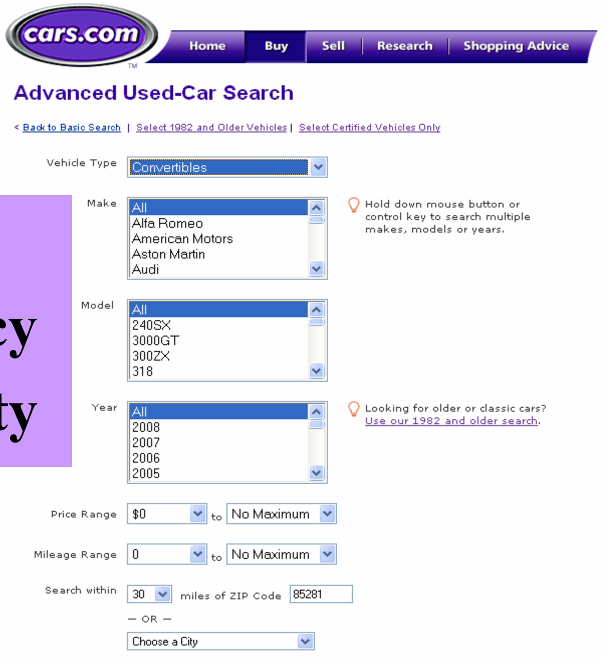
Data Provenance for Query Result Explanation

Yi Chen
Arizona State University

Joint work with Subbarao Kambhampati, Garrett Wolf, Hemal Khatri, Bhaumik Chokshi, Jianchun Fan, Ullas Nambiar

Challenges in Querying Web: Incomplete Data

- Incomplete Entry
- Extraction Inaccuracy
- Schema Heterogeneity



Advanced Used-Car Search

< Back to Basic Search | Select 1982 and Older Vehicles | Select Certified Vehicles Only

Vehicle Type: Convertibles

Make: All, Alfa Romeo, American Motors, Aston Martin, Audi

Hold down mouse button or control key to search multiple makes, models or years.

Model: All, 240SX, 3000GT, 300ZX, 318

Year: All, 2008, 2007, 2006, 2005

Looking for older or classic cars? Use our 1982 and older search.

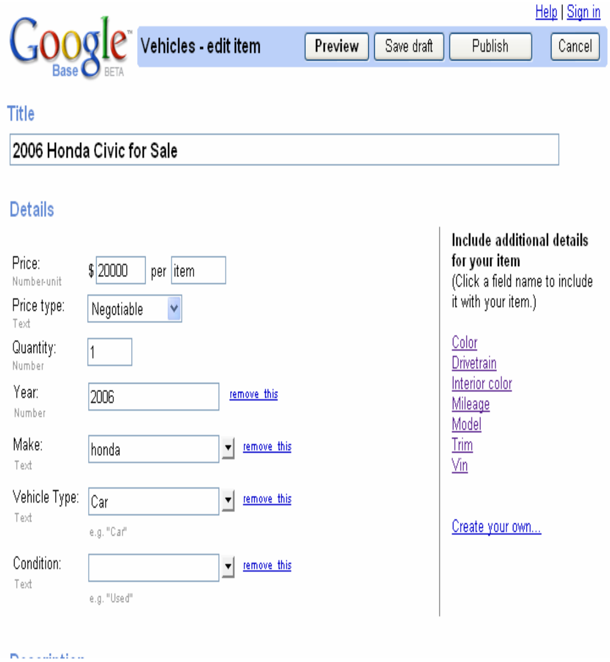
Price Range: $0 to No Maximum

Mileage Range: 0 to No Maximum

Search within: 30 miles of ZIP Code 85281

— OR —

Choose a City



Title

2006 Honda Civic for Sale

Details

Price: (Number-unit) $ 20000 per item

Price type: (Text) Negotiable

Quantity: (Number) 1

Year: (Number) 2006 remove this

Make: (Text) honda remove this

Vehicle Type: (Text) Car remove this
e.g. "Car"

Condition: (Text) remove this
e.g. "Used"

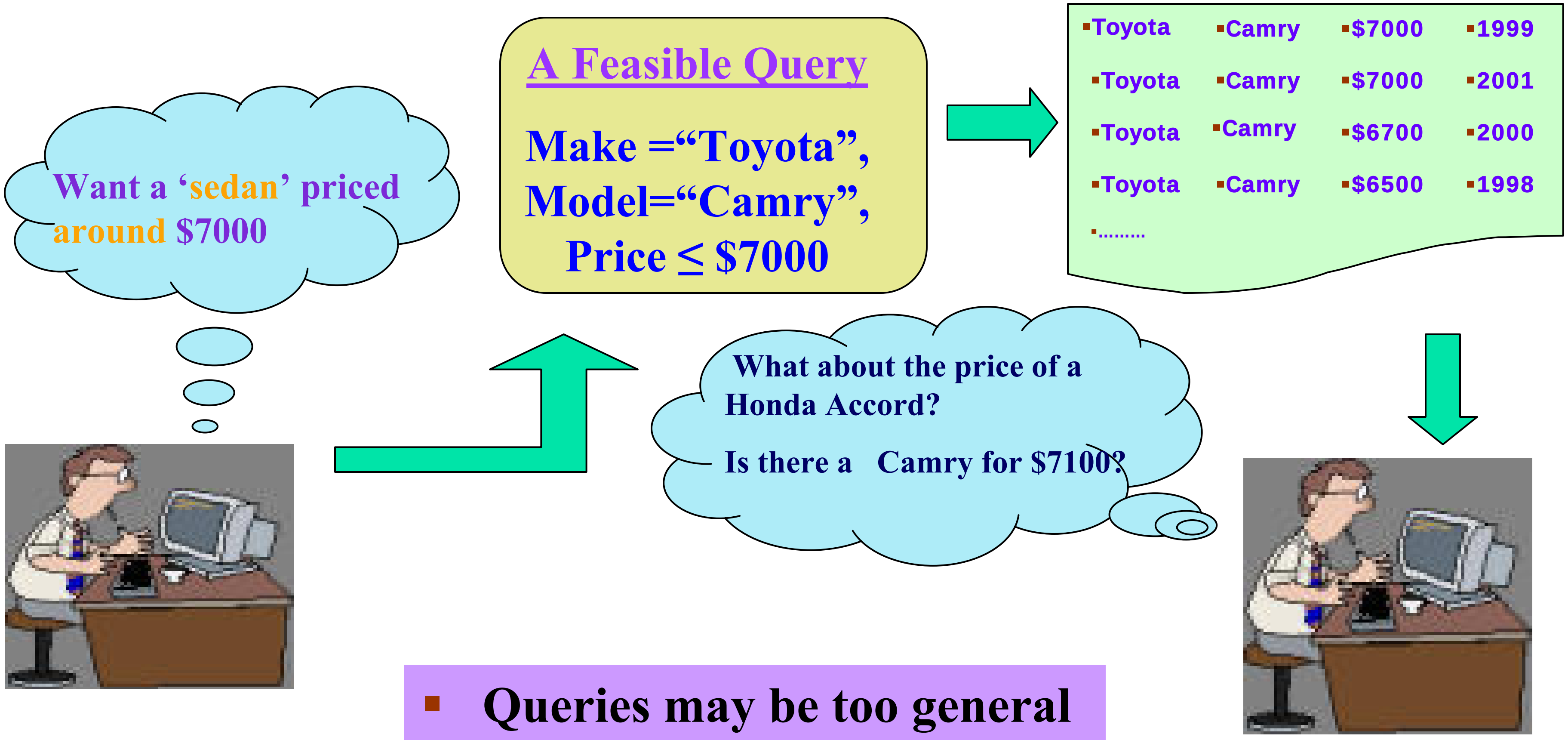
Include additional details for your item
(Click a field name to include it with your item.)

Color
Drivetrain
Interior color
Mileage
Model
Trim
Vin

Create your own...

Website	# of attributes	# of tuples	incomplete tuples	body style	engine
autotrader.com	13	25127	33.67%	3.6%	8.1%
carsdirect.com	14	32564	98.74%	55.7%	55.8%

Challenges in Querying Web: Imprecise Queries



Want a 'sedan' priced around $7000

A Feasible Query

Make ="Toyota", Model="Camry", Price ≤ $7000

Toyota	Camry	$7000	1999
Toyota	Camry	$7000	2001
Toyota	Camry	$6700	2000
Toyota	Camry	$6500	1998
………			

What about the price of a Honda Accord?

Is there a Camry for $7100?

- Queries may be too general
- Queries may be too specific

QUIC System [VLDB 07, ICDE 07, CIDR 07]

Imprecise Queries

Incomplete Data

Query results are no longer exactly satisfying user queries.
Rank them based on Expected Relevance Ranking

$$\mathcal{ER}(\hat{t}|Q,U,D) = \sum_{t \in C(\hat{t})} \mathcal{R}(t|Q,U)\mathcal{P}(t|\hat{t},D)$$

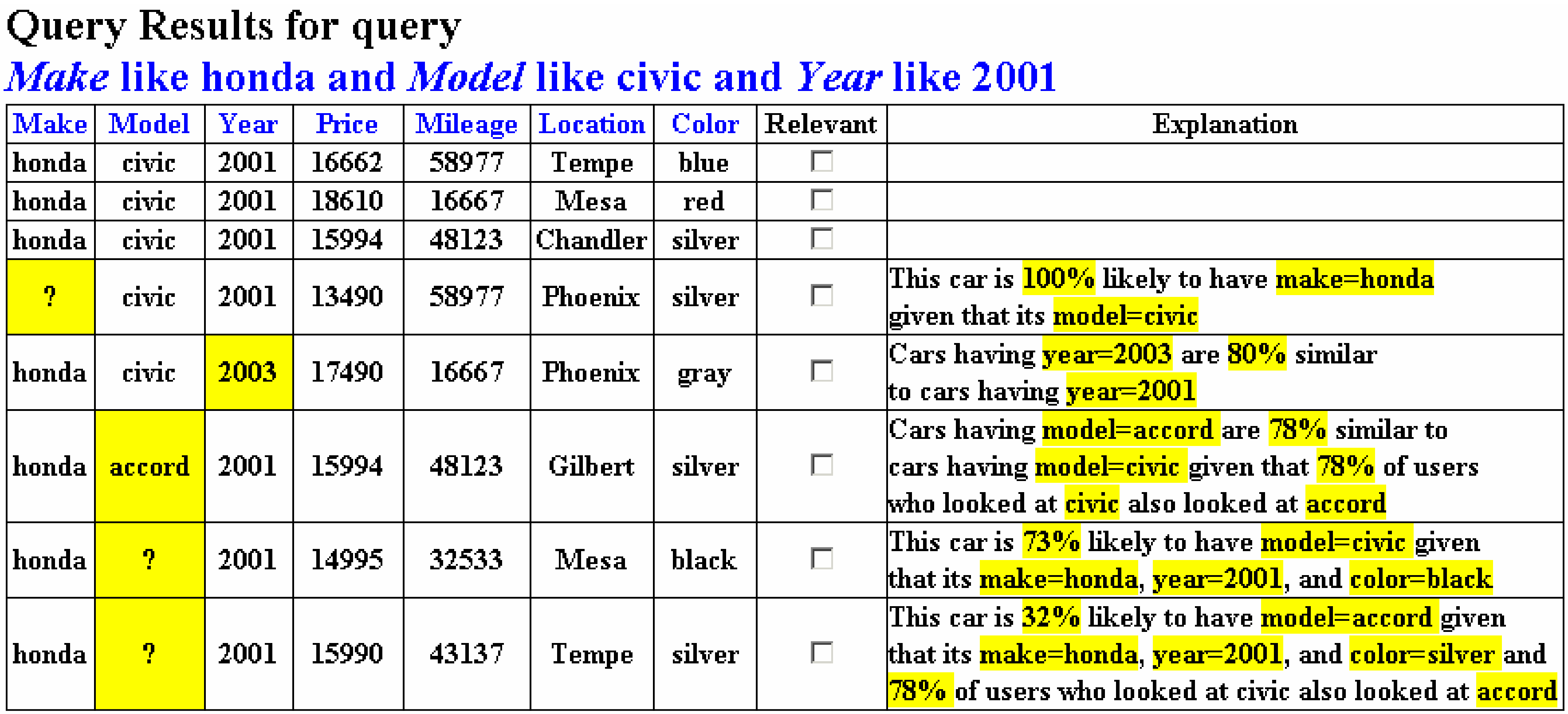
Relevance Function

Density Function

Automated & Non-intrusive assessment of
Relevance and Density functions

Query rewriting to retrieve similar/incomplete tuples
in the order of their relevance

Challenge: How Should User Believe it ?



Query Results for query

Make like honda and *Model* like civic and *Year* like 2001

Make	Model	Year	Price	Mileage	Location	Color	Relevant	Explanation
honda	civic	2001	16662	58977	Tempe	blue	□	
honda	civic	2001	18610	16667	Mesa	red	□	
honda	civic	2001	15994	48123	Chandler	silver	□	
?	civic	2001	13490	58977	Phoenix	silver	□	This car is 100% likely to have make=honda given that its model=civic
honda	civic	2003	17490	16667	Phoenix	gray	□	Cars having year=2003 are 80% similar to cars having year=2001
honda	accord	2001	15994	48123	Gilbert	silver	□	Cars having model=accord are 78% similar to cars having model=civic given that 78% of users who looked at civic also looked at accord
honda	?	2001	14995	32533	Mesa	black	□	This car is 73% likely to have model=civic given that its make=honda, year=2001, and color=black
honda	?	2001	15990	43137	Tempe	silver	□	This car is 32% likely to have model=accord given that its make=honda, year=2001, and color=silver and 78% of users who looked at civic also looked at accord

Discussions

- A recommendation system: A query processor that provides results beyond the ones that exactly satisfy a user query
- Explanation is important to gain user's trust
- We need to record data provenance about the reasoning of query processing.
  - What is provenance information?

    How the answers are derived (not just computed)?

    Need to know the reasoning and evidence (knowledge base, statistical information, etc)
  - How should we obtain the provenance information?

    Proactively recording
  - How should we display the provenance information to users?

    display (a) in a hierarchical form of different granularities; (b) a chain of provenance step by step
  - Can we query the provenance itself to inspect the system? How?

Questions?