Generations of Knowledge Graphs: The Crazy Ideas and The Business

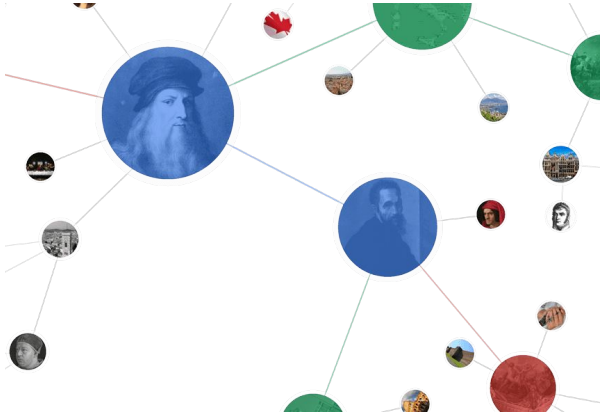
Xin Luna Dong

1/2023

This talk does not represent the company's point of view

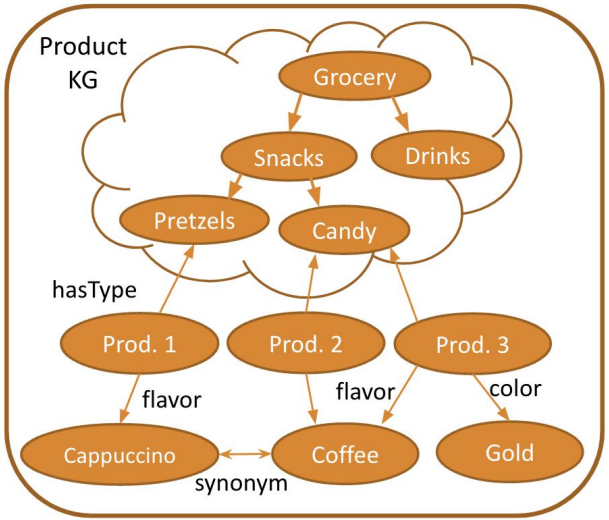
Theme I. Three Generations of Knowledge Graphs

1. Entity-Based KGs



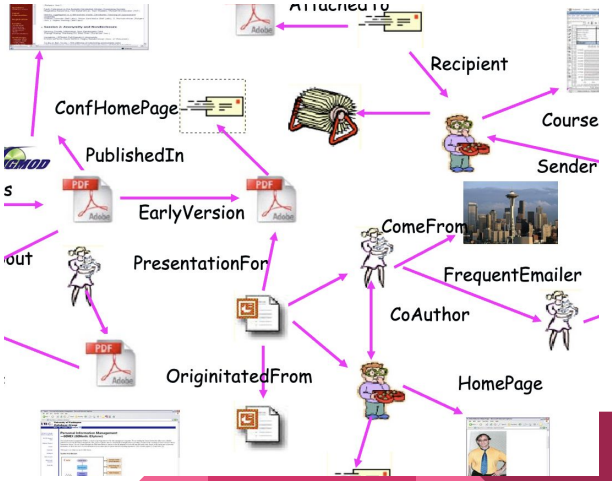
Google Generic KGs

2. Text-Rich KGs



Amazon Product KGs

3. Media-Rich Time-Based KGs



Personal KGs

Theme II. The Recipe from Innovation to Practice

1,000,000,000,000

Feasibility

A prototype, an experiment, to show it is possible

Quality

Production quality, E2E MVP experiences

Repeatability

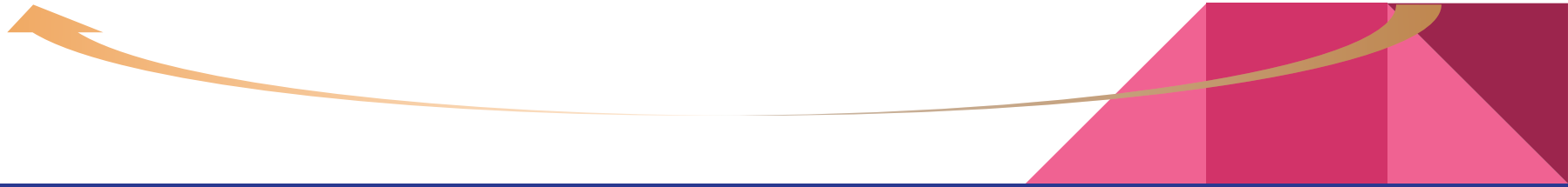
E2E pipelines, automations to allow for extensions to more domains

Scalability

Significant cost reduction for scale ups

Ubiquity

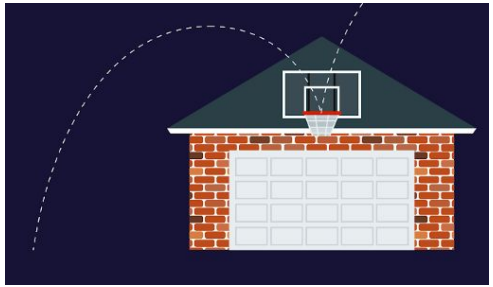
High coverage, long-tail use cases, assumption removal, to next cycle of inventions



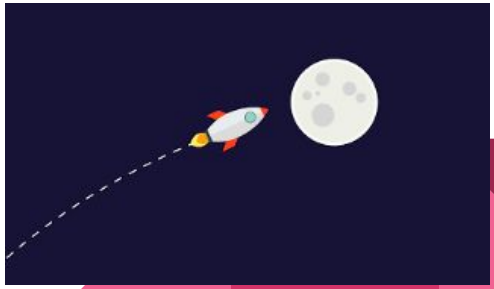
From Roofshots to Moonshots

1,000,000,000,000

Feasibility Quality Repeatability Scalability Ubiquity



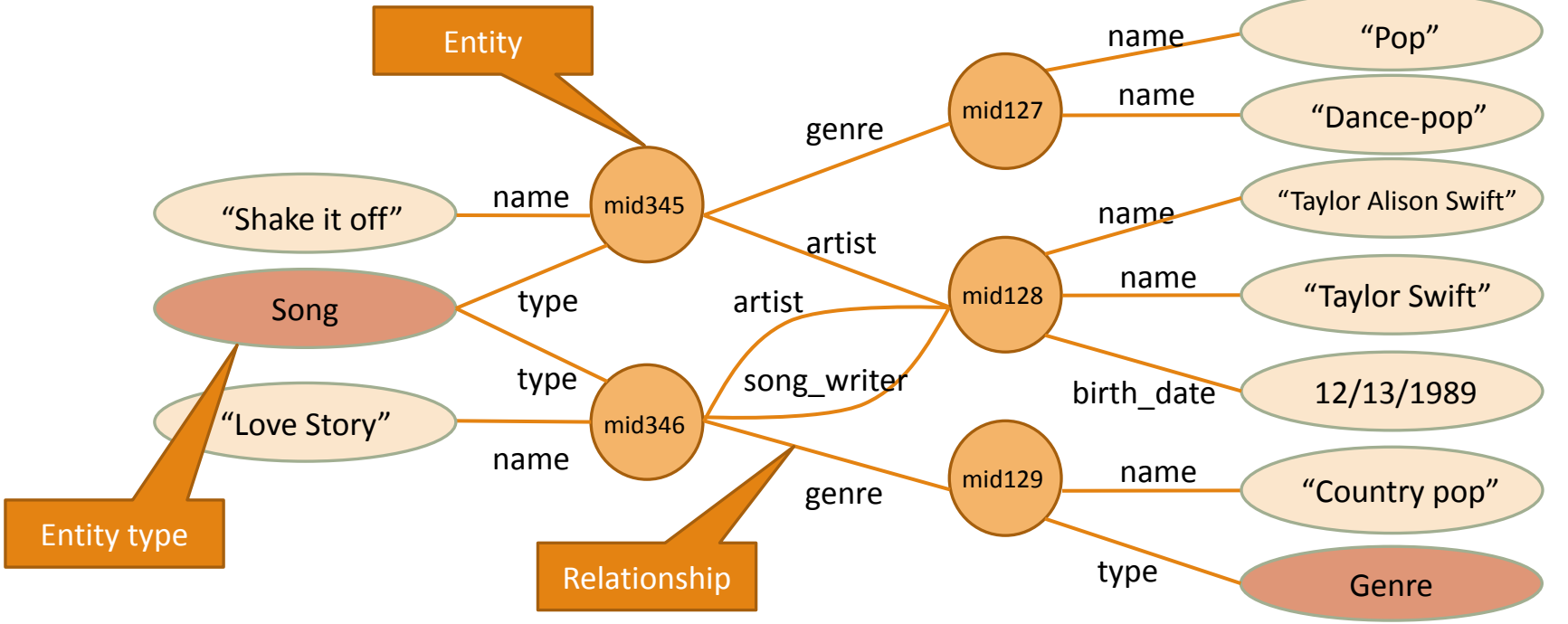
Roofshots



Moonshots

Generation #1: Entity-Based Knowledge Graphs

Entity-Based KG Example



Entity-Based KGs

Characteristics of Entity-Based KGs

- Ontology (types, relationships) manually defined w. clear semantics
- Entities are named-entities, w. no overlap

Crazy Idea

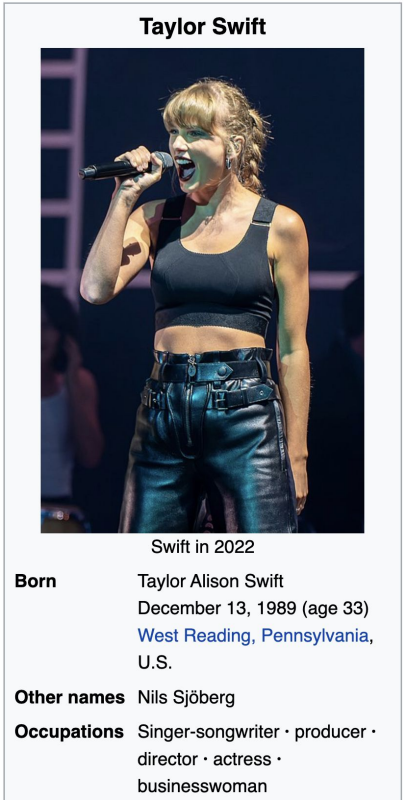
Create a graph of entities and relationships to represent the world

Transforming Wikipedia to A Knowledge Graph

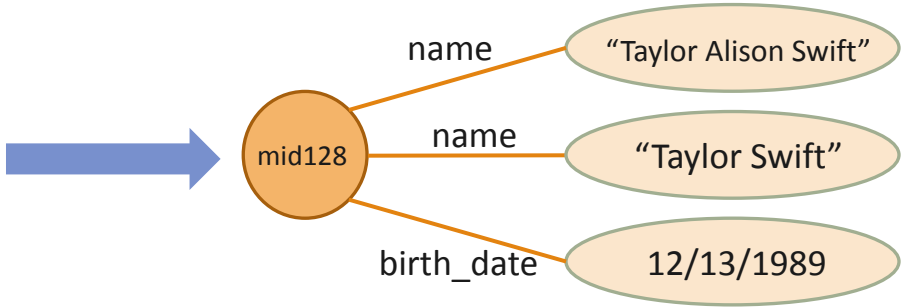
1

Feasibility

A prototype, an experiment, to show it is possible



First pot of gold: Transform Wikipedia Infoboxes to knowledge entities and relationships



Transforming Wikipedia to A Knowledge Graph

1K

Quality

Production quality, E2E MVP experiences

High quality of Wikipedia data guarantees production quality. Common practice in industry

Well-known Examples:

Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum. Yago: A core of semantic knowledge. WWW, 2007.

Integrating Data from Different Sources

1M

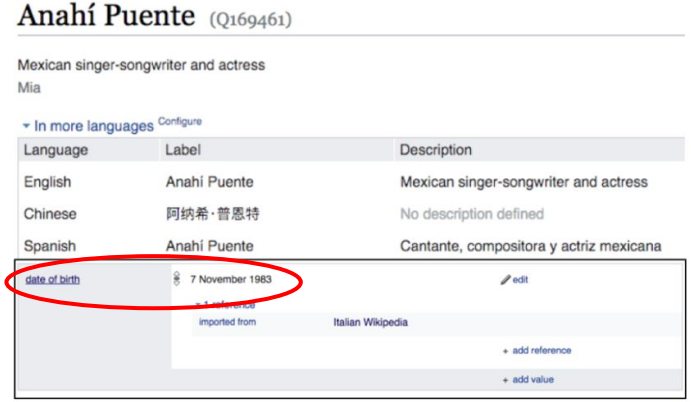
Repeatability

E2E pipelines, automations to allow for extensions to more domains

Tools for more data sources, long-tail domains

IMDB

WikiData



Are they the same person? → **Entity Linkage**

Are "Born" and "date of birth" the same? → **Schema Alignment**

Why "May 14, 1982" vs "7 November 1983"? → **Data fusion**

Heterogeneity, Heterogeneity, Heterogeneity

Integrating Data from Different Sources

1M

Repeatability

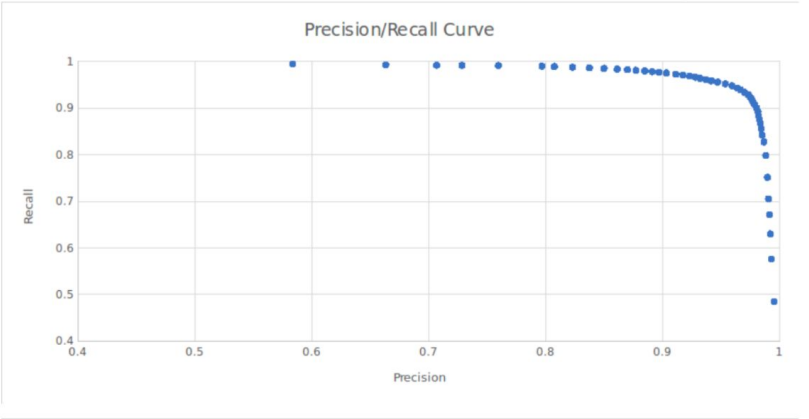
E2E pipelines, automations to allow for extensions to more domains

Biggest Challenge: Entity Linkage

- Random forest on attribute-wise similarity
- Results between Freebase and IMDb movies

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

1.5M labels



Integrating Data from Different Sources

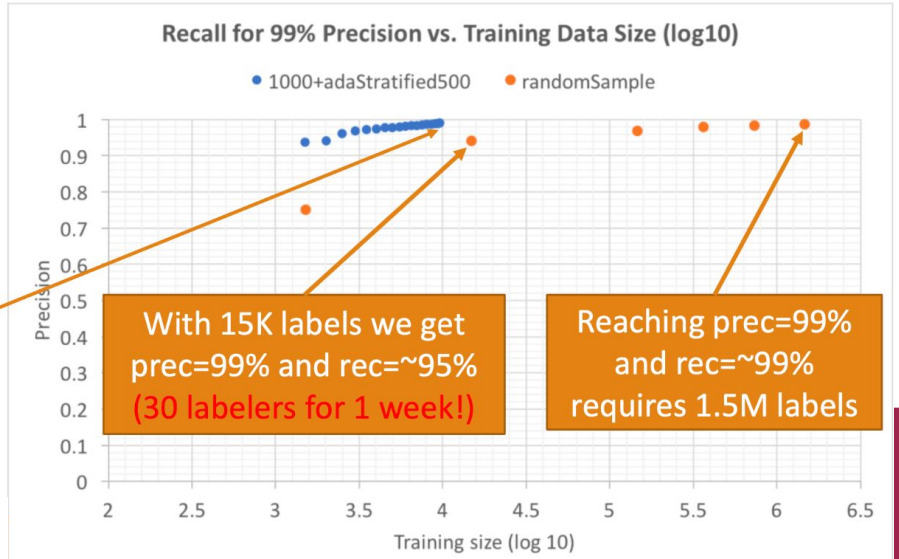
Biggest Challenge: Entity Linkage

- Apply active learning to minimize #labels

Repeatability

E2E pipelines, automations to allow for extensions to more domains

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



Zhu et al., Collective multi-type entity alignment between knowledge graphs, WebConf 2020.
Zhang et al., AutoBlock: A hands-off blocking framework for entity matching, WSDM 2020.
Zhang et al., OpenKI: Integrating open information extraction and knowledge bases with relation inference, NAACL 2019.
Trivedi et al., LinkNBed: Multi-Graph representation learning with entity linkage, ACL 2018.

Extracting Data from Semi-Structured Websites

1B

Scalability

Significant cost reduction for scale ups

Two-stage extraction based on distant supervision

- Identify subject
- Identify (attribute, value) pairs

Lockard et al., Ceres: Distantly supervised relation extraction from the semi-structured web. VLDB, 2018.

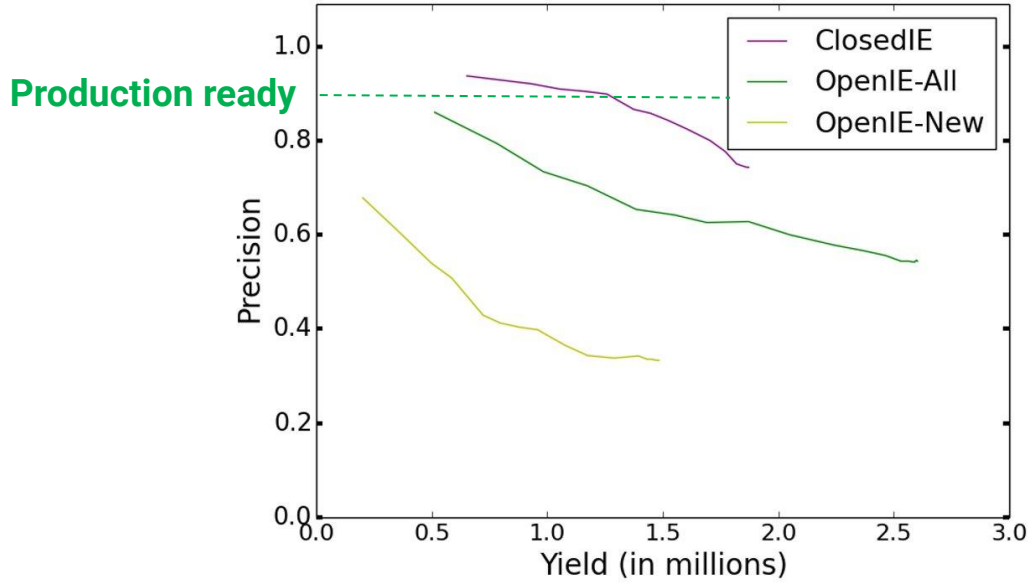
Extracting Data from Semi-Structured Websites

1B

Scalability

Significant cost reduction for scale ups

Latest progress: Extraction on 30+ long-tail movie websites



Lockard et al., Ceres: Distantly supervised relation extraction from the semi-structured web. VLDB, 2018.

Lockard et al., OpenCeres: When open information extraction meets the semi-structured web. NAACL, 2019.

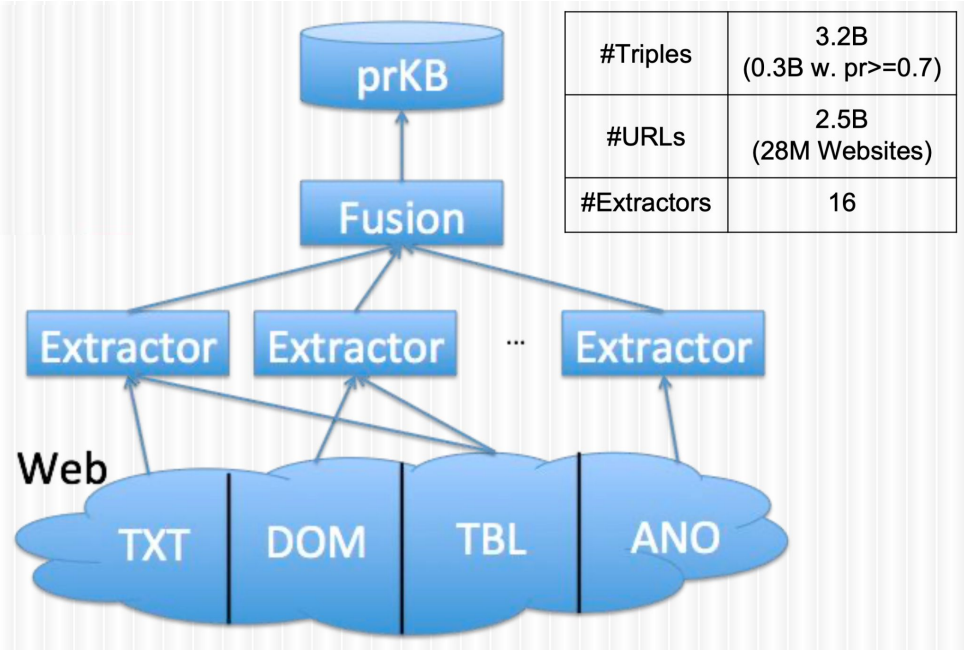
Lockard et al., ZeroShotCeres: Zero-shot relation extraction from semi-structured webpages, ACL 2020.

Web Knowledge Extractions & Fusion

1T

Ubiquity

High coverage, long-tail use cases, assumption removal, to next cycle of inventions



#Triples	3.2B (0.3B w. pr>=0.7)
#URLs	2.5B (28M Websites)
#Extractors	16

Solution: Extract knowledge from different types of web sources, and apply knowledge fusion to remove noises and generate probabilistic knowledge facts

Dong et al., Knowledge Vault: A Web-scale approach to probabilistic knowledge fusion. SIGKDD2014.

Web Knowledge Extractions & Integration

Ubiquity

High coverage, long-tail use cases, assumption removal, to next cycle of inventions

Why NOT in Google Knowledge Graph?

- Didn't reach **production quality**
  - Accuracy=0.7 is far less than the production requirement (Accuracy=0.99)
- Didn't find an **E2E MVP experience**
  - The 0.3B facts (vs. 70B in KG) are very "long-tail" to support meaningful use cases
- But underlying techs applied in long-tail knowledge collection etc.

#Triples	3.2B (0.3B w. pr>=0.7)
#URLs	2.5B (28M Websites)
#Extractors	16

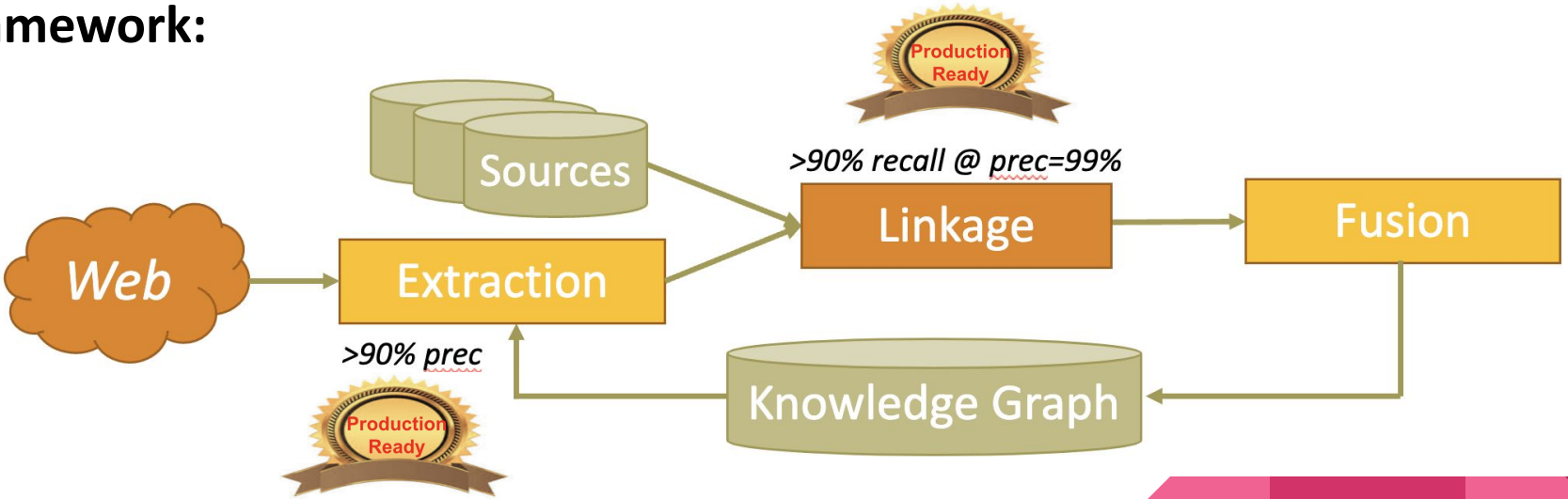
Dong et al., From data fusion to knowledge fusion. PVLDB2014.
Dong et al., Knowledge Vault: A Web-scale approach to probabilistic knowledge fusion. SIGKDD2014.
Dong et al., Knowledge-based trust: Estimating the trustworthiness of web sources. PVLDB2015.

Entity-Based KG Summary

Crazy idea: A graph of entities and relationships to represent the world

Main challenges: **Heterogeneous** data everywhere

Framework:



Crazy idea not successful yet: Web-scale Extraction & Fusion

What About the Product Domain?

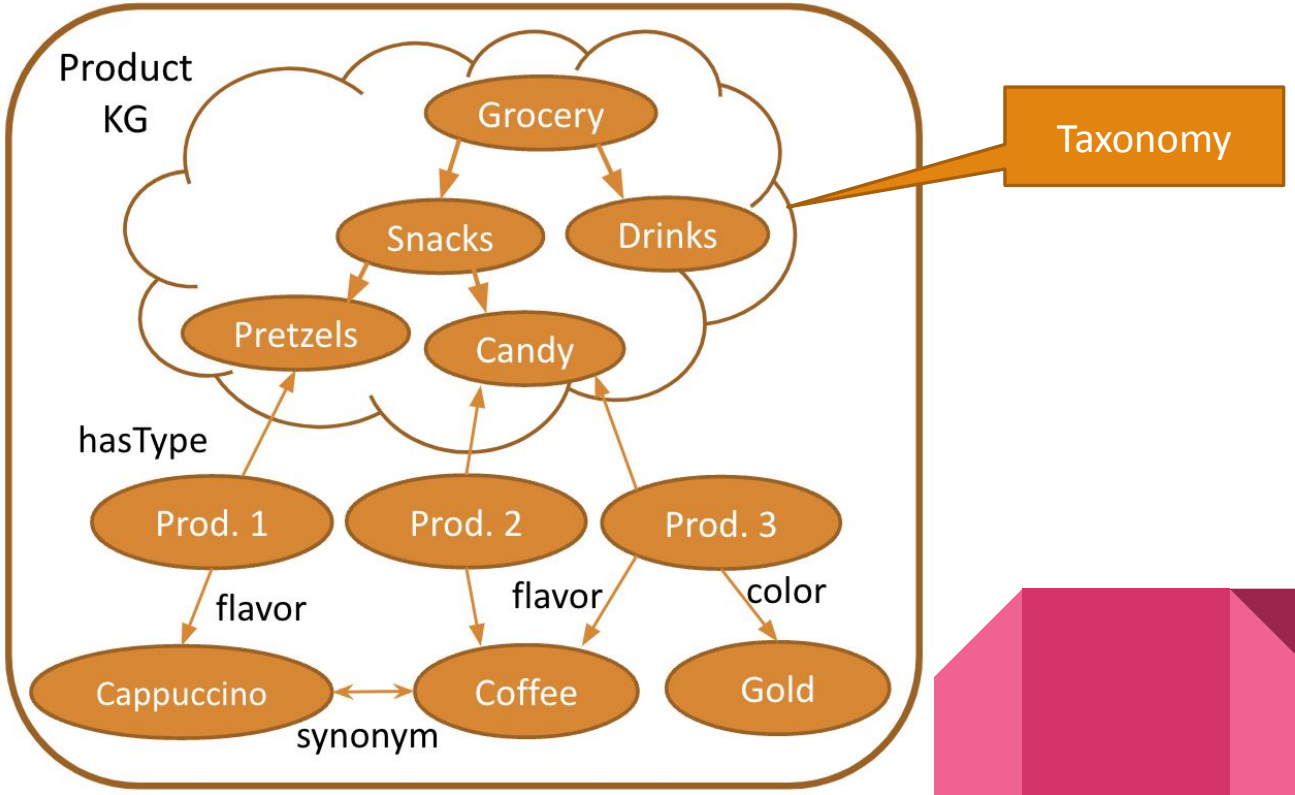
Ubiquity

High coverage, **long-tail use cases**, assumption removal, to next cycle of inventions

Can we do the same for products?

Generation #2: Text-Rich Knowledge Graphs

Text-Rich KG Example



Text-Rich KGs

Characteristics of Text-Rich KGs

- Ontology (types, relationships) very complex with overlaps and ambiguities; E.g., millions of product types
- Entities may not be named-entities, such as products
  E.g., “Onus 2 Colors Highlighter Stick, Shimmer Cream Powder Waterproof Light Face Cosmetics, creamy Self Sharpening Crayon STick Highlighter” vs. “Xin Luna Dong”
- Attribute values are oftentimes texts, with overlaps and ambiguities
  E.g., “Coffee” vs “Cappuccino” as icecream flavors

Crazy Idea:

Finding structure and modeling ambiguity from text sources

Use Case I: Providing Information

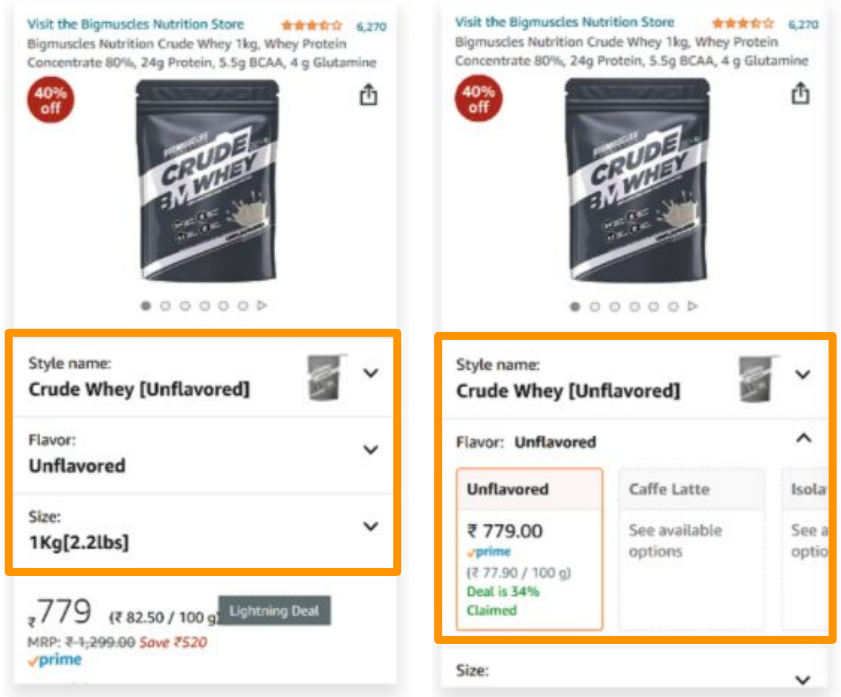
Roll over image to zoom in

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

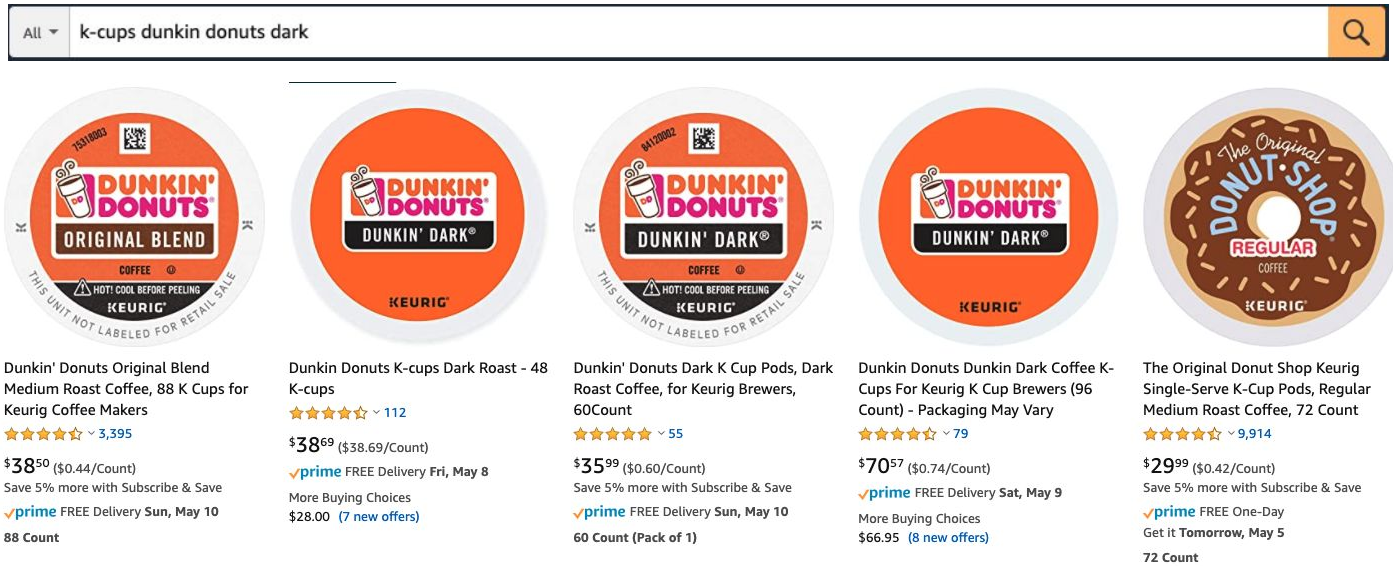
About this item

- 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

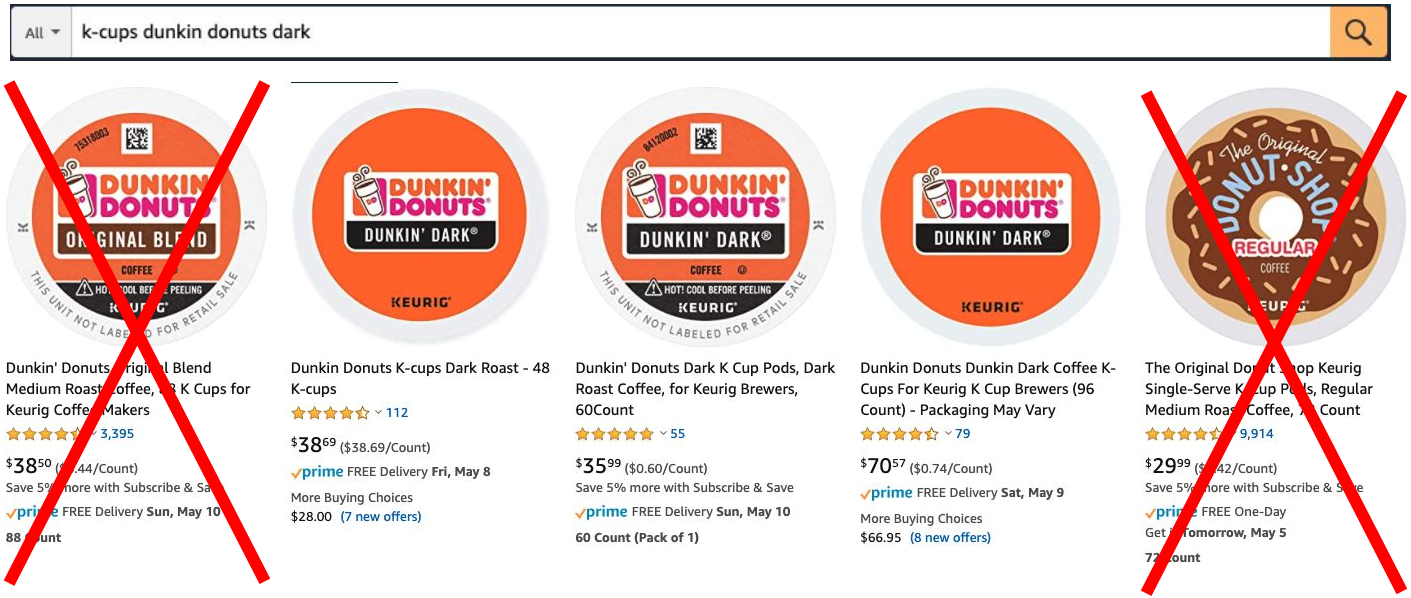
Use Case II: Providing Choices



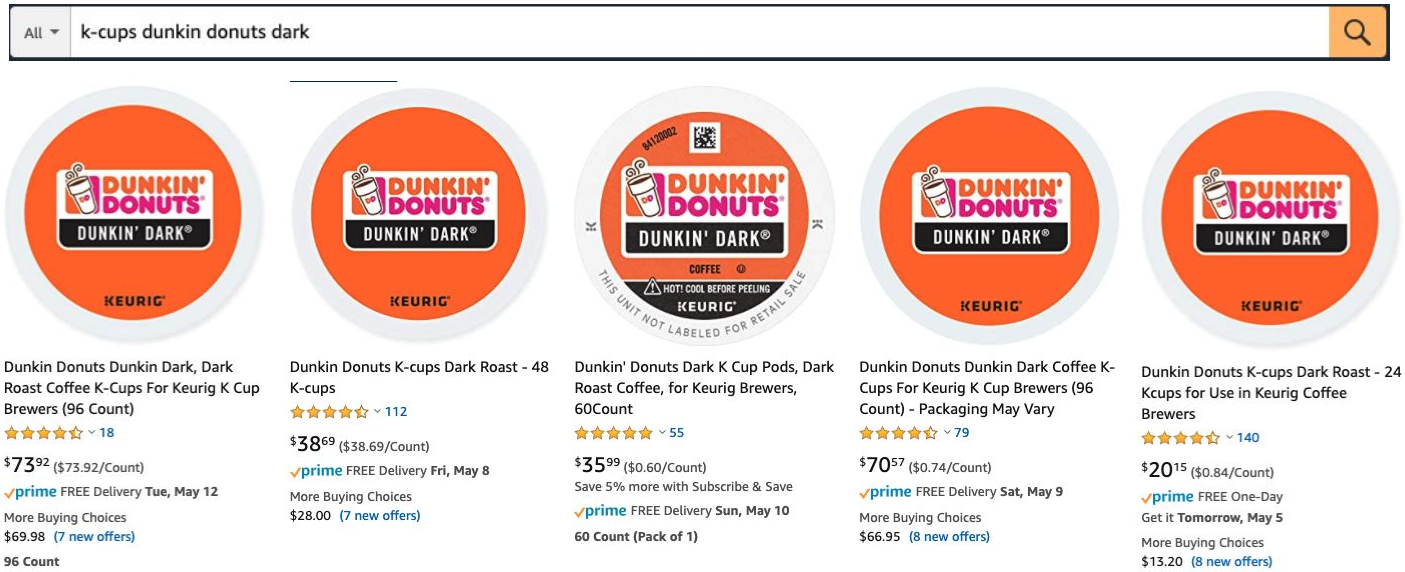
Use Case III: Improving Search



Use Case III: Improving Search



Use Case III: Improving Search



Dunkin Donuts Dunkin Dark, Dark Roast Coffee K-Cups For Keurig K Cup Brewers (96 Count)

18

$73.92 ($73.92/Count)

prime FREE Delivery **Tue, May 12**

More Buying Choices
$69.98 (7 new offers)

96 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)

Dunkin Donuts K-cups Dark Roast - 24 Kcups for Use in Keurig Coffee Brewers

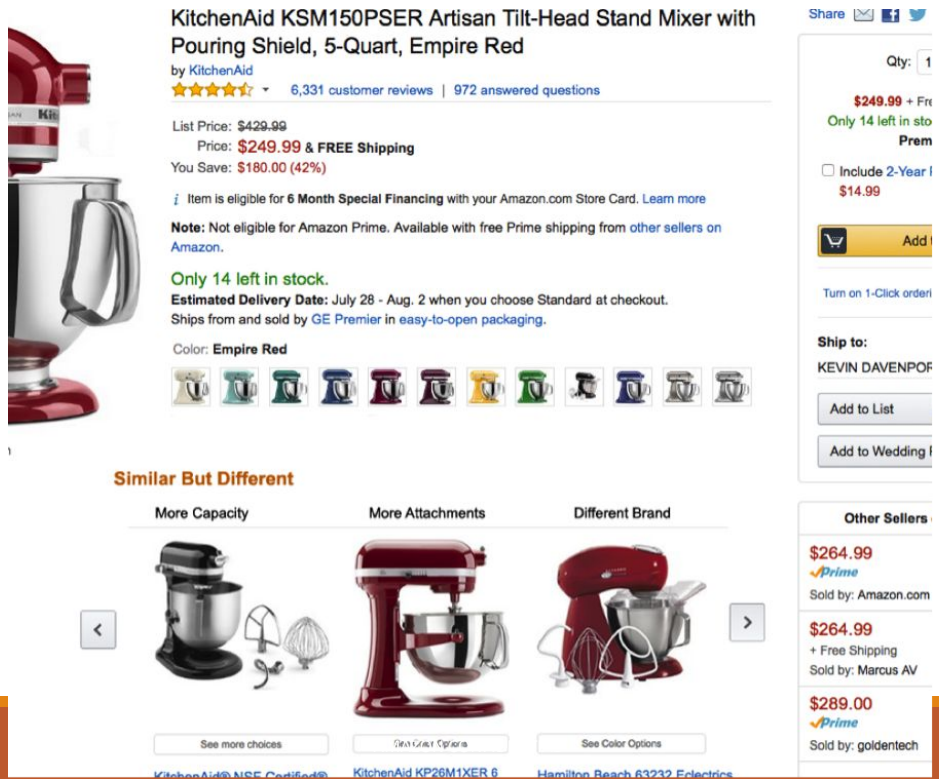
140

$20.15 ($0.84/Count)

prime FREE One-Day
Get it **Tomorrow, May 5**

More Buying Choices
$13.20 (8 new offers)

Use Case IV: Improving Recommendation

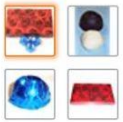


Do We Need Different Techniques?

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.

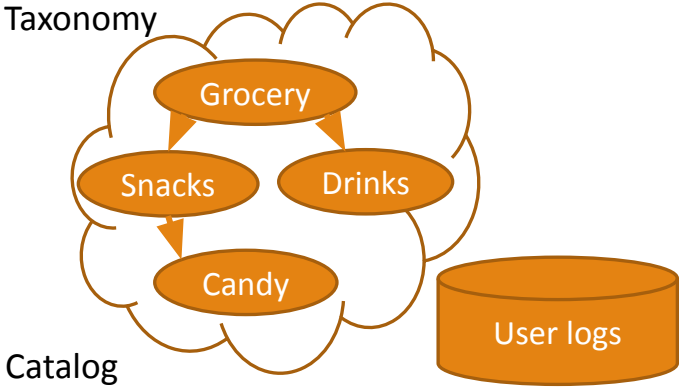
Deliver to Yaqing - Seattle 98109

Qty: 1

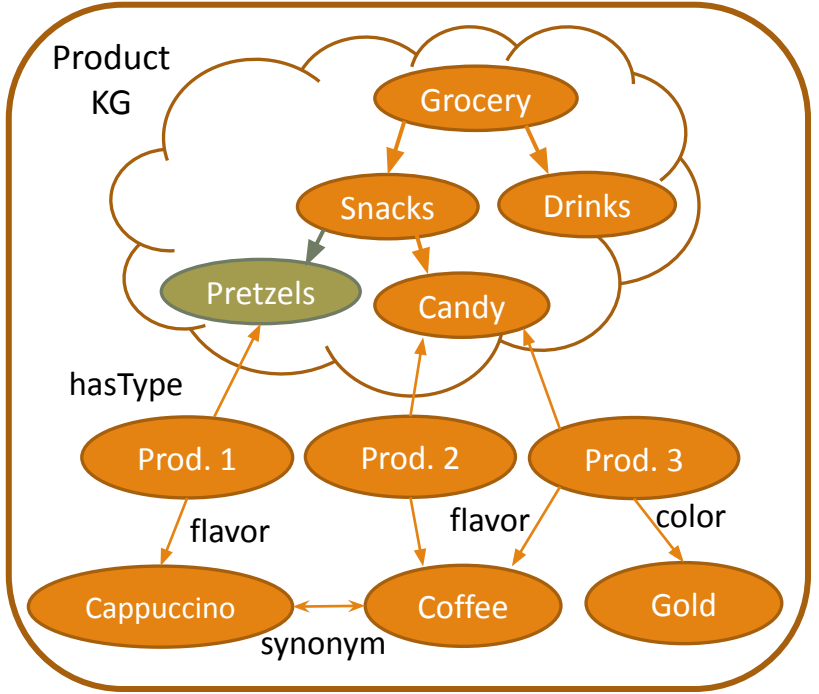
Turn on 1-click ordering

Different challenges: Unstructured and Noisy product data

AutoKnow: Self-Driving Product Knowledge Collection

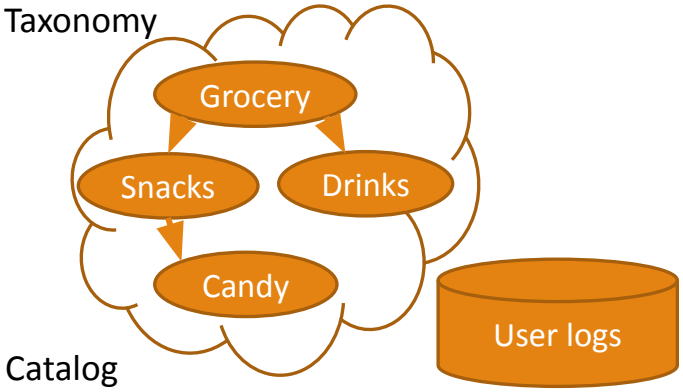


Product	Type	Flavor	Color
Product 1	Snacks	*Cherry*	
Product 2	Candy	?	?
Product 3	Candy	Choc.	Gold

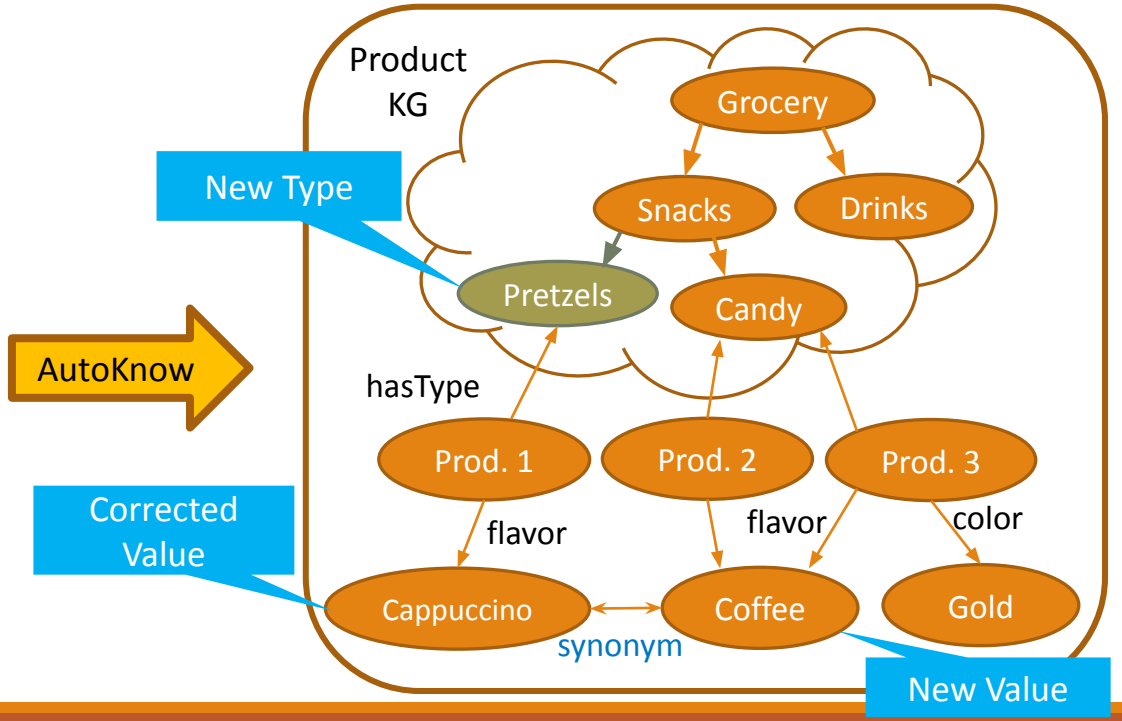


Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.

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.

Extracting Product Attribute Values

Feasibility

A prototype, an experiment, to show it is possible

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		

Generic Knowledge Extraction

Bill Gates founded Microsoft in 1975.

isFounder

Microsoft

Generic KE v.s. Product-Specific KE

Bill Gates founded Microsoft in 1975.

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)

Objects are often not entities

Subject is given

Solution: *OpenTag*— Applying deep tagging to identify structured attributes from product titles, descriptions, and bullets

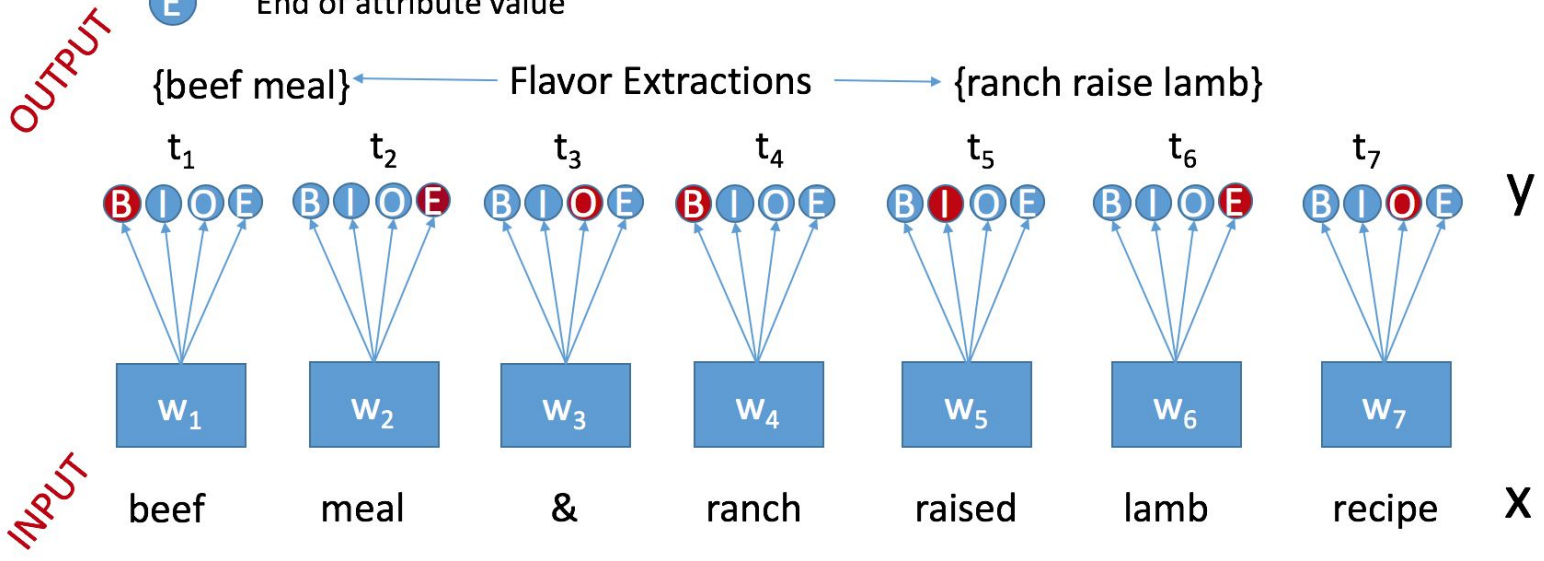
Zheng et al., OpenTag: Open attribute value extraction from product profiles, KDD 2018.

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

CRF

Captures correlations between BIOE tags

Attention

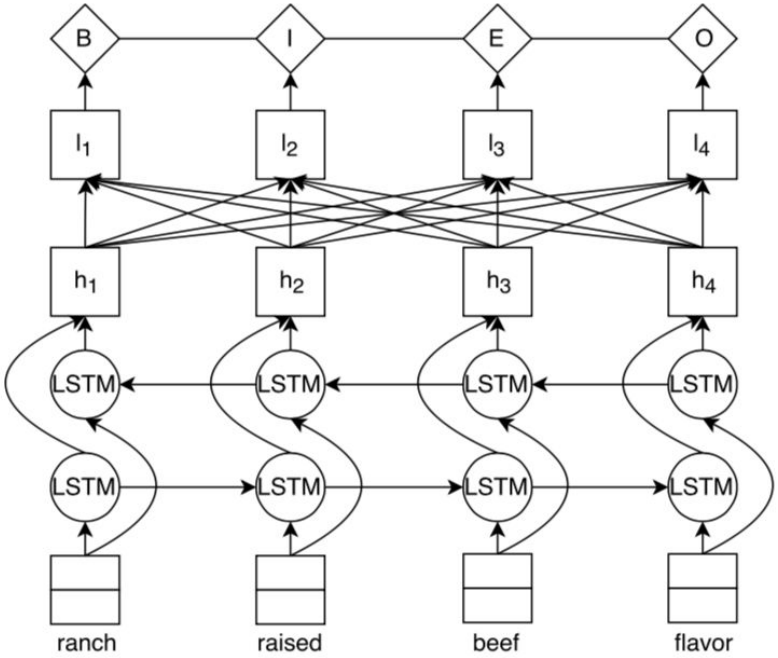
Identifies important terms leading to attribute values

Bi-LSTM

Captures sequence info

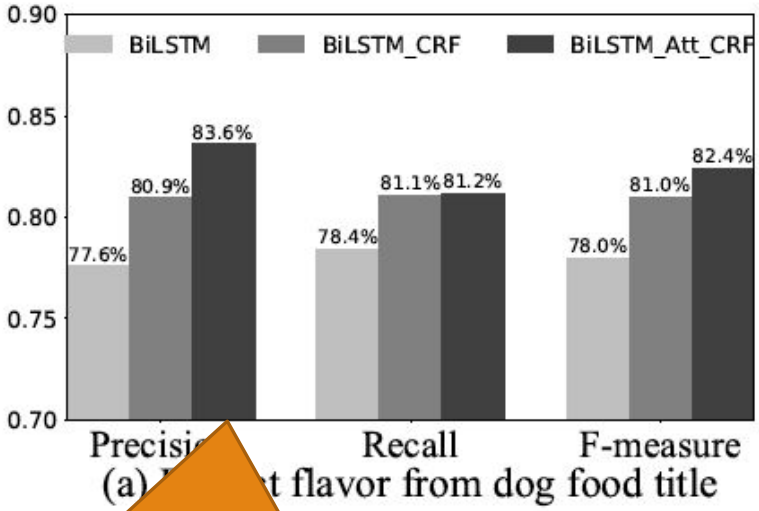
Word Embedding

Captures semantics of each token

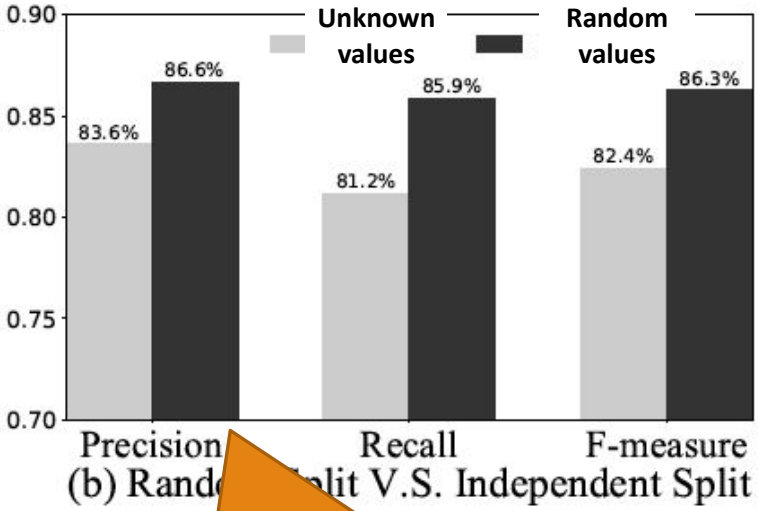


Zheng et al., OpenTag: Open attribute value extraction from product profiles, KDD 2018.

OpenTag Extraction from Product Profiles



(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

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

Building a Product Knowledge Extraction Pipeline

1K

Quality

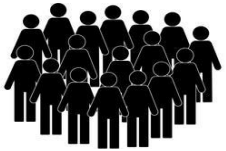
Production quality, E2E MVP experiences

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

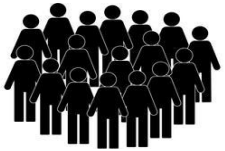
Quality goal: 90% accuracy for product attributes

Building a Product Knowledge Extraction Pipeline

1M

Repeatability

E2E pipelines, automations to allow for extensions to more domains

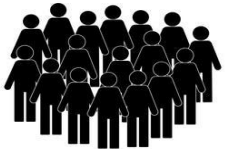


Understand domain and attributes, and generate LOTS OF training data

Train and fine-tune models

OpenTag

Postprocess extraction results to further improve data quality



Pre-publish evaluation as gatekeeper to guarantee high quality data

Building a Product Knowledge Extraction Pipeline

1M

Repeatability

E2E pipelines, automations to allow for extensions to more domains

Automatic Training Data Generation

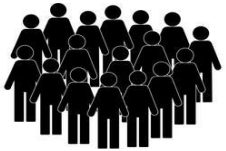
Weak Learning

Train and fine-tune models

OpenTag

Benchmarking

Postprocess extraction results to further improve data quality



Pre-publish evaluation as gatekeeper to guarantee high quality data

Building a Product Knowledge Extraction Pipeline

1M

Repeatability

E2E pipelines, automations to allow for extensions to more domains

Automatic Training Data Generation

Weak Learning

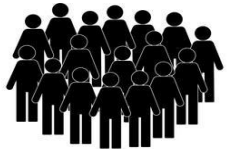
Train and fine-tune models

OpenTag

Benchmarking

Deep Learning Data Cleaning

Postprocess extraction results to further improve data quality



Pre-publish evaluation as gatekeeper to guarantee high quality data

Building a Product Knowledge Extraction Pipeline

1M

Repeatability

E2E pipelines, automations to allow for extensions to more domains

Train and fine-tune models

Automatic Training Data Generation

Weak Learning

OpenTag

Benchmarking

Deep Learning Data Cleaning

Postprocess extraction results to further improve data quality

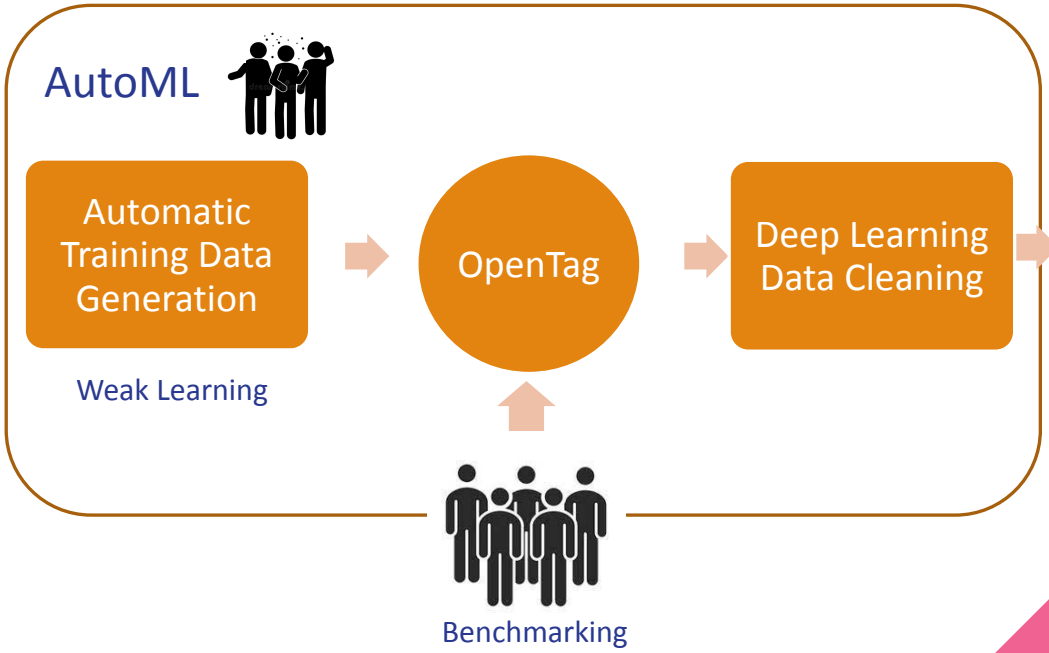
Scale-up pre-publish evaluation w. lower labeling needs

Building a Product Knowledge Extraction Pipeline

1M

Repeatability

E2E pipelines, automations to allow for extensions to more domains



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

4.3 out of 5 stars 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.

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

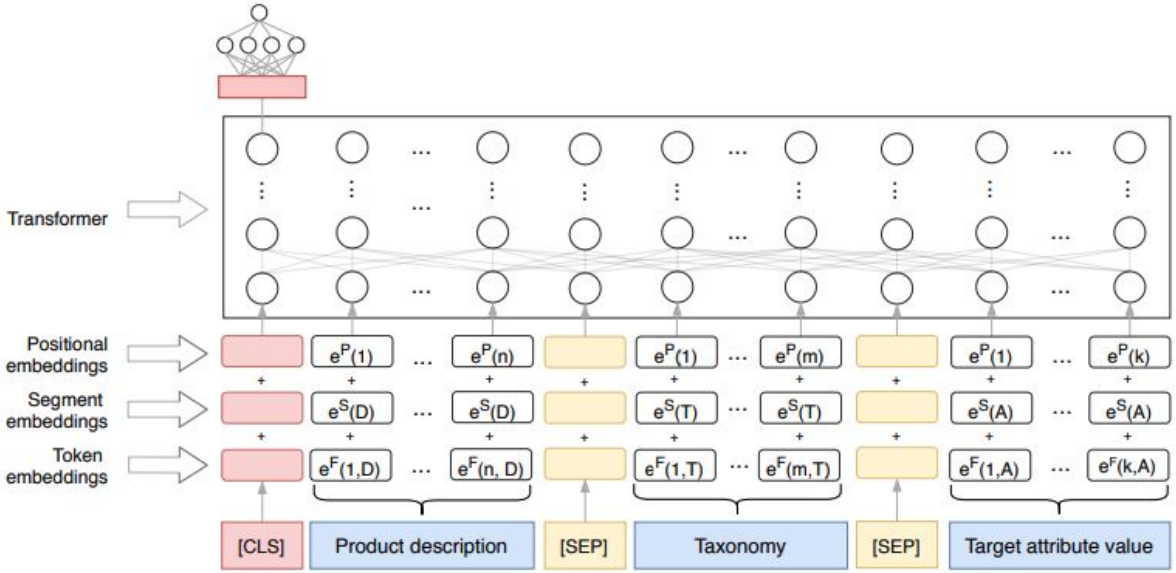
Size: **6 Pound**

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

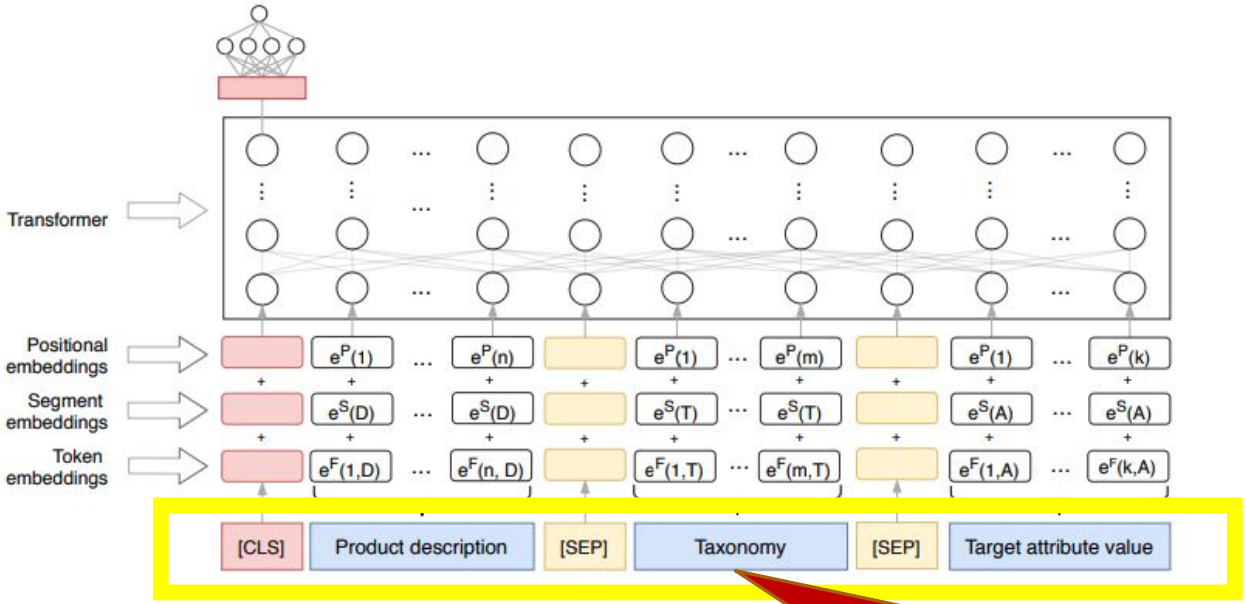
- Love of Candy's huge selection of bulk candy now includes Premium Mint Chocolate Lentils in a variety of bold & striking colors. Available in small to large sizes ranging from 1 to 10 lb bags. These beautiful chocolate morsels feature gourmet, dairy free dark chocolate coated in a crispy and crunchy 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"?

Transformer-Based Anomaly Detection



Transformer-Based Anomaly Detection



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)

Scaling Up Product Knowledge Extraction

1B

Scalability

Significant cost reduction for scale ups

Millions of categories

Thousands of attributes

Hundreds of languages

Solution: One-size-fits-all models

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.
Yan et al., AdaTag: Multi-Attribute Value Extraction from Product Profiles with Adaptive Decoding, ACL 2021.

Scaling Up for 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

Scaling Up for Millions of Categories

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

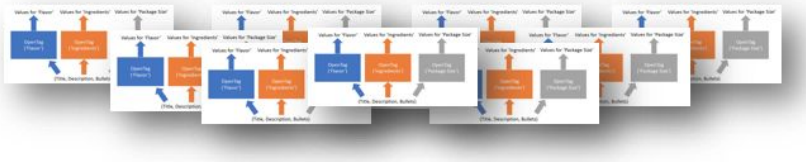
Samsung UN58RU7100FXZA Flat 58-Inch 4K UHD 7 Series Ultra HD Smart TV with HDR and Alexa Compatibility (2019 Model)

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

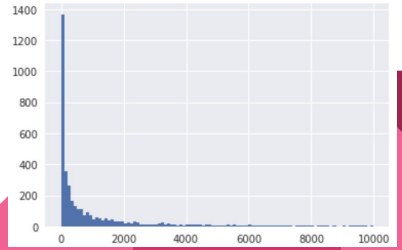
Caribou Coffee Caribou Blend, Medium Roast Ground Coffee, 20 Ounce Bag, Rainforest Alliance Certified

Option 2. Train a model for each category?

Store/orchestrate 100K+ OpenTag



Most categories are very sparse



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

Scaling Up for Millions of Categories

1B

Scalability

Significant cost reduction for scale ups

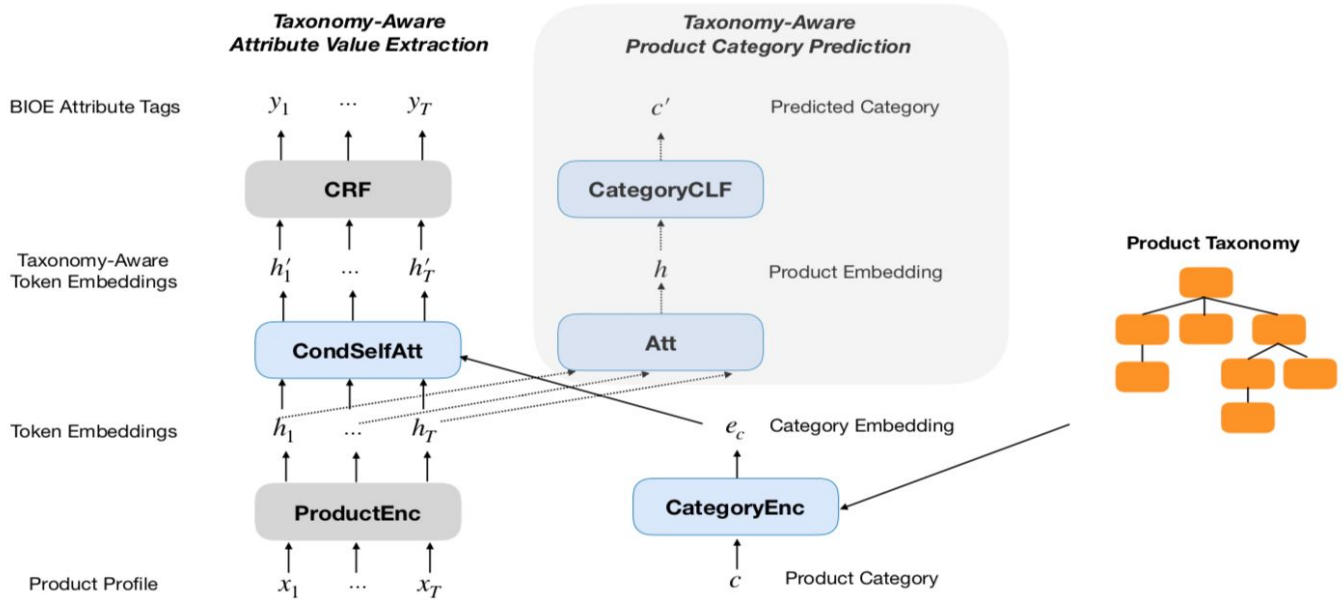


Figure 2: Our TXtract architecture for hierarchical multi-task learning.

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

Scaling Up for Millions of Categories

1B

Scalability

Significant cost reduction for scale ups

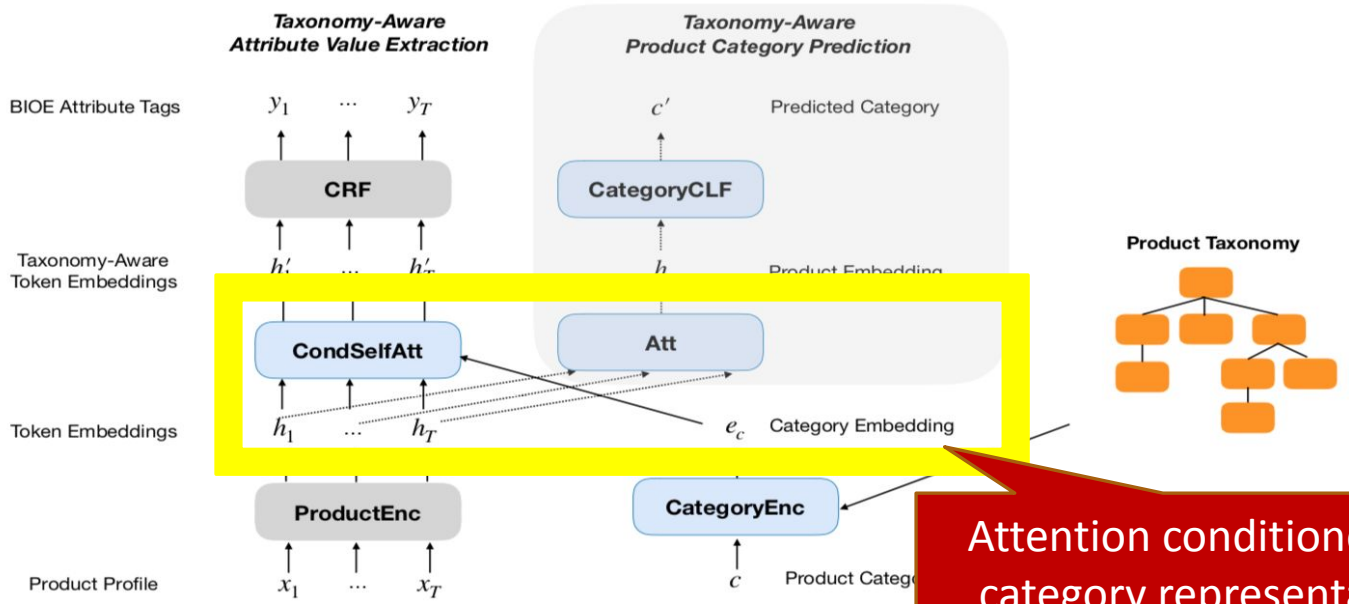


Figure 2: Our TXtract architecture for hierarchical multi-task learning.

Scaling Up for Millions of Categories

1B

Scalability

Significant cost reduction for scale ups

Multi-task Learning

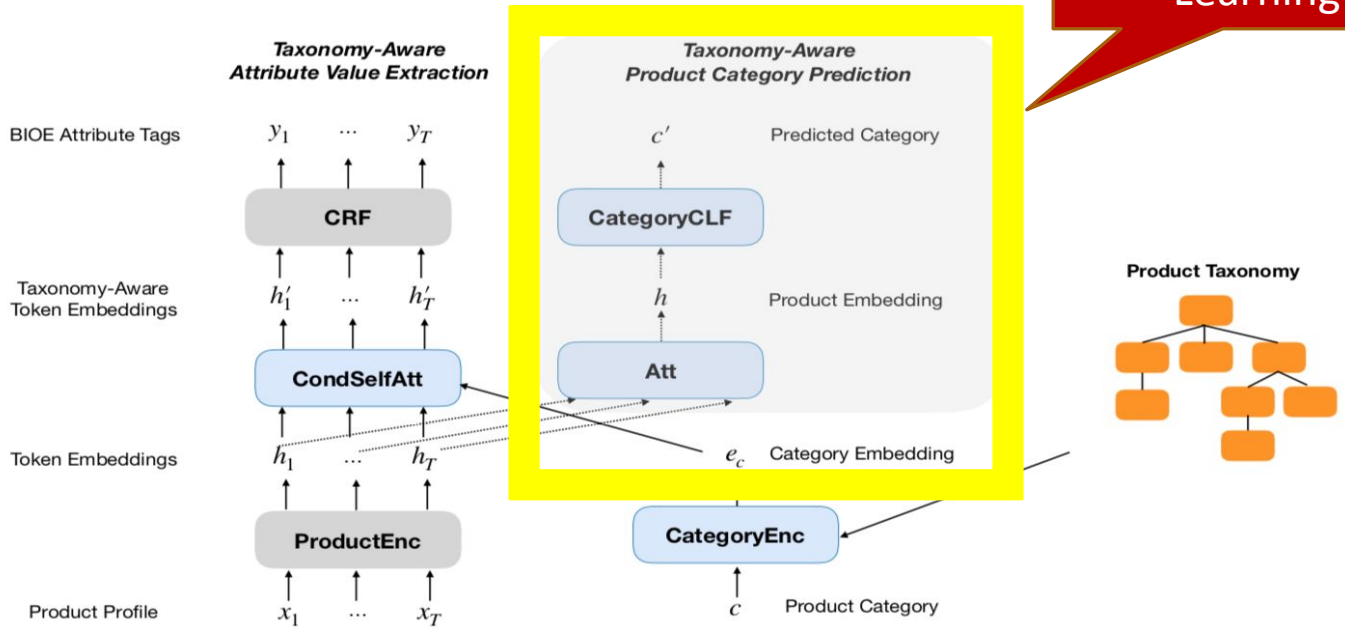


Figure 2: Our TXtract architecture for hierarchical multi-task learning.

Scaling Up for 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.

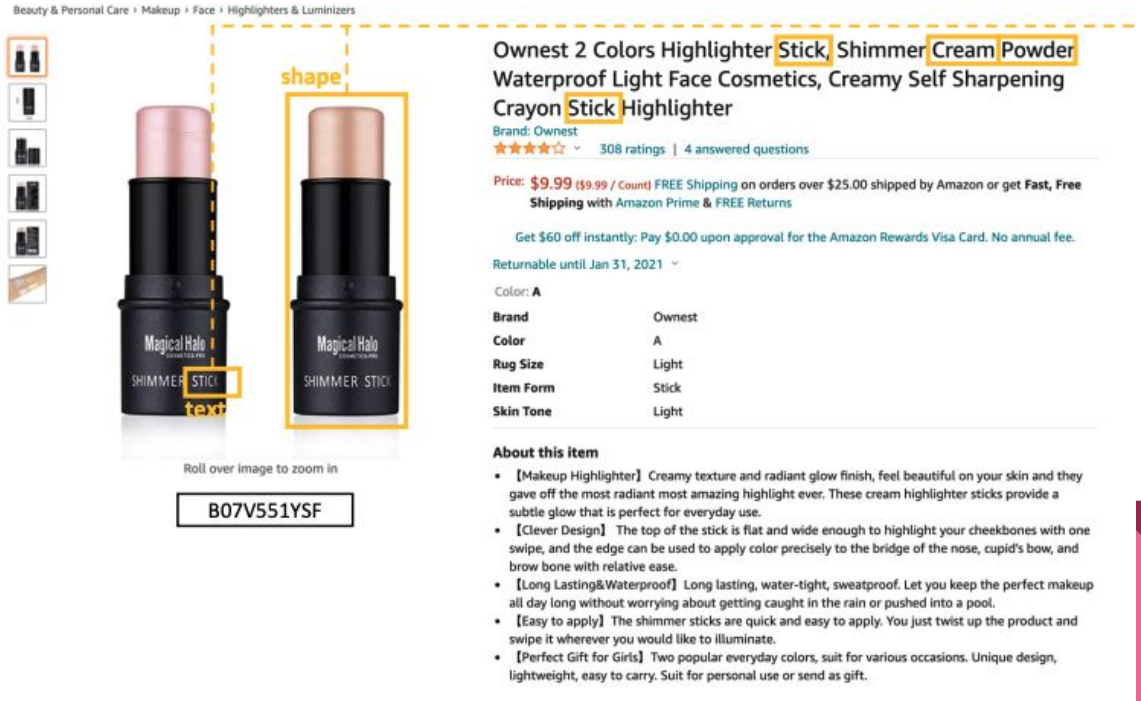
Product Knowl. Extraction from Broader Sources

1T

Ubiquity

High coverage, long-tail use cases, assumption removal, to next cycle of inventions

Product knowledge extraction from images, reviews, etc.



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

Product Knowl. Extraction from Broader Sources

1T

Ubiquity

High coverage, long-tail use cases, assumption removal, to next cycle of inventions

Multi-modal knowledge extraction

Text still plays the most important role

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

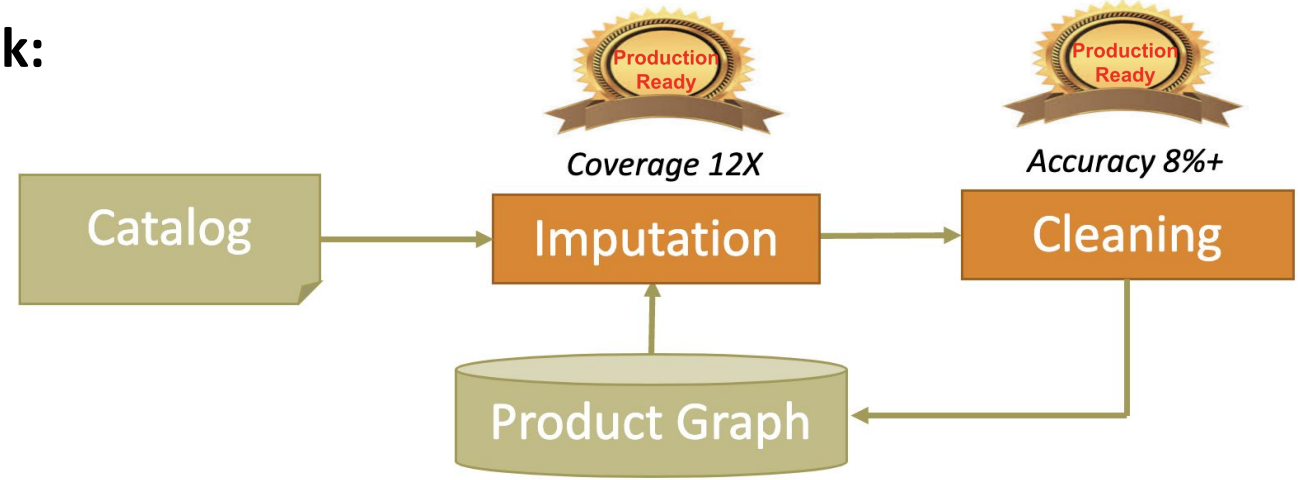
Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.

Text-Rich KG Summary

Crazy idea: Finding structure and modeling ambiguity from text-rich sources

Main challenges: **Sparse & Noisy** structured data everywhere

Framework:



Unmentioned crazy idea: Automatic taxonomy extraction & construction

Mao et al., Octet: Online catalog taxonomy enrichment with self-supervision. SigKDD, 2020.
Zhang et al., Minimally supervised structure-rich text categorization via learning on text-rich networks.
Zhang et al., OA-Mine: Open-World attribute mining for e-Commerce products with weak supervision. Webconf, 2022

Generation #3: Media-Rich Time-Based KGs

Background: Virtual Intelligent Assistant

Respond to commands

“Hey Siri, set a timer to 7pm”

“Ok, added to today’s reminders”

Background: Virtual Intelligent Assistant

Control devices

“Hey Alexa, turn off bedroom lights”

Background: Virtual Intelligent Assistant

Provide information

“Hey, Google, when did winter start?”

“Winter started on Wed, December 21”

Meta's Assistant

Empowering connection to people and experiences in your life

Meta Quest 2

"Hey Facebook" (double press the button on your controller)

"Who's online?"--meet up with friends

"Open Beat Saber"--jump straight in the game, and more.

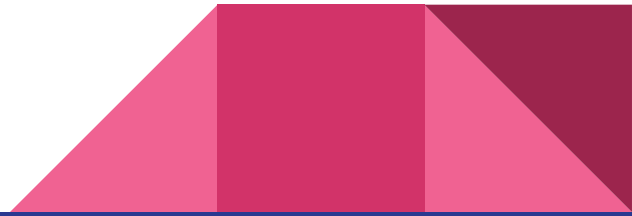
Ray-Ban Stories

"Hey Facebook, take a picture" --capture moments hands-free

"Hey Facebook"--call friends on Messenger, manage device settings, and more.

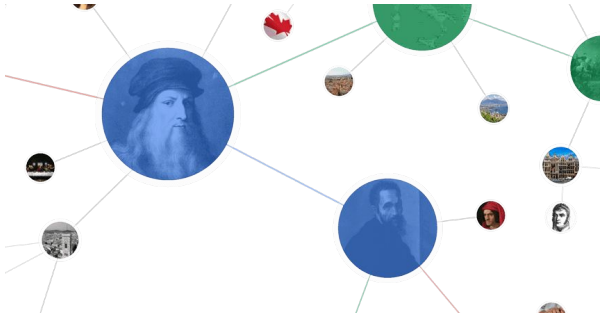
What is An Ideal Virtual Intelligent Assistant?

An *intelligent assistant* should be an agent that **knows you and the world**, can **receive your requests** or **predict your needs**, and provide you the **right services at the right time** with your permission.

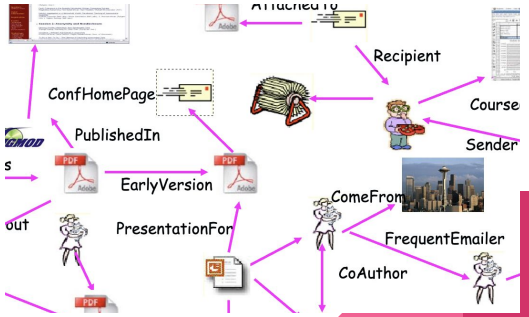


What is An Ideal Virtual Intelligent Assistant?

An *intelligent assistant* should be an agent that **knows you and the world**, can **receive your requests** or **predict your needs**, and provide you the **right services at the right time** with your permission.



Public KGs



Personal KGs

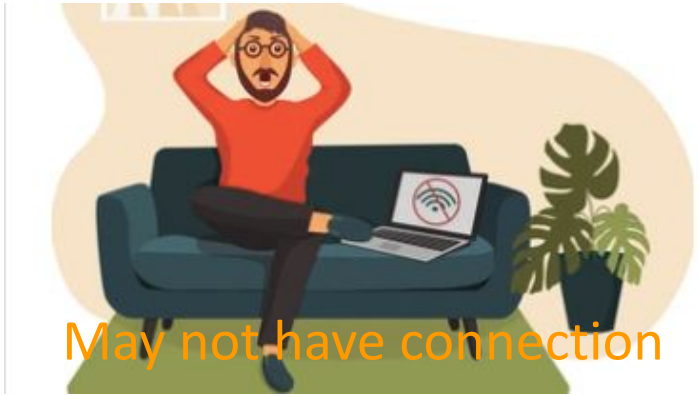
Evolution of Intelligent Assistant

Chatbot
Text input

Voice Asst
Voice input

AR/VR Asst
Voice + Visual + Context

What Is Different for An AR/VR Assistant?



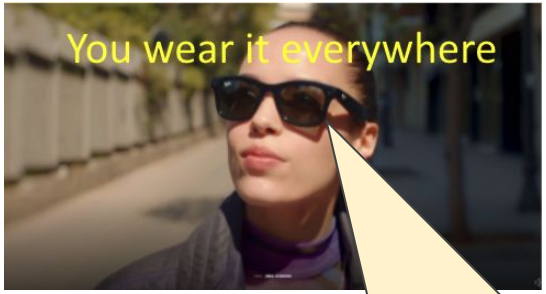
From Voice-Only to Multi-Modal

“How tall is Empire State Building?”

“What’s the name of this building and how tall is it?”

From Context-Agnostic to Context-Aware

"Show my shopping list"



"Remember to buy apples and bananas at the grocery store around the corner"

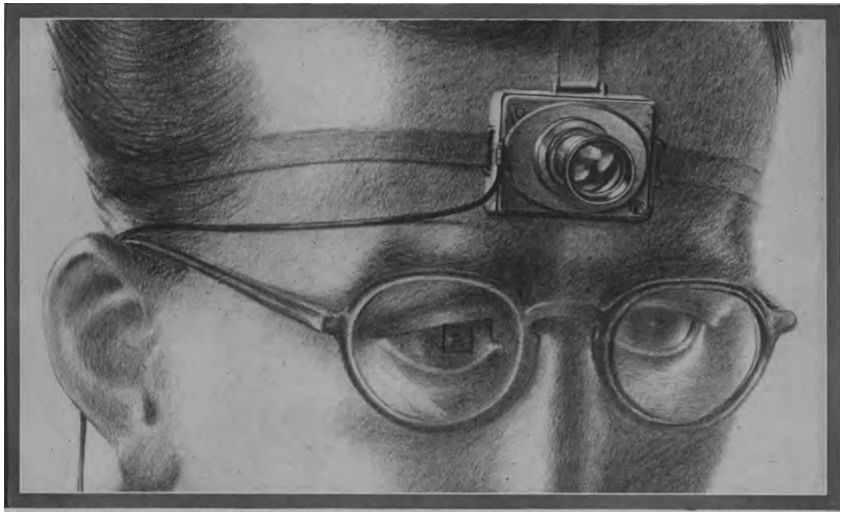
From Reactive to Proactive (Recommendation!)

“What’s the weather today?”

“Today is sunny, 70 degree. Would you like to play your favorite morning music?”

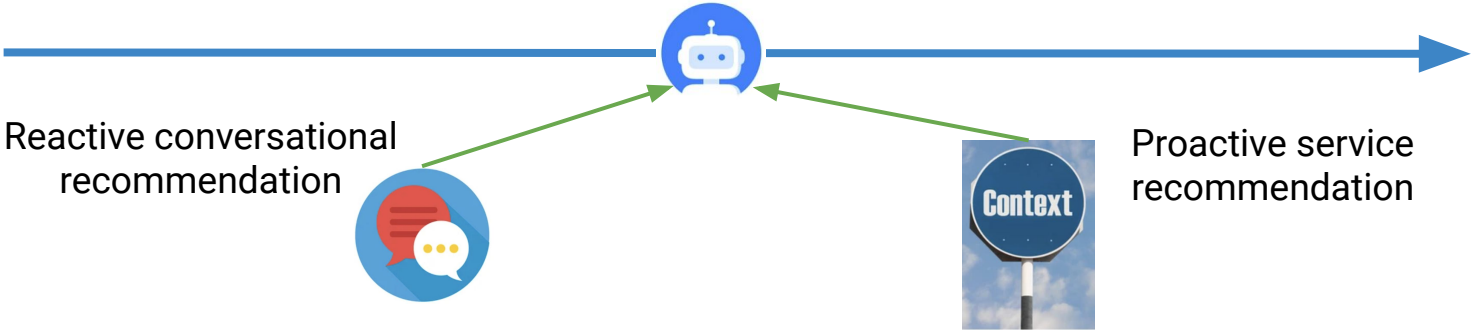
From Server-Side to On-Device

Two Sides of One Coin (1): Great Vehicle for Life Recording



MEMEX (MEMory & EXpansion)
by Vannevar Bush (1945)

Two Sides of One Coin (2): Great Vehicle for Personal Assistant Recommendation



Personal knowledge
(preferences, routines, etc.)

Personal KG and Memovoir

Utility

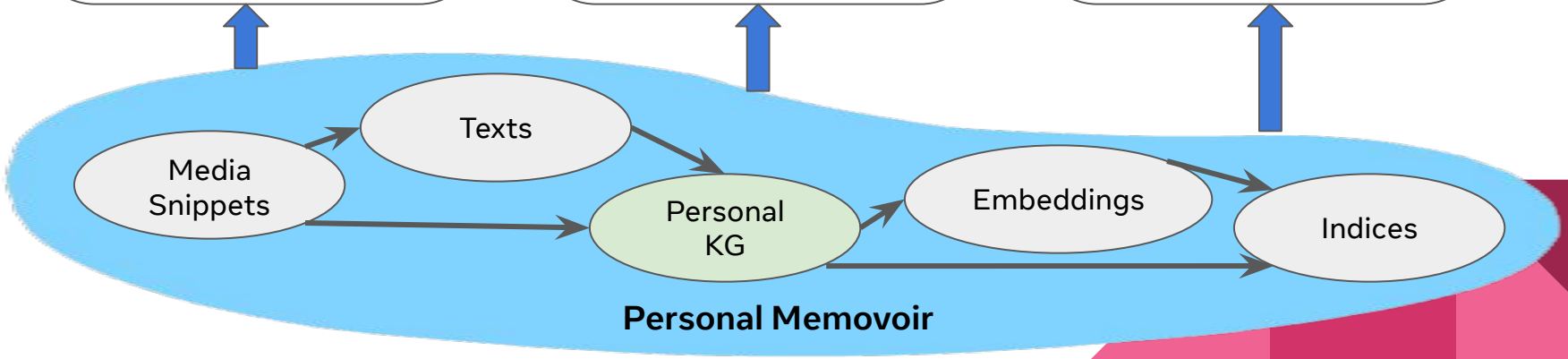
Where did I put my key?

I must have seen this lady before but when and where?

Memoir

Inspiration

At Lyon you can visit the statue of Saint-Exupéry with the Little Prince by his side. You read that book in 2018 and loved it.



Time-Based KGs & Media-Rich Memovoir

Characteristics of a Personal KG & Memovoir

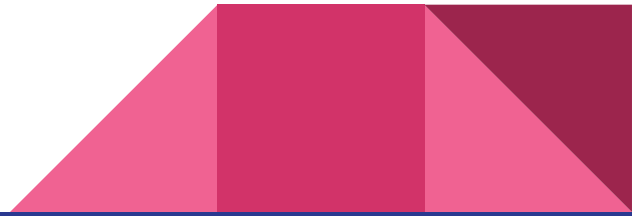
- Rich audio/video, associated w. time, location, etc.
- Modeling
  - entities at an abstract level; e.g., latte art coffee (from different stores), my key for front door
  - activities; e.g., dancing, watching ballet
- Historical–each activity is associated with a timestamp

Crazy Idea

Trace and abstract one's life from rich audios/videos, and use it for life experience Q&A, life journal creation, and life recommendation.

Research Problems for Personal KG and Memovoir

- How to record one's life with hardware (memory, battery) constraints?
- How to extract personal knowledge from the recordings?
- What is the best frequency, granularity, and domain richness to capture one's life?
- How to leverage the Personal KG and Memovoir for utility, memoir, and inspiration applications?
- How to leverage Personal KGs to best understand contexts?
- How to combine public and personal KGs for context-aware recommendation?



Take-Aways I. 3 Generations of KGs

1. Entity-Based KGs

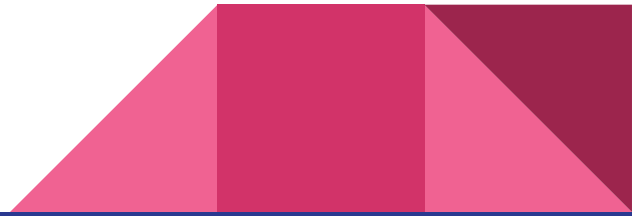
Resolving heterogeneity with **entity linkage** and **web knowledge extraction**

2. Text-Rich KGs

Extractions and **cleanings** from sparse and noisy source data, and handling semantics ambiguities

3. Media-Rich Time-Based KGs

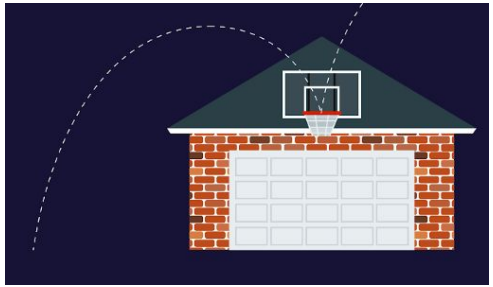
Many new challenges for knowledge collection and applications



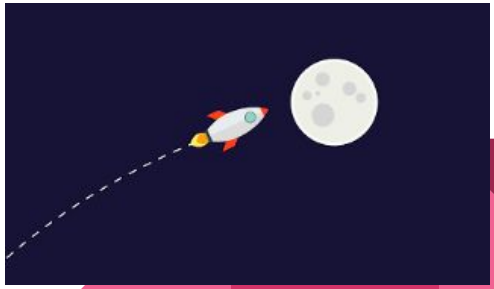
Take-Away 2. From Roofshots to Moonshots

1,000,000,000,000

Feasibility Quality Repeatability Scalability Ubiquity



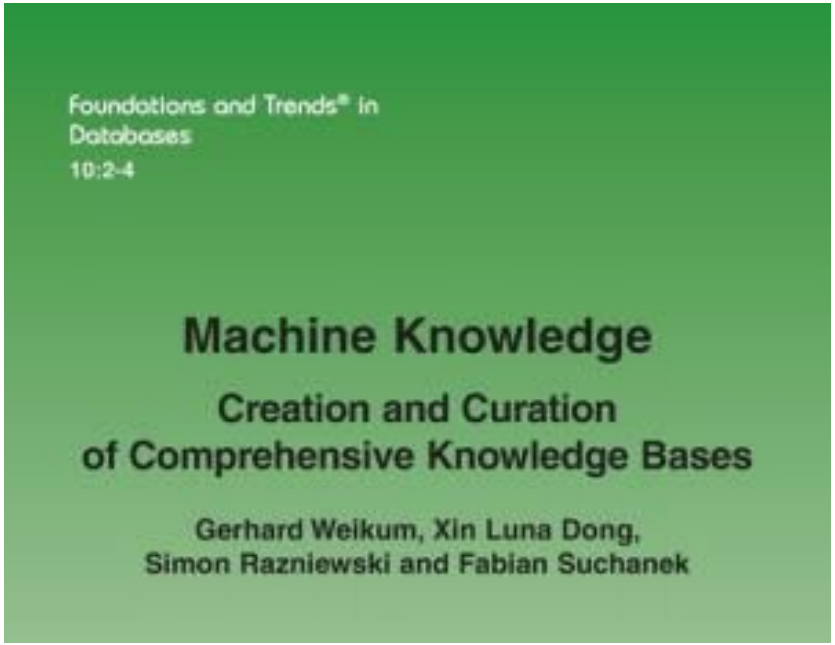
Roofshots



Moonshots

Shameless Advertisements

Book (2021)



Two benchmark datasets

- DI2KG Challenge: http://di2kg.dia.uniroma3.it/#challenge
- Extended SWDE benchmark: https://homes.cs.washington.edu/lockardc/expanded_swde.html

Thank You

Q&A?