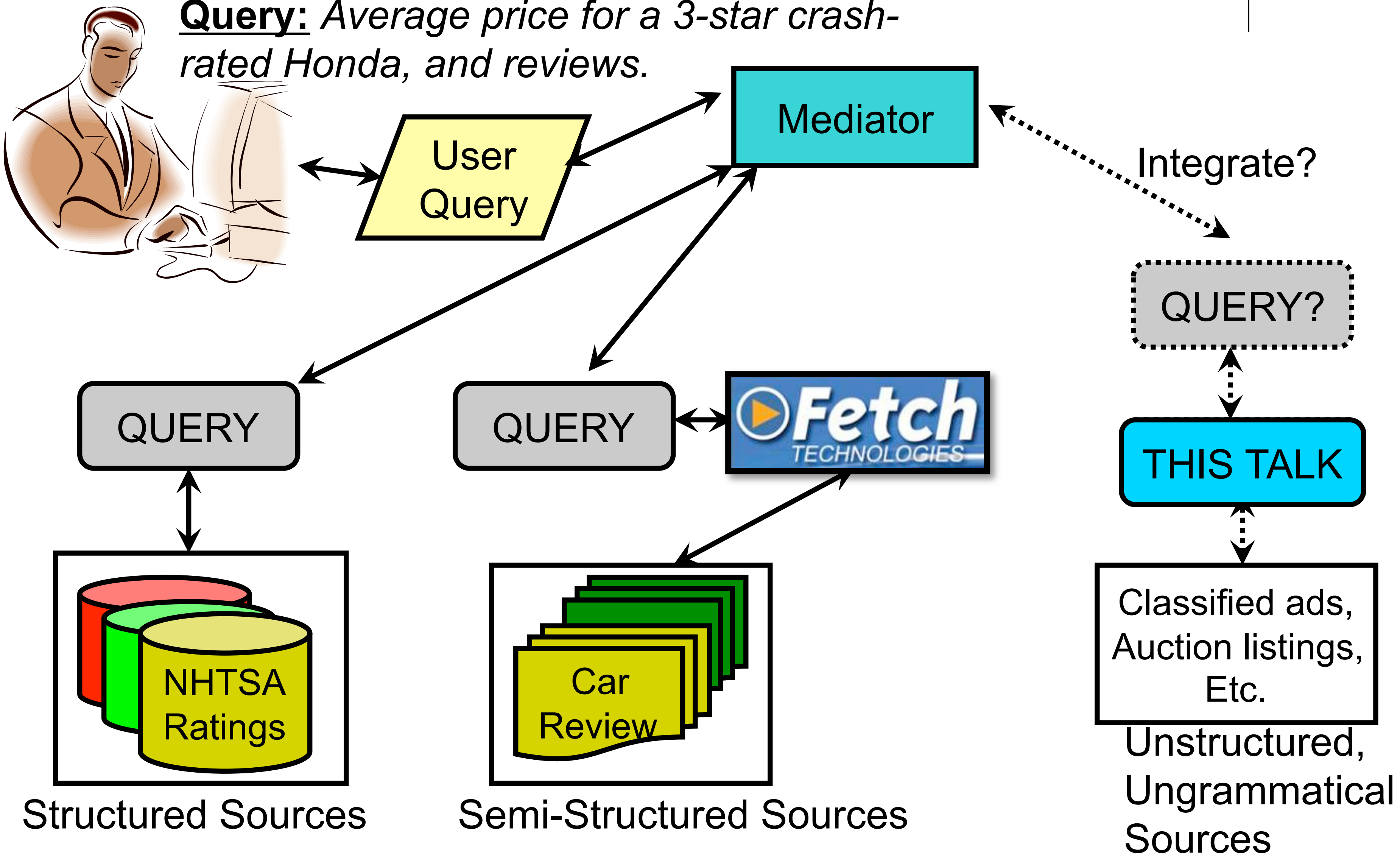
A Reference-Set Approach to Information Extraction from Unstructured, Ungrammatical Data Sources

Craig Knoblock
University of Southern California

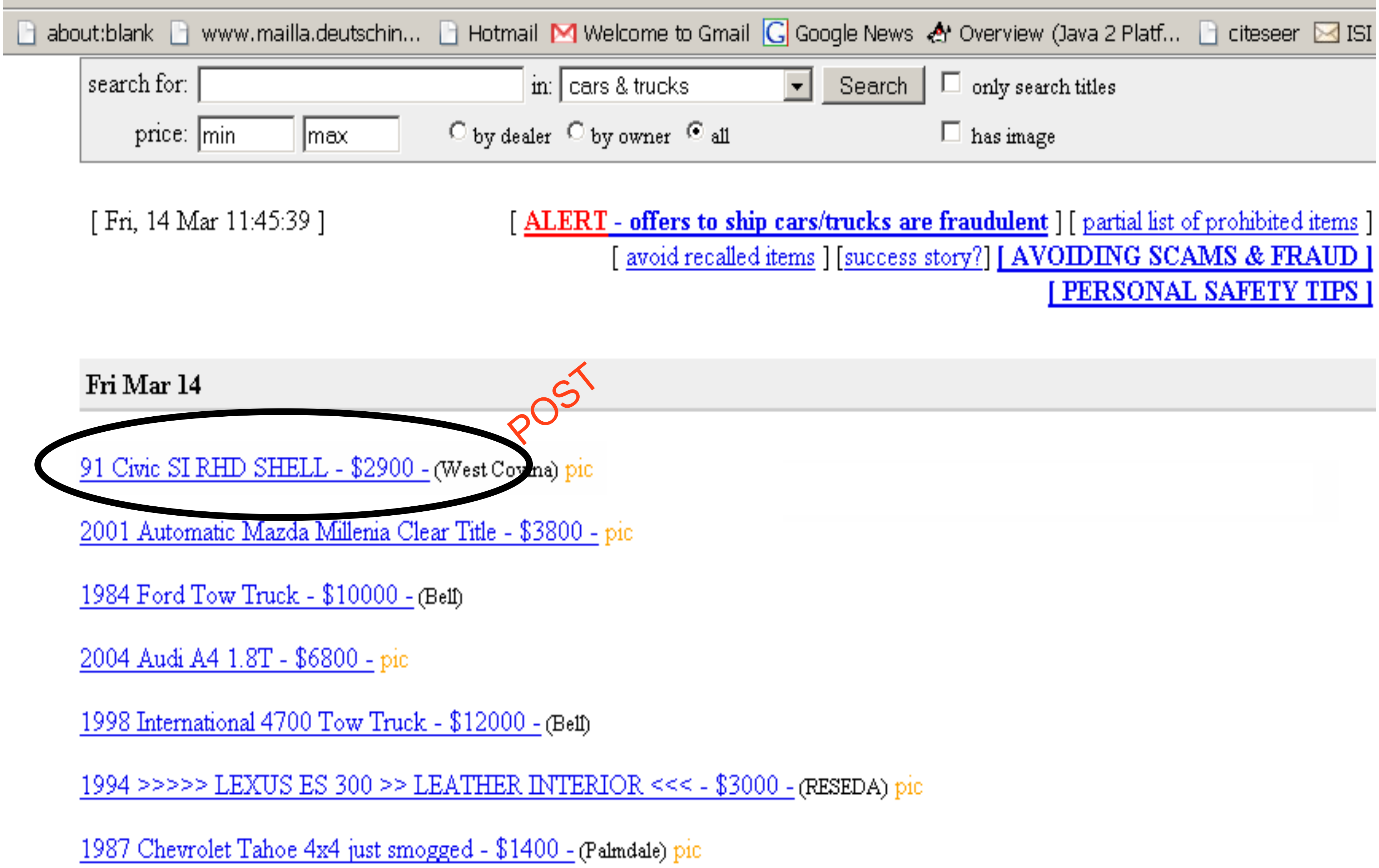
This is joint work with
Matthew Michelson
Fetch Technologies

Motivation: Data Integration

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



Unstructured, Ungrammatical Data:



[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)

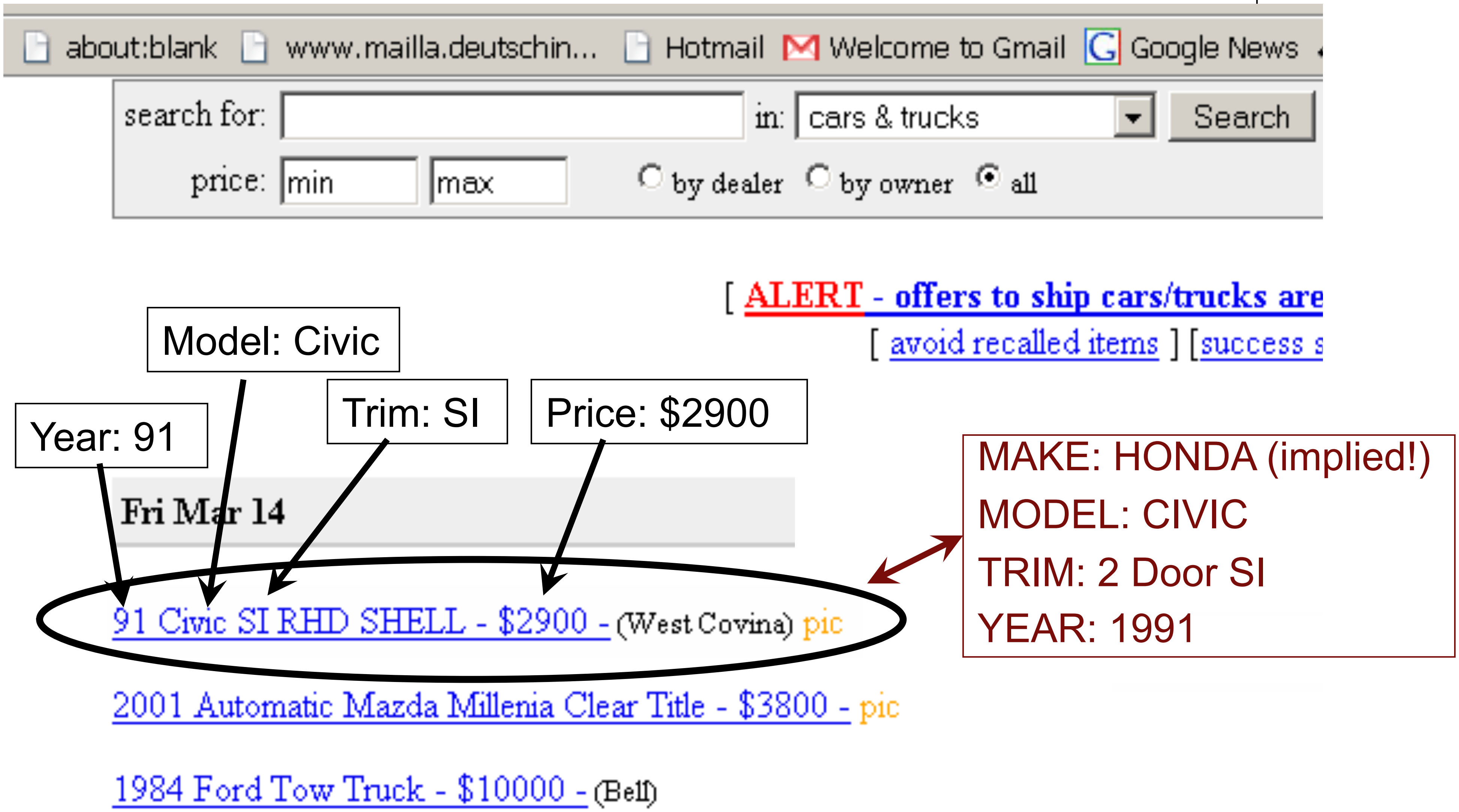
2004 Audi A4 1.8T - $6800 - pic

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

Structured Queries? … Information Extraction/Annotation!



about:blank www.mailla.deutschin... Hotmail Welcome to Gmail Google News

search for: in: cars & trucks Search

price: min max ○ by dealer ○ by owner ● all

[ALERT - offers to ship cars/trucks are

[avoid recalled items] [success s

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)

Model: Civic

Year: 91

Trim: SI

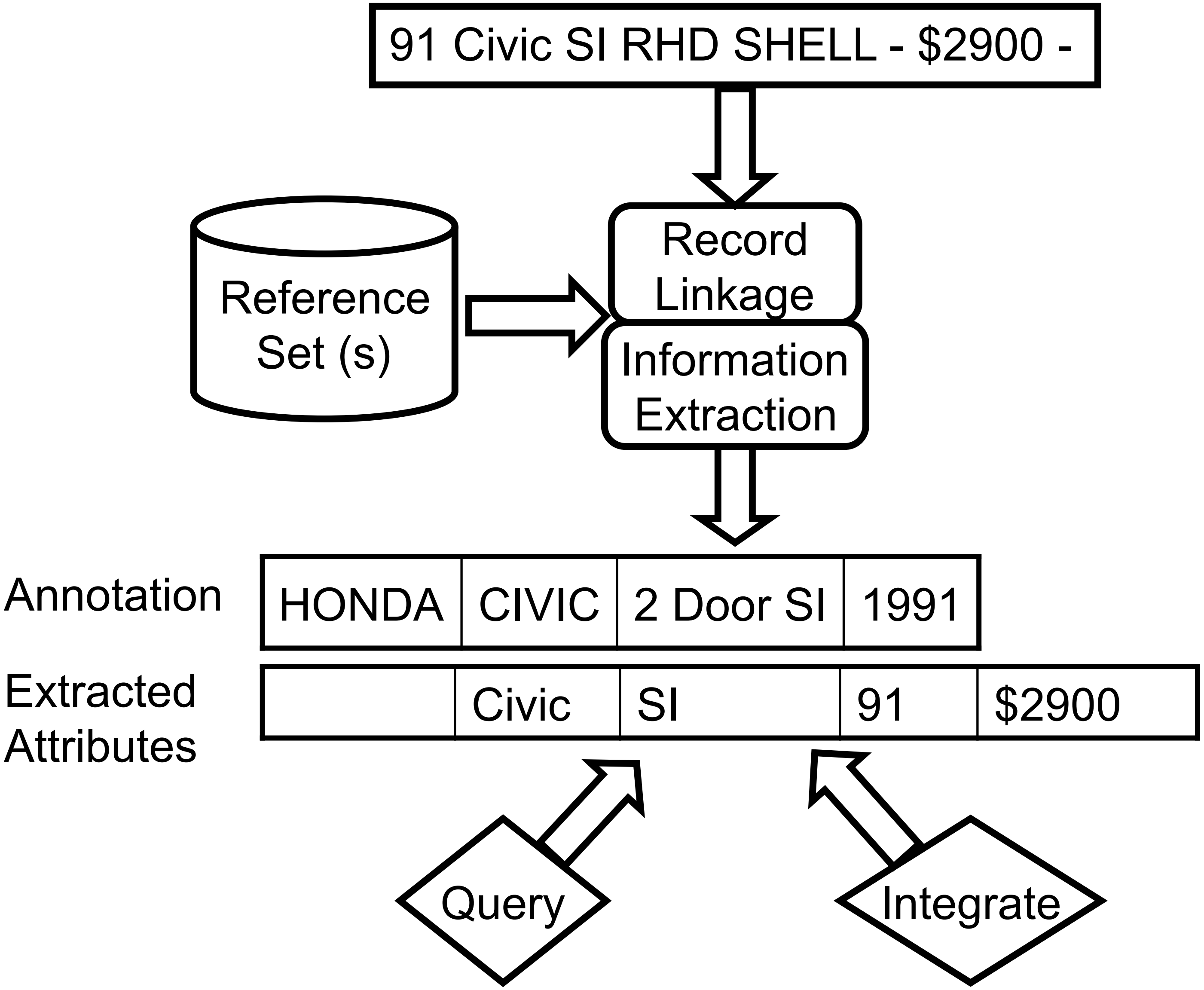
Price: $2900

MAKE: HONDA (implied!)
MODEL: CIVIC
TRIM: 2 Door SI
YEAR: 1991

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

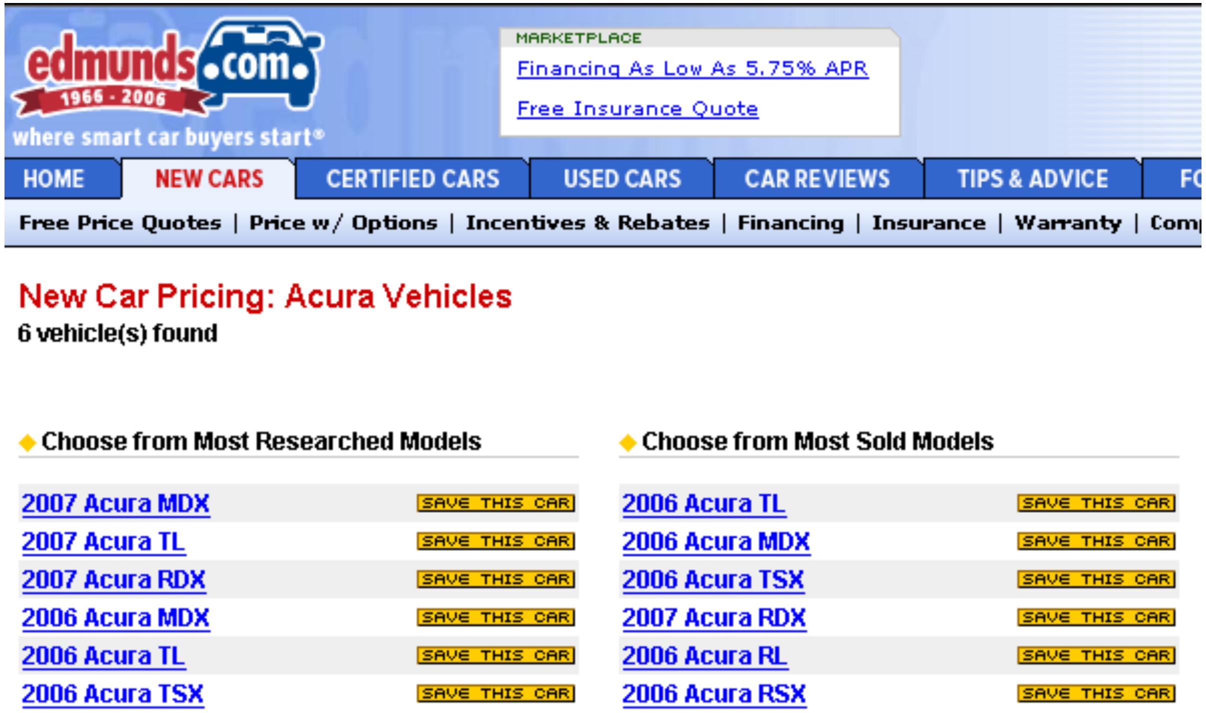


91 Civic SI RHD SHELL - $2900 -

Annotation	HONDA	CIVIC	2 Door SI	1991	
Extracted Attributes		Civic	SI	91	$2900

Reference Sets

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


edmunds.com
1966 - 2006
where smart car buyers start®
MARKETPLACE
Financing As Low As 5.75% APR
Free Insurance Quote
HOME
NEW CARS
CERTIFIED CARS
USED CARS
CAR REVIEWS
TIPS & ADVICE
FO
Free Price Quotes | Price w/ Options | Incentives & Rebates | Financing | Insurance | Warranty | Com

New Car Pricing: Acura Vehicles

6 vehicle(s) found

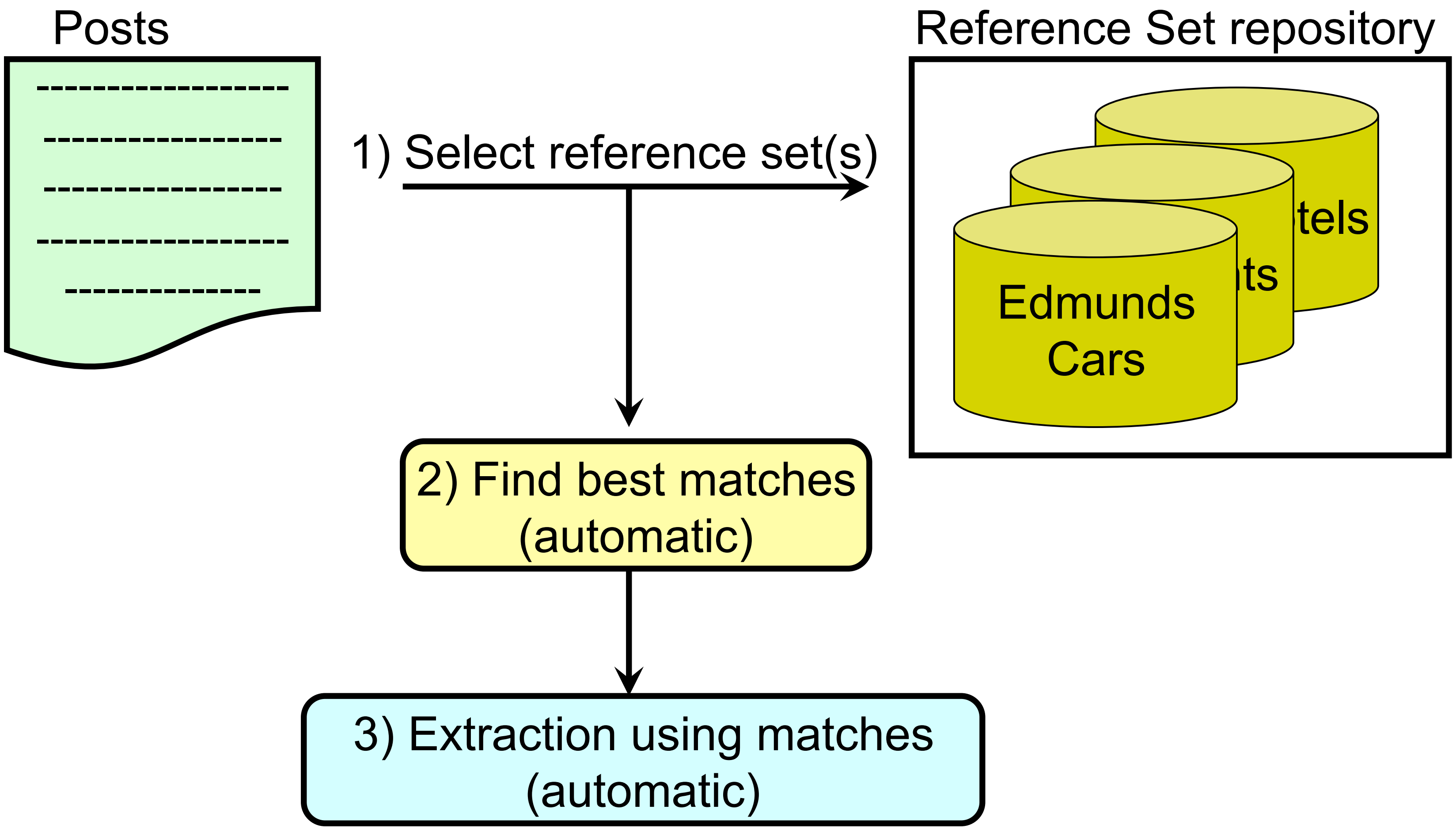
◆ Choose from Most Researched Models		◆ Choose from Most Sold Models	
2007 Acura MDX	SAVE THIS CAR	2006 Acura TL	SAVE THIS CAR
2007 Acura TL	SAVE THIS CAR	2006 Acura MDX	SAVE THIS CAR
2007 Acura RDX	SAVE THIS CAR	2006 Acura TSX	SAVE THIS CAR
2006 Acura MDX	SAVE THIS CAR	2007 Acura RDX	SAVE THIS CAR
2006 Acura TL	SAVE THIS CAR	2006 Acura RL	SAVE THIS CAR
2006 Acura TSX	SAVE THIS CAR	2006 Acura RSX	SAVE THIS CAR

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

Talk Topics

- Automatic matching and extraction using reference sets
  - Michelson & Knoblock, IJDAR, 2007
  - Code @ mmichelson.com
- Automatically building reference sets from the posts
  - Michelson & Knoblock, IJCAI, 2009
  - Michelson & Knoblock, JAIR, 2010
- Supervised machine learning w/ reference sets
  - Michelson & Knoblock, IJCAI, 2005
  - Michelson & Knoblock, JAIR, 2008
  - Code @ mmichelson.com

Automatic method: Three steps



Posts

Reference Set repository

1) Select reference set(s)

2) Find best matches (automatic)

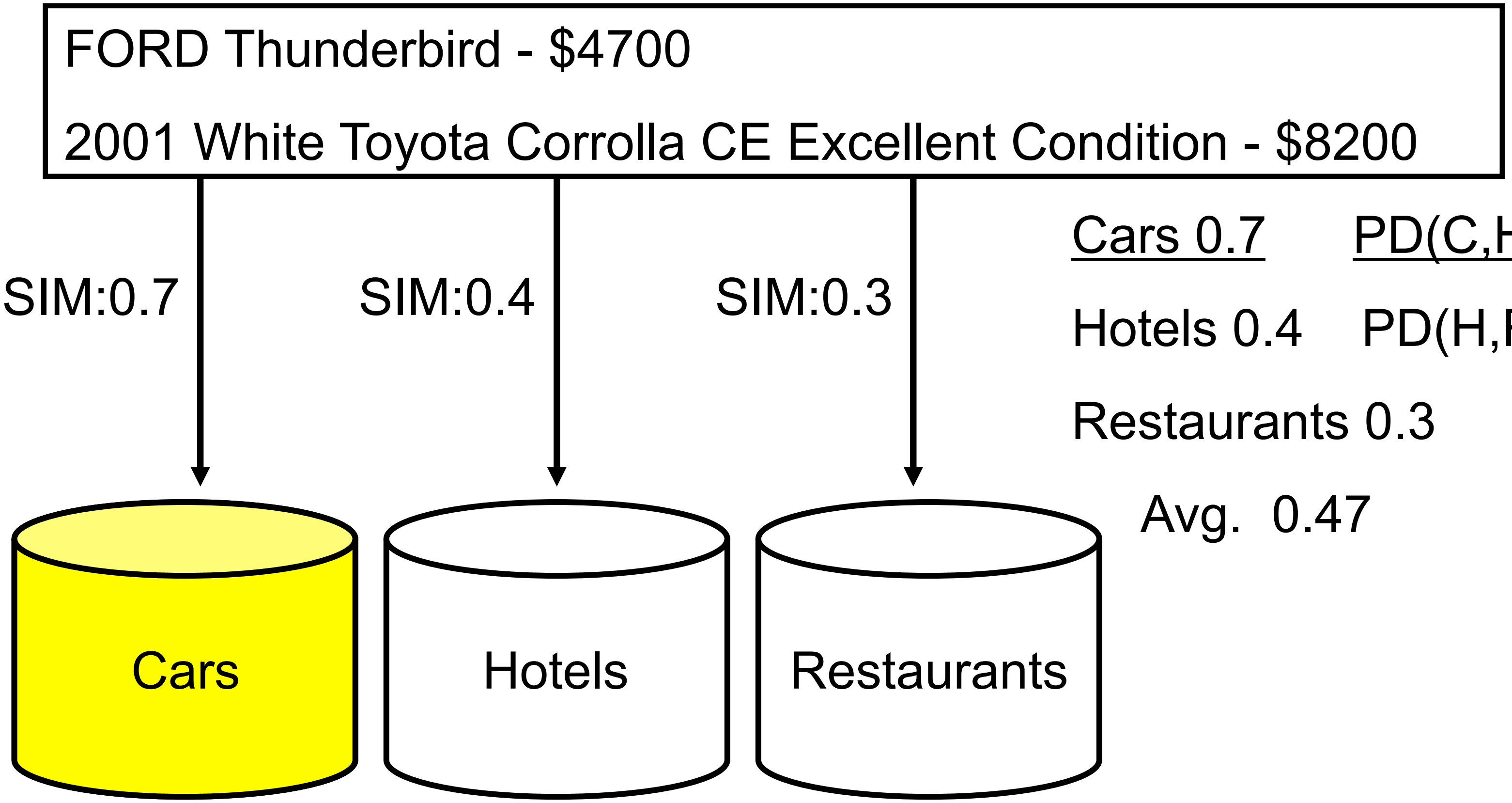
3) Extraction using matches (automatic)

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

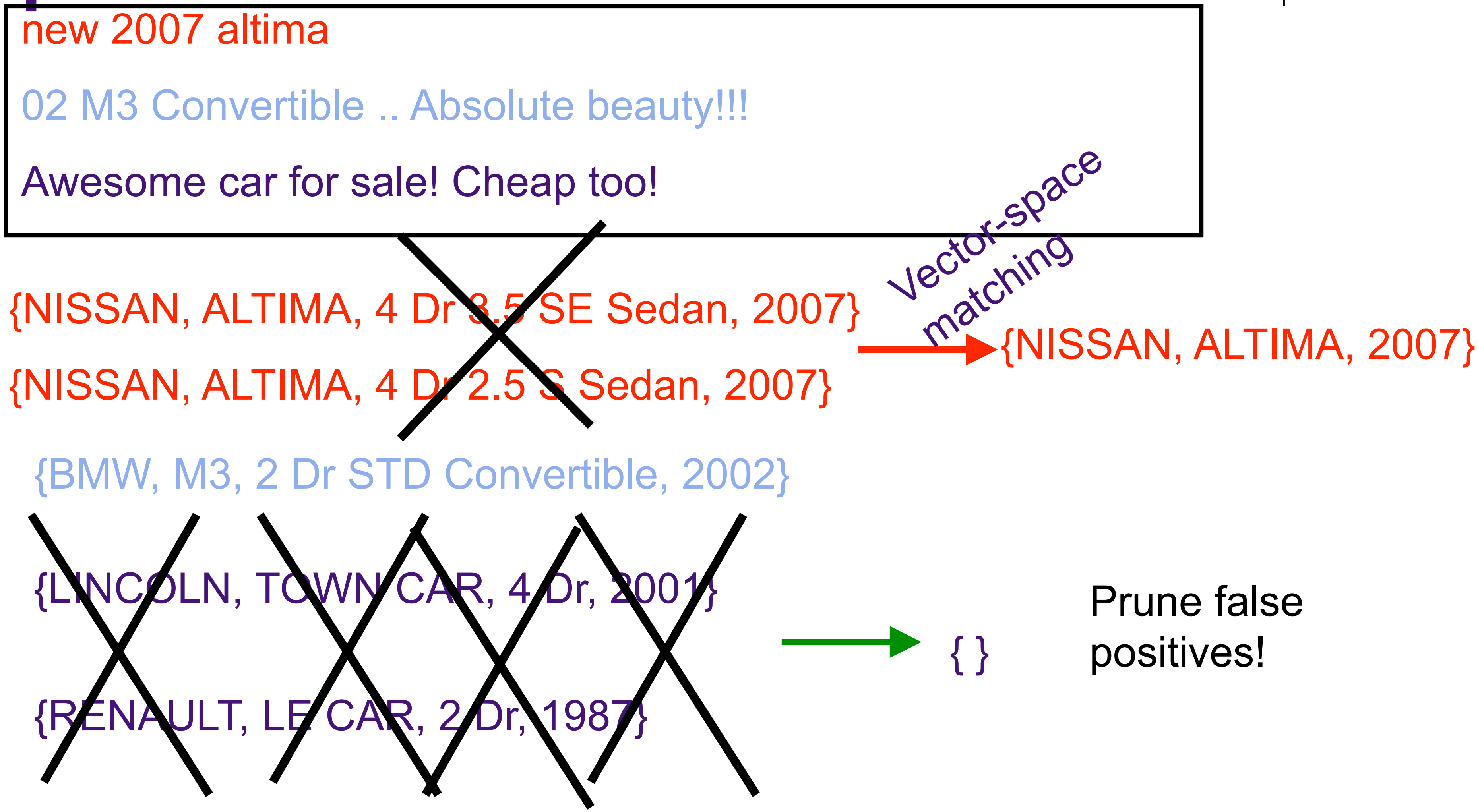
Cars 0.7 PD(C,H) = 0.75 > T

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

Restaurants 0.3

Avg. 0.47

Automatic 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}

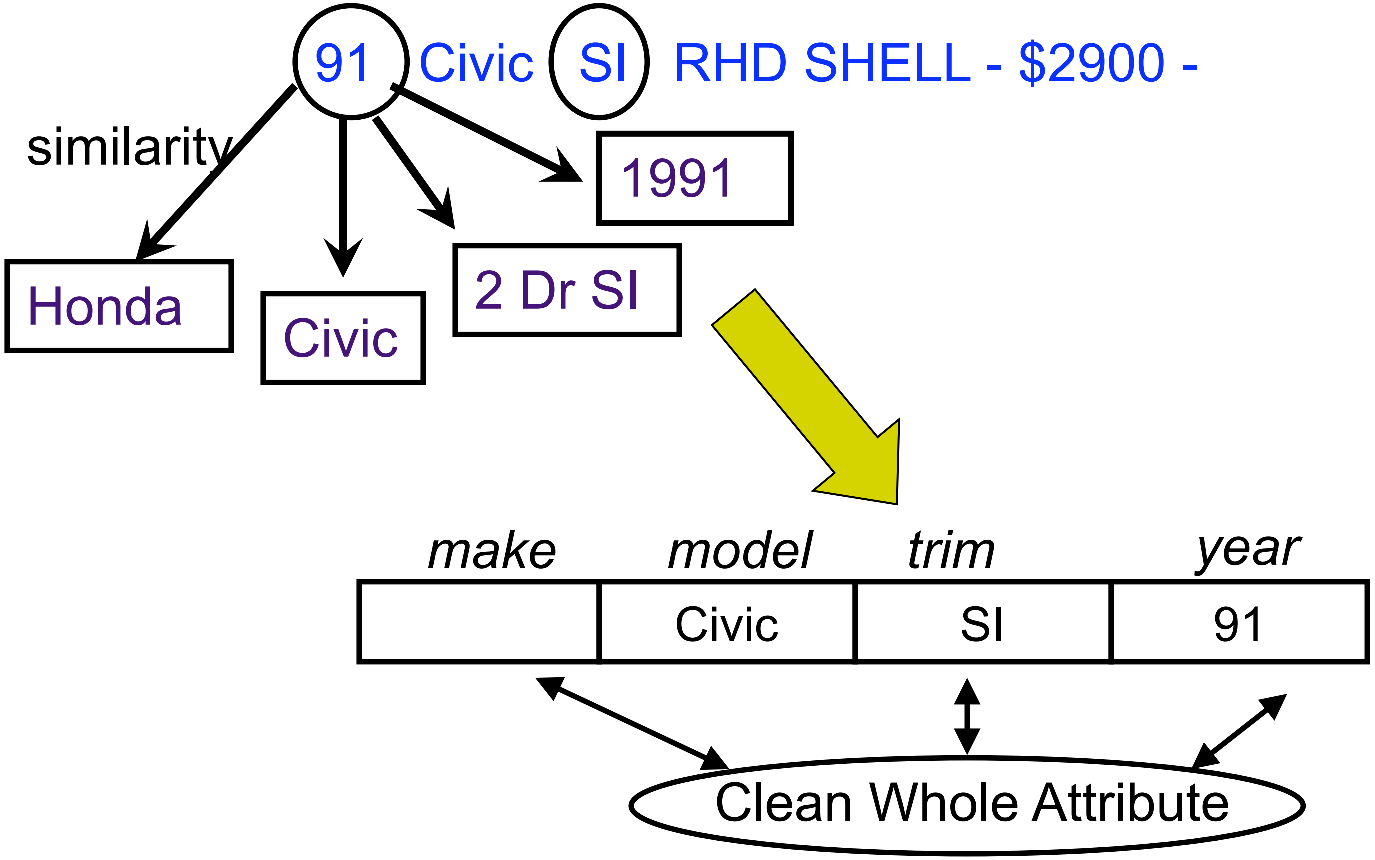
{BMW, M3, 2 Dr STD Convertible, 2002}

{LINCOLN, TOWN CAR, 4 Dr, 2001}

{RENAULT, LE CAR, 2 Dr, 1987}

Prune false positives!

Automatic Extraction



91 Civic SI RHD SHELL - $2900 -

similarity

1991

Honda

Civic

2 Dr SI

make	*model*	*trim*	*year*
	Civic	SI	91

Clean Whole Attribute

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!

Construction of Reference Sets

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

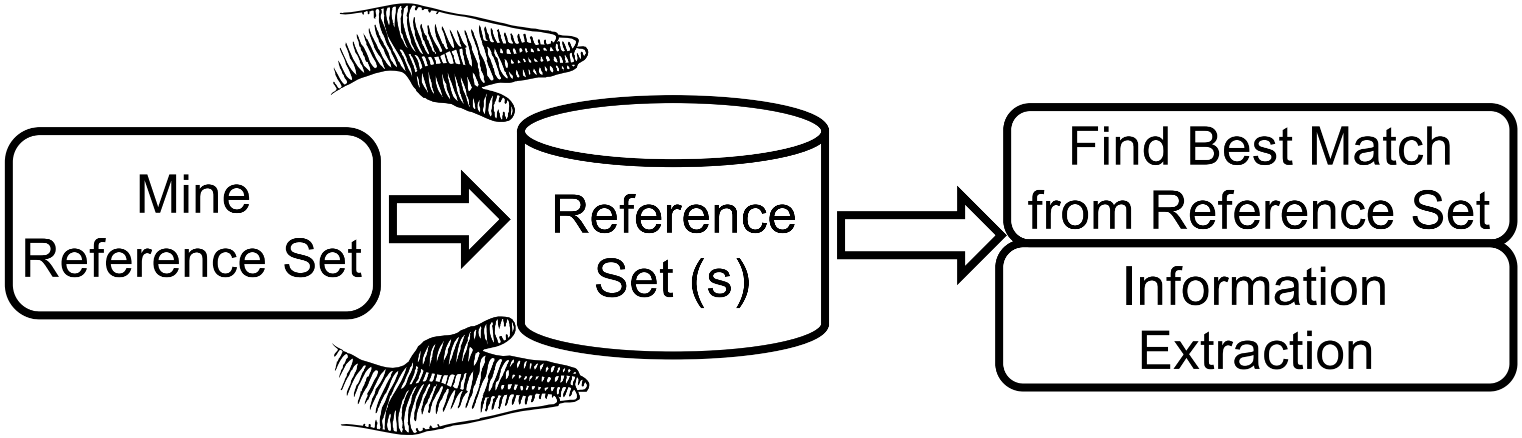
HP Pavillion DV2000 laptop
Gateway ML6230, Intel Cel …

- What about coverage?

Ford	Focus
Dodge	Caravan

?

ACURA TL 3.2 VTEC - 1999


Mine Reference Set

Reference Set (s)

Find Best Match from Reference Set

Information Extraction

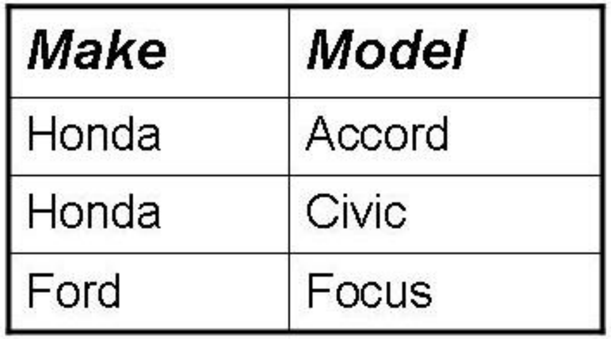
Seed-Based Reference Set Construction

- Use posts themselves
  - Overcome difficulty in finding full reference sets
    - Enumeration
    - Dynamic data
  - Overcome coverage issues
    - Using posts guarantees coverage

Seed-Based Reference Set Construction

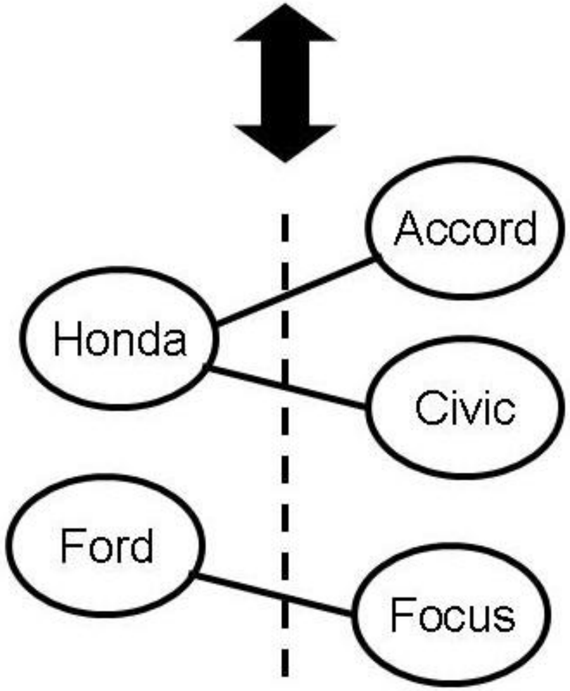
- Seeds
  - Smallest (most obvious) domain knowledge
    - Computer Makers: Apple, Dell, Lenovo
    - Easy to enumerate
  - Constrains tuples constructed (roots)
    - Cleaner reference set
  - Relatively static
    - Less change to worry about
- Posts themselves to fill in details
  - Computer Models, Model Nums…

Entity Trees



Make	*Model*
Honda	Accord
Honda	Civic
Ford	Focus

Reference Set



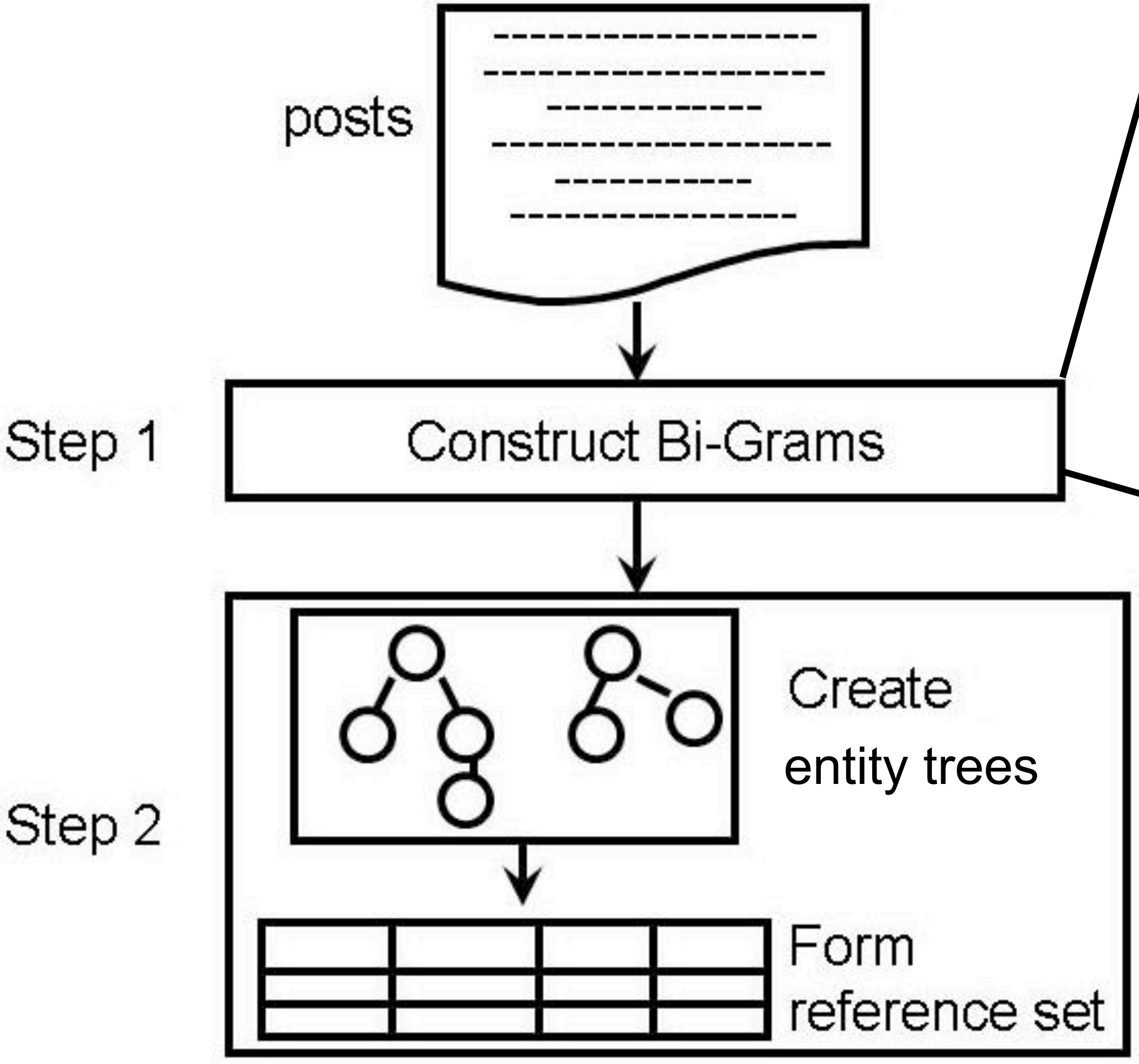
Forest of “Entity Trees”

Reference Set Construction

=

Constructing this forest

Entity Trees from Posts



91 Civic SI RHD …

↓

{91 Civic}
{Civic SI}
{SI RHD}
…

Seeds = roots

Fill in rest using posts

Constructing Entity Trees

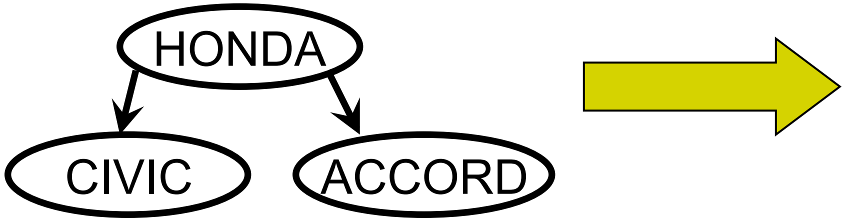
- Sanderson & Croft heuristic
  - x SUBSUMES y *IF* $P(x|y) \geq 0.75$ & $P(y|x) \leq P(x|y)$
- Merge heuristic
  - MERGE(x,y) *IF* x SUBSUMES y & $P(y|x) \geq 0.75$

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

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

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

- Construct hierarchies, then flatten



HONDA	CIVIC
HONDA	ACCORD

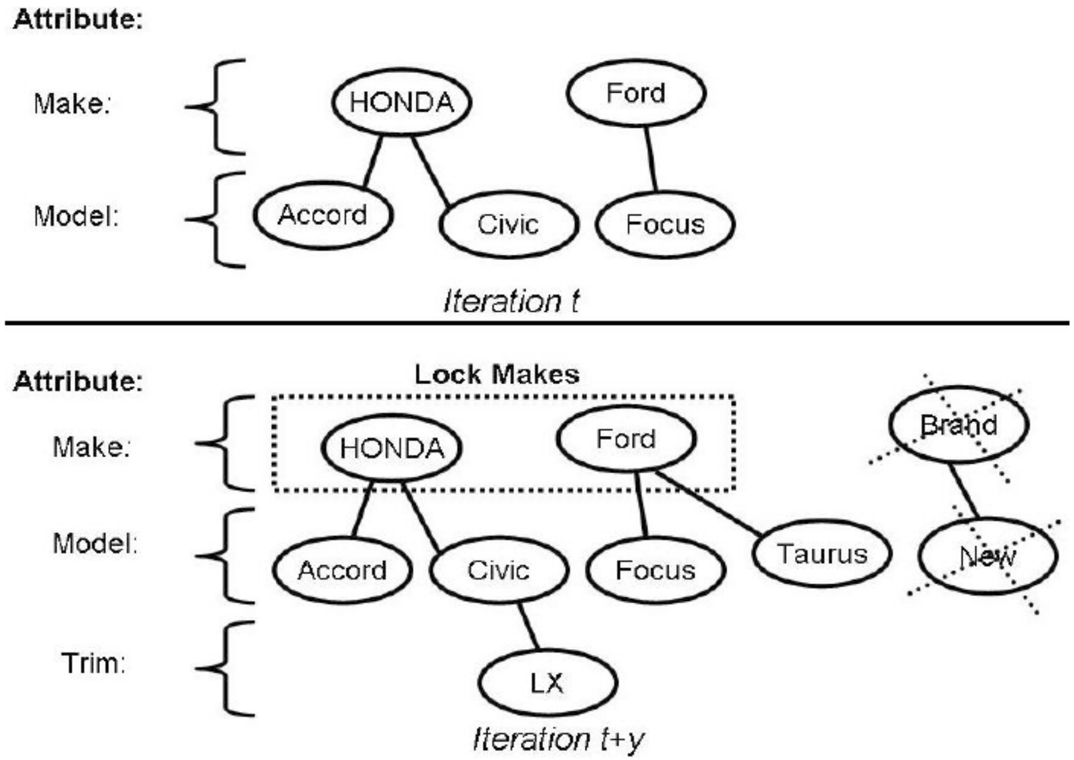
General Tokens

- {a, y}, {b, y}, {c, y} → y is "general token"
  - Occurs across entity trees…
  - Instead use P({a U b U c } | y)
  - e.g. car trims: Pathfinder LE, Corolla LE, …
  - Build entity trees
    - Do 1 Scan
      - Build initial trees
    - Iterate
      - Find "general tokens"

No seeds?

- “Iterative Locking Algorithm”
  - Instead of seeds, “lock” levels of the tree
  - Entropy of finding current leaves
    - Uncertainty labeling attributes
    - Compare % diff across # posts
  - Locks out noise
- **How many posts are enough**?
  - When you lock all levels

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



Experiments & Results

- Goal
  - How to compare reference sets?
    - Ontology comparison is rather open…
    - Might not take into account utility of reference set…
  - Extraction = proxy task to compare reference sets
    - Poor coverage → poor recall
    - Noise → bad extractions → worse results
- Compare extraction (use ARX)
  - Constructed using seeds ("Seed-based")
  - Constructed without seeds ("Auto")
  - Manually constructed reference sets ("Manual")

Experiments & Results

Experimental Domains:

Name	Source	Attributes	Num. Posts
Cars	Craigslist	make, model, trim	2,568
Laptops	Craigslist	maker, model, model num.	2,921
Skis	eBay	brand, model, model spec.	4,981

Name	Source	Num. Records
Cars	Edmunds	~27,000
Laptops	Overstock	279
Skis	Skis.com	213

“Manual” reference sets

Name	Source	Num. Seeds
Cars	Edmunds	102 makes
Laptops	Wikipedia	40 makers
Skis	Skis.com	18 brands

Seed sets

Experiments & Results (seed based)

	vs. Auto	vs. Manual
Outperforms	9/9	5/9
Within 5%	9/9	7/9

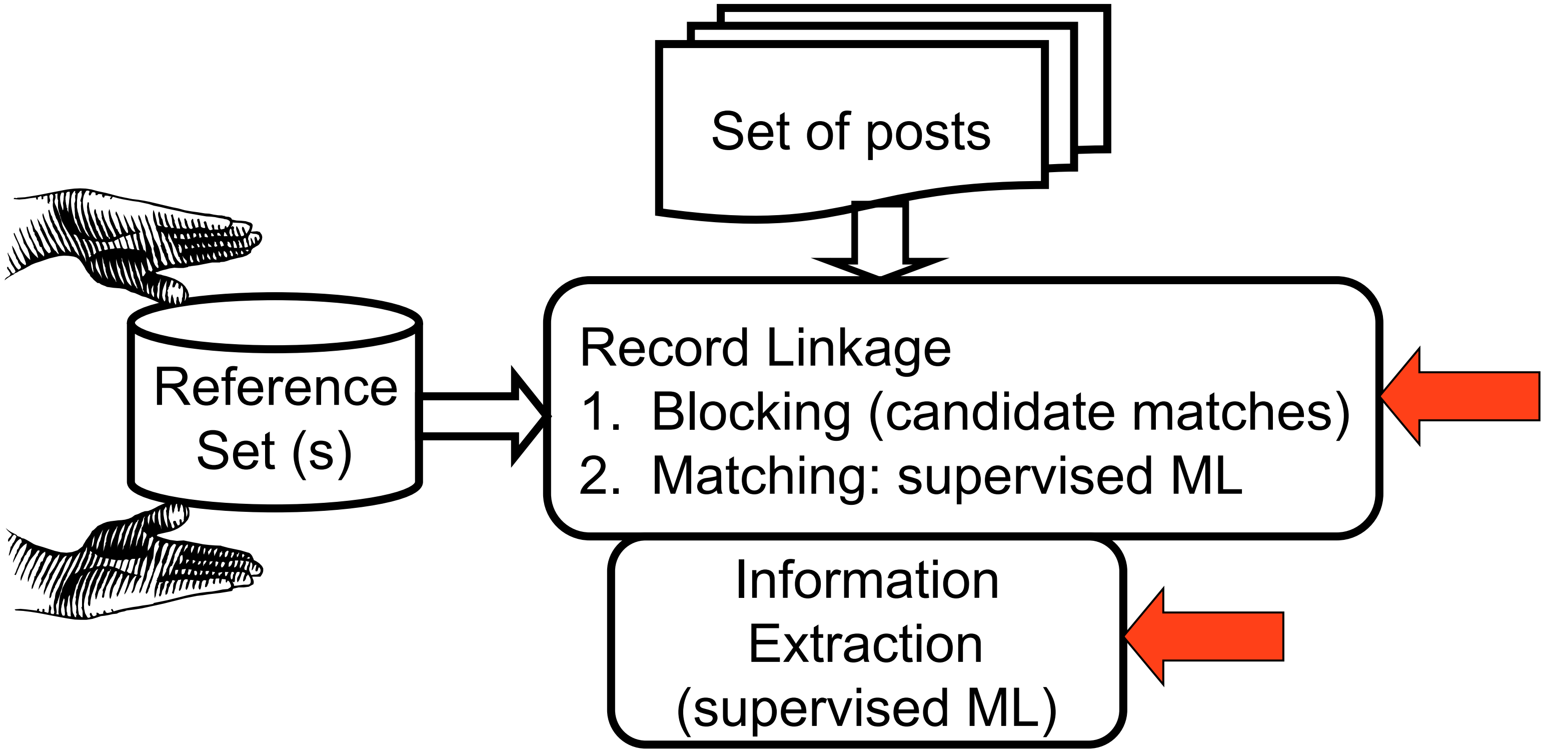
- Seed-based vs. Manual
  - Outperforms on majority of attributes / Competitive on most
    - # seeds << # records in manual reference set
  - Does best on hard to cover attributes
    - Ski model & model spec., Laptop model & model num.
      - Only 53.15% of values for these exist in manual sets!
      - Overstock = New computers, Craigslist = old computers
  - Poor performance vs. manual
    - Car trim: missing tokens (didn’t mine)
      - E.g. Manual = 4 Dr DX 4WD, Seed = DX
      - Miss “4 Dr” part of extraction → wrong in field-level results

Experiments & Results (locking based)

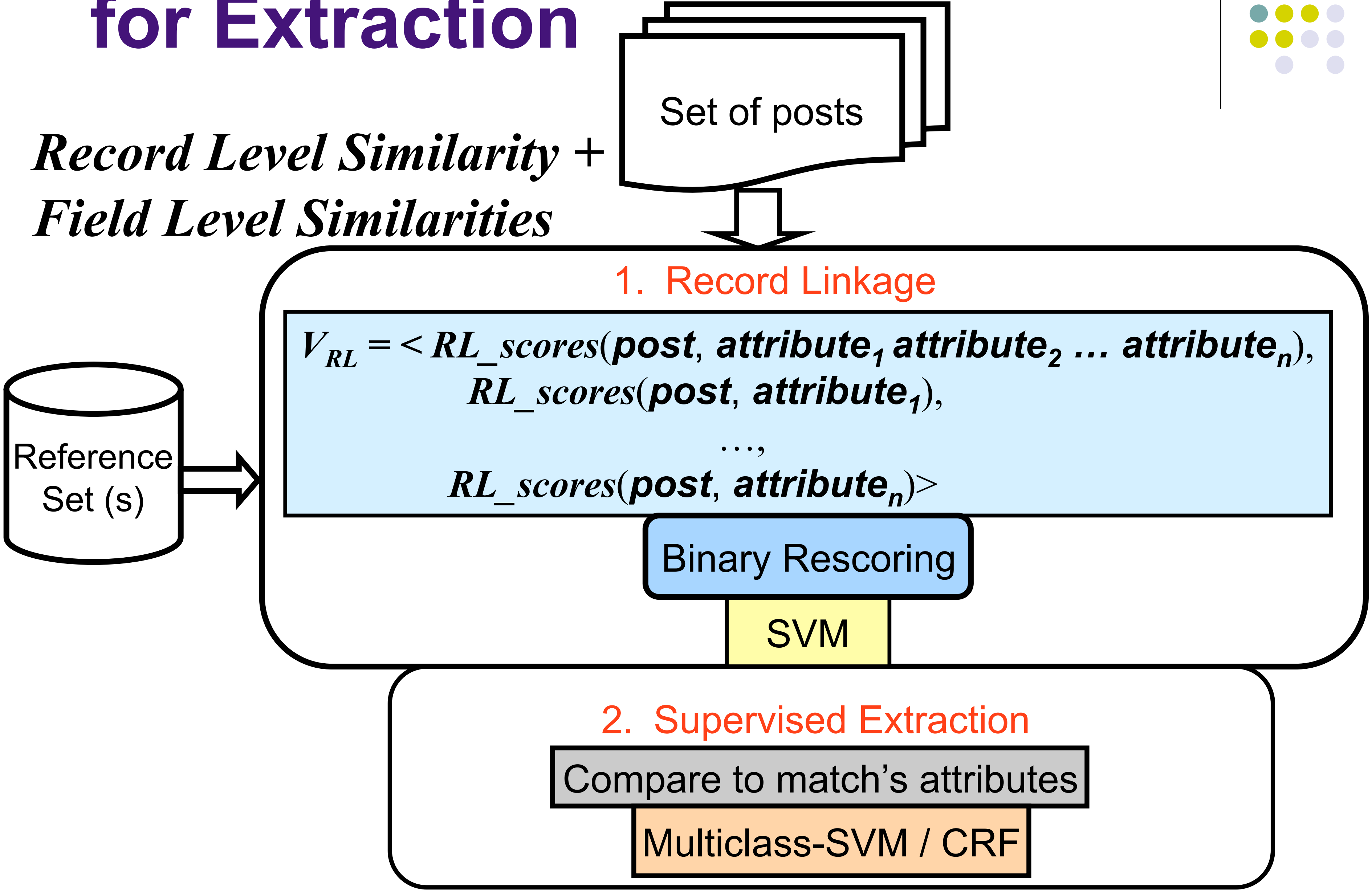
- Converges in all domains
  - E.g., locks before seen all posts
- Outperforms "Auto" on all Laptop attributes
  - Stat sig. 95%
- Cars/Skis
  - Only 1 significant difference vs. "Auto"
- → Should try to lock
  - Can't hurt you (only 1 significant drop), and in best case can help a lot (laptop)

Supervised Machine Learning for Extraction from Posts

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


Set of posts
Reference Set (s)
Record Linkage
1. Blocking (candidate matches)
2. Matching: supervised ML
Information Extraction (supervised ML)

Supervised Machine Learning for Extraction



Record Level Similarity + Field Level Similarities

Set of posts

Reference Set (s)

1. Record Linkage

$$\begin{aligned} V_{RL} = <\ & RL\_scores(\textbf{post}, \textbf{attribute}_1\ \textbf{attribute}_2 \ldots \textbf{attribute}_n), \\ & RL\_scores(\textbf{post}, \textbf{attribute}_1), \\ & \ldots, \\ & RL\_scores(\textbf{post}, \textbf{attribute}_n)> \end{aligned}$$

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)
- Blocking
  - 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)

- See papers for more comprehensive RW…

Conclusion: Topics Covered

- 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

Questions?

Code & Data:
mmichelson.com