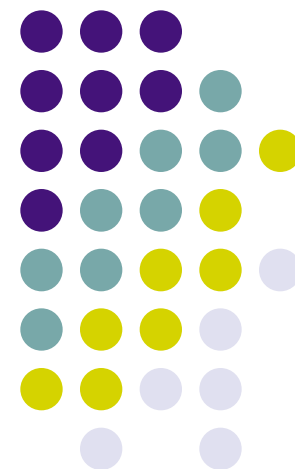
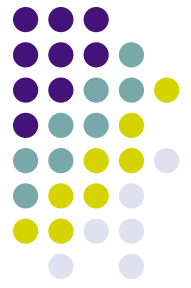


# A Reference-Set Approach to Information Extraction from Unstructured, Ungrammatical Data Sources

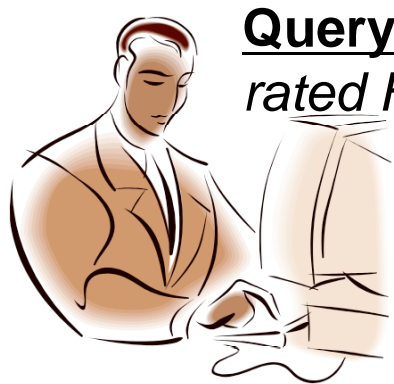
---

Matthew Michelson  
Ph.D. Defense  
Nov. 3<sup>rd</sup>, 2008

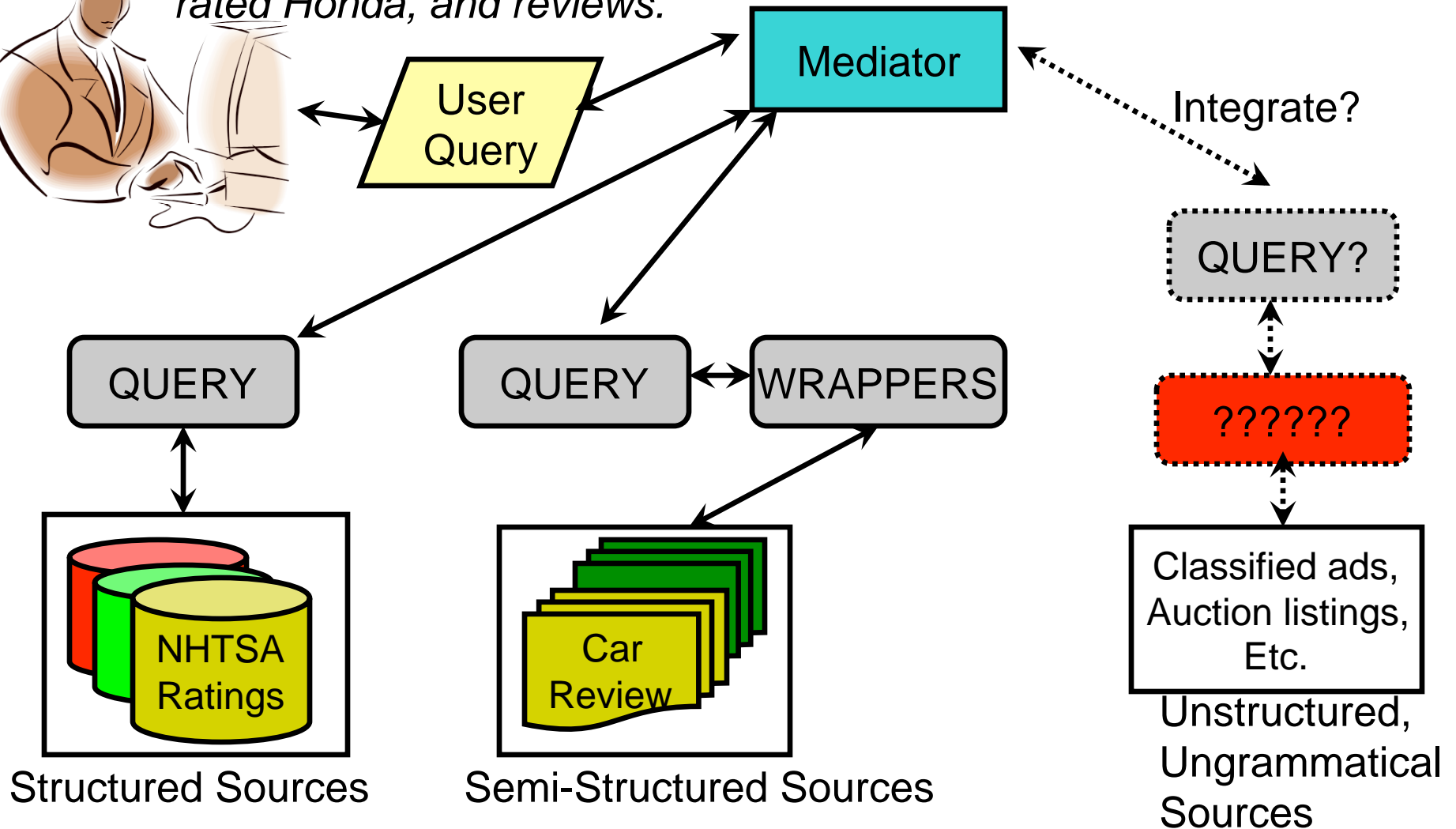


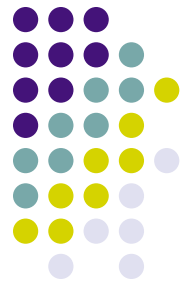


# Motivation: Data Integration

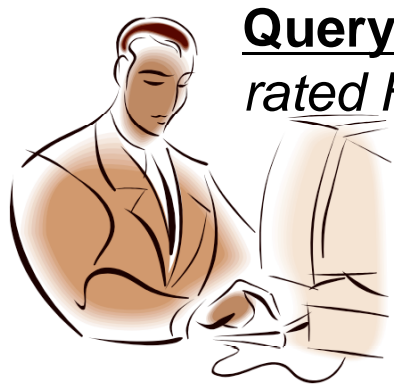


**Query:** *Average price for a 3-star crash-rated Honda, and reviews.*

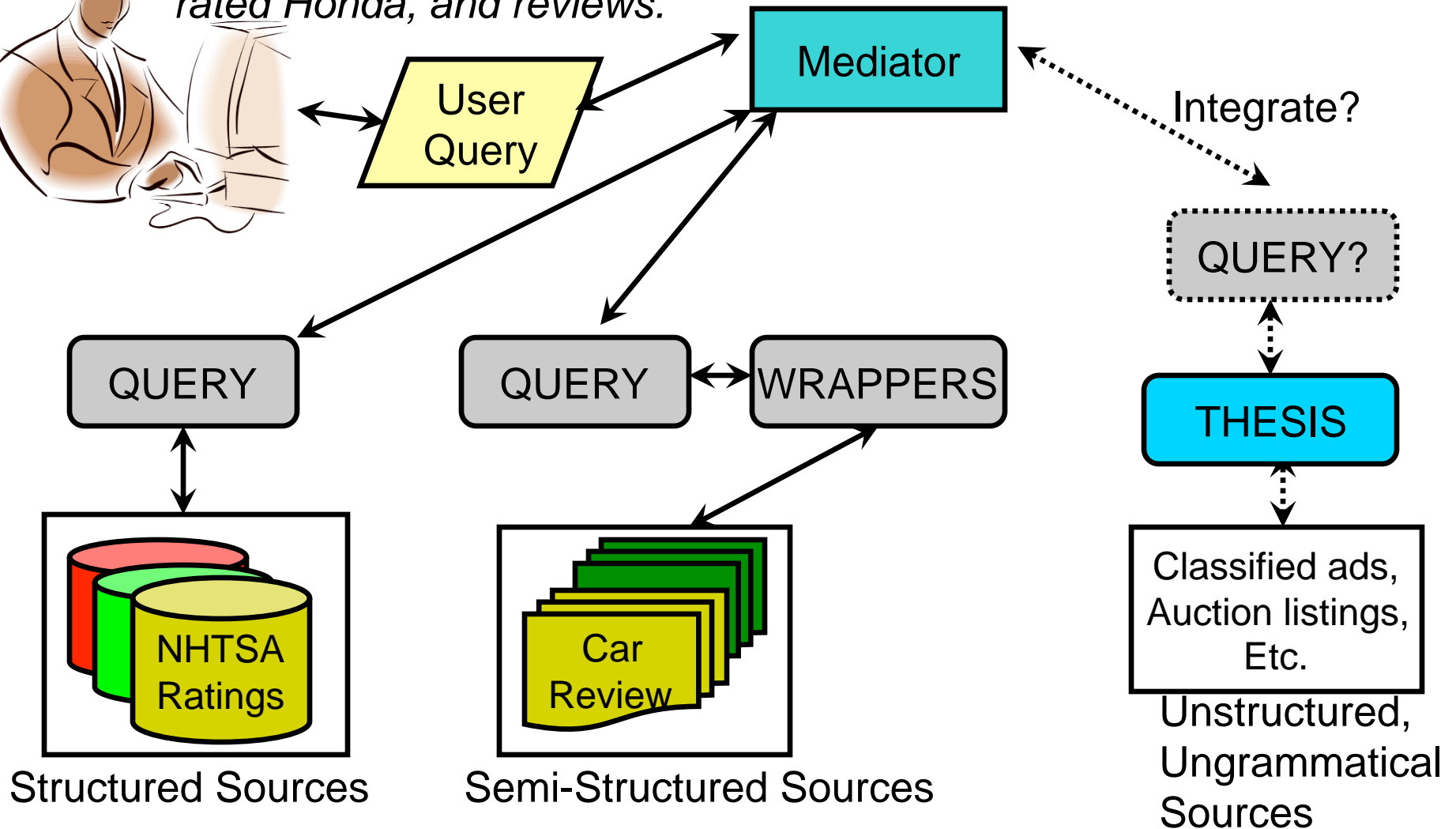




# Motivation: Data Integration



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





# Unstructured, Ungrammatical Data: “Posts”

Downhill Skis

Brand  
K2 (299)  
[Back to all options](#)

Length (cm)  
[Less than 140cm \(14\)](#)  
[140-150cm \(22\)](#)  
[151-160cm \(73\)](#)  
[161-170cm \(79\)](#)  
[171-180cm \(62\)](#)  
[181-190cm \(37\)](#)  
[More than 190cm \(6\)](#)

Condition  
[New](#)  
[Used](#)











Search Options

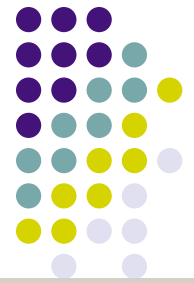
Location  
 Worldwide  
 Items within  miles of

Show only  
 Items listed with PayPal  
 Buying Options  
  
 More Buying Options

2007 items found in: Skis (save this search)

List View | [Picture Gallery](#) Sort by:  [Customize Display](#)

Item Title	Bids	Price*	Shipping to USA	Time Left
<b>Featured Items</b>				
 NEW VOLKL AC SKIS 170 MARKER 3 MOTION 11 BDGS UNLIMITED		<i>Buy It Now</i> \$385.94	\$30.00	2d 20h 14m
Optimize your selling success! Find out how to <a href="#">promote your items</a>				
 *NEW Line Anthem Alpine Ski 161cm Twin Tip + Bonus!!! RP=\$360, New, Never Used, With 1 Y Warranty! 1 Day Ship!		<i>Buy It Now or Best Offer</i> \$244.88	\$28.99	6m
 <b>2007 SALOMON SIAM 400 LADIES SKIS W/BINDING 160 CM NEW</b>		<i>Buy It Now</i> \$169.00	\$35.00	15m
 Rossignol Bandit B1 170 cm Used Snow Skis w/ Bindings		-	\$179.99 \$34.00	1h 09m
 New Atomic Nomad Blackeye Skis 171 New k2 volkl		-	\$379.99 \$20.00	1h 12m



# Unstructured, Ungrammatical Data: “Posts”

about:blank www.mailla.deutschin... Hotmail Welcome to Gmail Google News Overview (Java 2 Platf... citeseer ISI

search for:  in: cars & trucks Search  only search titles

price:  min  max  by dealer  by owner  all  has image

[ Fri, 14 Mar 11:45:39 ]

[ [ALERT - offers to ship cars/trucks are fraudulent](#) ] [ [partial list of prohibited items](#) ]  
[ [avoid recalled items](#) ] [ [success story?](#) ] [ [AVOIDING SCAMS & FRAUD](#) ]  
[ [PERSONAL SAFETY TIPS](#) ]

Fri Mar 14

[91 Civic SI RHD SHELL - \\$2900 - \(West Covina\)](#) [pic](#)

[2001 Automatic Mazda Millenia Clear Title - \\$3800 -](#) [pic](#)

[1984 Ford Tow Truck - \\$10000 - \(Bell\)](#)

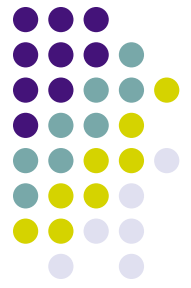
[2004 Audi A4 1.8T - \\$6800 -](#) [pic](#)

[1998 International 4700 Tow Truck - \\$12000 - \(Bell\)](#)

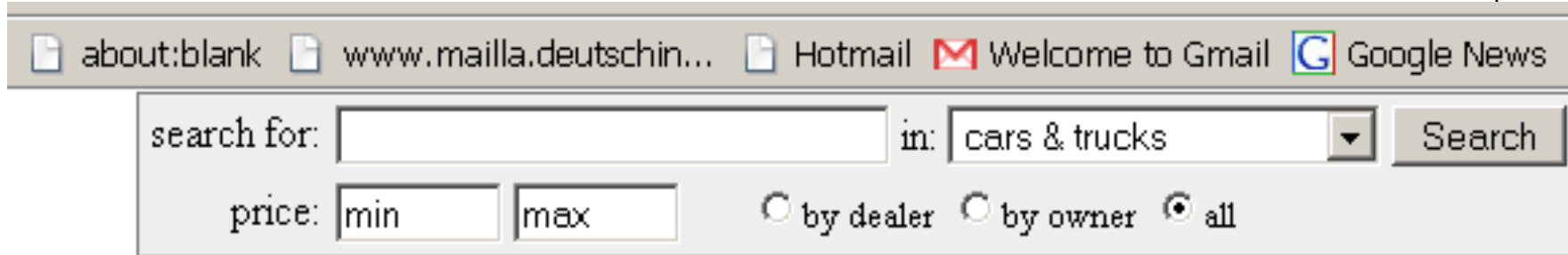
[1994 >>>> LEXUS ES 300 >> LEATHER INTERIOR <<< - \\$3000 - \(RESEDA\)](#) [pic](#)

[1987 Chevrolet Tahoe 4x4 just smogged - \\$1400 - \(Palmdale\)](#) [pic](#)

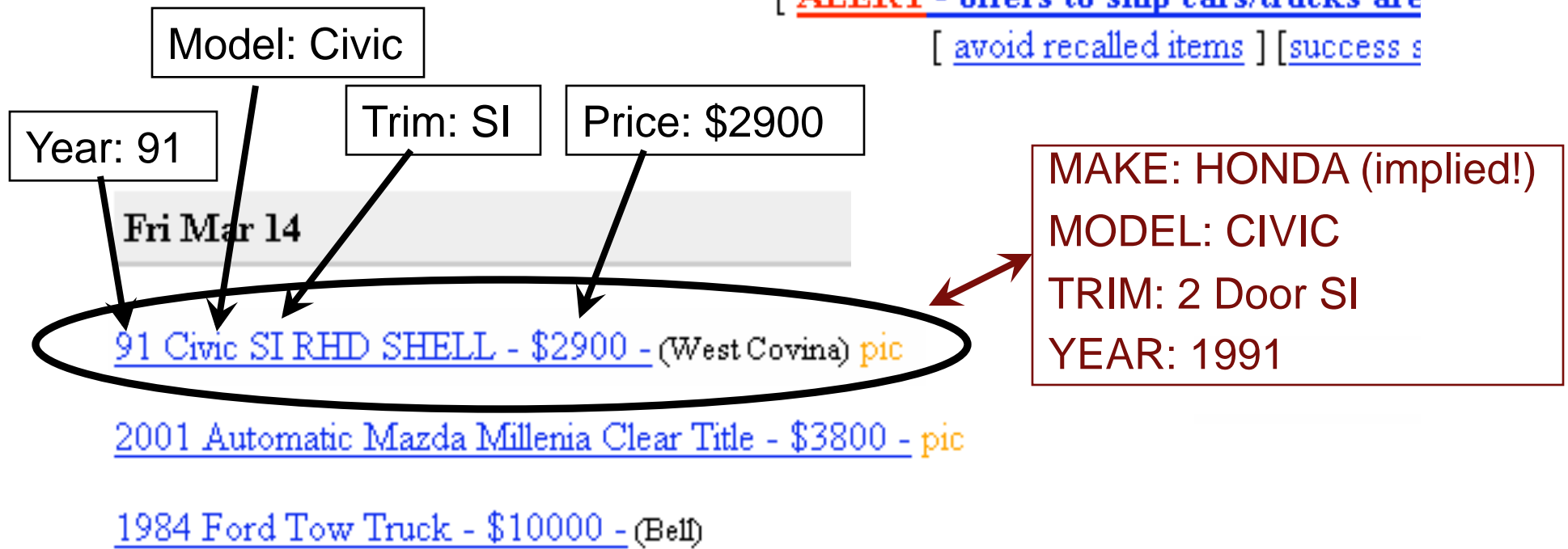
POST



# Query? ... Information Extraction/Annotation!



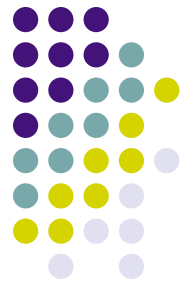
[ ALERT - offers to ship cars/trucks are  
[ avoid recalled items ] [ success s



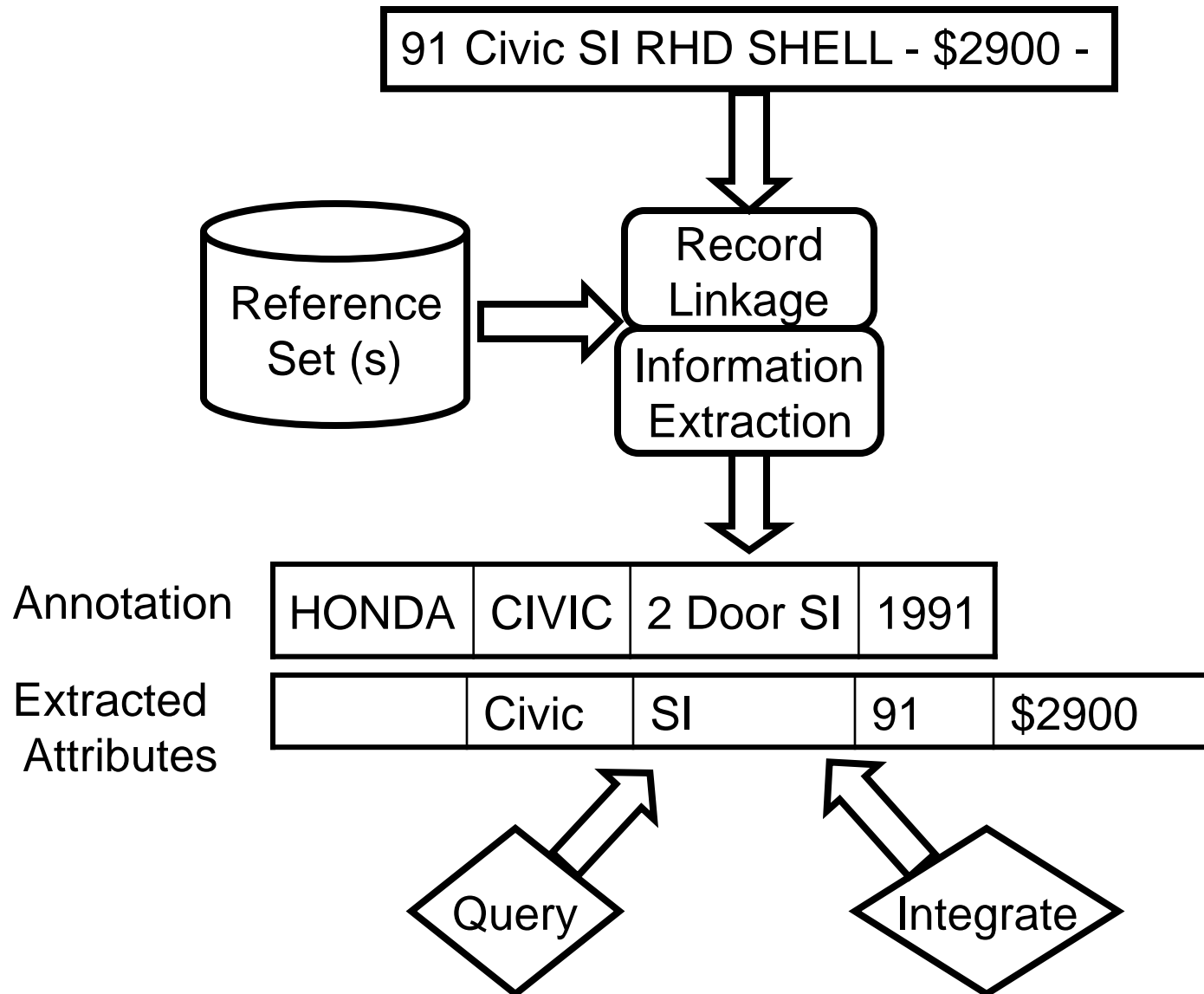


# Difficulties

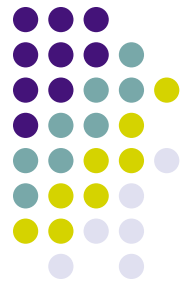
- Unstructured
  - No assumptions on structure
  - “Rule/Pattern” based techniques unsuited
- Ungrammatical
  - Does not conform to English grammar
  - Natural-Language Processing techniques unsuited



# Reference-Set Based Extraction/Annotation







# Reference Sets

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



## New Car Pricing: Acura Vehicles

6 vehicle(s) found

### ◆ Choose from Most Researched Models

<a href="#">2007 Acura MDX</a>	<a href="#">SAVE THIS CAR</a>
<a href="#">2007 Acura TL</a>	<a href="#">SAVE THIS CAR</a>
<a href="#">2007 Acura RDX</a>	<a href="#">SAVE THIS CAR</a>
<a href="#">2006 Acura MDX</a>	<a href="#">SAVE THIS CAR</a>
<a href="#">2006 Acura TL</a>	<a href="#">SAVE THIS CAR</a>
<a href="#">2006 Acura TSX</a>	<a href="#">SAVE THIS CAR</a>

### ◆ Choose from Most Sold Models

<a href="#">2006 Acura TL</a>	<a href="#">SAVE THIS CAR</a>
<a href="#">2006 Acura MDX</a>	<a href="#">SAVE THIS CAR</a>
<a href="#">2006 Acura TSX</a>	<a href="#">SAVE THIS CAR</a>
<a href="#">2007 Acura RDX</a>	<a href="#">SAVE THIS CAR</a>
<a href="#">2006 Acura RL</a>	<a href="#">SAVE THIS CAR</a>
<a href="#">2006 Acura RSX</a>	<a href="#">SAVE THIS CAR</a>

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



# Contributions

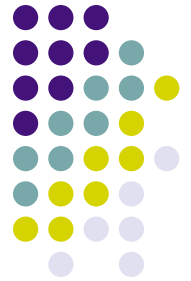
- Automatic matching and extraction algorithm that exploits a given reference set
  - Automatically select the appropriate reference sets from a repository of reference sets
- Automatic method for building reference sets from the posts themselves
  - Suggest the number of posts required to sufficiently 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 ↑

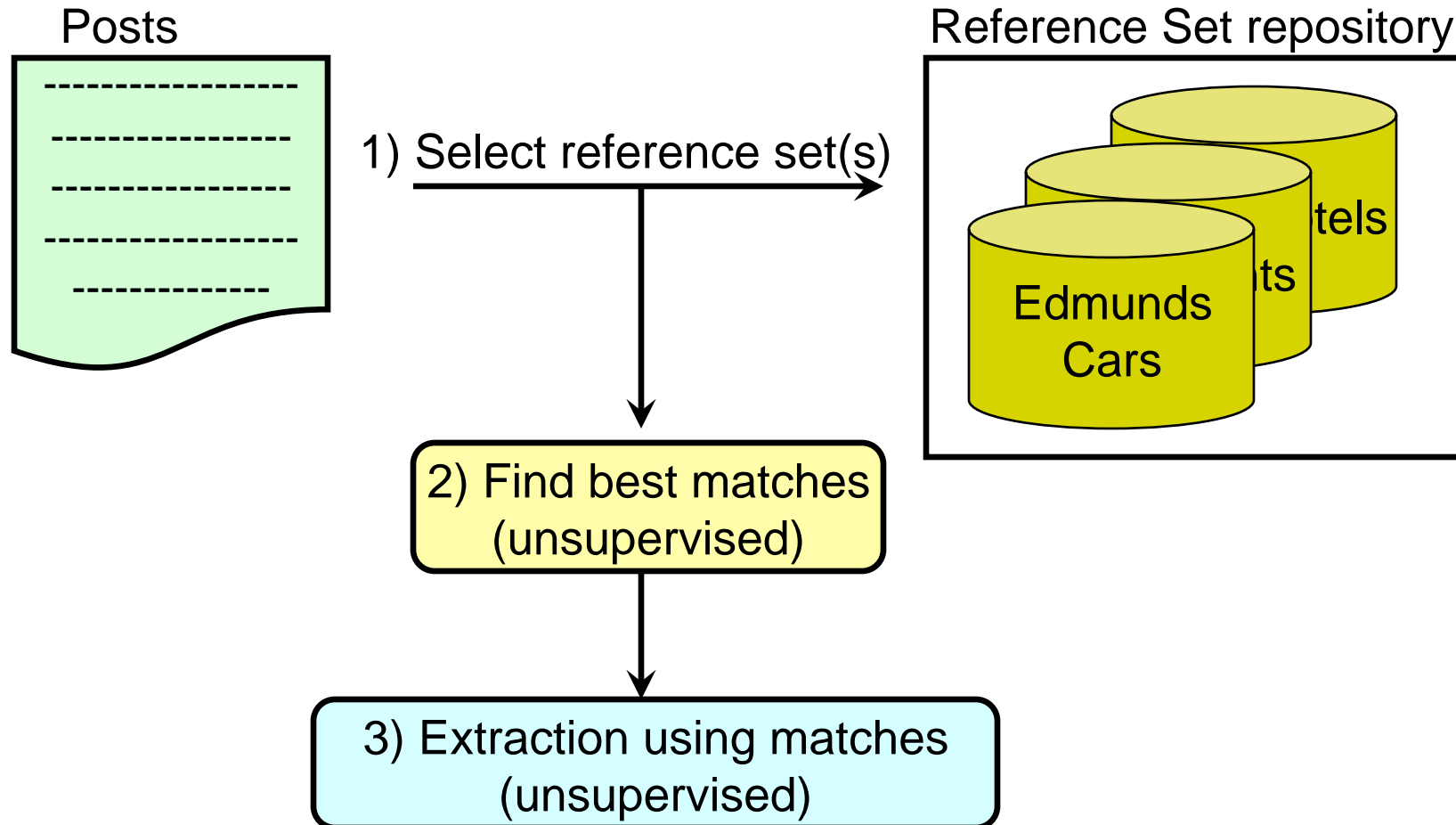
3 reference-set based extraction methods

	<i>Summary</i>	<i>Advantages</i>
Method 1 (ARX)  [IJ DAR 07]	<ol style="list-style-type: none"> <li>1. Automatically select reference set from repository</li> <li>2. Automatic extraction</li> </ol>	<ul style="list-style-type: none"> <li>• State-of-the-art extraction</li> <li>• Automatic, given reference set</li> </ul>
Method 2 (ILA)  [JAIR, review]	<ol style="list-style-type: none"> <li>1. Automatically build reference set</li> </ol>	<ul style="list-style-type: none"> <li>• Cannot build reference set (difficult attributes)</li> <li>• Fully automatic</li> <li>• Competitive state-of-the-art</li> </ul>
Method 3 (Phoebus)  [JAIR, 08]	<ol style="list-style-type: none"> <li>1. Supervised approach to extraction</li> </ol>	<ul style="list-style-type: none"> <li>• Highest-accuracy extraction</li> <li>• Deals with ambiguity</li> </ul>



# Automatic method: Three steps

IJDAR, 2007



ARX: Automatic Reference-set based eXtraction



# Selecting the Reference Set(s)

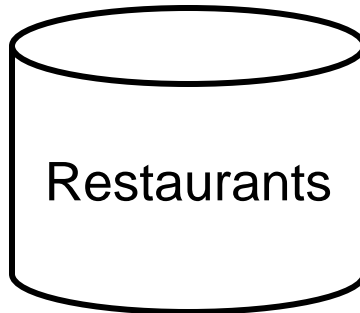
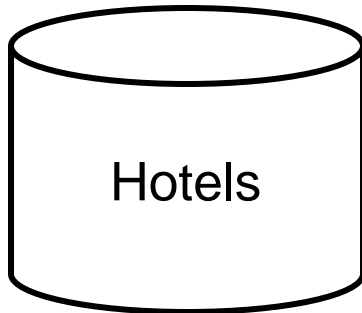
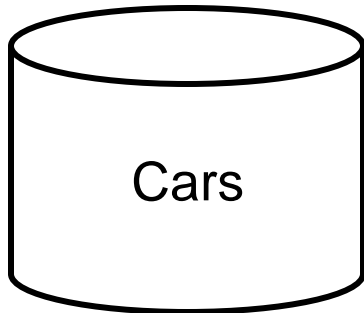
Vector space model: set of posts are 1 doc, reference sets are 1 doc

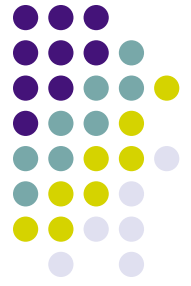
Select reference set most similar to the set of posts...

FORD Thunderbird - \$4700

2001 White Toyota Corrolla CE Excellent Condition - \$8200

SIM:0.7

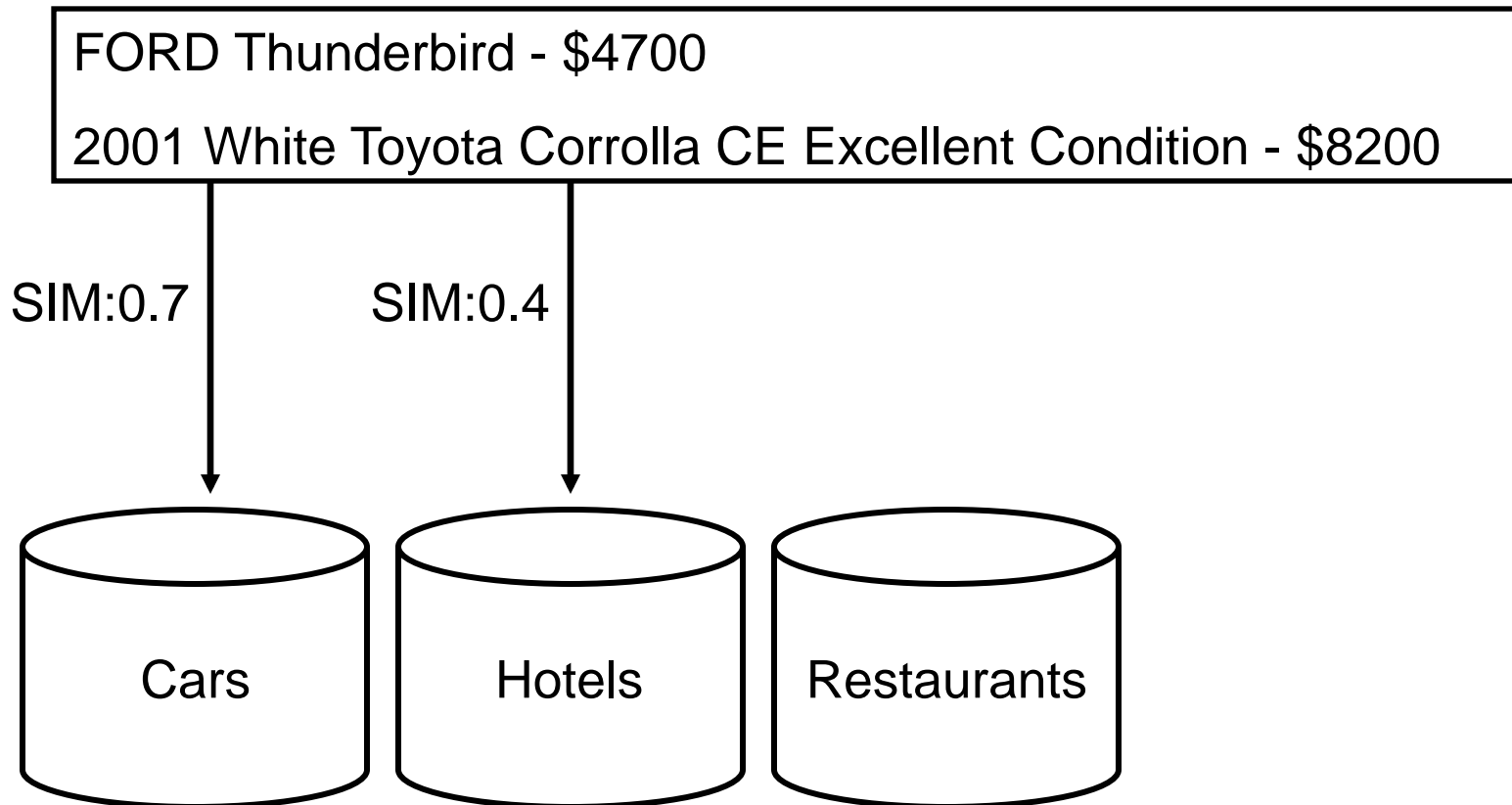




# Selecting the Reference Set(s)

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

Select reference set most similar to the set of posts...

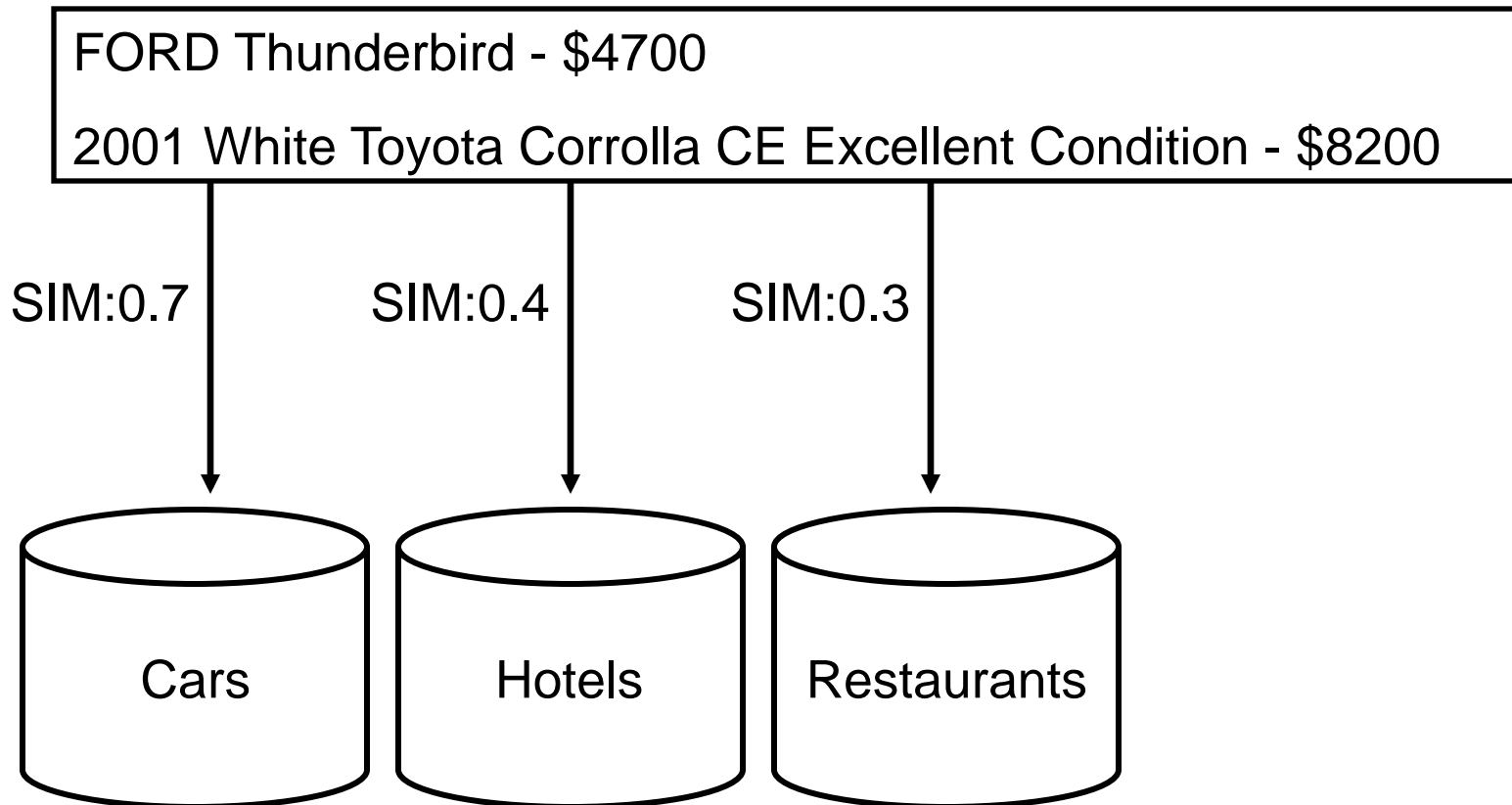


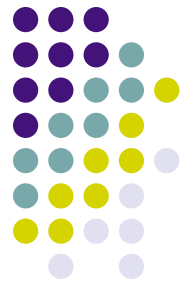


# Selecting the Reference Set(s)

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

Select reference set most similar to the set of posts...





# 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

SIM:0.7

SIM:0.4

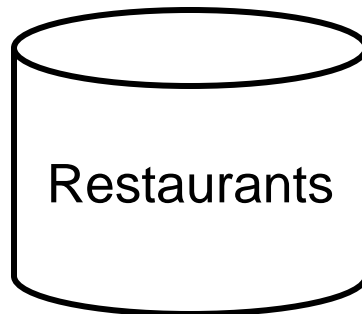
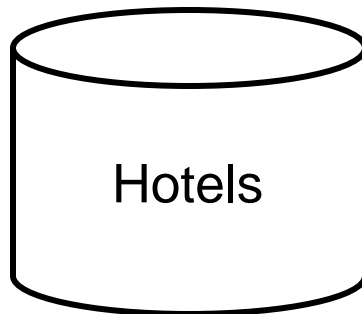
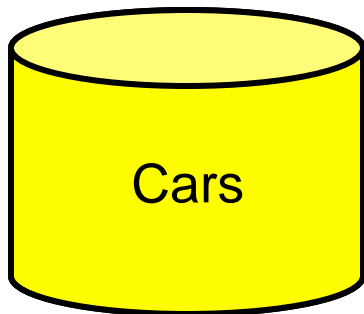
SIM:0.3

Cars 0.7    PD(C,H) = 0.75 > T

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

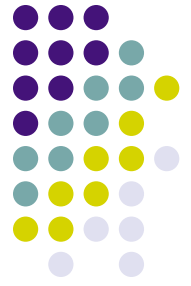
02 M3 Convertible .. Absolute beauty!!!

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!

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

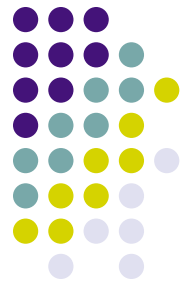
~~{NISSAN, ALTIMA, 4 Dr 2.5 S Sedan, 2007}~~

{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!

~~{NISSAN, ALTIMA, 4 Dr 3.5 SE Sedan, 2007}~~  
~~{NISSAN, ALTIMA, 4 Dr 2.5 S Sedan, 2007}~~

Vector-based matching

→ {NISSAN, ALTIMA, 2007}

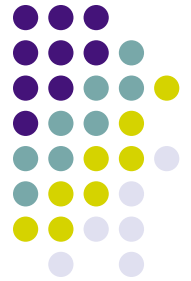
{BMW, M3, 2 Dr STD Convertible, 2002}

~~{LINCOLN, TOWN CAR, 4 Dr, 2001}~~

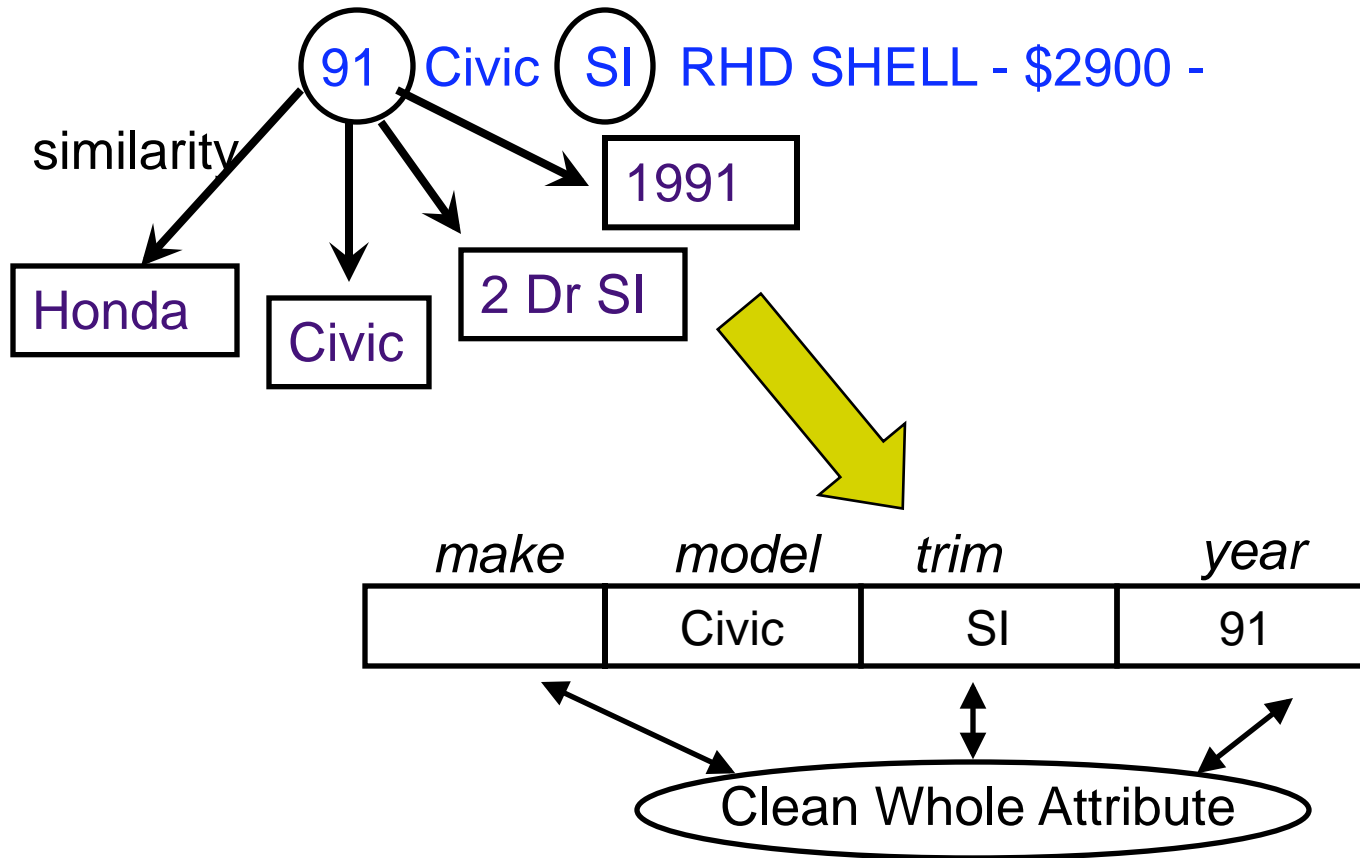
~~{RENAULT, LE CAR, 2 Dr, 1987}~~

→ {}

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)			
	<i>ARX</i>	<i>CRF-Orth</i>	<i>CRF-Win</i>	<i>Amilcare</i>
Make	<b>97.95</b>	83.66	78.67	94.57
Model	<b>88.61</b>	74.25	68.72	81.24
Trim	<b>49.70</b>	47.88	38.75	35.94
Year	86.47	88.04	84.52	<b>88.97</b>

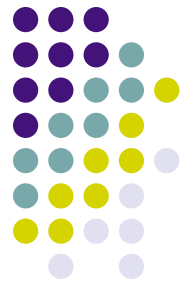
~27,000 cars: Edmunds/ Super Lamb Auto

	BFT Posts (biddingfortravel.com)			
	<i>ARX</i>	<i>CRF-Orth</i>	<i>CRF-Win</i>	<i>Amilcare</i>
Star Rating	91.03	94.77	94.21	<b>96.46</b>
Hotel Name	<b>73.46</b>	67.47	41.33	62.91
Local Area	<b>71.98</b>	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

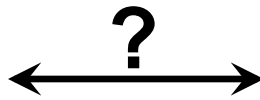
- What if there **isn't already a reference set?**

HP Pavillion DV2000 laptop
----------------------------

Gateway ML6230, Intel Cel ...
-------------------------------

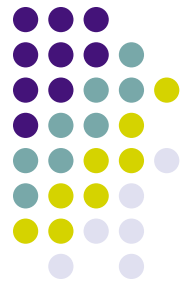
- What about **coverage?**

Ford	Focus
Dodge	Caravan



ACURA TL 3.2 VTEC - 1999
--------------------------



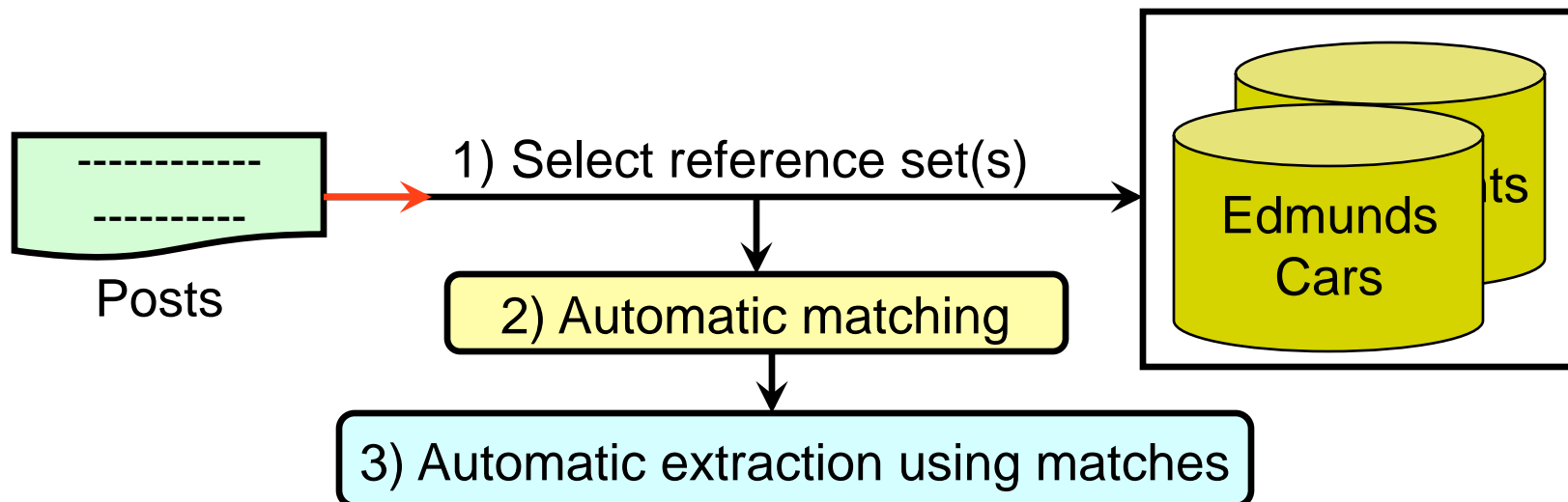
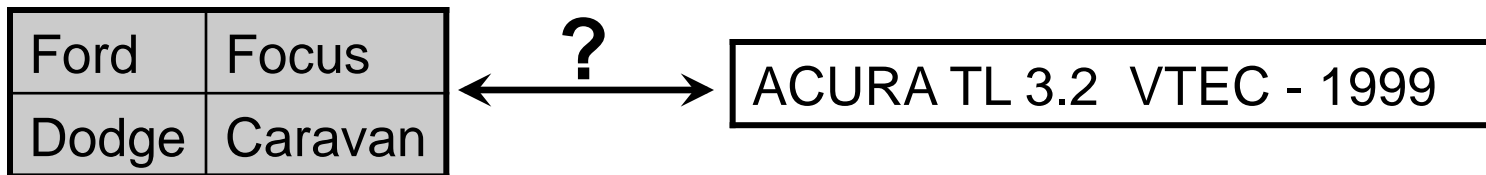


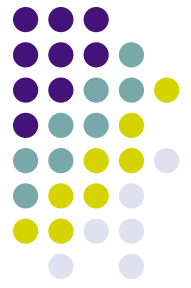
# Automatic construction of reference sets

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

HP Pavillion DV2000 laptop
Gateway ML6230, Intel Cel ...

- What about **coverage?**



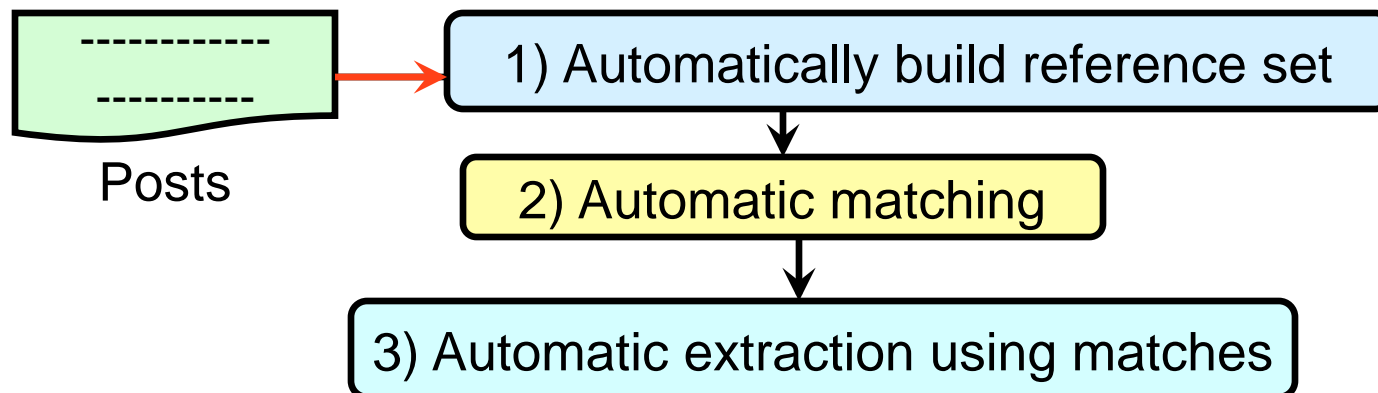
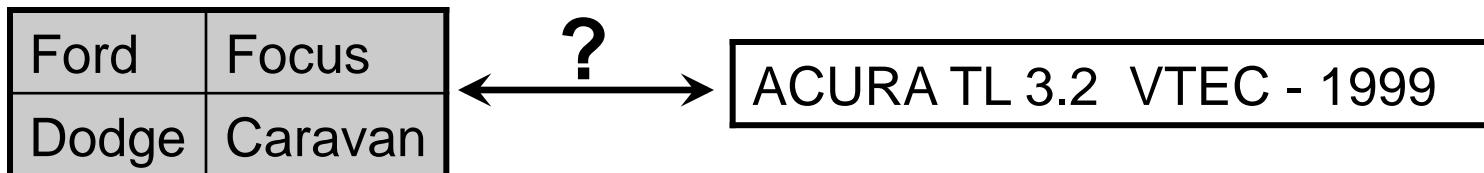


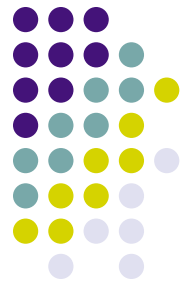
# Automatic construction of reference sets

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

HP Pavillion DV2000 laptop
Gateway ML6230, Intel Cel ...

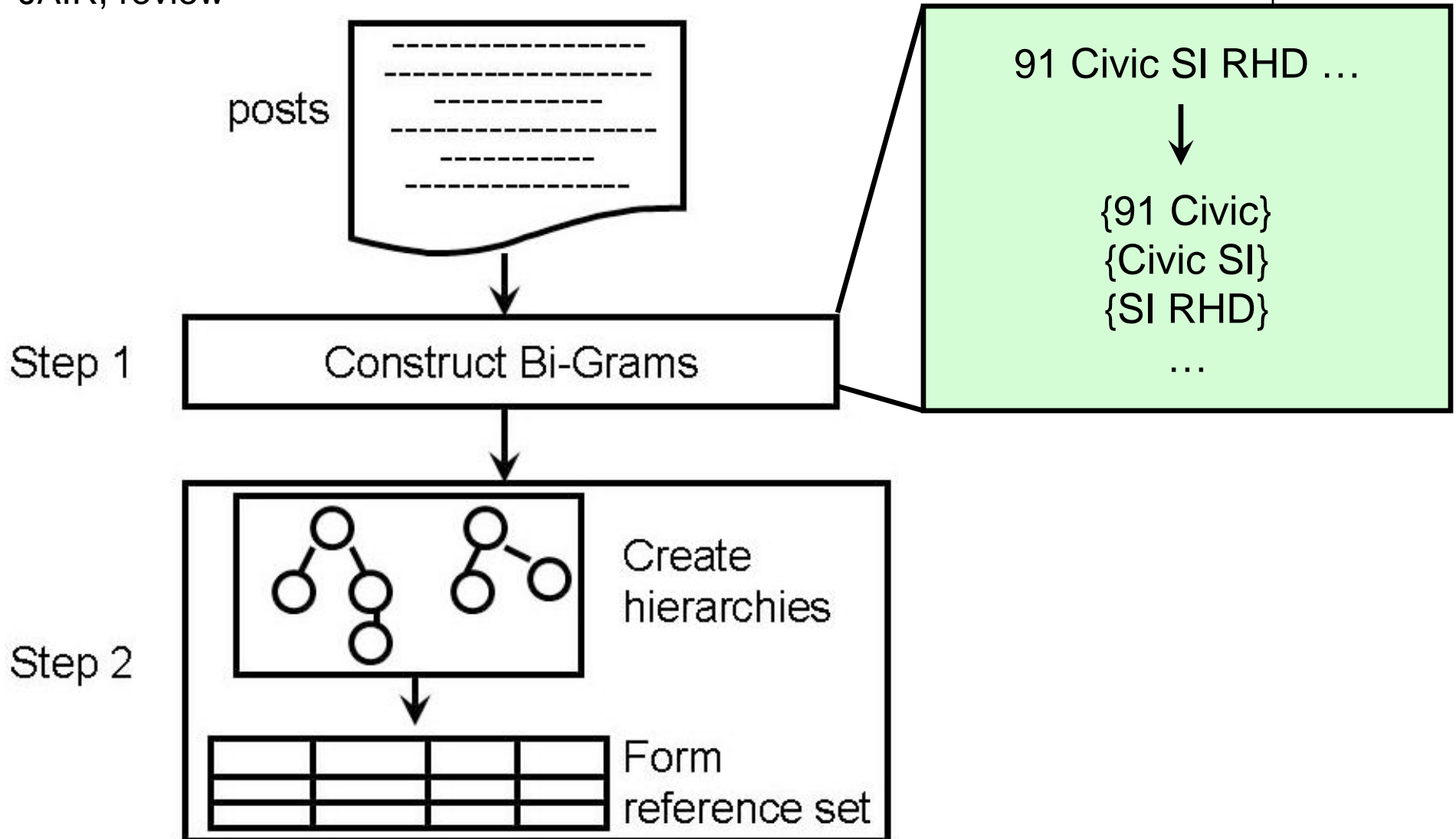
- What about **coverage?**





# Build reference sets from posts

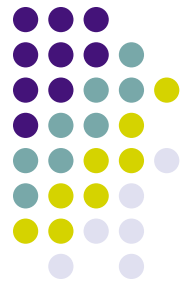
JAIR, review





# 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  $y$  &  $P(y|x) \geq 0.75$



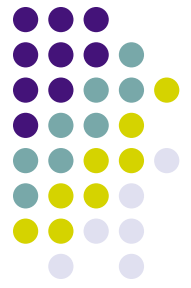
# Constructing entity hierarchies

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

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

$$P(\text{Honda}|\text{civic}) = 2/2 = 1$$

$$P(\text{civic}|\text{Honda}) = 2/4 = 0.5 \rightarrow \text{SUBSUME}, \text{ not MERGE}$$



# 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  $y$  &  $P(y|x) \geq 0.75$

Honda civic is cool

$$P(\text{Honda}|\text{civic}) = 2/2 = 1$$

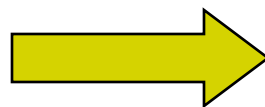
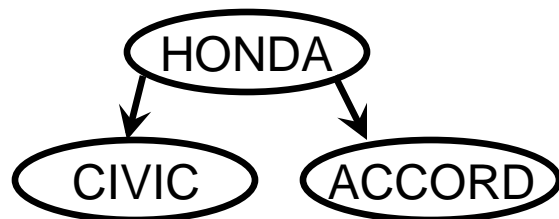
Honda civic is nice

Honda accord rules

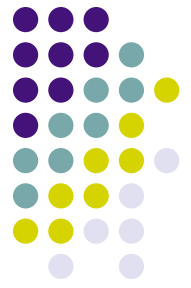
$$P(\text{civic}|\text{Honda}) = 2/4 = 0.5 \rightarrow \text{SUBSUME}, \text{ not MERGE}$$

Honda accord 4 u!

- Construct hierarchies, then flatten



HONDA	CIVIC
HONDA	ACCORD



# Construction issues

- $\{a, y\}, \{b, y\}, \{c, y\} \rightarrow y$  is “general token”
  - Instead use  $P(\{a \cup b \cup c\} | y)$
  - e.g. car trims: Pathfinder LE, Corolla LE, ...

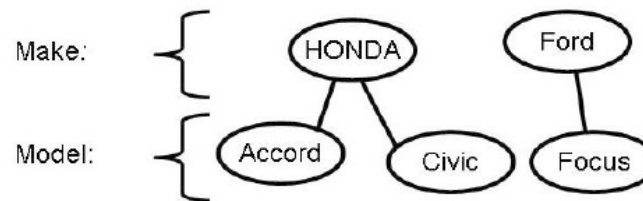
- **How many posts are enough?**

- Lock attributes (tree levels)

- Lock out noise
- Need only enough posts until lock all levels

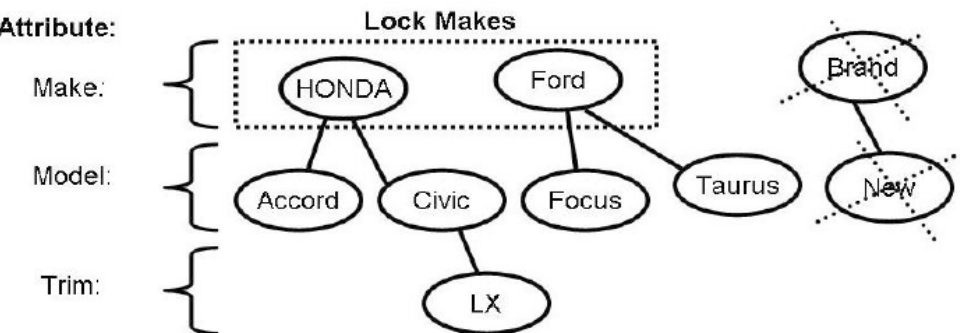
**Key: redundancy.** At some point you’ve gotten all you can from the posts

Attribute:

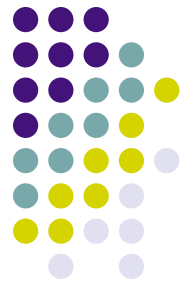


Iteration t

Attribute:



Iteration t+y

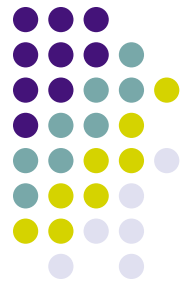


# Results: Information Extraction

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

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





# Results: Information Extraction

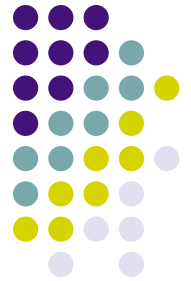
Iterative Locking Algorithm (ILA) vs. manual reference set (ARX for extraction)

Number of reference set tuples discovered



27,000 → wasted effort!

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



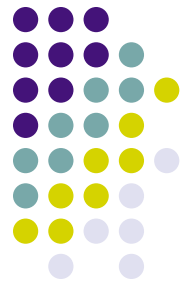
# Results: Information Extraction

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

Determined by locking



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

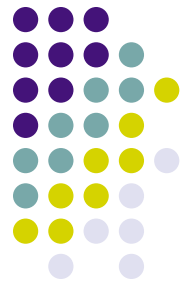


# Results: Information Extraction

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

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

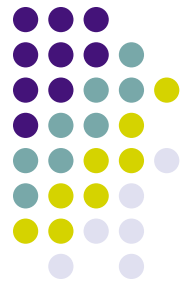
Competitive: fully automatic...



# Results: Information Extraction

Laptops (Craigslist): 2,400 posts			
<i>Manufacturer</i>	Recall	Prec.	F-Mes.
ILA (295)	60.42	74.35	66.67
Overstock (279)	84.41	95.59	89.65
<i>Model</i>	Recall	Prec.	F-Mes.
ILA (295)	61.91	76.18	68.31
Overstock (279)	43.19	80.88	56.31
<i>Model Num.</i>	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			
<i>Brand</i>	Recall	Prec.	F-Mes.
ILA (1,392)	60.84	55.26	57.91
Skis.com (213)	83.62	87.05	85.30
<i>Model</i>	Recall	Prec.	F-Mes.
ILA (1,392)	51.33	48.93	50.10
Skis.com (213)	28.12	67.95	39.77
<i>Model Spec.</i>	Recall	Prec.	F-Mes.
ILA (1,392)	39.14	56.35	46.29
Skis.com (213)	18.28	59.44	27.96



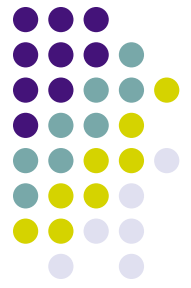
# Results: Information Extraction

Laptops (Craigslist): 2,400 posts			
<i>Manufacturer</i>	Recall	Prec.	F-Mes.
ILA (295)	60.42	74.35	66.67
Overstock (279)	84.41	95.59	89.65
<i>Model</i>	Recall	Prec.	F-Mes.
ILA (295)	61.91	76.18	68.31
Overstock (279)	43.19	80.88	56.31
<i>Model Num.</i>	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			
<i>Brand</i>	Recall	Prec.	F-Mes.
ILA (1,392)	60.84	55.26	57.91
Skis.com (213)	83.62	87.05	85.30
<i>Model</i>	Recall	Prec.	F-Mes.
ILA (1,392)	51.33	48.93	50.10
Skis.com (213)	28.12	67.95	39.77
<i>Model Spec.</i>	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			
<i>Manufacturer</i>	Recall	Prec.	F-Mes.
ILA (295)	60.42	74.35	66.67
Overstock (279)	84.41	95.59	89.65
<i>Model</i>	Recall	Prec.	F-Mes.
ILA (295)	61.91	76.18	68.31
Overstock (279)	43.19	80.88	56.31
<i>Model Num.</i>	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			
<i>Brand</i>	Recall	Prec.	F-Mes.
ILA (1,392)	60.84	55.26	57.91
Skis.com (213)	83.62	87.05	85.30
<i>Model</i>	Recall	Prec.	F-Mes.
ILA (1,392)	51.33	48.93	50.10
Skis.com (213)	28.12	67.95	39.77
<i>Model Spec.</i>	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”,



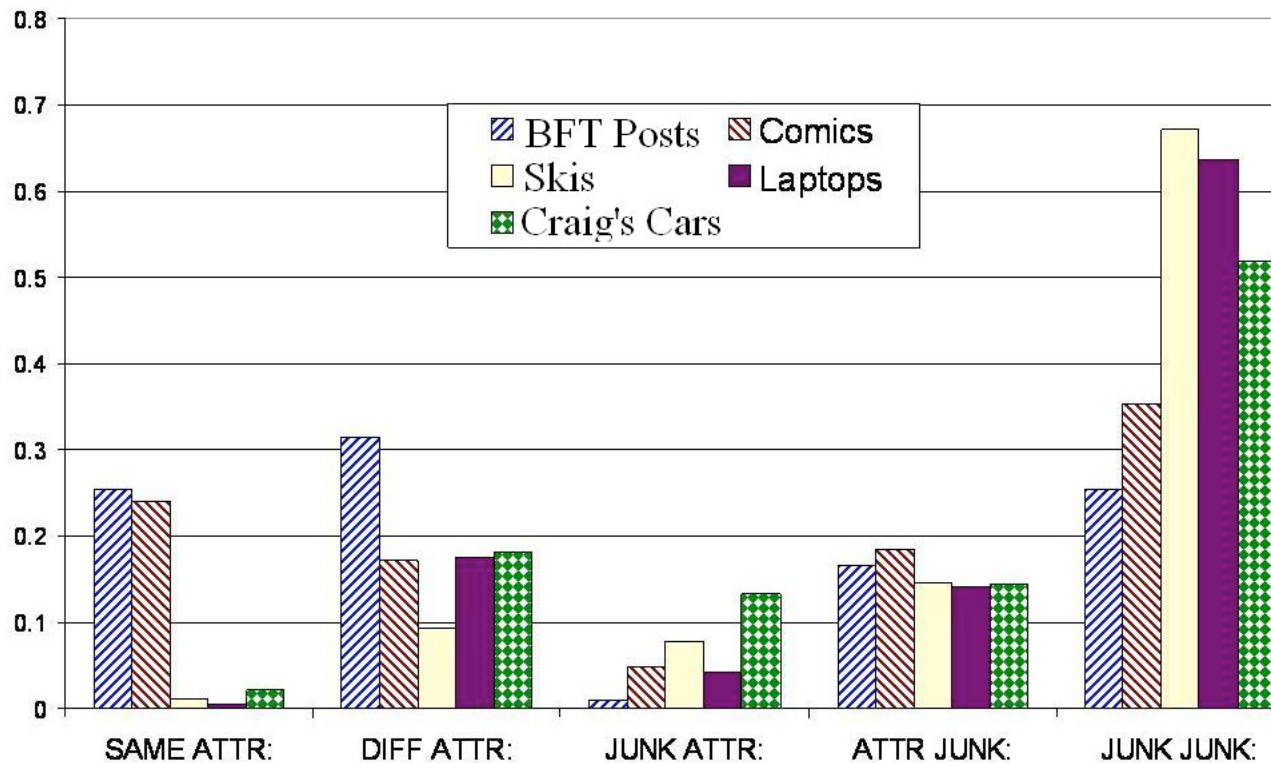
## ILA's Applicability

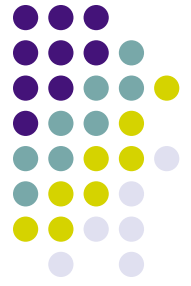
- Difficulty: multi-token, multi-attribute domains
    - BFT: 2.5\* Courtyard Rancho Cordova Marriott ...
      - “Boundary” issue
  - 5 bigram-types:
    - ... brand new Land Rover Discovery for...
- “DIFF ATTR”, “SAME ATTR”, “ATTR JUNK”, “JUNK ATTR”, “JUNK JUNK”



# ILA's Applicability

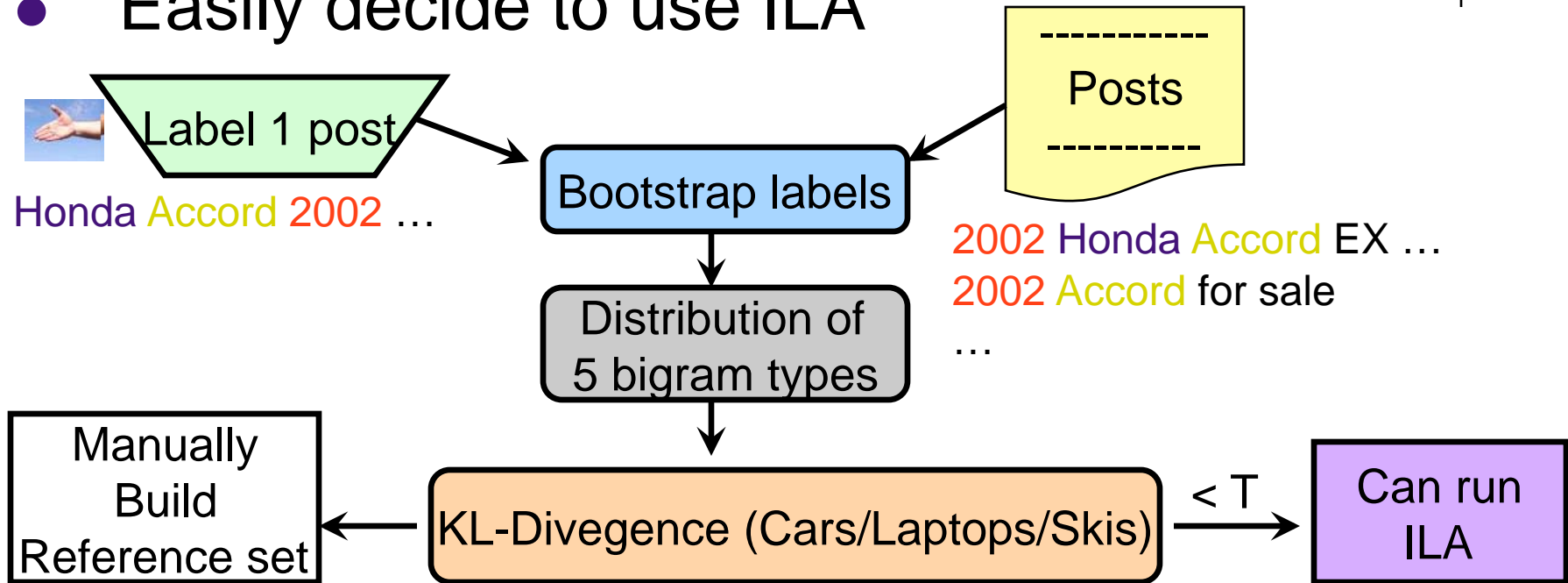
- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5\* Courtyard Rancho Cordova Marriott ...
    - “Boundary” issue
- 5 bigram-types:





# “Bootstrap-Compare”

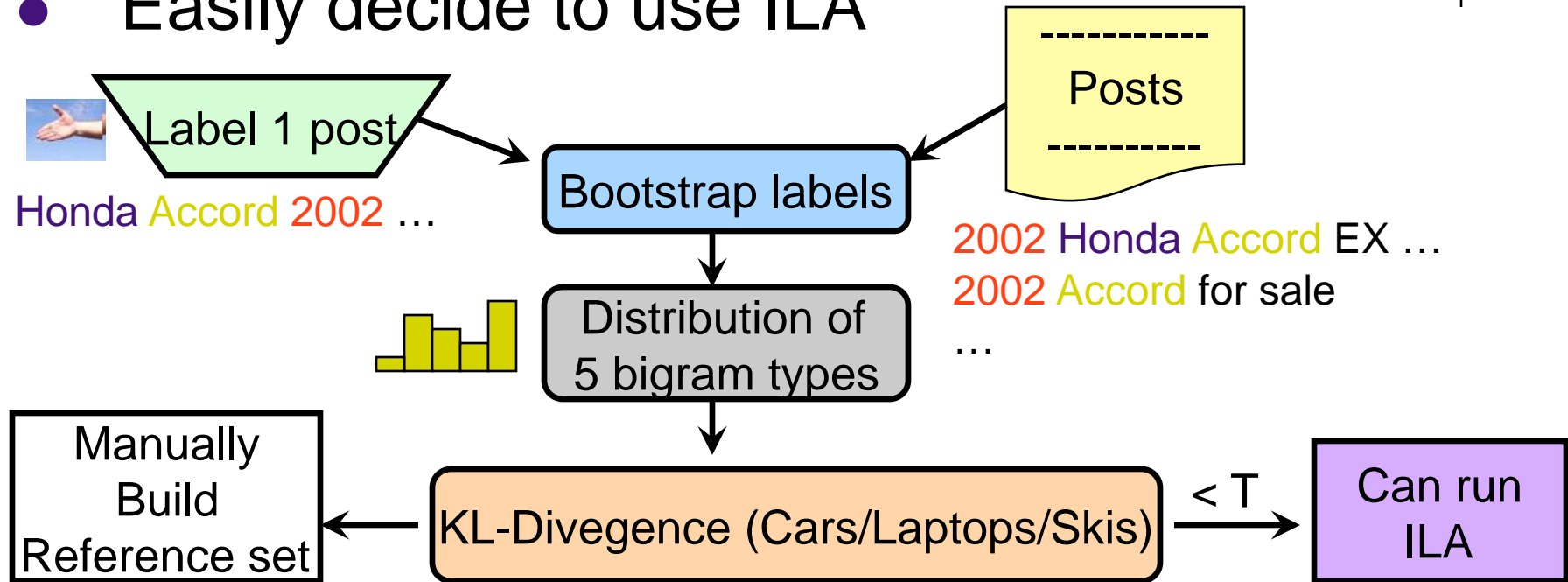
- Easily decide to use ILA

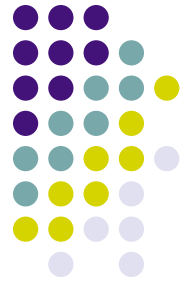




# “Bootstrap-Compare”

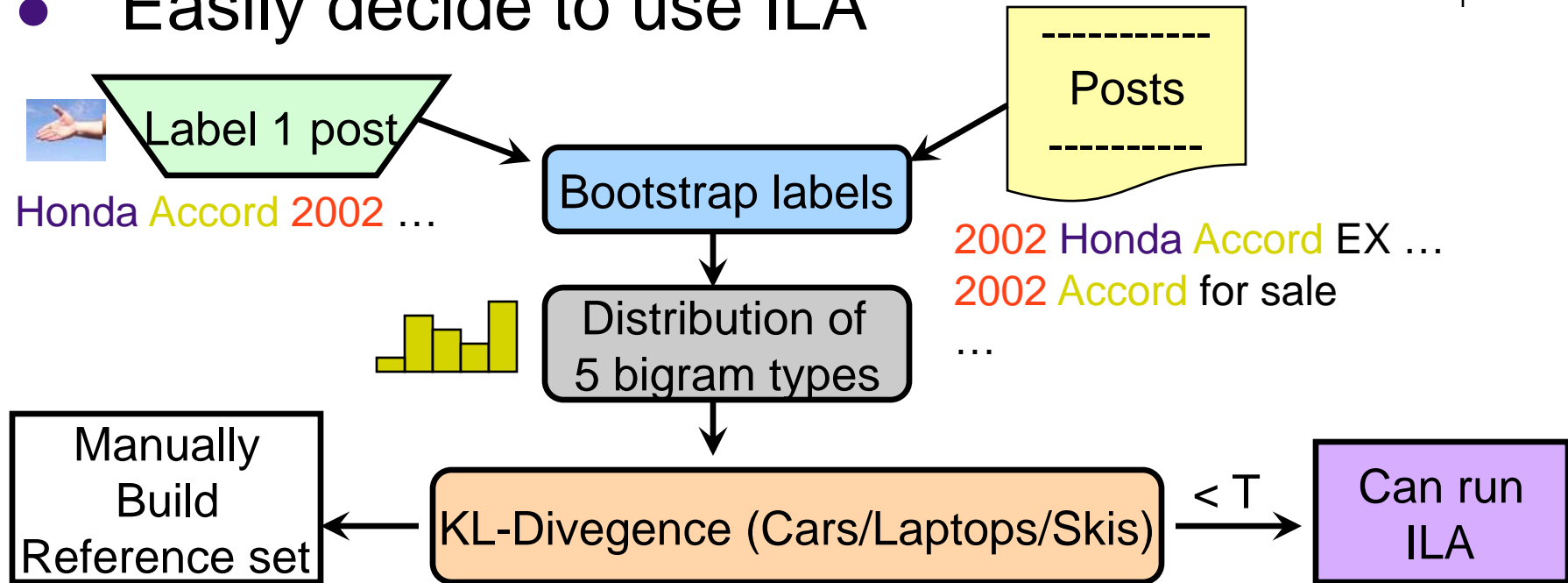
- Easily decide to use ILA





# “Bootstrap-Compare”

- Easily decide to use ILA



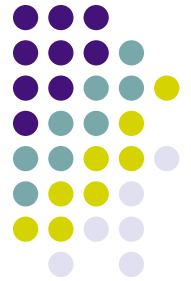
- Experiments

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

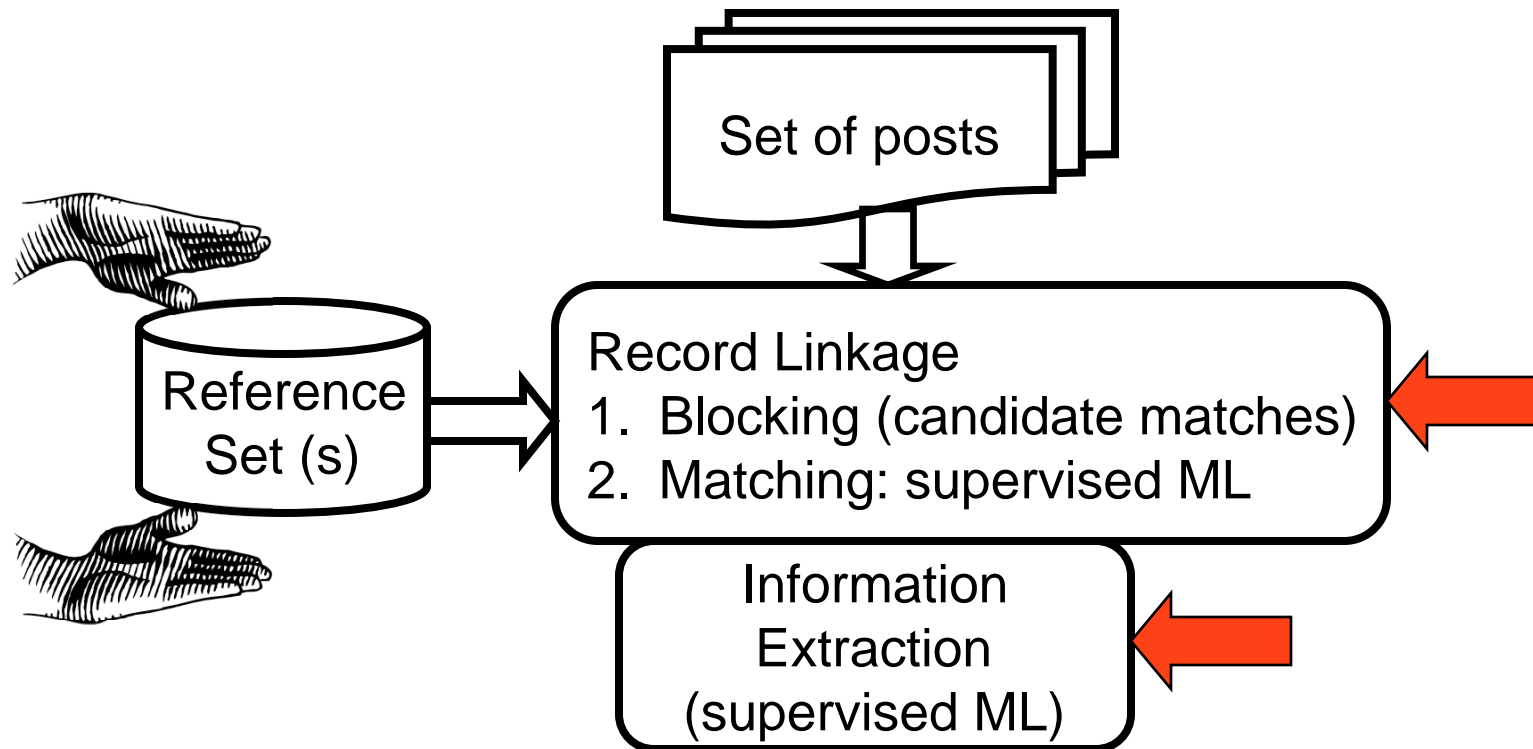


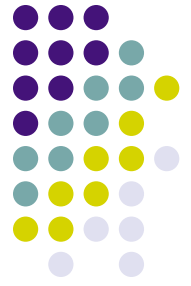
# Supervised Machine Learning for Extraction from Posts

JAIR, 2008



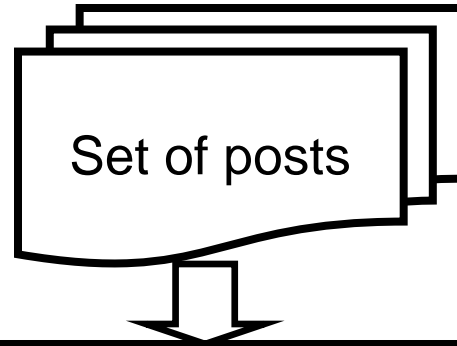
- Require **highest-accuracy** extraction
  - Ambiguity: **626**, Mazda or car price?





# Supervised Machine Learning for Extraction

*Record Level Similarity +  
Field Level Similarities*



## 1. Record Linkage

$$V_{RL} = \langle RL\_scores(\mathbf{post}, \mathbf{attribute}_1, \mathbf{attribute}_2, \dots, \mathbf{attribute}_n), \\ RL\_scores(\mathbf{post}, \mathbf{attribute}_1), \\ \dots, \\ RL\_scores(\mathbf{post}, \mathbf{attribute}_n) \rangle$$

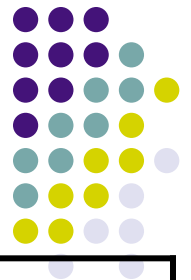
Binary Rescoring

SVM

## 2. Supervised Extraction

Compare to match's attributes

Multiclass-SVM / CRF



# 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?

