production impacts

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

I. Integrating Knowledge from Structured Sources

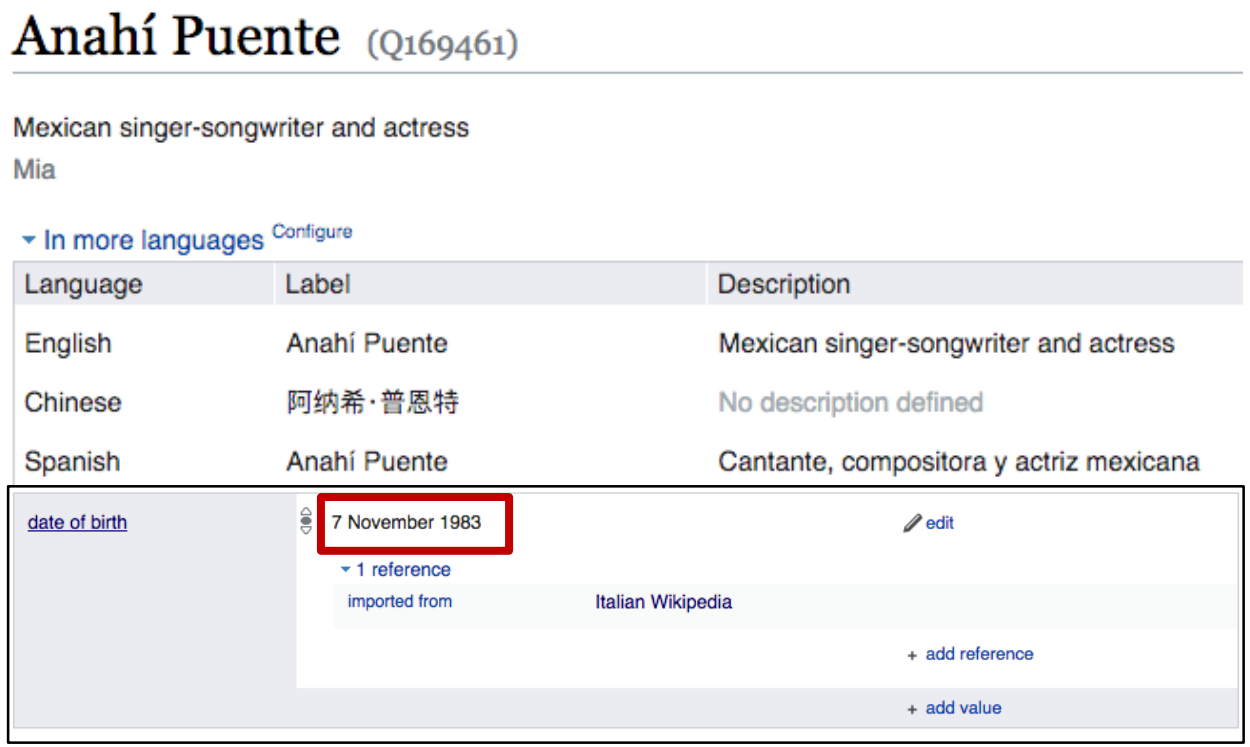


Challenges: Linkage & Quality

IMDB

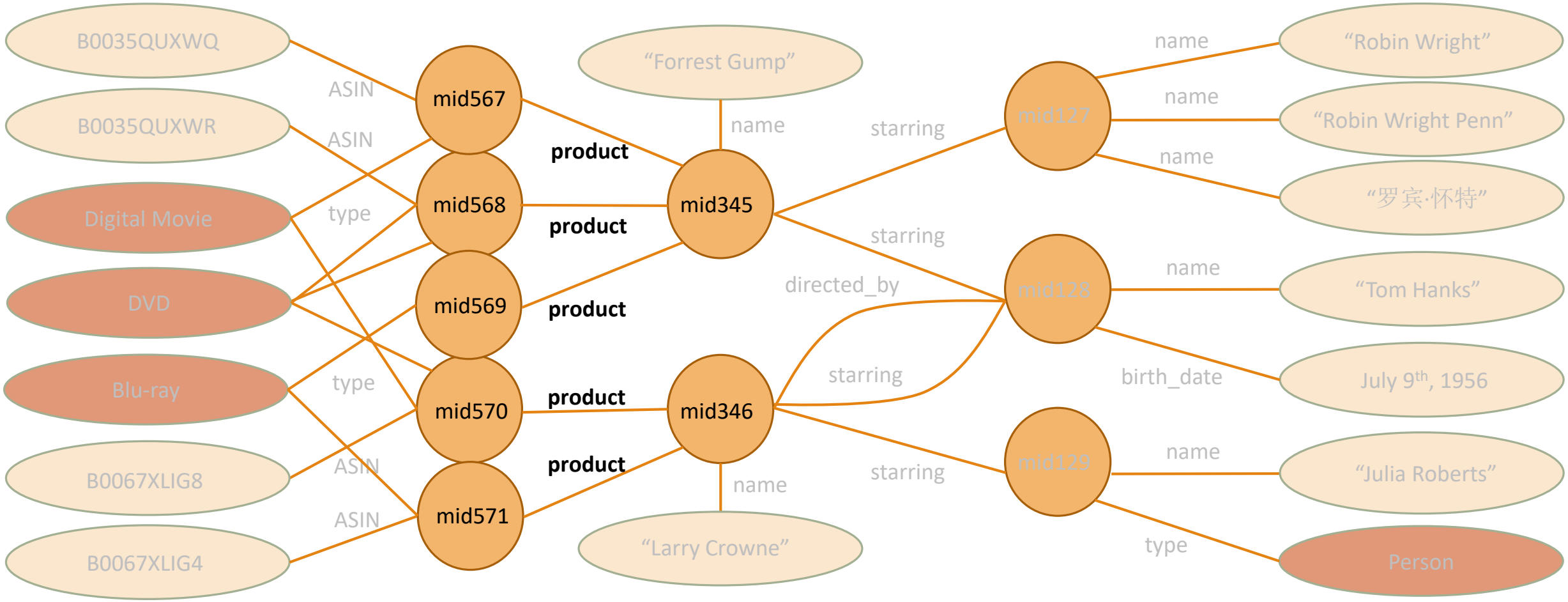
Same entity?

WikiData

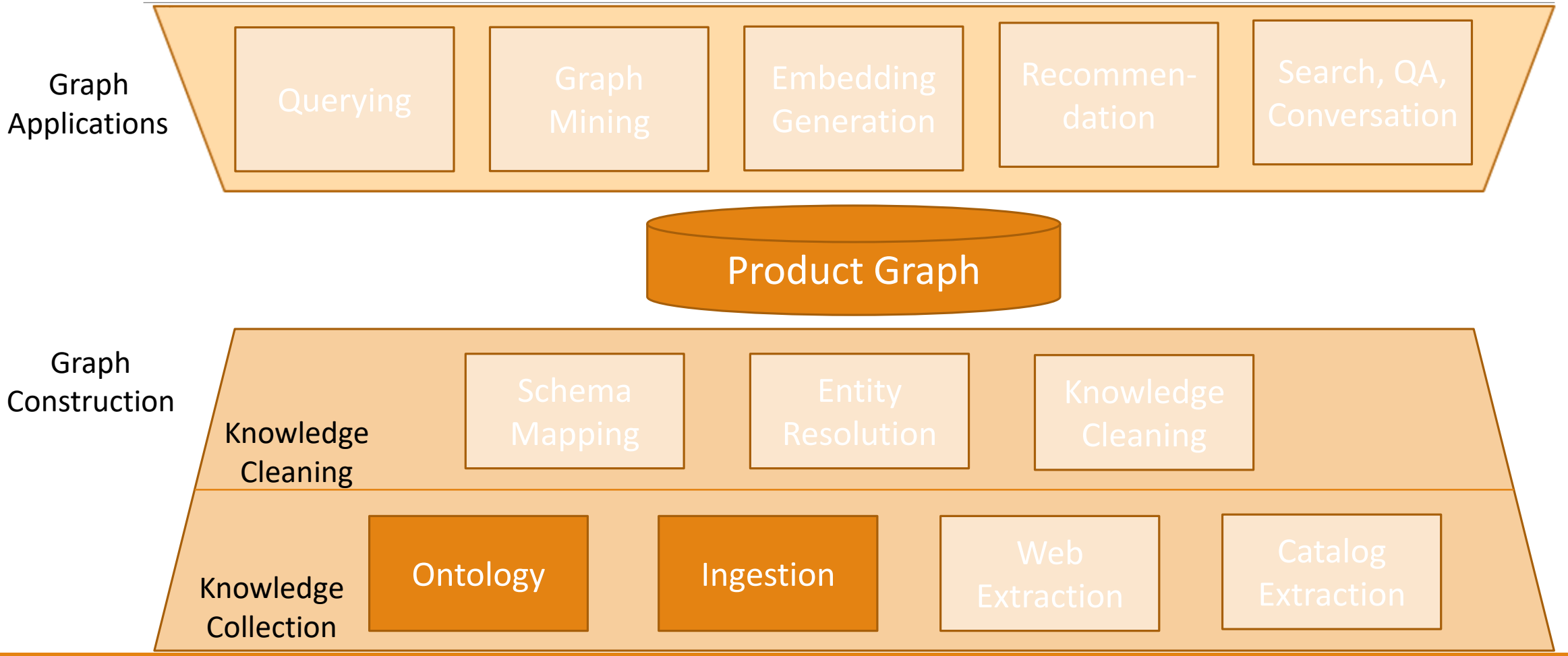


Which BirthDate is correct?

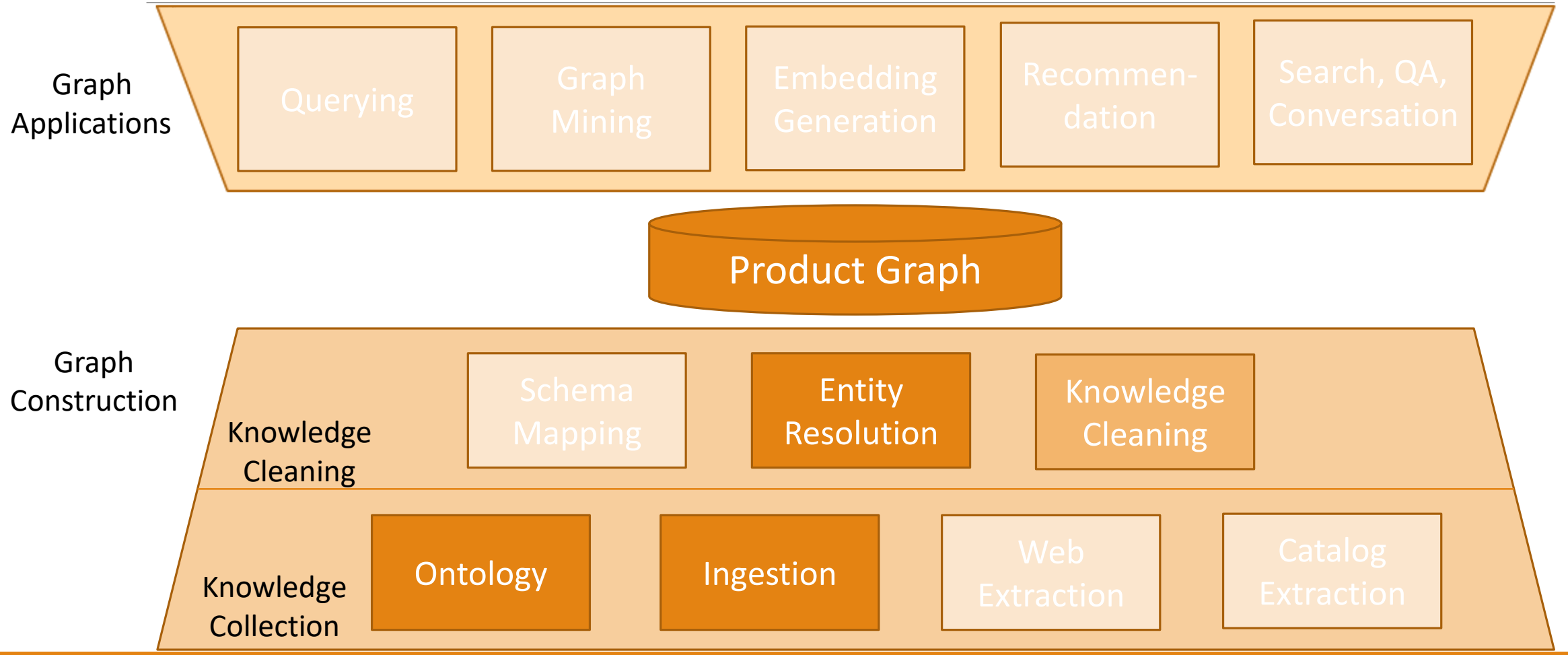
Challenges: Linkage



I. Integrating Knowledge from Structured Sources



I. Integrating Knowledge from Structured Sources

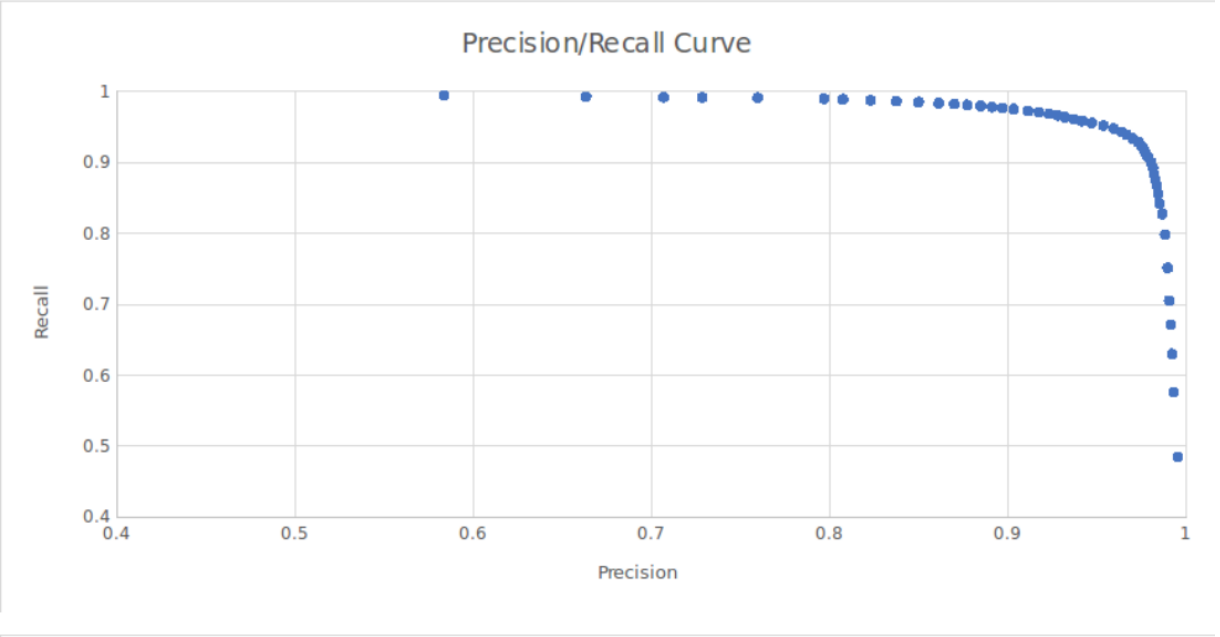


I. Integrating Knowledge from Structured Sources—Entity Resolution

- 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

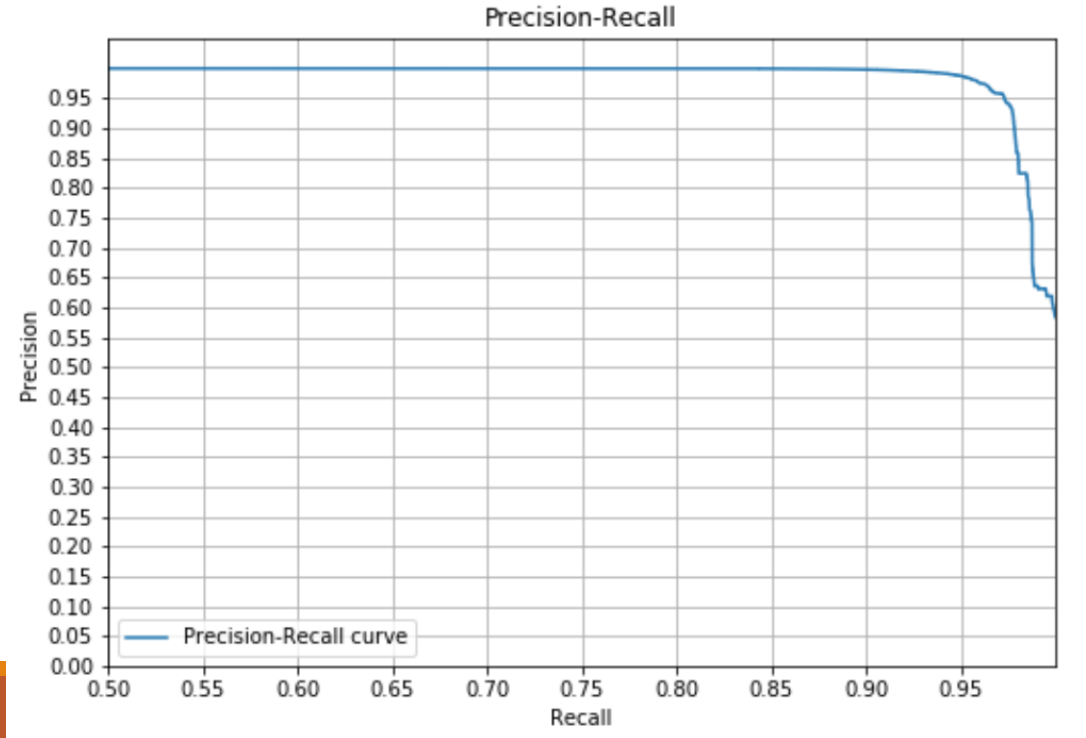
- Our method: Random forest on attribute-wise similarity
- Results between Amazon Movies and IMDb:

  200K labels

  ~150 features

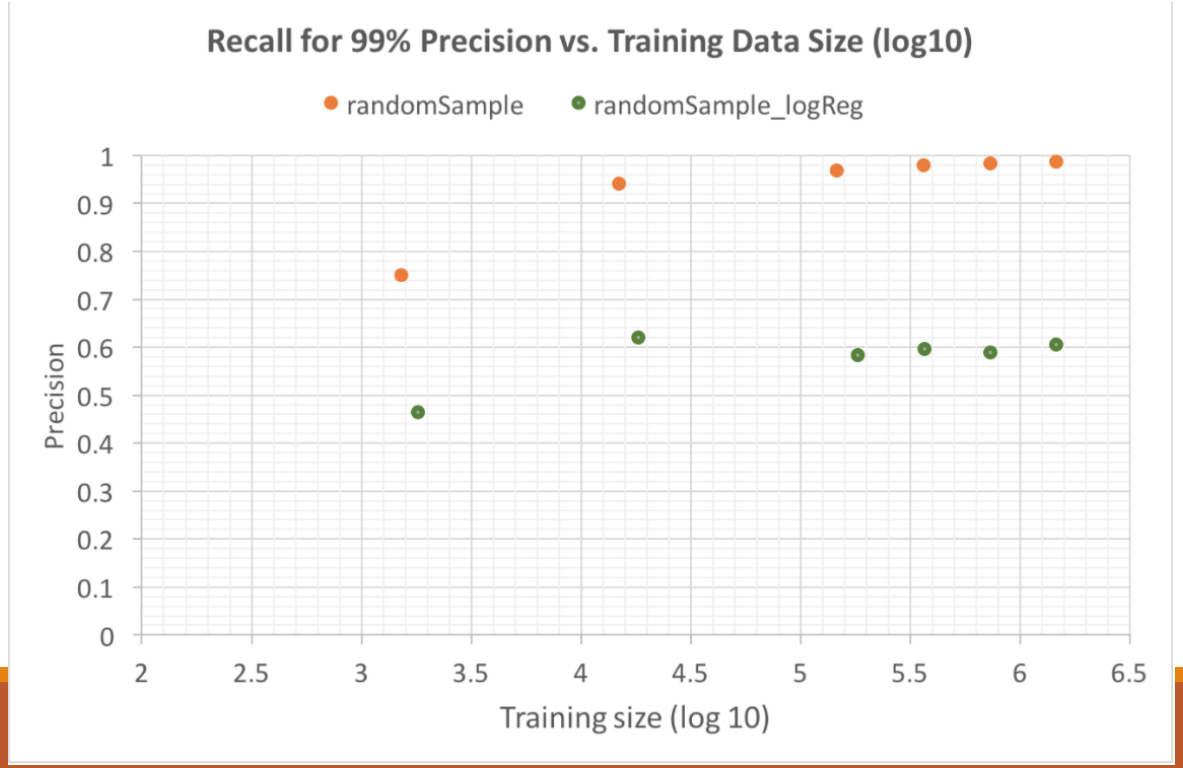
  AUPRC=0.9923

  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



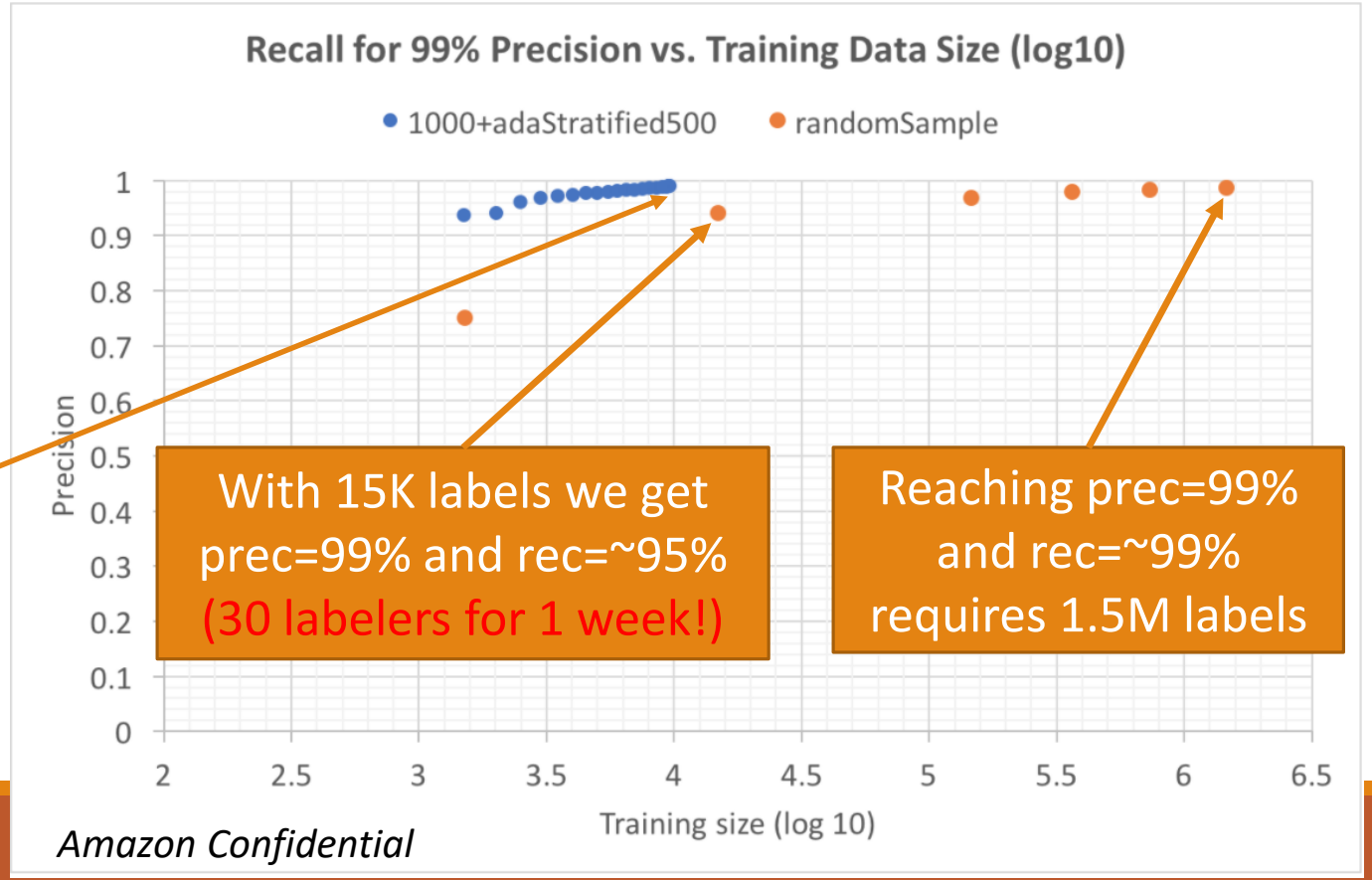
I. Integrating Knowledge—Entity Resolution Which ML Model Works Best?

- 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

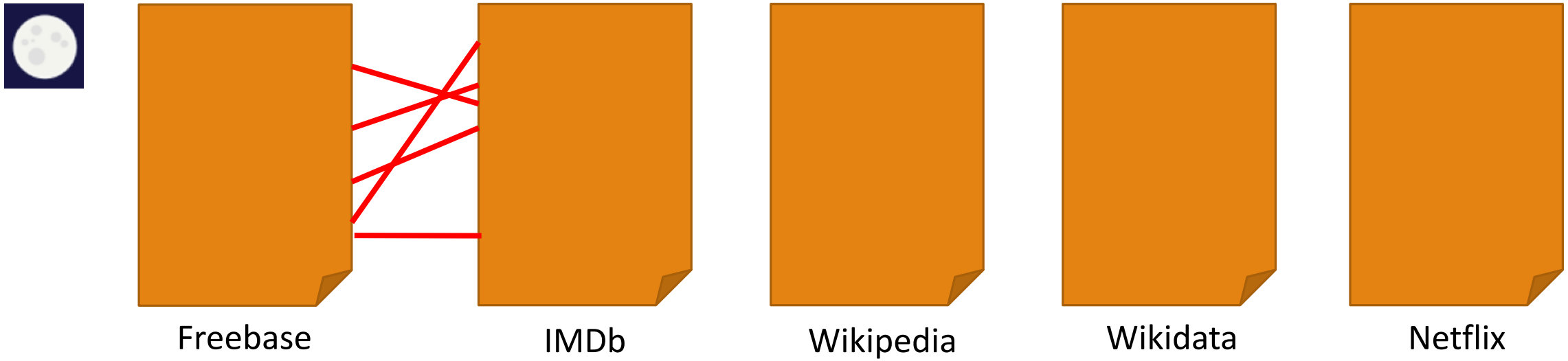
- Moonshot: Apply active learning to minimize #labels

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



I. Integrating Knowledge from Structured Sources—Entity Resolution

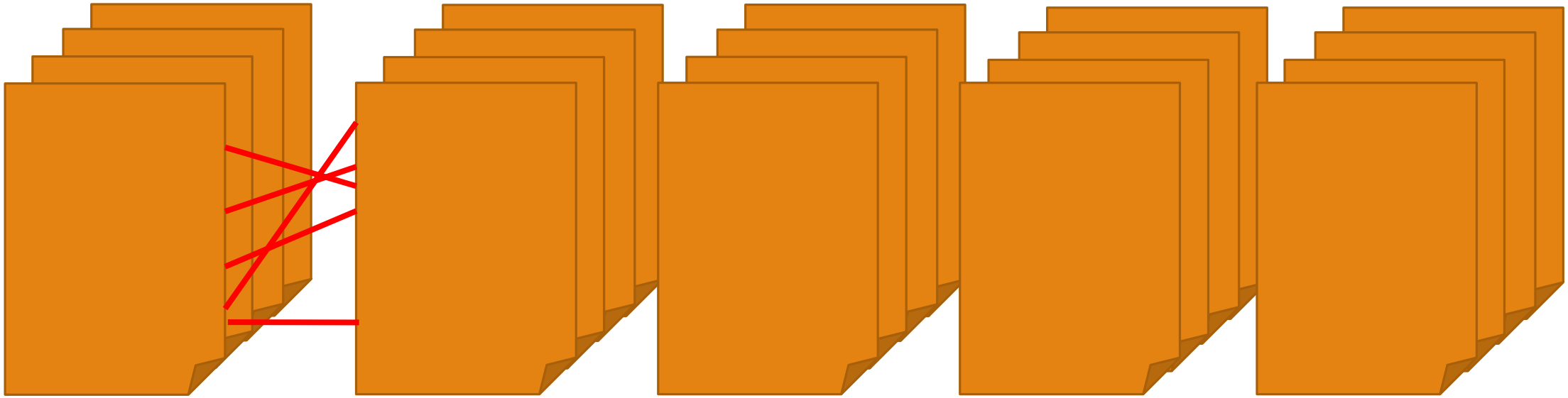
- Moonshots: Seamless incremental graph linkage with high precision and recall



  - 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

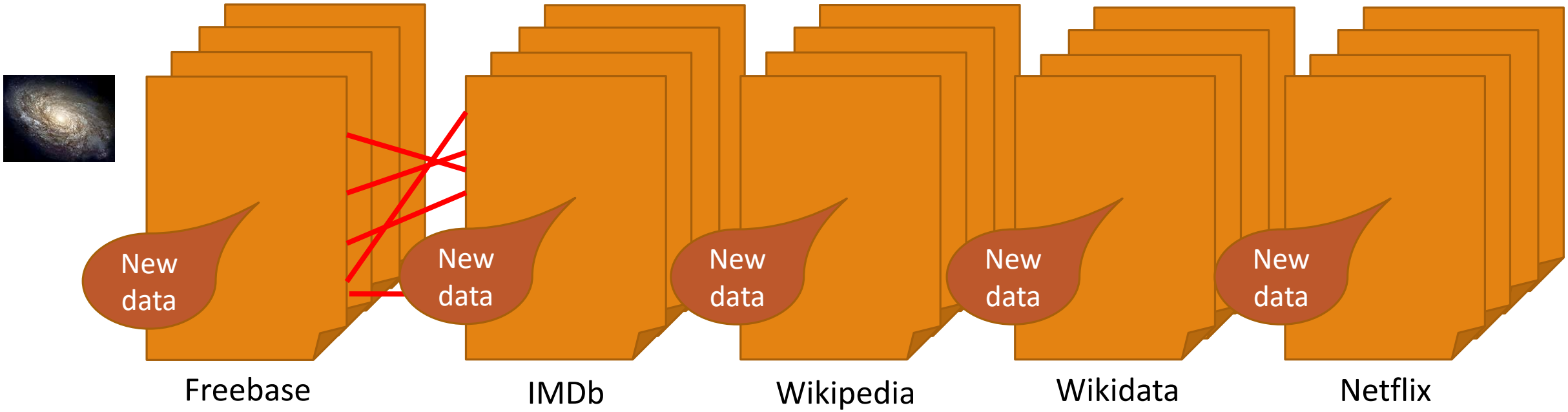
- Moonshots: Seamless incremental graph linkage with high precision and recall



  - 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?*

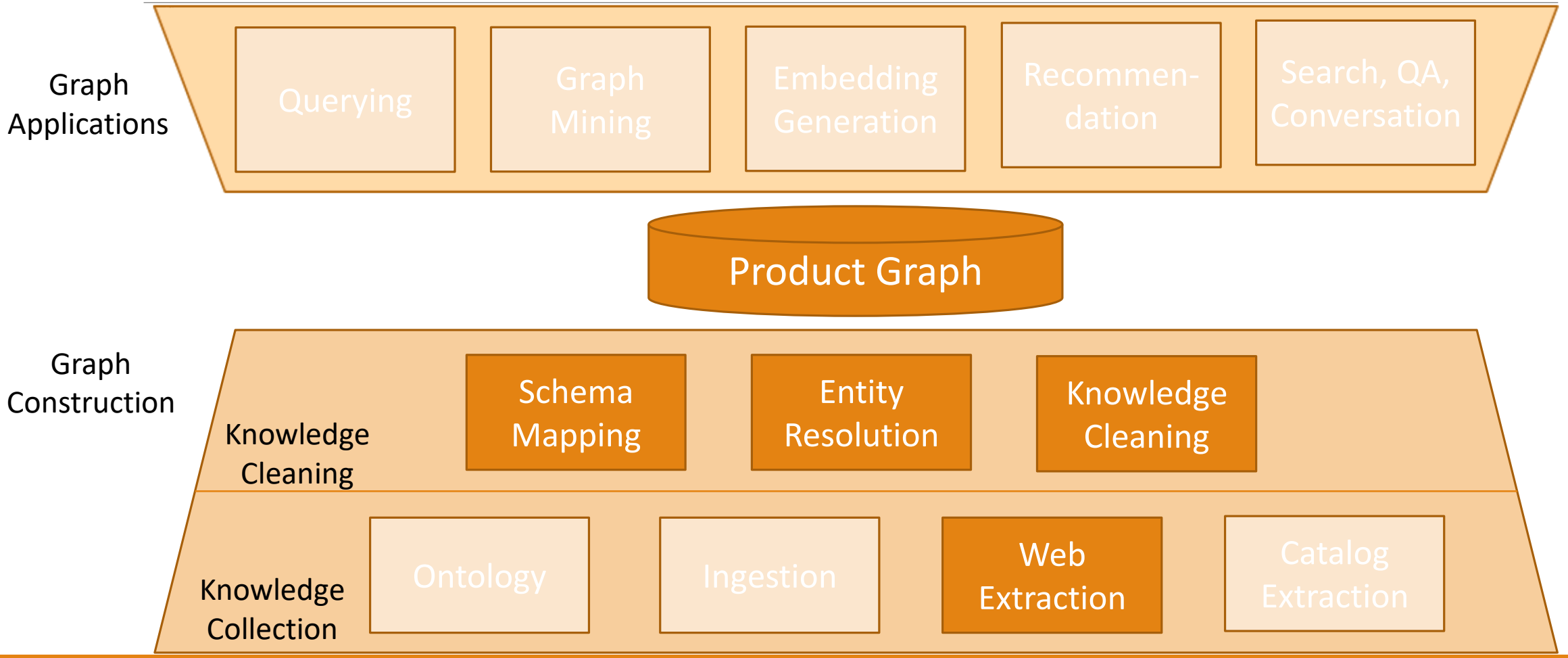
I. Integrating Knowle f s i Structured Sources—

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

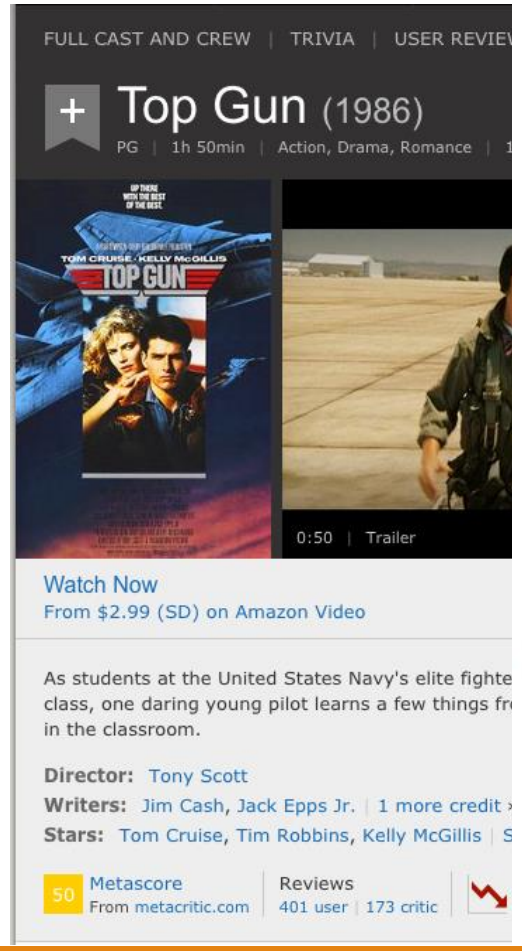


  - New data are arriving over time, requiring incremental linkage and model evolution
  - *How to perform incremental linkage and evolve the model?*

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



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



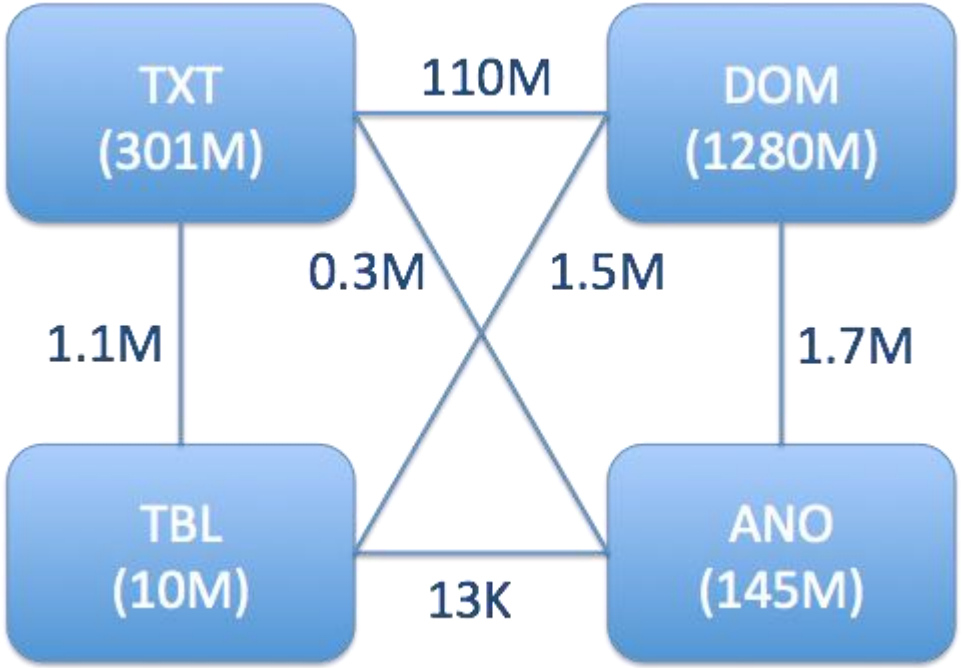
Aamir Khan is receiving rave reviews for Dangal.

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

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

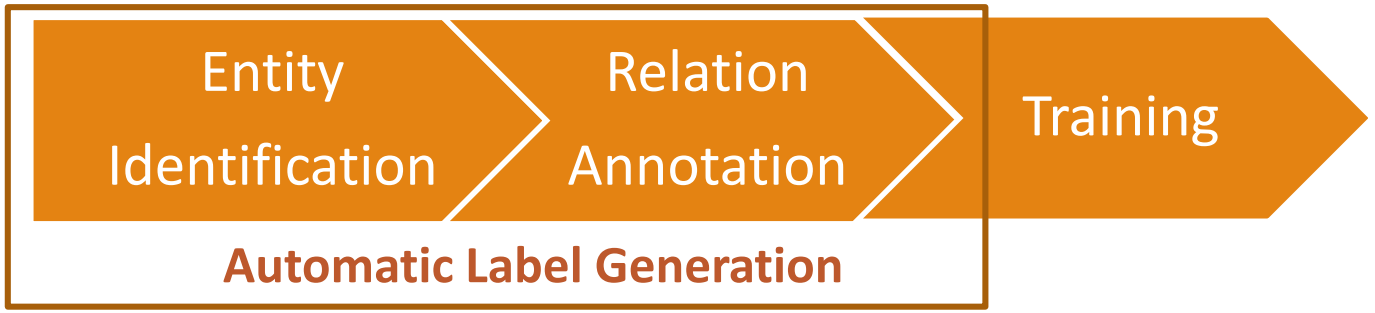
- Knowledge Vault @ Google showed big potential from DOM-tree extraction [Dong et al., KDD'14][Dong et al., VLDB'14]

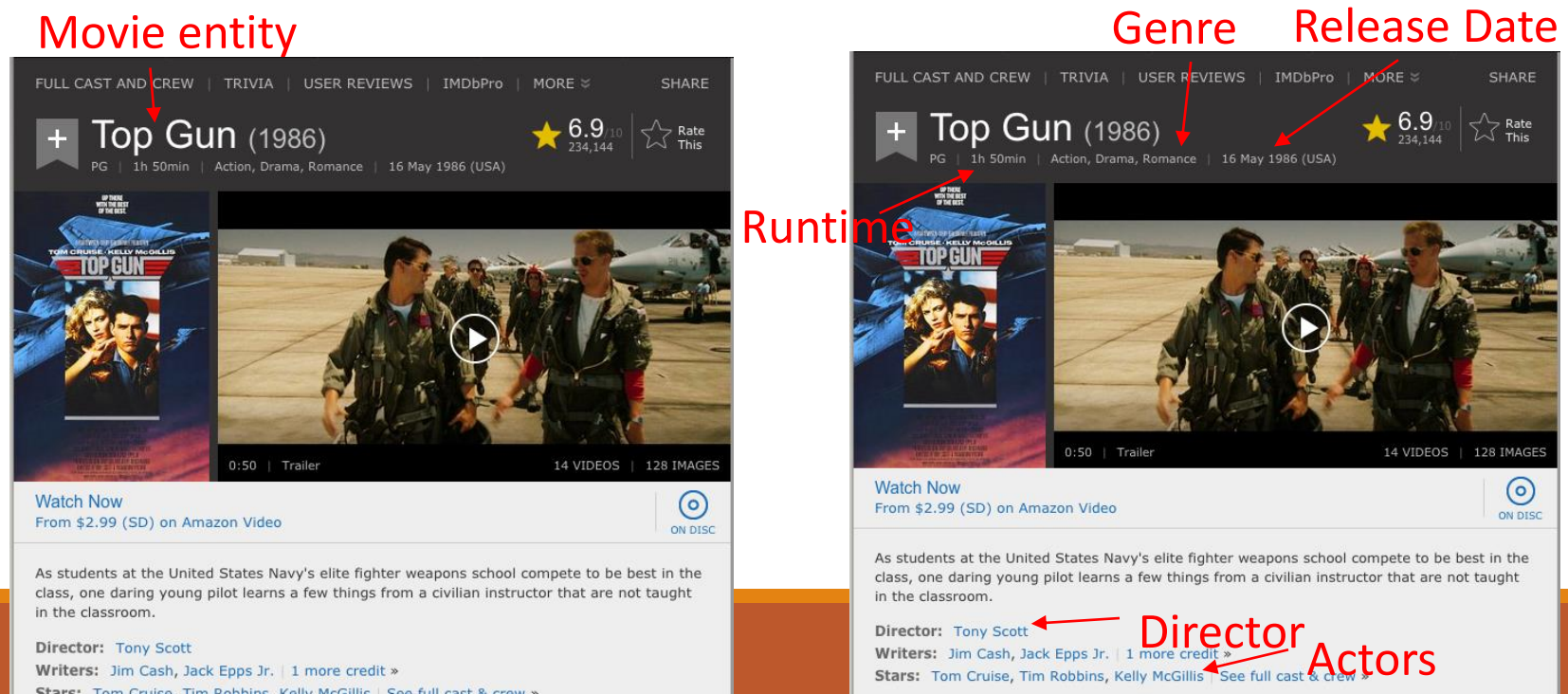
Accu	Accu (conf $\geq$ .7)
0.36	0.52



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

II. Extracting Knowledge from Web—Distantly Supervised DOM Extraction





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

- 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

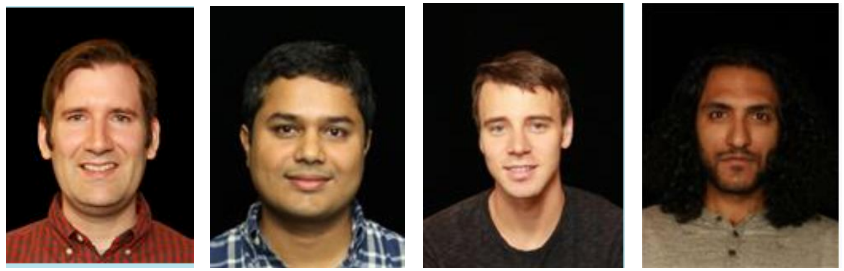
- Extraction on long-tail movie websites

#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 : **3.2**
Annotated : Extracted #triples	1 : **4.1**
# Extractions	1.7 M
Precision	83%

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



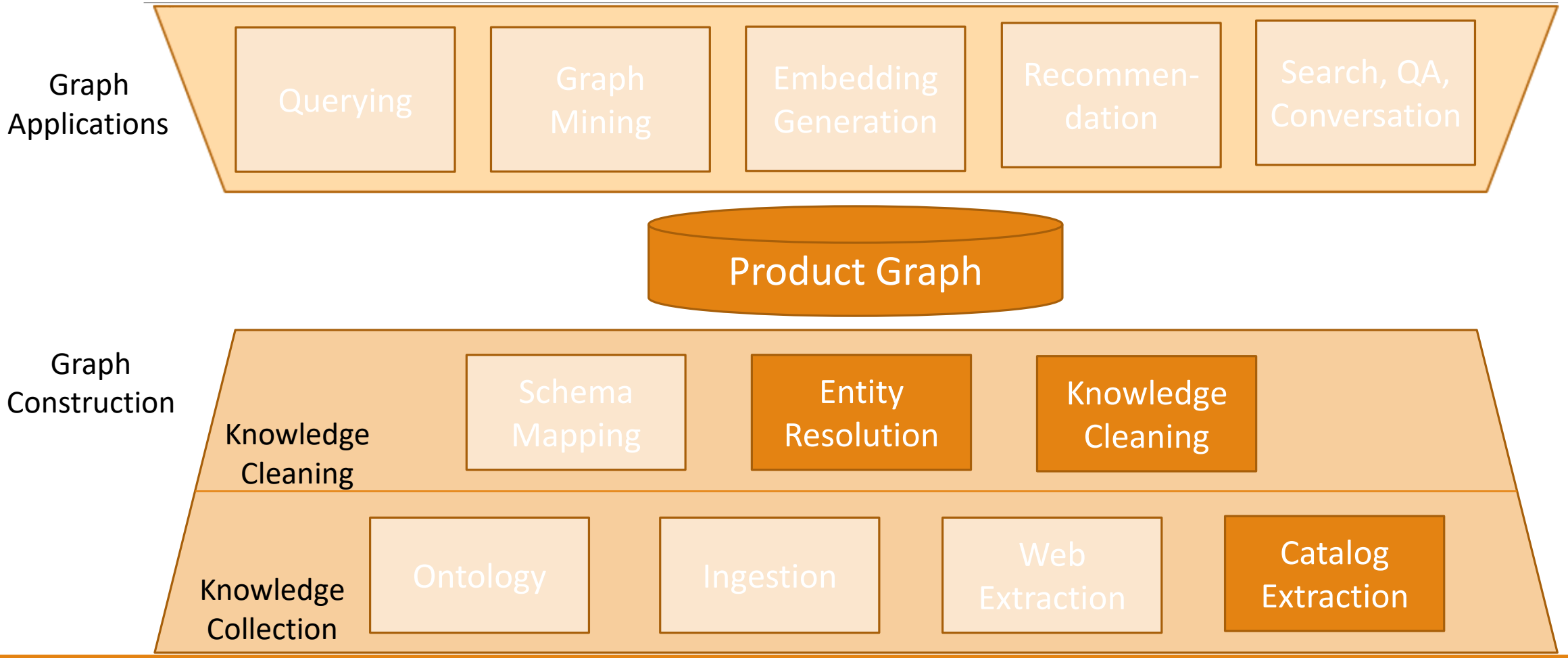
Annotation-based knowledge extraction

Distantly supervised web extraction

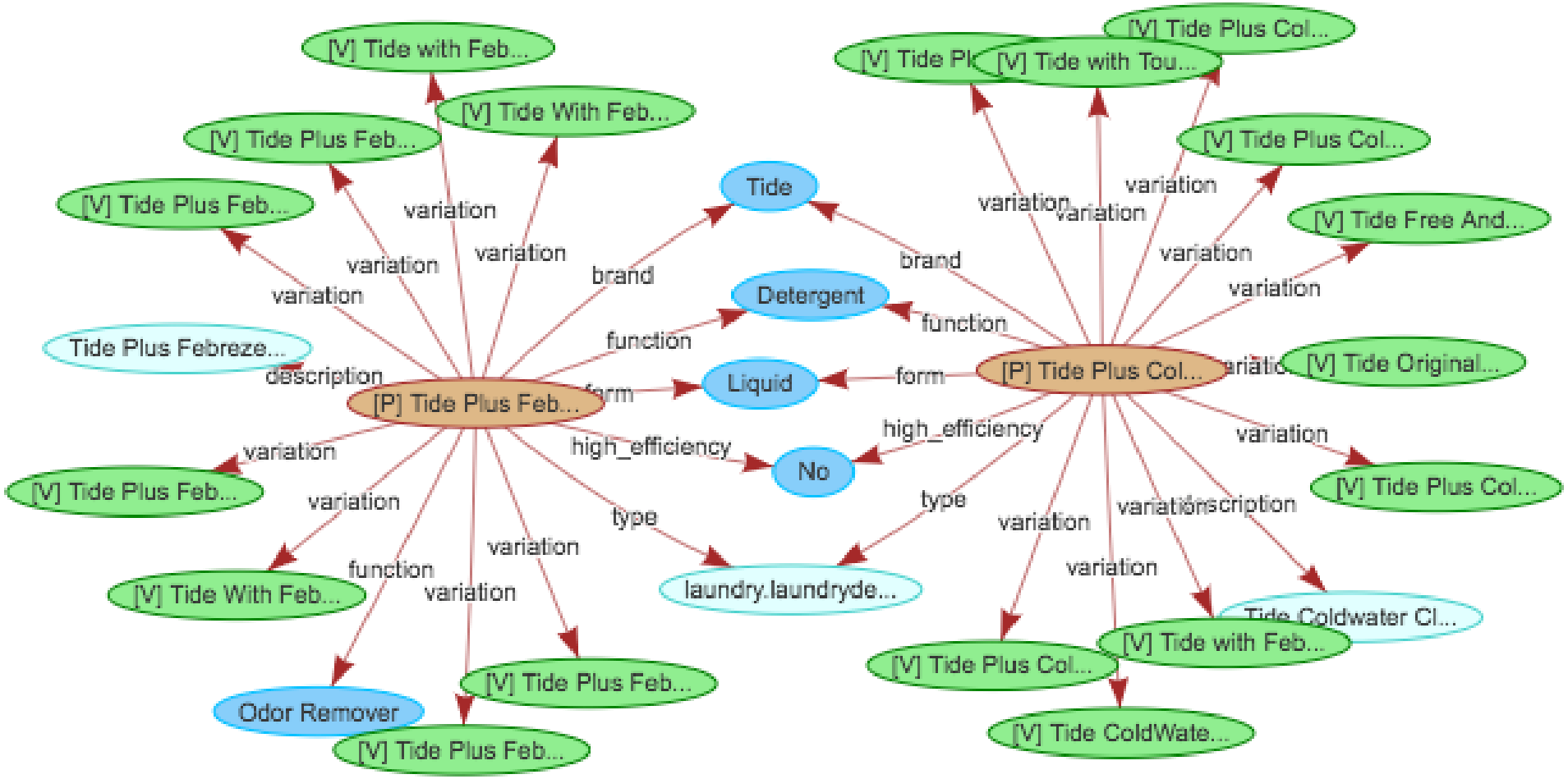
OpenIE DOM extraction

Nearly-automatic interactive extraction on any new vertical

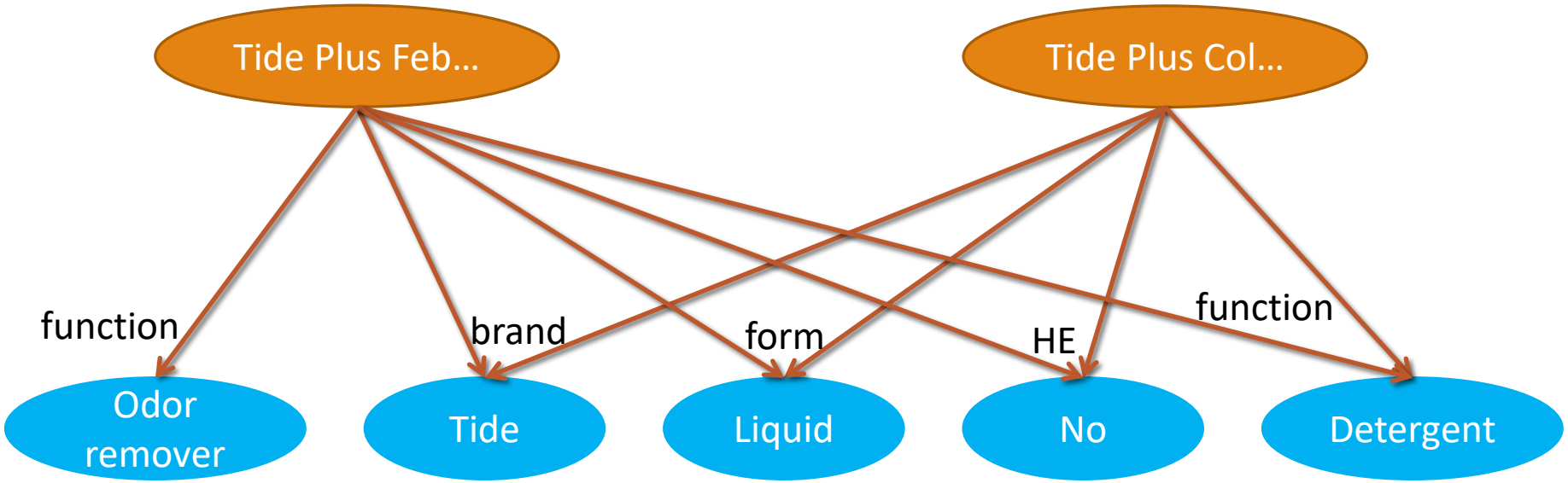
III. Building a Shallow Graph from Product Profiles in Amazon Catalog



Another Example of Product Graph



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



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

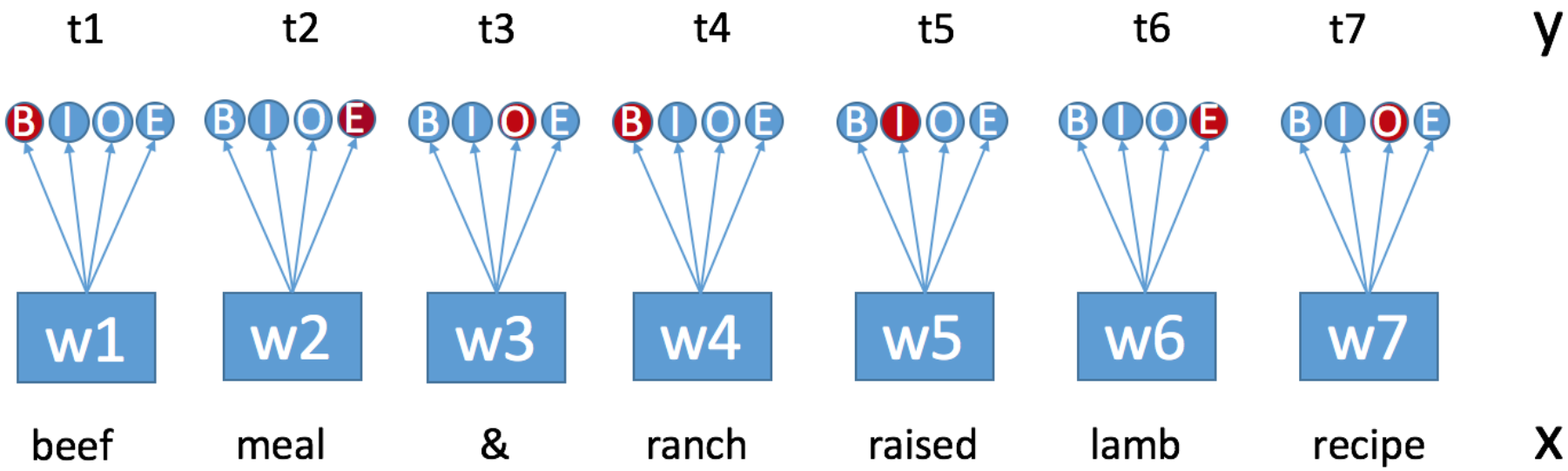
name	form	scent
Tide Detergent with Febreze Freshness		
Gain Apple Mango Tango Liquid Laundry Detergent		
Gain Joyful Expression Powder Detergent		
Tide PODS Original Scent HE Turbo Laundry Detergent Pacs 81-load Tub		
Tide PODS Free & Gentle HE Turbo Laundry Detergent Pacs 35-load Bag		

III. Open Attribute Extraction by Named Entity Recognition

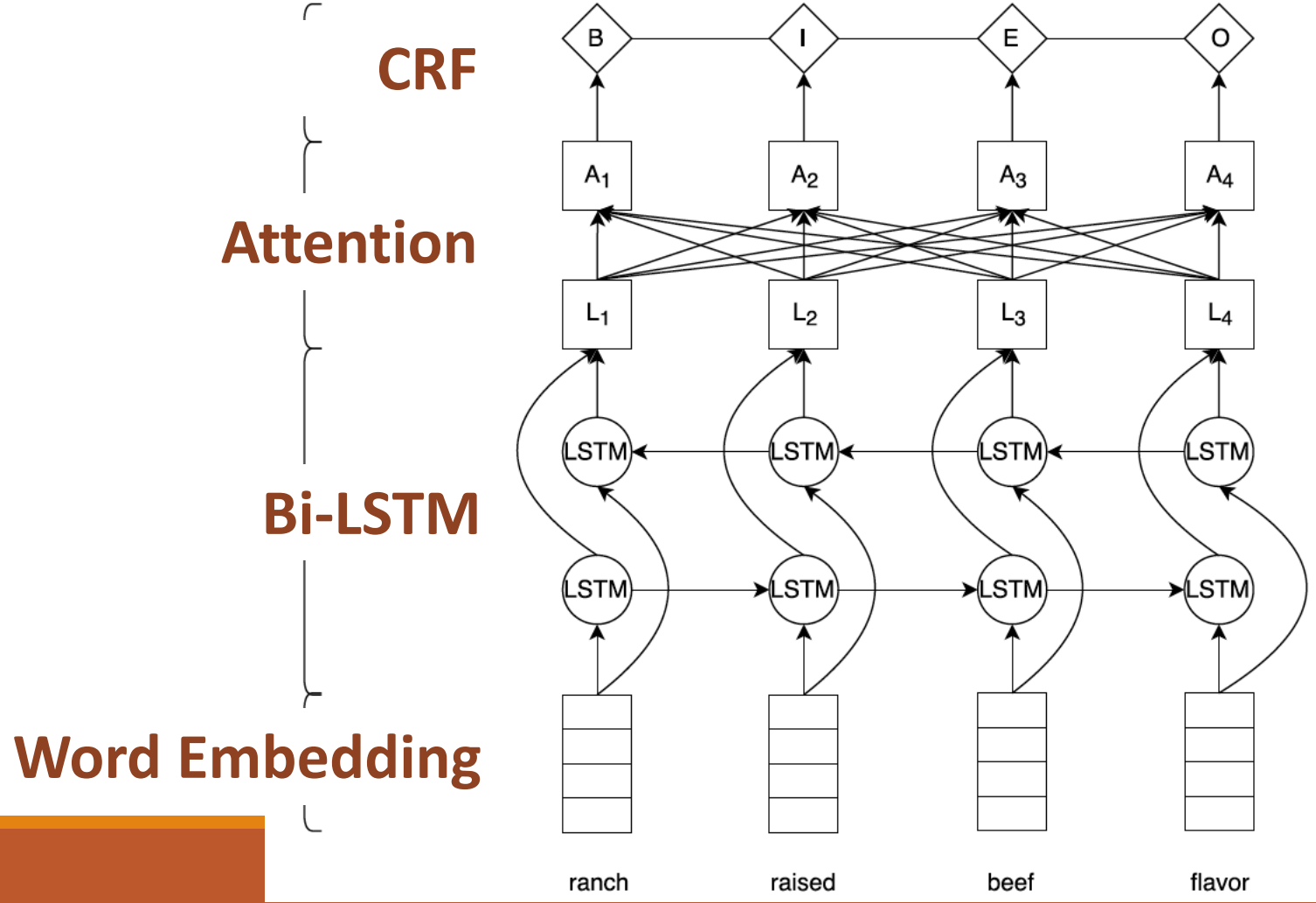
- B — Beginning of entity
- I — Inside of entity
- O — Outside of entity
- E — End of entity

x={w1,w2,…,wn} input sequence

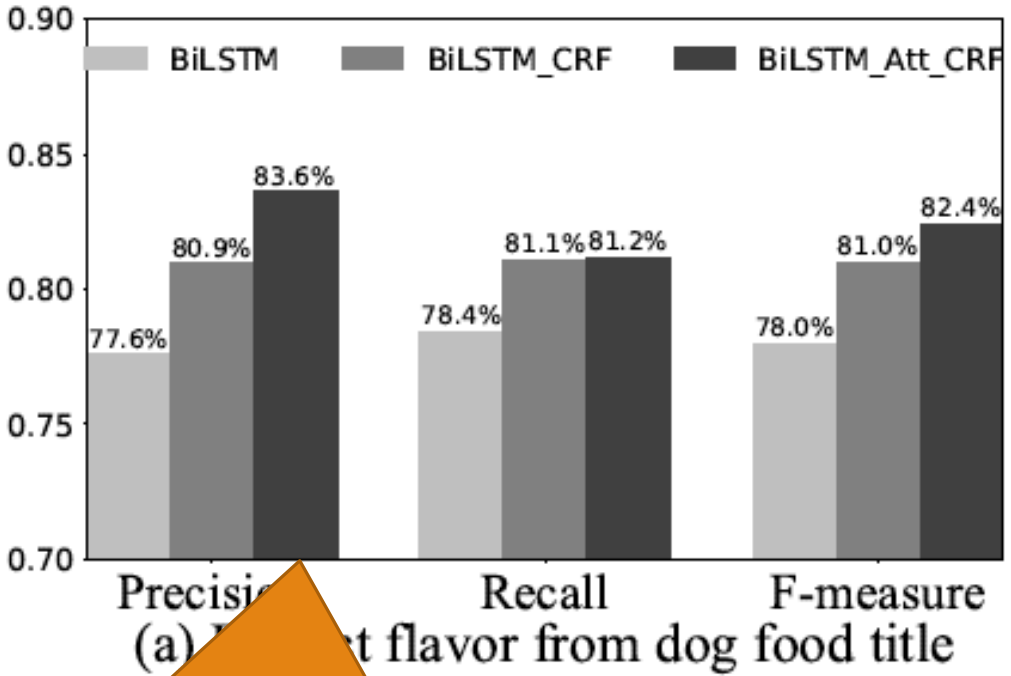
y={t1,t2,…,tn} tagging decision



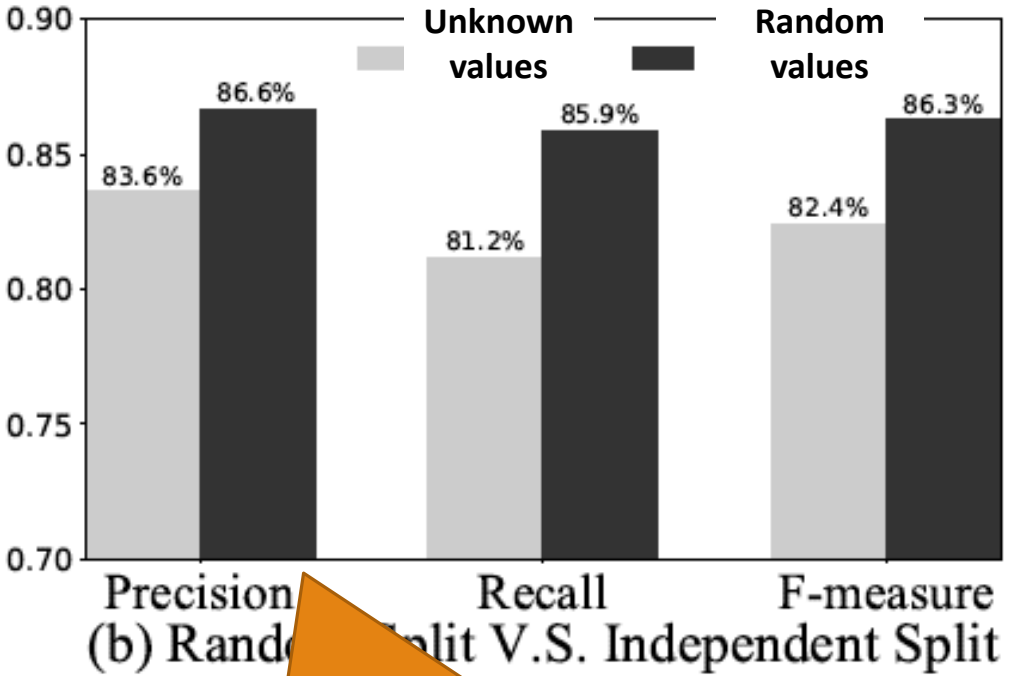
III. Open Attribute Extraction by Named Entity Recognition



III. Open Attribute Extraction by Named Entity Recognition



(a) t flavor from dog food title



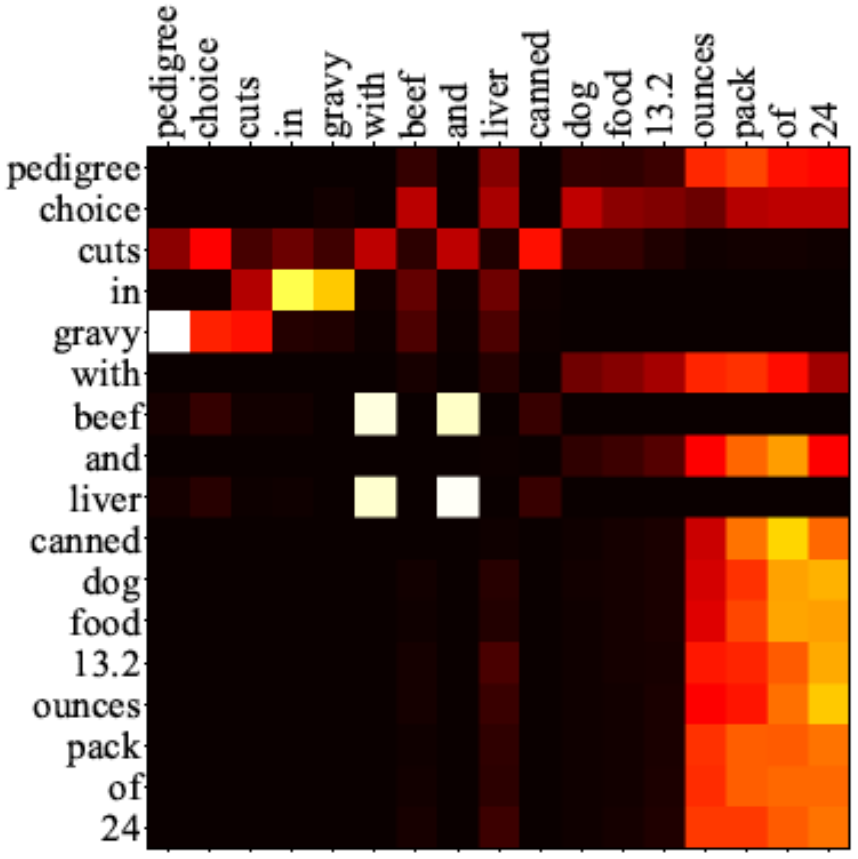
(b) Rand lit V.S. Independent Split

BiLSTM+CRF+Attention obtains best results

Extraction on new values is comparable to already known values

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

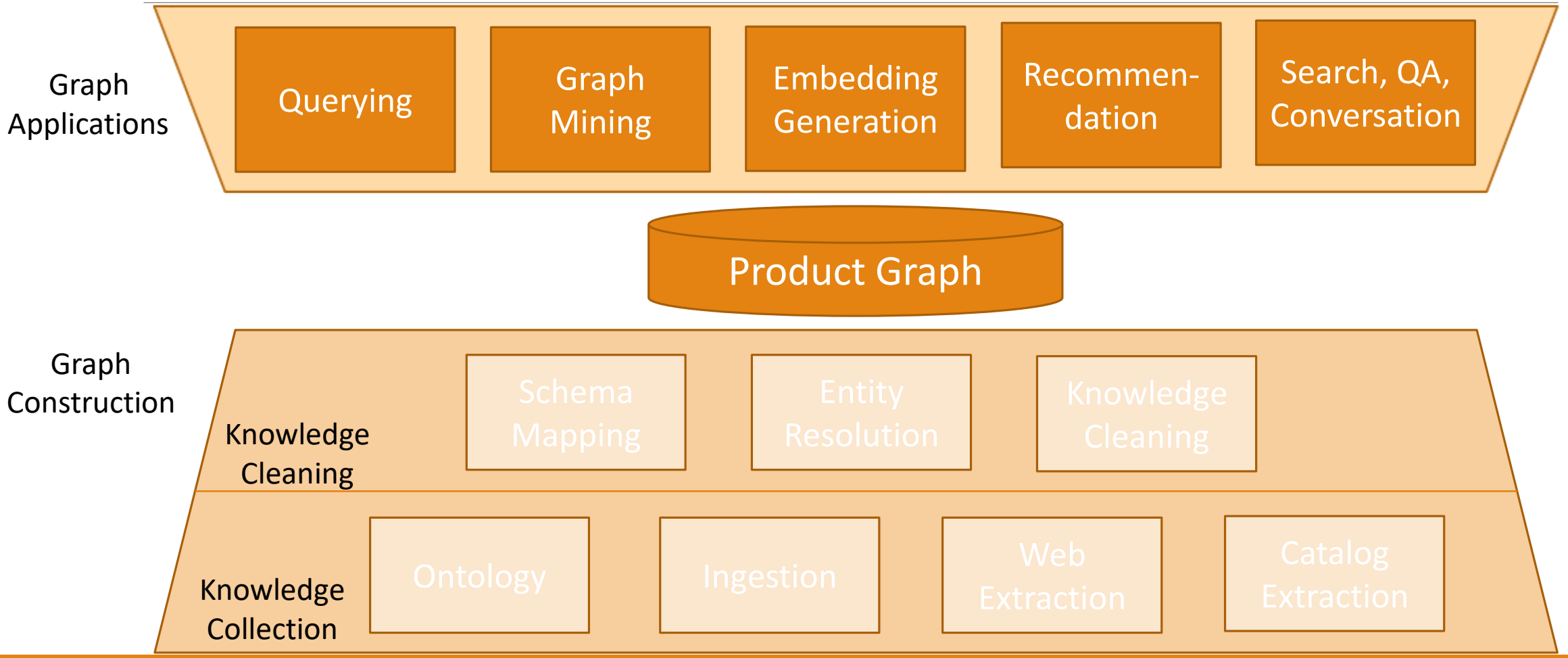

Product profile extraction

Automatically building a shallow KG

Open aspect extraction

Review extraction & sentiment analysis

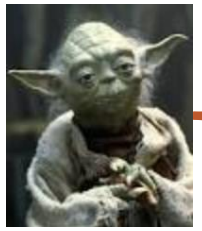
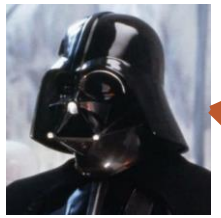
IV. Graph Mining and Embedding

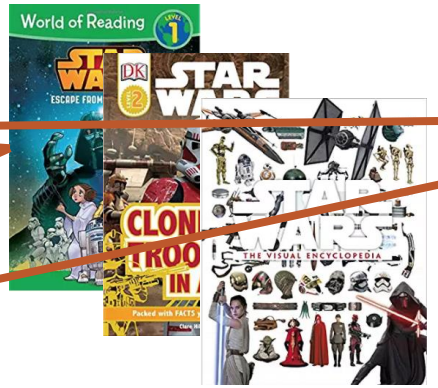


IV. Graph Mining



Which char more important?

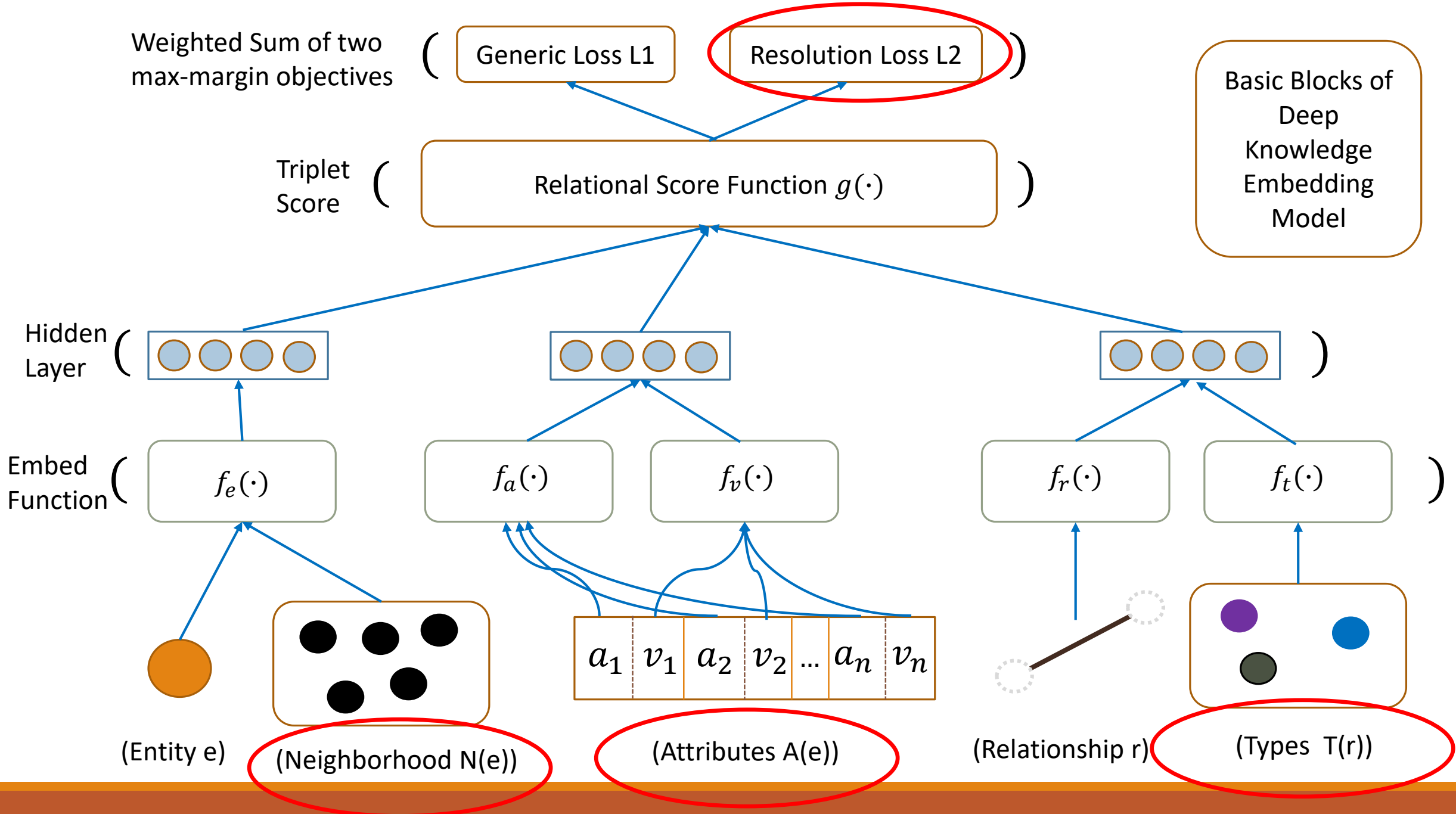





Trend of Darth Vader lamps?

Why people who bought this lamp also bought this chair?

IV. Graph Embedding


Weighted Sum of two max-margin objectives
Generic Loss L1
Resolution Loss L2
Triplet Score
Relational Score Function $g(\cdot)$
Basic Blocks of Deep Knowledge Embedding Model
Hidden Layer
Embed Function
$f_e(\cdot)$
$f_a(\cdot)$
$f_v(\cdot)$
$f_r(\cdot)$
$f_t(\cdot)$
a_1 v_1 a_2 v_2 ... a_n v_n
(Entity e)
(Neighborhood N(e))
(Attributes A(e))
(Relationship r)
(Types T(r))

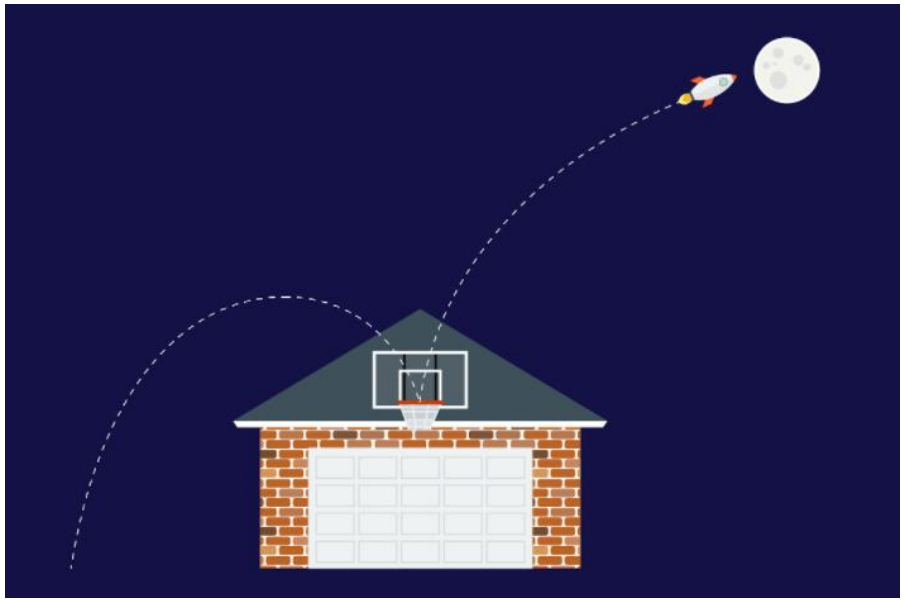
IV. Graph Embedding for




	Link prediction		Entity resolution
	IMDB-MRR	FB-MRR	AUPRC
RASCAL	0.574	0.147	0.32
TransR	0.621	0.132	0.29
HolE	0.712	0.224	0.42
GAKE	0.421	0.112	0.45
Ours	0.763	0.310	0.58

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
- There are many exciting research problems that we are tackling



Thank You!