A Reference-Set Approach to Information Extraction from Unstructured, Ungrammatical Data Sources

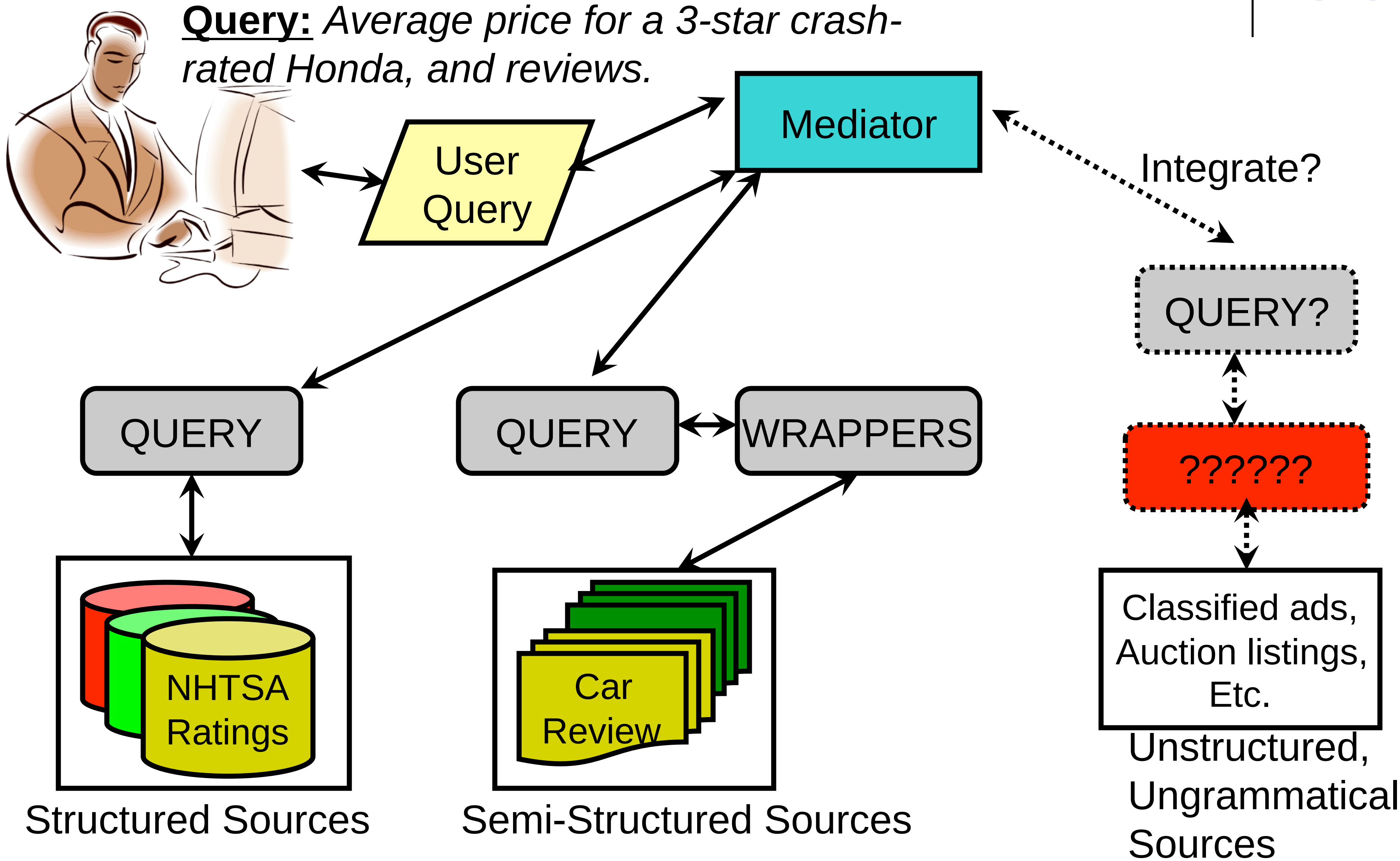
Matthew Michelson

Ph.D. Defense

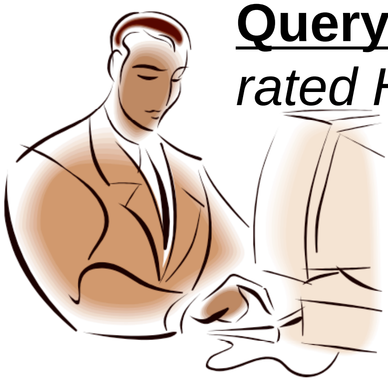
Nov. 3rd, 2008

Motivation: Data Integration

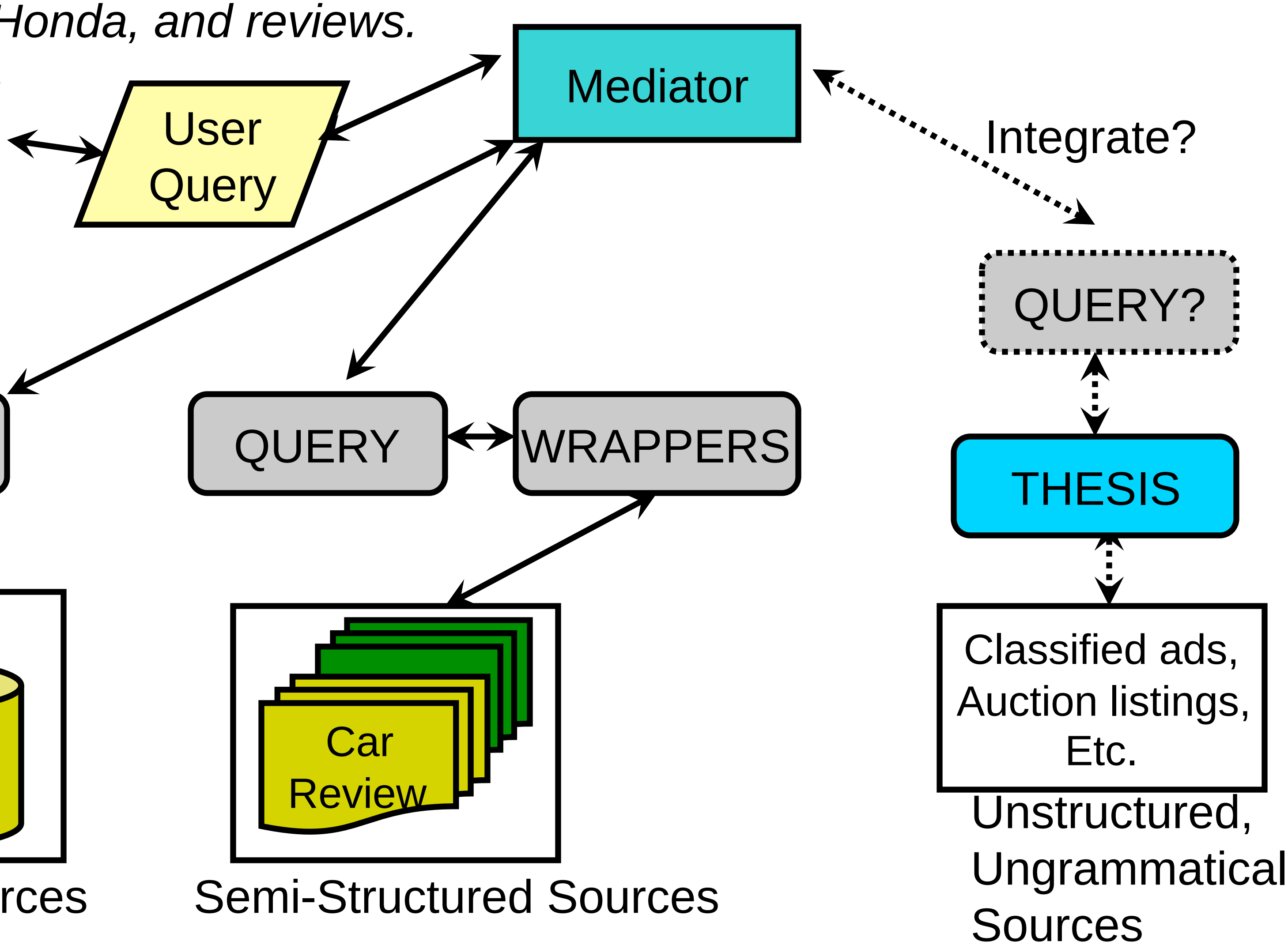
Query: *Average price for a 3-star crash-rated Honda, and reviews.*



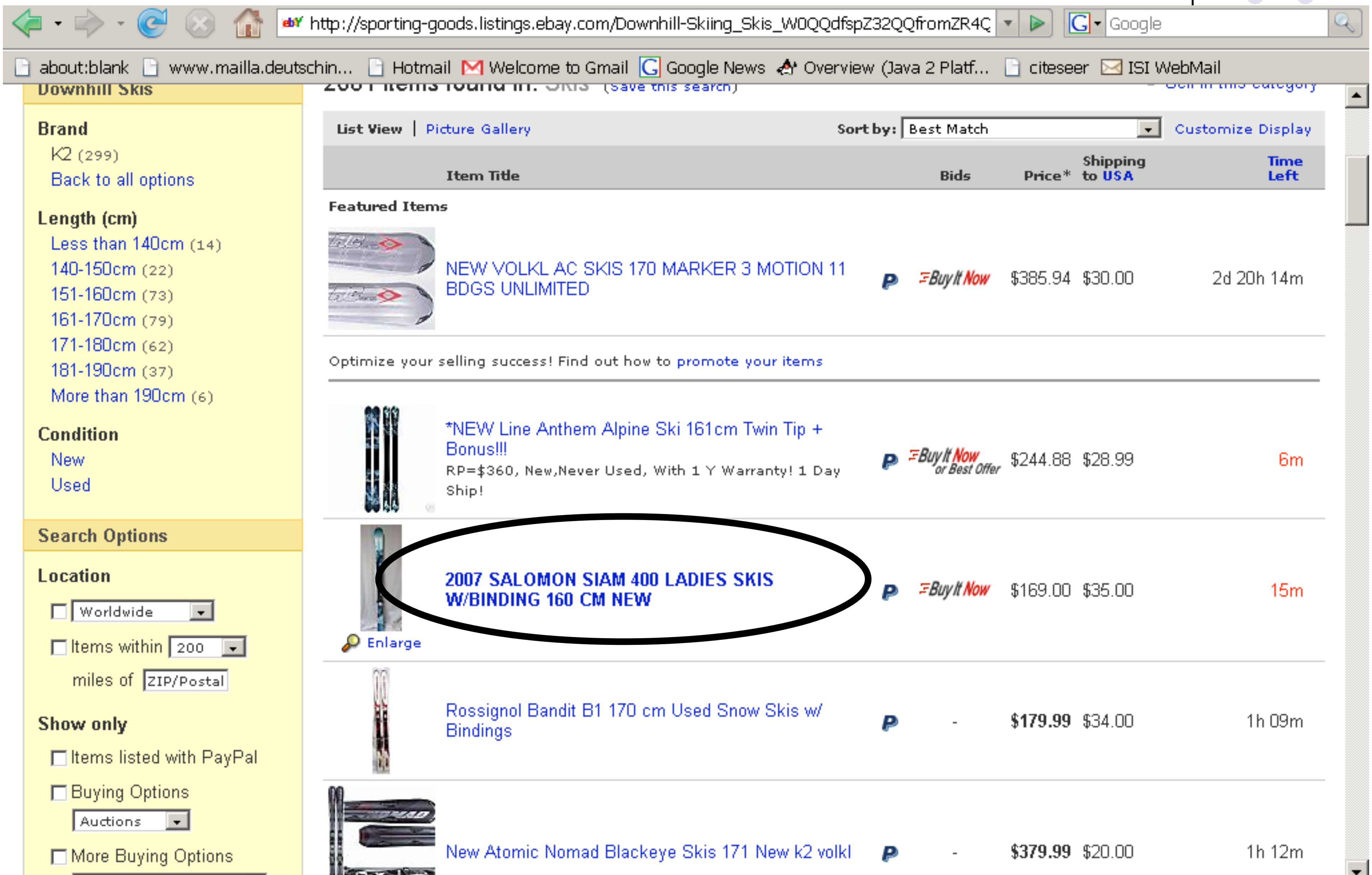
Motivation: Data Integration



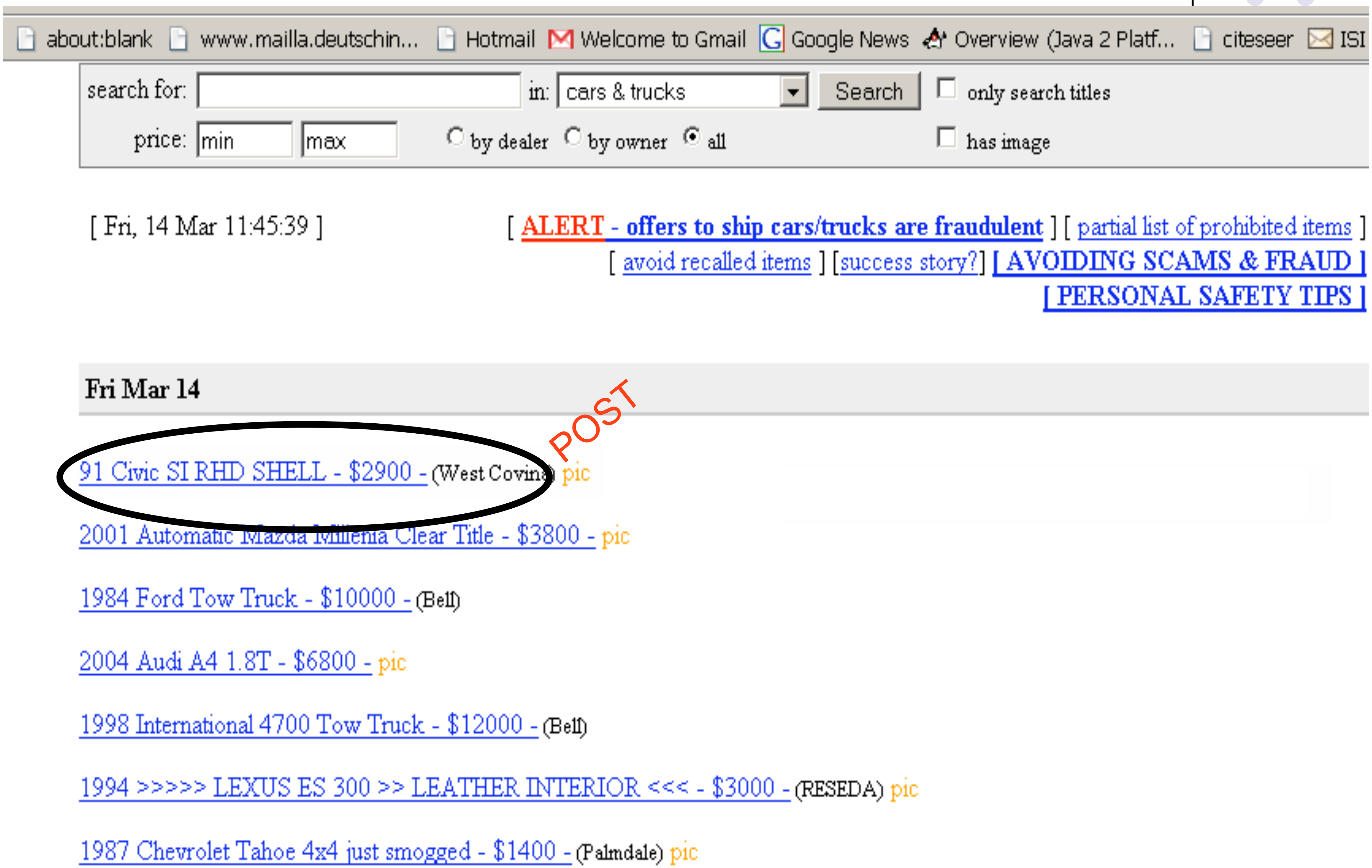
Query: *Average price for a 3-star crash-rated Honda, and reviews.*



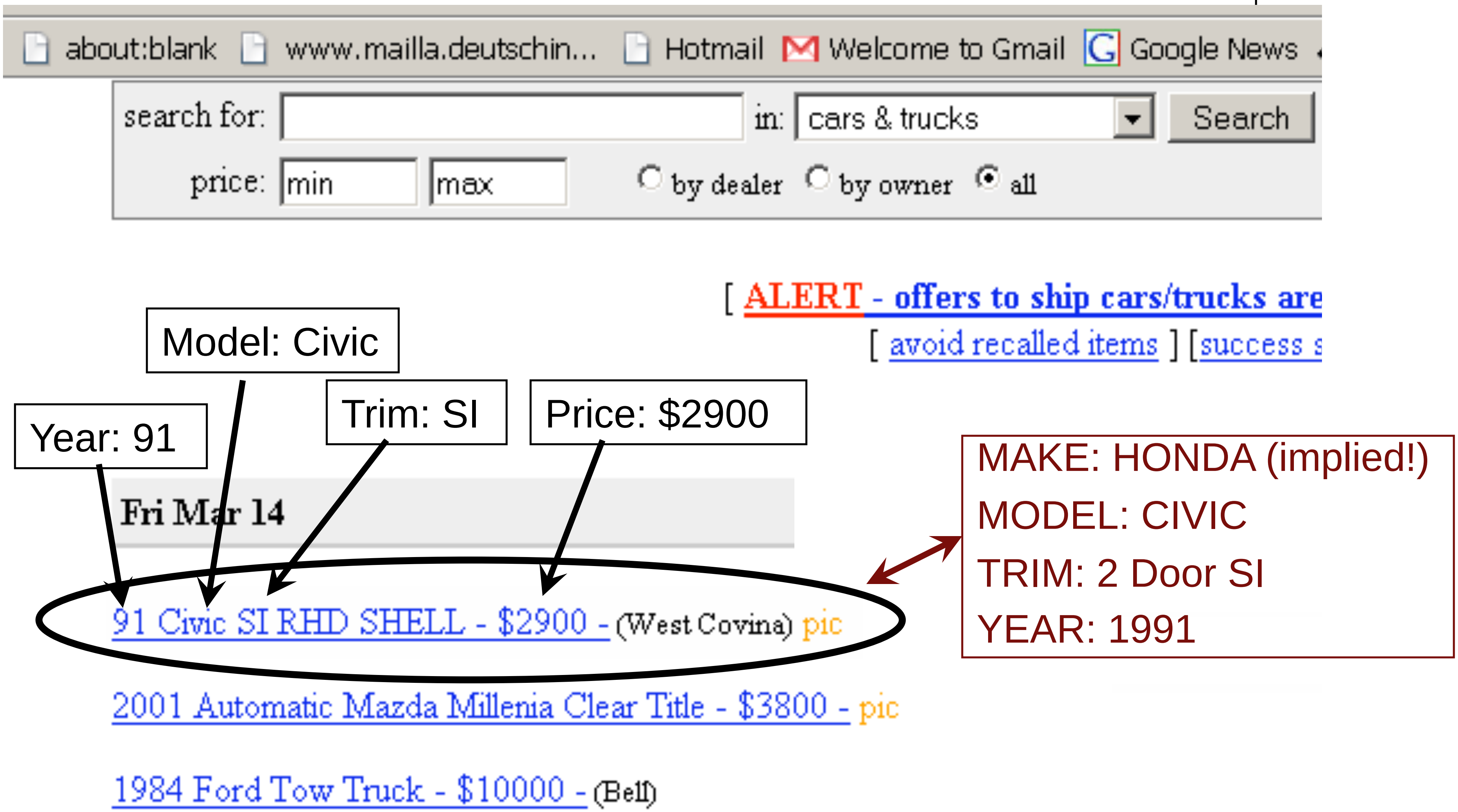
Unstructured, Ungrammatical Data: "Posts"



Unstructured, Ungrammatical Data: "Posts"



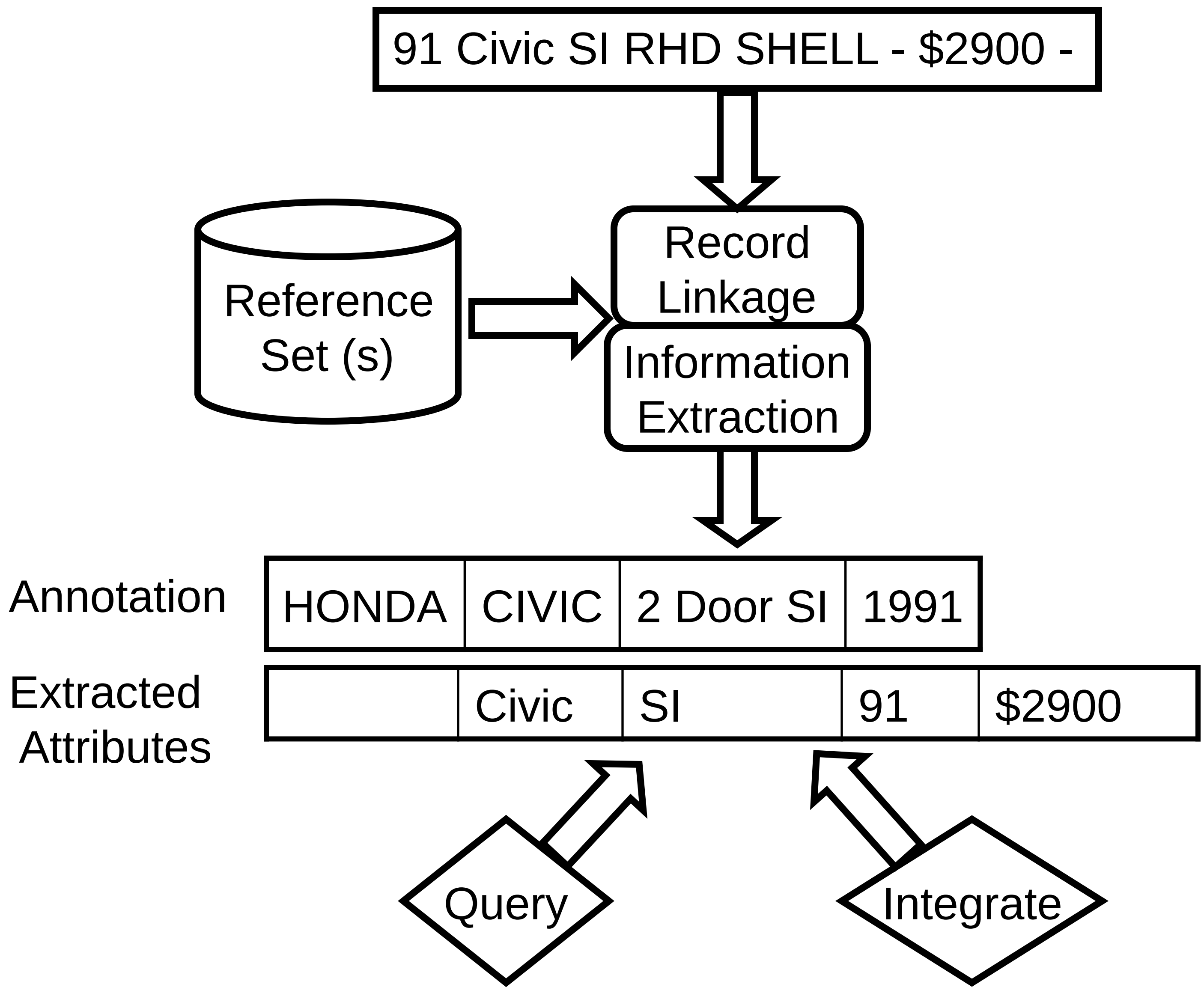
Query? …
Information Extraction/Annotation!



Difficulties

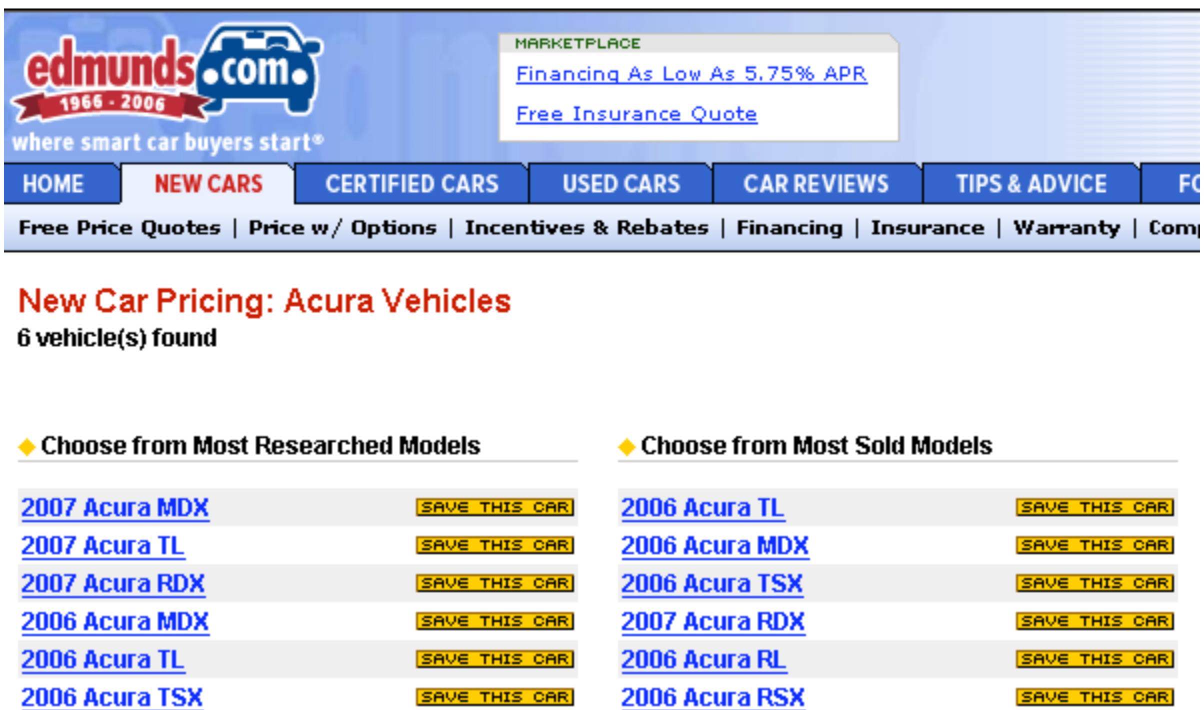
- Unstructured
  - No assumptions on structure
  - “Rule/Pattern” based techniques unsuited
- Ungrammatical
  - Does not conform to English grammar
  - Natural-Language Processing techniques unsuited

Reference-Set Based Extraction/ Annotation



Reference Sets

- Collections of entities and their attributes
  - List cars →<make, model, trim, …>



Scrape make, model, trim, year for all cars from 1990-2005…

Contributions

- Automatic matching and extraction algorithm that exploits a given reference set
  - Automatically select the appropriate reference sets from a repository of reference sets
- Automatic method for building reference sets from the posts themselves
  - Suggest the number of posts required to sufficiently build reference set
  - Algorithm to determine whether automatic method will work, or user should create reference set
- Supervised machine learning for high-accuracy
  - High accuracy, even in the face of ambiguity

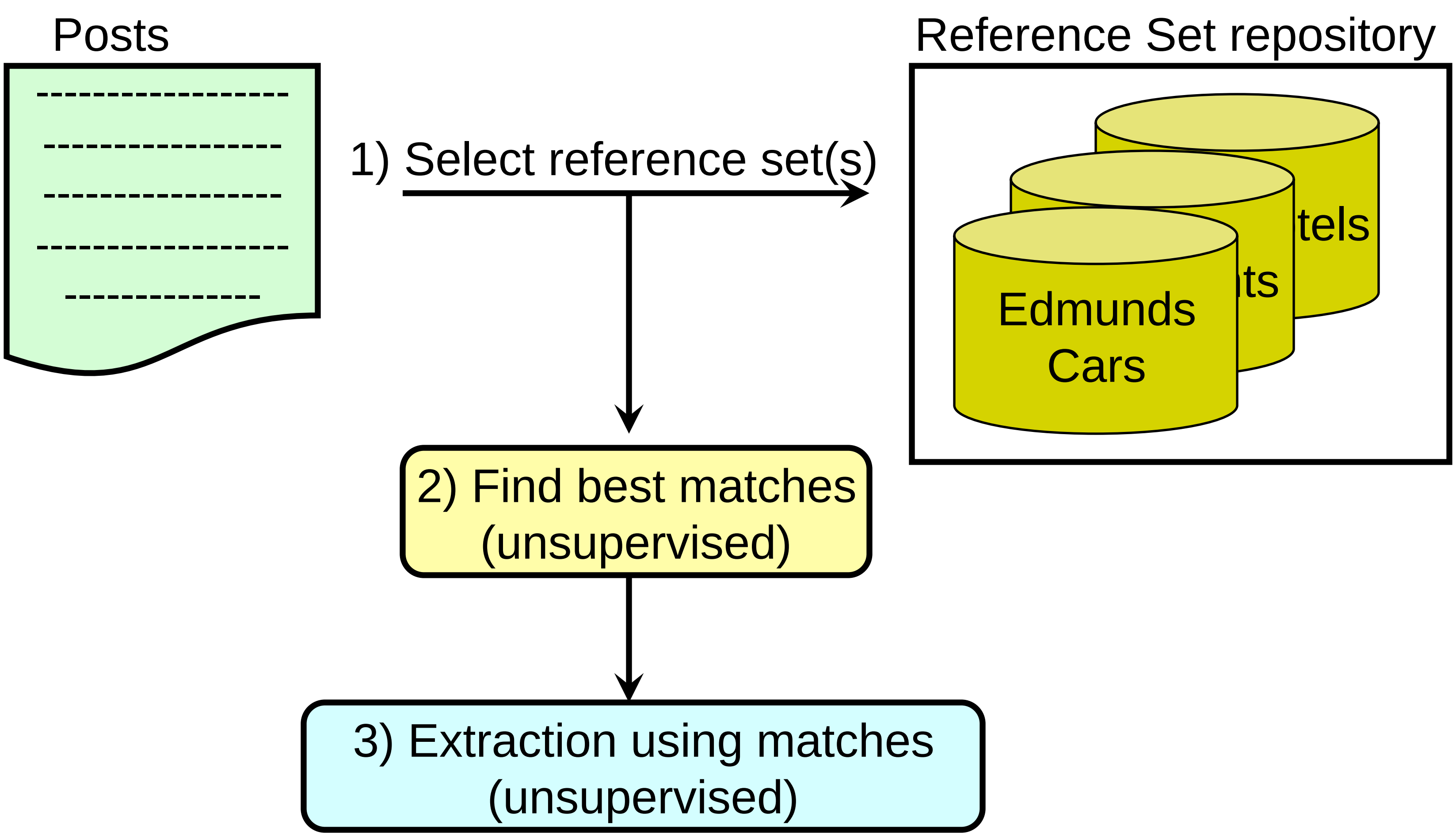
Contributions

3 reference-set based extraction methods

	Summary	***Advantages***
Method 1 (ARX) [IJDAR 07]	1. Automatically select reference set from repository 2. Automatic extraction	• State-of-the-art extraction • Automatic, given reference set
Method 2 (ILA) [JAIR, review]	1. Automatically build reference set	• Cannot build reference set (difficult attributes) • Fully automatic • Competitive state-of-the-art
Method 3 (Phoebus) [JAIR, 08]	1. Supervised approach to extraction	• Highest-accuracy extraction • Deals with ambiguity

Automatic method: Three steps

IJDAR, 2007



ARX: Automatic Reference-set based eXtraction

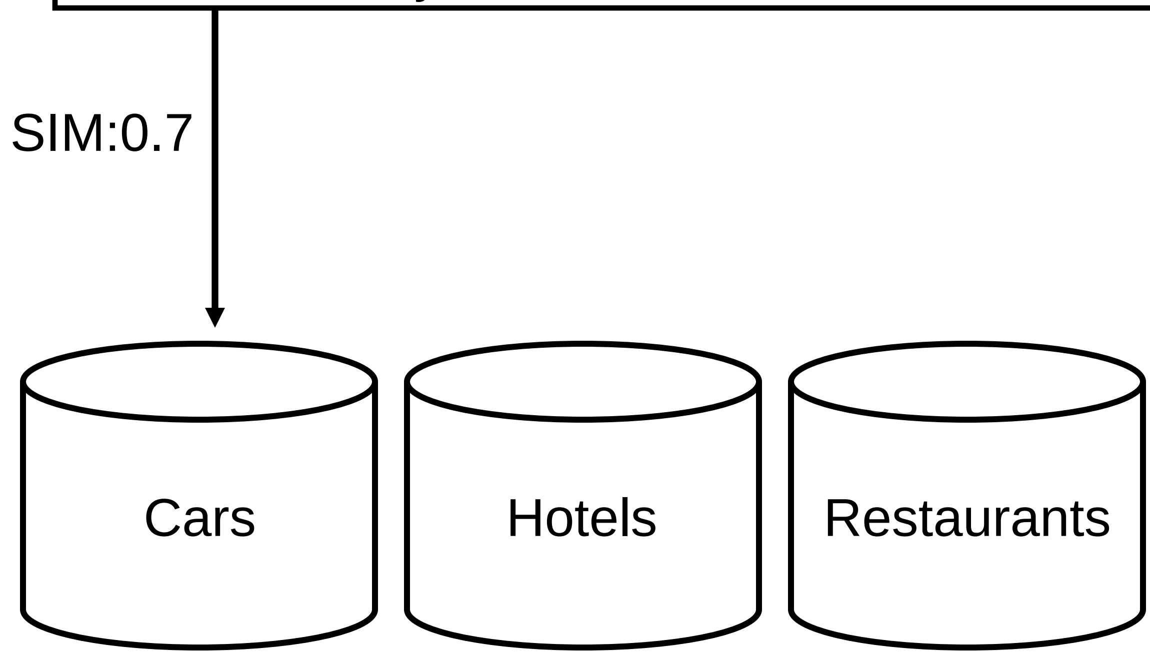
Selecting the Reference Set(s)

Vector space model: set of posts are 1 doc, reference sets are 1 doc

Select reference set most similar to the set of posts…

FORD Thunderbird - $4700

2001 White Toyota Corrolla CE Excellent Condition - $8200



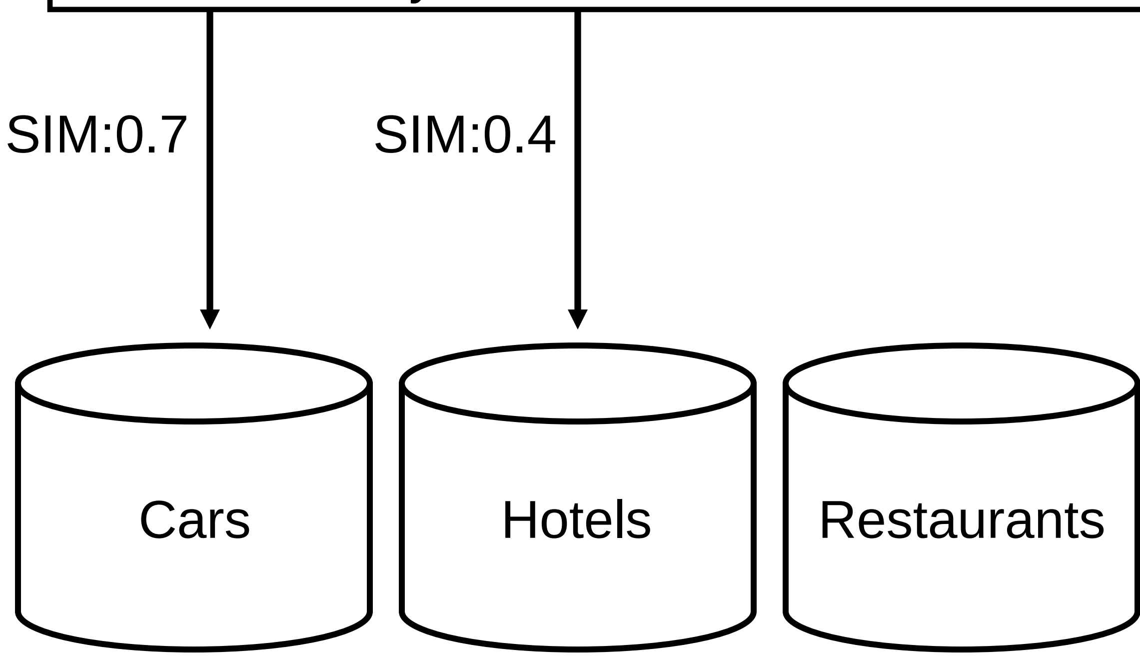
Selecting the Reference Set(s)

Vector space model: set of posts are 1 doc, reference sets are 1 doc

Select reference set most similar to the set of posts…

FORD Thunderbird - $4700

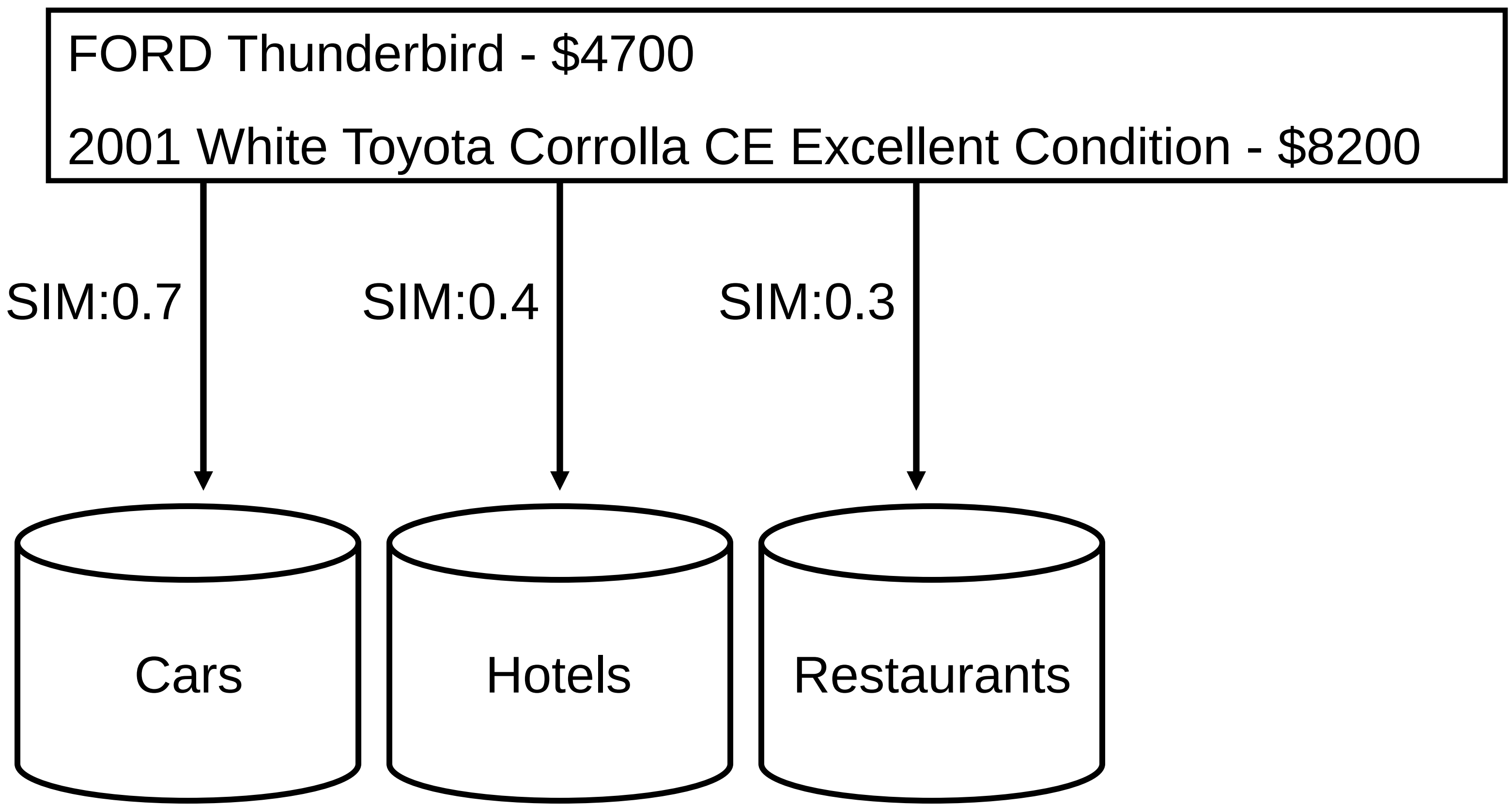
2001 White Toyota Corrolla CE Excellent Condition - $8200



Selecting the Reference Set(s)

Vector space model: set of posts are 1 doc, reference sets are 1 doc

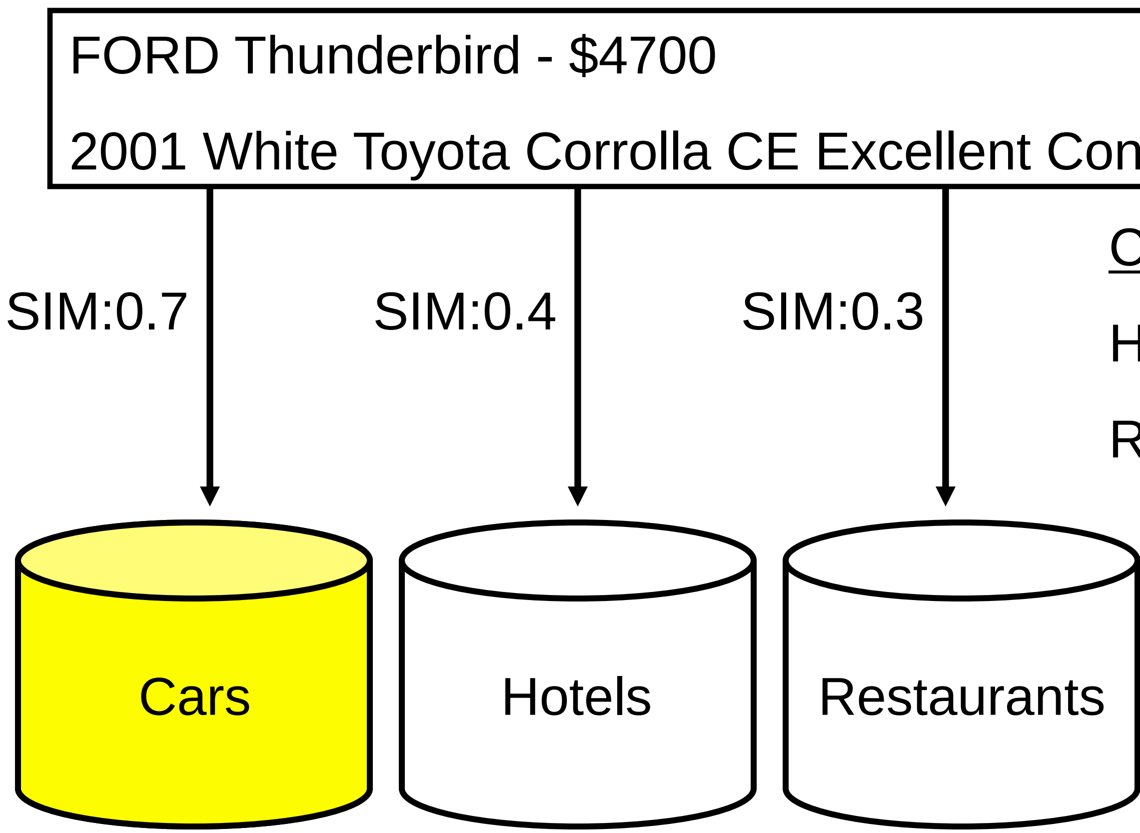
Select reference set most similar to the set of posts…



Selecting the Reference Set(s)

Vector space model: set of posts are 1 doc, reference sets are 1 doc

Select reference set most similar to the set of posts…



<u>Cars 0.7</u> <u>PD(C,H) = 0.75 > T</u>

Hotels 0.4 PD(H,R) = 0.33 < T

Restaurants 0.3

Avg. 0.47

Unsupervised matching between the posts and reference set

new 2007 altima

02 M3 Convertible .. Absolute beauty!!!

Awesome car for sale! Cheap too!

{NISSAN, ALTIMA, 4 Dr 3.5 SE Sedan, 2007}

{NISSAN, ALTIMA, 4 Dr 2.5 S Sedan, 2007}

→ {NISSAN, ALTIMA, 2007}

Unsupervised matching between the posts and reference set

new 2007 altima

02 M3 Convertible .. Absolute beauty!!!

Awesome car for sale! Cheap too!

{NISSAN, ALTIMA, 4 Dr 3.5 SE Sedan, 2007}

{NISSAN, ALTIMA, 4 Dr 2.5 S Sedan, 2007}

Vector-based matching

{NISSAN, ALTIMA, 2007}

Unsupervised matching between the posts and reference set

new 2007 altima

02 M3 Convertible .. Absolute beauty!!!

Awesome car for sale! Cheap too!

Vector-based matching

~~{NISSAN, ALTIMA, 4 Dr 3.5 SE Sedan, 2007}~~ → {NISSAN, ALTIMA, 2007}

~~{NISSAN, ALTIMA, 4 Dr 2.5 S Sedan, 2007}~~

{BMW, M3, 2 Dr STD Convertible, 2002}

{LINCOLN, TOWN CAR, 4 Dr, 2001}

{RENAULT, LE CAR, 2 Dr, 1987}

Unsupervised matching between the posts and reference set

new 2007 altima

02 M3 Convertible .. Absolute beauty!!!

Awesome car for sale! Cheap too!

Vector-based matching

{NISSAN, ALTIMA, 4 Dr 3.5 SE Sedan, 2007}

{NISSAN, ALTIMA, 4 Dr 2.5 S Sedan, 2007}

→ {NISSAN, ALTIMA, 2007}

{BMW, M3, 2 Dr STD Convertible, 2002}

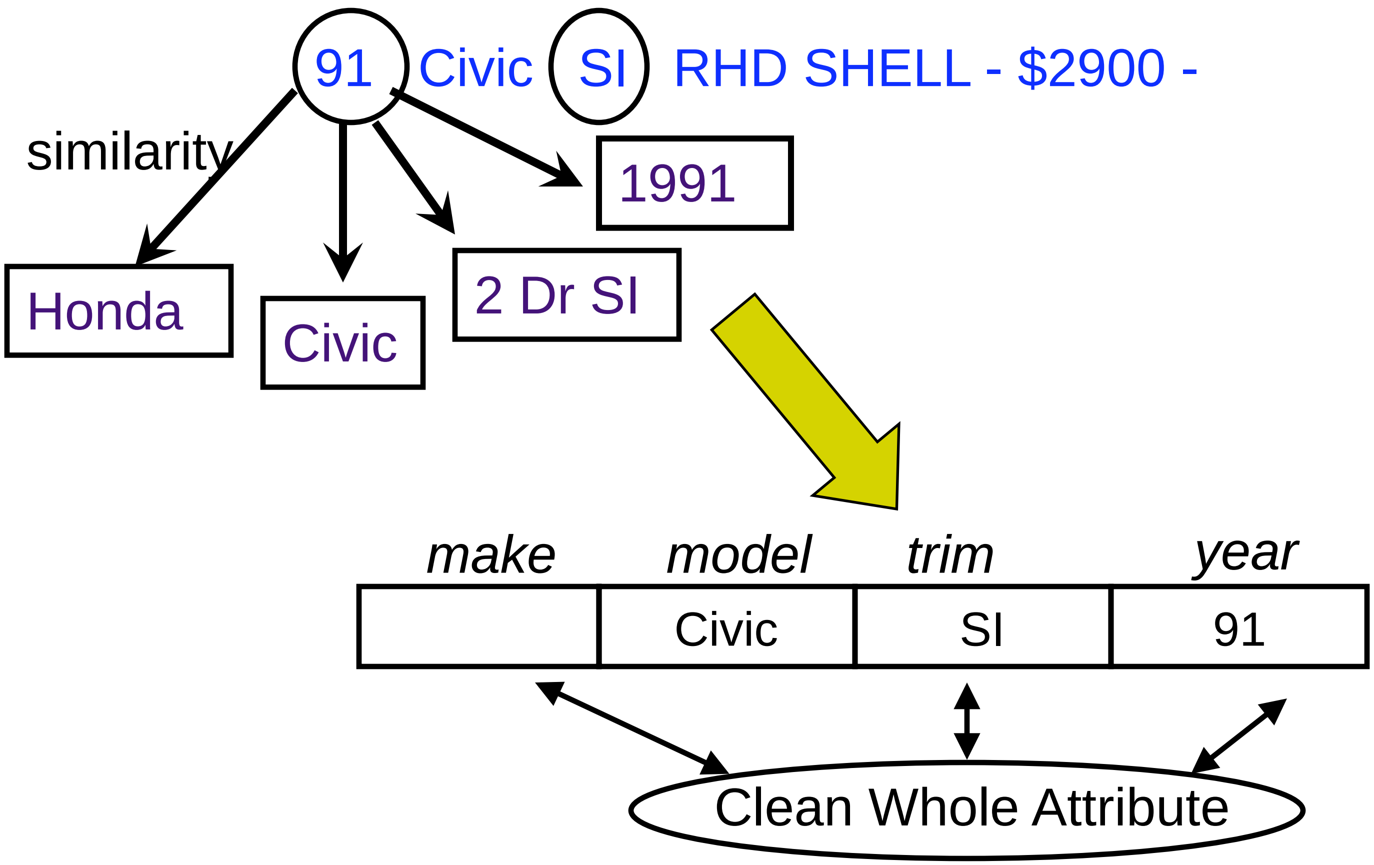
{LINCOLN, TOWN CAR, 4 Dr, 2001}

{RENAULT, LE CAR, 2 Dr, 1987}

→ { }

Prune false positives!

Unsupervised Extraction



Results: Information Extraction

- State-of-the-art comparison
  1. Conditional Random Field (structure)
     1. CRF-Orth
        - Orthographic features: cap, start-num, etc.
     2. CRF-Win
        - CRF-Orth + 2-word sliding window
          - more structure!
  2. Amilcare
     - NLP
     - "Gazetteers" (list of hotels, etc.)
- ARX = automatic, others = supervised
- Field-level extractions
  - All tokens required, no extras (strict!)

Results: Information Extraction

	Craigs Cars Posts (Craigslist)			
	ARX	*CRF-Orth*	*CRF-Win*	*Amilcare*
Make	**97.95**	83.66	78.67	94.57
Model	**88.61**	74.25	68.72	81.24
Trim	**49.70**	47.88	38.75	35.94
Year	86.47	88.04	84.52	**88.97**

~27,000 cars: Edmunds/ Super Lamb Auto

	BFT Posts (biddingfortravel.com)			
	ARX	*CRF-Orth*	*CRF-Win*	*Amilcare*
Star Rating	91.03	94.77	94.21	**96.46**
Hotel Name	**73.46**	67.47	41.33	62.91
Local Area	**71.98**	70.19	33.07	68.01

~130 hotels: BiddingForTravel.com

Automatic, state-of-the-art extraction on posts

- ARX
  - Automatic & better than supervised on 5/7 attributes
  - Cases where ARX underperforms
    - w/in 5%
    - Strong numeric component
  - Recall issue
- CRF-Win
  - Worst on 6/7
  - Can’t rely on structure!

Automatic construction of reference sets

- What if there isn't already a reference set?

HP Pavillion DV2000 laptop
Gateway ML6230, Intel Cel …

- What about coverage?

Ford	Focus
Dodge	Caravan

?

ACURA TL 3.2 VTEC - 1999

Automatic construction of reference sets

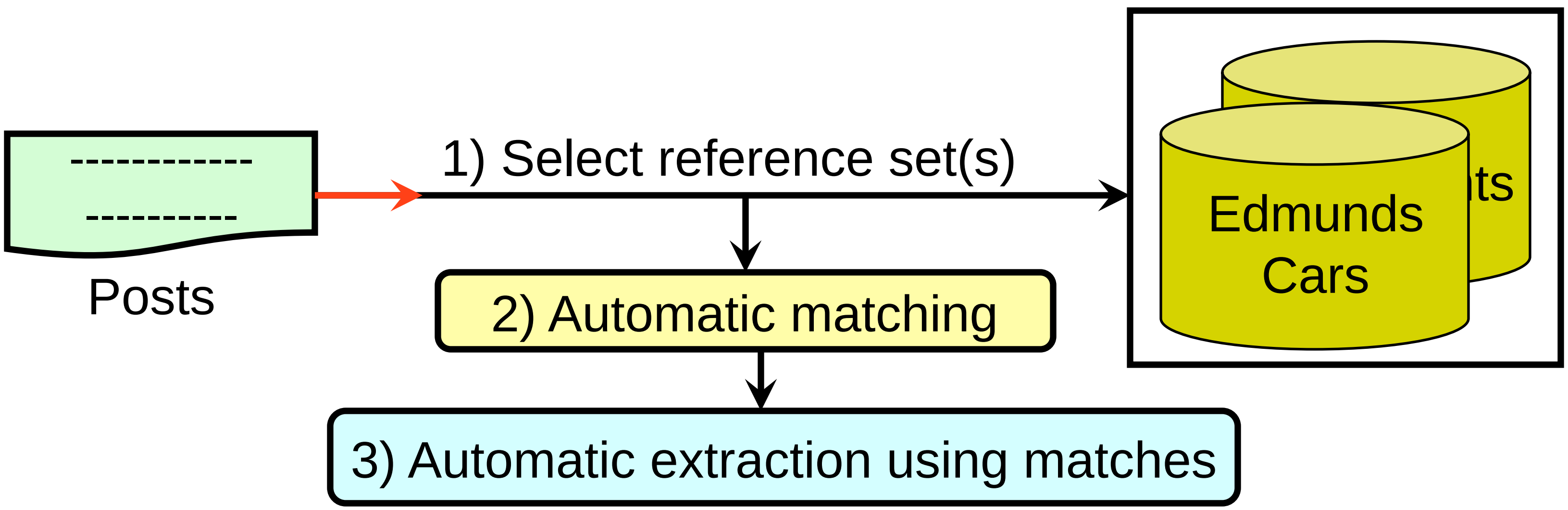
- What if there isn't already a reference set?

HP Pavillion DV2000 laptop
Gateway ML6230, Intel Cel …

- What about coverage?

Ford	Focus
Dodge	Caravan

? ↔ ACURA TL 3.2 VTEC - 1999



Automatic construction of reference sets

- What if there isn't already a reference set?

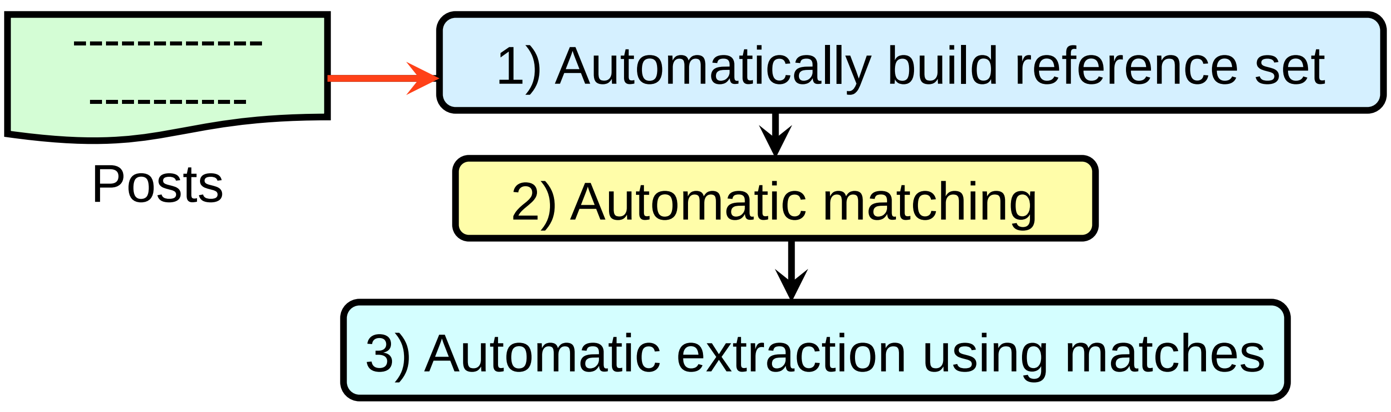
HP Pavillion DV2000 laptop	
Gateway ML6230, Intel Cel …	

- What about coverage?

Ford	Focus
Dodge	Caravan

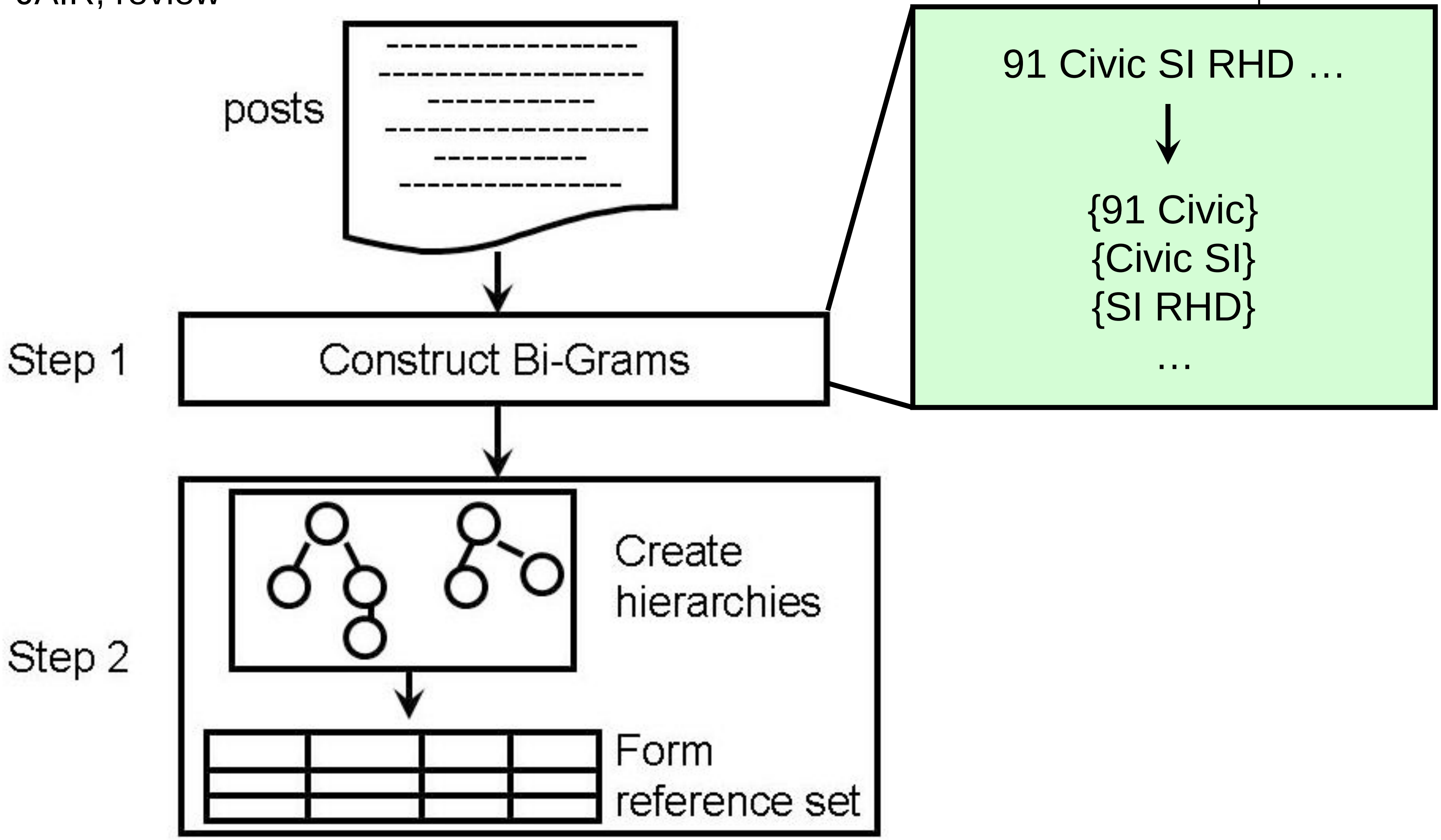
?

ACURA TL 3.2 VTEC - 1999



Build reference sets from posts

JAIR, review



Constructing entity hierarchies

- Sanderson & Croft heuristic
  - x SUBSUMES y *IF* $P(x|y) \geq 0.75$ & $P(y|x) \leq P(x|y)$
- Merge heuristic
  - MERGE(x,y) *IF* x SUBSUMES y & $P(y|x) \geq 0.75$

Constructing entity hierarchies

- Sanderson & Croft heuristic
  - x SUBSUMES y *IF* P(x|y) ≥ 0.75 & P(y|x) ≤ P(x|y)
- Merge heuristic
  - MERGE(x,y) *IF* x SUBSUMES y & P(y|x) ≥ 0.75

Honda civic is cool
Honda civic is nice
Honda accord rules
Honda accord 4 u!

P(Honda|civic) = 2/2 = 1

P(civic|Honda) = 2/4 = 0.5 → SUBSUME, not MERGE

Constructing entity hierarchies

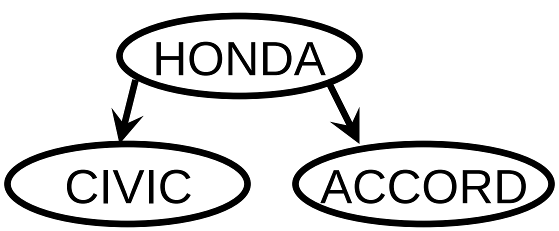
- Sanderson & Croft heuristic
  - x <u>SUBSUMES</u> y *IF* $P(x|y) \geq 0.75$ & $P(y|x) \leq P(x|y)$
- Merge heuristic
  - <u>MERGE</u>(x,y) *IF* x <u>SUBSUMES</u> y & $P(y|x) \geq 0.75$

Honda civic is cool
Honda civic is nice
Honda accord rules
Honda accord 4 u!

P(Honda|civic) = 2/2 = 1

P(civic|Honda) = 2/4 = 0.5 → <u>SUBSUME</u>, not <u>MERGE</u>

- Construct hierarchies, then flatten



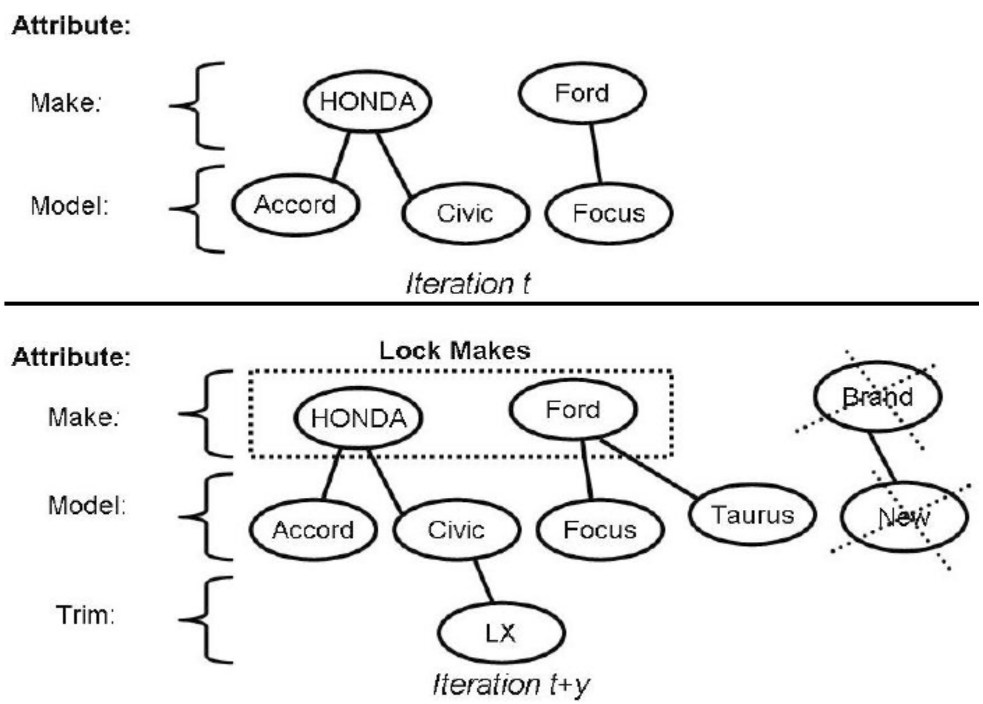
HONDA	CIVIC
HONDA	ACCORD

Construction issues

- {a, y}, {b, y}, {c, y} → y is "general token"
  - Instead use P({a U b U c } | y)
  - e.g. car trims: Pathfinder LE, Corolla LE, …

- **How many posts are enough**?
- Lock attributes (tree levels)
  - Lock out noise
  - Need only enough posts until lock all levels

Key: redundancy. At some point you've gotten all you can from the posts



Results: Information Extraction

Iterative Locking Algorithm (ILA) vs. manual reference set (ARX for extraction)

Craig's Cars: 4,400 posts			
Make	Recall	Prec.	F-Mes.
ILA (580)	78.19	84.52	81.23
Edmunds (27,006)	92.51	99.52	95.68
Model	Recall	Prec.	F-Mes.
ILA (580)	64.25	82.79	72.35
Edmunds (27,006)	79.50	91.86	85.23
Trim	Recall	Prec.	F-Mes.
ILA (580)	23.45	52.17	32.35
Edmunds (27,006)	38.01	63.69	47.61

Results: Information Extraction

Iterative Locking Algorithm (ILA) vs. manual reference set

(ARX for extraction)

Number of reference set tuples discovered

Craig's Cars: 4,400 posts			
Make	Recall	Prec.	F-Mes.
ILA (580)	78.19	84.52	81.23
Edmunds (27,006)	92.51	99.52	95.68
Model	Recall	Prec.	F-Mes.
ILA (580)	64.25	82.79	72.35
Edmunds (27,006)	79.50	91.86	85.23
Trim	Recall	Prec.	F-Mes.
ILA (580)	23.45	52.17	32.35
Edmunds (27,006)	38.01	63.69	47.61

27,000 → wasted effort!

Results: Information Extraction

Iterative Locking Algorithm (ILA) vs. manual reference set

(ARX for extraction)

Determined by locking

Craig's Cars: 4,400 posts			
Make	Recall	Prec.	F-Mes.
ILA (580)	78.19	84.52	81.23
Edmunds (27,006)	92.51	99.52	95.68
Model	Recall	Prec.	F-Mes.
ILA (580)	64.25	82.79	72.35
Edmunds (27,006)	79.50	91.86	85.23
Trim	Recall	Prec.	F-Mes.
ILA (580)	23.45	52.17	32.35
Edmunds (27,006)	38.01	63.69	47.61

Results: Information Extraction

Iterative Locking Algorithm (ILA) vs. manual reference set

(ARX for extraction)

Craig's Cars: 4,400 posts			
Make	Recall	Prec.	F-Mes.
ILA (580)	78.19	84.52	81.23
Edmunds (27,006)	92.51	99.52	95.68
Model	Recall	Prec.	F-Mes.
ILA (580)	64.25	82.79	72.35
Edmunds (27,006)	79.50	91.86	85.23
Trim	Recall	Prec.	F-Mes.
ILA (580)	23.45	52.17	32.35
Edmunds (27,006)	38.01	63.69	47.61

Competitive: fully automatic…

Results: Information Extraction

Laptops (Craigslist): 2,400 posts			
Manufacturer	Recall	Prec.	F-Mes.
ILA (295)	60.42	74.35	66.67
Overstock (279)	84.41	95.59	89.65
Model	Recall	Prec.	F-Mes.
ILA (295)	61.91	76.18	68.31
Overstock (279)	43.19	80.88	56.31
Model Num.	Recall	Prec.	F-Mes.
ILA (295)	27.91	81.08	41.52
Overstock (279)	6.05	78.79	11.23

Skis (eBay): 4,600 posts			
Brand	Recall	Prec.	F-Mes.
ILA (1,392)	60.84	55.26	57.91
Skis.com (213)	83.62	87.05	85.30
Model	Recall	Prec.	F-Mes.
ILA (1,392)	51.33	48.93	50.10
Skis.com (213)	28.12	67.95	39.77
Model Spec.	Recall	Prec.	F-Mes.
ILA (1,392)	39.14	56.35	46.29
Skis.com (213)	18.28	59.44	27.96

Results: Information Extraction

Laptops (Craigslist): 2,400 posts			
Manufacturer	Recall	Prec.	F-Mes.
ILA (295)	60.42	74.35	66.67
Overstock (279)	84.41	95.59	89.65
Model	Recall	Prec.	F-Mes.
ILA (295)	61.91	76.18	68.31
Overstock (279)	43.19	80.88	56.31
Model Num.	Recall	Prec.	F-Mes.
ILA (295)	27.91	81.08	41.52
Overstock (279)	6.05	78.79	11.23

Overstock: new laptops do not cover used ones for sale

Skis (eBay): 4,600 posts			
Brand	Recall	Prec.	F-Mes.
ILA (1,392)	60.84	55.26	57.91
Skis.com (213)	83.62	87.05	85.30
Model	Recall	Prec.	F-Mes.
ILA (1,392)	51.33	48.93	50.10
Skis.com (213)	28.12	67.95	39.77
Model Spec.	Recall	Prec.	F-Mes.
ILA (1,392)	39.14	56.35	46.29
Skis.com (213)	18.28	59.44	27.96

Ski Brands: Many models found as brands. Again, specific attributes

Results: Information Extraction

Laptops (Craigslist): 2,400 posts			
Manufacturer	Recall	Prec.	F-Mes.
ILA (295)	60.42	74.35	66.67
Overstock (279)	84.41	95.59	89.65
Model	Recall	Prec.	F-Mes.
ILA (295)	61.91	76.18	68.31
Overstock (279)	43.19	80.88	56.31
Model Num.	Recall	Prec.	F-Mes.
ILA (295)	27.91	81.08	41.52
Overstock (279)	6.05	78.79	11.23

Skis (eBay): 4,600 posts			
Brand	Recall	Prec.	F-Mes.
ILA (1,392)	60.84	55.26	57.91
Skis.com (213)	83.62	87.05	85.30
Model	Recall	Prec.	F-Mes.
ILA (1,392)	51.33	48.93	50.10
Skis.com (213)	28.12	67.95	39.77
Model Spec.	Recall	Prec.	F-Mes.
ILA (1,392)	39.14	56.35	46.29
Skis.com (213)	18.28	59.44	27.96

Overstock: new laptops do not cover used ones for sale

Ski Brands: Many models found as brands. Again, specific attributes

Fully automatic method that is competitive with supervised methods

ILA vs. CRF-Win	
Outperforms	Within 10%
4/9	7/9

ILA vs. CRF-Ortho	
Outperforms	Within 10%
1/9	4/9

ILA's Applicability

- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5* Courtyard Rancho Cordova Marriott …
    - "Boundary" issue
- 5 bigram-types:
  - … brand new Land Rover Discovery for…

ILA’s Applicability

- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5* Courtyard Rancho Cordova Marriott …
    - “Boundary” issue
- 5 bigram-types:
  - … brand new Land Rover Discovery for…

“DIFF ATTR”,

ILA's Applicability

- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5* Courtyard Rancho Cordova Marriott …
    - "Boundary" issue
- 5 bigram-types:
  - … brand new Land Rover Discovery for…

"DIFF ATTR","SAME ATTR",

ILA's Applicability

- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5* Courtyard Rancho Cordova Marriott …
    - "Boundary" issue
- 5 bigram-types:
  - … brand new Land Rover Discovery for…

"DIFF ATTR", "SAME ATTR", "ATTR JUNK",

ILA's Applicability

- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5* Courtyard Rancho Cordova Marriott …
    - "Boundary" issue
- 5 bigram-types:
  - … brand new Land Rover Discovery for…

"DIFF ATTR", "SAME ATTR", "ATTR JUNK", "JUNK ATTR",

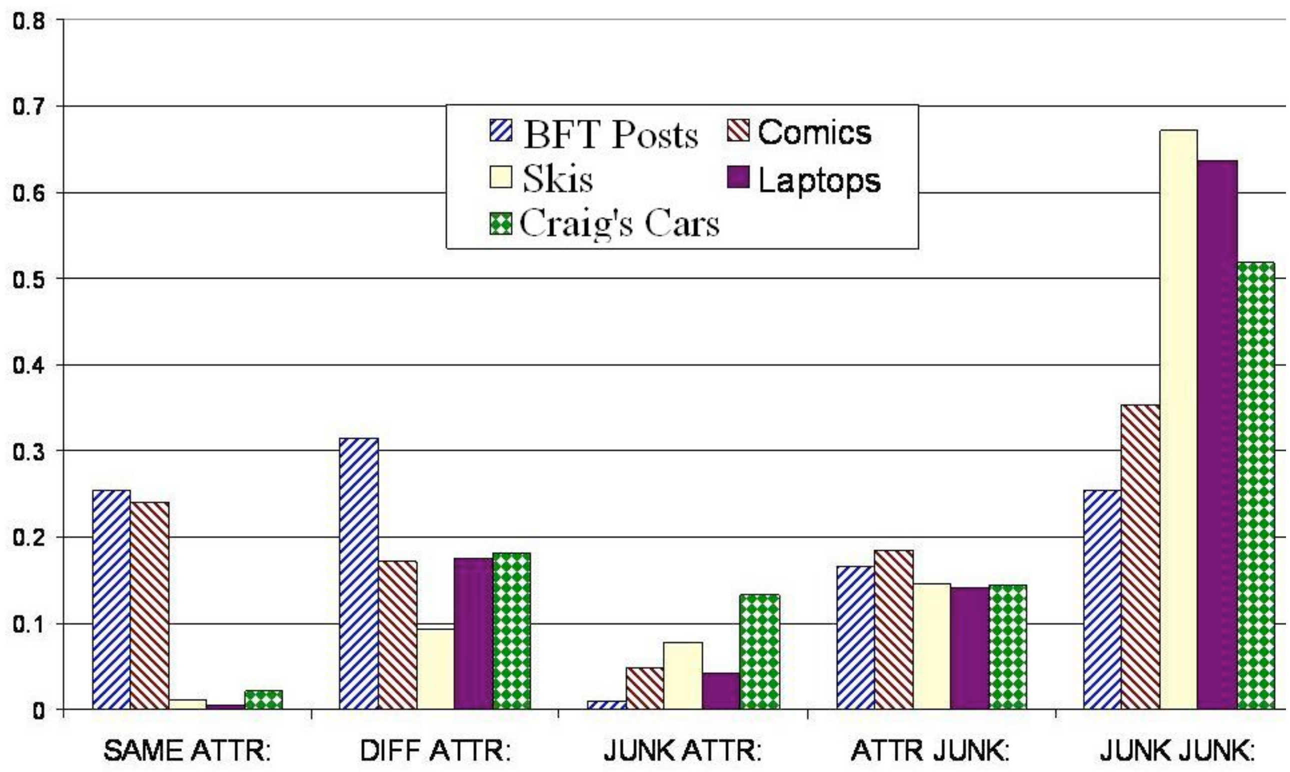
ILA's Applicability

- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5* Courtyard Rancho Cordova Marriott …
    - "Boundary" issue
- 5 bigram-types:
  - … brand new Land Rover Discovery for…

"DIFF ATTR","SAME ATTR","ATTR JUNK", "JUNK ATTR","JUNK JUNK"

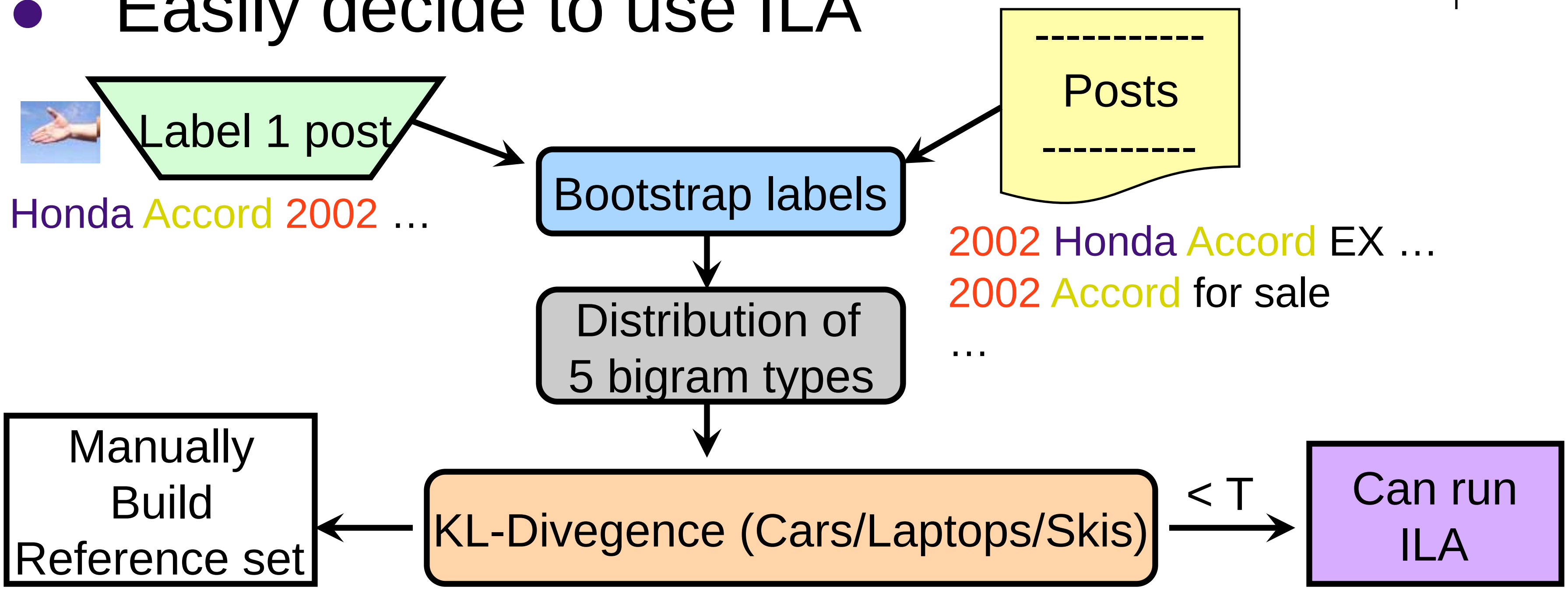
ILA's Applicability

- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5* Courtyard Rancho Cordova Marriott …
    - "Boundary" issue
- 5 bigram-types:



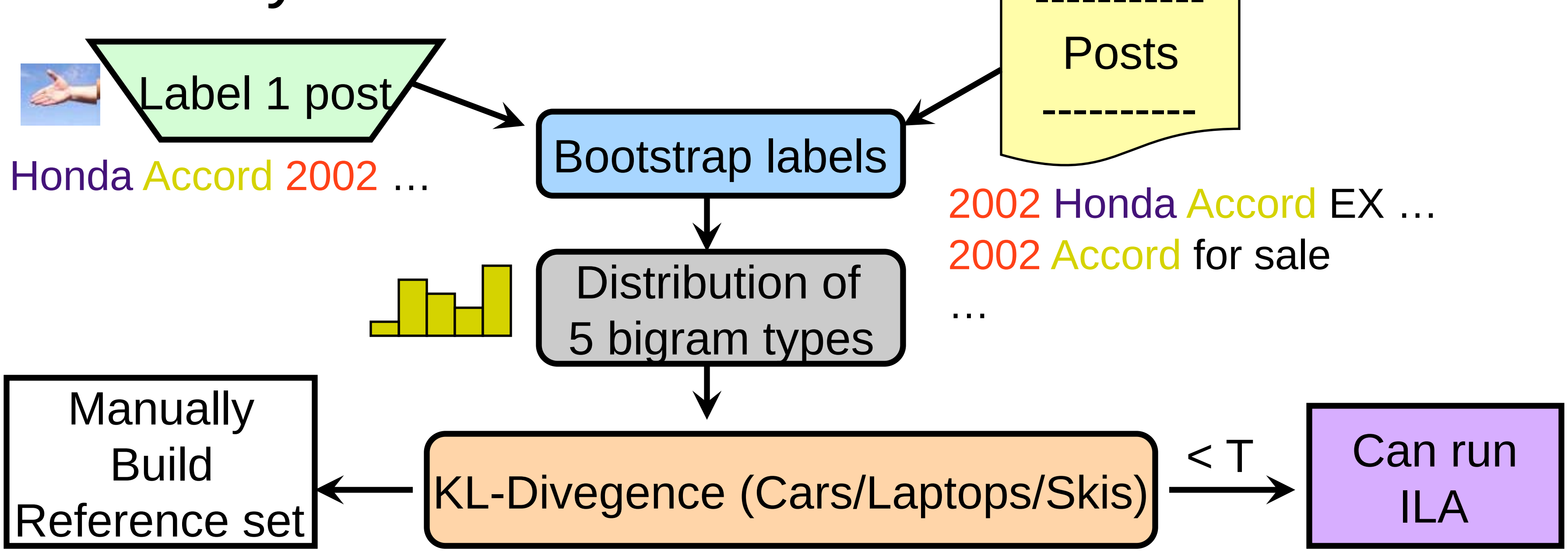
"Bootstrap-Compare"

- Easily decide to use ILA



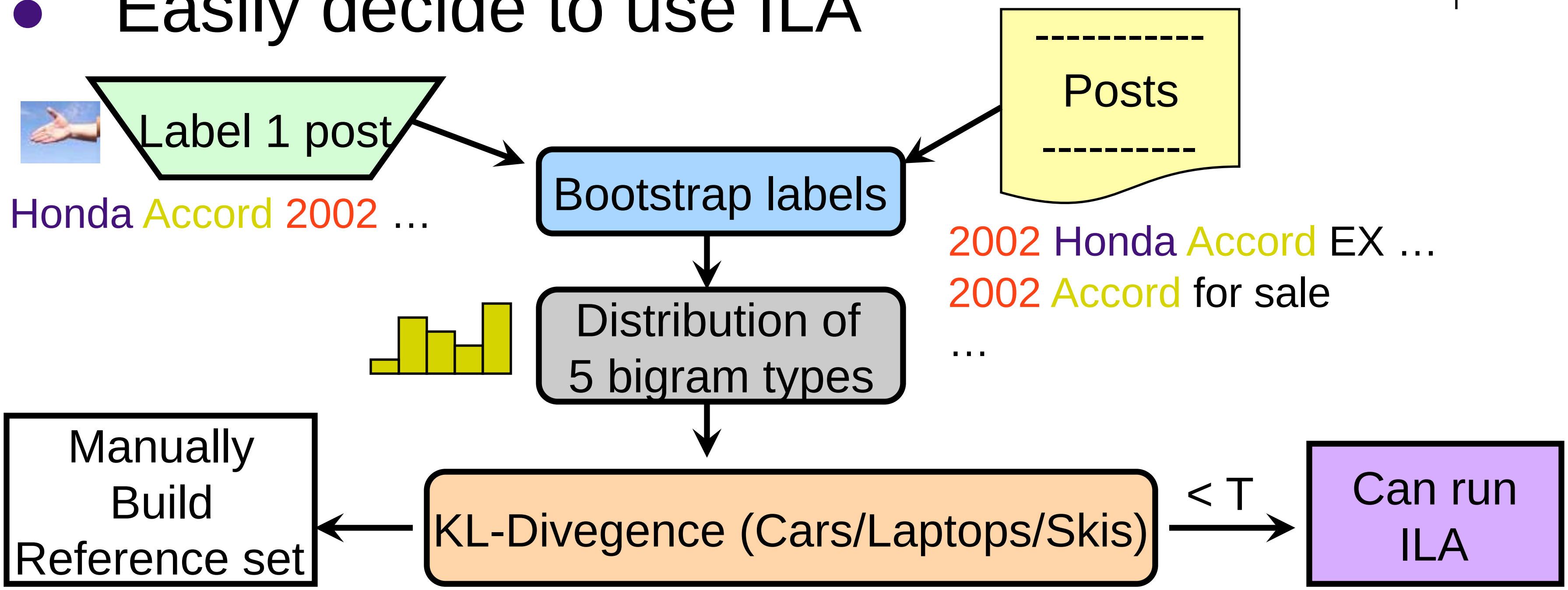
"Bootstrap-Compare"

- Easily decide to use ILA



"Bootstrap-Compare"

- Easily decide to use ILA



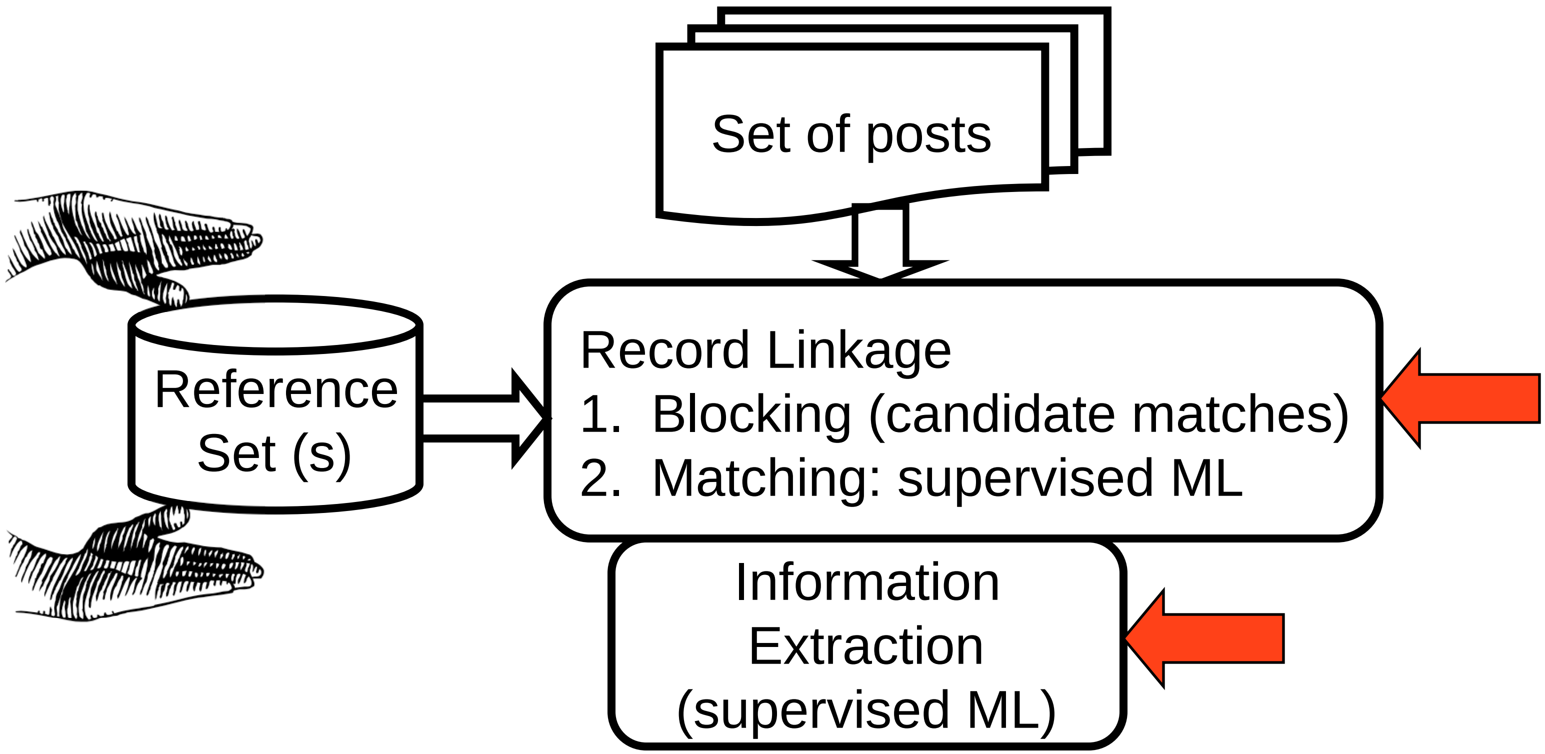
- Experiments

Source	*Can build?*	*Classification*
Digicams (eBay)	Yes, good extraction	ILA: 18/20
Cora (references)	No, poor extraction	Manual: 20/20

Supervised Machine Learning for Extraction from Posts

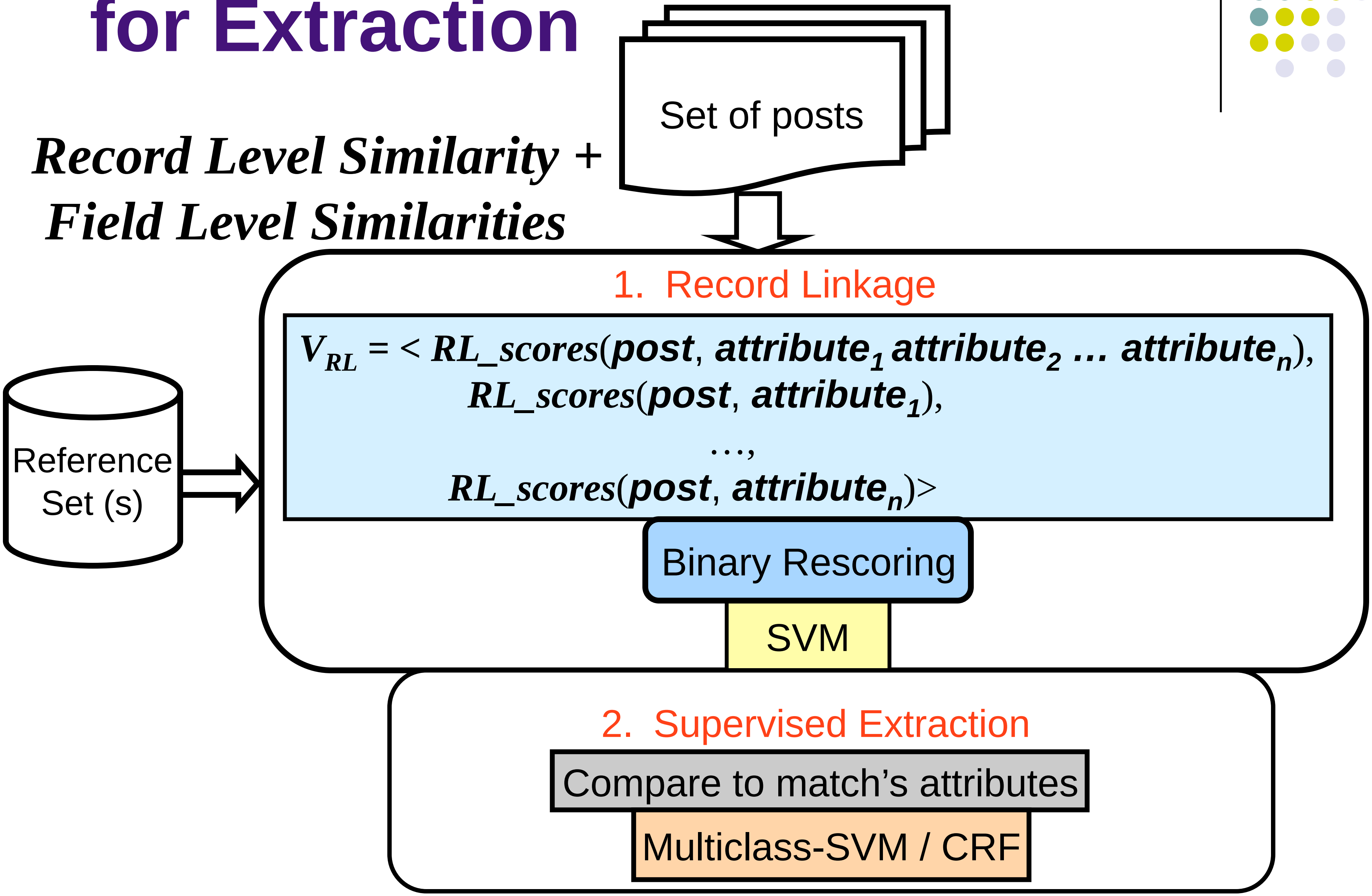
JAIR, 2008

- Require highest-accuracy extraction
  - Ambiguity: 626, Mazda or car price?



Supervised Machine Learning for Extraction

Record Level Similarity + Field Level Similarities



Results: Information Extraction

Domain	Num. of Attributes with Max F-Mes.						Total Attributes
	Phoebus	PhoebusCRF	ARX	Amilcare	CRF-Win	CRF-Orth	
BFT	**2**	**2**	0	1	0	0	5
eBay Comics	**2**	**1**	1	1	1	0	6
Craig's Cars	**5**	**0**	0	0	0	0	5
All	**9**	**3**	1	2	1	0	16

- Phoebus/PhoebusCRF
  - Best 12/16 attributes (> ARX > other methods)
  - Different extraction methods → reference set makes difference
- CRF-Win max: Comics price attribute
  - Not statistically significant…
  - CRFs outperformed
    - No structure to rely on!
- Amilcare/ARX use reference sets
  - Every max F-mes. used reference set

Related Work

- Semantic Annotation
  - Require grammar/structure (Cimiano, Handschuh & Staab, 2004; Dingli, Ciravegna, & Wilks, 2003; Handschuh, Staab & Ciravegna, 2002; Vargas-Vera, et. al., 2002)
- Record Linkage
  - Decomposed attributes (Fellegi & Sunter, 1969; Bilenko & Mooney, 2003)
  - WHIRL (Cohen, 2000): simple matching
- Data Cleaning
  - Tuple-to-Tuple (Lee, et. al., 1999; Chaudhuri, et. al., 2003)
- BSL
  - Other work focuses on methods, not choosing attributes (Baxter, Christen, & Churches, 2003; McCallum, Nigam, & Ungar, 2000; Winkler, 2005)
  - Bilenko, Kamath, & Mooney, 2006: graphical set covering

Related Work (2)

- Unstructured information extraction
  - DataMold (Borkar, Deshmukh, & Sarawagi, 2001), CRAM (Agichtein & Ganti, 2004): no junk tokens
  - Semi-CRF methods (Cohen & Sarawagi, 2004) : dictionary component, but look-up
- Ontology based IE
  - requires ontology management (Embley, et. al., 1999; Ding, Embley & Liddle, 2006; Muller, et. al., 2004)
- Ontology creation
  - Use web pages to build single hierarchies (Sanderson & Croft, 1999; Schmitz, 2006; Comiano, Hotho & Staab, 2004; Dupret & Piwowarski, 2006; Makrehchi & Kamel, 2007)
    - I build many and flatten them

Conclusion: Contributions

- Automatic, state-of-the-art extraction on posts given reference set(s)
- Automatically build reference set for cases where difficult to do so manually
- Supervised extraction on posts with highest accuracy

Conclusion: Future Work

- Applications
  - Information Retrieval
    - Source classification → page of “cars”
  - Ontology alignment
    - Match 2 ontologies to posts, then transitive closure
  - Semantic Web mark-up
- Research
  - More robust automatic creation
  - Weakly (semi?) supervised approach to IE
  - Information Fusion
    - Larger documents? NER?
  - Data mining the results
    - Create portals
    - User decision support

Questions?