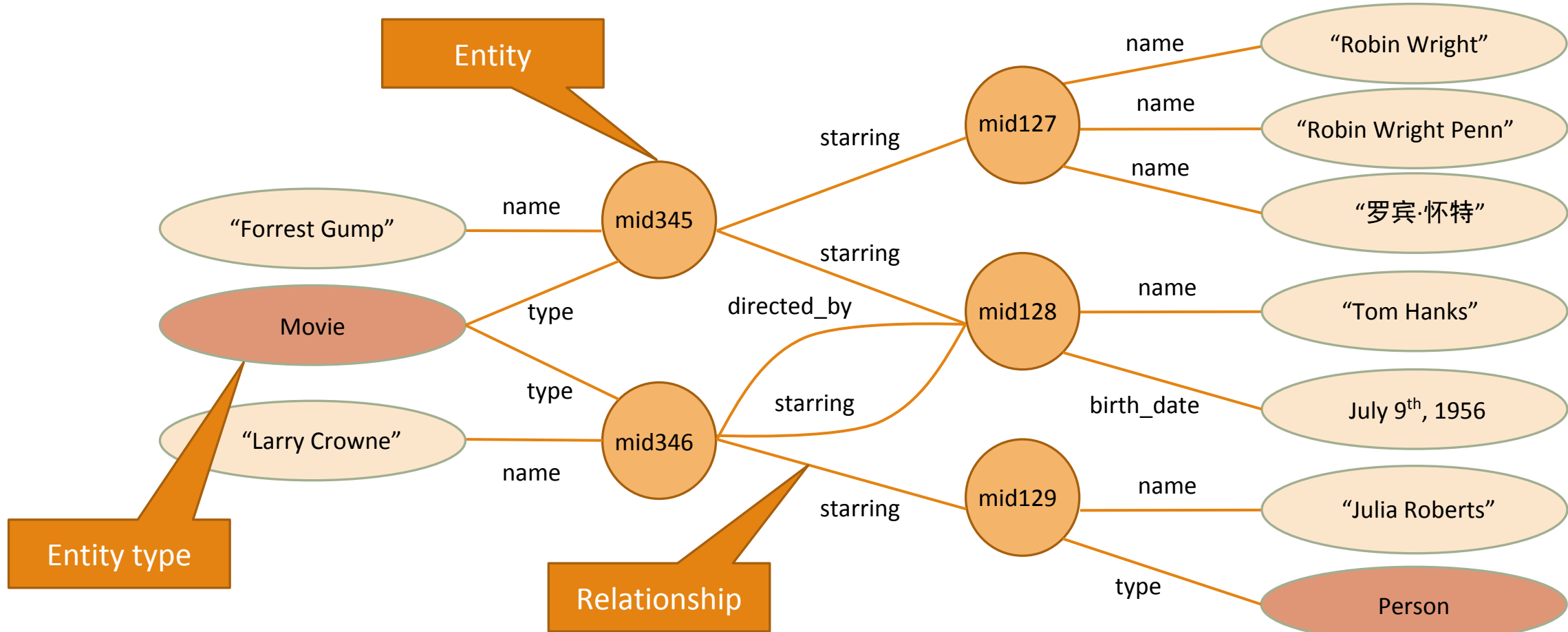# Challenges and Innovations in Building a Product Knowledge Graph

XIN LUNA DONG, AMAZON

JANUARY, 2018
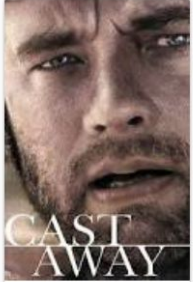
Product Graph vs. Knowledge Graph

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
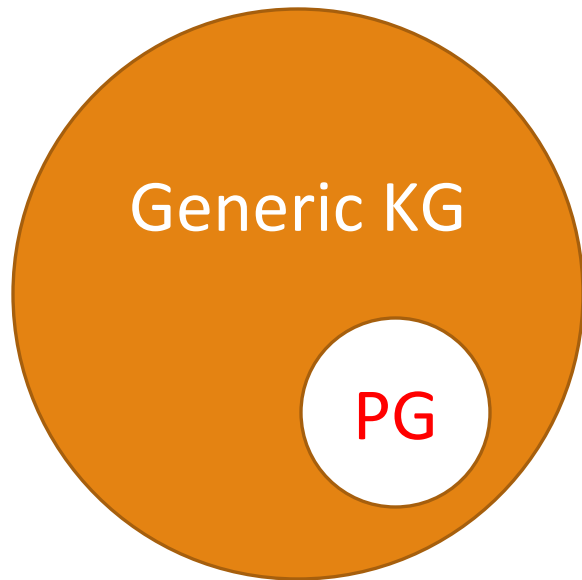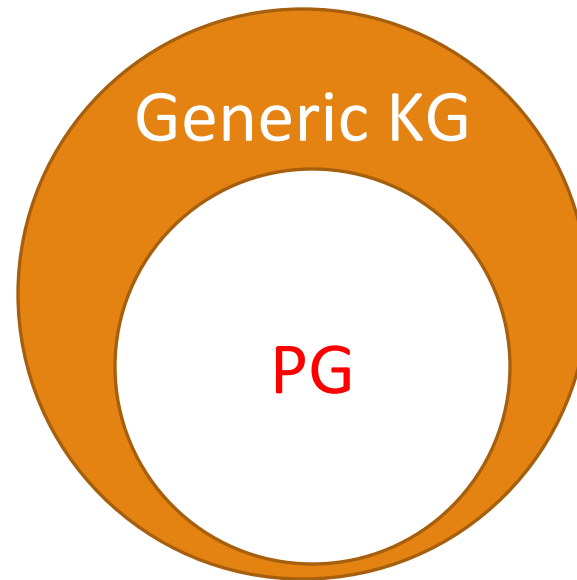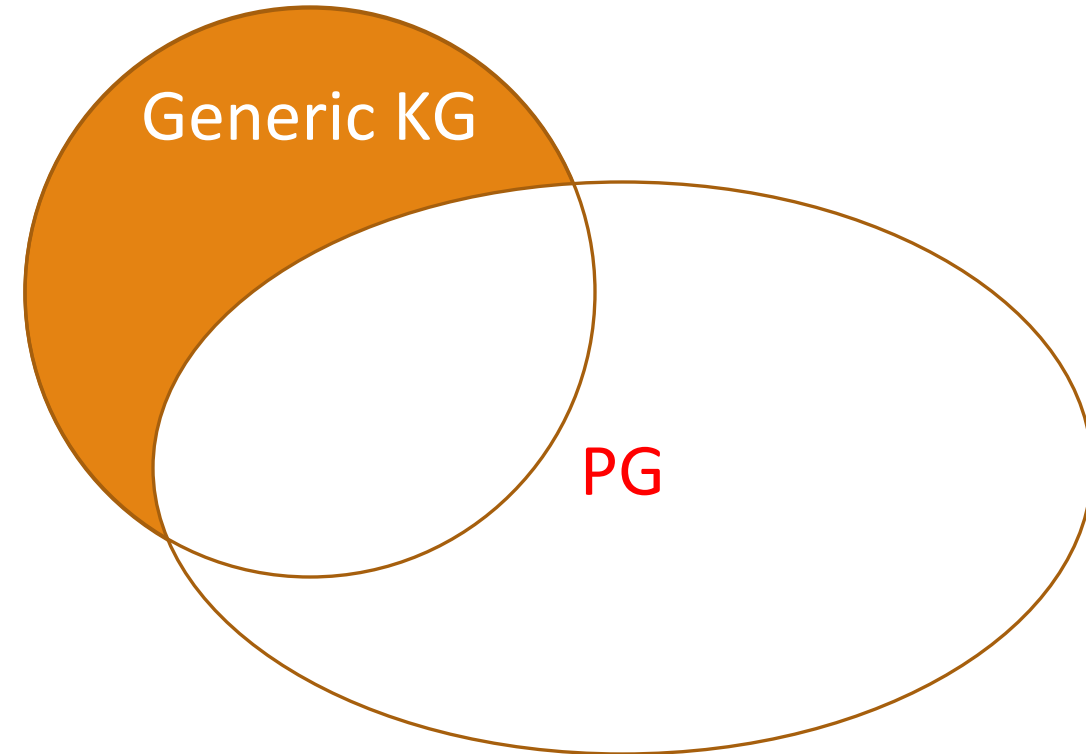
Knowledge Graph Example for 2 Movies
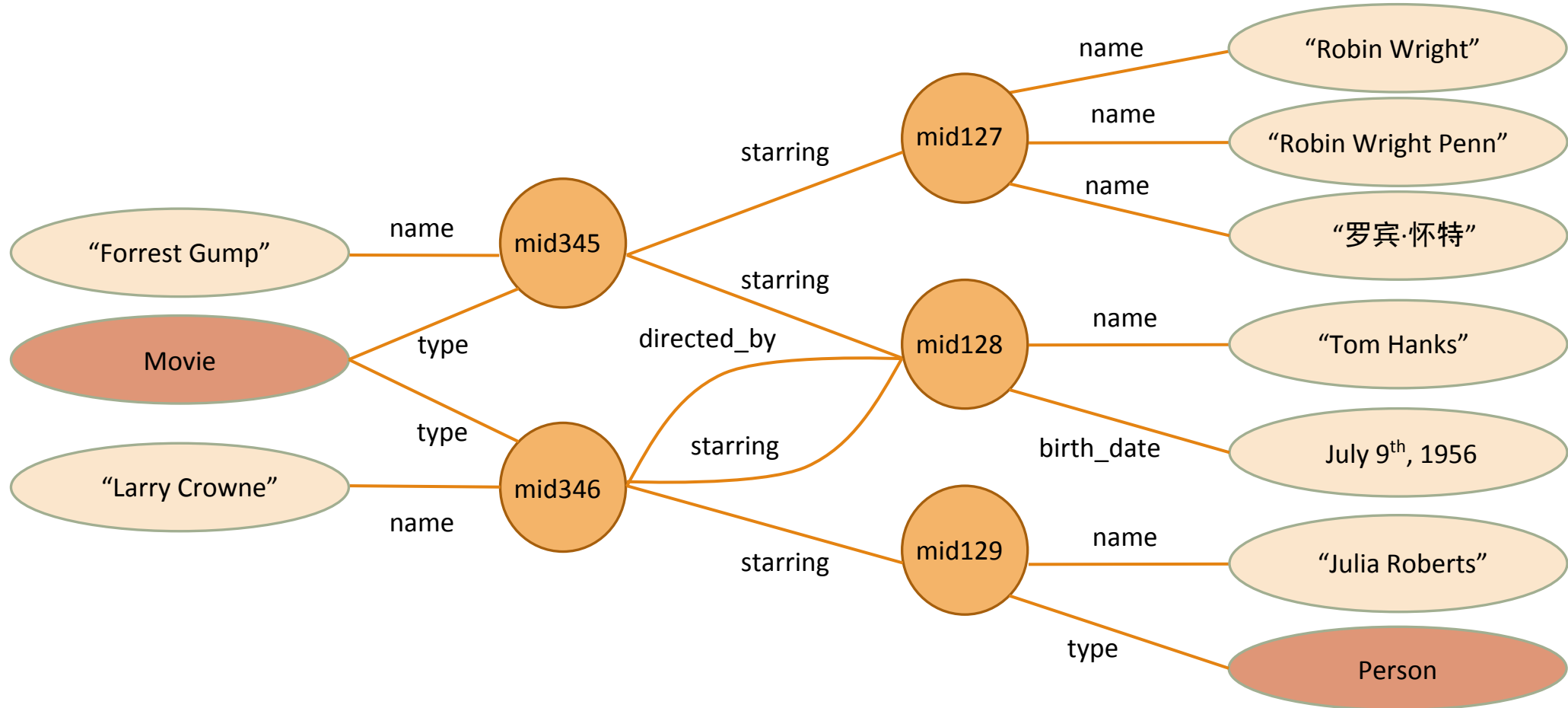
Product Graph vs. Knowledge Graph
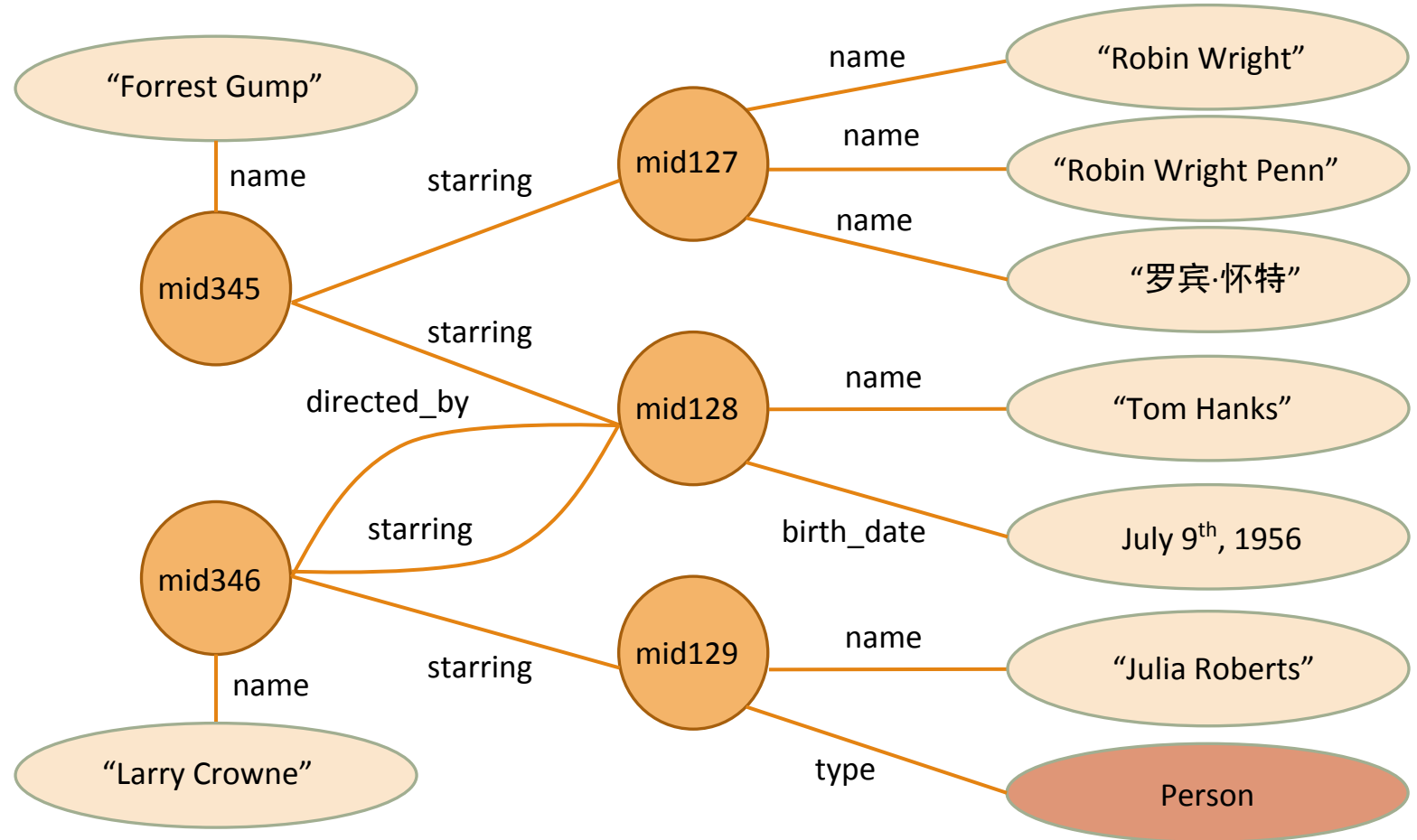
Product Graph vs. Knowledge Graph
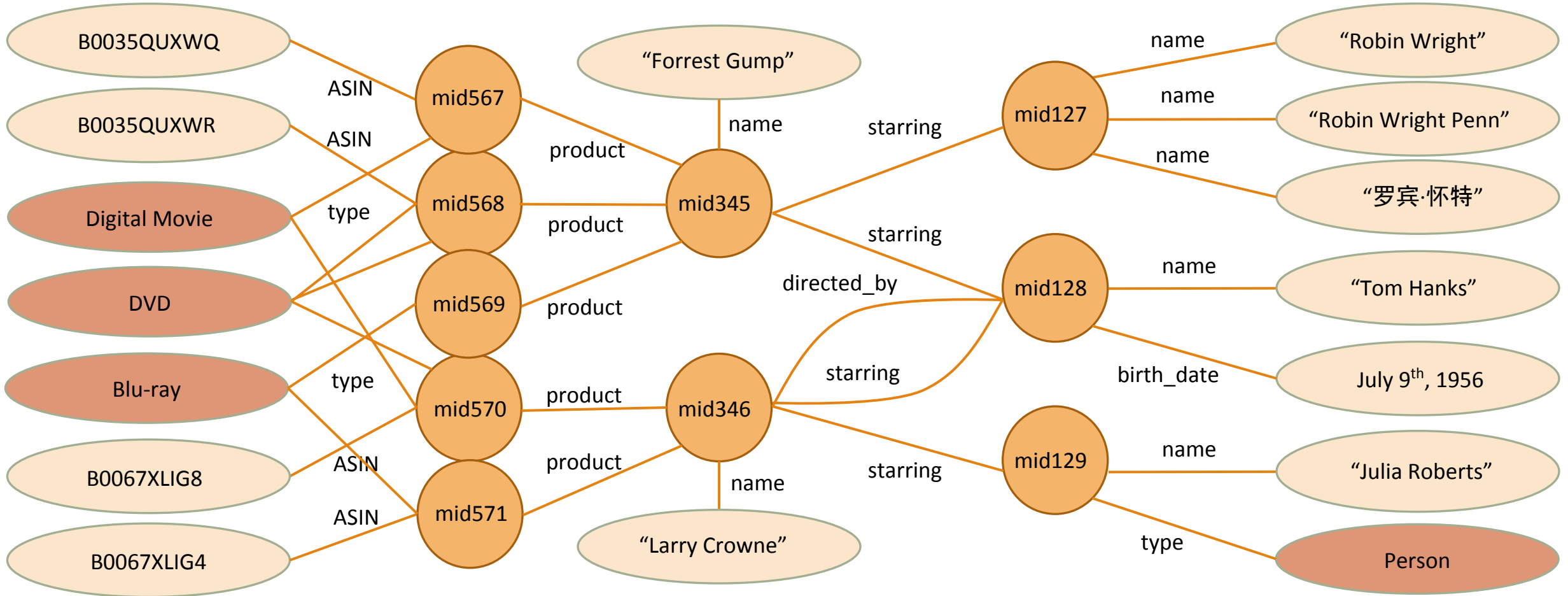
Another Example of Product Graph
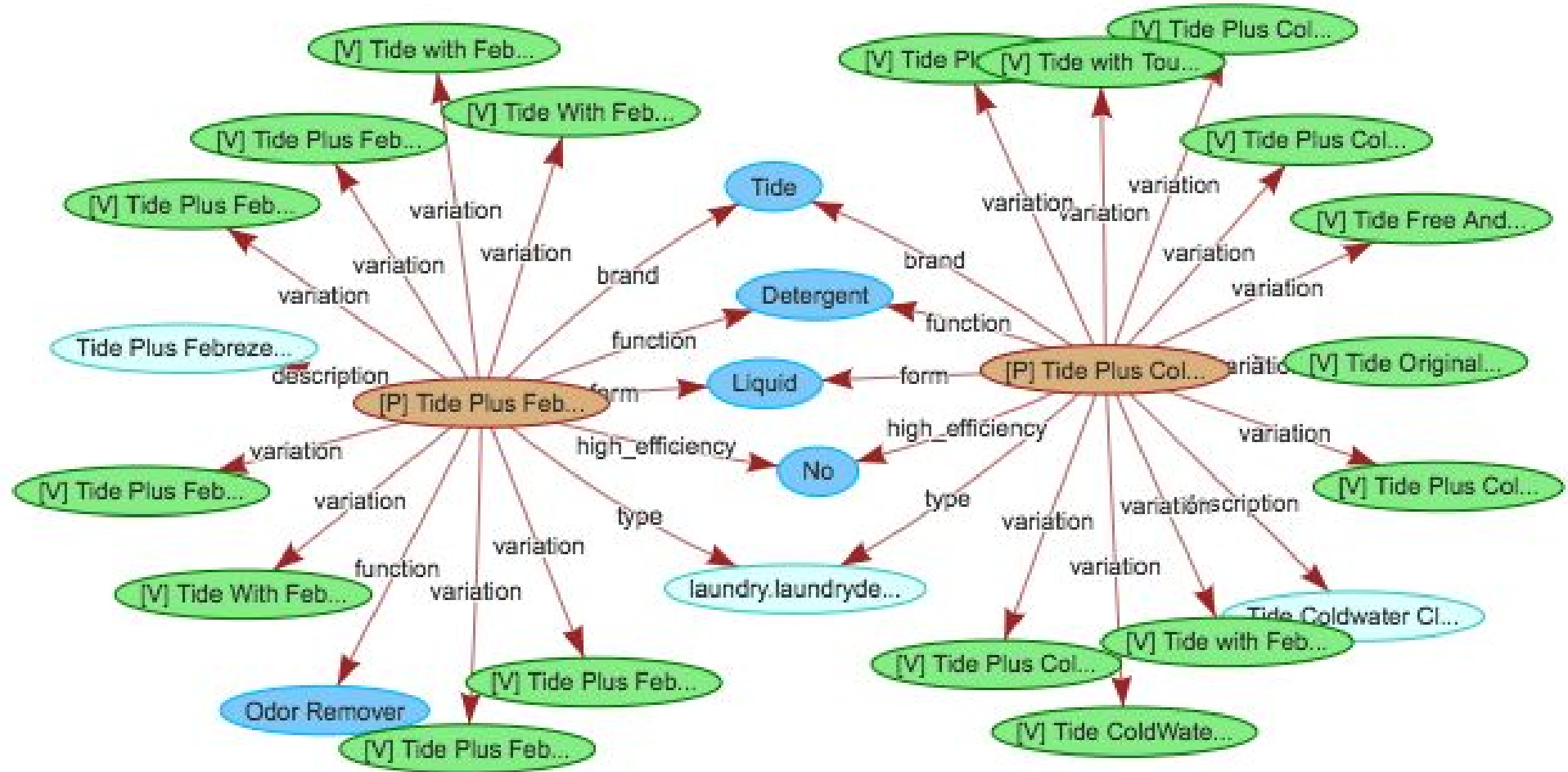
Knowledge Graph vs. Product Graph

(A)

Generic KG

PG

(B)

Generic KG

PG
(Movie,
Music,
Book)

(C) ✓

Generic KG

Movie,
Music,
Book,
etc.

Product
Graph

Generic KG

Movie, Music, Book, etc.
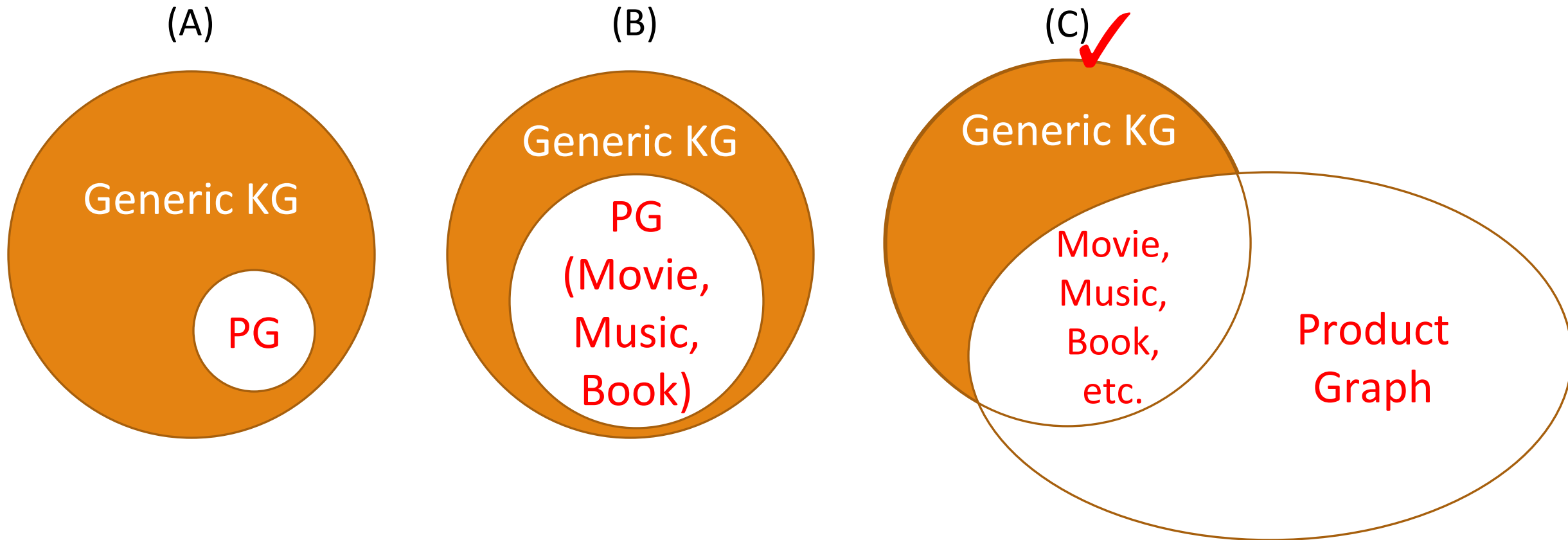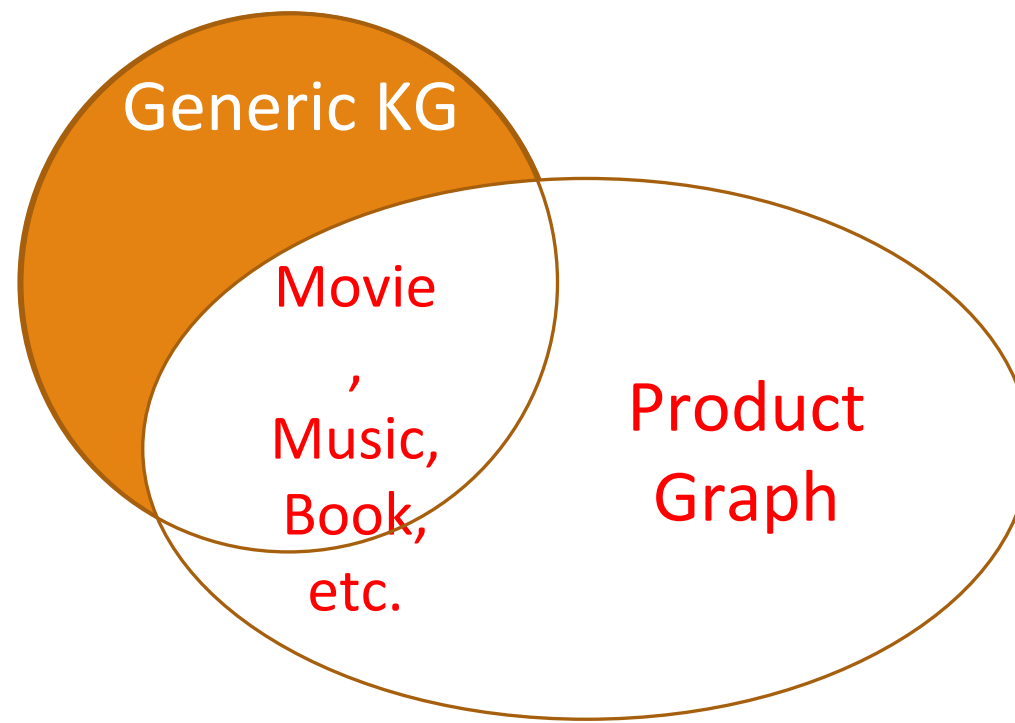
Product Graph

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

How to Build a Product Graph?

Where is Knowledge from?

Product Graph

amazon

WWW

Structured Data
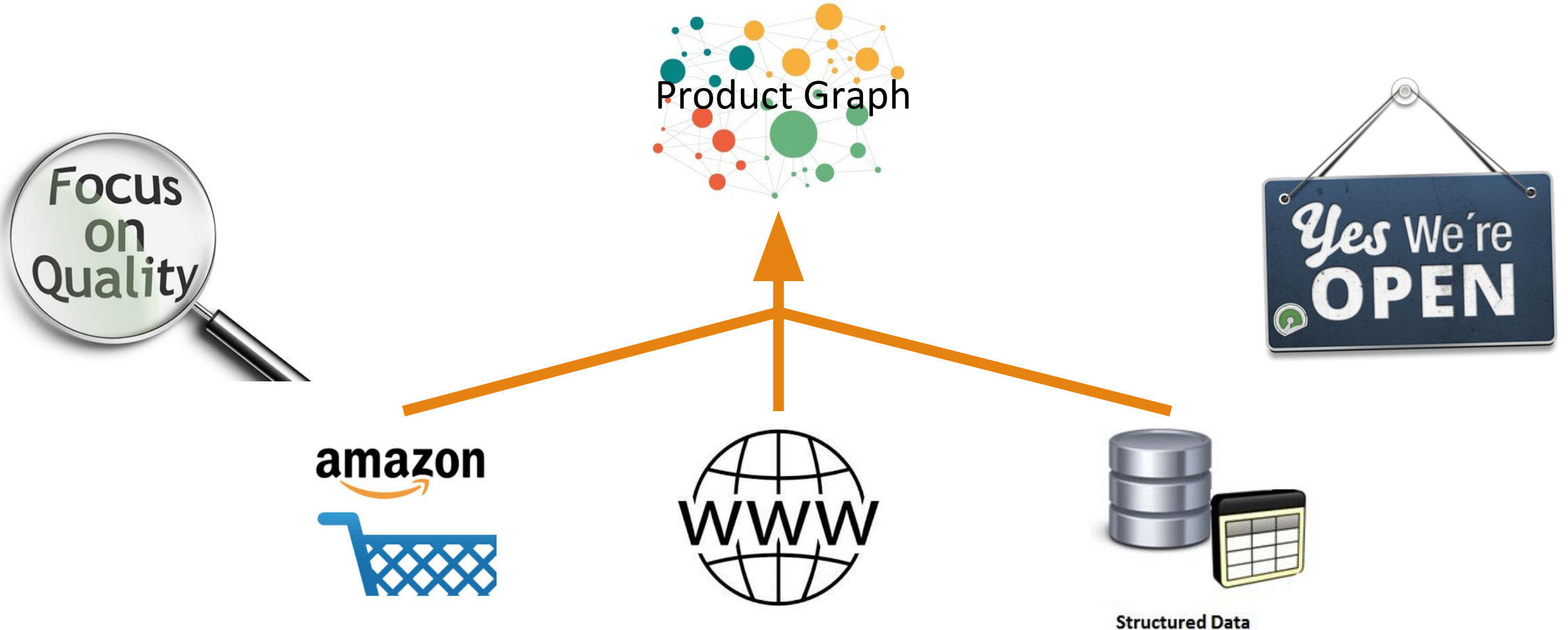
Architecture
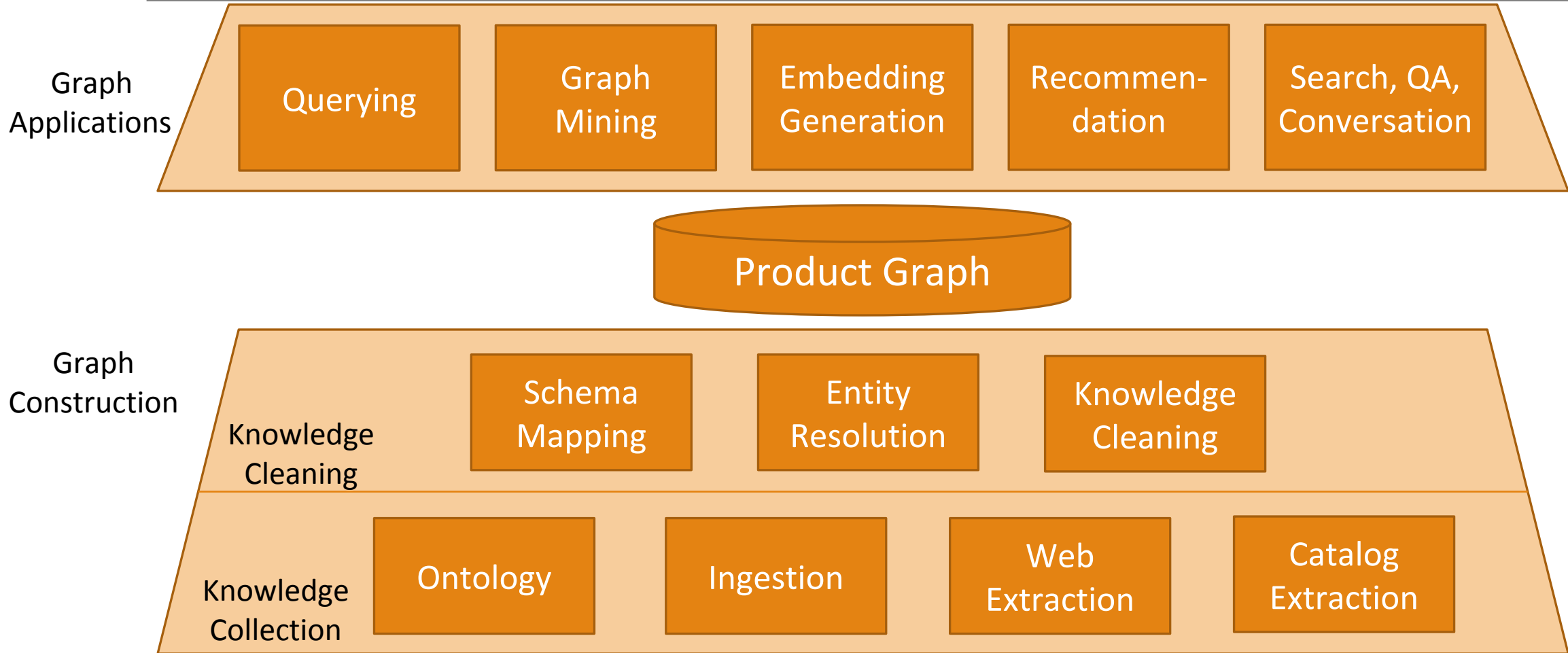
Which ML Model Works Best?

Which ML Model Works Best?

Tree-based models

??

Neural network

Research Philosophy

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

Roofshots: Deliver incrementally and make production impacts

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

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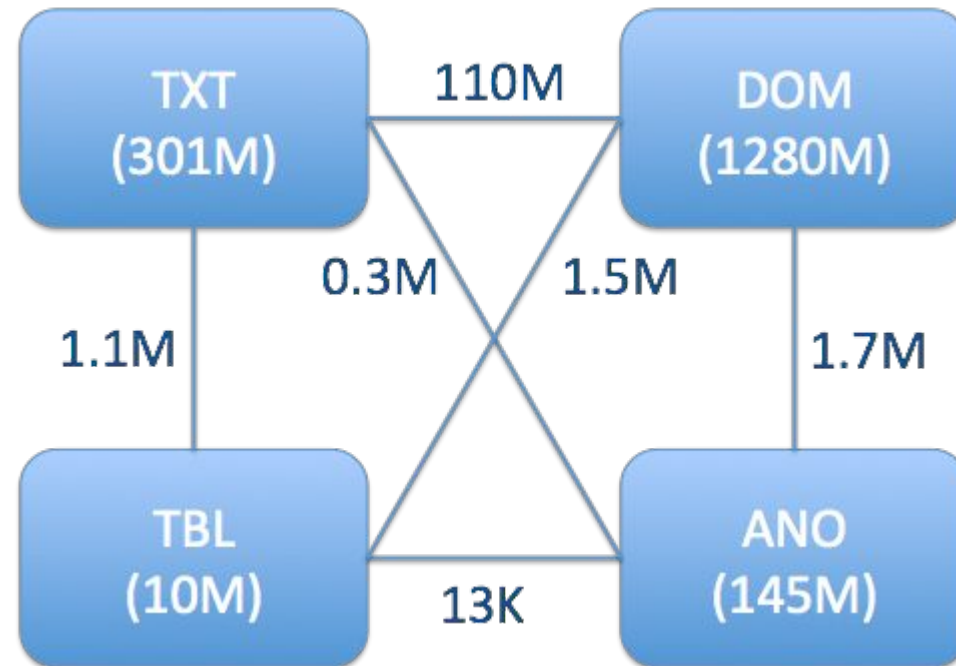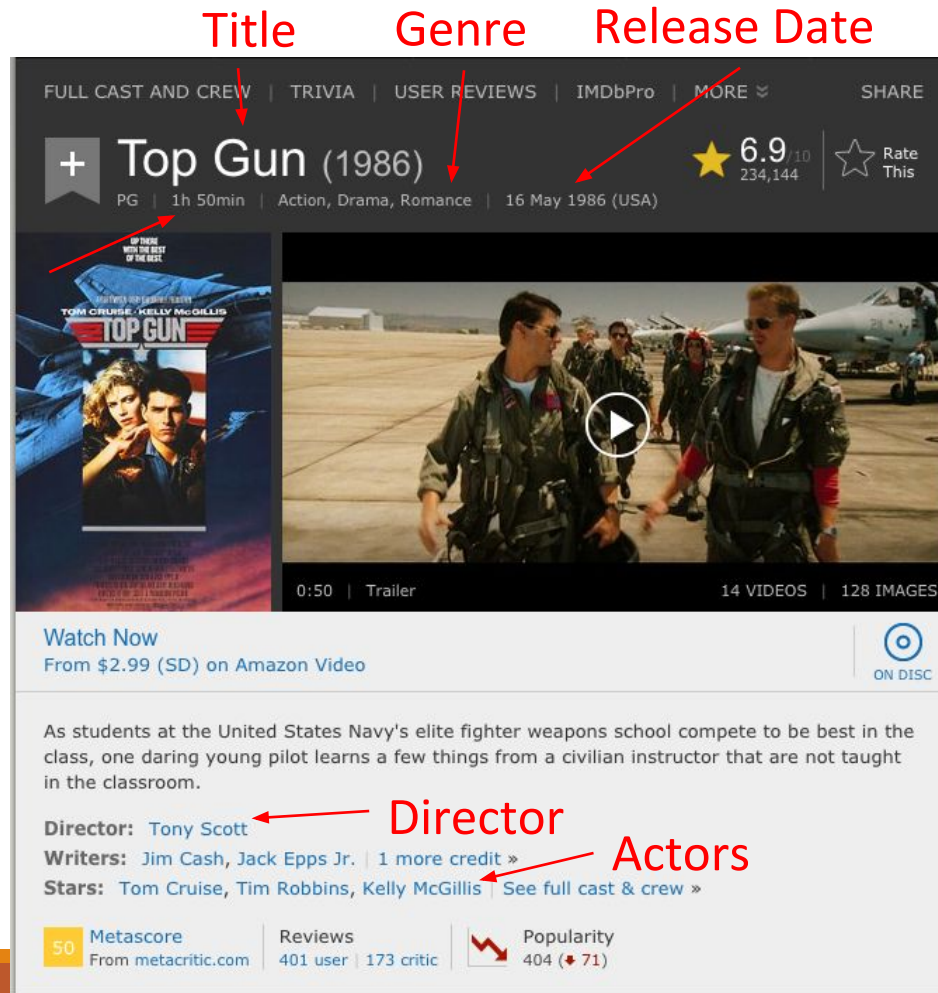
Accu	Accu (conf $\geq$ .7)
0.43	0.63
0.09	0.62

TXT (301M)
DOM (1280M)
TBL (10M)
ANO (145M)

110M
0.3M
1.5M
1.1M
1.7M
13K

I. Extracting Knowledge from Web—Annotation-Based DOM Extraction

Title   Genre   Release Date

Runtime

Annotation-based knowledge extraction

Director

Actors

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")

I. Extracting Knowledge from Web—Annotation-Based DOM Extraction

Annotation-based knowledge extraction

Alexa, When did Padme Amidala die?
What model is R2D2?
Who is Luke Skywalker's master?
Where is Boba Fett from?
Who is Darth Vader's apprentice?

I. Extracting Knowledge from Web—Distantly Supervised DOM Extraction

Distantly supervised web extraction

Annotation-based knowledge extraction

I. Extracting Knowledge from Web—Distantly Supervised DOM Extraction

Entity Identification ➤ Automatic Annotation ➤ Training

Automatic Label Generation
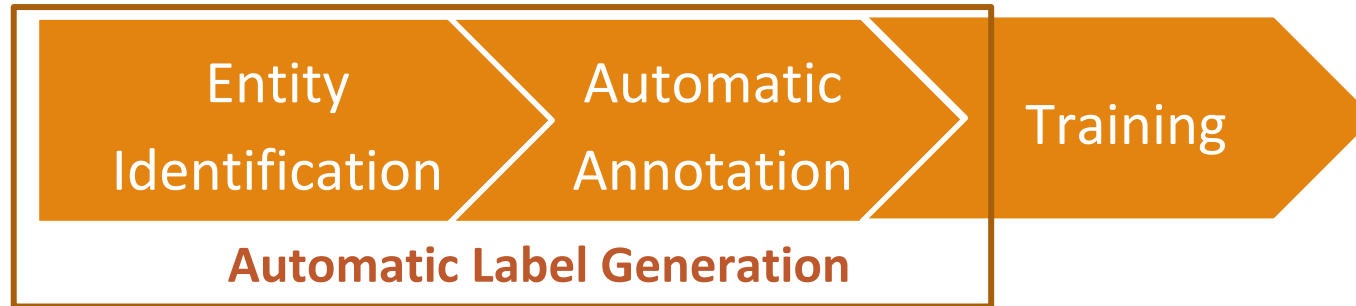
Movie entity

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")

I. Extracting Knowledge from Web—Distantly Supervised DOM Extraction

☐ Extraction on IMDb

1. Very high extraction precision
2. Extracting triples with new entities

Predicate	Precision	Recall
Type.object.name ("name")	1	1
People.person.place_of_birth	1	1
Common.topic.alias	1	1
Film.actor.film	0.98	0.47
Film.director.film	0.98	0.91
Film.producer.film	0.89	0.57
Film.writer.film	0.96	0.60

Predicate	Precision	Recall
Type.object.name ("title")	0.97*	0.97*
Tv.tv_series_episode.episode_number	1	1
Tv.tv_series_episode.season_number	1	1
Film.film.directed_by	0.99	1
Film.film.written_by	1	0.98
Film.film.genre	0.90*	1
Film.film.starring	1	0.97
Tv.tv_series_episode.series	1	1

*Ground truth is incomplete. Manual inspection suggests close to 100% accuracy.

I. Extracting Knowledge from Web—Distantly Supervised DOM Extraction

❑ Extraction experiments on http://swde.codeplex.com/ (2011)

	Title		Director(s)		Genre(s)	
Site	P	R	P	R	P	R
allmovies	1	1	1	1	0.71	0.96
amctv	1	1	0.98	0.97	0.95	0.91
boxofficemojo	1	1	1	0.98	0.67*	0.91
hollywood	1	1	0.94	1	1	0.97
iheartmovies	1	1	1	1	1	1
IMDB	1	1	1	0.98	1	1
metacritic	1	1	1	1	1	1
MSN	1	1	1	1	1	1
rottentomatoes	1	1	1	1	1	0.91
yahoo	1	1	1	0.99	0.99	0.94

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

❑ Logistic regression: Best results (20K features on one website)

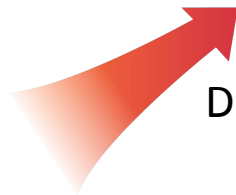❑ Random forest: lower precision and recall
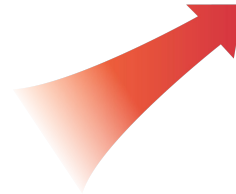
I. Extracting Knowledge from Semi-Structured Data on the Web

Annotation-based knowledge extraction

Distantly supervised web extraction

OpenIE DOM extraction

Nearly-automatic interactive extraction on any new vertical

II. Extracting Knowledge 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		

II. Open Attribute Extraction by Named Entity Recognition

B Beginning of entity

I Inside of entity
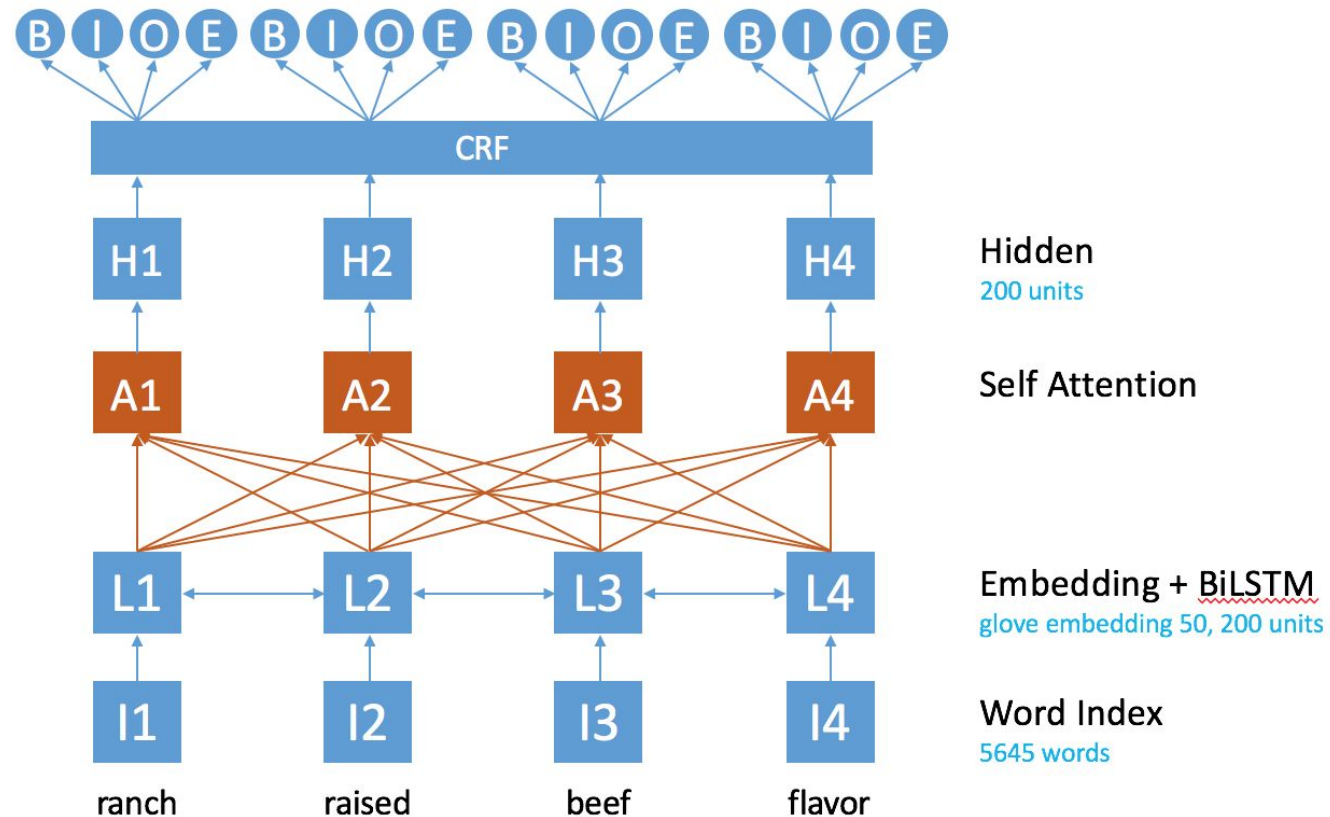
O Outside of entity

E End of entity

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

$y=\{t_1,t_2,...,t_n\}$ tagging decision
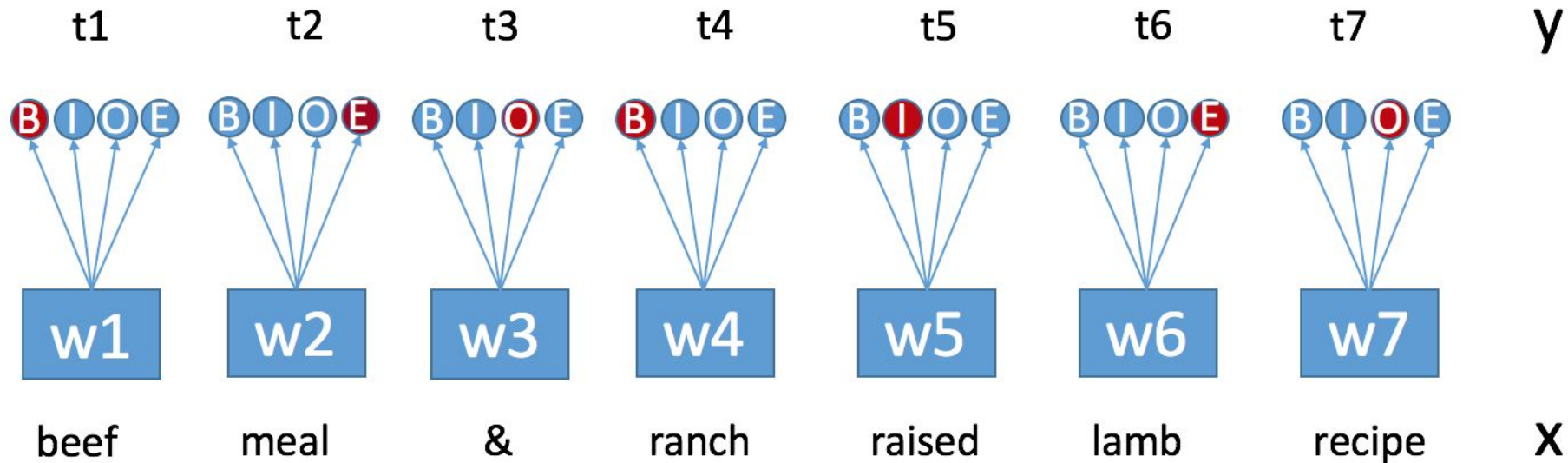
II. Open Attribute Extraction by NER —Which ML Model Works Best?

❑Recurrent Neural Network, CRF, Attention

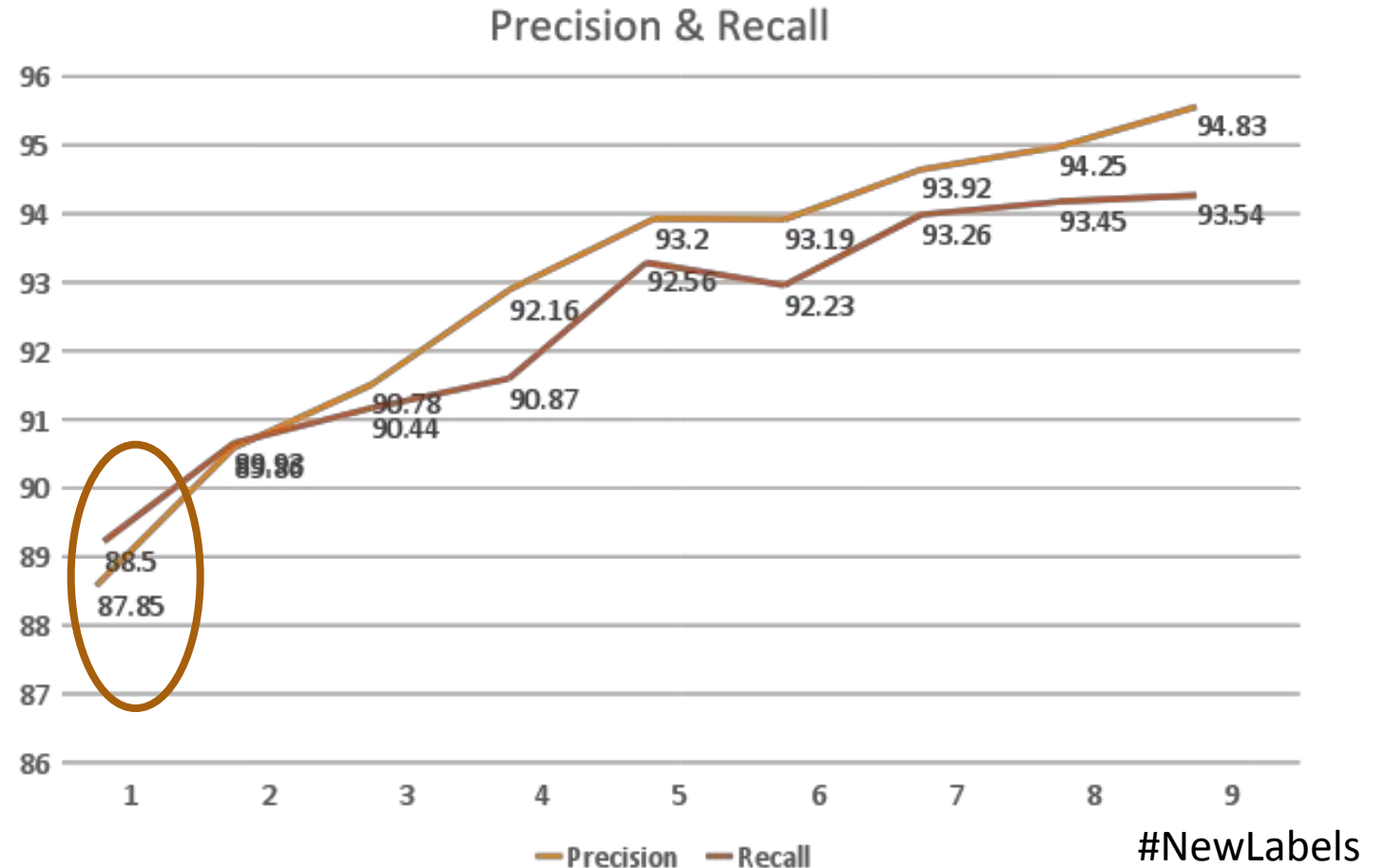# II. Open Attribute Extraction by NER —Adding Active Learning
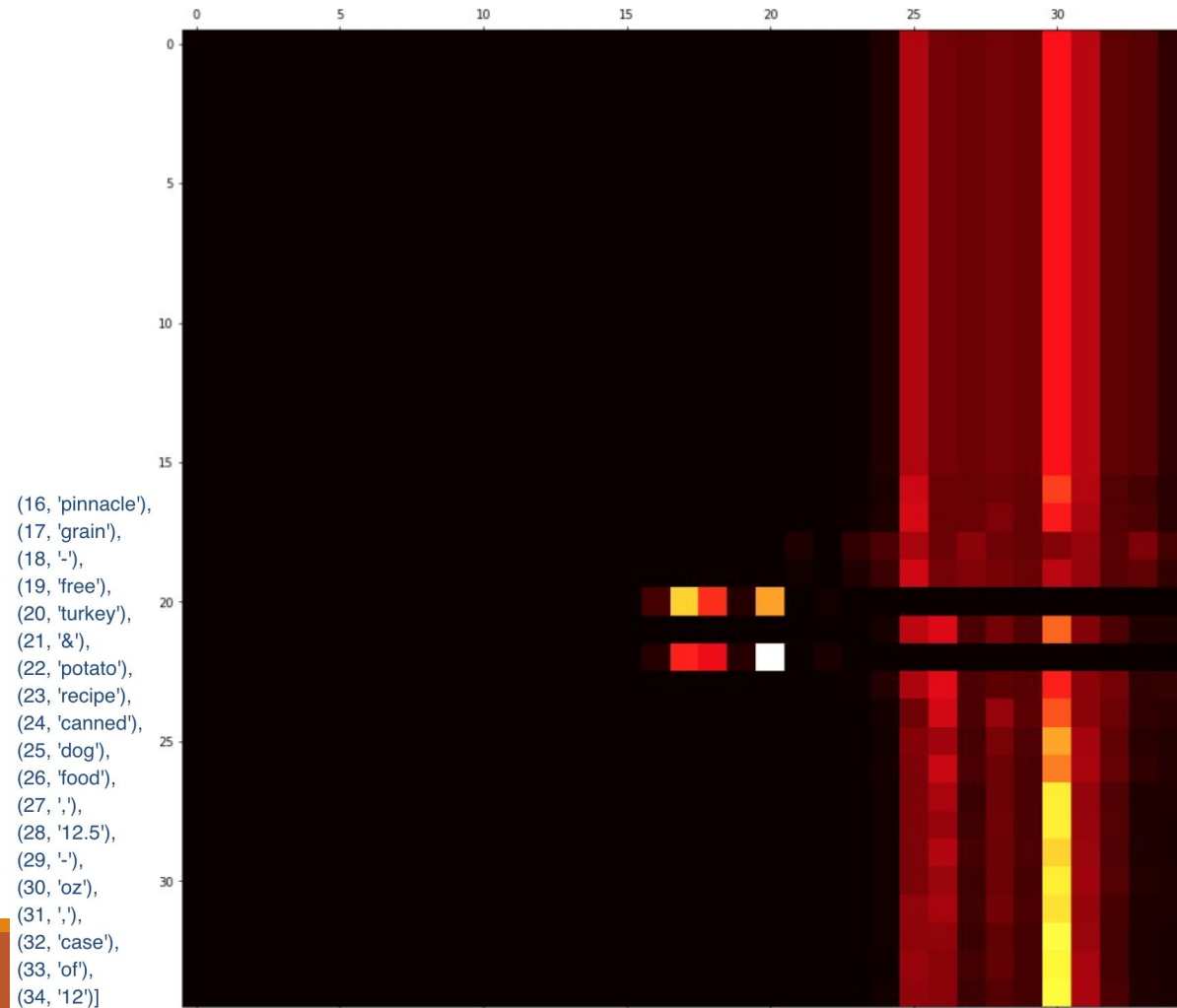

Training
500 Sentences
7927 Words
944 Flavors

Testing
600 Sentences
7896 Words
786 Flavors

Different flavors from Training data

Precision & Recall

Precision   Recall

#NewLabels

II. Open Attribute Extraction by NER —Attentions Help Find Contexts

(16, 'pinnacle'),
(17, 'grain'),
(18, '-'),
(19, 'free'),
(20, 'turkey'),
(21, '&'),
(22, 'potato'),
(23, 'recipe'),
(24, 'canned'),
(25, 'dog'),
(26, 'food'),
(27, ','),
(28, '12.5'),
(29, '-'),
(30, 'oz'),
(31, ','),
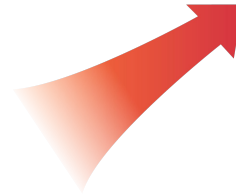(32, 'case'),
(33, 'of'),
(34, '12')]

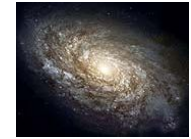# II. Extracting Knowledge from Product Profiles in Amazon Catalog

Product profile
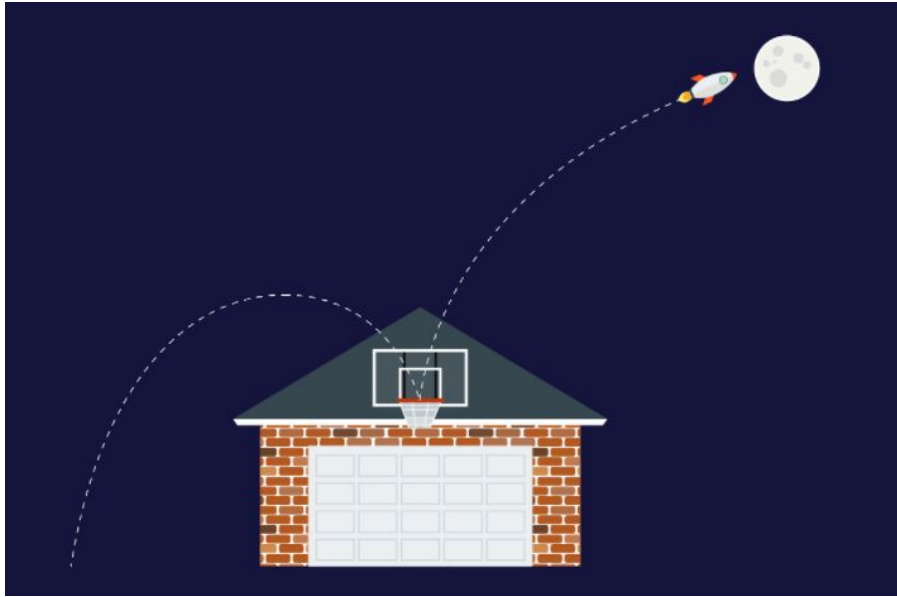extraction
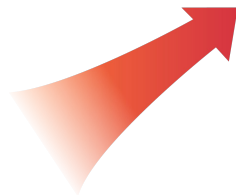
Automatically
building a
shallow KG

Open aspect
extraction

Review extraction
& sentiment analysis

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!