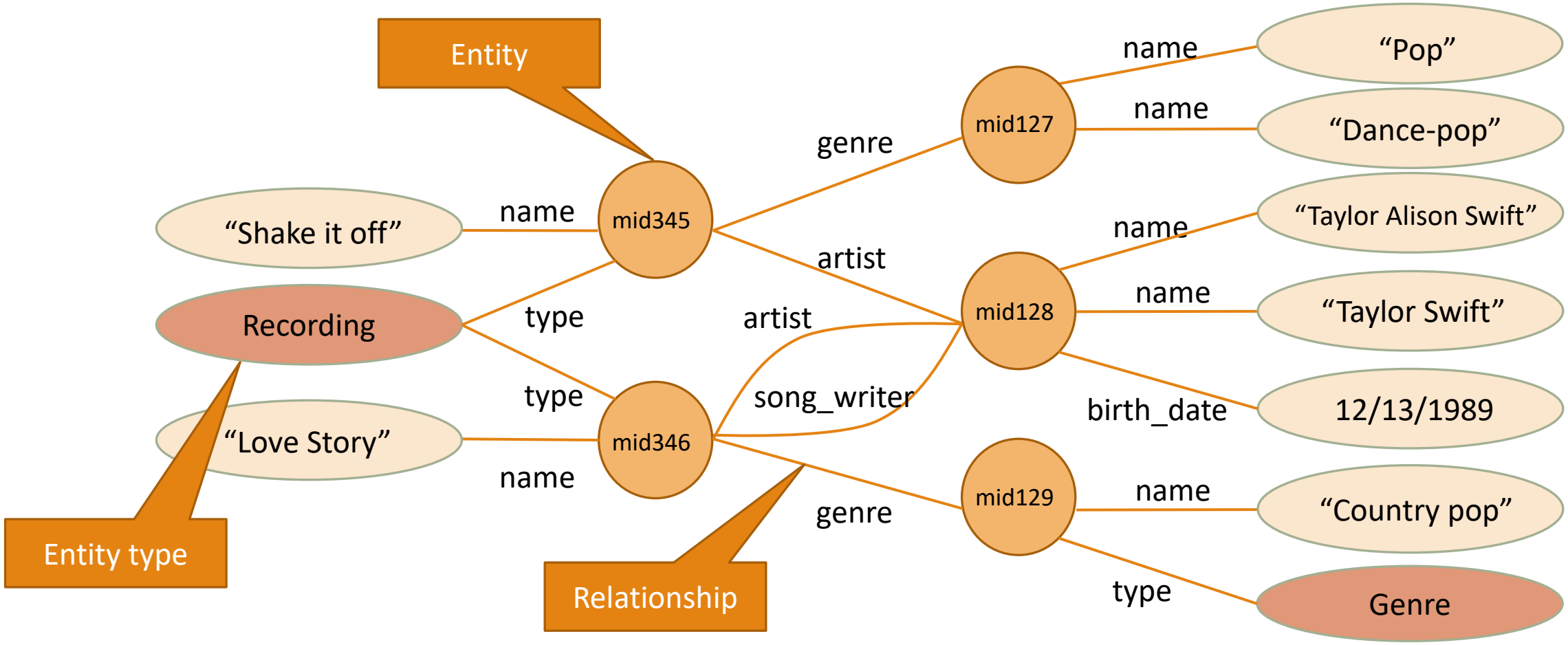
Zero to One Billion: The Path for a Rich Product Graph

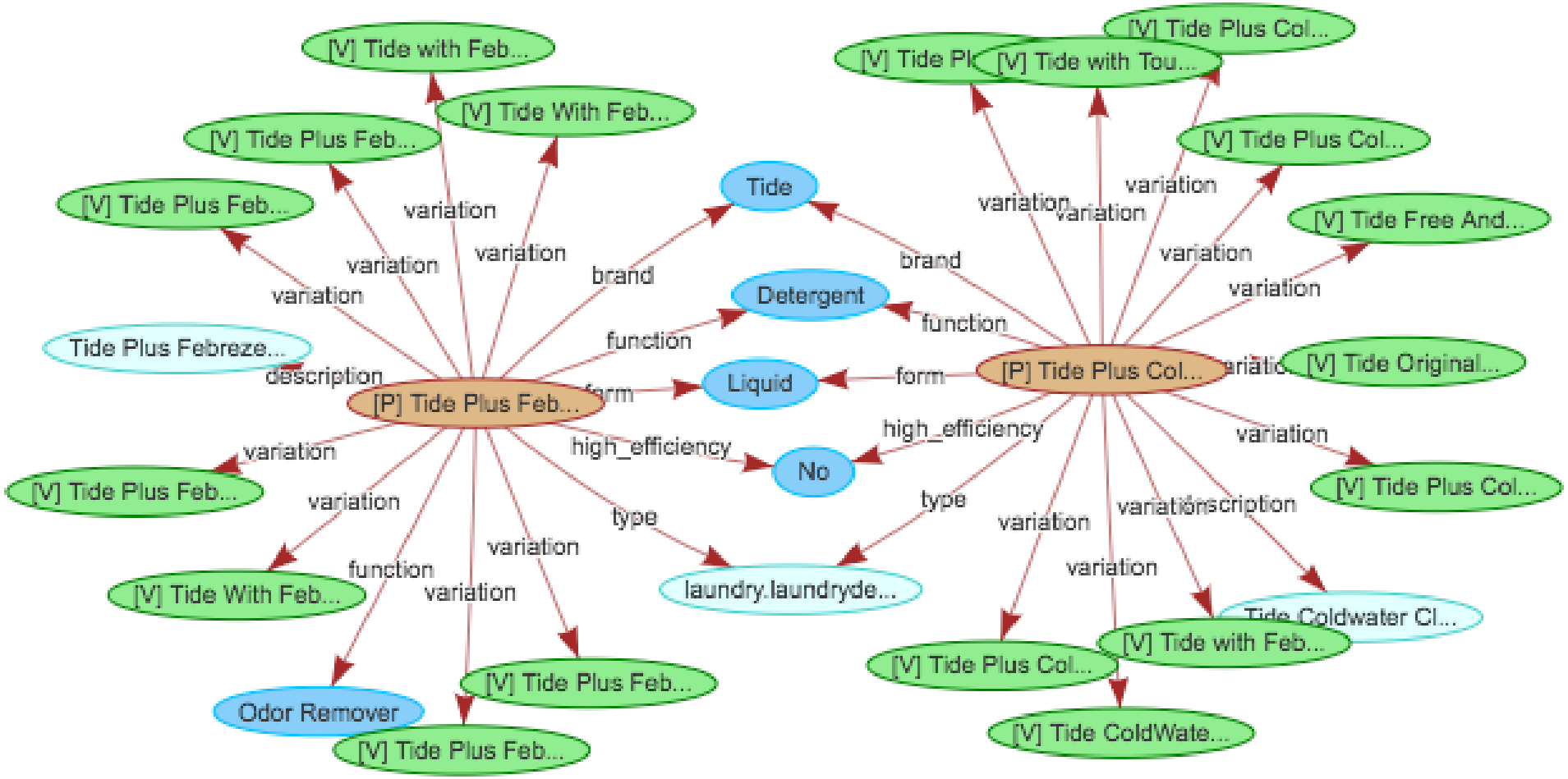
XIN LUNA DONG, AMAZON

5/2021

Knowledge Graph Example for 2 Songs

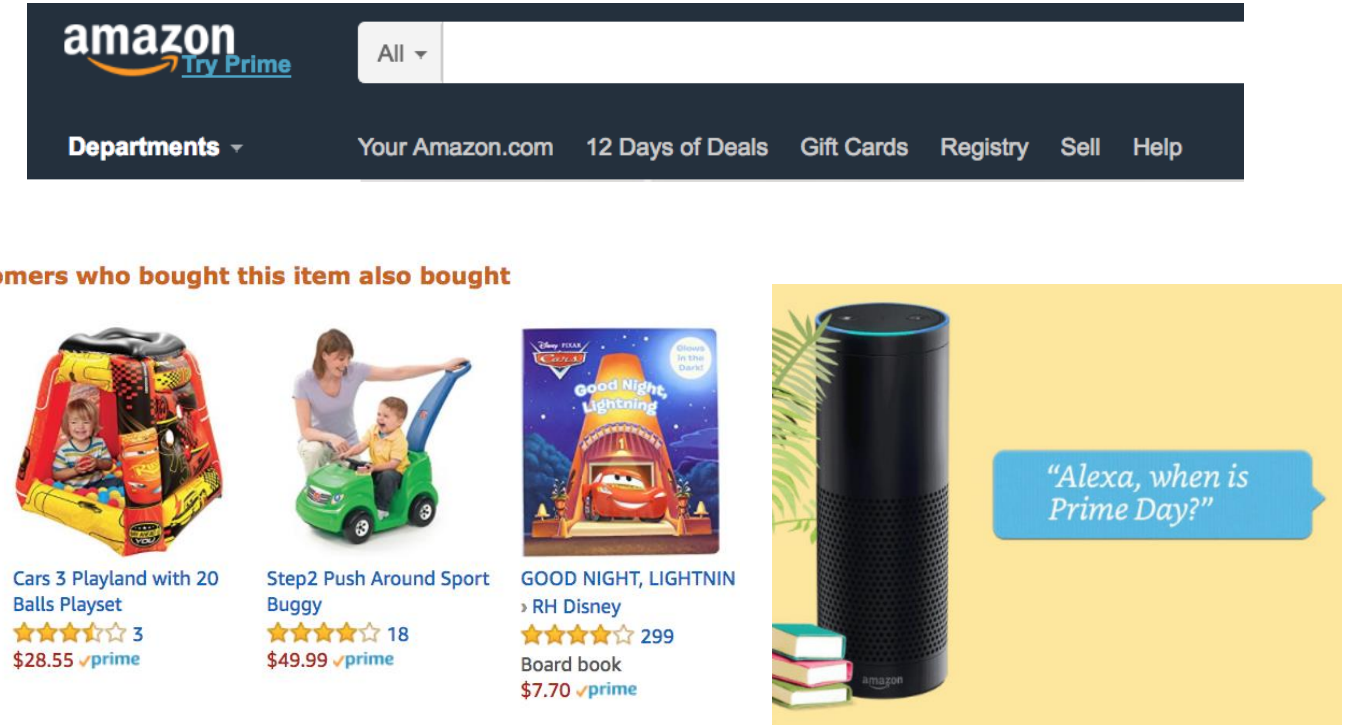


Product Graph Example for 2 Products



Product Graph

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



Use Case I: Providing Information

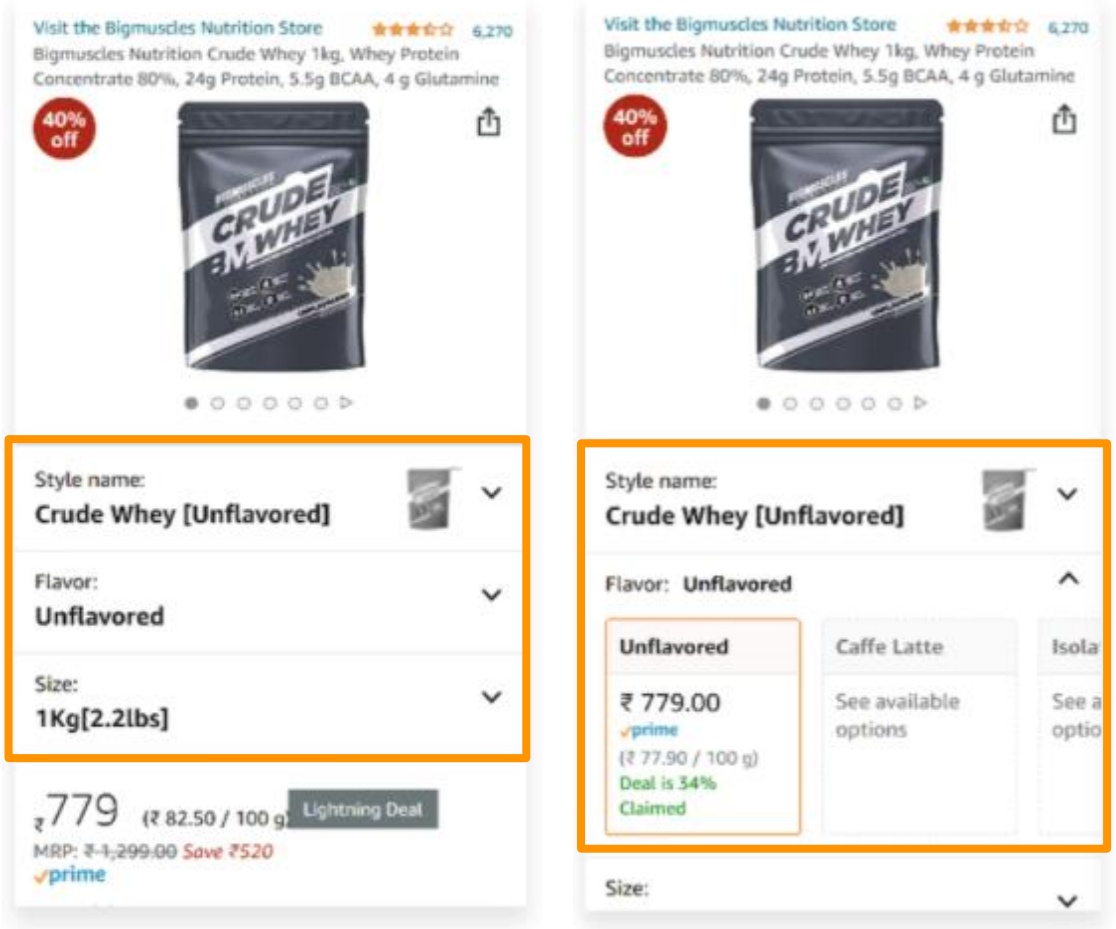
Roll over image to zoom in

Brand	Cetaphil
Ingredients	Water, Cetyl Alcohol, Propylene Glycol, Iodopropynyl Butylcarbamate, 2-Bromo-2-Nitropropane-1, 3-Diol, Sodium Lauryl Sulfate, Stearyl Alcohol, Methylparaben, Propylparaben, Sodium Citrate, Butylparaben, Allantoin, Zinc Gluconate.
Scent	Fragrance free
Additional Item Information	Non-Comedogenic, Fragrance-free, Natural
Skin Type	Sensitive

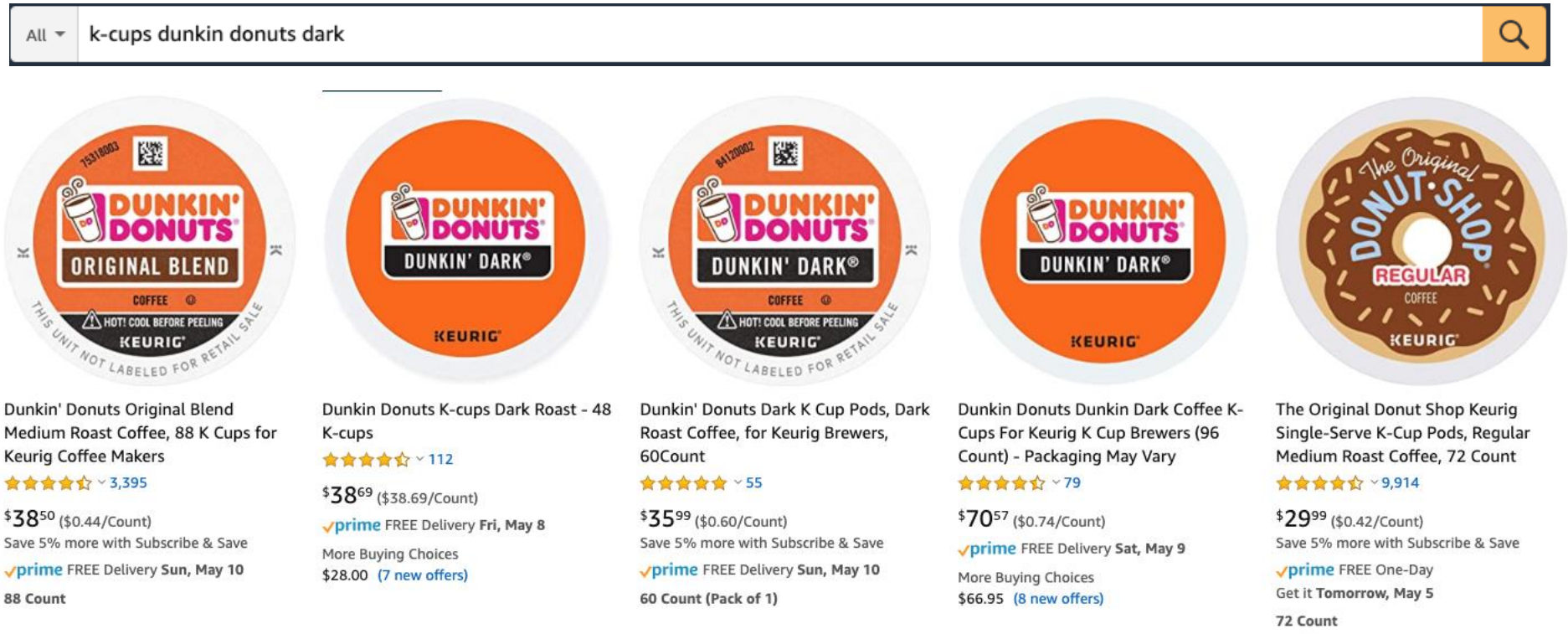
About this item

- Gentle for everyday use; Cetaphil gentle skin cleansing cloths will leave your skin feeling clean, refreshed and balanced after every use
- Removes makeup & dirt: Thoroughly remove makeup and dirt, leaving skin clean
- Mild & non irritating: Soap free formulation won't strip skin of its natural protective oils and emollients

Use Case II: Providing Choices



Use Case III: Improving Search



Dunkin' Donuts Original Blend Medium Roast Coffee, 88 K Cups for Keurig Coffee Makers

3,395

$38.50 ($0.44/Count)

Save 5% more with Subscribe & Save

prime FREE Delivery **Sun, May 10**

88 Count

Dunkin Donuts K-cups Dark Roast - 48 K-cups

112

$38.69 ($38.69/Count)

prime FREE Delivery **Fri, May 8**

More Buying Choices

$28.00 (7 new offers)

Dunkin' Donuts Dark K Cup Pods, Dark Roast Coffee, for Keurig Brewers, 60Count

55

$35.99 ($0.60/Count)

Save 5% more with Subscribe & Save

prime FREE Delivery **Sun, May 10**

60 Count (Pack of 1)

Dunkin Donuts Dunkin Dark Coffee K-Cups For Keurig K Cup Brewers (96 Count) - Packaging May Vary

79

$70.57 ($0.74/Count)

prime FREE Delivery **Sat, May 9**

More Buying Choices

$66.95 (8 new offers)

The Original Donut Shop Keurig Single-Serve K-Cup Pods, Regular Medium Roast Coffee, 72 Count

9,914

$29.99 ($0.42/Count)

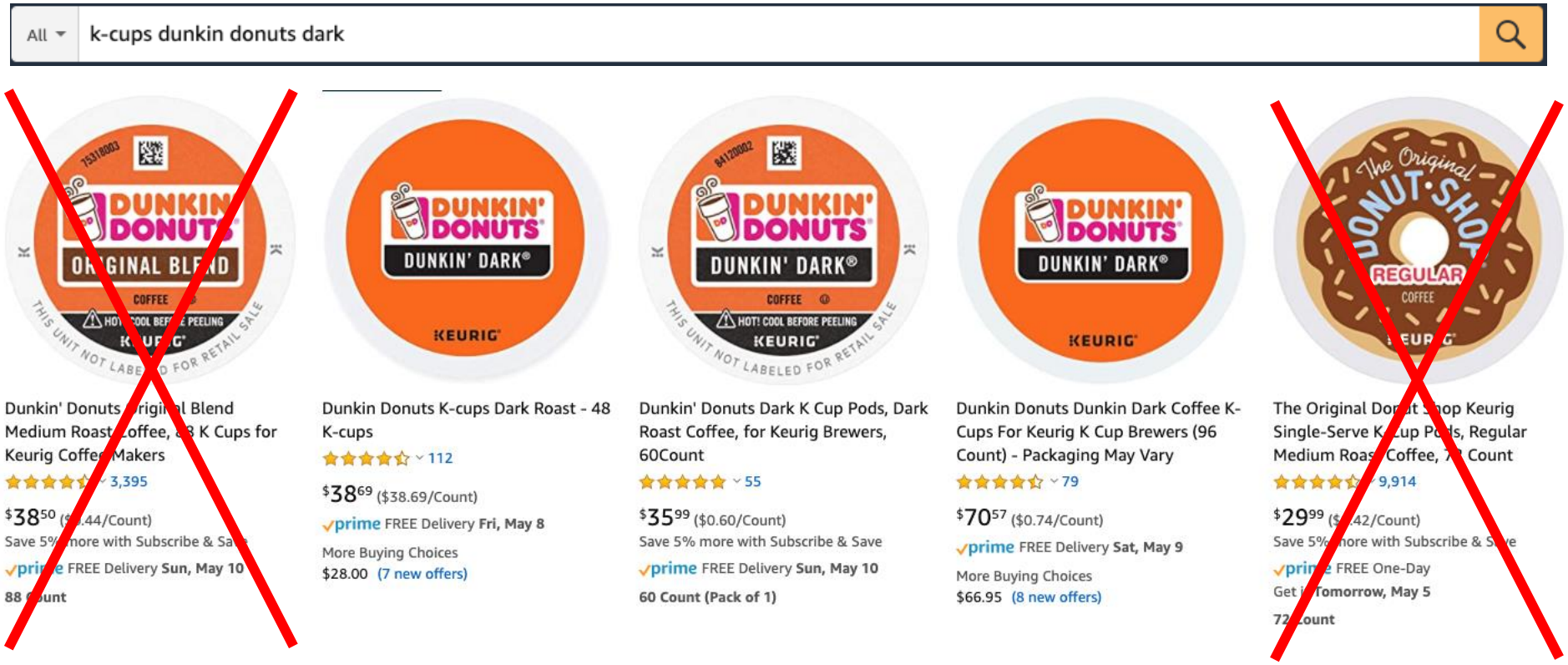
Save 5% more with Subscribe & Save

prime FREE One-Day

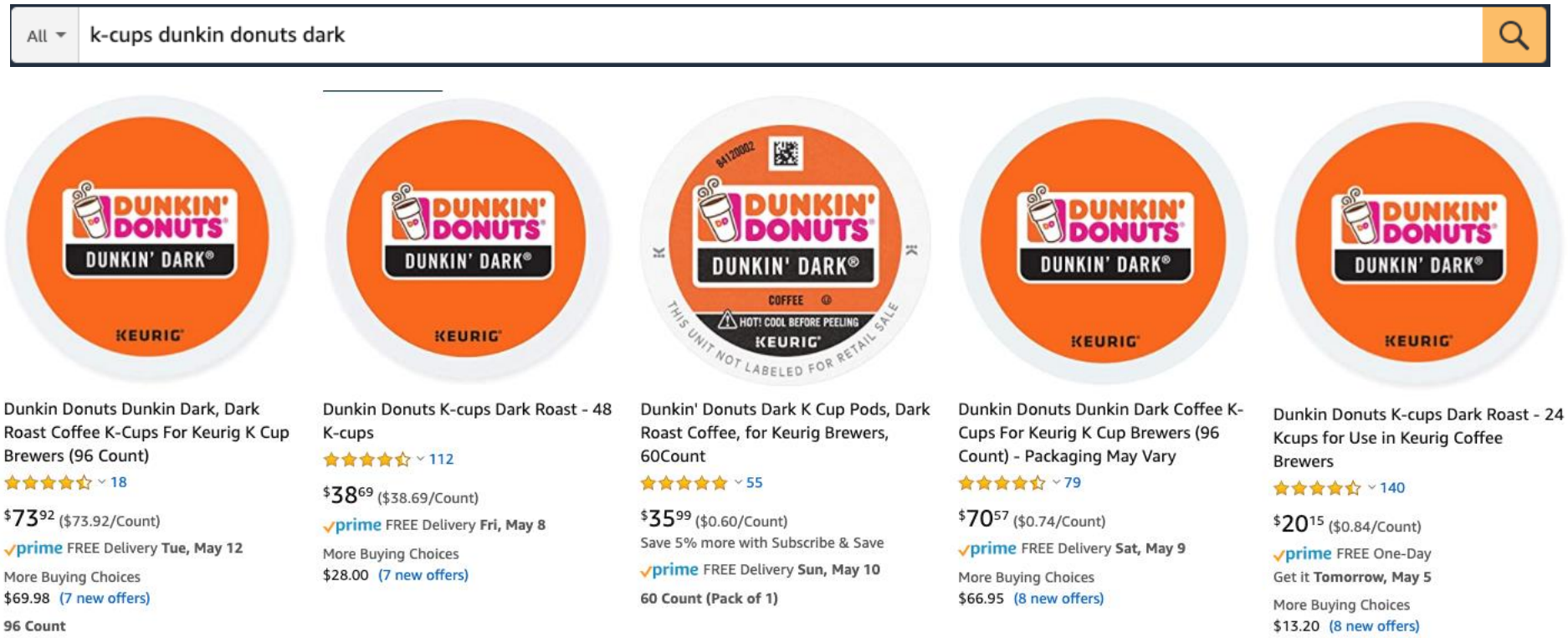
Get it **Tomorrow, May 5**

72 Count

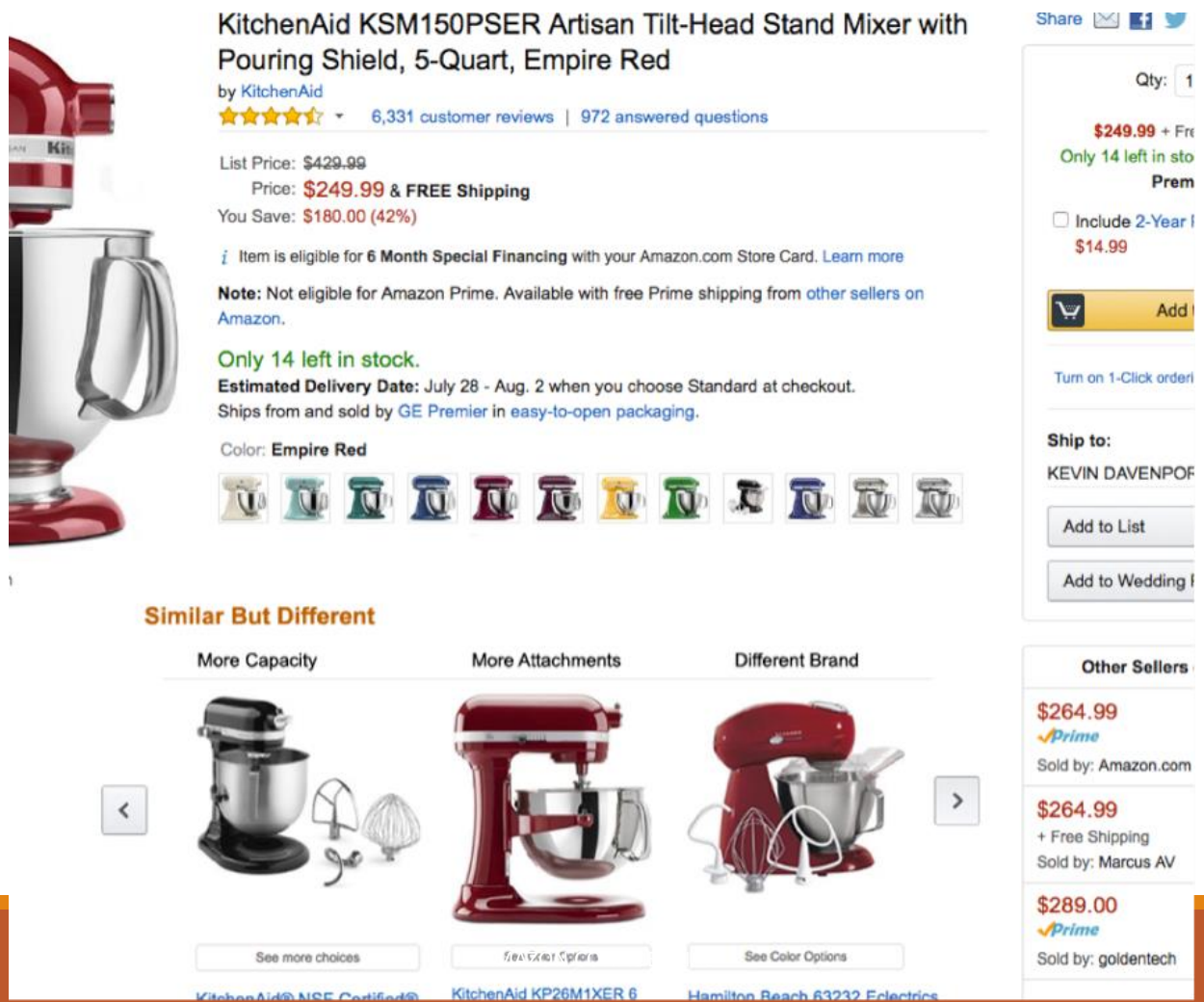
Use Case III: Improving Search



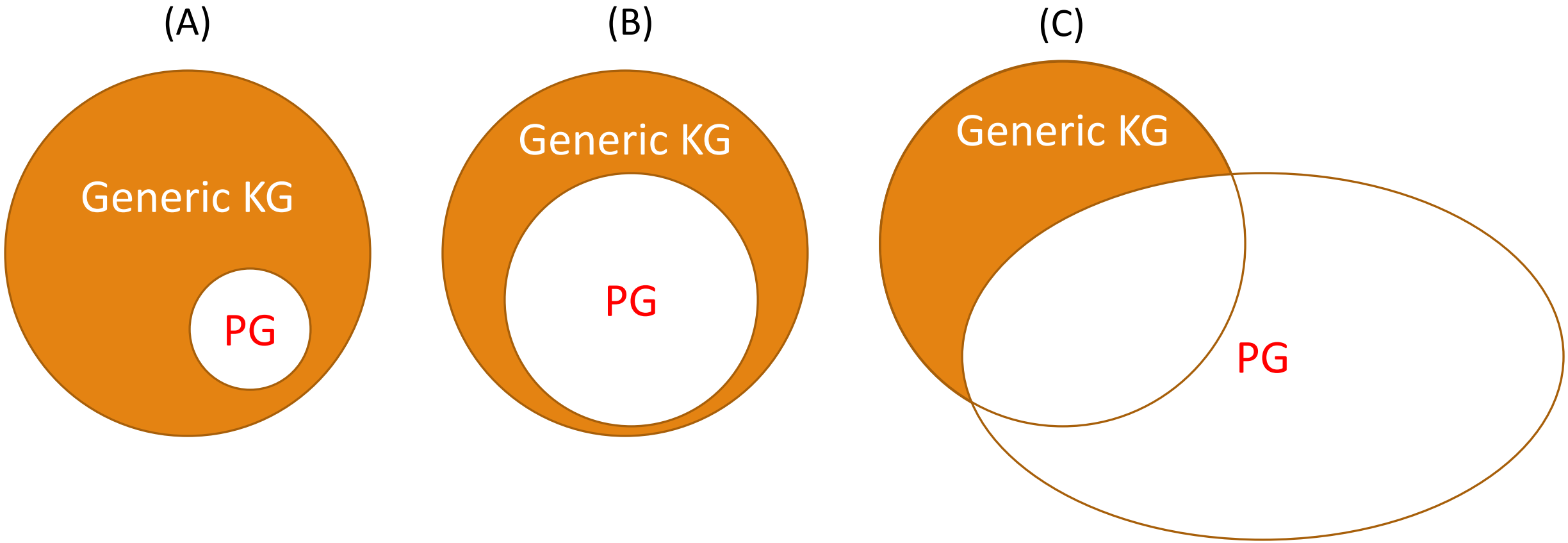
Use Case III: Improving Search



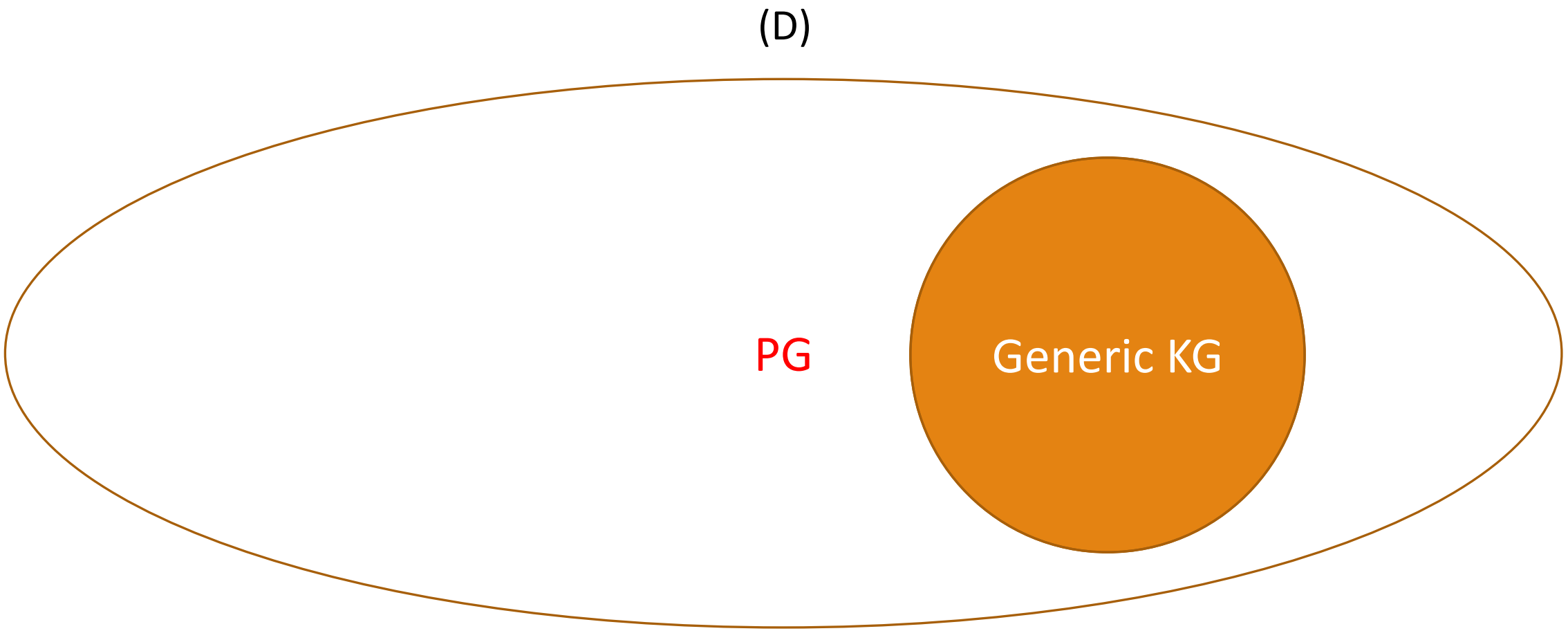
Use Case IV: Improving Recommendation



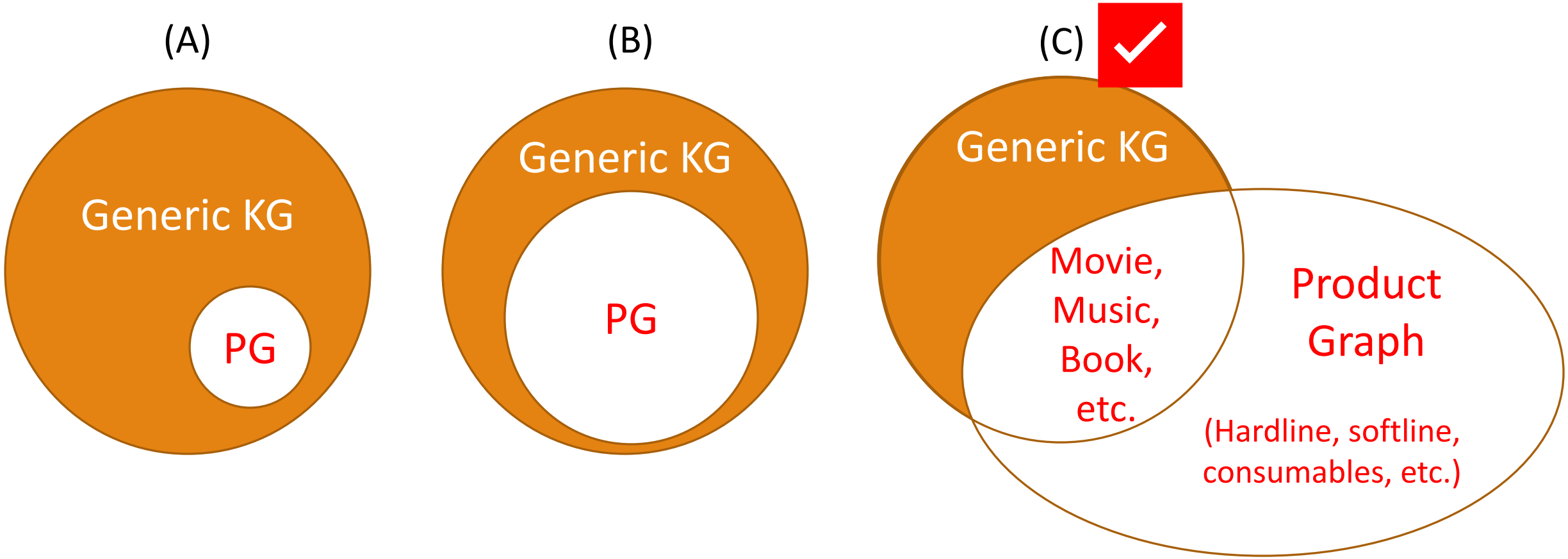
Product Graph vs. Knowledge Graph



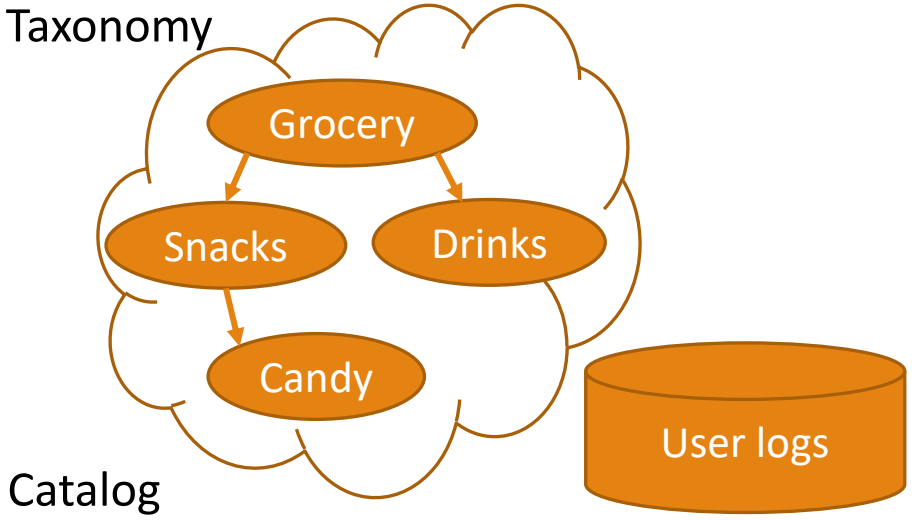
Product Graph vs. Knowledge Graph



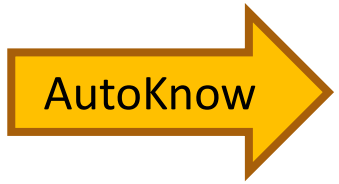
Knowledge Graph vs. Product Graph

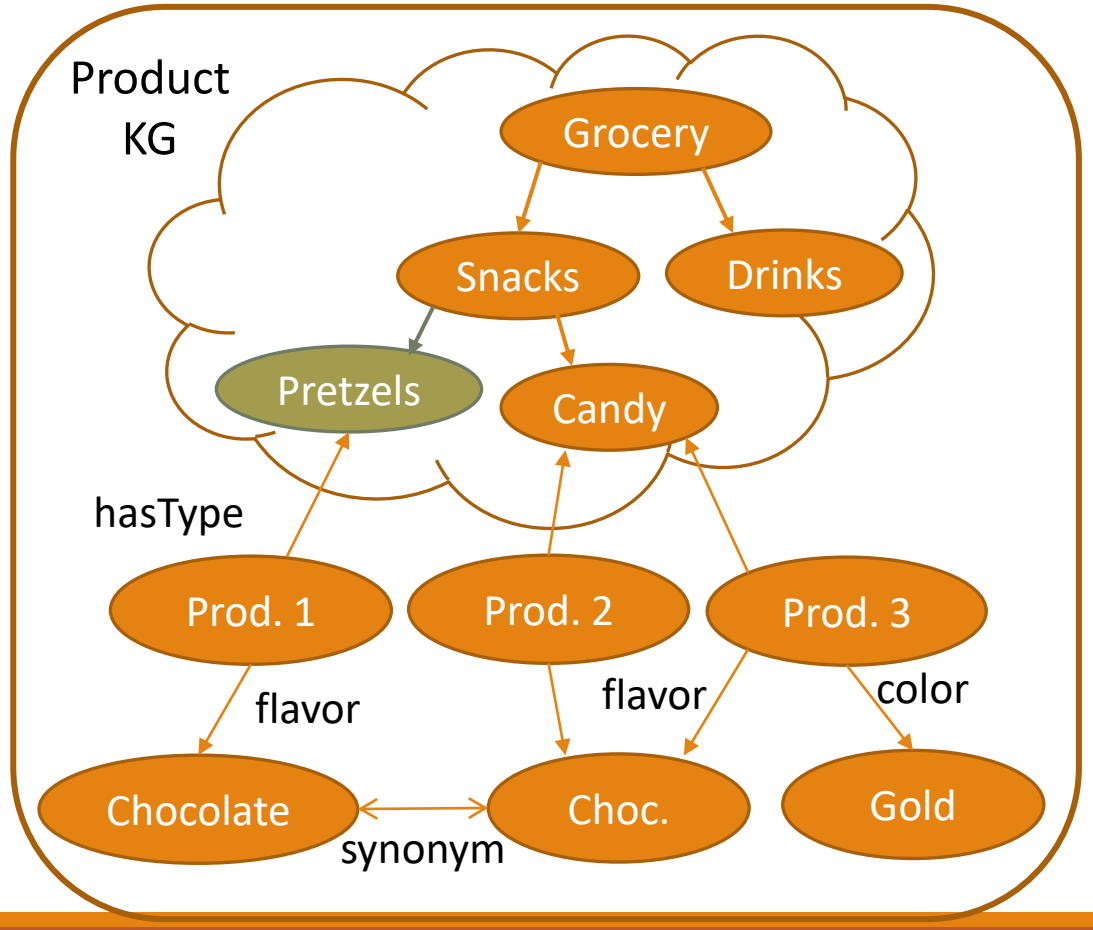


AutoKnow: Self-Driving Product Knowledge Collection



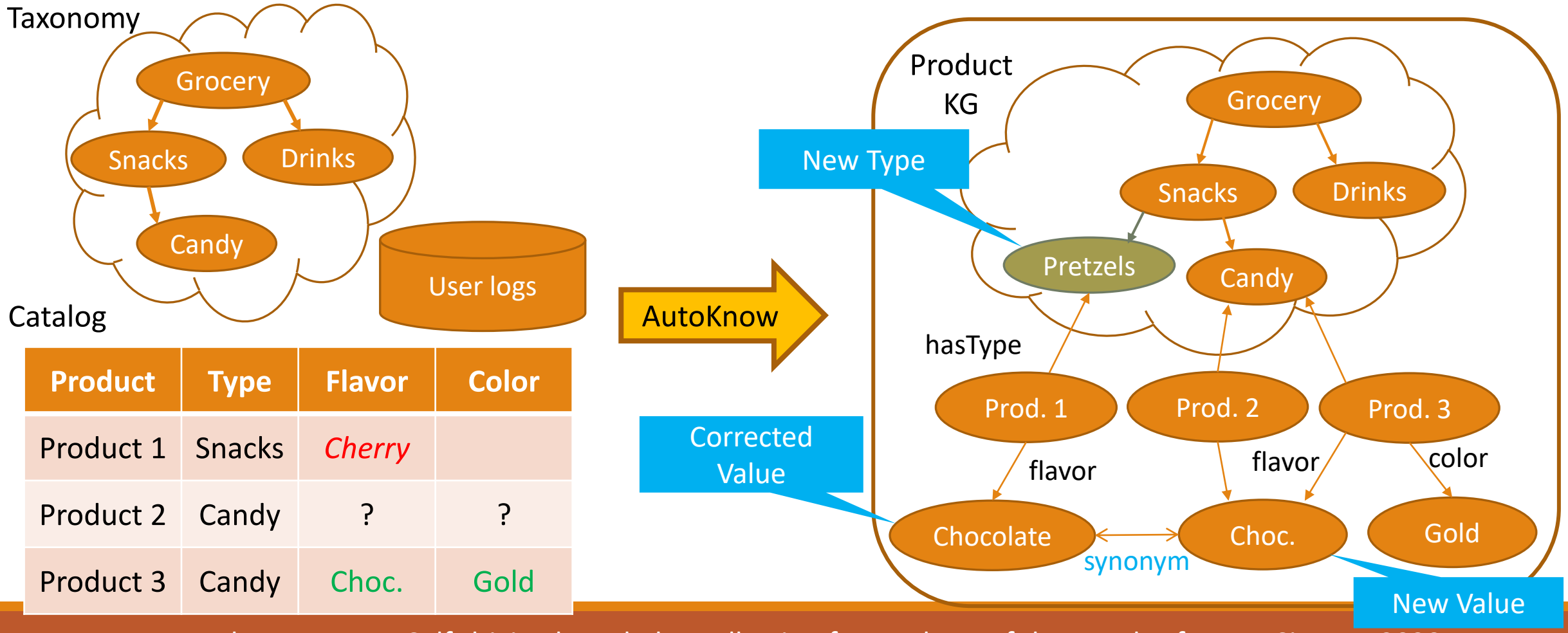
Product	Type	Flavor	Color
Product 1	Snacks	*Cherry*	
Product 2	Candy	?	?
Product 3	Candy	Choc.	Gold





Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.

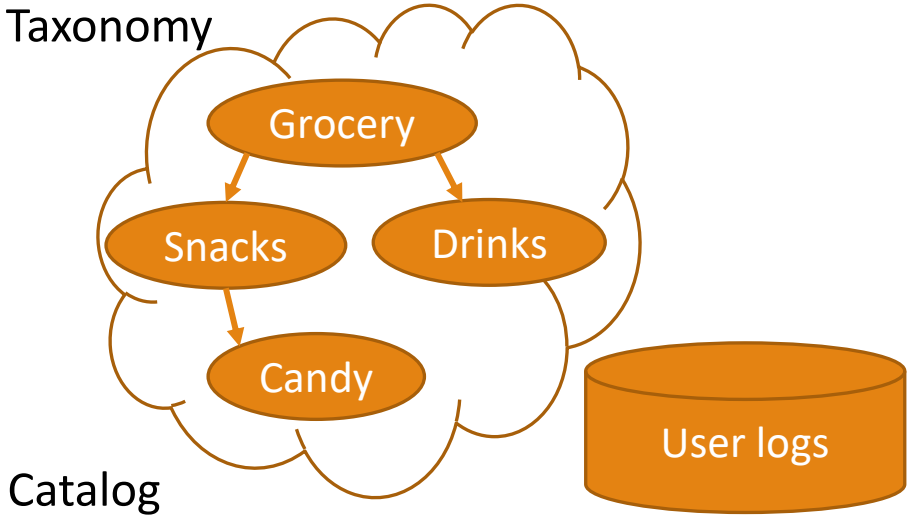
AutoKnow: Self-Driving Product Knowledge Collection



Product	Type	Flavor	Color
Product 1	Snacks	*Cherry*	
Product 2	Candy	?	?
Product 3	Candy	Choc.	Gold

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.

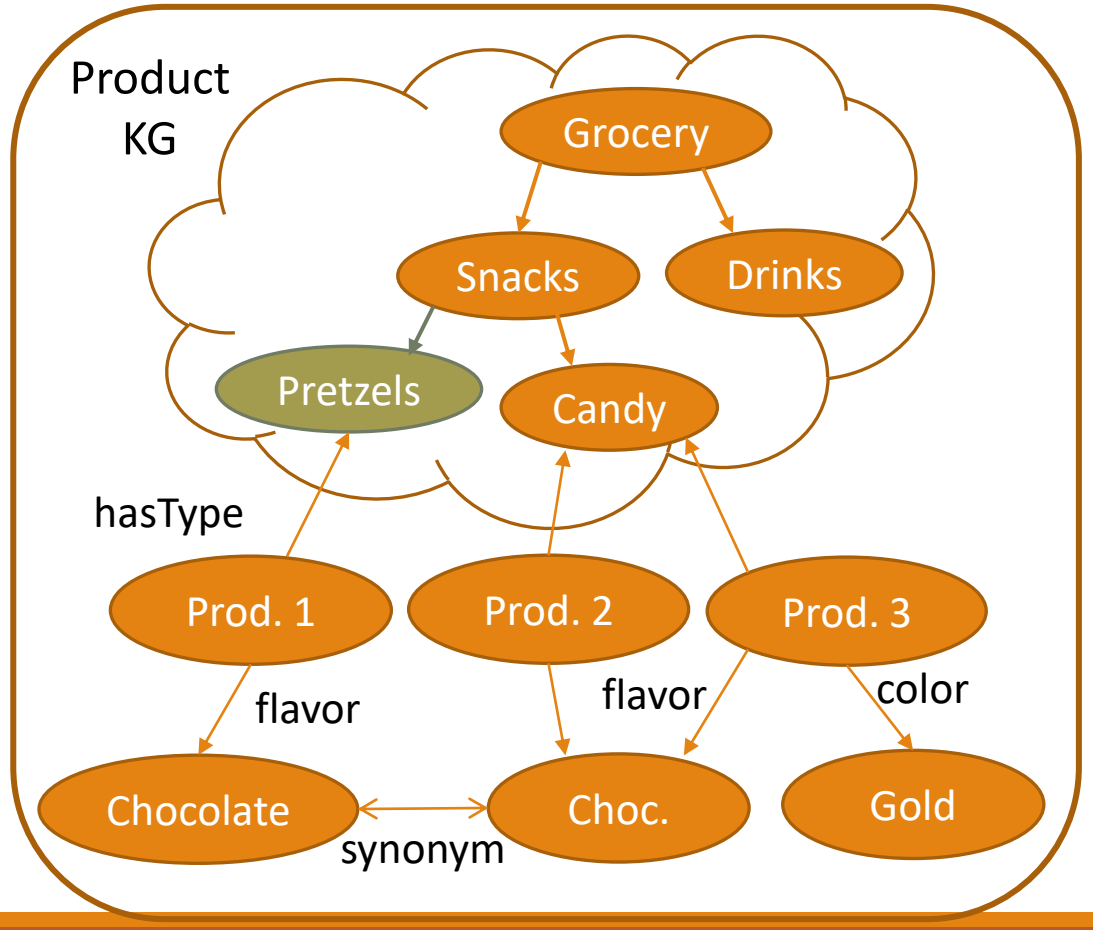
AutoKnow: Self-Driving Product Knowledge Collection



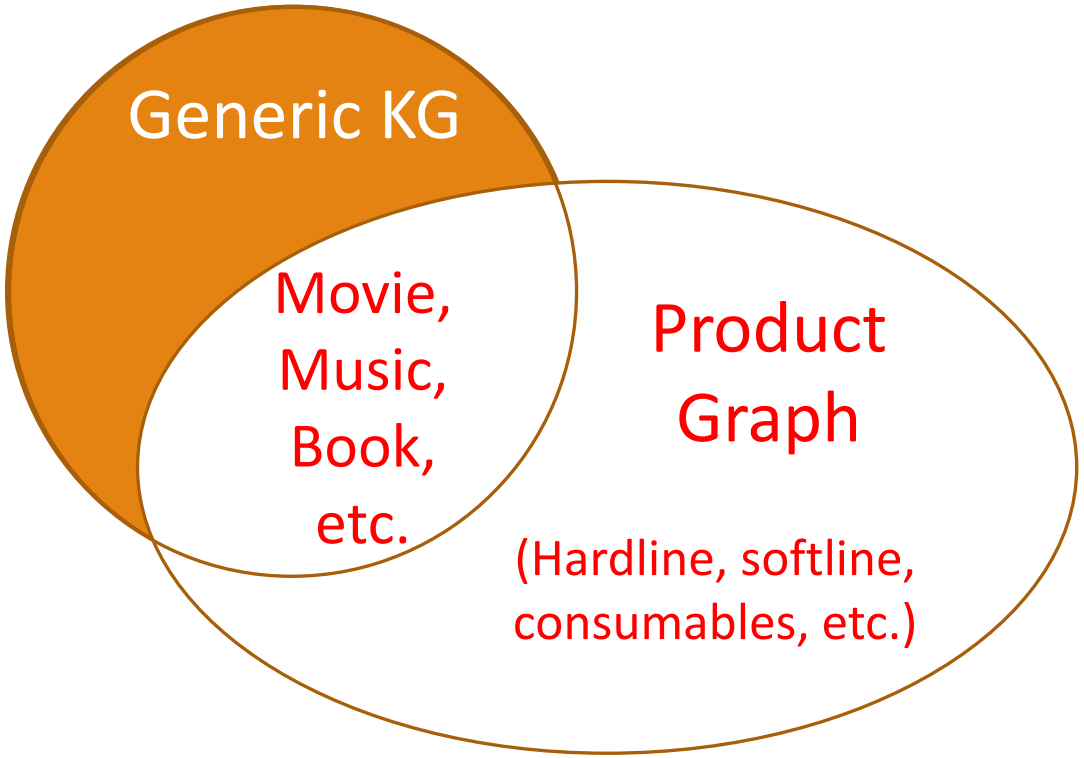
Product	Type	Flavor	Color
Product 1	Snacks	*Cherry*	
Product 2	Candy	?	?
Product 3	Candy	Choc.	Gold

AutoKnow

- **#Types ↑ 3X**
- **Defect rate ↓ up to 68 percent points**



Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.



But, Is The Problem Harder?

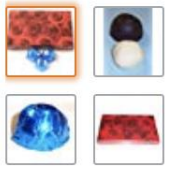
Challenges in Building Product Graph I

- Sparse and noisy structured data

Scott's Cakes Dark Chocolate Toffee Cream Filling Candies with Dark Blue Foils in a 1 Pound Red Roses Box

by Scott's Cakes

Be the first to review this item



Price: $19.95 + $14.95 shipping

You can get 5% back on all Amazon.com purchases with the Amazon Prime Store Card. No annual fee.

Note: Not eligible for Amazon Prime.

In Stock. Ships from and sold by Scott's Cakes.

38 Flavors: 1 lb. Red Rose Box

1 lb. Christma…	1 lb. Daisy Box	1 lb. Red Rose Box
$19.95	$19.95	$19.95

19 Sizes: Dark Blue Foils

Aqua Foils	Burgandy Foils	Dark Blue Foils
$19.95	$19.95	$19.95

Get it Thu, Aug 22 - Tue, Aug 27

Get it Tue, Aug 20 - Fri, Aug 23 if you choose paid shipping at checkout.

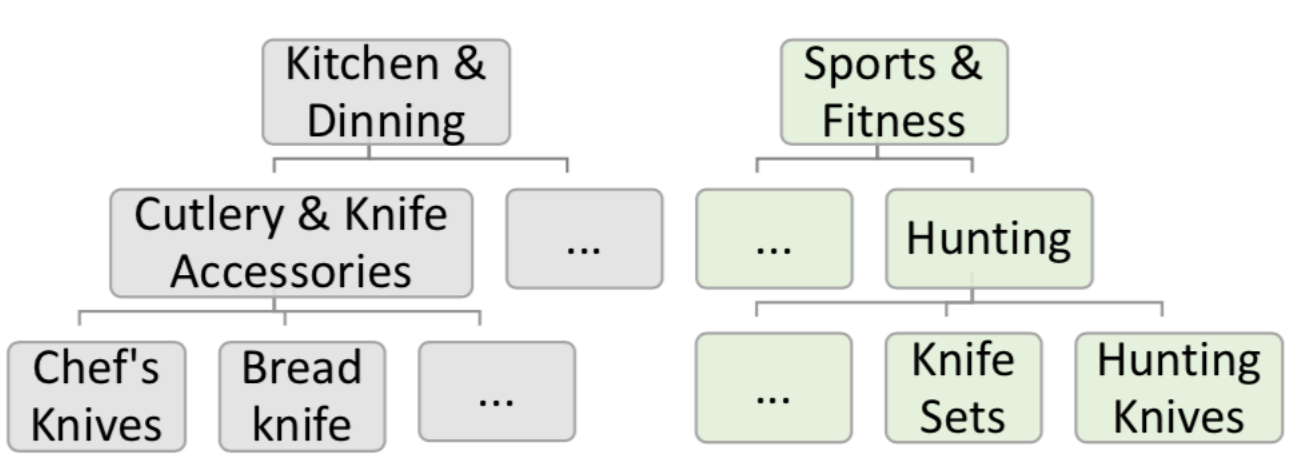
Deliver to Yaqing - Seattle 98109

Qty: 1

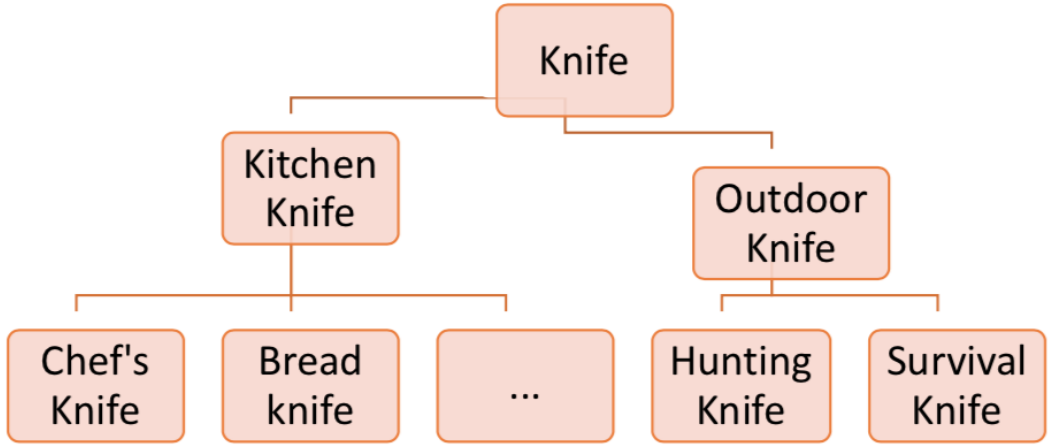
Turn on 1-click ordering

Challenges in Building Product Graph II

- Extremely complex domains
  - How to identify the millions of product types?
  - How to organize types into a taxonomy tree?



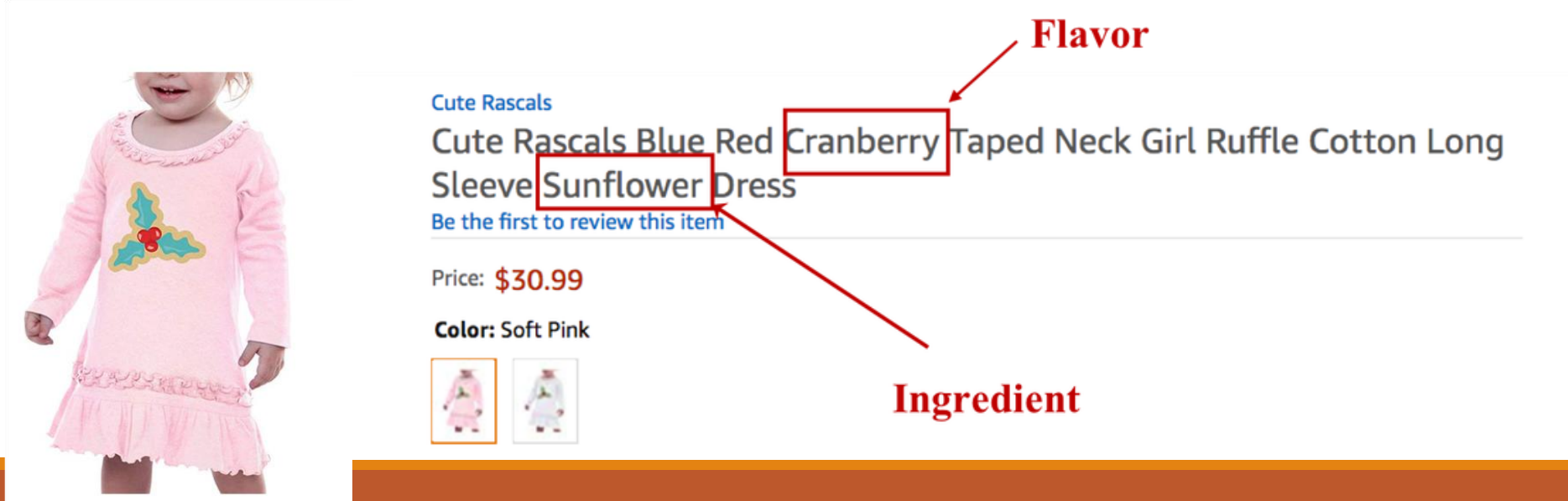
Sellers' view



Buyers' view

Challenges in Building Product Graph III

- Big variety across product types
  - Different attributes apply to different product types
  - Different value vocabularies and different patterns



Challenges in Building Product Graph III

- Big variety across product types
  - Different attributes apply to different product types
  - Different value vocabularies and different patterns

Samsung UN58RU7100FXZA Flat 58-Inch 4K UHD 7 Series Ultra HD Smart TV with HDR and Alexa Compatibility (2019 Model)

Taylors of Harrogate Classic Tea Variety Box, 48 Count (Pack of 1)

Caribou Coffee Caribou Blend, Medium Roast Ground Coffee, 20 Ounce Bag, Rainforest Alliance Certified

Scale Up in 3 Dimensions

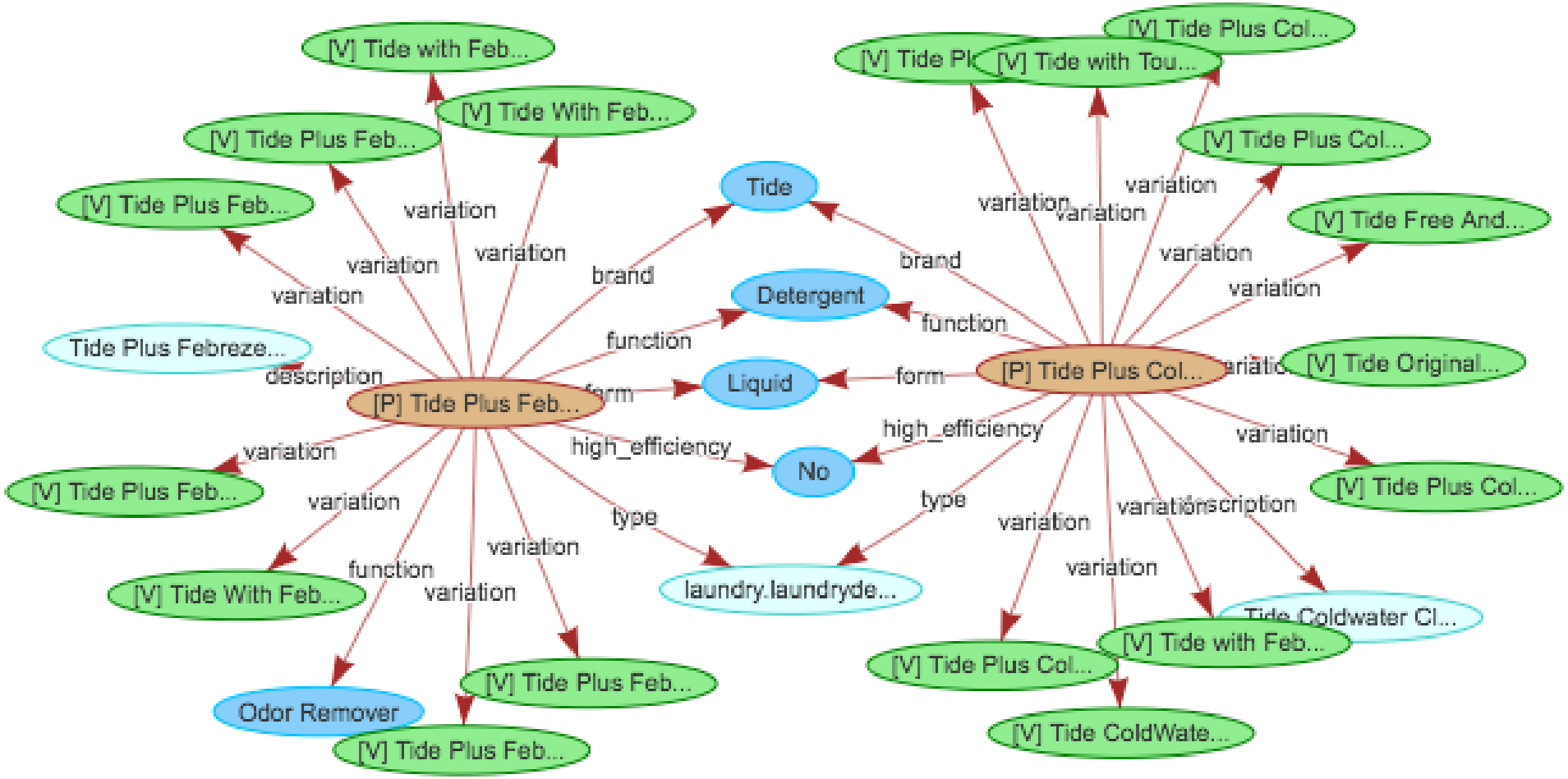
Millions of categories

Thousands of attributes

Hundreds of languages

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

A 100-Year Project



Deliver the Data Business

1,000,000,000

Deliver the Data Business

High precision models

Deliver the Data Business

1,000

High precision models

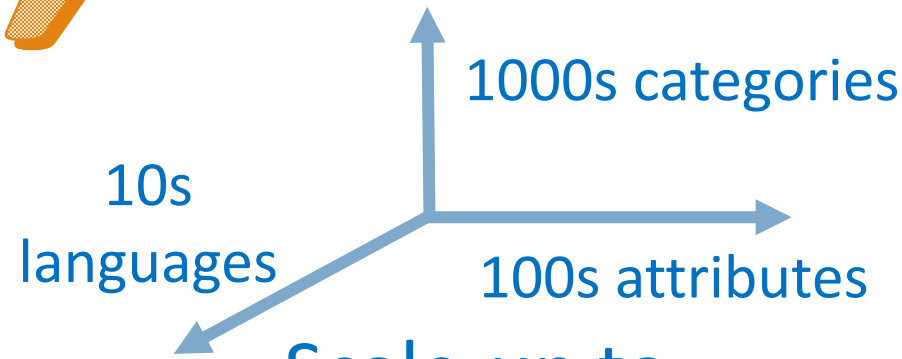
E2E pipeline + AutoML to reduce modeling cost

Deliver the Data Business

1,000,000

High precision models

E2E pipeline + AutoML to reduce modeling cost



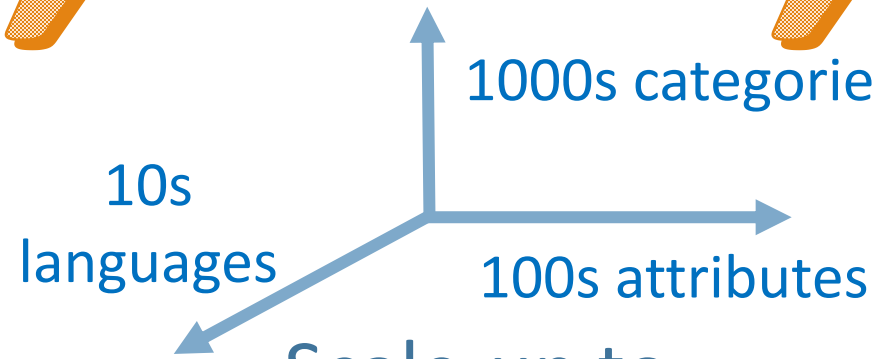
Scale-up to reduce #models

Deliver the Data Business

1,000,000,000

High precision models

E2E pipeline + AutoML to reduce modeling cost



Scale-up to reduce #models

Higher yield from multi-modal models

From 0 to 1: High-Quality Data

From Zero to One: A Core Algorithm

High precision models

OpenTag Extraction from Product Profiles

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		

Zheng et al., OpenTag: Open attribute value extraction from product profiles, KDD 2018.

OpenTag Extraction from Product Profiles

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

Zheng et al., OpenTag: Open attribute value extraction from product profiles, KDD 2018.

OpenTag Extraction from Product Profiles

B Beginning of attribute value

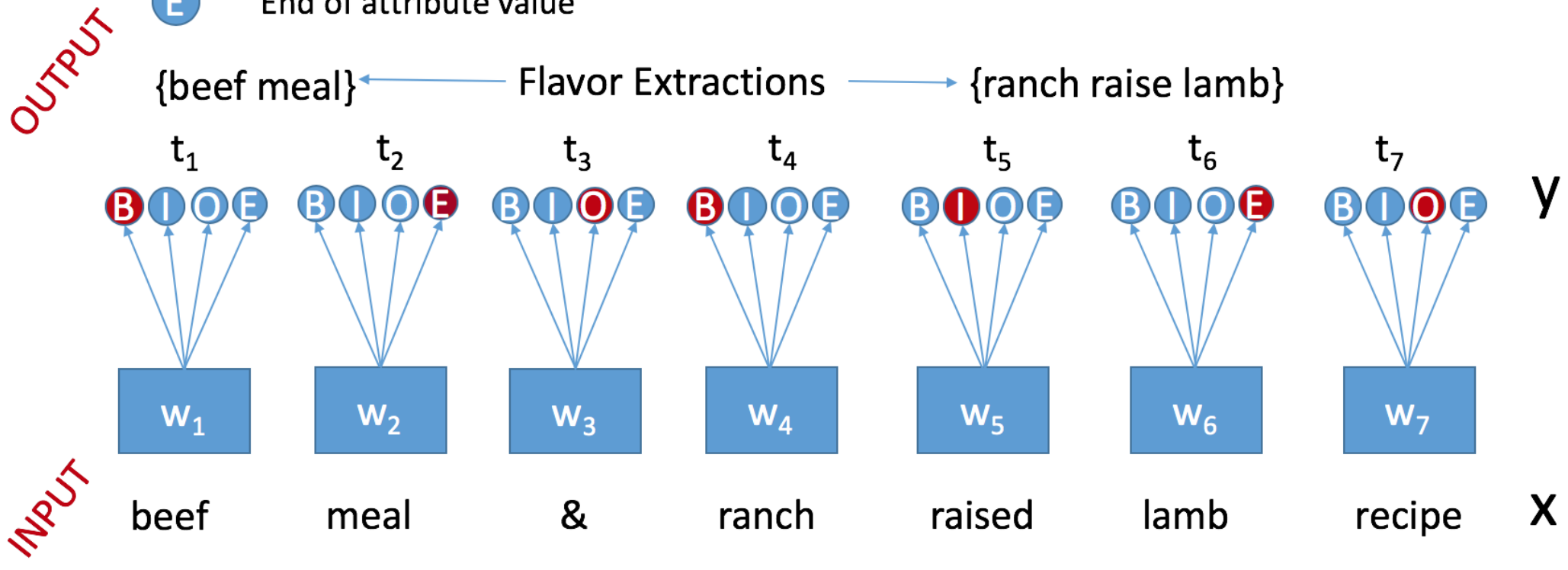
I Inside of attribute value

O Outside of attribute value

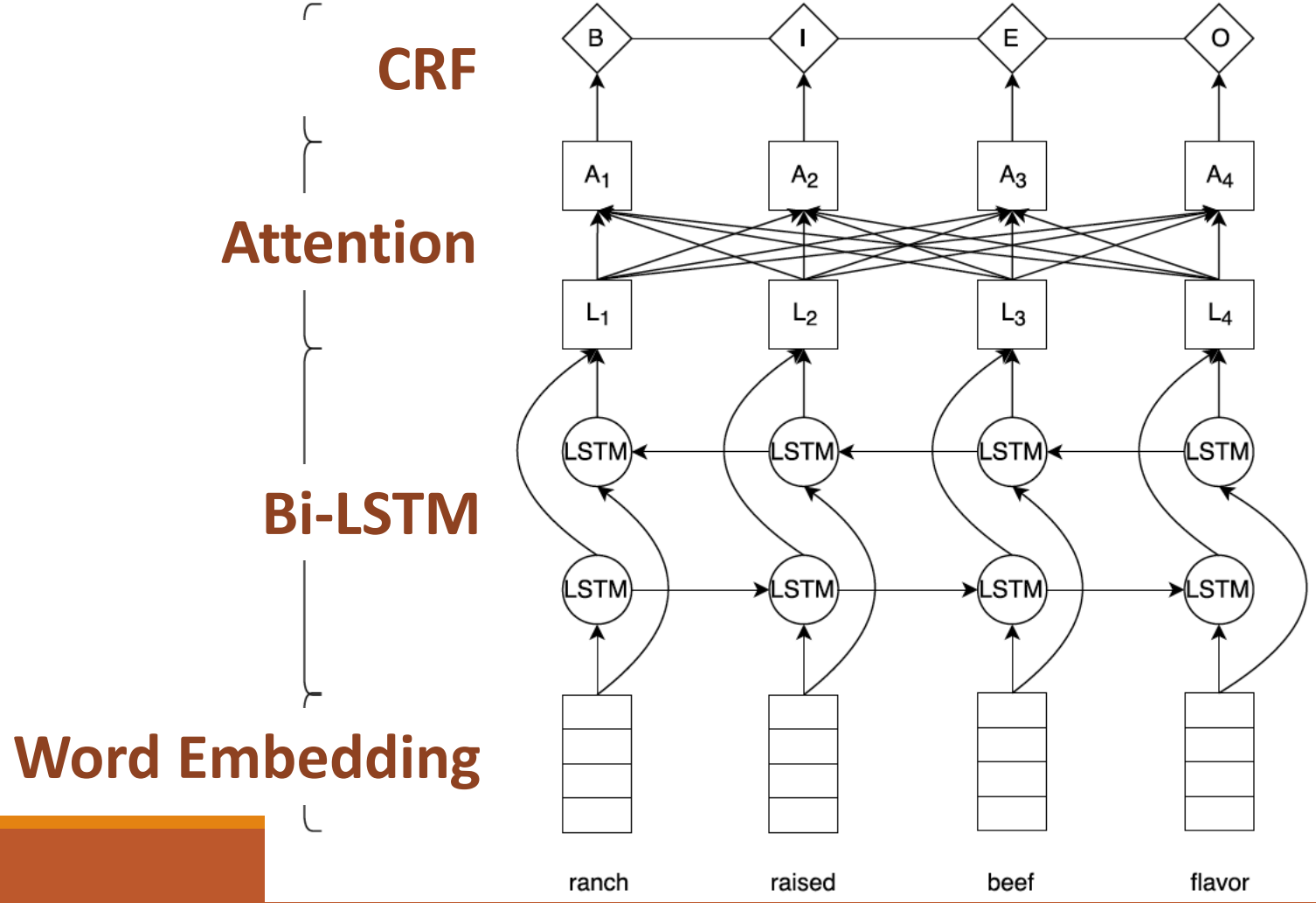
E End of attribute value

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

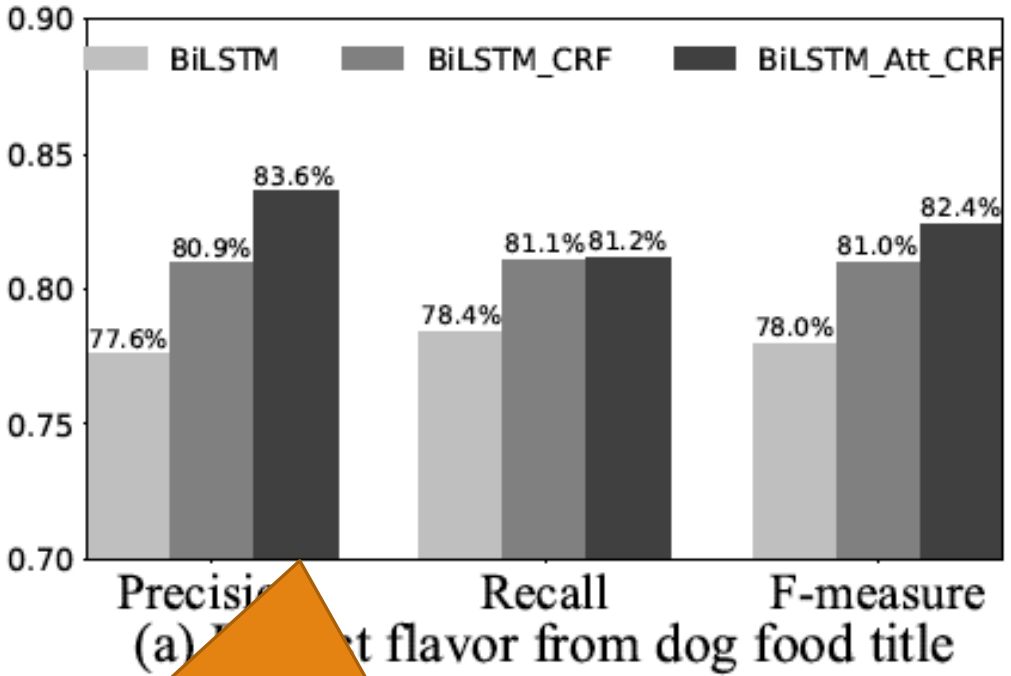
$y=\{t_1,t_2,\ldots,t_n\}$ tagging decision



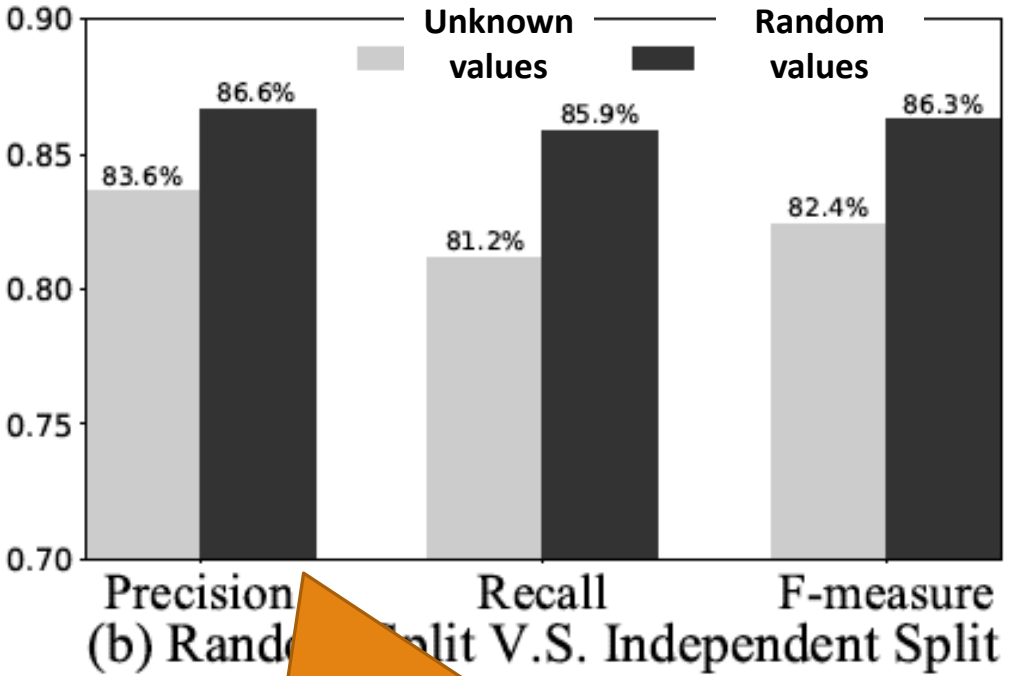
OpenTag Extraction from Product Profiles



OpenTag Extraction from Product Profiles



(a) ... flavor from dog food title



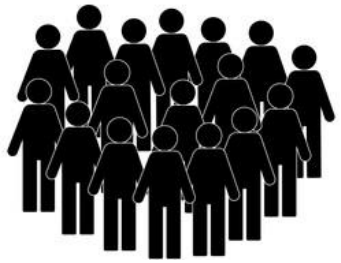
(b) Rand... Split V.S. Independent Split

BiLSTM+CRF+Attention obtains best results

Extraction on new values is comparable to already known values

Zheng et al., OpenTag: Open attribute value extraction from product profiles, KDD 2018.

OpenTag in Practice

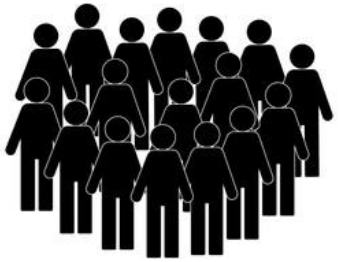


Understand domain and attributes, and generate LOTS OF training data

Train and fine-tune models

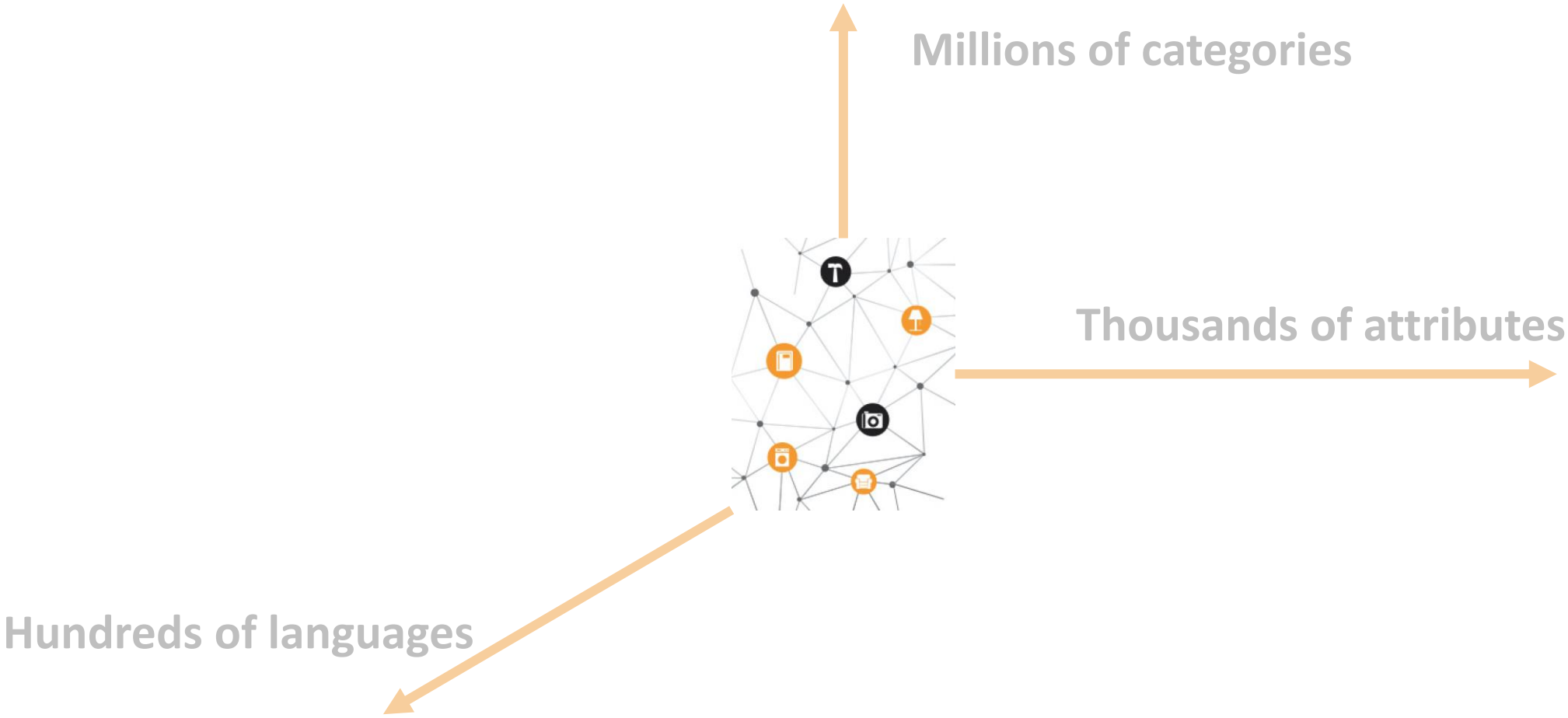
OpenTag

Postprocess extraction results to further improve data quality



Pre-publish evaluation as gatekeeper to guarantee high quality data

Still the Origin Point



From 1 to 1K: Reducing Modeling Cost

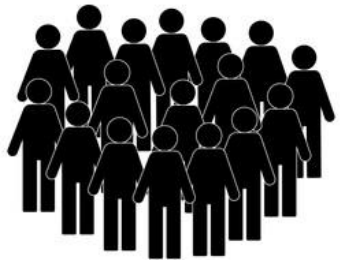
From 1 to 1K: E2E AutoML Pipeline

1,000

High precision models

E2E pipeline + AutoML to reduce modeling cost

An End-to-End Pipeline

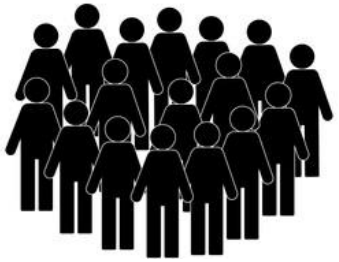


Understand domain and attributes, and generate LOTS OF training data

Train and fine-tune models

OpenTag

Postprocess extraction results to further improve data quality



Pre-publish evaluation as gatekeeper to guarantee high quality data

An End-to-End Pipeline

Train and fine-tune models

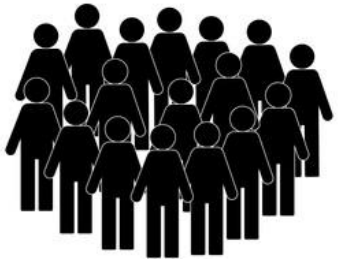
Automatic Training Data Generation

Distant supervision, Data programming

OpenTag

Benchmarking

Postprocess extraction results to further improve data quality



Pre-publish evaluation as gatekeeper to guarantee high quality data

An End-to-End Pipeline

Train and fine-tune models

Automatic Training Data Generation

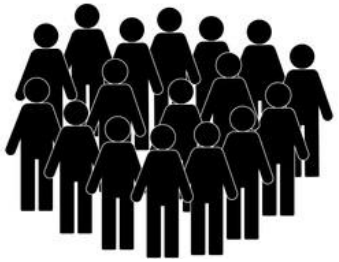
Distant supervision, Data programming

OpenTag

Benchmarking

Deep Learning Data Cleaning

Postprocess extraction results to further improve data quality



Pre-publish evaluation as gatekeeper to guarantee high quality data

An End-to-End Pipeline

Train and fine-tune models

Automatic Training Data Generation

Distant supervision, Data programming

OpenTag

Benchmarking

Deep Learning Data Cleaning

Postprocess extraction results to further improve data quality

Scale-up pre-publish evaluation w. lower labeling needs

An End-to-End Pipeline

AutoML

Automatic Training Data Generation

Distant supervision, Data programming

OpenTag

Deep Learning Data Cleaning

Benchmarking

Scale-up pre-publish evaluation w. lower labeling needs

Transformer-Based Anomaly Detection

Grocery & Gourmet Food › Candy & Chocolate › Mints

Love of Candy Bulk Candy - Pink Mint Chocolate Lentils - 6lb Bag

Brand: Love of Candy

14 ratings | 3 answered questions

Price: $84.99 ($0.89 / Ounce) + $16.92 shipping

Pay $14.17/month for 6 months, interest-free with your Amazon Prime Rewards Visa Card

Flavor Name: **Pink**

Blue | Green | Orange | Pastel Assortment | **Pink** | Red | White | Yellow

Size: **6 Pound**

1 Pound | 2 Pound | 3 Pound | 4 Pound | 5 Pound | **6 Pound** | 7 Pound | 8 Pound | 9 Pound | 10 Pound

- Love of Candy's huge selection of bulk candy now includes Premium Mint Chocolate Lentils in a variety of bold & striking colors. Available in small to large sizes ranging from 1 to 10 lb bags. These beautiful chocolate morsels feature gourmet, dairy free dark chocolate coated in a crispy and crunchy mint candy shell. Similar to M&M's, these mint chocolate candy lentils are fun, bite-sized snacks that can be enjoyed during any occasion.
- Sourced from the most esteemed candy makers from around the world, we've put together an extremely broad collection of wholesale candy to fulfill your every need. Whether you're in need of candy for vending machines, piñatas or candy buffets, you can trust that Love of Candy's got you covered. Our consistent product quality and unmatched customer satisfaction have quickly made Love of Candy the market's most trusted source of high quality, wholesale bulk candy.

Wang et al., Automatic validation of textual attribute values in eCommerce Catalog by learning with limited labeled data, KDD'20

Transformer-Based Anomaly Detection

Grocery & Gourmet Food › Candy & Chocolate › Mints

Love of Candy Bulk Candy - Pink Mint Chocolate Lentils - 6lb Bag

Brand: Love of Candy

14 ratings | 3 answered questions

Price: $84.99 ($0.89 / Ounce) + $16.92 shipping

Pay $14.17/month for 6 months, interest-free with your Amazon Prime Rewards Visa Card

Flavor Name: **Pink**

Blue | Green | Orange | Pastel Assortment | **Pink** | Red | White | Yellow

Size: **6 Pound**

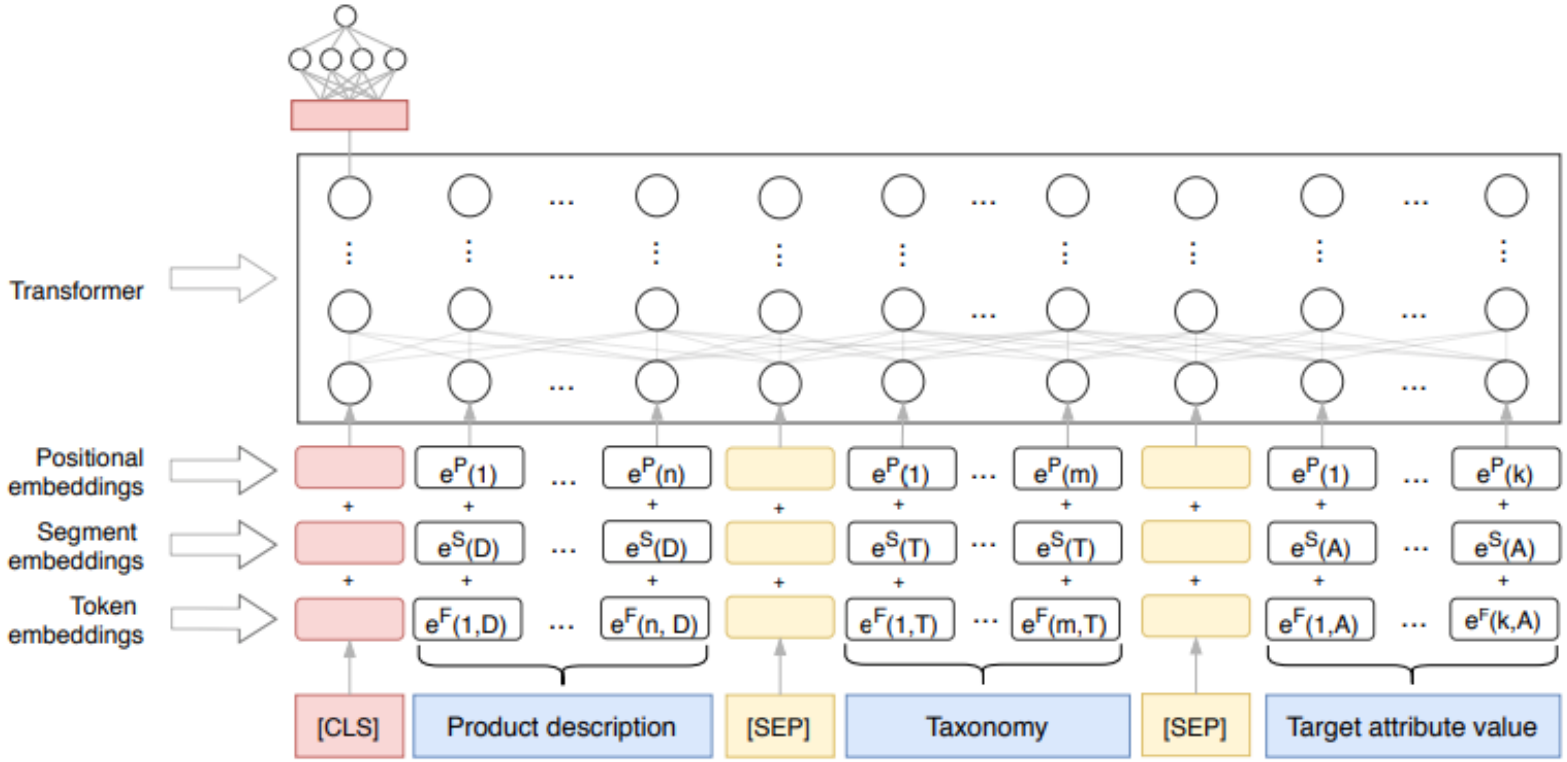
1 Pound | 2 Pound | 3 Pound | 4 Pound | 5 Pound | **6 Pound** | 7 Pound | 8 Pound | 9 Pound | 10 Pound

- Love of Candy's huge selection of bulk candy now includes Premium Mint Chocolate Lentils in a variety of bold & striking colors. Available in small to large sizes ranging from 1 to 10 lb bags. These beautiful chocolate morsels feature gourmet, dairy free dark chocolate coated in a crispy and crunchy mint candy shell. Similar to M&Ms, these mint chocolate candy lentils are fun, bite-sized snacks that can be enjoyed during any occasion.
- Sourced from the most esteemed candy makers from around the world, we've put together an extremely broad collection of wholesale candy to fulfill your every need. Whether you're in need of candy for vending machines, piñatas or candy buffets, you can trust that Love of Candy's got you covered. Our consistent product quality and unmatched customer satisfaction have quickly made Love of Candy the market's most trusted source of high quality, wholesale bulk candy.

Is the flavor "Pink"?

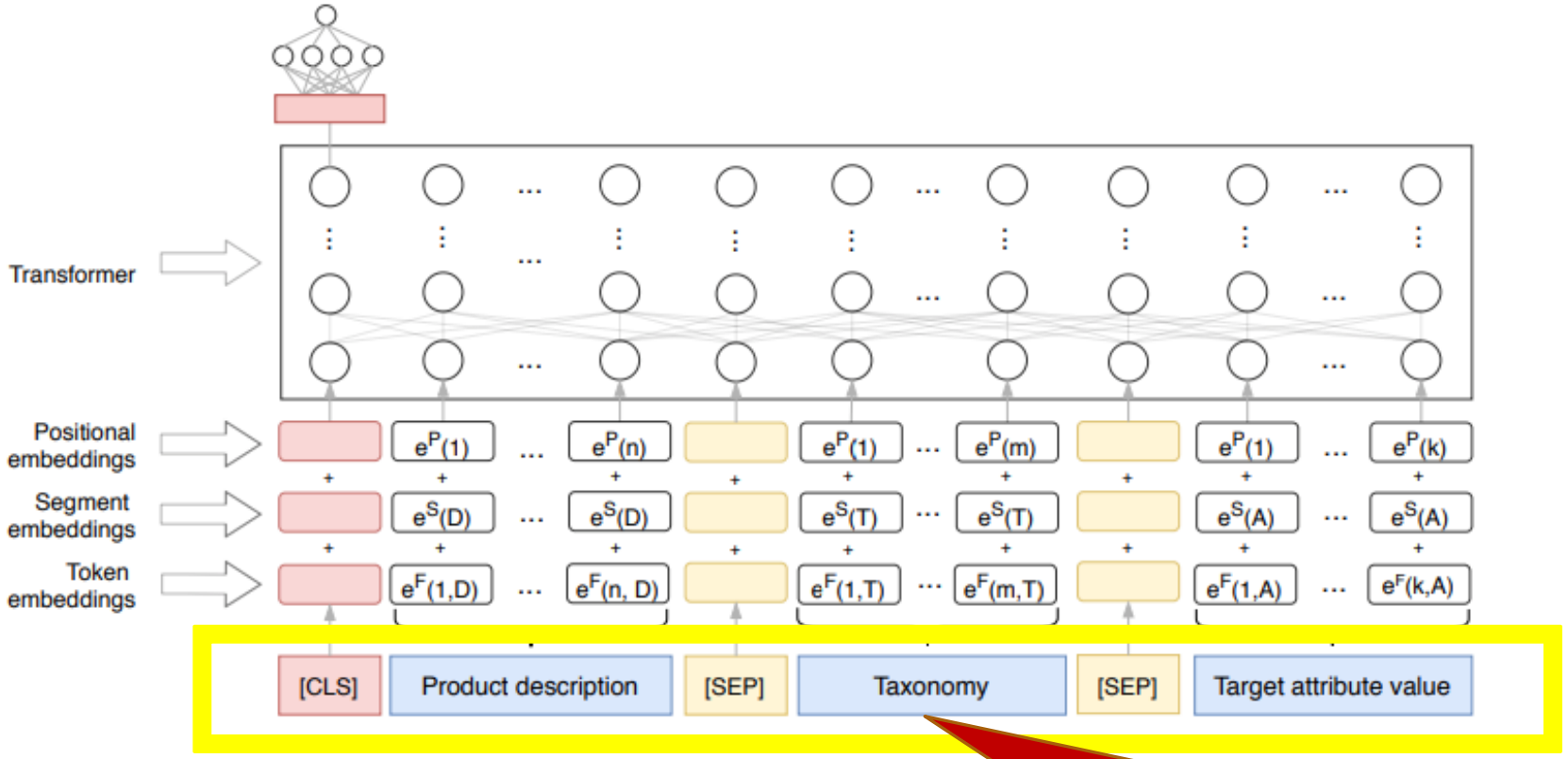
Wang et al., Automatic validation of textual attribute values in eCommerce Catalog by learning with limited labeled data, KDD'20

Transformer-Based Anomaly Detection



Wang et al., Automatic validation of textual attribute values in eCommerce Catalog by learning with limited labeled data, KDD'20

Transformer-Based Anomaly Detection



Category as input for model training

Transformer-Based Anomaly Detection

- Identify **1.77MM** incorrect values for Flavor and Scent for Consumables with **90% precision**

Product	Attr	Value
Love of Candy Bulk Candy - Pink Mint Chocolate Lentils - 6lb Bag	Flavor	Pink
Scott's Cakes Dark Chocolate Fruit & Nut Cream Filling Candies with Burgandy Foils in a 1 Pound Snowflake Box	Flavor	1 lb. snowflake box
Lucky Baby - Baby Blanket Envelope Swaddle Winter Wrap Coral Fleece Newborn Blanket Sleeper Infant Stroller Wrap Toddlers Baby Sleeping Bag (color 1)	Flavor	color 1
ASUTRA Himalayan Sea Salt Body Scrub Exfoliator + Body Brush (Vitamin C), 12 oz | Ultra Hydrating, Gentle, Moisturizing | All Natural & Organic Jojoba, Sweet Almond, Argan Oils	Scent	vitamin c body scrub - 12oz & body brush
Folgers Simply Smooth Ground Coffee, 2 Count (Medium Roast), 31.1 Ounce	Scent	2Packages (Breakfast Blend, 31.1 oz)

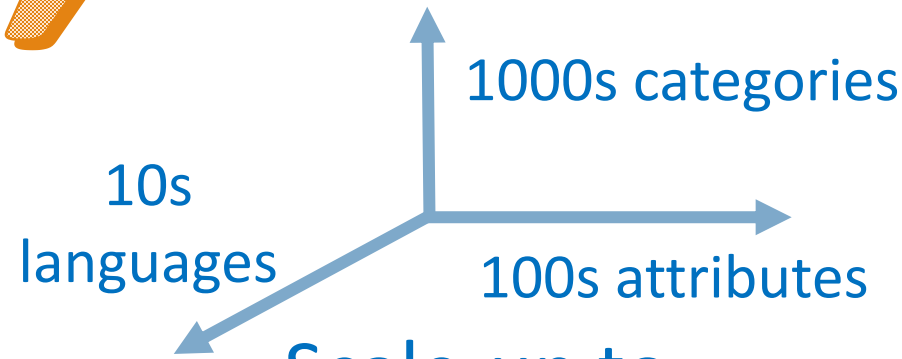
From 1K to 1M: Scaling Up

From 1K to 1M: One Size Fits All

1,000,000

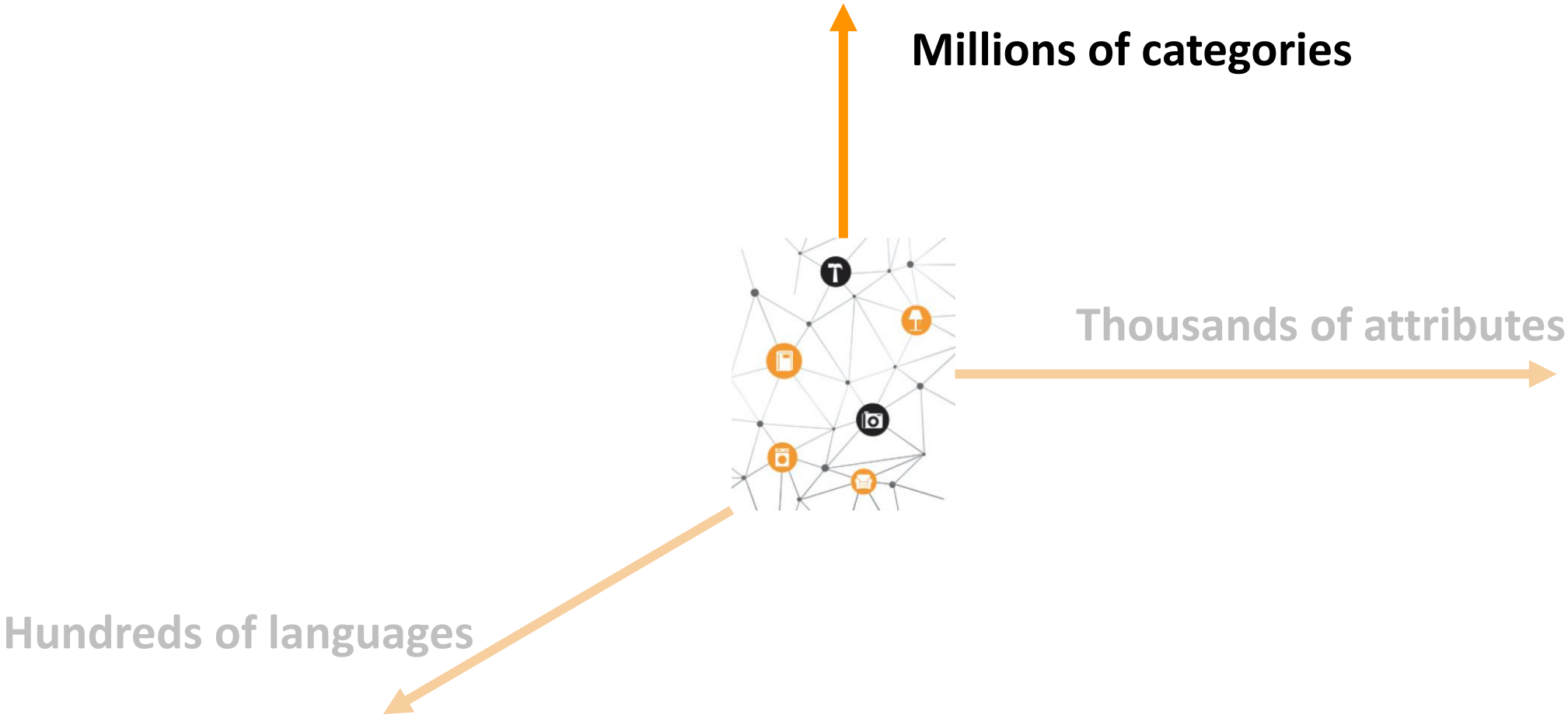
High precision models

E2E pipeline + AutoML to reduce modeling cost



Scale-up to reduce #models

Scale up for Millions of Categories



Scale-up I: Millions of Categories

Option 1. Train a single model? *Train/Test Distribution shift -> Invalid predictions*

Samsung UN58RU7100FXZA Flat 58-Inch 4K UHD 7 Series Ultra HD Smart TV with HDR and Alexa Compatibility (2019 Model)

Taylors of Harrogate Classic Tea Variety Box, 48 Count (Pack of 1)

Caribou Coffee Caribou Blend, Medium Roast Ground Coffee, 20 Ounce Bag, Rainforest Alliance Certified

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.

Scale-up I: Millions of Categories

Option 1. Train a single model? *Train/Test Distribution shift -> Invalid predictions*

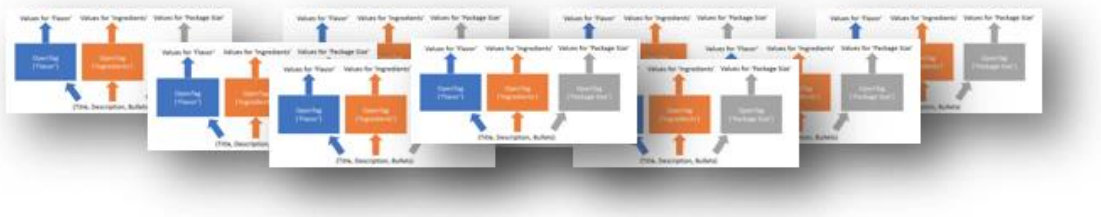
Samsung UN58RU7100FXZA Flat 58-Inch 4K UHD 7 Series Ultra HD Smart TV with HDR and Alexa Compatibility (2019 Model)

Taylors of Harrogate Classic Tea Variety Box, 48 Count (Pack of 1)

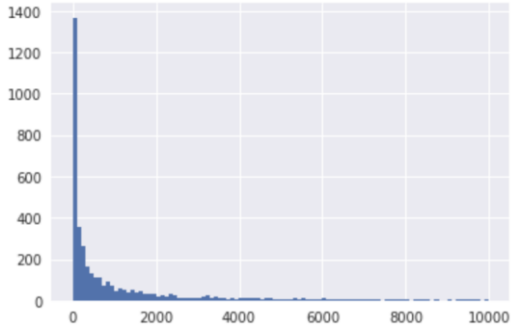
Caribou Coffee Caribou Blend, Medium Roast Ground Coffee, 20 Ounce Bag, Rainforest Alliance Certified

Option 2. Train a model for each category?

Store/orchestrate 100K+ OpenTag models



Most categories are very sparse



Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.

Scale-up I: Millions of Categories

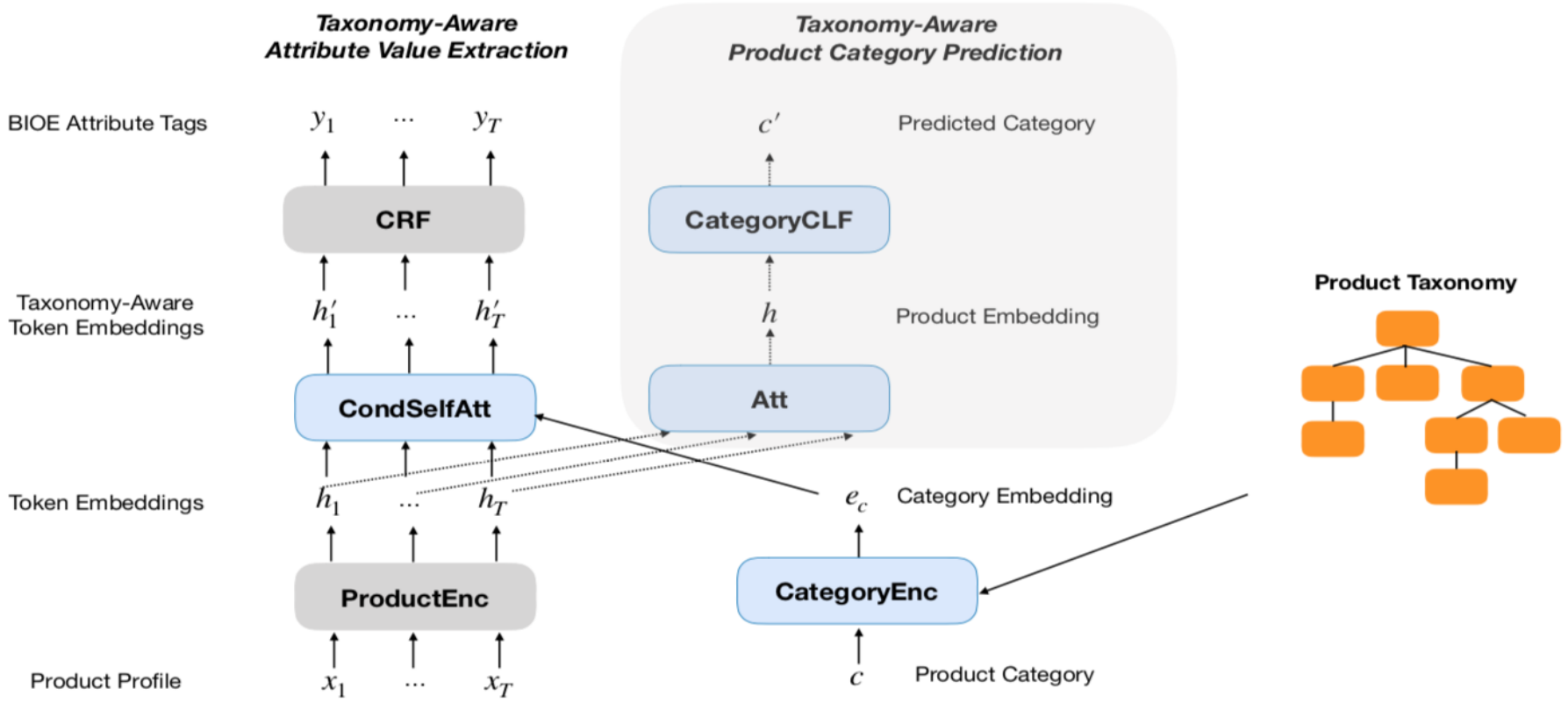


Figure 2: Our TXtract architecture for hierarchical multi-task learning.

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.

Scale-up I: Millions of Categories

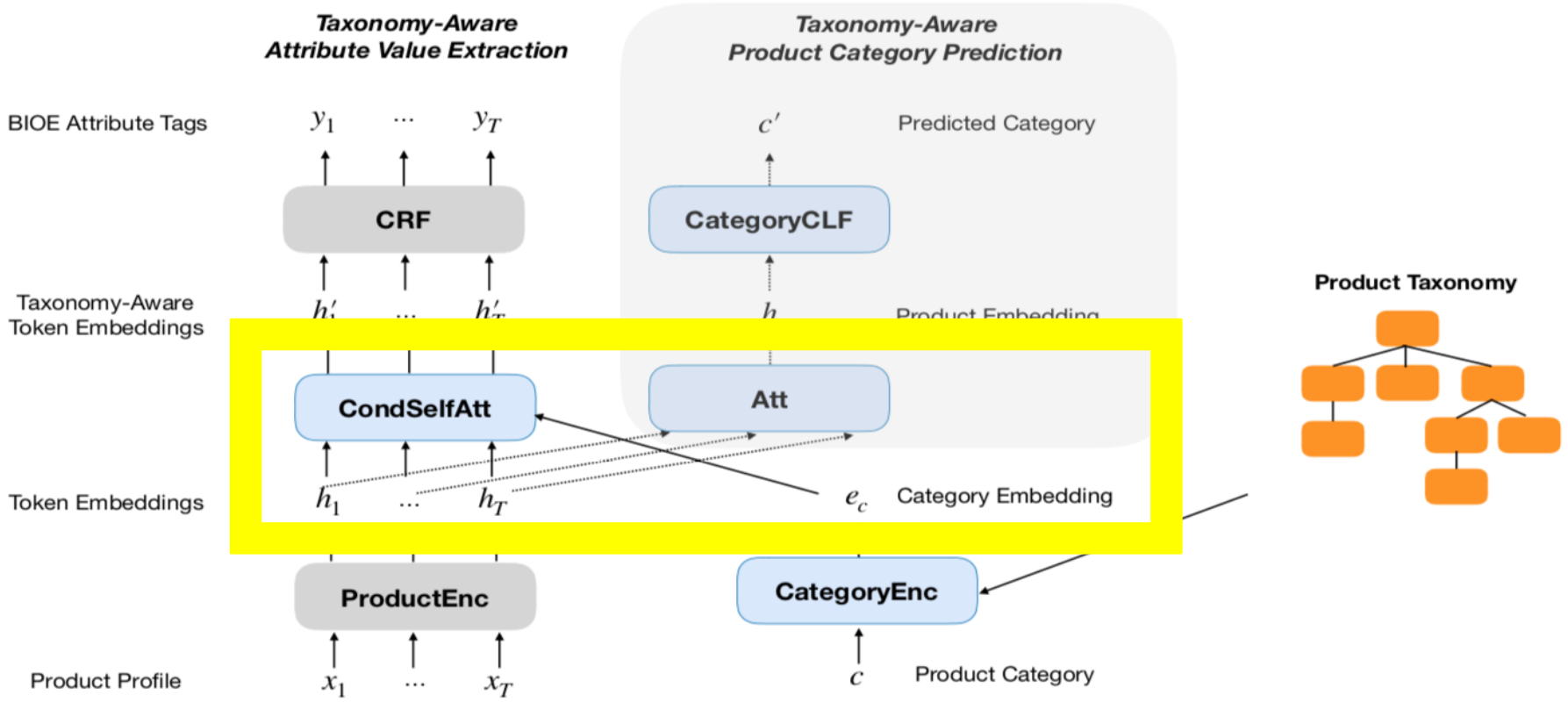


Figure 2: Our TXtract architecture for hierarchical multi-task learning.

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.

Scale-up I: Millions of Categories

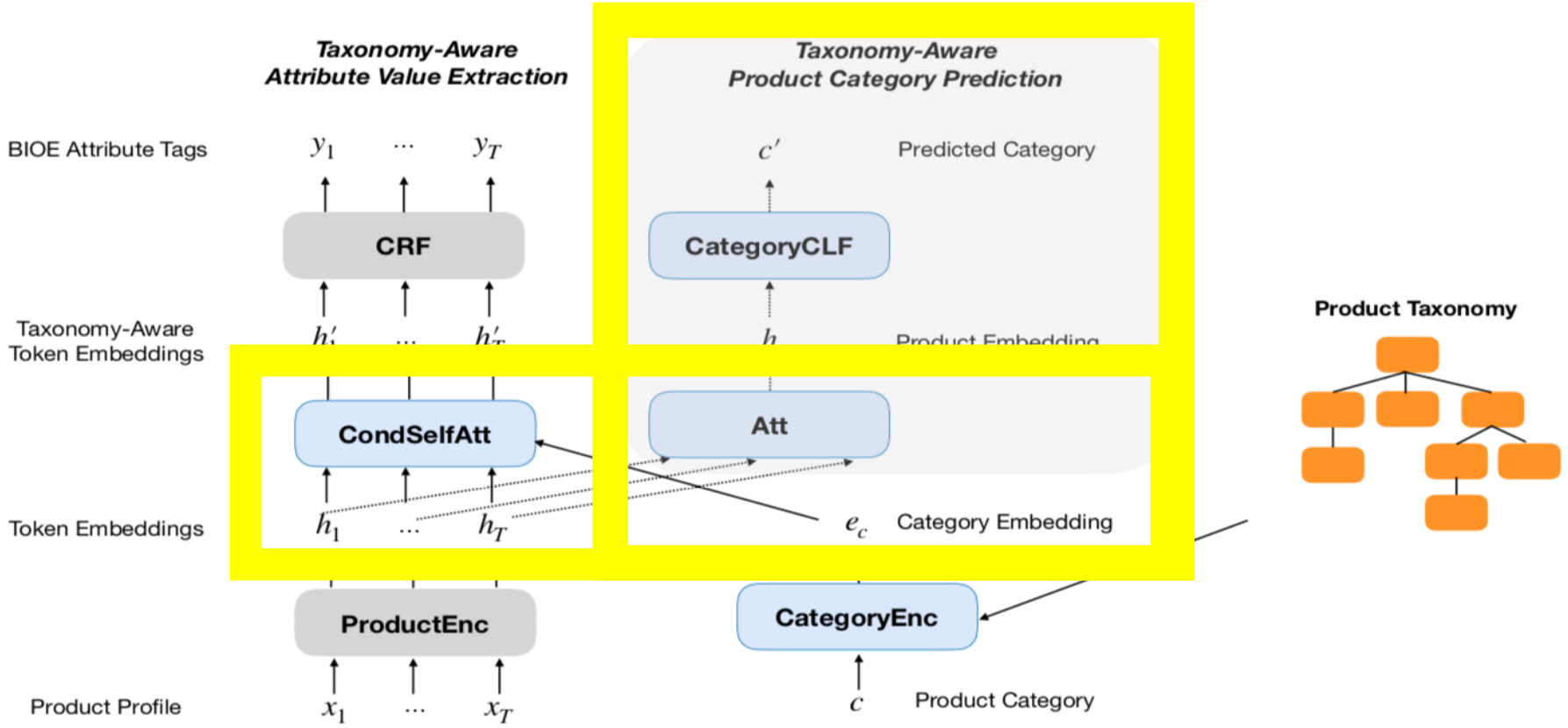
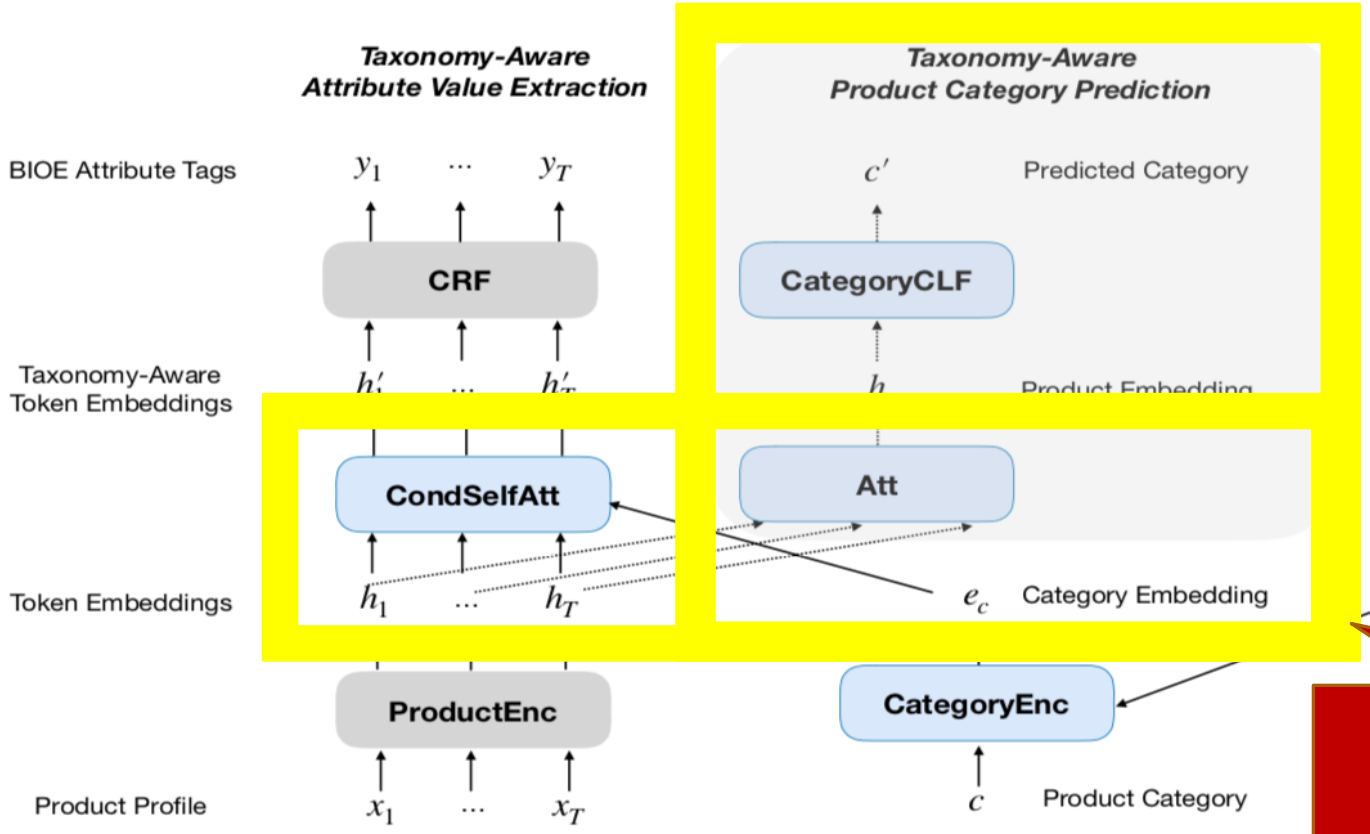


Figure 2: Our TXtract architecture for hierarchical multi-task learning.

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.

Scale-up I: Millions of Categories



Attention conditioned on category representation

Figure 2: Our TXtract architecture for hierarchical multi-task learning.

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.

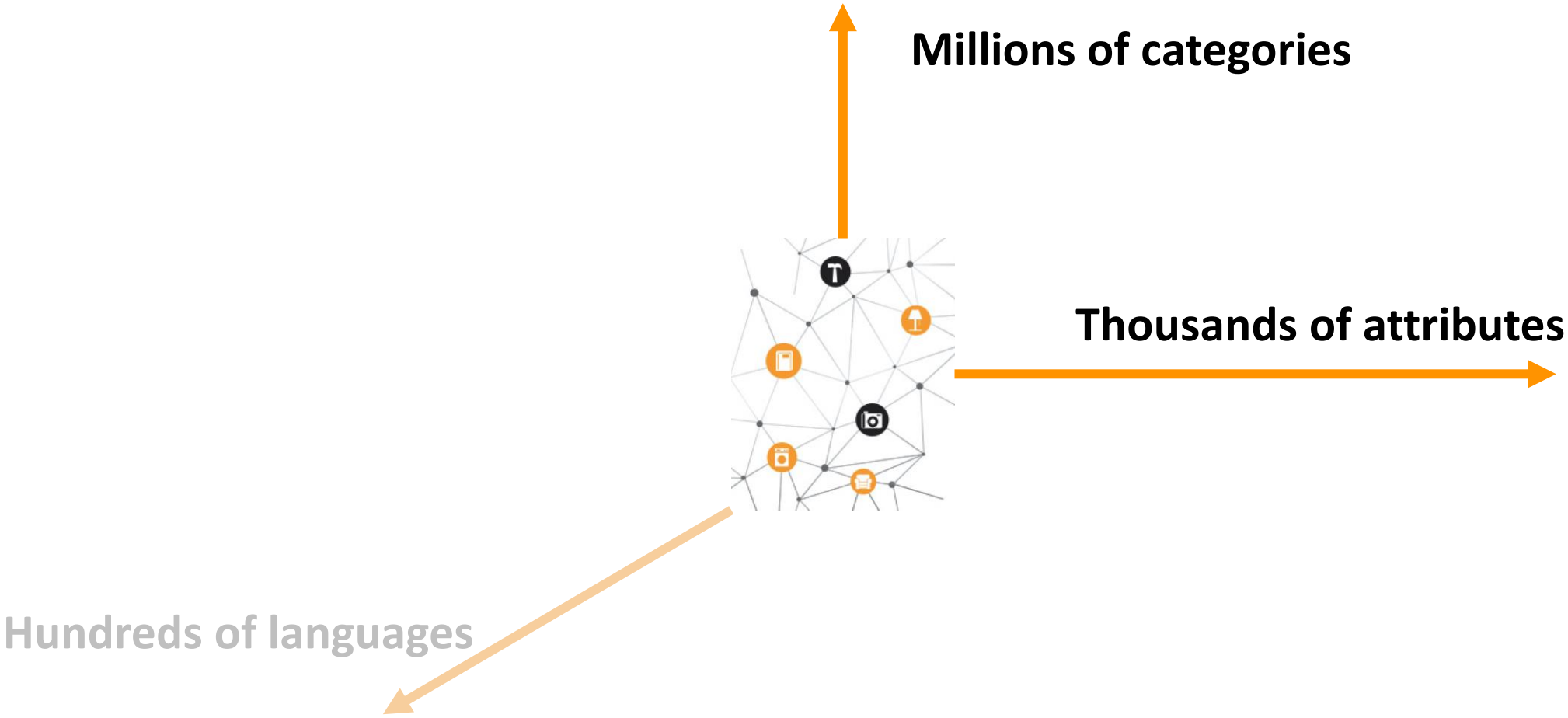
Scale-up I: Millions of Categories

- Train one model on 4K categories, and improve state-of-the-art by 10.4% in F1, and by 11.7% in coverage

	Title	OpenTag	TXtract
1	Controlled Labs Purple Wraath 90 Servings - Purple Lemonade	*flavor*: -	*flavor*: purple lemonade
2	Click - Espresso Protein Drink Vanilla Latte - 16 oz.	*flavor*: espresso	*flavor*: vanilla latte
3	Mason Vitamins Melatonin 500 mcg Fast Meltz Tablets, Fruit, 60 Count	*flavor*: -	*flavor*: fruit
4	Fashion Glitter Matte Eye Shadow Powder Palette Single Shimmer Eyeshadow	*scent*: palette	*scent*: -
5	Baby car seat cover, Nursing covers Breastfeeding cover carseat canopy (Style5)	*scent*: style5	*scent*: -

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.

Scale up for Thousands of Attributes



Scale-up II: Thousands of Attributes

First Aid Beauty Ultra Repair Cream: Vegan and Gluten-Free Intense Moisturizer for Dry Sensitive Skin. Perfect for Skin Conditions and Eczema. Pink Grapefruit (14 ounce)

About this item

- HEAD-TO-TOE: Head-to-toe moisturizer that provides instant relief and long-term hydration for dry, distressed skin, even eczema. The beautiful, whipped texture is instantly absorbed with no greasy after-feel. Grapefruit has a bright citrus fruit scent that is fresh, juicy and sparkling.
- CLINICALLY PROVEN: Formulated with Colloidal Oatmeal, Shea Butter, Ceramide 3 and the FAB Antioxidant Booster, it provides immediate relief and visible improvement for parched skin and it is clinically proven to increase hydration by 169% immediately upon application.

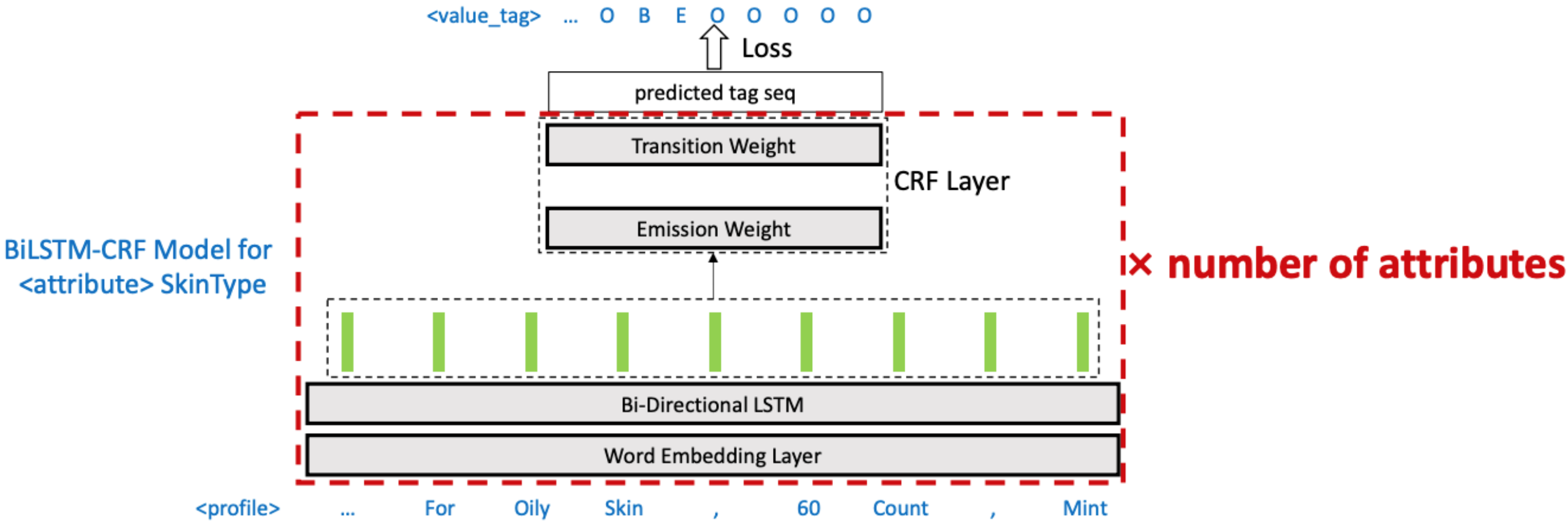
Product description

Banish dry skin with First Aid Beauty's Ultra Repair Cream. Suitable for all skin types, especially dry, flaky skin, this hydration wonder leaves skin feeling smooth, hydrated and comfortable after just a single use.

Mentioned Attributes: **Brand** **SkinType** **Scent** **Quantity**

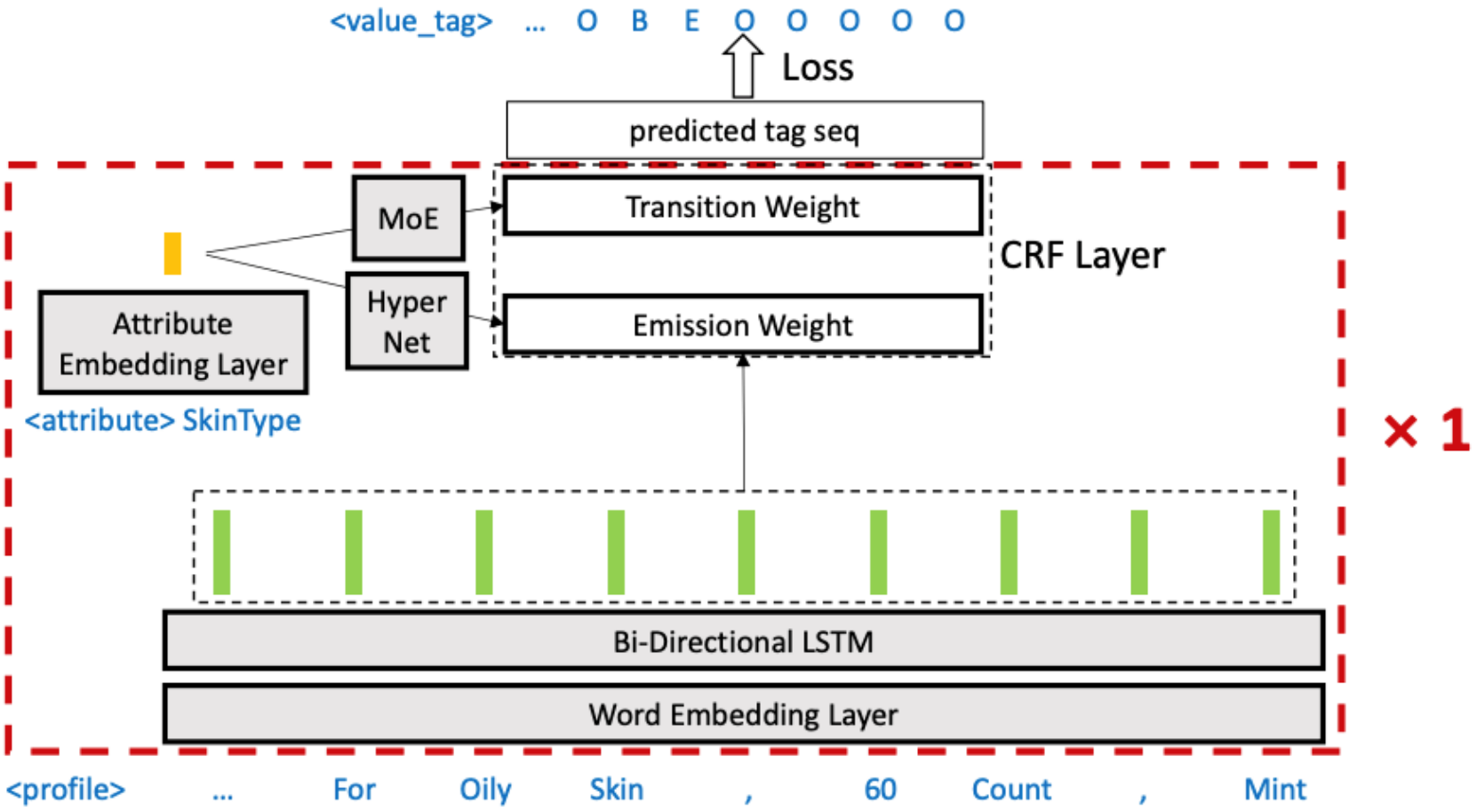
Yan et al., AdaTag: Multi-Attribute Value Extraction from Product Profiles with Adaptive Decoding, ACL 2021.

Scale-up II: Thousands of Attributes



Yan et al., AdaTag: Multi-Attribute Value Extraction from Product Profiles with Adaptive Decoding, ACL 2021.

Scale-up II: Thousands of Attributes

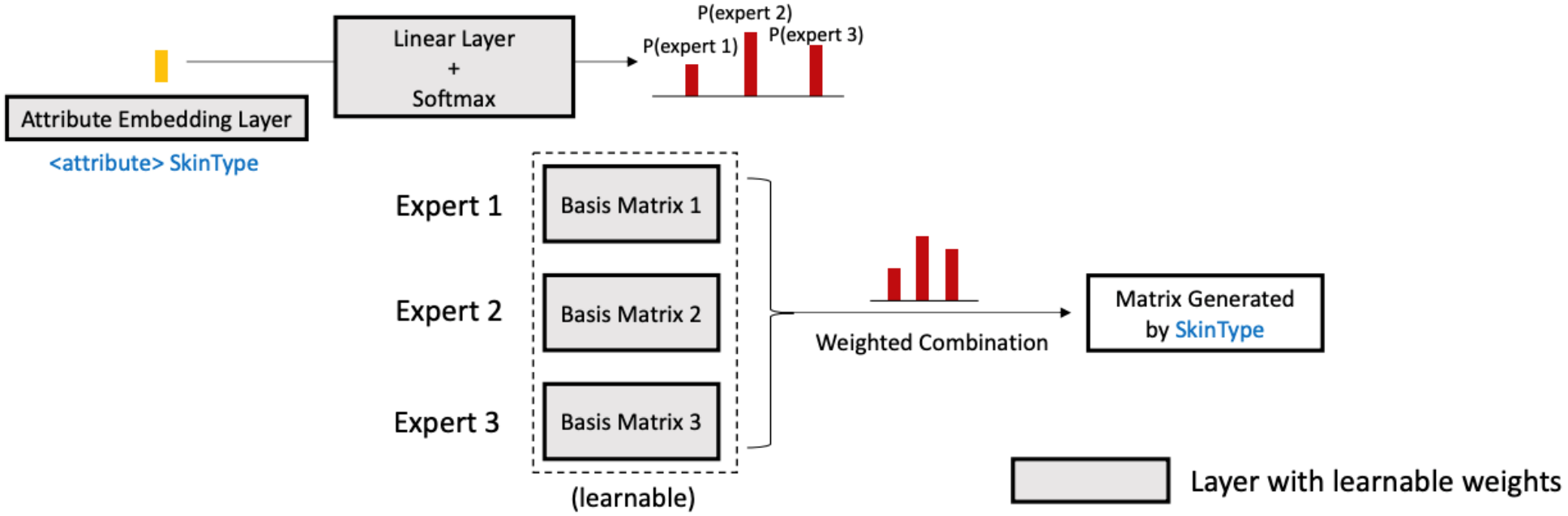


Yan et al., AdaTag: Multi-Attribute Value Extraction from Product Profiles with Adaptive Decoding, ACL 2021.

Scale-up II: Thousands of Attributes

MoE (Mixture-of-Experts)

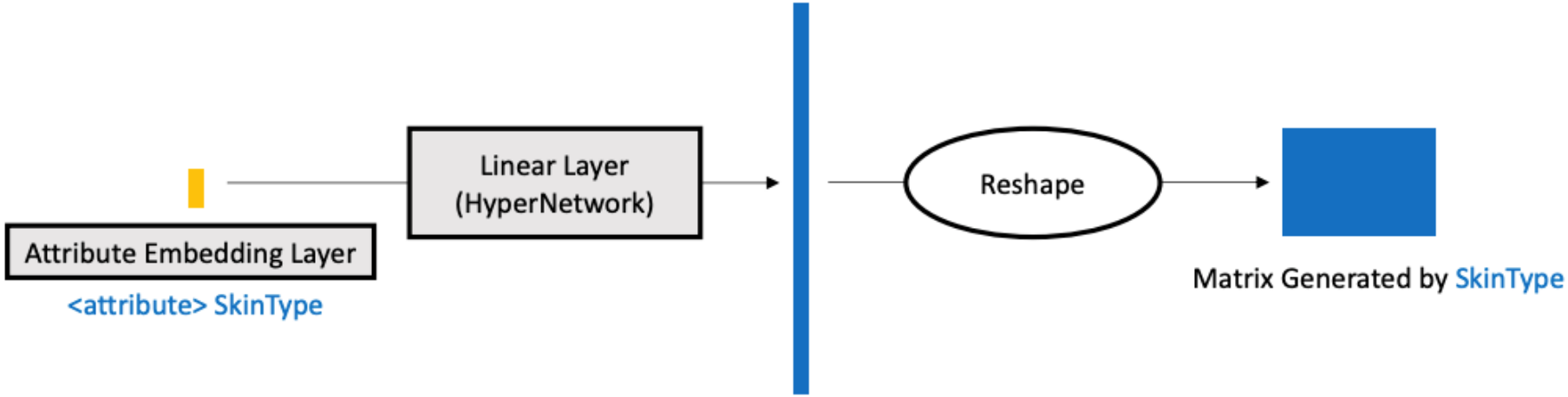
- Question: Attribute -> Layer Weight (Matrix)
  - Use a weighted combination of basis matrices to generate the weight.



Scale-up II: Thousands of Attributes

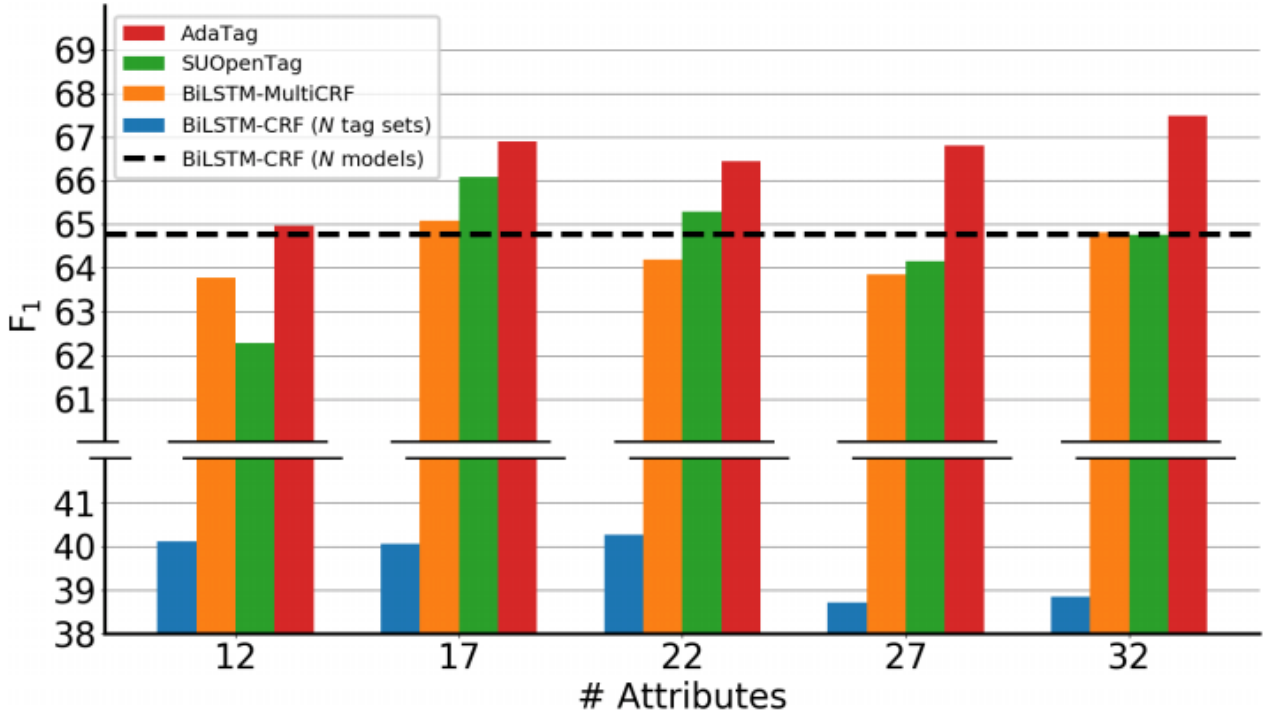
HyperNet (Hyper Network)

- Question: Attribute -> Layer Weight (Matrix)
  - Use a network to generate the weight of another network.



Scale-up II: Thousands of Attributes

- Train one model for 32 attributes, obtaining higher quality than single-attribute models



Yan et al., AdaTag: Multi-Attribute Value Extraction from Product Profiles with Adaptive Decoding, ACL 2021.

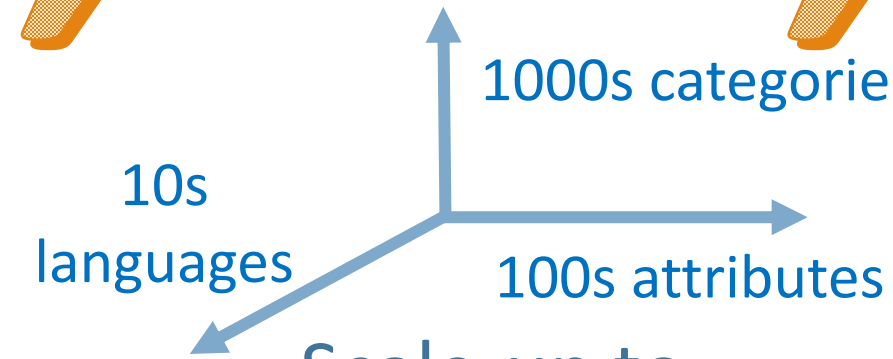
From 1M to 1B: Increasing the Yield

From 1M to 1B: Multi-Modal Extraction

1,000,000,000

High precision models

E2E pipeline + AutoML to reduce modeling cost

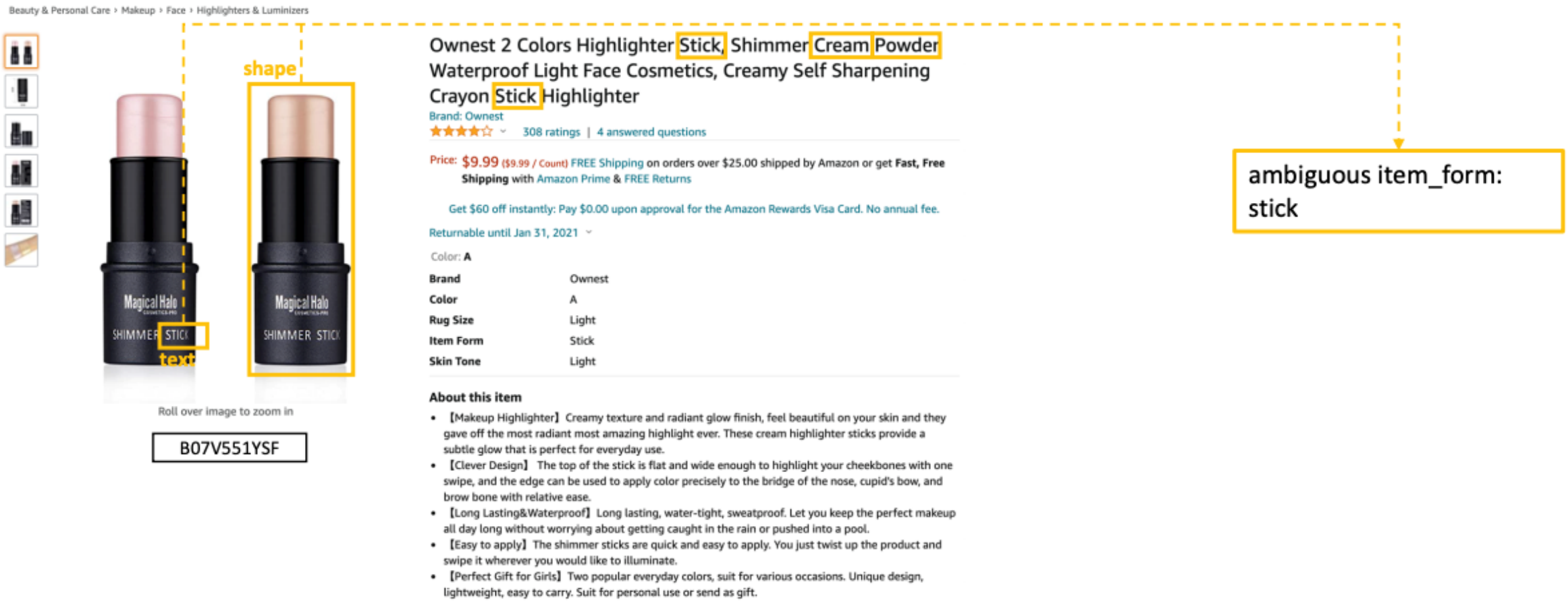


Scale-up to reduce #models

Higher yield from multi-modal models

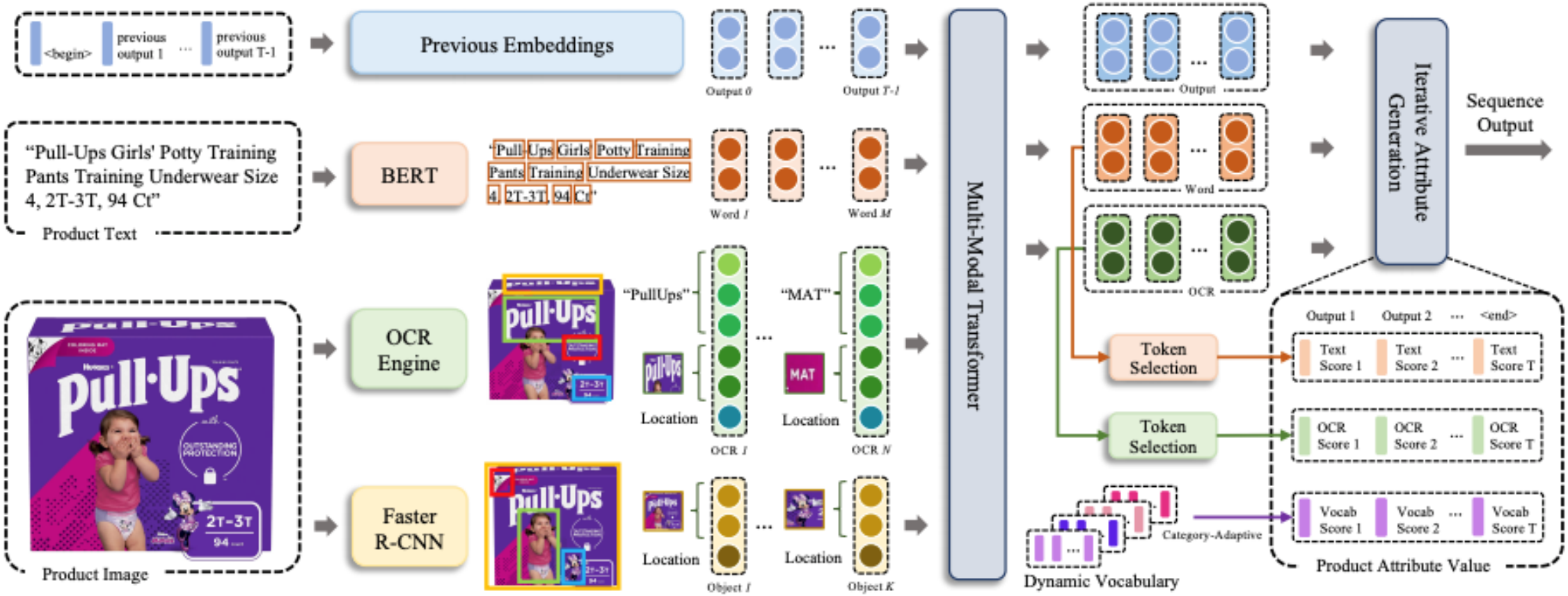
Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.

Multi-Modal Signals



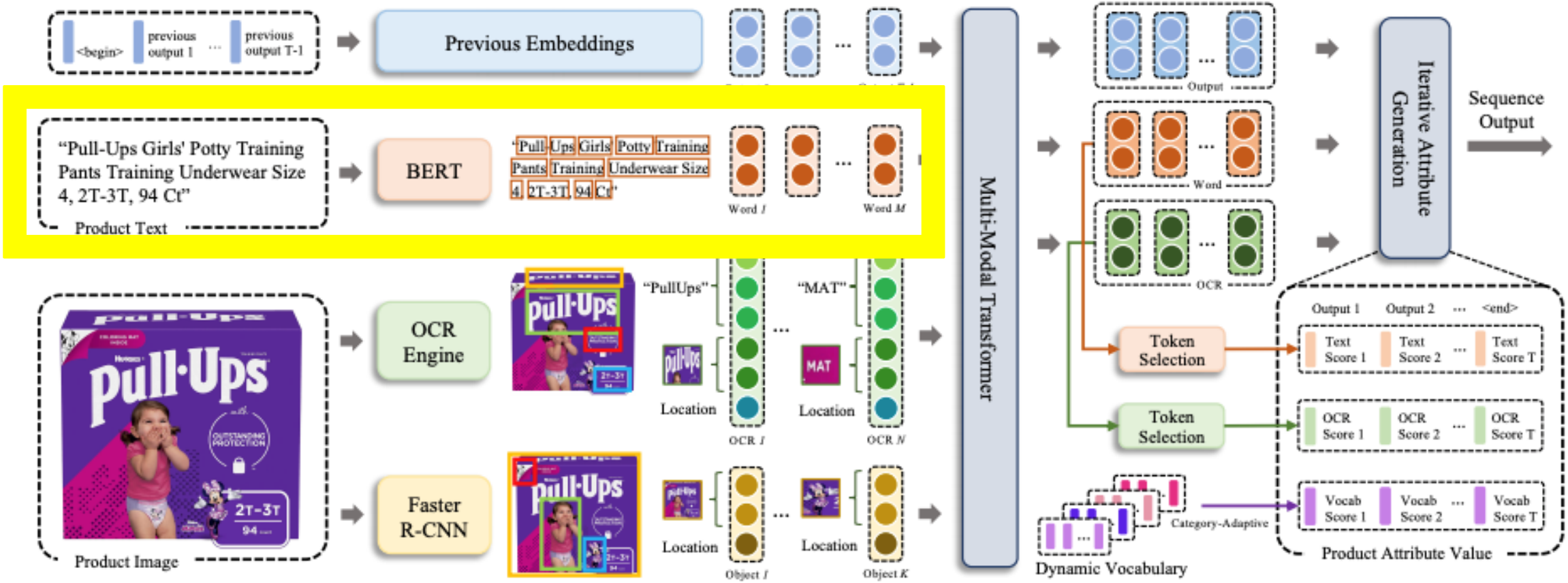
Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.

Multi-Modal Extraction



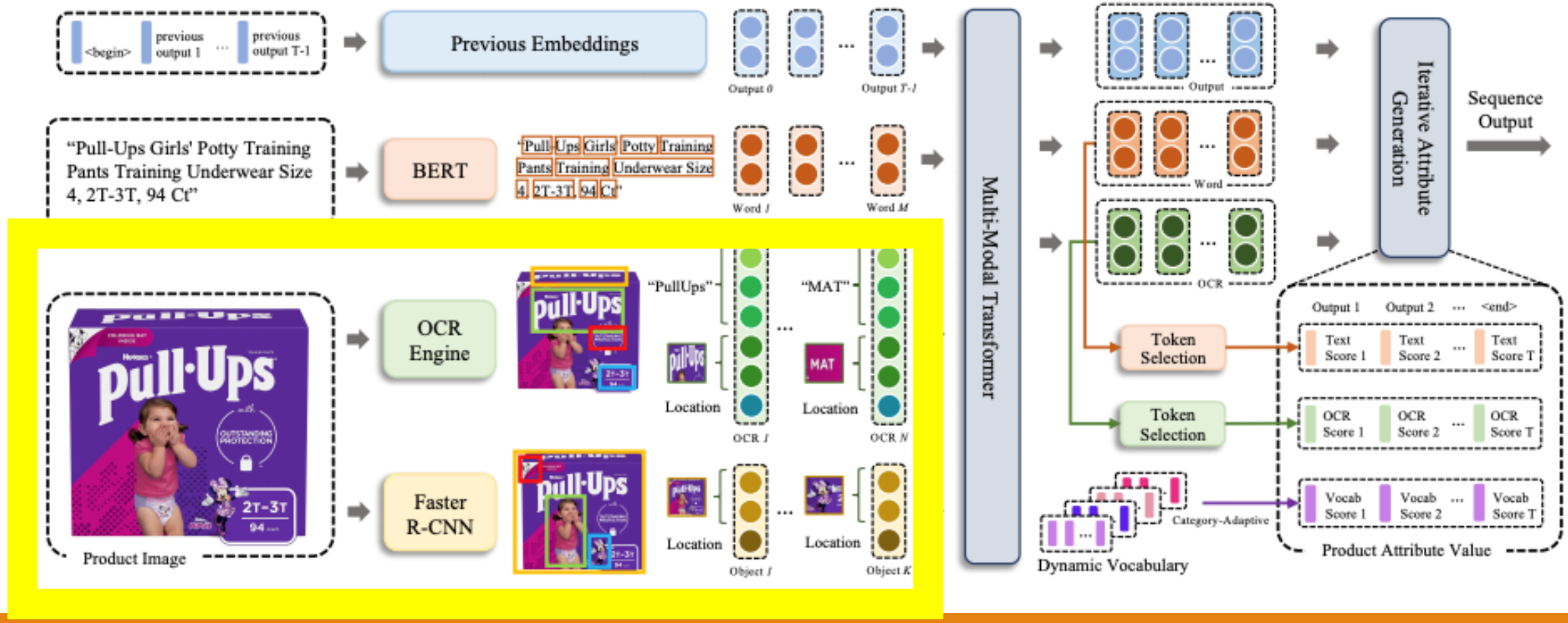
Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.

Multi-Modal Extraction



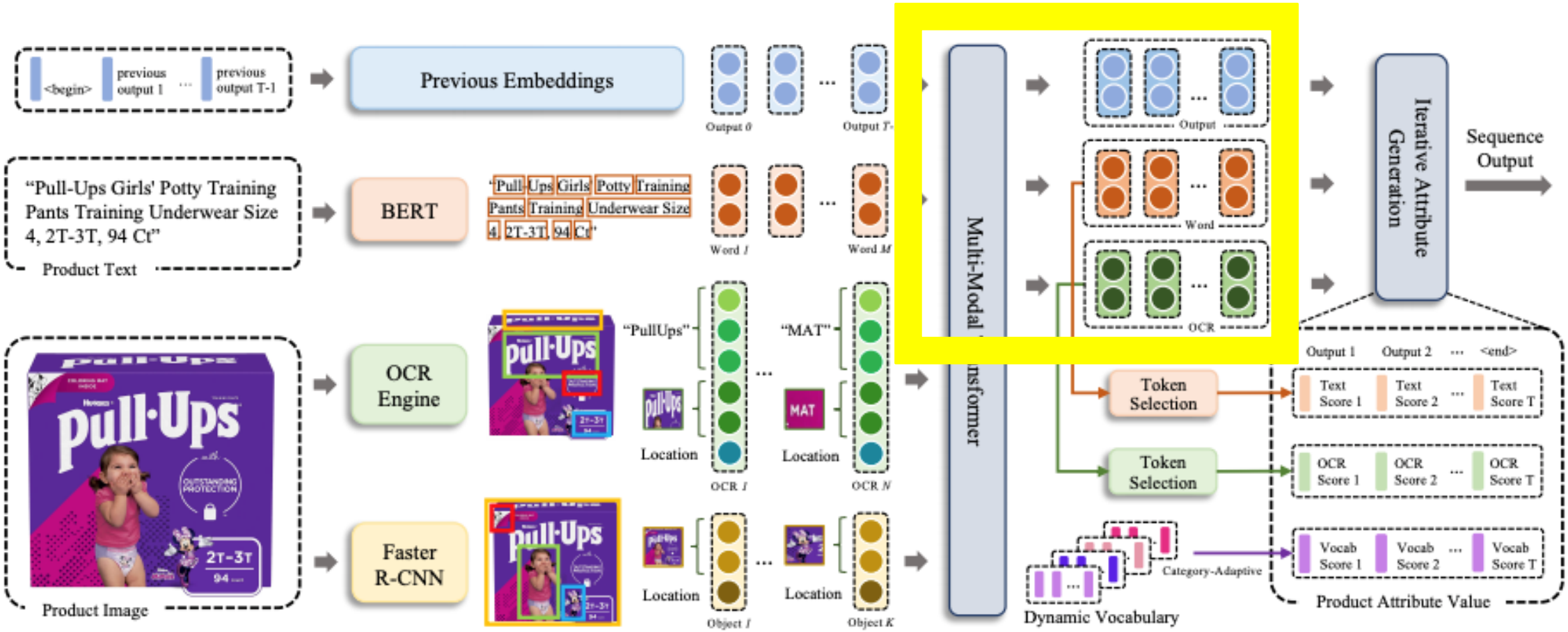
Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.

Multi-Modal Extraction



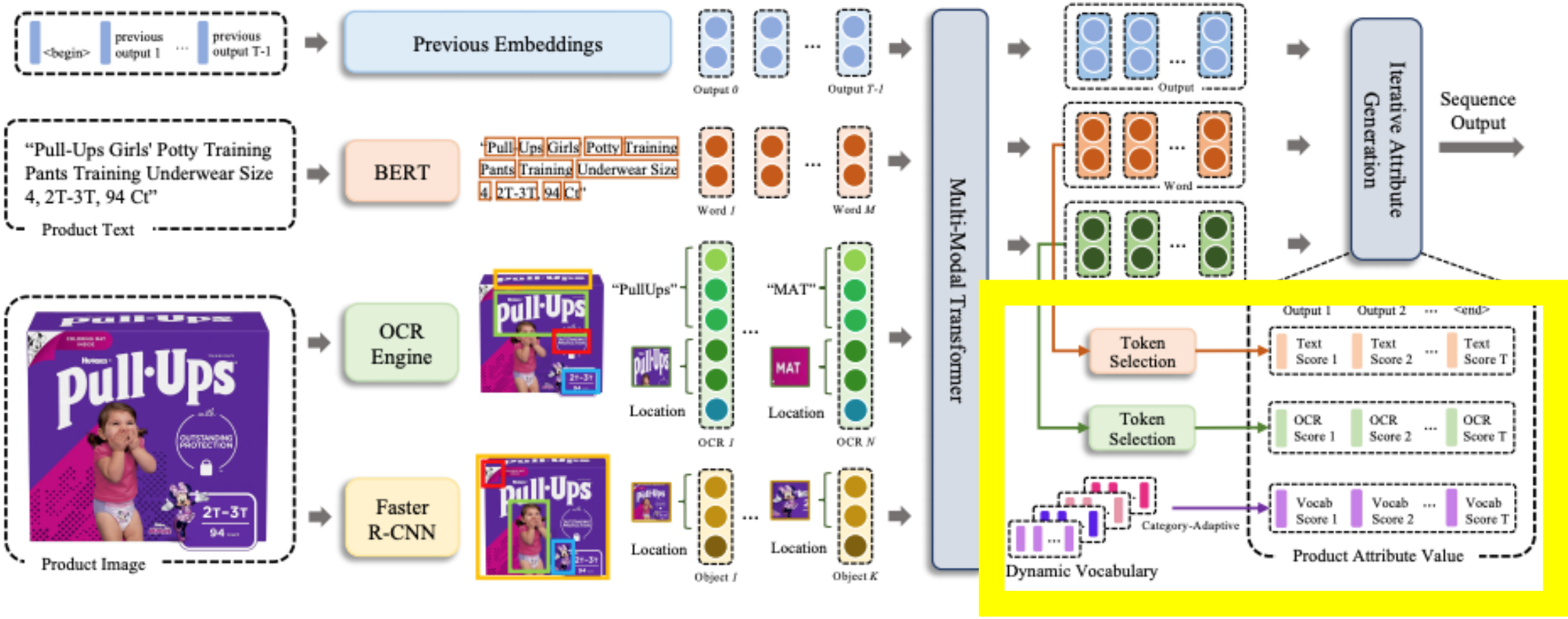
Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.

Multi-Modal Extraction



Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.

Multi-Modal Extraction



Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.

Multi-Modal Extraction

- Improve over OpenTag (text-only) by 11% on F-measure

Models	*P(%)*	*R(%)*	*F1(%)*
PAM *w/o* text	79.9	63.4	70.7
PAM *w/o* image	88.7	72.1	79.5
PAM *w/o* OCR	82.0	69.4	75.1
PAM	91.3	75.3	82.5

Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.

Web Extraction

Laundry > Liquid Laundry Detergent

ItemForm

ARM & HAMMER™

Plus OxiClean™, Fresh Scent

Scent

★★★★½ 4.8 126 REVIEWS

Add your own Review

THE SUPERCHARGED STAIN-FIGHTING DUO.

- Removes over 101 tough stains
- Powers out dirt and odors using the power of ARM & HAMMER™ Baking Soda and supercharged with OxiClean™ Stain Fighters
- Long-lasting scent
- Clean rinse with low-suds formula
- Works in standard and High Efficiency (HE) models
- Powered by OxiClean™, America's #1 stain fighter Additive Brand*
- Long-lasting fresh scent

Size: 122.5 oz

WHERE TO BUY

+ Product Ingredients

+ FAQs

Dong, Ceres: Harvesting Knowledge from Semi-Structured Web, Keynote at CIKM, 2020.

Successful vs. Not-Yet-Successful Fields in Industry

SUCCESS

- Entity linkage
- Knowledge extraction (ClosedIE)
- Knowledge cleaning
- Knowledge-based Q&A

1. One cannot live without it
2. The techniques are ready

SUCCESS NOT-YET

- Schema mapping
  - Small scale: Manual
  - Large scale: “Replaced” by ClosedIE
- OpenIE
  - “Replaced” by Reading comprehension
- Knowledge fusion
  - Not fully needed yet
- Knowledge inference
  - Quality not high enough yet

Take Aways

- We are building an authoritative product knowledge graph for millions of categories, thousands of attributes, and hundreds of languages
- High-accuracy modeling is the first step for building an authoritative knowledge graph
- AutoML E2E pipelines, one-size-fits-all solutions, and multi-modal models are critical for enriching the knowledge

Thank You!