Exploiting Background Knowledge to Build Reference Sets for Information Extraction

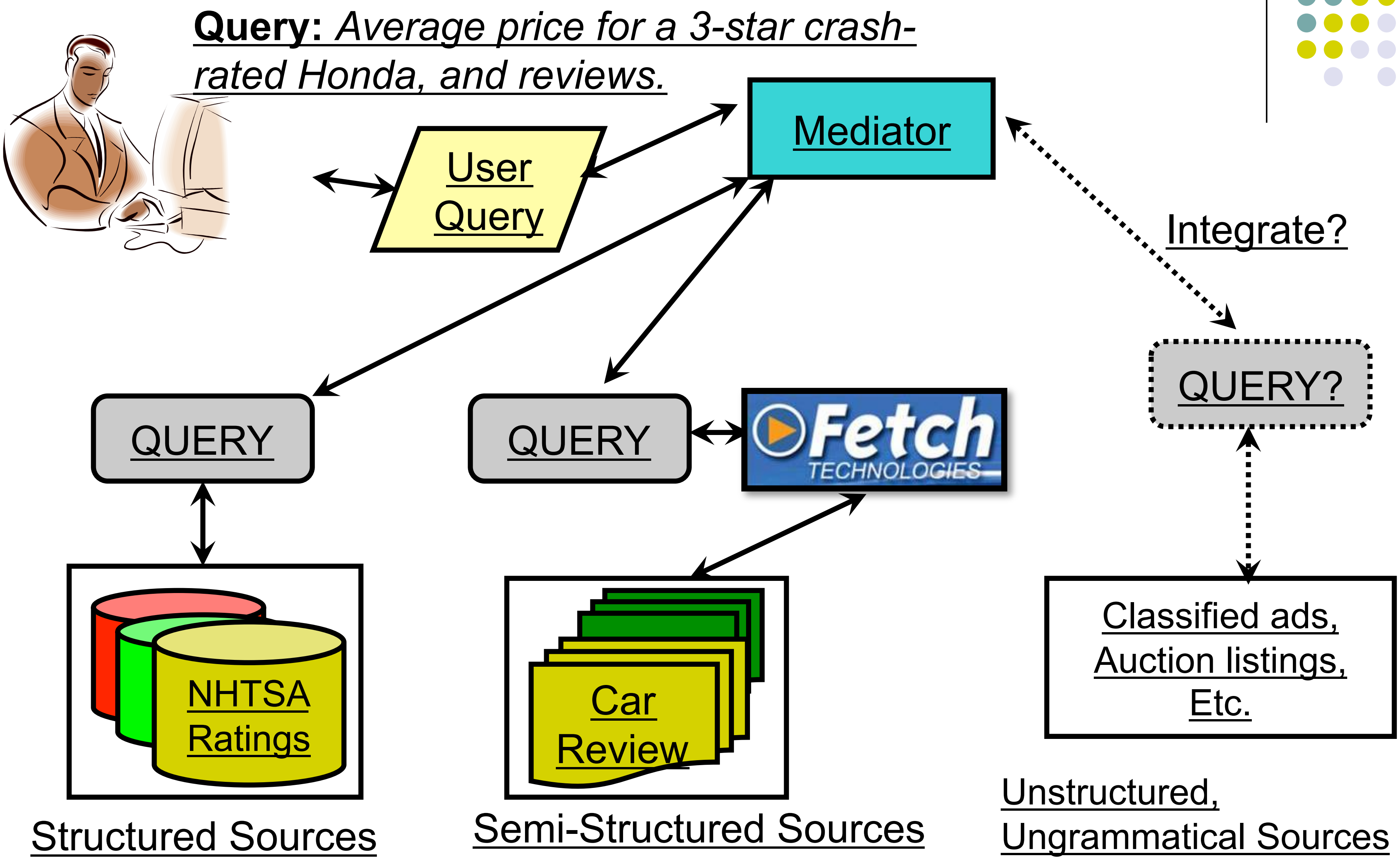
Matthew Michelson & Craig A. Knoblock

Fetch Technologies*

USC Information Sciences Institute

* Work done while at USC Information Sciences Institute

Motivation: Data Integration



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

User Query

Mediator

Integrate?

QUERY

QUERY

QUERY?

NHTSA Ratings

Car Review

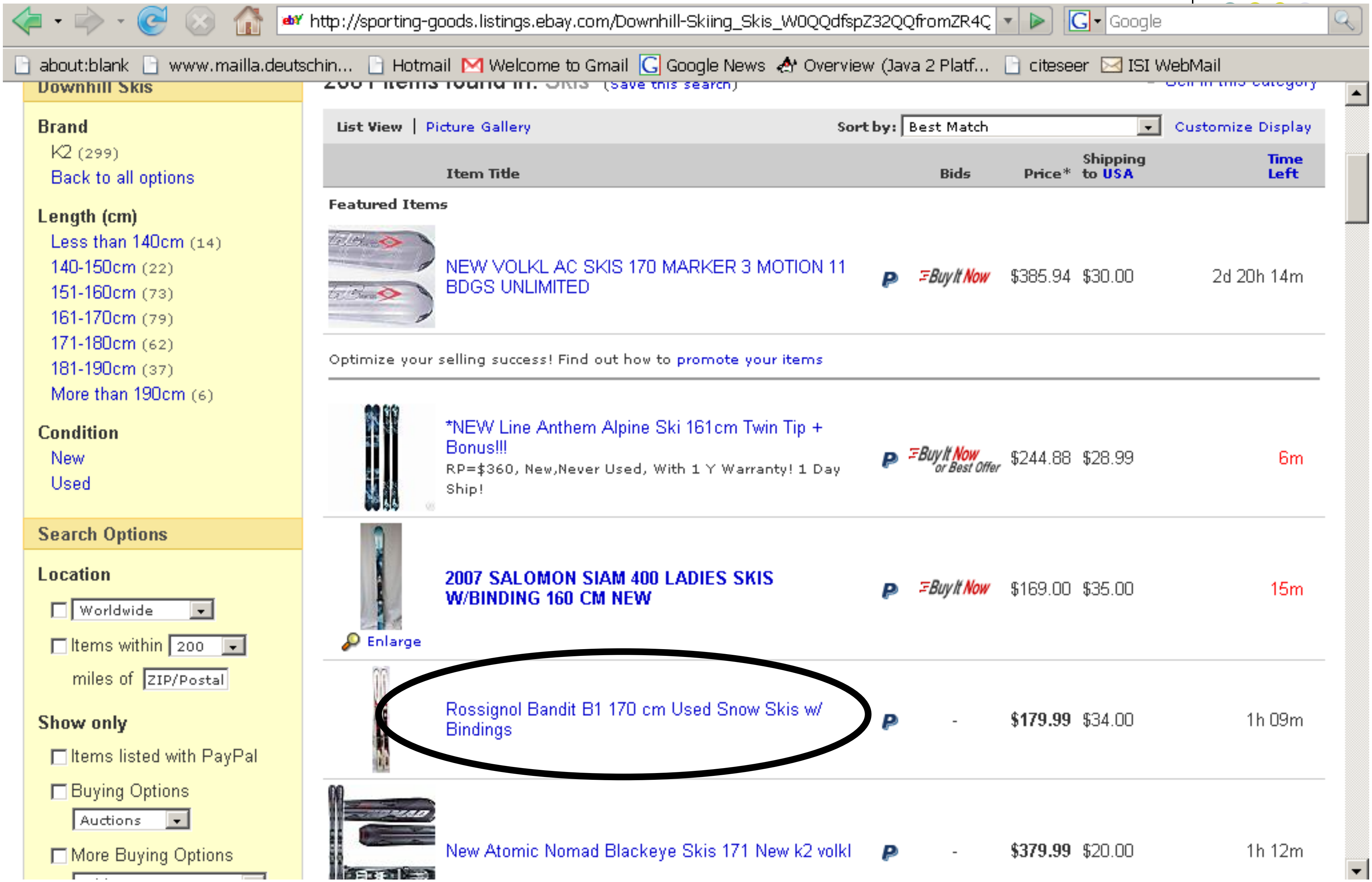
Classified ads, Auction listings, Etc.

Structured Sources

Semi-Structured Sources

Unstructured, Ungrammatical Sources

Unstructured, Ungrammatical Data: “Posts”



http://sporting-goods.listings.ebay.com/Downhill-Skiing_Skis_W0QQdfspZ32QQfromZR4C

Downhill Skis

(save this search)

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 200 miles of ZIP/Postal

Show only

□ Items listed with PayPal

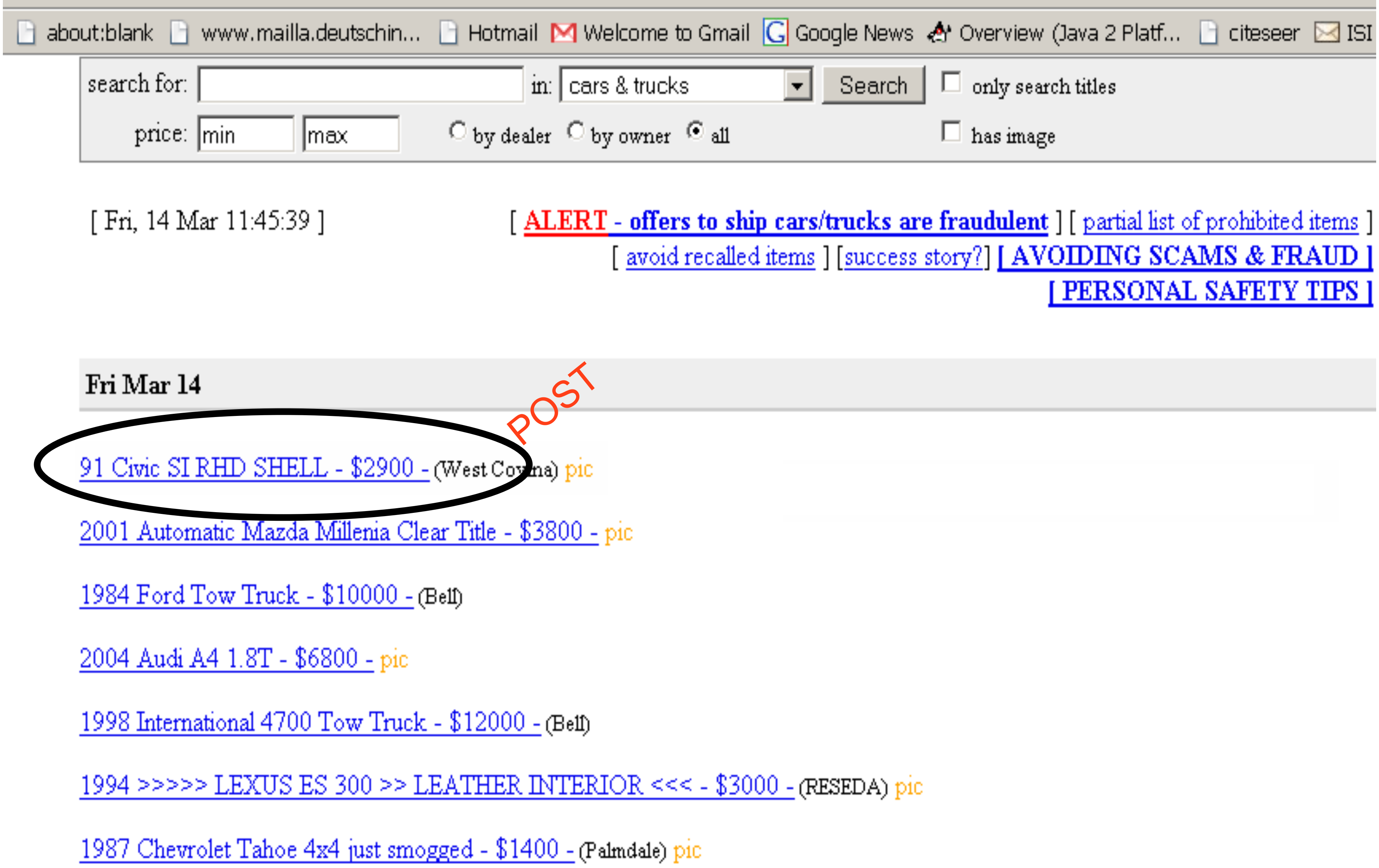
□ Buying Options
Auctions

□ More Buying Options

List View | Picture Gallery   Sort by: Best Match   Customize Display

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

Unstructured, Ungrammatical Data: “Posts”



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

POST

91 Civic SI RHD SHELL - $2900 - (West Covina) pic

2001 Automatic Mazda Millenia Clear Title - $3800 - pic

1984 Ford Tow Truck - $10000 - (Bell)

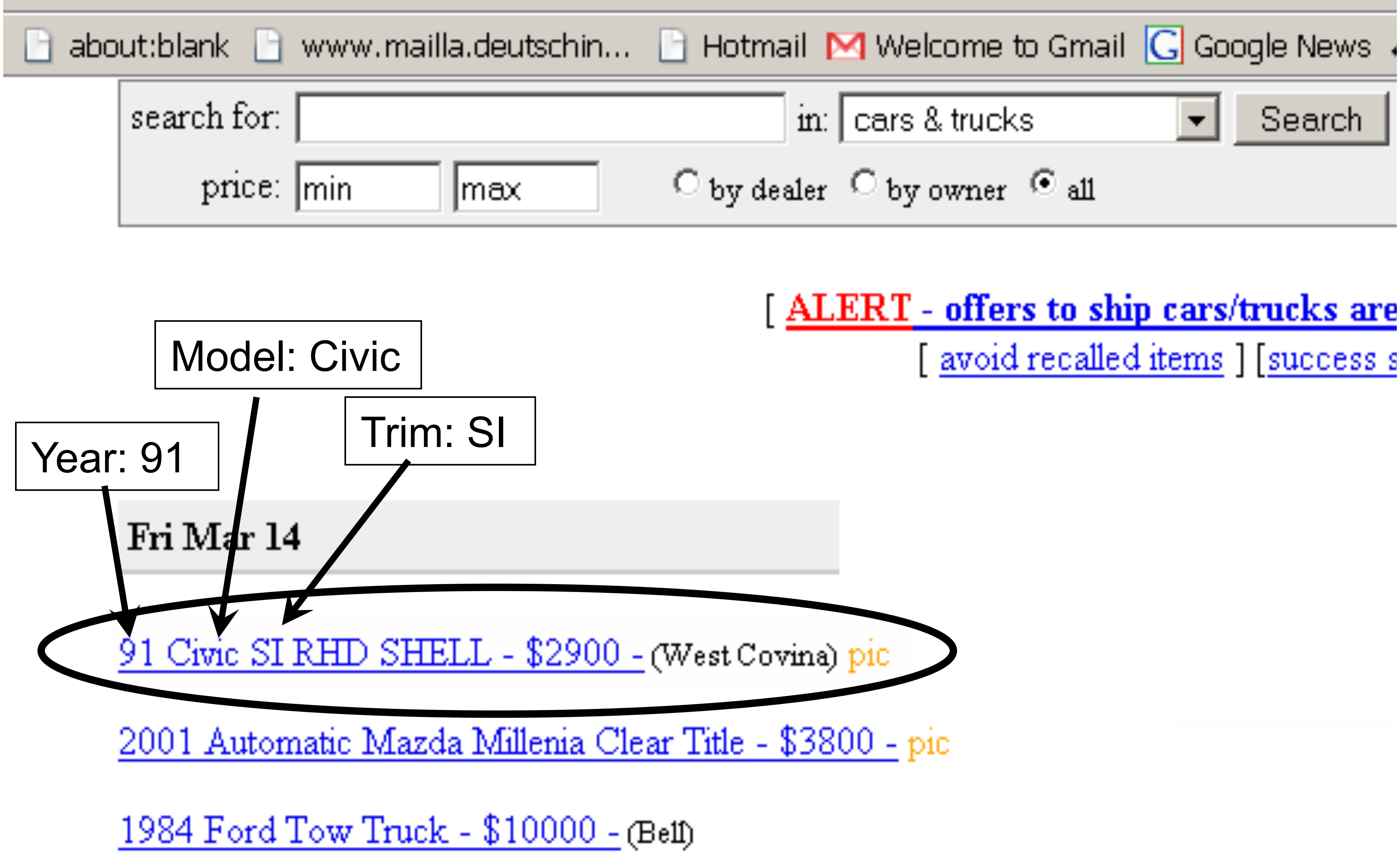
2004 Audi A4 1.8T - $6800 - pic

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

Query? …
Information Extraction!



search for: in: cars & trucks

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

[ALERT - offers to ship cars/trucks are

[avoid recalled items] [success s

Model: Civic

Trim: SI

Year: 91

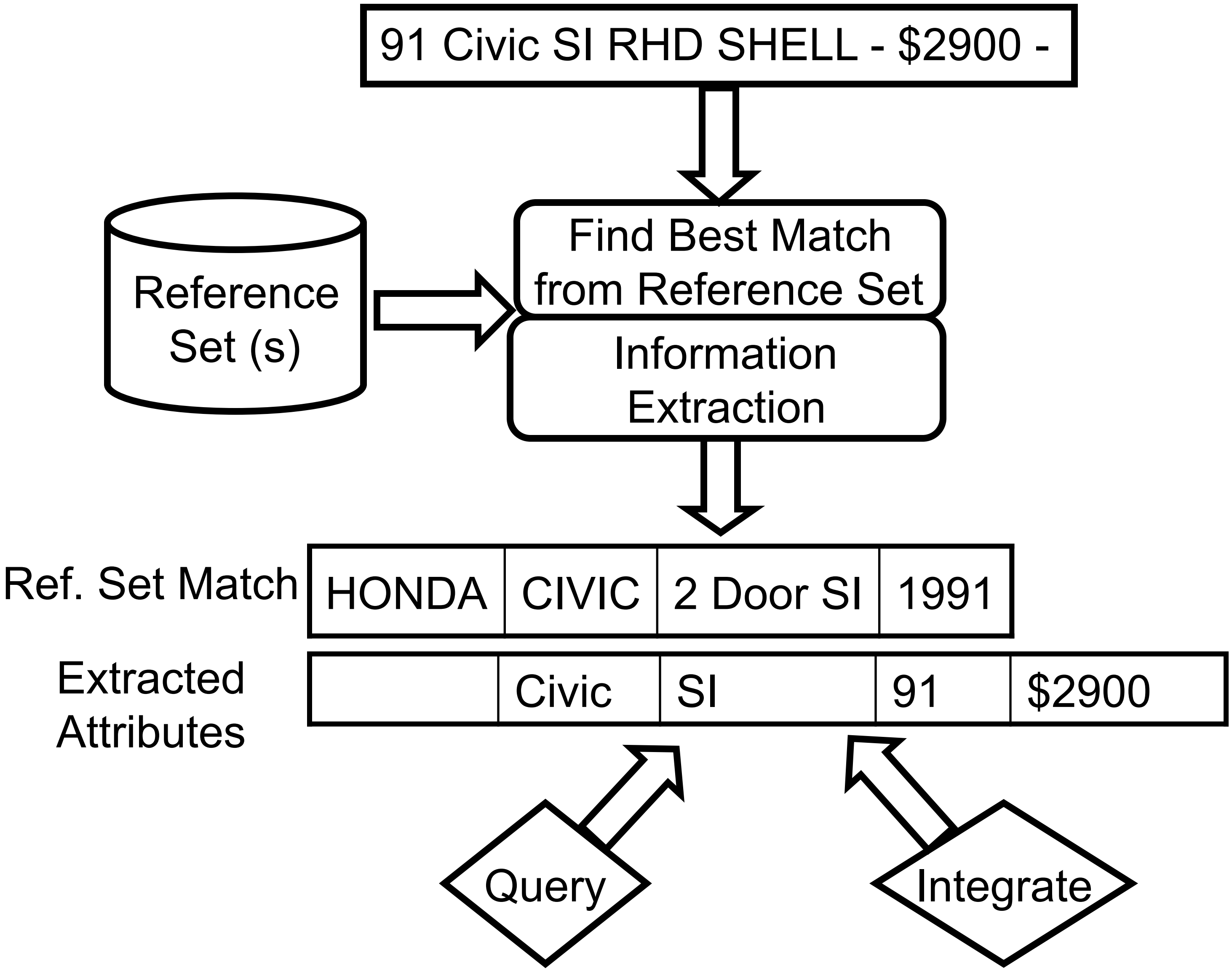
Fri Mar 14

91 Civic SI RHD SHELL - $2900 - (West Covina) pic

2001 Automatic Mazda Millenia Clear Title - $3800 - pic

1984 Ford Tow Truck - $10000 - (Bell)

Reference-Set Based Extraction/ Annotation



91 Civic SI RHD SHELL - $2900 -

Reference Set (s)

Find Best Match from Reference Set

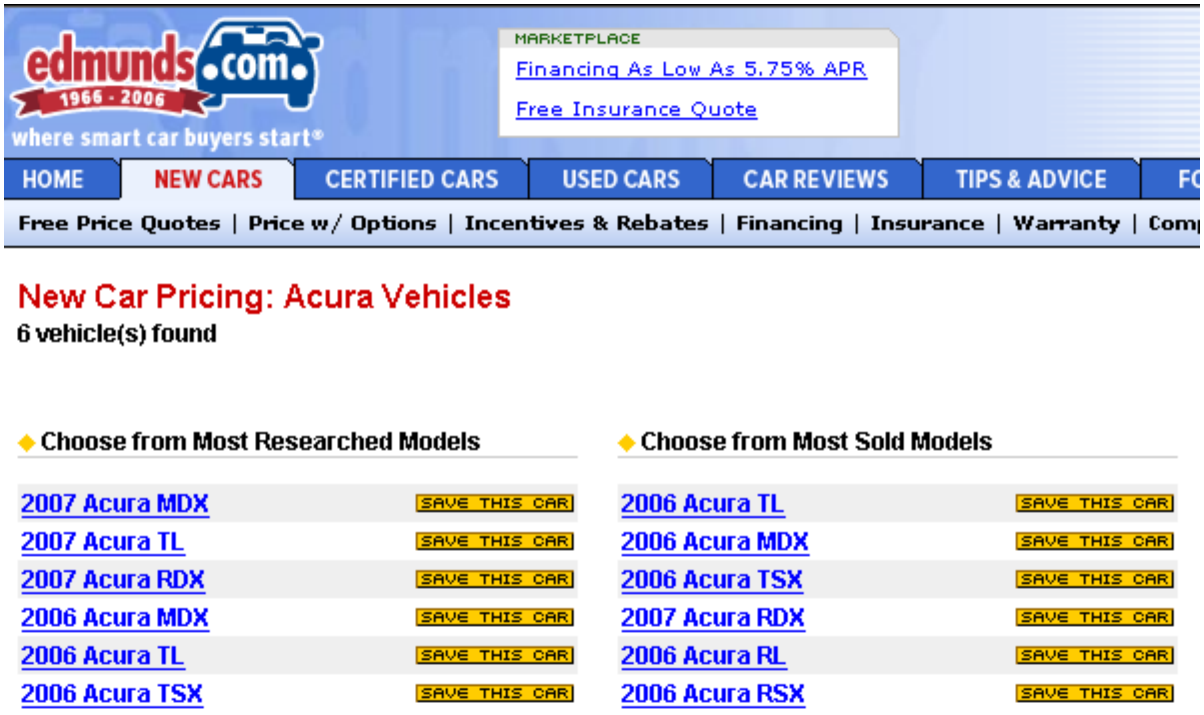
Information Extraction

Ref. Set Match	HONDA	CIVIC	2 Door SI	1991	
Extracted Attributes		Civic	SI	91	$2900

M+K, JAIR, 2008,
M+K, IJDAR, 2007,
M+K, IJCAI, 2005

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
SAVE THIS CAR

New Car Pricing: Acura Vehicles

6 vehicle(s) found

◆ Choose from Most Researched Models

2007 Acura MDX
2007 Acura TL
2007 Acura RDX
2006 Acura MDX
2006 Acura TL
2006 Acura TSX

◆ Choose from Most Sold Models

2006 Acura TL
2006 Acura MDX
2006 Acura TSX
2007 Acura RDX
2006 Acura RL
2006 Acura RSX

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

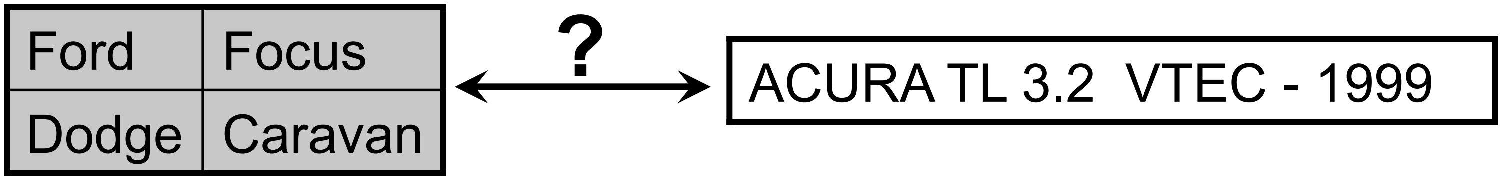
Construction of Reference Sets

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

HP Pavillion DV2000 laptop

Gateway ML6230, Intel Cel …

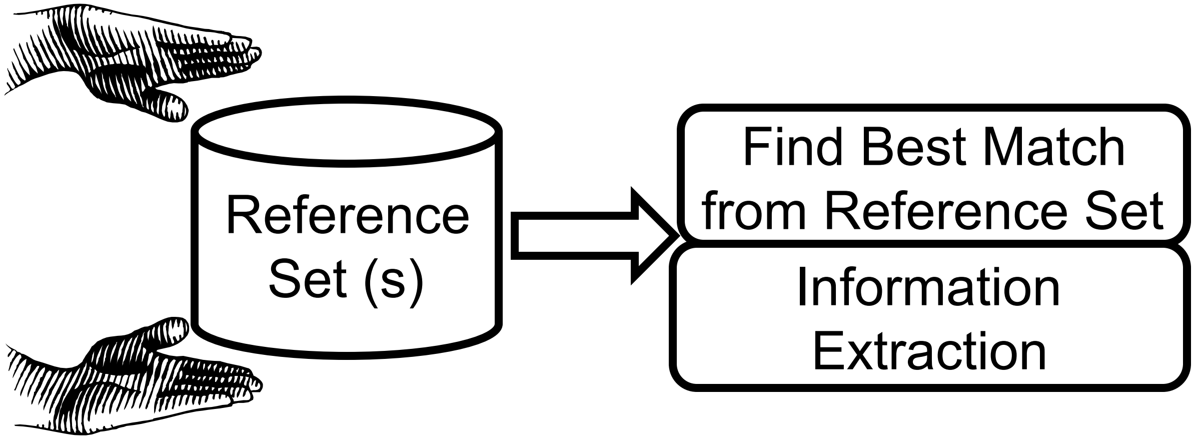
- What about coverage?



Ford	Focus
Dodge	Caravan

?

ACURA TL 3.2 VTEC - 1999



Reference Set (s)

Find Best Match from Reference Set

Information Extraction

Construction of Reference Sets

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

HP Pavillion DV2000 laptop

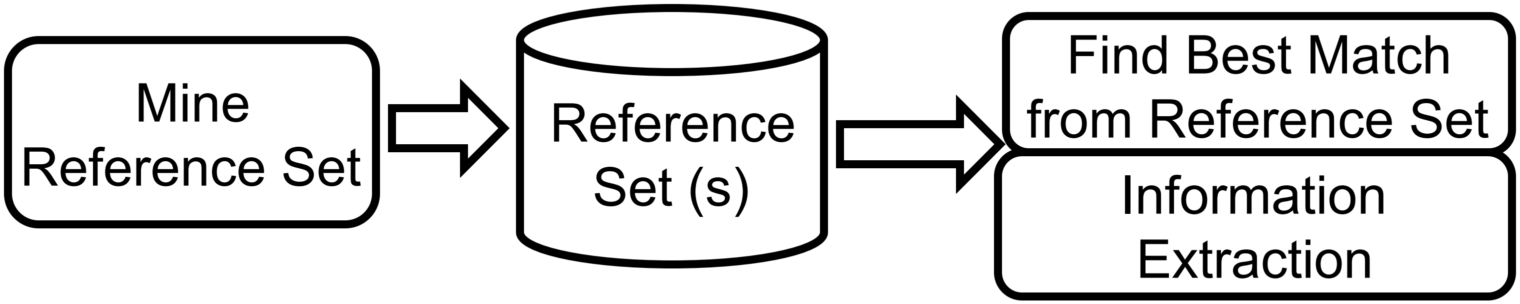
Gateway ML6230, Intel Cel …

- 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

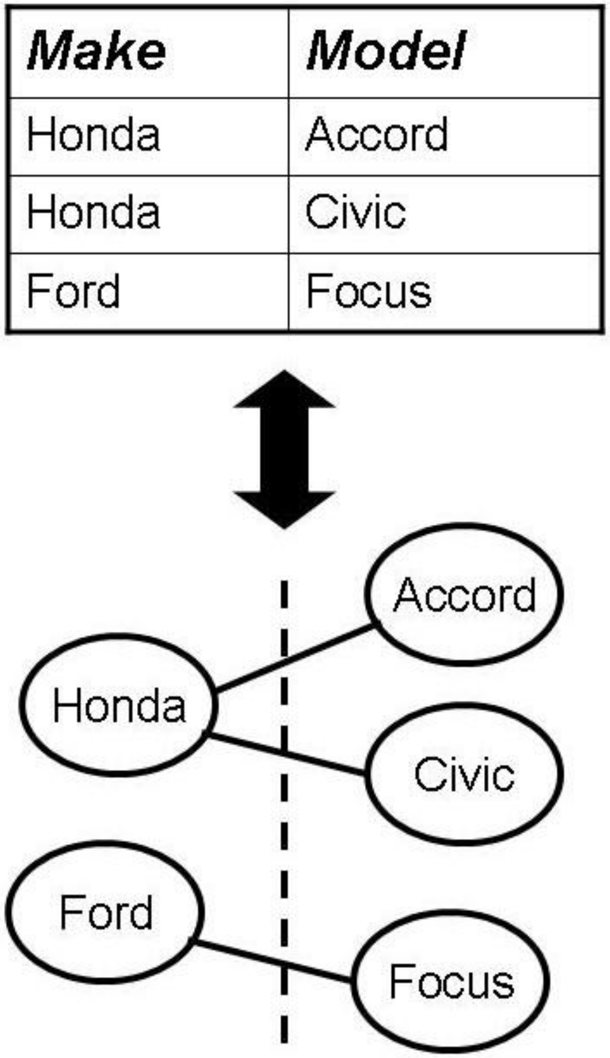
Seed-Based Reference Set Construction

- 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

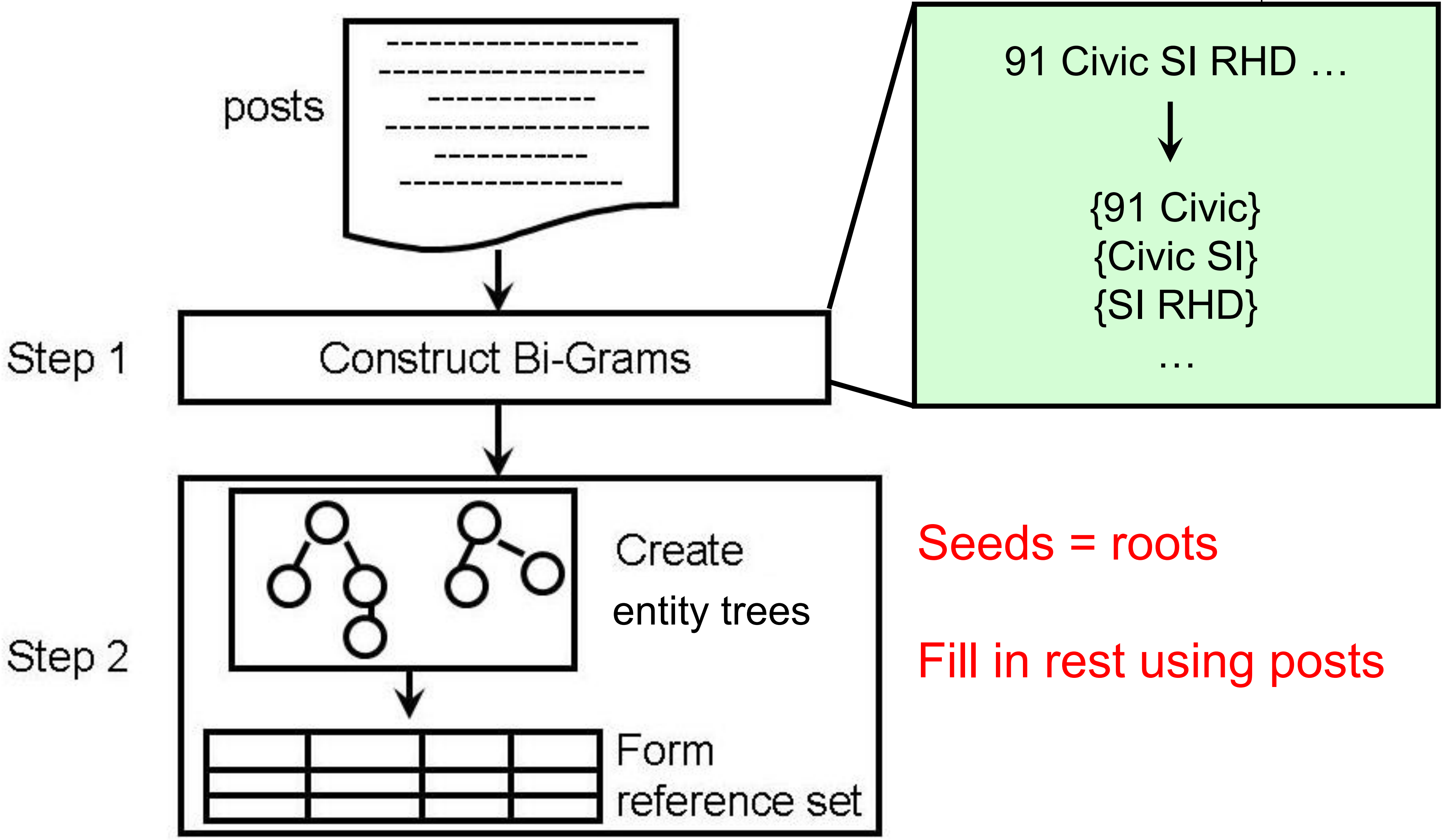
Forest of “Entity Trees”

Reference Set Construction

=

Constructing this forest

Entity Trees from Posts



posts

Step 1

Construct Bi-Grams

91 Civic SI RHD …

{91 Civic}
{Civic SI}
{SI RHD}
…

Step 2

Create entity trees

Form reference set

Seeds = roots

Fill in rest using posts

Constructing Entity Trees

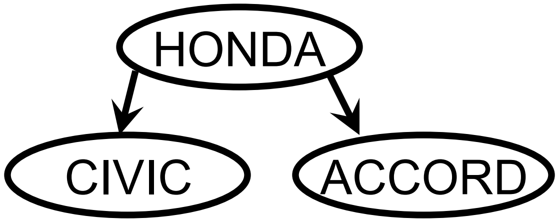
- Sanderson & Croft heuristic
  - x SUBSUMES y *IF* $P(x|y) \geq 0.75$ & $P(y|x) \leq P(x|y)$
- Merge heuristic
  - MERGE(x,y) *IF* x SUBSUMES y & $P(y|x) \geq 0.75$

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

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

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

- Construct hierarchies, then flatten



HONDA	CIVIC
HONDA	ACCORD

General Tokens

- {a, y}, {b, y}, {c, y} → y is “general token”
  - 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”

Experiments & Results

- Goal
    - Build reference sets for information extraction
    - Extraction = task to compare reference sets
        - Poor coverage → poor recall
        - Noise → bad extractions → worse results
- Compare extraction (M+K, IJDAR, 2007)
    - 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

	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

Related Work

- Unsupervised Information Extraction
  - Finds relations, uses patterns
- Ontology creation
  - NLP based
  - Single, large concept hierarchies

Conclusions / Future Work

- Seed-based reference set construction
  - Seeds provide roots
    - More static foundation
    - Cleaner entity trees
  - Posts provide rest of entity-trees
    - Capture dynamic data
    - Better Coverage

- Future directions
  - More background knowledge
    - Google sets? Partial reference sets?
  - Siblings in entity trees
    - Roles? Identify? Combine?

Questions?