Knowledge Cleaning

Overview and Introduction

Knowledge Extraction

Knowledge Cleaning **30 min**

Q&A

Break

Ontology Mining

Applications

Conclusion and Future Directions

Q&A

Section Structure

- Problem Definition

  What are unique challenges for PG beyond generic KGs?
- Short answer -- key intuition

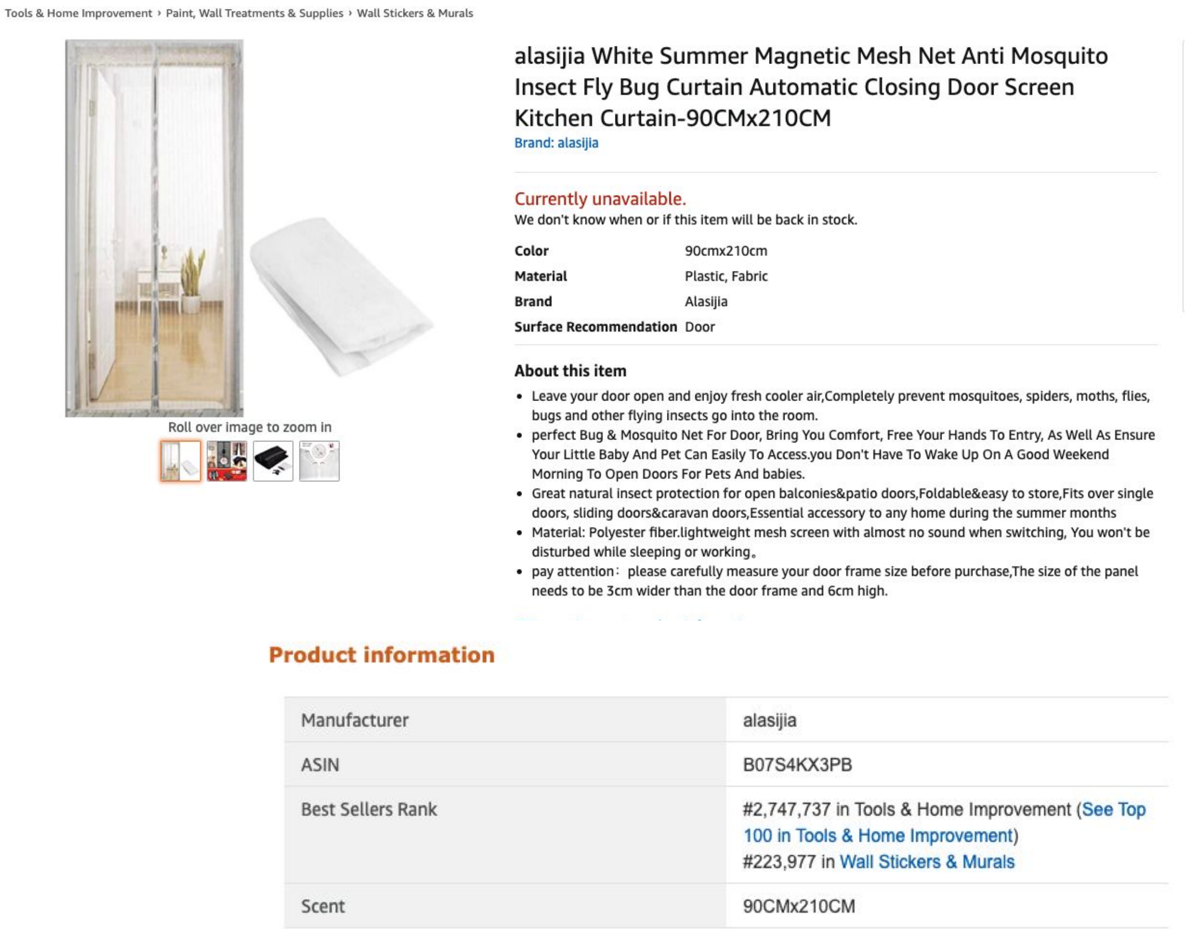
  What are key intuitions for building product KGs?
- Long answer -- details

  What are practical tips?
- Reflection/short-answer

  Can we apply the techniques to other domains?

Why Knowledge Cleaning?

From the eyes of customers



Tools & Home Improvement › Paint, Wall Treatments & Supplies › Wall Stickers & Murals

alasijia White Summer Magnetic Mesh Net Anti Mosquito Insect Fly Bug Curtain Automatic Closing Door Screen Kitchen Curtain-90CMx210CM

Brand: alasijia

Currently unavailable.

We don't know when or if this item will be back in stock.

Color	90cmx210cm
Material	Plastic, Fabric
Brand	Alasijia
Surface Recommendation	Door

Roll over image to zoom in

About this item

- Leave your door open and enjoy fresh cooler air,Completely prevent mosquitoes, spiders, moths, flies, bugs and other flying insects go into the room.
- perfect Bug & Mosquito Net For Door, Bring You Comfort, Free Your Hands To Entry, As Well As Ensure Your Little Baby And Pet Can Easily To Access.you Don't Have To Wake Up On A Good Weekend Morning To Open Doors For Pets And babies.
- Great natural insect protection for open balconies&patio doors,Foldable&easy to store,Fits over single doors, sliding doors&caravan doors,Essential accessory to any home during the summer months
- Material: Polyester fiber.lightweight mesh screen with almost no sound when switching, You won't be disturbed while sleeping or working。
- pay attention： please carefully measure your door frame size before purchase,The size of the panel needs to be 3cm wider than the door frame and 6cm high.

Product information

Manufacturer	alasijia
ASIN	B07S4KX3PB
Best Sellers Rank	#2,747,737 in Tools & Home Improvement (See Top 100 in Tools & Home Improvement) #223,977 in Wall Stickers & Murals
Scent	90CMx210CM

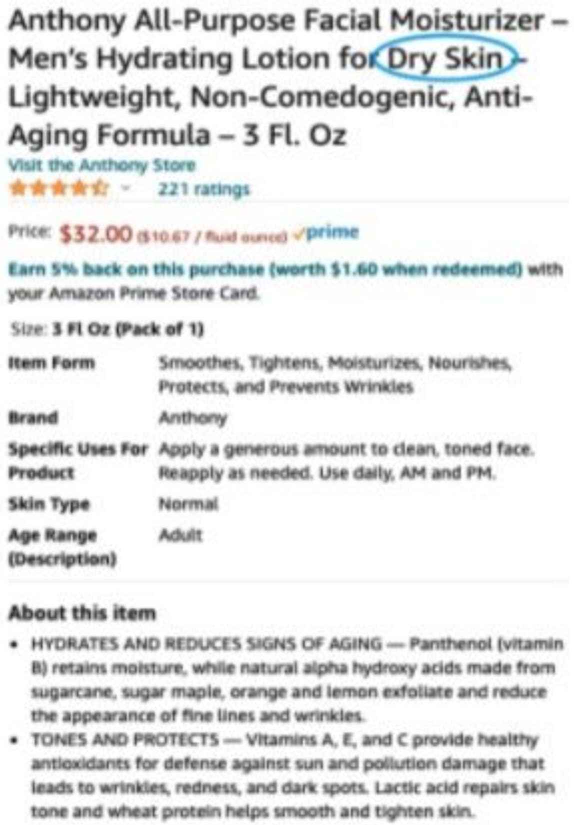
Backend data storage

Attribute	Attribute Value
Title	alasijia White Summer Magnetic Mesh Net Anti Mosquito Insect Fly Bug Curtain Automatic Closing Door Screen Kitchen Curtain-90CMx210CM
Brand	Alasijia
Color	**90cm X 210cm**
Material	Plastic Fabric
Scent	**90cm X 210cm**

Why Knowledge Cleaning?

From the eyes of customers

All Purpose
Facial Moisturizer
All Skin Types
Objective: Face it, it's your future. Moisturize wisely.
Strategy: Natural AHAs – Sugarcane, Sugar Maple, Orange and Lemon Extracts – work hard to help erase and prevent wrinkles. Lactic Acid and Licorice Root Extract help to even skin tone and texture. Panthenol helps retain moisture, while Vitamins A, C and E provide antioxidant protection against free radicals from environmental aggressors.
90 ml / 3 fl oz
Anthony



Anthony All-Purpose Facial Moisturizer – Men's Hydrating Lotion for Dry Skin – Lightweight, Non-Comedogenic, Anti-Aging Formula – 3 Fl. Oz

Visit the Anthony Store

★★★★☆ 221 ratings

Price: $32.00 ($10.67 / fluid ounce) ✓prime

Earn 5% back on this purchase (worth $1.60 when redeemed) with your Amazon Prime Store Card.

Size: **3 Fl Oz (Pack of 1)**

Item Form	Smoothes, Tightens, Moisturizes, Nourishes, Protects, and Prevents Wrinkles
Brand	Anthony
Specific Uses For Product	Apply a generous amount to clean, toned face. Reapply as needed. Use daily, AM and PM.
Skin Type	Normal
Age Range (Description)	Adult

About this item

- HYDRATES AND REDUCES SIGNS OF AGING — Panthenol (vitamin B) retains moisture, while natural alpha hydroxy acids made from sugarcane, sugar maple, orange and lemon exfoliate and reduce the appearance of fine lines and wrinkles.
- TONES AND PROTECTS — Vitamins A, E, and C provide healthy antioxidants for defense against sun and pollution damage that leads to wrinkles, redness, and dark spots. Lactic acid repairs skin tone and wheat protein helps smooth and tighten skin.

Backend data storage

Attribute	Attribute Value
Title	Anthony All-Purpose Facial Moisturizer – Men's Hydrating Lotion for **Dry Skin** – Lightweight, Non-Comedogenic, Anti-Aging Formula – 3 Fl. Oz
Brand	Anthony
Skin Type	**Normal**
Age Range	Adult

What is Knowledge Cleaning?

- Problem definition
  - Given a fact **t** = {**e, a, v**}, where
    - **e**: the product
    - **a**: an attribute of the product e
    - **v**: the attribute value of e
  - Identify if **t** states the true fact about **e**

Generic Solution

- Key intuition: detecting data inconsistency
  - Column-wise: among values of the same attribute
  - Row-wise: among values of different attributes of the same entity
  - Graph-wise: among values of the entire data set
  - Across-source: among different data sources

Generic Solution

Column-wise inconsistency

- Key intuition: detecting data inconsistency

Product	Brand	Color	Scent	Skin type
Anthony All-Purpose Facial Moisturizer – Men’s Hydrating Lotion for Dry Skin – Lightweight, Non-Comedogenic, Anti-Aging Formula – 3 Fl. Oz	Anthony		scented	**normal**
CeraVe Moisturizing Cream | Body and Face Moisturizer for Dry Skin | Body Cream with Hyaluronic Acid and Ceramides | 19 Ounce	CeraVe		lavender	dry skin
White Summer Magnetic Mesh Net Anti Mosquito Insect Fly Bug Curtain Automatic Closing Door Screen Kitchen Curtain-90CMx210CM	Alasijia	**unscented**	**90cm X 210cm**	
Insulated Door Curtain-Magnetic Thermal Door Cover, Screen Door Self-Closing Privacy Screen Door Hands Free for Patio, Kitchen, Bedroom, Air Conditioner Room, Fits Doors up to 34" x 80"	SANJIANKER	transparent	unscented	
…	…	…	…	…

Generic Solution

- Key intuition: detecting data inconsistency

Row-wise inconsistency

Product	Brand	Color	Scent	Skin type
Anthony All-Purpose Facial Moisturizer – Men’s Hydrating Lotion for **Dry Skin** – Lightweight, Non-Comedogenic, Anti-Aging Formula – 3 Fl. Oz	Anthony		scented	**normal**
CeraVe Moisturizing Cream | Body and Face Moisturizer for Dry Skin | Body Cream with Hyaluronic Acid and Ceramides | 19 Ounce	CeraVe		lavender	dry skin
White Summer Magnetic Mesh Net Anti Mosquito Insect Fly Bug Curtain Automatic Closing Door Screen Kitchen Curtain-90CMx210CM	Alasijia	**unscented**	**90cm X 210cm**	
Insulated Door Curtain-Magnetic Thermal Door Cover, Screen Door Self-Closing Privacy Screen Door Hands Free for Patio, Kitchen, Bedroom, Air Conditioner Room, Fits Doors up to 34" x 80"	SANJIANKER	transparent	unscented	
...	...	...	...	...

Generic Solution

- Key intuition: detecting data inconsistency

Source-wise inconsistency

Product	Brand	Color	Scent	n type
Anthony All-Purpose Facial Moisturizer – Men's Hydrating Lotion for **Dry Skin** – Lightweight, Non-Comedogenic, Anti-Aging Formula – 3 Fl. Oz	Anthony		scen	**rmal**
CeraVe Moisturizing Cream | Body and Face Moisturizer for Dry Skin | Body Cream with Hyaluronic Acid and Ceramides | 19 Ounce	CeraVe		la ender	ry skin
White Summer Magnetic Mesh Net Anti Mosquito Insect Fly Bug Curtain Automatic Closing Door Screen Kitchen Curtain-90CMx210CM	Alasijia	**unscented**	**90cm X 210cm**	
Insulated Door Curtain-Magnetic Thermal Door Cover, Screen Door Self-Closing Privacy Screen Door Hands Free for Patio, Kitchen, Bedroom, Air Conditioner Room, Fits Doors up to 34" x 80"	SANJIANKER	transparent	unscented	
...	...	...	...	...

Generic Solution

- Key intuition: detecting data inconsistency

Across source inconsistency

	Product	Brand	Color	Scent	Skin type
Source A	Anthony All-Purpose Facial Moisturizer – Men’s Hydrating Lotion for Dry Skin – Lightweight, Non-Comedogenic, Anti-Aging Formula – 3 Fl. Oz	Anthony		scented	**normal**
	…	…	…	…	…
Source B	Anthony All-Purpose Facial Moisturizer – 3 Fl. Oz, Lightweight, Men’s Hydrating Lotion for Dry Skin	Anthony		scented	dry skin
	…	…	…	…	…
Source C	Anthony All-Purpose Facial Moisturizer – Men’s Hydrating Lotion for Dry Skin (3 Fl. Oz)	Anthony		scented	dry skin
	…	…	…	…	…

Unique Challenges for Product Knowledge Cleaning and Solutions

- Noisy structured data and rich unstructured textual data

  Leverage unstructured textual attribute as context to identify errors

- Big variety across product types

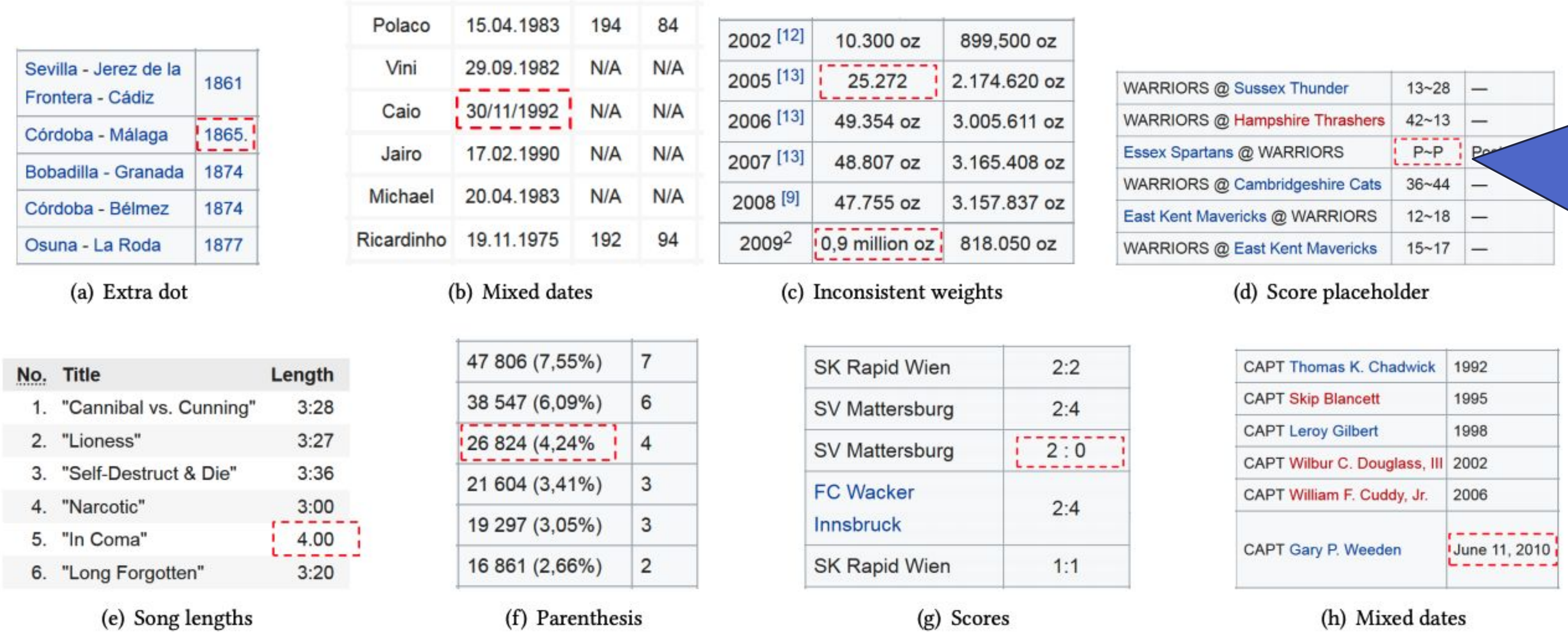
  Predict attribute correctness conditioned on product types

- Limited training labels for large-scale, rich data

  Distant supervision and few-shot learning setting

Generic Solution: Detect Column-wise Inconsistency

- Auto-Detect [SIGMOD 2018]
  - Automatically detect incompatible values by leveraging an ensemble of judiciously selected generalization languages



Sevilla - Jerez de la Frontera - Cádiz	1861
Córdoba - Málaga	1865.
Bobadilla - Granada	1874
Córdoba - Bélmez	1874
Osuna - La Roda	1877

(a) Extra dot

Polaco	15.04.1983	194	84
Vini	29.09.1982	N/A	N/A
Caio	30/11/1992	N/A	N/A
Jairo	17.02.1990	N/A	N/A
Michael	20.04.1983	N/A	N/A
Ricardinho	19.11.1975	192	94

(b) Mixed dates

2002 [12]	10.300 oz	899,500 oz
2005 [13]	25.272	2.174.620 oz
2006 [13]	49.354 oz	3.005.611 oz
2007 [13]	48.807 oz	3.165.408 oz
2008 [9]	47.755 oz	3.157.837 oz
2009[2]	0,9 million oz	818.050 oz

(c) Inconsistent weights

WARRIORS @ Sussex Thunder	13~28	—
WARRIORS @ Hampshire Thrashers	42~13	—
Essex Spartans @ WARRIORS	P~P	Po
WARRIORS @ Cambridgeshire Cats	36~44	—
East Kent Mavericks @ WARRIORS	12~18	—
WARRIORS @ East Kent Mavericks	15~17	—

(d) Score placeholder

No.	Title	Length
1.	"Cannibal vs. Cunning"	3:28
2.	"Lioness"	3:27
3.	"Self-Destruct & Die"	3:36
4.	"Narcotic"	3:00
5.	"In Coma"	4.00
6.	"Long Forgotten"	3:20

(e) Song lengths

47 806 (7,55%)	7
38 547 (6,09%)	6
26 824 (4,24%	4
21 604 (3,41%)	3
19 297 (3,05%)	3
16 861 (2,66%)	2

(f) Parenthesis

SK Rapid Wien	2:2
SV Mattersburg	2:4
SV Mattersburg	2 : 0
FC Wacker Innsbruck	2:4
SK Rapid Wien	1:1

(g) Scores

CAPT Thomas K. Chadwick	1992
CAPT Skip Blancett	1995
CAPT Leroy Gilbert	1998
CAPT Wilbur C. Douglass, III	2002
CAPT William F. Cuddy, Jr.	2006
CAPT Gary P. Weeden	June 11, 2010

(h) Mixed dates

Each language captures "**local**" compatibility and is "**sensitive**" to different types of errors

Huang et al.,Auto-Detect: Data-Driven Error Detection in Tables, SigKDD, 2020.

Generic Solution: Detect Column-wise Inconsistency

- Auto-Detect [SIGMOD 2018]

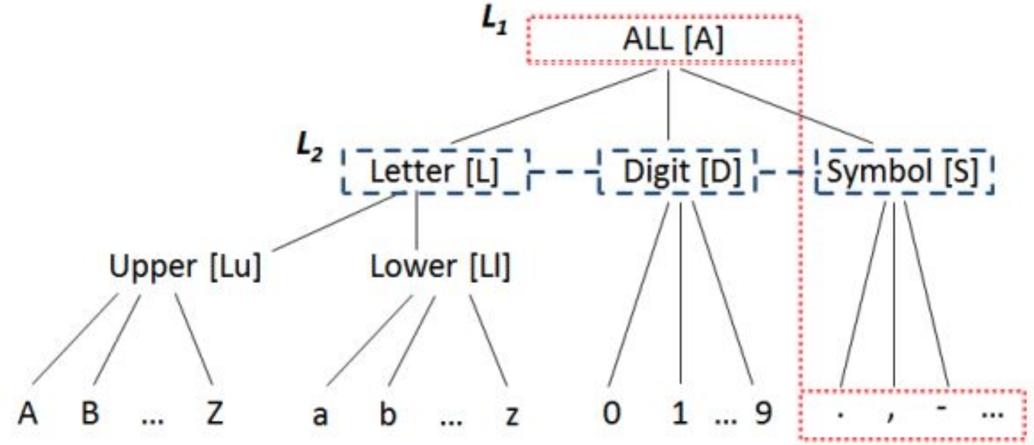
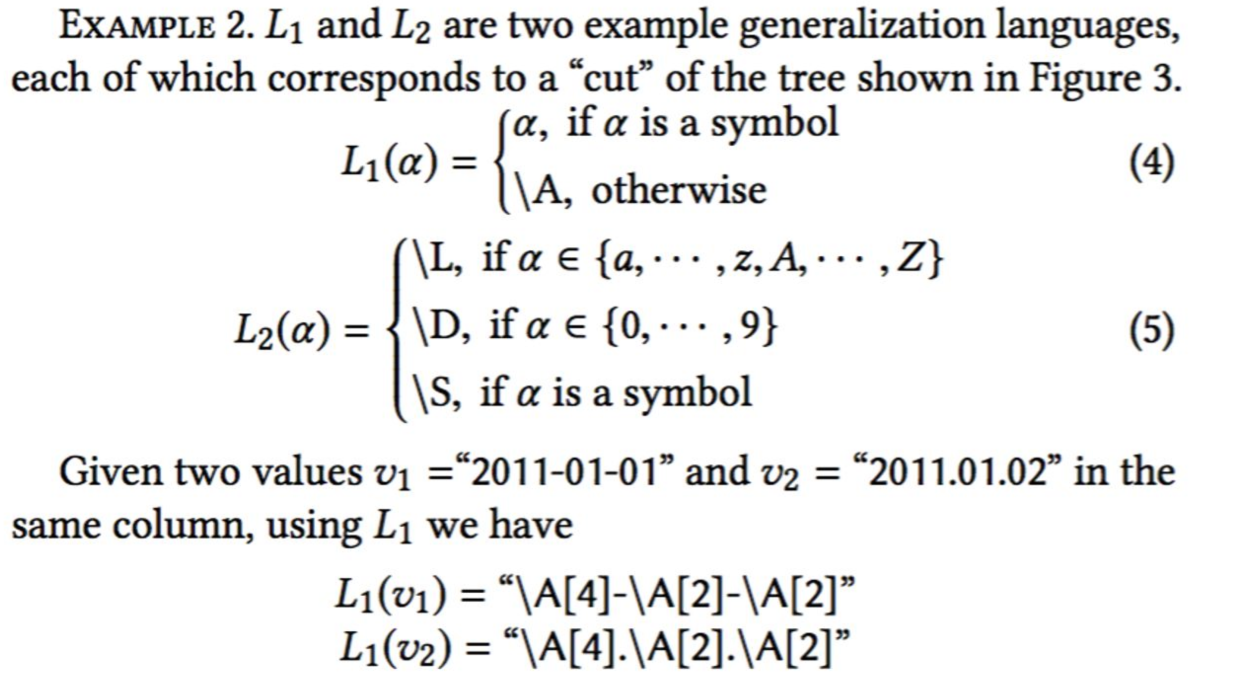


Figure 3: A generalization tree



EXAMPLE 2. L_1 and L_2 are two example generalization languages, each of which corresponds to a “cut” of the tree shown in Figure 3.

$$L_1(\alpha) = \begin{cases} \alpha, & \text{if } \alpha \text{ is a symbol} \\ \backslash\text{A}, & \text{otherwise} \end{cases} \tag{4}$$

$$L_2(\alpha) = \begin{cases} \backslash\text{L}, & \text{if } \alpha \in \{a, \cdots, z, A, \cdots, Z\} \\ \backslash\text{D}, & \text{if } \alpha \in \{0, \cdots, 9\} \\ \backslash\text{S}, & \text{if } \alpha \text{ is a symbol} \end{cases} \tag{5}$$

Given two values v_1 =“2011-01-01” and v_2 = “2011.01.02” in the same column, using L_1 we have

$$L_1(v_1) = \text{“}\backslash\text{A[4]-}\backslash\text{A[2]-}\backslash\text{A[2]”}$$
$$L_1(v_2) = \text{“}\backslash\text{A[4].}\backslash\text{A[2].}\backslash\text{A[2]”}$$

Huang et al.,Auto-Detect: Data-Driven Error Detection in Tables, SigKDD, 2020.

Generic Solution: Detect Column-wise Inconsistency

- Auto-Detect [SIGMOD 2018]
  - Capture "**global**" compatibility

Edward	55
Smith	388
Adam	783
Mike	792
Jane	1,000
Andrew	874

Easy to detect the error with one "local" generalization language

Huang et al.,Auto-Detect: Data-Driven Error Detection in Tables, SigKDD, 2020.

Generic Solution: Detect Column-wise Inconsistency

- Auto-Detect [SIGMOD 2018]
  - Ensemble generalization languages to capture "**global**" compatibility

Edward	55
Smith	388
Adam	783
Mike	792
Jane	1,000
Andrew	874

Derek	1,394
Jennifer	487
Mike	2,499
Andrew	1,983
Jane	1,000
Ethan	874

Jennifer	1155
Mike	31,388
Andrew	648
Edward	11,562
Smith	556
Adam	874

Numbers with separator "," co-occur often with numbers containing no separator

Huang et al.,Auto-Detect: Data-Driven Error Detection in Tables, SigKDD, 2020.

Generic Solution: Detect Column-wise Inconsistency

- Auto-Detect [SIGMOD 2018]

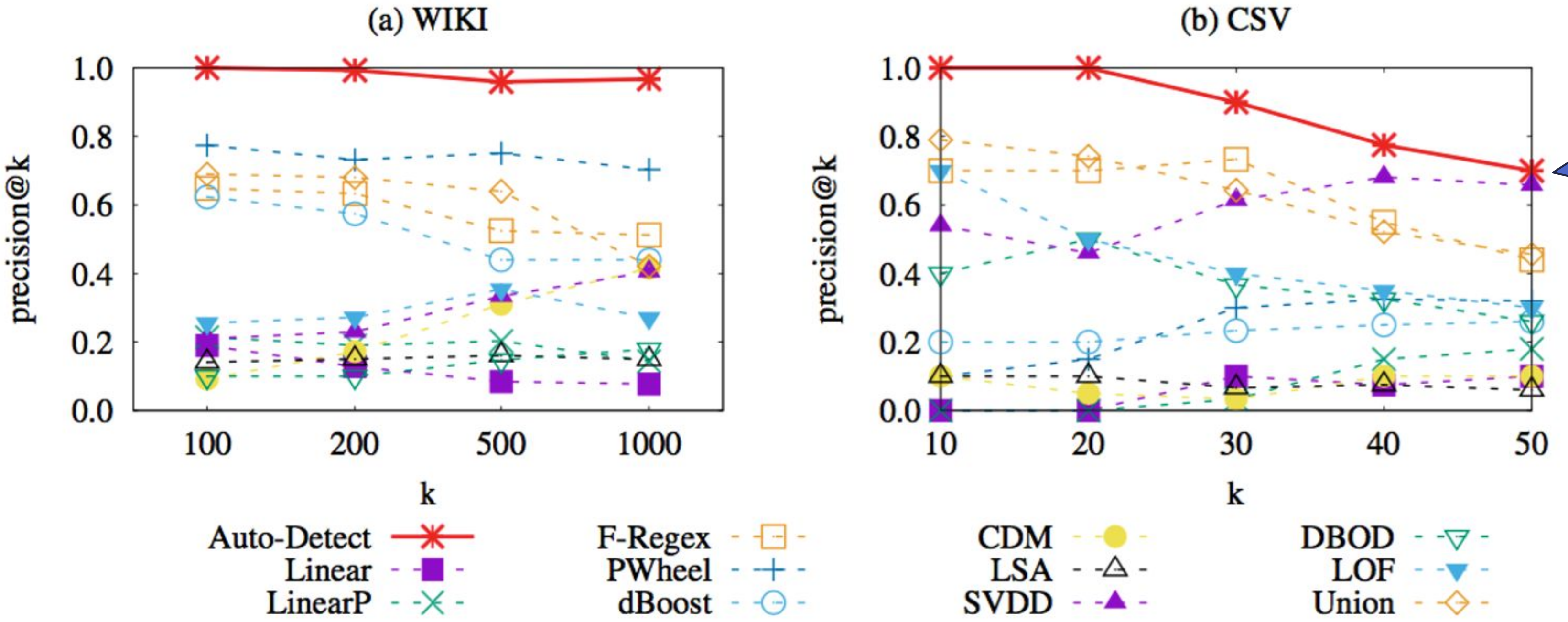


Figure 4: Quality results using manually labeled ground truth

Auto-Detect can find errors with high precision

Huang et al.,Auto-Detect: Data-Driven Error Detection in Tables, SigKDD, 2020.

Product Specific Challenges

ID	Flavor
1	cherry bbq
2	hazelnut & vanilla
3	black olives
4	apple b-b-q
5	dark almond chocolate
6	caperberries 2kg
7	sugar 2kg
8	8 1/2 x 11
9	134 lb
10	4 oz

- More noisy structured data, less of a formatting issue
- Need simpler and less sensitive cleaning solution

Product Specific: Syntactic based Clustering

ID	Flavor
1	cherry bbq
2	hazelnut & vanilla
3	black olives
4	apple b-b-q
5	dark almond chocolate
6	caperberries 2kg
7	sugar 2.0lb
8	8 1/2 x 11
9	134 lb
10	4 oz

Cluster the values based on the **similarity of their syntactic structure**

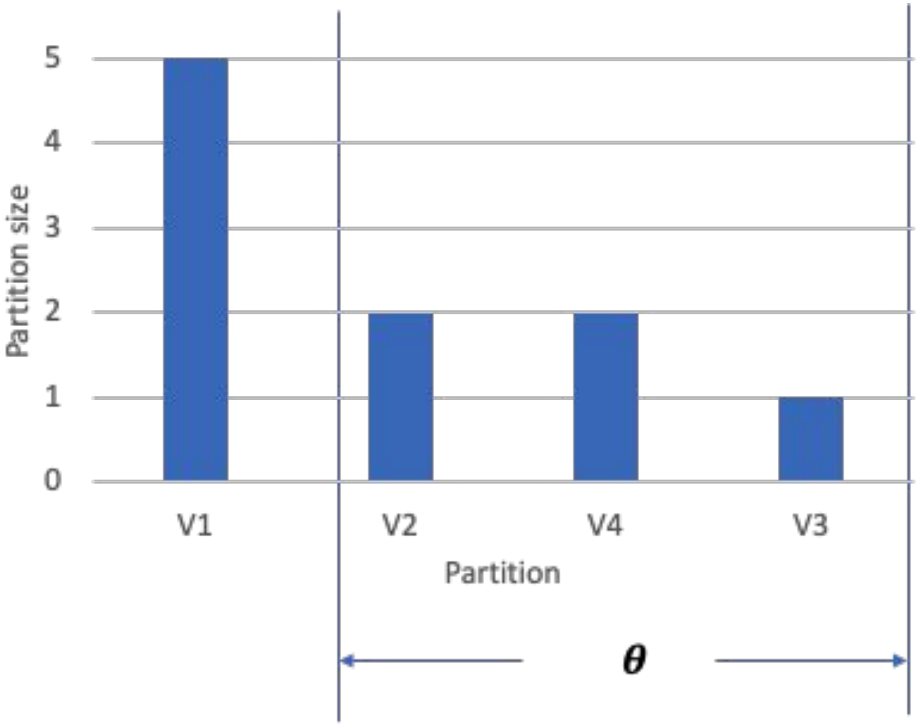
Distance function: Use **descriptive length** to quantify the “generality” of regex pattern *

ID	Flavor values	Partition
1	cherry bbq	V1
2	hazelnut & vanilla	V1
3	black olives	V1
4	apple b-b-q	V1
5	dark almond chocolate	V1
6	caperberries 2kg	V2
7	sugar 2.0lb	V2
8	8 1/2 x 11	V3
9	134 lb	V4
10	4 oz	V4

Product Specific: Syntactic based Clustering

ID	Flavor values	Partition
1	cherry bbq	V1
2	hazelnut & vanilla	V1
3	black olives	V1
4	apple b-b-q	V1
5	dark almond chocolate	V1
6	caperberries 2kg	V2
7	sugar 2.0lb	V2
8	8 1/2 x 11	V3
9	134 lb	V4
10	4 oz	V4

Identify the **tail partitions** as outliers



PG Specific: Syntactic based Clustering

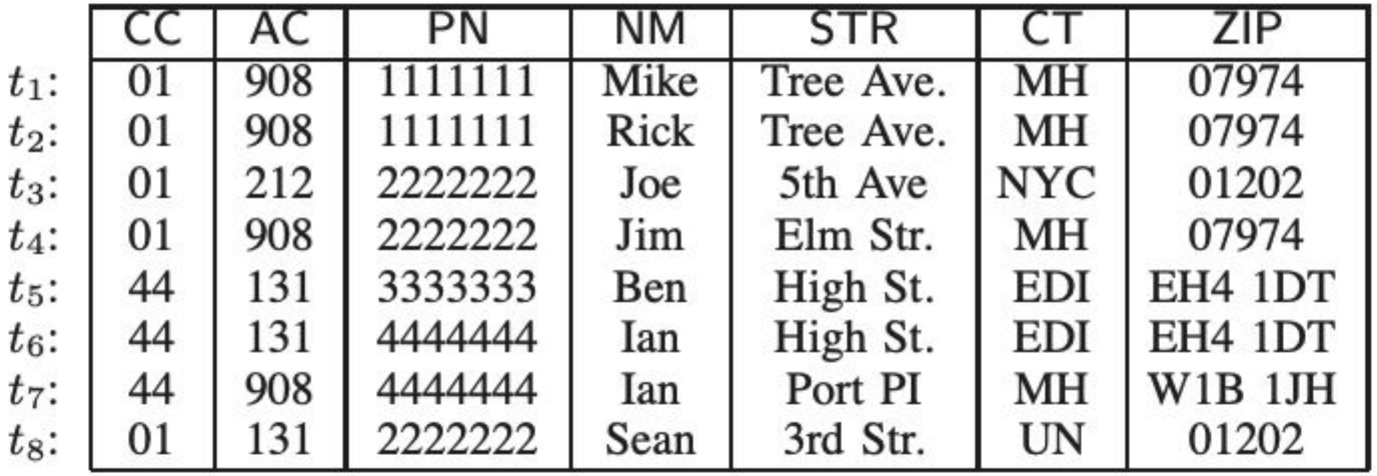
	Values	Outliers found	% Outliers	Precision
Attribute A	90K	4K	2%	81%
Attribute B	80K	3K	4%	89%

Promising precision

- **Unsupervised** model requires no training data
- Detect data errors with **promising precision**
  - Erroneous value like Scent = “90CM X 210CM” will be identified

Generic Solution: Detect Row-wise Inconsistency

- Discover conditional functional dependency [TKDD 2011]



	CC	AC	PN	NM	STR	CT	ZIP
t_1:	01	908	1111111	Mike	Tree Ave.	MH	07974
t_2:	01	908	1111111	Rick	Tree Ave.	MH	07974
t_3:	01	212	2222222	Joe	5th Ave	NYC	01202
t_4:	01	908	2222222	Jim	Elm Str.	MH	07974
t_5:	44	131	3333333	Ben	High St.	EDI	EH4 1DT
t_6:	44	131	4444444	Ian	High St.	EDI	EH4 1DT
t_7:	44	908	4444444	Ian	Port PI	MH	W1B 1JH
t_8:	01	131	2222222	Sean	3rd Str.	UN	01202

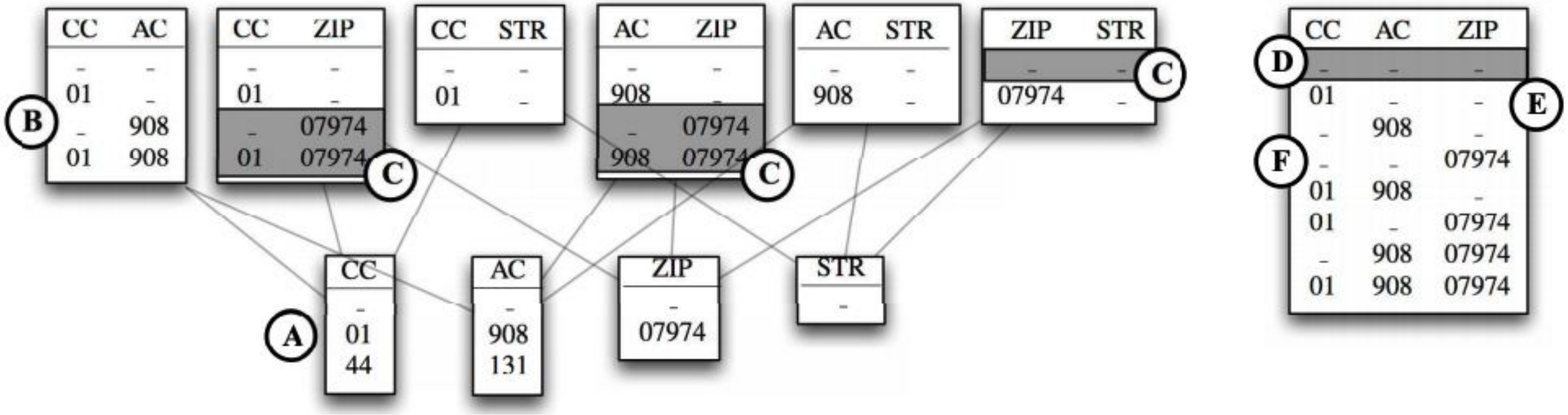
ϕ_0: ([CC, ZIP] → STR, (44, _ ‖ _))
ϕ_1: ([CC, AC] → CT, (01, 908 ‖ MH))
ϕ_2: ([CC, AC] → CT, (44, 131 ‖ EDI))
ϕ_3: ([CC, AC] → CT, (01, 212 ‖ NYC))

CC (country_code) as “01” + AC (area_code) as “212” determine CT (city) as “NYC”

- CFD is the form of (X->A, t_p), X->A is an FD and t_p is a patten tuple with attributes in X and A. t_p is either a constant or an unnamed variable “_”

Fan et al.,Discover Conditional Functional Dependency, TKDE, 2011.

Generic Solution: Detect Row-wise Inconsistency



CC	AC
_	_
01	_
_	908
01	908

CC	ZIP
_	_
01	_
_	07974
01	07974

CC	STR
_	_
01	_

AC	ZIP
_	_
908	_
_	07974
908	07974

AC	STR
_	_
908	_

ZIP	STR
_	_
07974	_

CC
_
01
44

AC
_
908
131

ZIP
_
07974

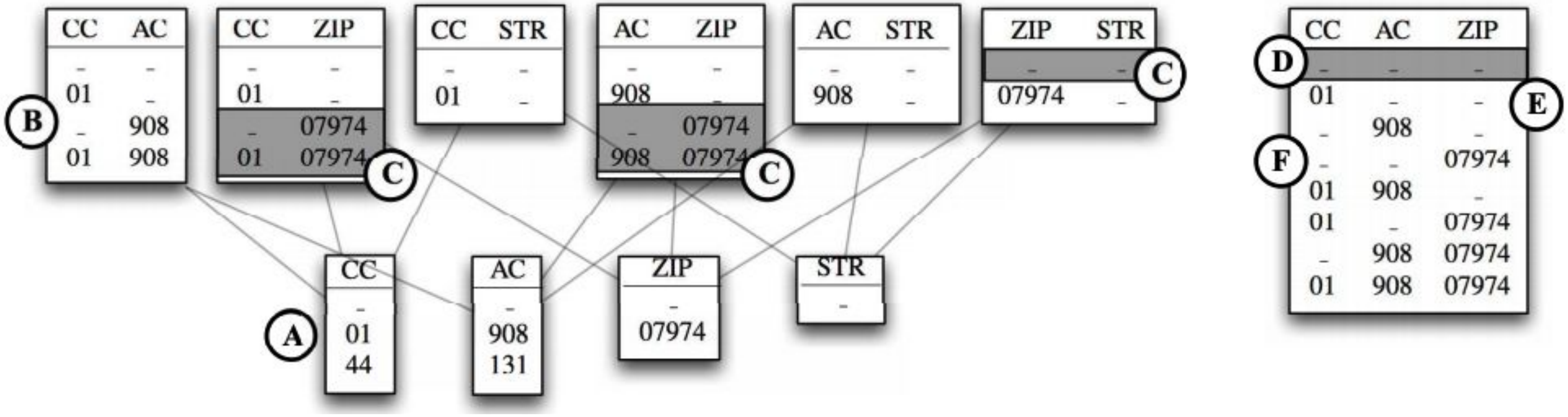
STR
_

CC	AC	ZIP
_	_	_
01	_	_
_	908	_
_	_	07974
01	908	_
01	_	07974
_	908	07974
01	908	07974

- A: Initially find all single attribute/value pairs that appear at least k times
- B: Pair attributes together and creates consistent patterns
- C: For the gray shaded patterns, finds valid CFDs
- D: creates triples of CFDs

Fan et al.,Discover Conditional Functional Dependency, TKDE, 2011.

Generic Solution: Detect Row-wise Inconsistency



CC	AC
_	_
01	_
_	908
01	908

CC	ZIP
_	_
01	_
_	07974
01	07974

CC	STR
_	_
01	_

AC	ZIP
_	_
908	_
_	07974
908	07974

AC	STR
_	_
908	_

ZIP	STR
_	_
07974	_

CC
_
01
44

AC
_
908
131

ZIP
_
07974

STR
_

CC	AC	ZIP
_	_	_
01	_	_
_	908	_
_	_	07974
01	908	_
01	_	07974
_	908	07974
01	908	07974

- E: Update support set, not only of the current pattern but also of those with a more specific pattern on the LHS-attributes
- F: Compute the pattern tuples

Fan et al.,Discover Conditional Functional Dependency, TKDE, 2011.

Product Specific Challenges

- Rich unstructured textual data
- Big variety across product types

Product	Brand	Color	Scent	Skin type
Anthony All-Purpose Facial Moisturizer – Men's Hydrating Lotion for **Dry Skin** – Lightweight, Non-Comedogenic, Anti-Aging Formula – 3 Fl. Oz	Anthony		scented	**normal**
CeraVe Moisturizing Cream | Body and Face Moisturizer for Dry Skin | Body Cream with Hyaluronic Acid and Ceramides | 19 Ounce	CeraVe		lavender	dry skin
…	…	…	…	…

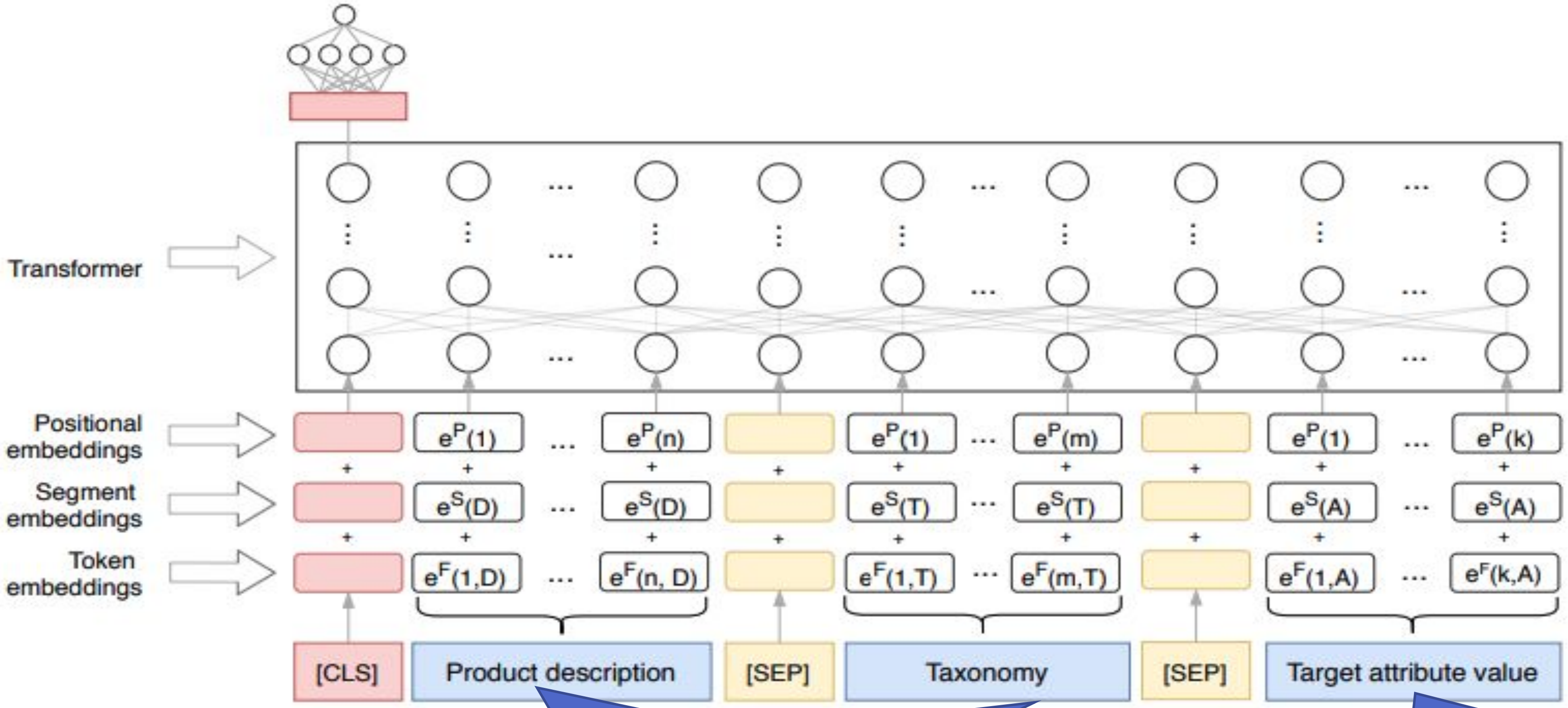
Functional Dependency is not sufficient to detect the inconsistency

Product Specific: Cleaning in Auto-Know

- Auto-Know [KDD 2020]
  - **Transformer-based** model jointly processing signals from product profile, product taxonomy via multi-head attention to decide if an attribute value is correct
  - Model is **taxonomy-aware**
  - Training Data: Use existing catalog data for **distant supervision**

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

Product Specific: Cleaning in Auto-Know



Product title, description, bullets

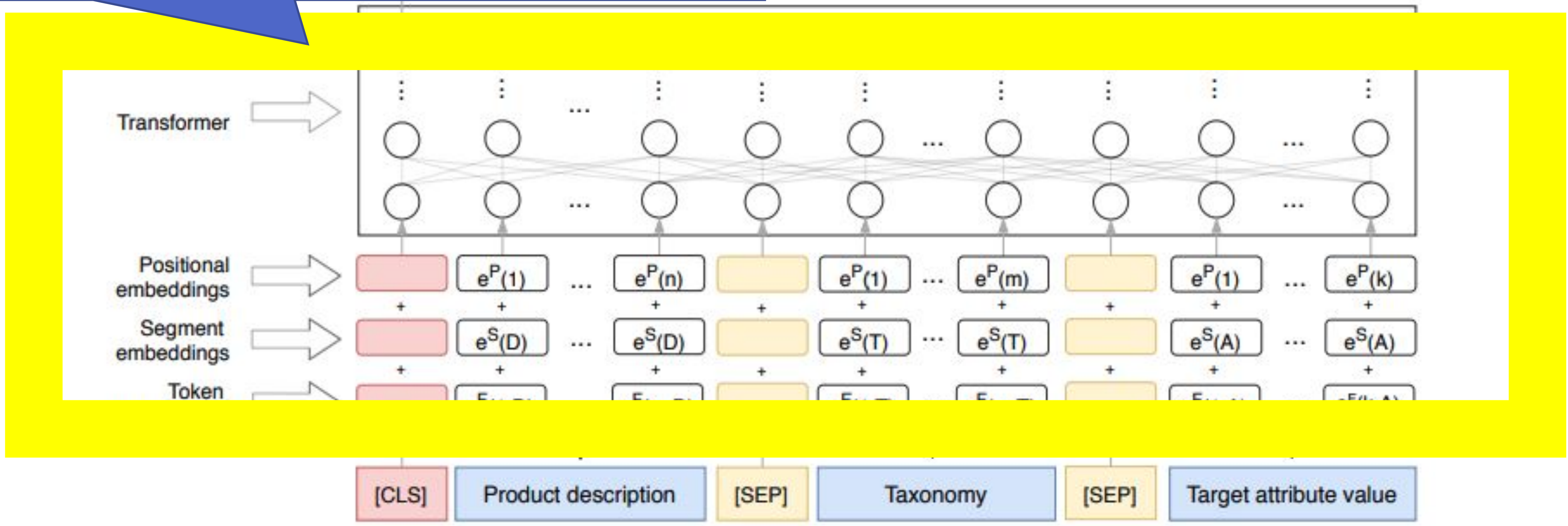
Category as input for model training

The attribute value

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

Product Specific: Cleaning in Auto-Know

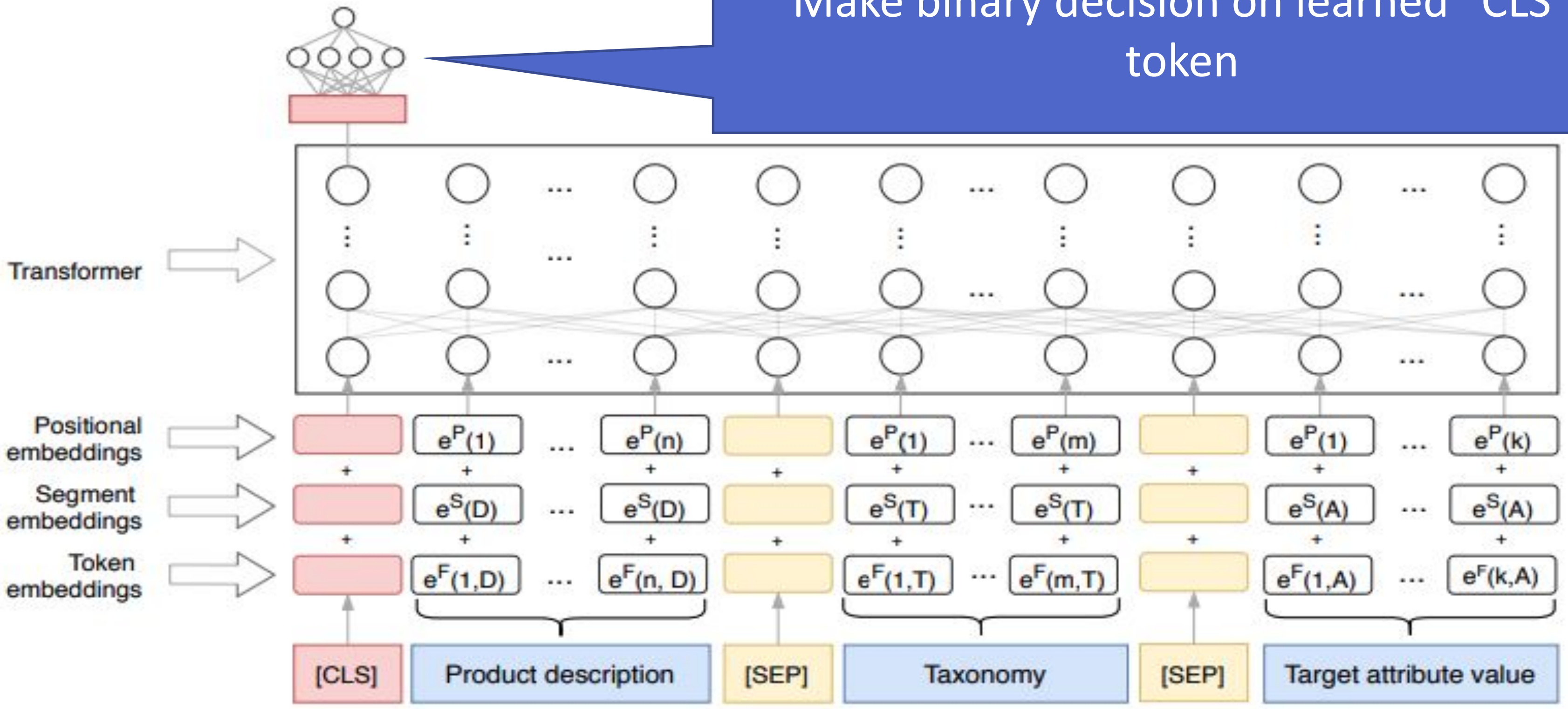
Learn the semantic consistency among product profile, taxonomy and attribute value with **Transformer model**



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

Product Specific: Cleaning in Auto-Know

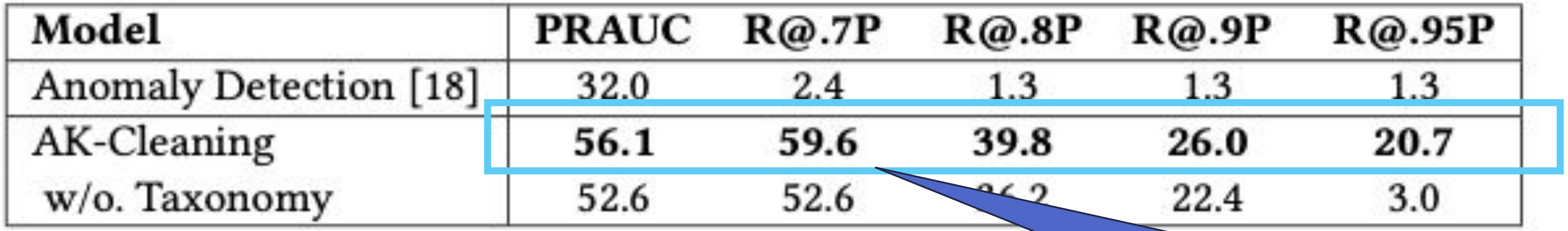
Make binary decision on learned “CLS” token



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

Product Specific: Cleaning in Auto-Know

- Experiment
  - Evaluated on 223 product categories



Model	PRAUC	R@.7P	R@.8P	R@.9P	R@.95P
Anomaly Detection [18]	32.0	2.4	1.3	1.3	1.3
AK-Cleaning	**56.1**	**59.6**	**39.8**	**26.0**	**20.7**
w/o. Taxonomy	52.6	52.6	36.2	22.4	3.0

- Rich text of unstructured attributes helps cleaning
- Taxonomy signal is critical

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

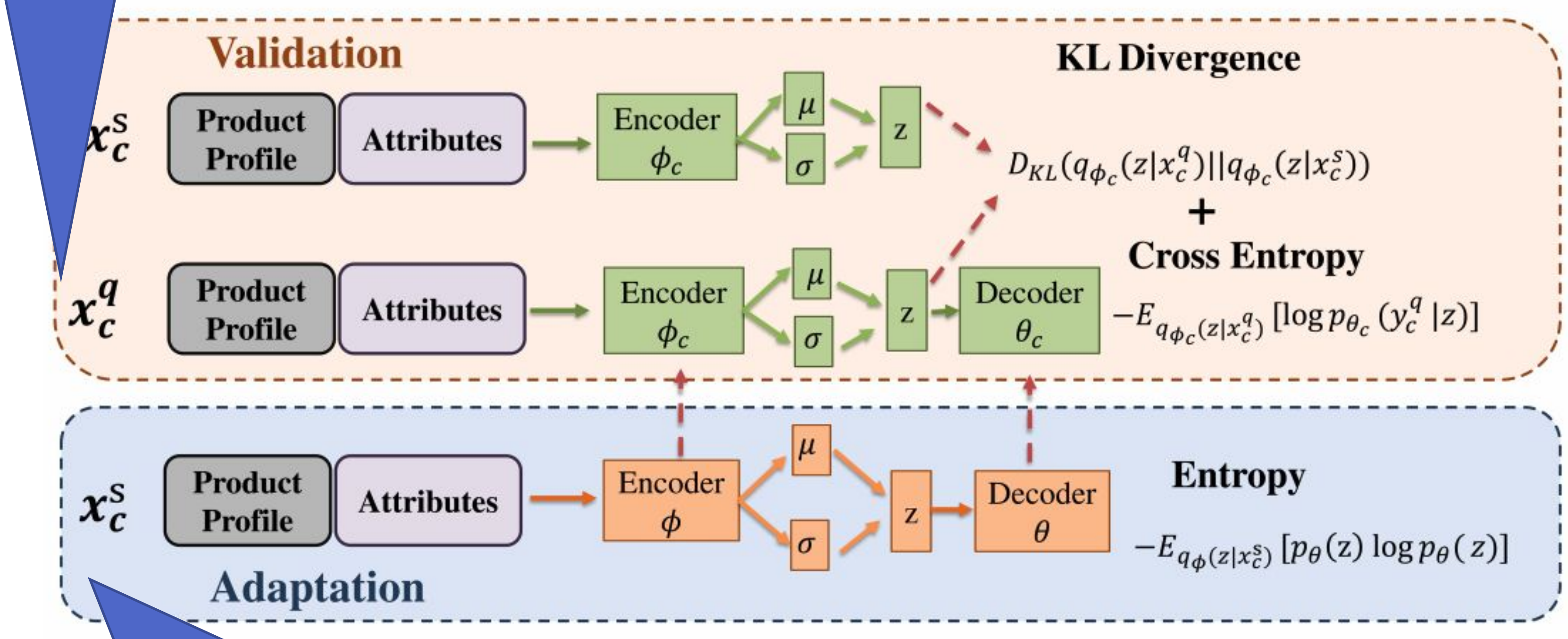
Product Specific: MetaBridge

- MetaBridge [KDD 2020]
  - **Few-shot learning** setting to address the lack of training data issue, especially to handle a large number of product categories
  - Meta-learning approach: leverage **labeled data** from a small number of categories for training category-agnostic models and utilize **unlabeled data** to capture category-specific information

Wang et al., Automatic Validation of Textual Attribute Values in ECommerce Catalog by Learning with Limited Labeled Data, SigKDD, 2020.

PG Specific: MetaBridge

Limited labeled data as query set



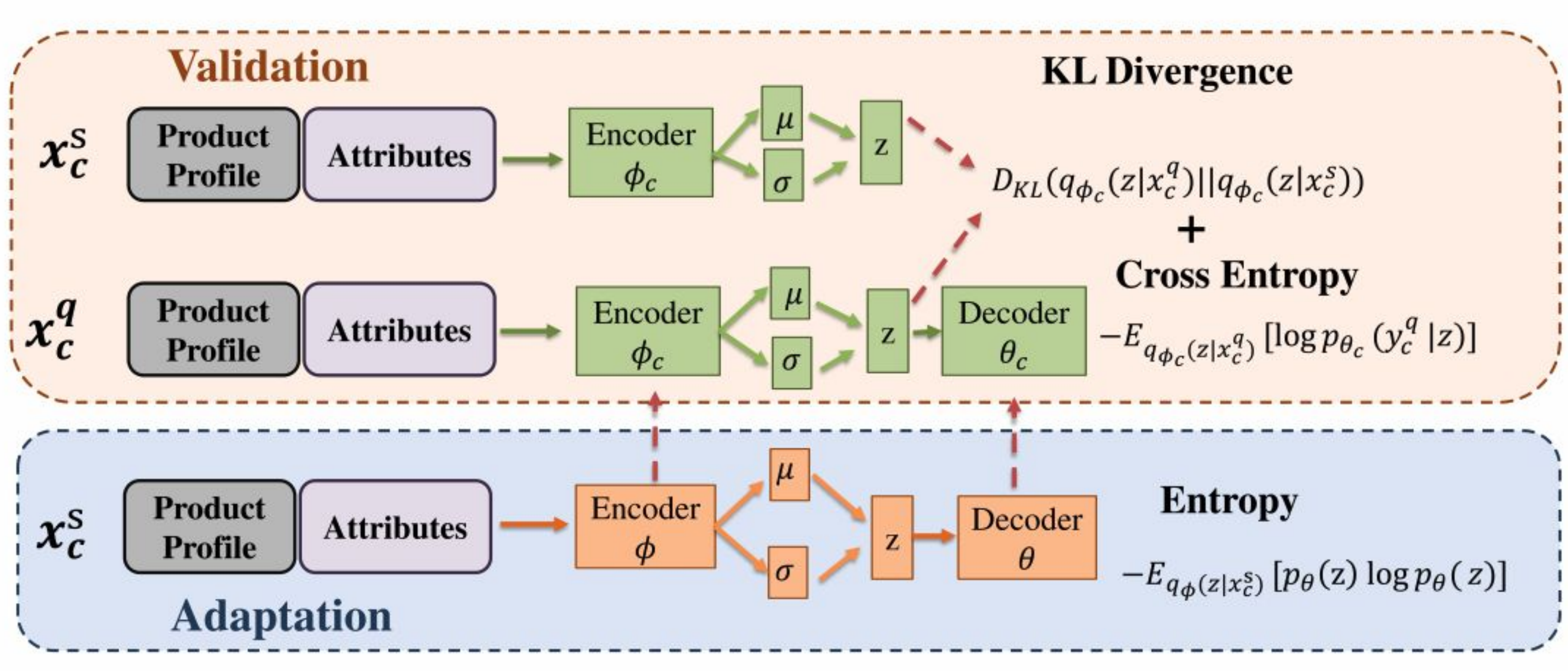
Validation: The adapted model is used to validate attributes for products in category c

Adaptation: Model parameter is updated to capture the uncertainty of category c

Unlabeled support set

Wang et al., Automatic Validation of Textual Attribute Values in ECommerce Catalog by Learning with Limited Labeled Data, SigKDD, 2020.

PG Specific: MetaBridge



The objective function includes: **supervised inference loss** and **bridging regularizer**

Define the loss function on the unlabeled support set by entropy minimization

Wang et al., Automatic Validation of Textual Attribute Values in ECommerce Catalog by Learning with Limited Labeled Data, SigKDD, 2020.

PG Specific: MetaBridge

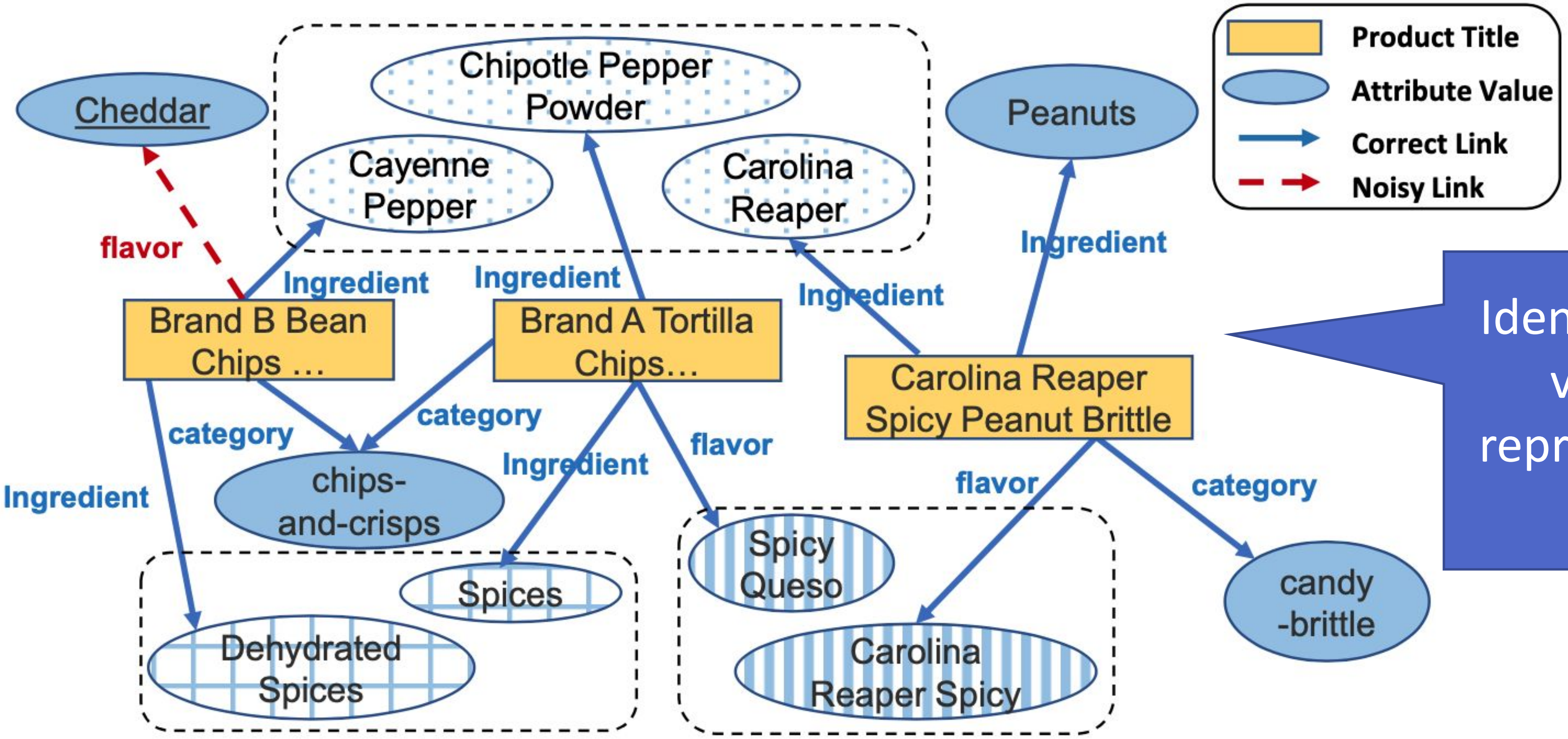
Setting	Method	Flavor		Ingredient	
		PRAUC	R@P=0.9	PRAUC	R@P=0.9
Supervised	RF	0.6986	4.43	0.4683	14.69
Fine-tune	BERT	0.7599	27.76	0.5292	17.00
Meta-Learning	MAML	0.7486	22.62	0.5289	22.48
Meta-Learning	MetaBridge	0.7852	30.77	0.5658	27.00

MetaBridge makes best use of training labels and unlabeled data, outperforms supervised and fine-tuning methods

658 categories. Each category has 5 labeled data as query set and 100 unlabeled data as support set

Wang et al., Automatic Validation of Textual Attribute Values in ECommerce Catalog by Learning with Limited Labeled Data, SigKDD, 2020.

Generic Solution: Source-wise Inconsistency



Identify inconsistency that violates the learned representations of entities and relations

Generic Solution: Source-wise Inconsistency

- Trans-E [NeurIPs 2013]
  - Treat relations as the translation operations between vectors corresponding to entities
  - Learn embeddings by minimizing a margin-based ranking criterion over the training set
  - Corrupt triples by replacing training triples with either head or tail replaced by a random entity

Bordes et al., Translating Embeddings for Modeling Multi-relational Data. NeurIPs 2013

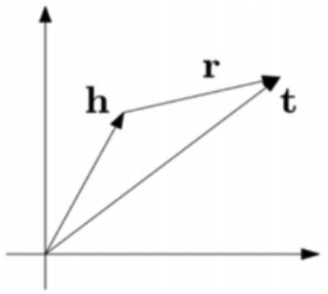
Generic Solution: Graph embedding

- Trans-E [NeurIPs 2013]
  - The score function of (h, r, t)

$$f_r(h,t) = - \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_{L_1/L_2}$$

  - Loss function

$$L = \underset{\substack{\downarrow \\ \text{Positive} \\ \text{triple set}}}{\sum_{(h,r,t)\in\triangle}} \; \underset{\substack{\downarrow \\ \text{Negative} \\ \text{triple set}}}{\sum_{(h',r,t')\in\triangle'}} \max(0, f_r(h,t) + \underset{\substack{\downarrow \\ \text{Optimal Margin}}}{M_{opt}} - f_r(h',t'))$$



h + r = t

China + Capital = Beijing

France + Capital = Paris

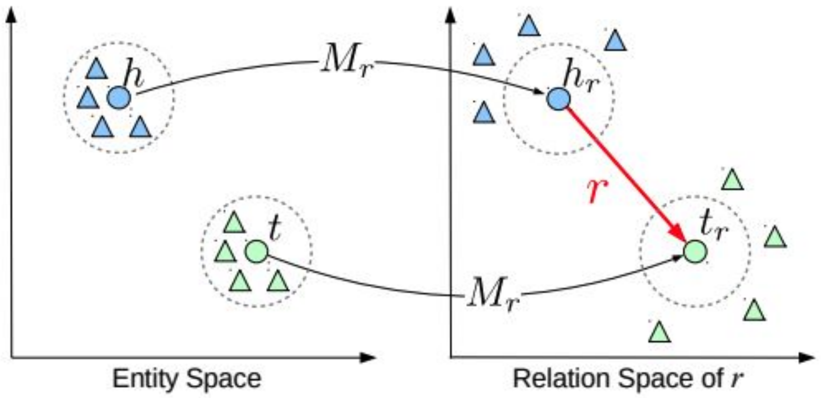
Bordes et al., Translating Embeddings for Modeling Multi-relational Data. NeurIPs 2013

Generic Solution: Graph embedding

- Trans-R [AAAI 2015]
  - For each triple (h, r, t), entities in the entity space are first projected into r-relation space as hr and tr with operation Mr, then h_r + r = t_r
  - Scoring function of (h, r, t)

$$\mathbf{h}_r = \mathbf{h}\mathbf{M}_r, \quad \mathbf{t}_r = \mathbf{t}\mathbf{M}_r.$$

$$f_r(h,t) = \|\mathbf{h}_r + \mathbf{r} - \mathbf{t}_r\|_2^2$$



Lin et al., Learning Entity and Relation Embeddings for Knowledge Graph Completion. In AAAI, 2015

Generic Solution: Graph embedding

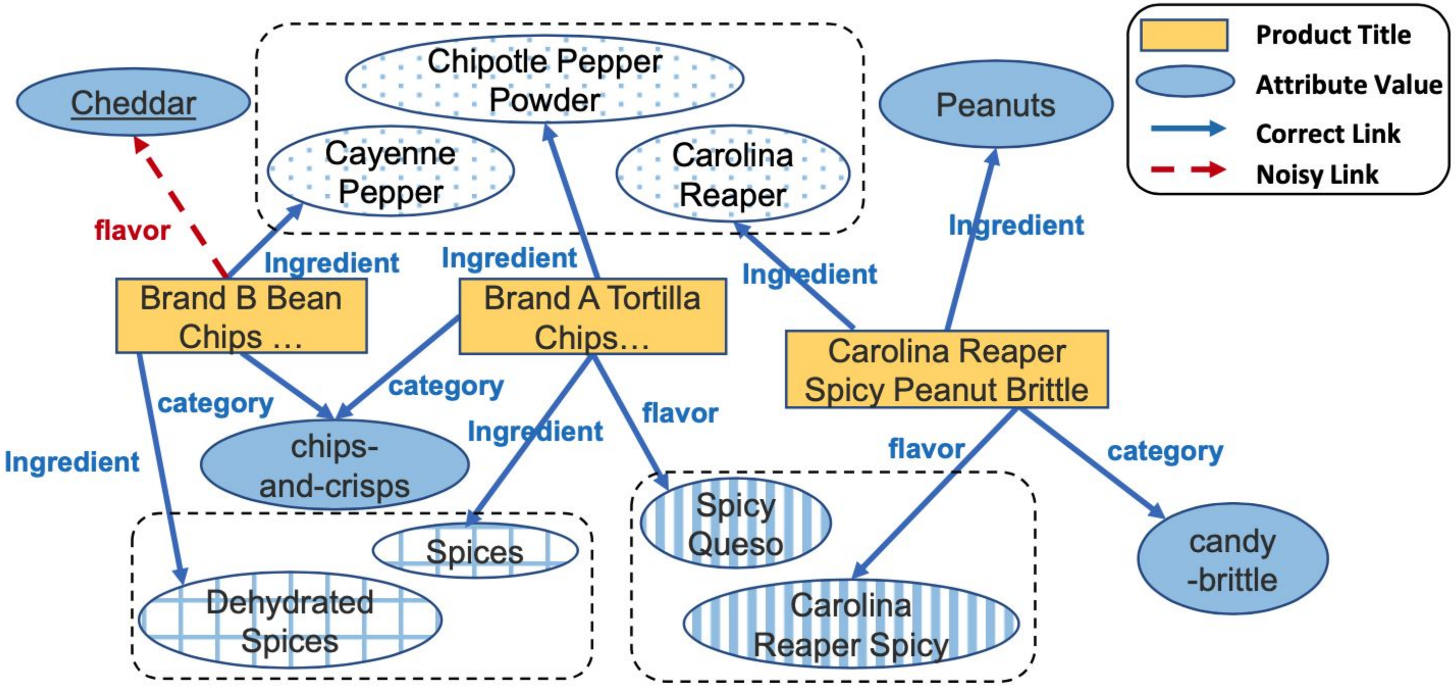
Data Sets	**WN 11**	**FB13**	**FB15K**
TransE	75.9	70.9	79.6
TransH	77.7	**76.5**	79.0
TransR	**85.5**	74.7	**81.7**

Knowledge Graph Embedding methods showed promising precision in detecting data errors

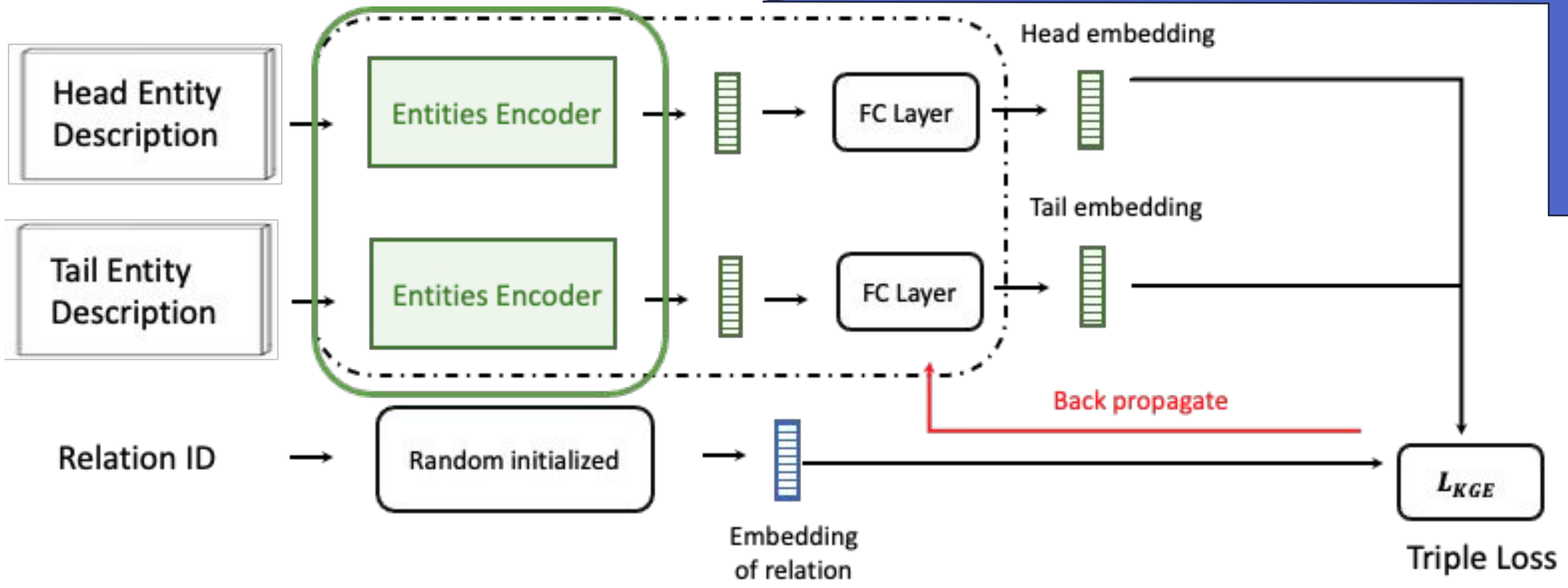
Lin et al., Learning Entity and Relation Embeddings for Knowledge Graph Completion. In AAAI, 2015

Product Specific Challenges

- Text data heavy instead of entity heavy graph



Product Specific: Semantic Knowledge Embedding



Add an entity encoder to encode the semantic information of entities

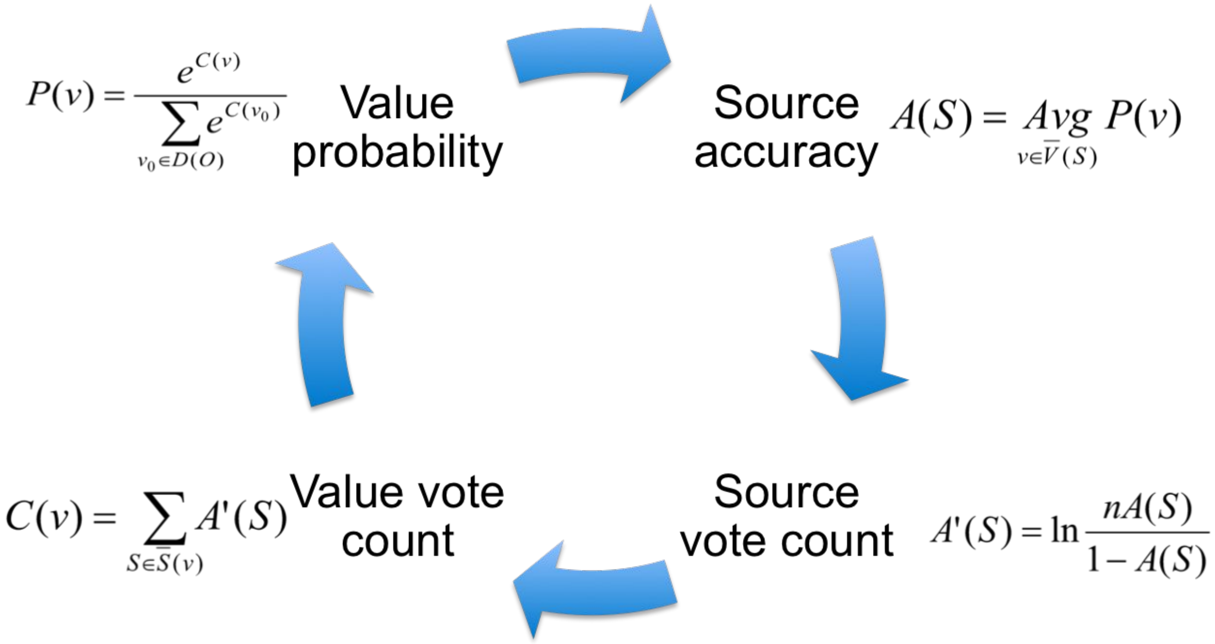
Product Specific: Semantic Knowledge Embedding

Methods	R@P=60	R@P=70	R@P=80	R@P=90
Vanilla KGE	0.466	0.390	0.308	0.213
Semantic KGE	**0.846**	**0.662**	**0.425**	**0.286**

- Incorporating rich semantic information to the graph embedding learning has significantly improved the performance of cleaning

Generic Solution: Knowledge Fusion

- ACCU [VLDB 2013]

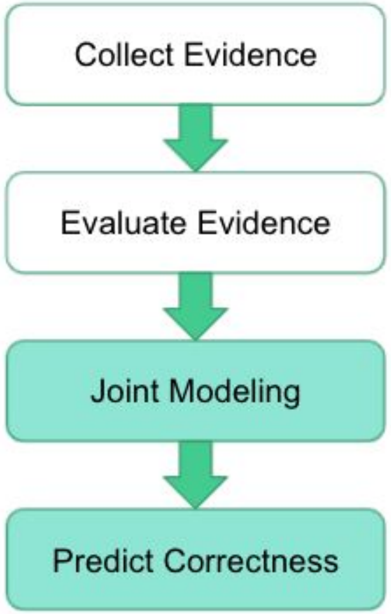


$$P(v) = \frac{e^{C(v)}}{\sum_{v_0 \in D(O)} e^{C(v_0)}}$$

$$A(S) = \underset{v \in \bar{V}(S)}{Avg}\ P(v)$$

$$A'(S) = \ln \frac{nA(S)}{1 - A(S)}$$

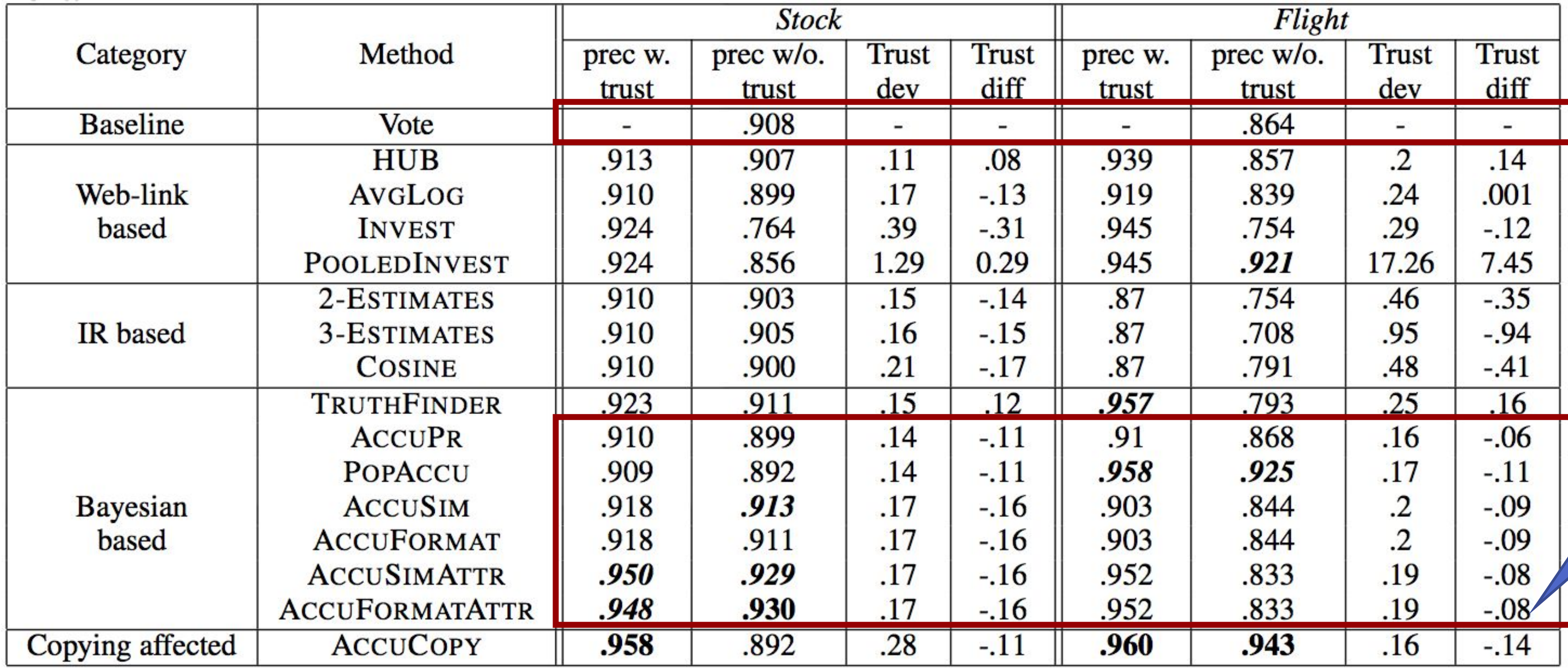
$$C(v) = \sum_{S \in \bar{S}(v)} A'(S)$$



Li et al., Truth finding on the Deep Web: Is the problem solved? In VLDB, 2013

Generic Solution: Knowledge Fusion

- ACCU [VLDB 2013]



Category	Method	*Stock*				*Flight*			
		prec w. trust	prec w/o. trust	Trust dev	Trust diff	prec w. trust	prec w/o. trust	Trust dev	Trust diff
Baseline	Vote	-	.908	-	-	-	.864	-	-
Web-link based	HUB	.913	.907	.11	.08	.939	.857	.2	.14
	AvgLog	.910	.899	.17	-.13	.919	.839	.24	.001
	Invest	.924	.764	.39	-.31	.945	.754	.29	-.12
	PooledInvest	.924	.856	1.29	0.29	.945	***.921***	17.26	7.45
IR based	2-Estimates	.910	.903	.15	-.14	.87	.754	.46	-.35
	3-Estimates	.910	.905	.16	-.15	.87	.708	.95	-.94
	Cosine	.910	.900	.21	-.17	.87	.791	.48	-.41
Bayesian based	TruthFinder	.923	.911	.15	.12	***.957***	.793	.25	.16
	AccuPr	.910	.899	.14	-.11	.91	.868	.16	-.06
	PopAccu	.909	.892	.14	-.11	***.958***	***.925***	.17	-.11
	AccuSim	.918	***.913***	.17	-.16	.903	.844	.2	-.09
	AccuFormat	.918	.911	.17	-.16	.903	.844	.2	-.09
	AccuSimAttr	***.950***	***.929***	.17	-.16	.952	.833	.19	-.08
	AccuFormatAttr	***.948***	***.930***	.17	-.16	.952	.833	.19	-.08
Copying affected	AccuCopy	**.958**	.892	.28	-.11	**.960**	**.943**	.16	-.14

Leverage source trustworthiness significantly improve the fact checking accuracy

Li et al., Truth finding on the Deep Web: Is the problem solved? In VLDB, 2013

Reflections/Short-answers

- **Definition**: Finding wrong attribute values
- **Recipe**: Identify data inconsistency column-wise, row-wise, source-wise and across sources
- **Key to Success for Products**:
  - Leverage rich textual information of unstructured data as context
  - Solution with aware of taxonomy
- **Applicability to Other Domains:**
  - Domains with heavy text data
  - Rich taxonomy information
  - Domains like: medical, legal, etc.

Future Directions

- **Ensemble** the methods that identify data inconsistency from different aspects
- Incorporate **common sense knowledge** like ConceptNet to clean the data
- Enhance the **interpretability** of knowledge cleaning decisions
- Distinguish data errors and **inapplicability**

Questions?

Overview and Introduction

Knowledge Extraction

Knowledge Cleaning

Q&A **10 min**

Break

Ontology Mining

Applications

Conclusion and Future Directions

Q&A

Break

Overview and Introduction

Knowledge Extraction

Knowledge Cleaning

Q&A

Break **20 min**

Ontology Mining

Applications

Conclusion and Future Directions

Q&A