Integrating Product Data from the Semantic Web using Deep Learning Techniques

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Prof. Dr. Christian Bizer

Ingénierie des Connaissances (IC@PFIA 2022)
Saint-Étienne, France

June 29, 2022

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Hello

- **Prof. Dr. Christian Bizer**
- Chair of Web-based Information Systems @ University of Mannheim

- Research Areas:
  - Large-scale data integration
  - Information extraction from semi-structured sources
  - Knowledge base construction
  - Analysis of the adoption of semantic web technologies

- eMail: christian.bizer@uni-mannheim.de

Integrating Product Data from the Semantic Web using Deep Learning Techniques

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Use Cases
  1. Price comparison in e-commerce
  2. Complementing product knowledge graphs
- Trends
  1. Wide adoption of schema.org annotations on the Web
  2. Success of deep learning techniques for entity matching

1. Product Data on the Semantic Web

Schema.org Annotations in HTML Pages


schema.org Search

Home Schemas Documentation

Thing > Organization > LocalBusiness

A particular physical business or branch of an organization. Examples of LocalBusiness include a restaurant, a particular branch of a restaurant chain, a branch of a bank, a medical practice, a club, a bowling alley, etc.

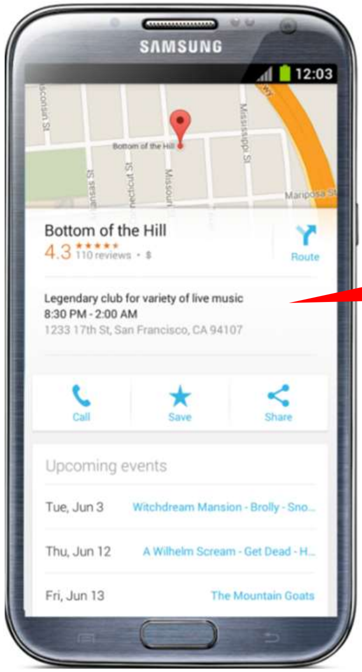
Property	Expected Type	Description
Properties from Thing		
description	Text	A short description of the item.
image	URL	URL of an image of the item.
name	Text	The name of the item.
url	URL	URL of the item.
Properties from Place		
address	PostalAddress	Physical address of the item.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
containedIn	Place	The basic containment relation between places.

- ask website owners since 2011 to annotate data within pages for enriching search results
- 675 Types: Event, Place, Local Business, **Product**, Review, Person
- Encoding: Microdata, RDFa, JSON-LD

Schema.org Product Annotations within HTML using the Microdata Syntax

```
<div  itemtype="http://schema.org/Product">
    <h1 itemprop="name">Coleman Signature Instant Dome 7 Orange</h1>
   Item # <span itemprop="gtin12">200734246807</span>
   <img itemprop="image" src="../20073424680.jpg">
  <p itemprop="description">The Coleman Signature 7-Person Instant Dome
       tent makes camping easy so you can enjoy every moment of your
       outdoor adventure. In about a minute you can have ...</p>
  <span itemprop="review" itemtype="http://schema.org/Review">
        <span itemprop="ratingValue">5</span>
        <span itemprop="reviewBody">Great waterproof tent!</span>
    </span>
</div>
```

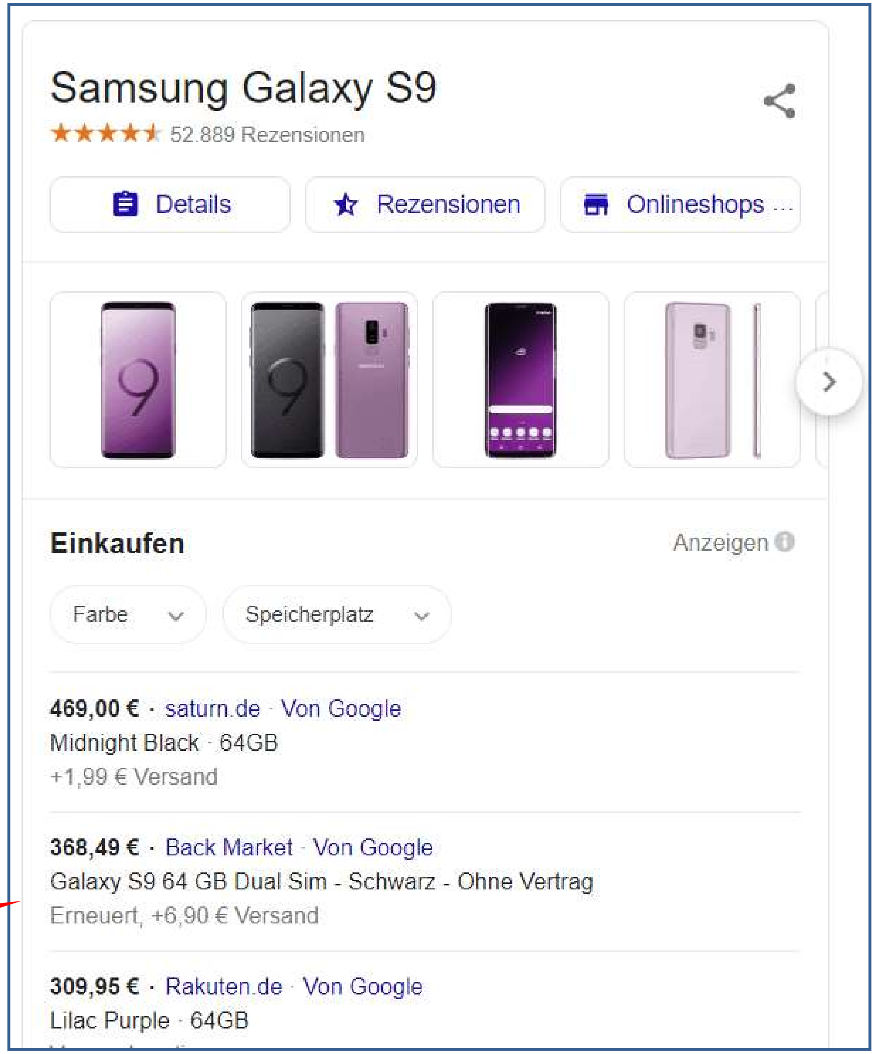
Usage of Schema.org Data @ Google



Data snippets within search results

Local businesses on maps

Product offers



https://developers.google.com/search/docs/advanced/structured-data/search-gallery

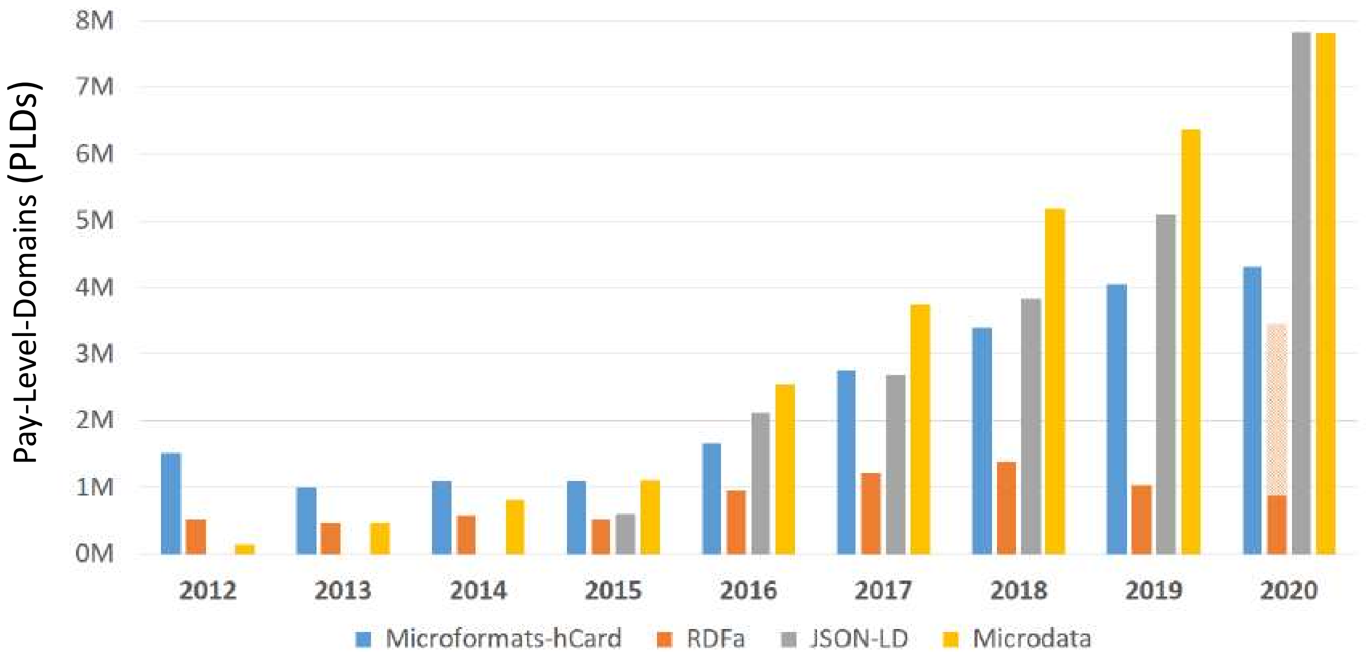
Web Data Commons Project

- extracts all Microformat, Microdata, RDFa, JSON-LD data from the Common Crawl (CC)
- analyzes and provides the extracted data for download
- statistics of some extraction runs
  - 2010 CC Corpus: 2.8 billion HTML pages → 5.1 billion RDF triples
  - 2013 CC Corpus: 2.2 billion HTML pages → 17.2 billion RDF triples
  - 2017 CC Corpus: 3.1 billion HTML pages → 38.2 billion RDF triples
  - 2020 CC Corpus: 3.4 billion HTML pages → 86.3 billion RDF triples

http://webdatacommons.org/structureddata/

Adoption of Semantic Annotations 2020

- 1.7 billion HTML pages out of the 3.4 billion pages provide semantic annotations (50.0%).
- 15.3 million pay-level-domains (PLDs) out of the 34.5 million PLDs (websites) provide semantic annotations (44.3%).



http://webdatacommons.org/structureddata/2020-12/stats/stats.html

Frequently used Schema.org Classes

Top Classes	# Websites (PLDs)	
	JSON-LD	Microdata
schema:WebPage	4,484,026	1,339,999
schema:Person	3,151,809	514,990
schema:BreadcrumbList	1,688,820	924,991
schema:Article	1,327,578	627,303
schema:Product	1,234,972	1,059,149
schema:Offer	1,182,855	946,725
schema:PostalAddress	863,243	585,417
schema:BlogPosting	529,020	552,338
schema:LocalBusiness	363,843	280,338
schema:AggregateRating	432,014	315,253
schema:Place	255,139	93,124
schema:Event	194,115	77,722
schema:Review	181,097	158,333
schema:JobPosting	28,759	8,520

http://webdatacommons.org/structureddata/2020-12/stats/schema_org_subsets.html

Use Case 1: Price Comparison

Questions to answer for a price comparison portal:

1. Which vendors offer a specific product?
2. Which prices do the vendors charge for the product?
3. Which shipping conditions and charges apply?

idealo Google Shopping

Use Case 1: Price Comparison

```
<html><head>
  <script type="application/ld+json">
  { "@context": "https://schema.org/",
    "@type": "Product",
    "sku": "23211",
    “gtin13": "4049998402140“,
    "name": "RYZEN 5 GAMING PC 6 x 4.10 GHz Turbo 16GB",
     "offers": {
          "@type": "Offer",
          "url": "https://americancomputers.al/shop/ryzen-5-gaming-pc/",
           "availability": "https://schema.org/InStock",
           "price": “749.00",
           "priceCurrency": “EUR",
           "shippingDetails": { "@type": "OfferShippingDetails",
               "shippingRate": { "@type": "MonetaryAmount",
                "value": "3.49", "currency": "USD“},
                "shippingDestination": { "@type": "DefinedRegion",
                "addressCountry": "US" }
     } } }   </script> </head></html>
```

Product identifiers allow grouping offers for the same product from different shops

Price information is available as part of schema:offer

Shipping details are available as part of schema:offer

Adoption of Schema.org Product Identifier Annotations

In 2020, over 60% of the e-commerce websites annotate product IDs

schema.org/ Product property	2013				2017				2020			
	Markedup entities		PLDs		Markedup entities		PLDs		Markedup entities		PLDs	
	# (in K)	%	# (in K)	%	# (in K)	%	# (in K)	%	# (in K)	%	# (in K)	%
gtin8	0.3	0.0	0.0	0.0	540.9	0.1	0.3	0.0	3,661.9	0.5	22.8	1.0
gtin12	0.3	0.0	0.0	0.0	507.6	0.1	0.6	0.1	4,052.6	0.5	24.5	1.0
gtin13	177.5	0.1	0.3	0.4	5,737.9	1.2	6.6	1.0	29,590.2	3.7	68.0	2.9
gtin14	10.9	0.0	0.0	0.0	578.1	0.1	0.8	0.1	2,784.0	0.3	18.0	0.8
identifier	273.4	0.2	0.2	0.2	425.3	0.1	0.6	0.1	4,197.5	0.5	14.1	0.6
productID	28,427.0	16.0	7.4	10.8	54,787.4	11.0	38.0	6.3	51,663.9	6.5	109.3	4.7
mpn	1,561.4	0.9	0.5	0.7	15,678.3	3.2	10.1	1.7	69,860.7	8.8	148.0	6.3
sku	14,513.1	8.2	1.3	1.9	49,732.8	10.0	150.4	25.3	241,700.5	30.4	1,291.1	56.2

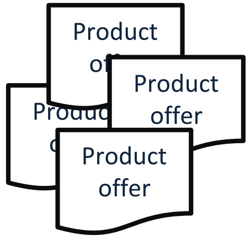
sku often contains merchant independent identifiers

http://webdatacommons.org/**structureddata/2020-12/stats/stats.html**

Grouping Offers by Product Identifier

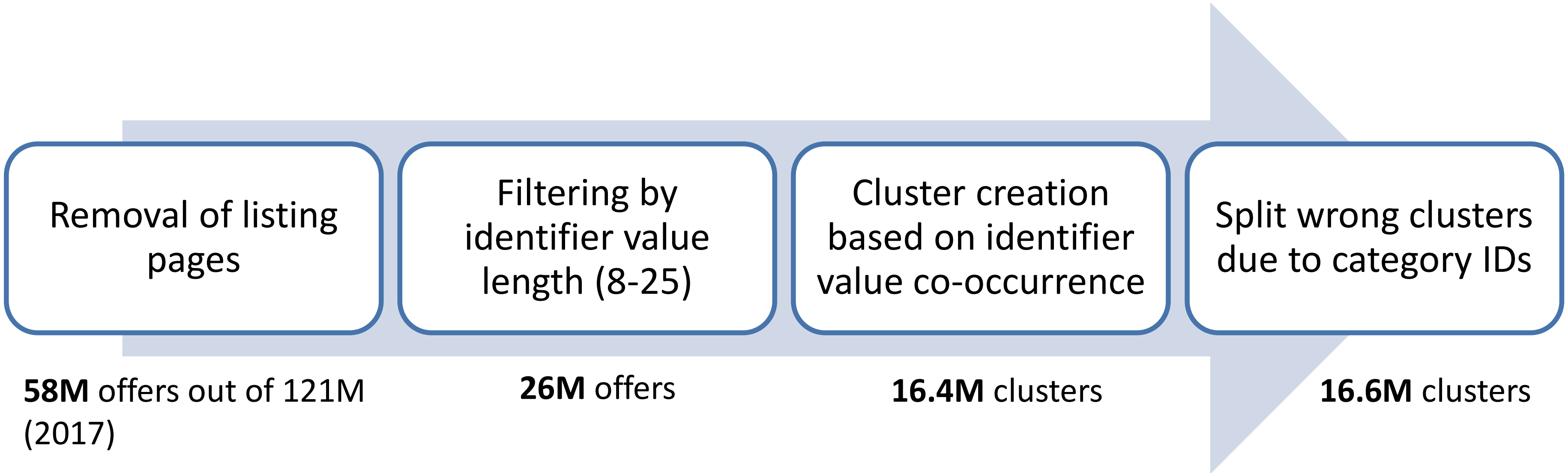
Group size	# Occurrences 2013	2017	2020
1	4,492,158	41,538,284	100,464,424
[2-5]	2,084,117	10,676,183	36,033,295
[6-10]	409,380	1,234,244	5,254,722
[11-20]	285,232	634,551	2,576,139
[21-50]	180,727	310,188	1,182,615
[51-100]	54,480	95,664	295,704
[101-200]	23,863	40,512	123,900
[201-500]	17,408	16,411	65,008
[501-800]	4,082	3,049	16,205
[801-1000]	1,208	917	5,057
[1001-5000]	2,436	2,127	11,623
[5001-10000]	125	208	852
>10000	161	497	465

Group of offers sharing the same ID



The CC 2020 contains least two offers for over 45 million products

Workflow for Cleansing the Product Clusters



Primpeli, Peeters, Bizer: **The WDC training dataset and gold standard for large-scale product matching.** WWW2019 Companion.

The WDC Product Corpora

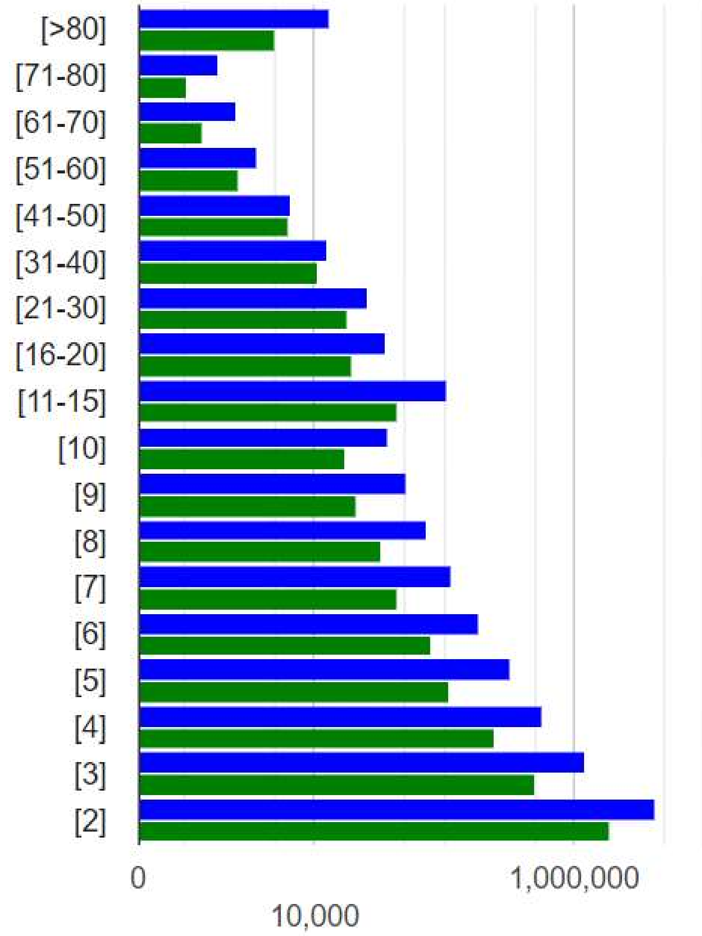
– extracted from CC 2017 and 2020

Version	# Offers	# Clusters >2	# Websites
2020	98 Million	7.1 Million	603,000
2017	26 Million	3.0 Million	78,000

– cluster quality: 93,4 %
(evaluated on sample of 900 pairs)

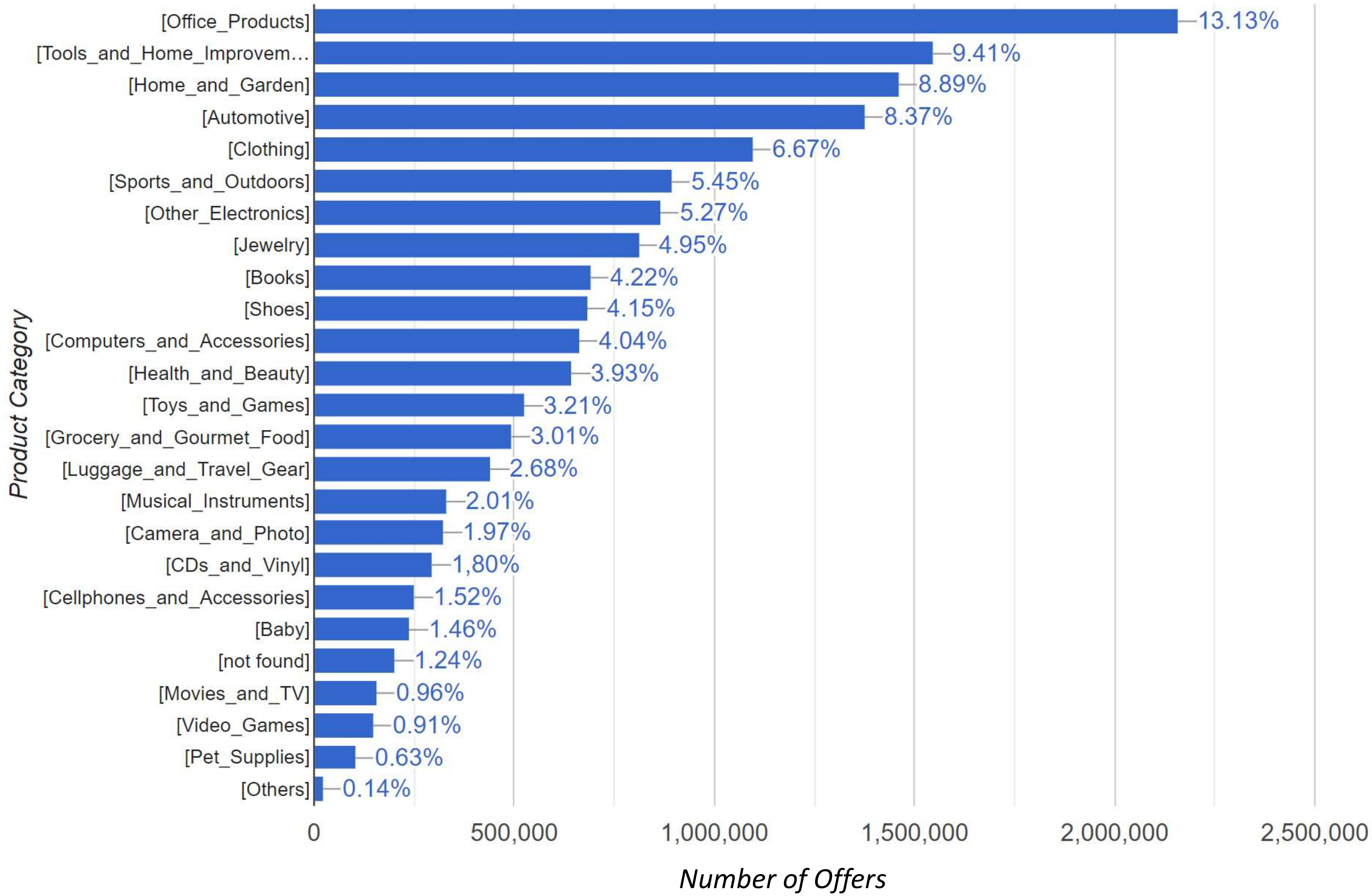
– available for download as JSON

Cluster Size Distribution
Blue: 2020, Green: 2017, log-scale



http://webdatacommons.org/**largescaleproductcorpus/v2020/**
http://webdatacommons.org/**largescaleproductcorpus/v2/**

Product Categories of the Offers in the WDC Product Corpus 2017



Use Case 1: Price Comparison

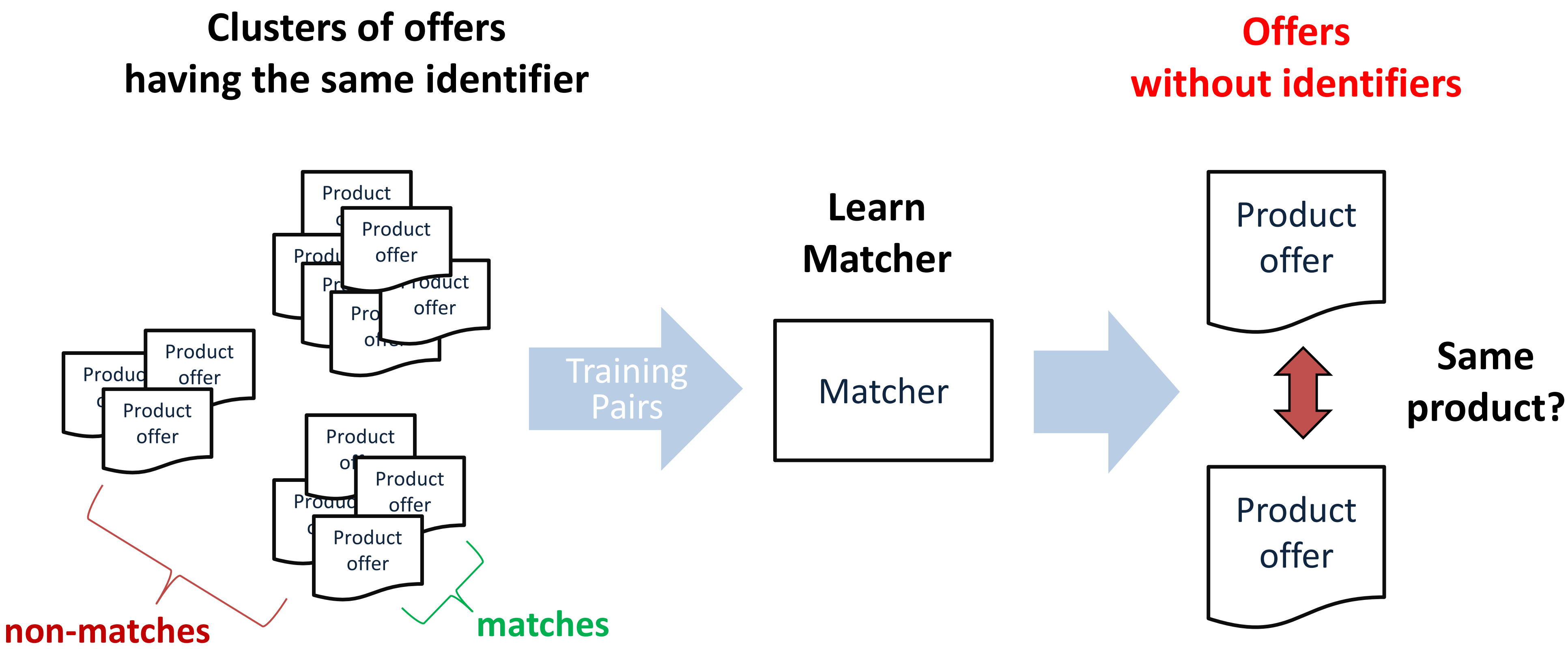
UNIVERSITY OF MANNHEIM
Data and Web Science Group

How to cover more e-shops?

603,000 websites in 2020 crawl offer reliable identifiers

1,691,00 websites offer product data but no reliable identifiers

Learn to Match Products using Schema.org Identifiers as Supervision



WDC Training Sets and Gold Standard for Large-Scale Product Matching

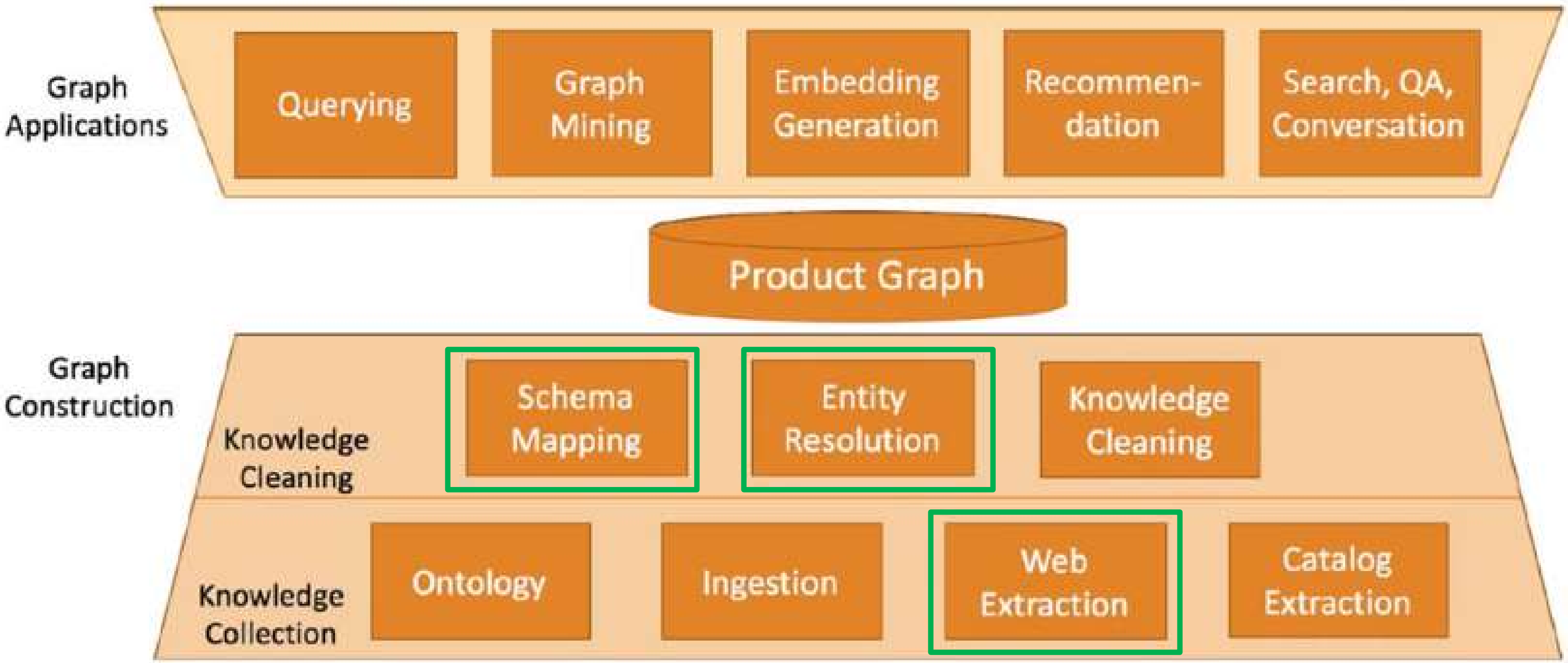
- covers four product categories
  - computers, cameras, watches, shoes
- Training sets of different sizes
  - 2,000 to 60,000 pairs of offers
- Manually verified test sets of 1100 pairs from each category

Category	Size	Positives	Negatives
Computers	XLarge	9.690	58.771
	Large	6.146	27.213
	Medium	1.762	6.332
	small	722	2.112
Cameras	XLarge	7.178	35.099
	Large	3.843	16.193
	Medium	1.108	4.147
	Small	486	1.400

Category	Size	Positives	Negatives
Watches	XLarge	9.264	52.305
	Large	5.163	21.864
	Medium	1.418	4.995
	small	580	1.675
Shoes	XLarge	4.141	38.288
	Large	3.482	19.507
	Medium	1.214	4.591
	Small	530	1.533

http://webdatacommons.org/**largescaleproductcorpus**/v2/

Use Case 2: Completing Product Knowledge Graphs



Luna Dong: **Challenges and Innovations in Building a Product Knowledge Graph.** KDD 2018.

Schema.org Attributes used to Describe Products in 2020

Attribute	# PLDs
schema:Product/name	99 %
schema:Product/offers	94 %
schema:Offer/price	95 %
schema:Offer/priceCurrency	95 %
schema:Product/description	84 %
schema:Offer/availability	72 %
schema:Product/sku	56 %
schema:Product/brand	30 %
schema:Product/image	26 %
schema:Product/aggregateRating	17 %
schema:Product/category	8 %
schema:Product/productID	5 %
…	…

New Samsung Galaxy S4 GT-19505 16GB 5.0 inches Android Smartphone with 2-Year Sprint Contract - White Frost

299 US$

The Galaxy S4 is among the earliest phones to feature a 1080p Full HD display. The various connectivity options on the Samsung include …

000214632623

Samsung

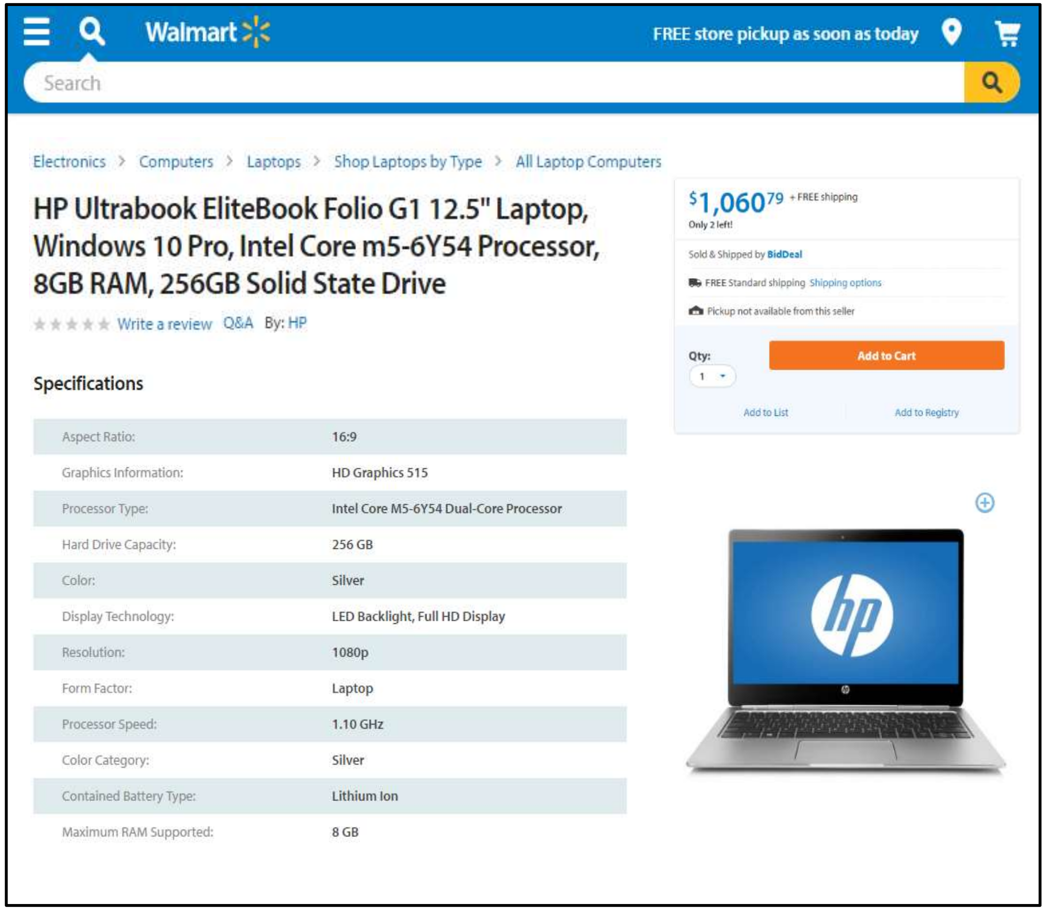
No category-specific attributes, such as memory or screen size ☹

http://webdatacommons.org/structureddata/schemaorgtables/

Option: Exploit HTML Tables Containing Product Specifications

schema:name

Product specifications as HTML Table



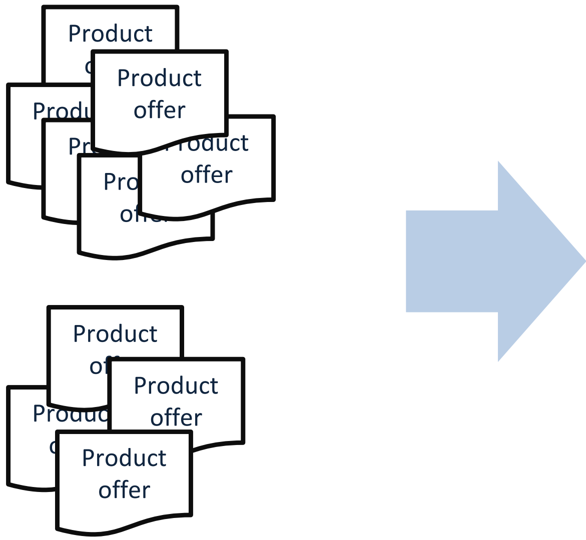
10% of the product pages in CC2017 contain specification tables

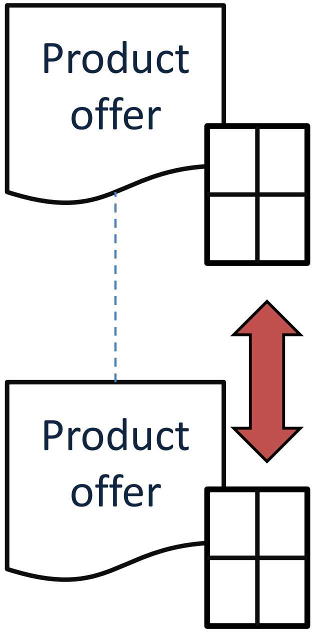
Qui, et al.: **DEXTER: Large-Scale Discovery and Extraction of Product Specifications on the Web.** VLDB 2015.
Petrovski, et al: **The WDC Gold Standards for Product Feature Extraction and Product Matching.** ECWeb 2016.

Option: Exploit Clusters to Combine Data from Multiple Offers for the Same Product

Clusters of offers having the same identifiers

Combine specification table content from multiple offers for same product





2. Product Matching using Deep Learning Techniques

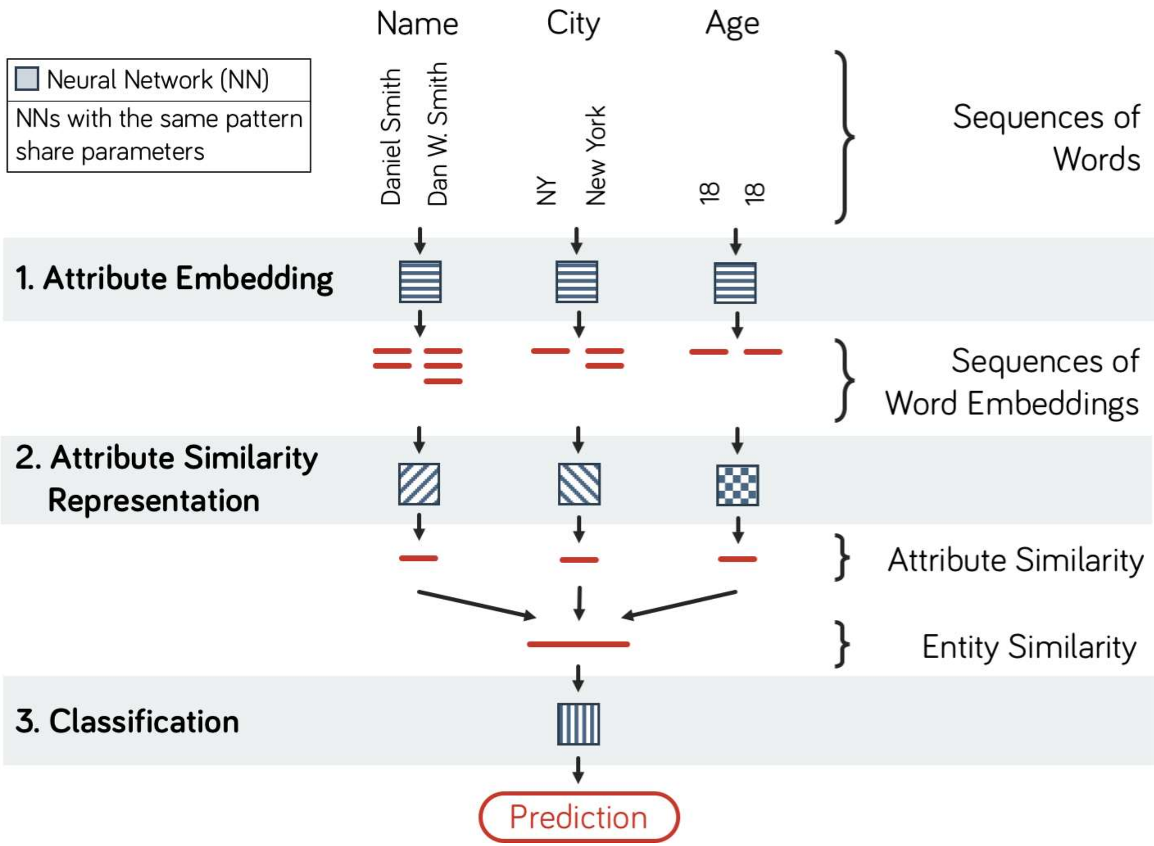
Performance of Magellan as an Example of a Traditional Matching Method

Type	Dataset	Topic	Magellan F1
Textual Data	WDC Computer - Large	Products	64.5
	WDC Computer - Small	Products	57.6
	Abt-Buy	Products	43.6
	Amazon-Google	Products	49.1
Structured Data	iTunes-Amazon	Music	91.2
	DBLP-ACM	Bibliographic	98.4
	DBLP-Scholar	Bibliographic	94.7

Product matching performance too low for practical applications ☹

Konda, et al.: **Magellan: Toward Building Entity Matching Management**. PVLDB 2016.
Anna Primpeli and Christian Bizer: **Profiling Entity Matching Benchmark Tasks**. CIKM 2020.

DeepMatcher (2018)



- Embeddings: FastText
- Summarization: Bi-RNN with attention
- Similarity computation: element-wise difference and multiplication, concatenation
- Classification: Fully connected neural net, cross entropy loss

Mudgal, Sidharth, et al.: **Deep Learning for Entity Matching: A Design Space Exploration.** SIGMOD 2018.

Evaluation: DeepMatcher versus Magellan

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Type	Dataset	DeepMatcher F1	Difference
Textual Data	WDC Computer - Large	89.5	+25.0
	WDC Computer - Small	70.5	+12.9
	Abt-Buy	62.8	+19.2
	Amazon-Google	69.3	+20.1
Structured Data	iTunes-Amazon	88.5	-2.7
	DBLP-ACM	98.4	+0.0
	DBLP-Scholar	92.3	+2.4

- DeepMatcher outperforms traditional methods on textual data
- mixed results on structured data

Mudgal, Sidharth, et al.: **Deep Learning for Entity Matching: A Design Space Exploration.** SIGMOD 2018.

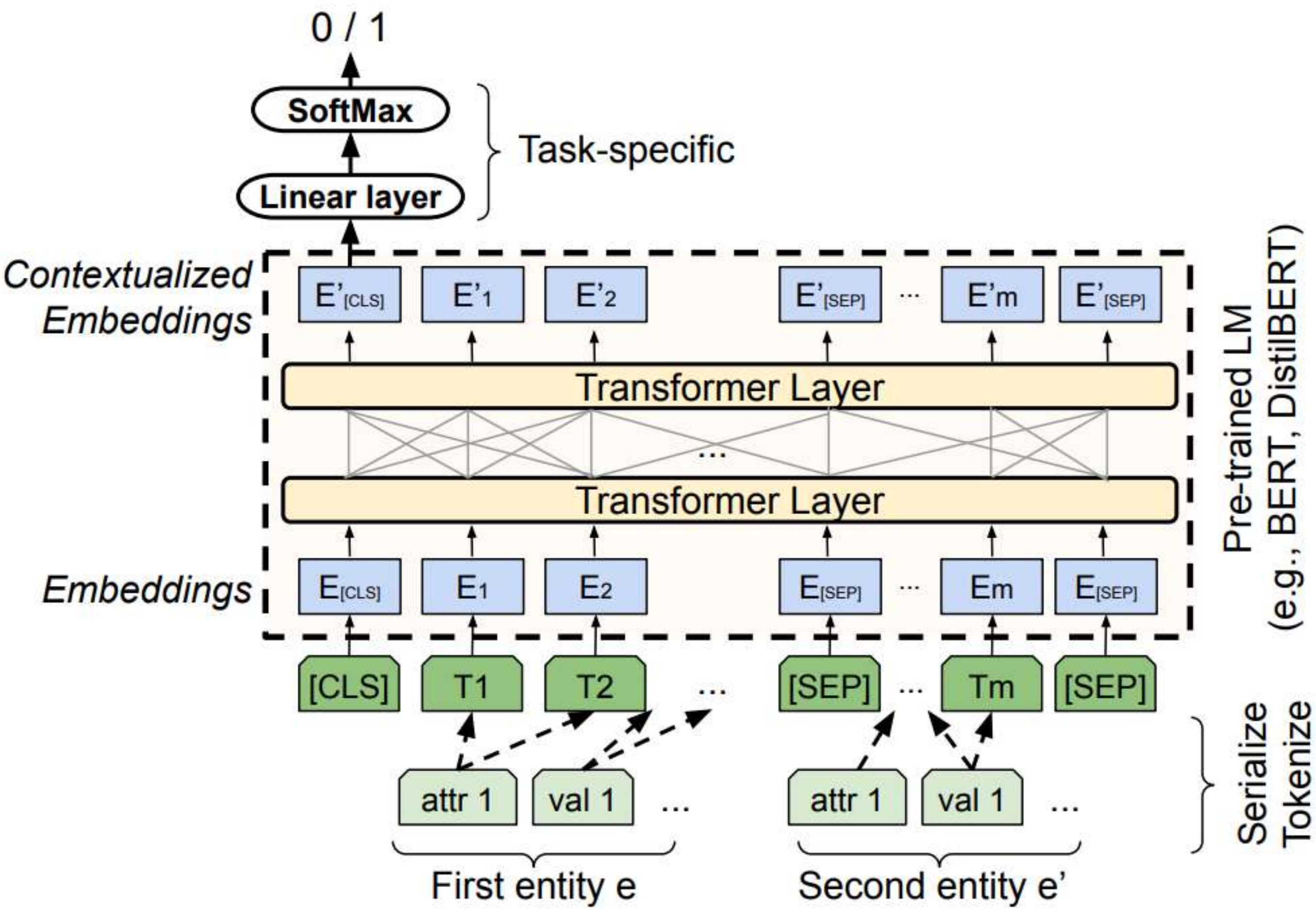
DITTO (2021)

- applies BERT, DistilBERT, RoBERTa for entity matching
- adds methods for entity summarization, highlighting matching clues, training data augmentation

- Entity serialization for BERT
  - Pair of entity descriptions are turned into single sequence
  - [CLS] Entity Description 1 [SEP] Entity Description 2 [SEP]
  - Entity Description = [COL] $attr_1$ [VAL] val_1 . . . [COL] $attr_k$ [VAL] val_k

Yuliang, et al: **Deep entity matching with pre-trained language models.** PVLDB 2021.

DITTO: Architecture



– [CLS] token summarizes the pair of entities
– linear layer on top of [CLS] token for matching decision

DITTO: Evaluation

Type	Dataset	DITTO F1	DeepMatcher F1	Magellan F1
Textual Data	WDC Computer - Large	91.7	89.5 +3.2	64.5 +27.2
	WDC Computer - Small	80.8	70.5 +10.3	57.6 +23.2
	Abt-Buy	89.3	62.8 +26.5	43.6 +45.7
	Amazon-Google	75.6	69.3 +6.3	49.1 +26.5
Structured Data	iTunes-Amazon	97.0	88.5 +8.5	91.2 +5.8
	DBLP-ACM	99.0	98.4 +0.6	98.4 +0.6
	DBLP-Scholar	95.6	92.3 +3.3	94.7 +0.9

– large performance gain for textual data
– constant improvement for structured data

Potential Reasons for the Performance Gain

- Serialization allows model to attend to all attributes
  - no strict separation between attributes
- WordPiece tokenizer breaks unknown terms into pieces
  - no problems with out of vocabulary terms
- Transfer learning from pre-training texts
  - different surface forms are already close in embedding space
- Contextualization of the token embeddings
  - potentially more suited for capturing differing semantics

Paganelli, et al: **Analyzing How BERT Performs Entity Matching.** PVLDB 2022.

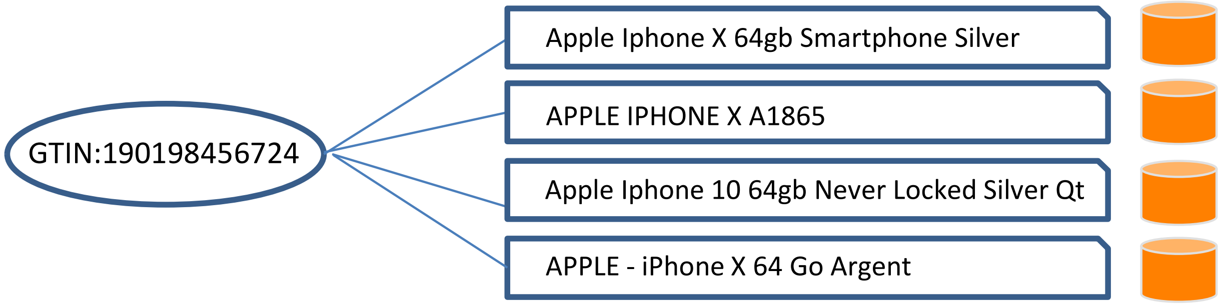
Seen versus Unseen during Training

1. model trained on WDC computers large
2. model tested using different test sets:

Test Set	RoBERTa F1	BERT F1	DeepMatcher F1	Magellan F1
a) Seen products - different offers	94.73	92.11	89.55	63.45
b) Unseen products - high similarity	83.18	84.53	58.64	27.89
c) Unseen products - low similarity	77.72	76.92	58.78	59.04
Δ between a) and c)	-17.01	-15.19	-30.91	-4.41

Peeters, Bizer: **Dual-Objective Fine-Tuning of BERT for Entity Matching**. PVLDB 2021

Back to schema.org Identifiers: Let’s exploit our Clusters!

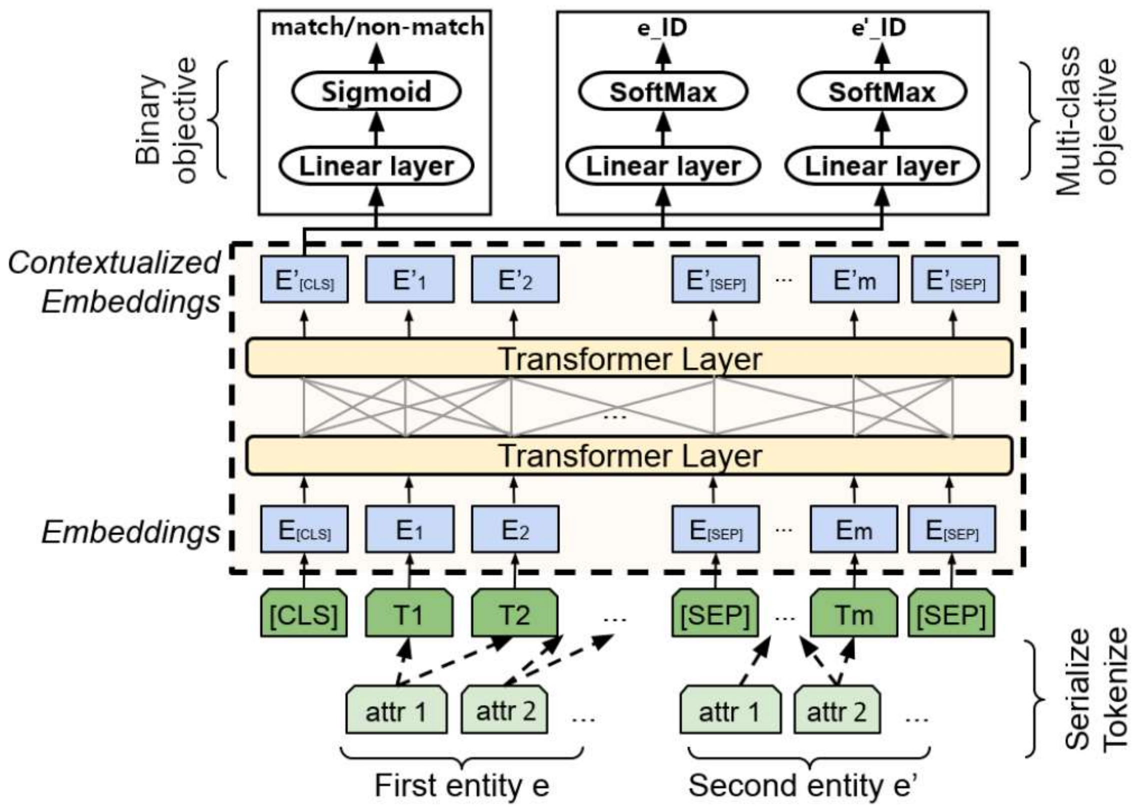


Clusters of product offers allow to

– learn recognizers for single products
– treat entity matching as a multi-class classification task

JointBERT (2021)

- combines pairwise matching and multi-class classification via multi-task learning
- hypothesis: Learning to recognize single entities might help to classify entity pairs



Overall loss:

$$L_i = BCEL\left(y_{b_i}, \hat{y}_{b_i}\right) + \left(CEL\left(y_{l_i}, \hat{y}_{l_i}\right) + CEL\left(y_{r_i}, \hat{y}_{r_i}\right)\right)$$

BCEL: Binary cross entropy loss
CEL: Cross entropy loss

Peeters, Bizer: **Dual-Objective Fine-Tuning of BERT for Entity Matching**. PVLDB 2021.

JointBERT: Evaluation

Testset	Training Set	Word Co-oc	Magellan	Deepmatcher	BERT	RoBERTa	Ditto	JointBERT
WDC computers	xlarge	82.39	63.16	88.95	94.57	94.73	96.53	**97.49**
	large	81.23	64.56	84.32	92.11	94.68	93.81	**96.90**
	medium	70.94	61.59	69.85	89.31	**91.90**	88.97	88.82
	small	62.69	57.60	61.22	80.46	**86.37**	81.52	77.55
WDC cameras	xlarge	73.33	51.70	84.88	91.42	94.39	94.74	**98.02**
	large	76.24	54.49	82.16	91.02	93.91	94.41	**96.51**
	medium	69.89	54.99	69.34	87.02	**90.20**	87.97	87.91
	small	64.86	52.78	59.65	77.47	**85.74**	78.67	78.30
WDC watches	xlarge	79.78	56.04	88.34	95.76	94.87	97.05	**97.09**
	large	79.64	60.59	86.03	95.23	93.93	97.17	**98.46**
	medium	69.54	66.62	67.92	89.00	**92.28**	89.16	87.46
	small	63.49	59.73	54.97	78.73	**87.16**	81.32	75.83
WDC shoes	xlarge	70.38	61.45	86.74	87.44	88.88	93.28	**97.88**
	large	71.18	60.48	83.17	87.37	86.60	90.07	**95.16**
	medium	72.43	59.80	74.40	79.82	81.12	**83.20**	82.61
	small	63.65	58.57	64.71	74.49	**80.29**	75.13	73.13

Increase of 1% to 5% F1 given enough training examples

Peeters, Bizer: **Dual-Objective Fine-Tuning of BERT for Entity Matching**. PVLDB 2021.

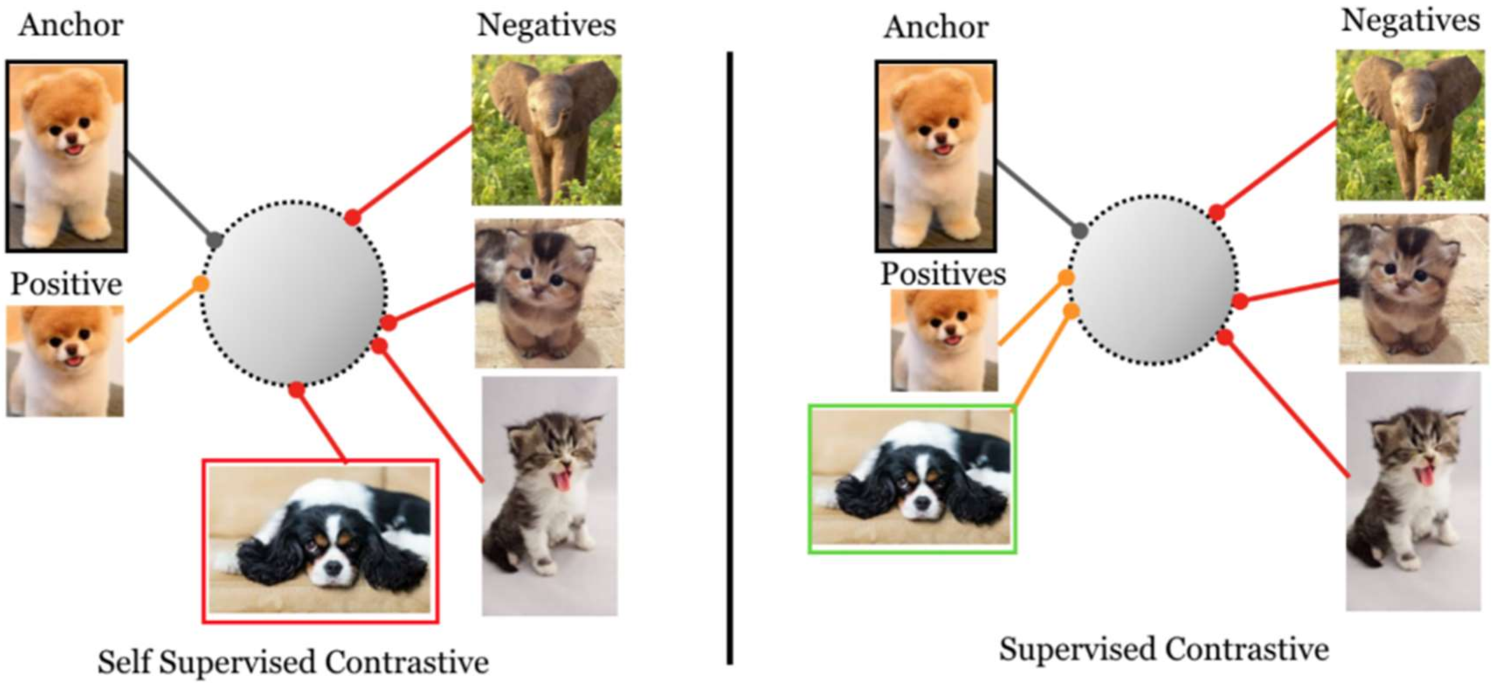
What about Long-Tail Products without many Training Examples?

Testset	Training Set	Word Co-oc	Magellan	Deepmatcher	BERT	RoBERTa	Ditto	JointBERT
WDC computers	xlarge	82.39	63.16	88.95	94.57	94.73	96.53	**97.49**
	large	81.23	64.56	84.32	92.11	94.68	93.81	**96.90**
	medium	70.94	61.59	69.85	89.31	**91.90**	88.97	88.82
	small	62.69	57.60	61.22	80.46	**86.37**	81.52	77.55
WDC cameras	xlarge	73.33	51.70	84.88	91.42	94.39	94.74	**98.02**
	large	76.24	54.49	82.16	91.02	93.91	94.41	**96.51**
	medium	69.89	54.99	69.34	87.02	**90.20**	87.97	87.91
	small	64.86	52.78	59.65	77.47	**85.74**	78.67	78.30
WDC watches	xlarge	79.78	56.04	88.34	95.76	94.87	97.05	**97.09**
	large	79.64	60.59	86.03	95.23	93.93	97.17	**98.46**
	medium	69.54	66.62	67.92	89.00	**92.28**	89.16	87.46
	small	63.49	59.73	54.97	78.73	**87.16**	81.32	75.83
WDC shoes	xlarge	70.38	61.45	86.74	87.44	88.88	93.28	**97.88**
	large	71.18	60.48	83.17	87.37	86.60	90.07	**95.16**
	medium	72.43	59.80	74.40	79.82	81.12	**83.20**	82.61
	small	63.65	58.57	64.71	74.49	**80.29**	75.13	73.13

Results for smaller training sets (~2-3K) are ~10% F1 below top results achieved using >50K training pairs.

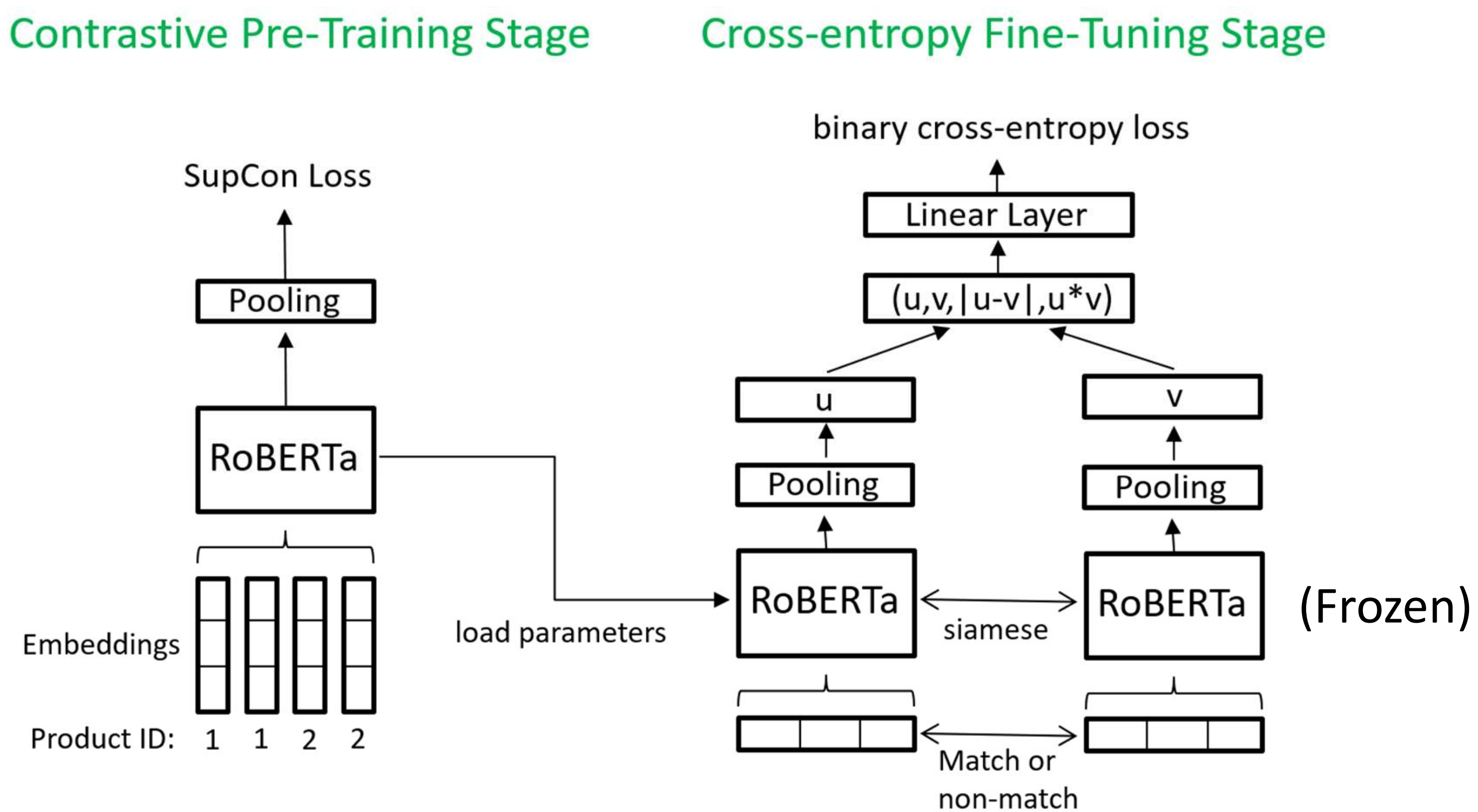
Contrastive Pretraining in Vision

- applies data augmentation to add positives
- uses large batches containing many positive and negative examples
- maximizes distance between classes in the embedding space



Khosla, et al.: **Supervised Contrastive Learning**. NeurIPS 2020.

Supervised Contrastive Pretraining for Entity Matching (2022)



Input: Batch of n product offers with product IDs

Input: Batch of product offer pairs with match/non-match labels

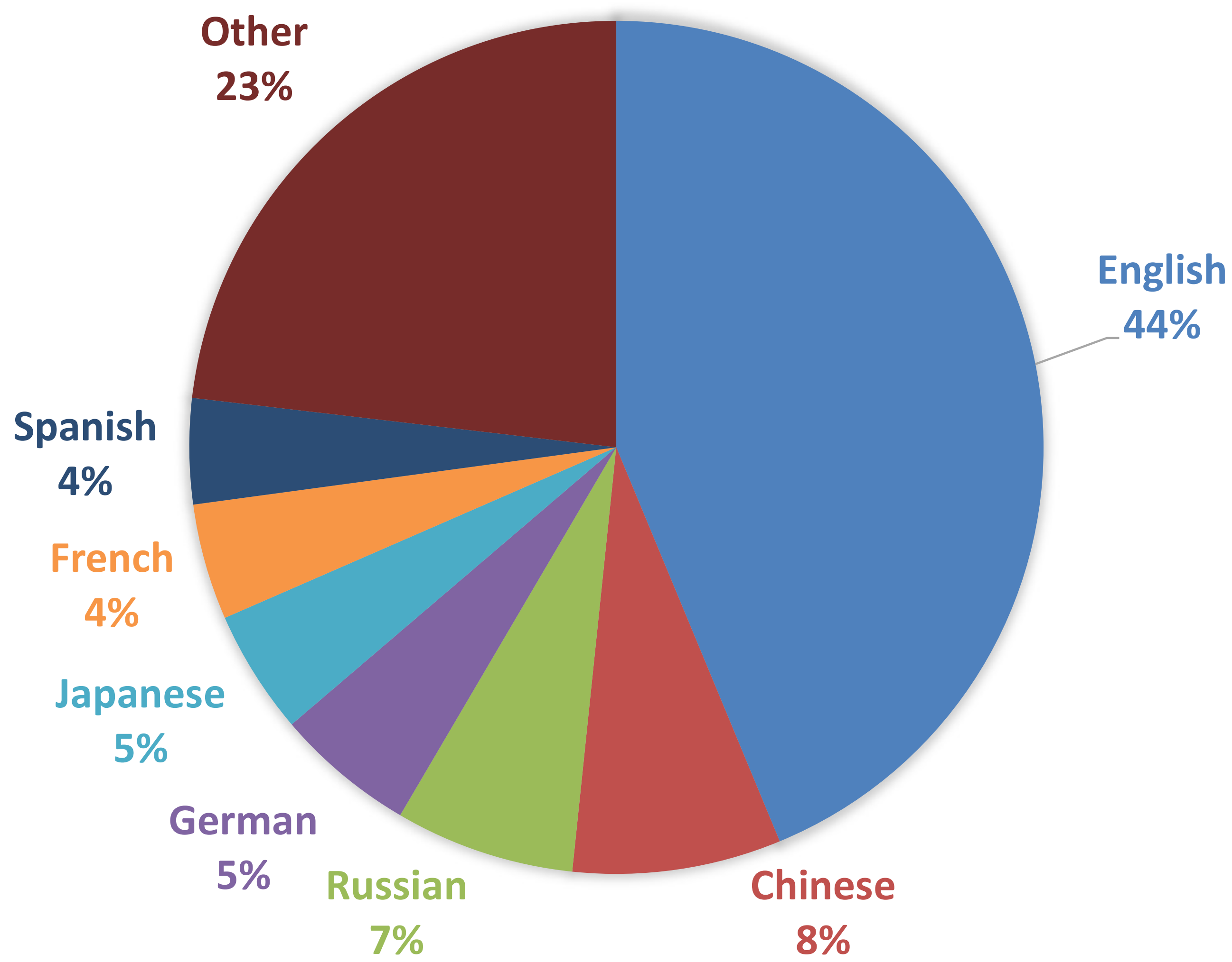
Peeters, Bizer: **Supervised Contrastive Learning for Product Matching**. WWW Companion 2022.

Evaluation: Supervised Contrastive Pretraining R-SupCon + Augmentation

	WDC Computers				Abt-Buy	Amazon-Google
# Training Pairs	~3K (small)	~8K (medium)	~23K (large)	~68K (xlarge)	~7.5K	~9K
DeepMatcher	61.22	69.85	84.32	88.95	62.80	70.70
RoBERTa	86.37	91.90	94.68	94.73	91.05	74.10
Ditto	80.76	88.62	91.70	95.45	89.33	75.58
JointBERT	77.55	88.82	96.90	97.49	83.44	-
R-SupCon	93.18	97.66	98.16	**98.33**	93.70	**79.28**
R-SupCon+augment	**95.21**	**98.50**	**98.50**	98.33	**94.29**	76.14
Δ to best baseline	+ 8.84	+ 6.60	+ 1.60	+ 0.84	+ 3.24	+ 3.70

F1 > 0.95 results also for small training sets!

Language of the Webpages in the Common Crawl



Cross-Language Learning for Product Matching

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- schema.org ID clusters may cover multiple languages
  - many offers in English as head language
  - less offers in languages such as French or German
- Potential for cross-language learning!
- Experiment: Fine-tuning mBERT

German training set extended with English pairs. F1s:

DE \ EN	0	450	900	1800	3600	7200	Δ 0-7200
450	67.11	72.79	75.44	80.83	86.82	87.97	20.86
900	75.76	75.10	74.00	87.67	88.92	88.19	12.43
1800	87.69	88.43	88.38	90.17	90.72	91.44	3.75
3600	93.63	92.98	92.46	93.97	93.25	94.46	0.83

Peeters, Bizer: **Cross-Language Learning for Product Matching**. WWW Companion 2022.

Conclusions: Deep Learning for Product Matching

1. Transformer-based matchers boost matching performance
   – F1 scores >0.95 in many cases
2. Contrastive pre-training and cross-language learning improve performance for long-tail products
3. All products should be covered by training examples
   – F1 scores for unseen products around 0.80
4. Reduced feature engineering effort
   – less information extraction effort due to serialization
   – less value normalization necessary due to pre-training

Conclusions: Schema.org Product Data on the Web

1. Schema.org annotations are a valuable source of product data
   – many sources, many languages, many products
   – UC1: annotations directly enable price comparison
   – UC2: augmenting KGs requires additional information extraction
2. Product identifiers provide distant supervision
   – rich source of training data for learning matchers
3. Similar potentials exist in other schema.org domains
   – Local businesses, job postings, reviews, movies, …
   – Tasks: Matching, hierarchical classification, sentiment analysis, pre-training of domain-specific language models, …

Hands On:
Entity Matching Methods

- The source code for all discussed methods is available
  - you can test if the methods work for your use cases

- Current benchmark results are collected on Papers with Code
  - good reference point for latest developments
  - provides links to the source code of the different matchers

https://paperswithcode.com/task/entity-resolution/

Hands On: Experimenting with Schema.org Data

- All mentioned datasets are available for download
- Product Data Corpora and Training/Test Sets (JSON)
  - http://webdatacommons.org/largescaleproductcorpus/v2/
  - https://webdatacommons.org/largescaleproductcorpus/v2020/
- Data for all Schema.org Classes (RDF quads)
  - LocalBusiness, Event, JobPosting, Review, …
  - http://webdatacommons.org/structureddata/
- Schema.org Table Corpus (JSON)
  - data from 42 schema.org classes grouped into 2 million tables: one table per schema.org class and website
  - http://webdatacommons.org/structureddata/schemaorgtables/

Thank you.