

# NADEEF/ER: INTERACTIVE ENTITY RESOLUTION

Ahmed Elmagarmid,<sup>1</sup> Ihab Ilyas,<sup>2</sup> Mourad Ouzzani,<sup>1</sup> Jorge Quiané,<sup>1</sup> Nan Tang,<sup>1</sup> Si Yin<sup>1</sup>




<sup>1</sup> Qatar Computing Research Institute

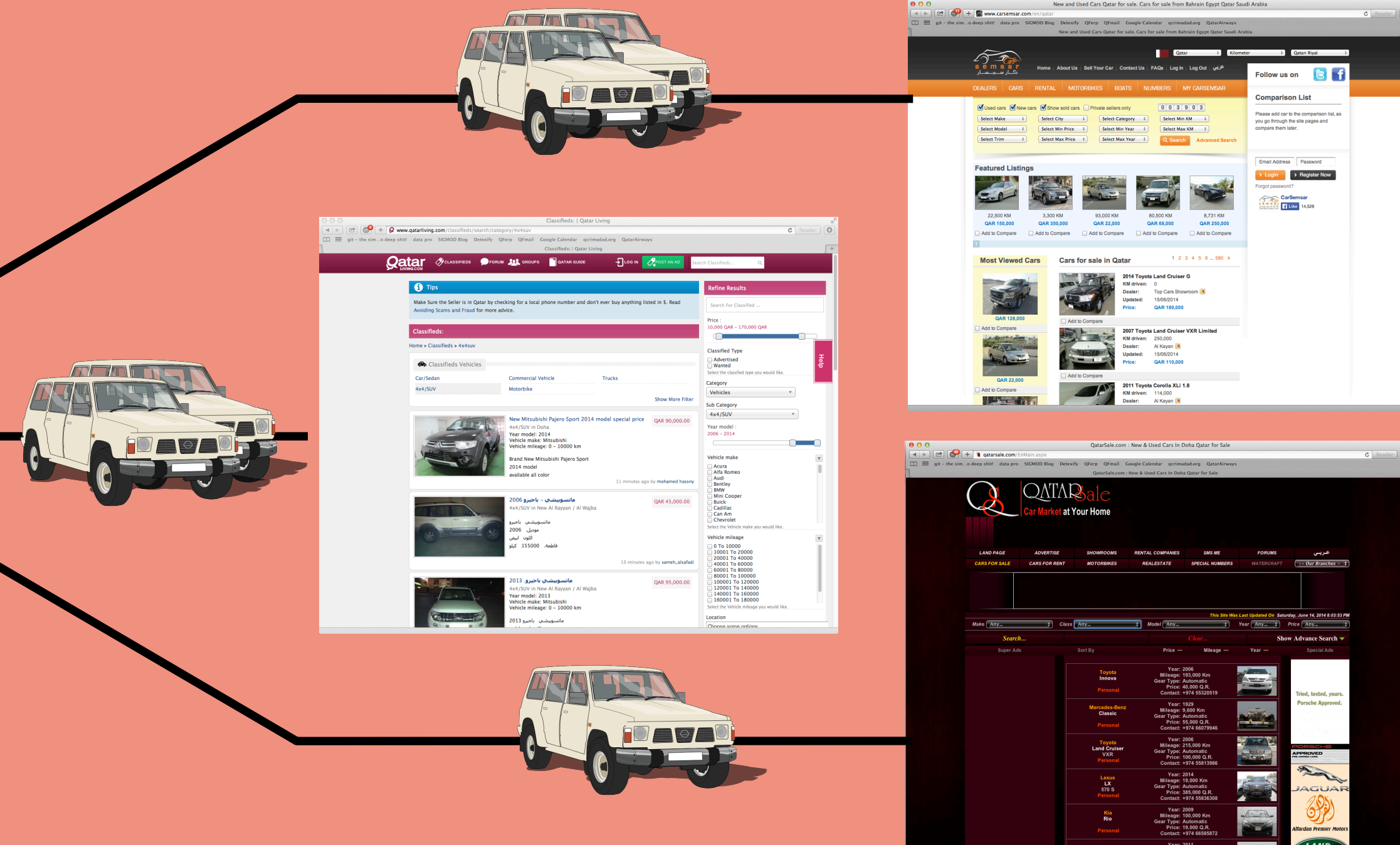
<sup>2</sup> University of Waterloo



I want a patrol nissan

### QCars

	<b>Model:</b> Patrol <b>Brand:</b> Nissan <b>Mileage:</b> 10,000km <b>Year:</b> 2011 <b>Price:</b> 220,000 QAR
	<b>Model:</b> patrol <b>Brand:</b> nissa <b>Mileage:</b> 10,000km <b>Year:</b> 2011 <b>Price:</b> 200,000 QAR
	<b>Model:</b> patrol <b>Brand:</b> nissan <b>Mileage:</b> 10,500km <b>Year:</b> 2011 <b>Price:</b> 215,000 QAR



## Limits of existing ER systems:

- (1) Hard to customize and specify ad-hoc ER rules
- (2) Lack of interactivity to guide developers in defining ER rules
- (3) Scalability problem in terms of data size and streaming data

## NADEEF/ER

### Easy specification

Rich graphical interface for rule specification

EditorCode

tb_google		tb_amazon
id		id
name		title
description		description
manufacturer		manufacturer
price		price
tid		tid

#	Column A	Operation	Column B	Comparator	Value	Action
1	tb_google.name	Levenshtein	tb_amazon.title	>	0.8	Del
2	tb_google.description	Equals	tb_amazon.description	=		Del
3	tb_google.tid	Edit distance	tb_amazon.tid	=	1	Del

GenerateSave

### Extensibility

Customizable ER process

EditorCode

Horizontal Scope / Vertical Scope / Block / Iterator / Detect

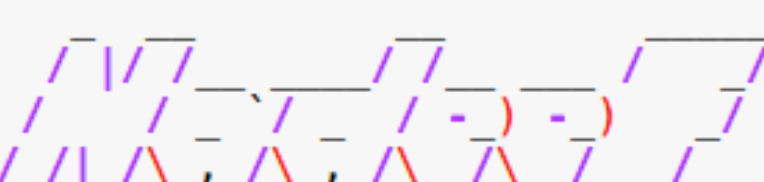
```
1- /** Code Generated by NADEEF */
2- import qa.qcri.nadeef.core.datamodel.*;
3- import qa.qcri.nadeef.core.datamodel.*;
4- import java.util.*;
5-
6- public class ER1 extends PairTupleRule {
7-     @Override
8-     public void initialize(String id, List<String> tableNames) {
9-         super.initialize(id, tableNames);
10-    }
11-
12-    @Override
13-    public Collection<Table> horizontalScope(Collection<Table> table) {
14-        return table;
15-    }
16-
17-    @Override
18-    public Collection<Table> verticalScope(Collection<Table> table) {
19-        return table;
20-    }
21-
22-    @Override
23-    public Collection<Table> block(Collection<Table> table) {
24-        return table;
25-    }
26-}
```

A+ A- Verify

### Scalable Ad-hoc Analytics

Spark, Python, and R support

File Edit View Insert Cell Kernel Help



Interactive Analytics for Violation

Type 'vt' to access violation graph  
Type 'sc' to access SparkContext  
Made by Qatar Computing Research Institute, 2014 (<http://da.qcri.org>).  
Enjoy!

In [2]: sc

Out[2]: <pyspark.context.SparkContext at 0x7f8f24135790>

### Interactivity

Better understanding of duplicates

NADEEF

Refresh New Data Source

Rules Discover Detect More

Details

Rule Type ER  
Table Name tb\_google  
Code L5tb\_google.name, tb\_amazon...

Progress No job is running

Overview Rule Distribution Sky Graph Triple Rank

Source Violation Audit


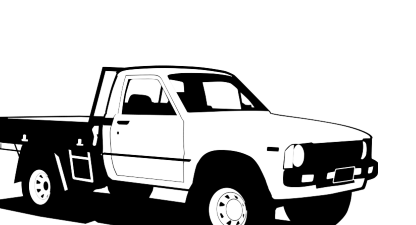
Showing 1 to 3 of 9 entries

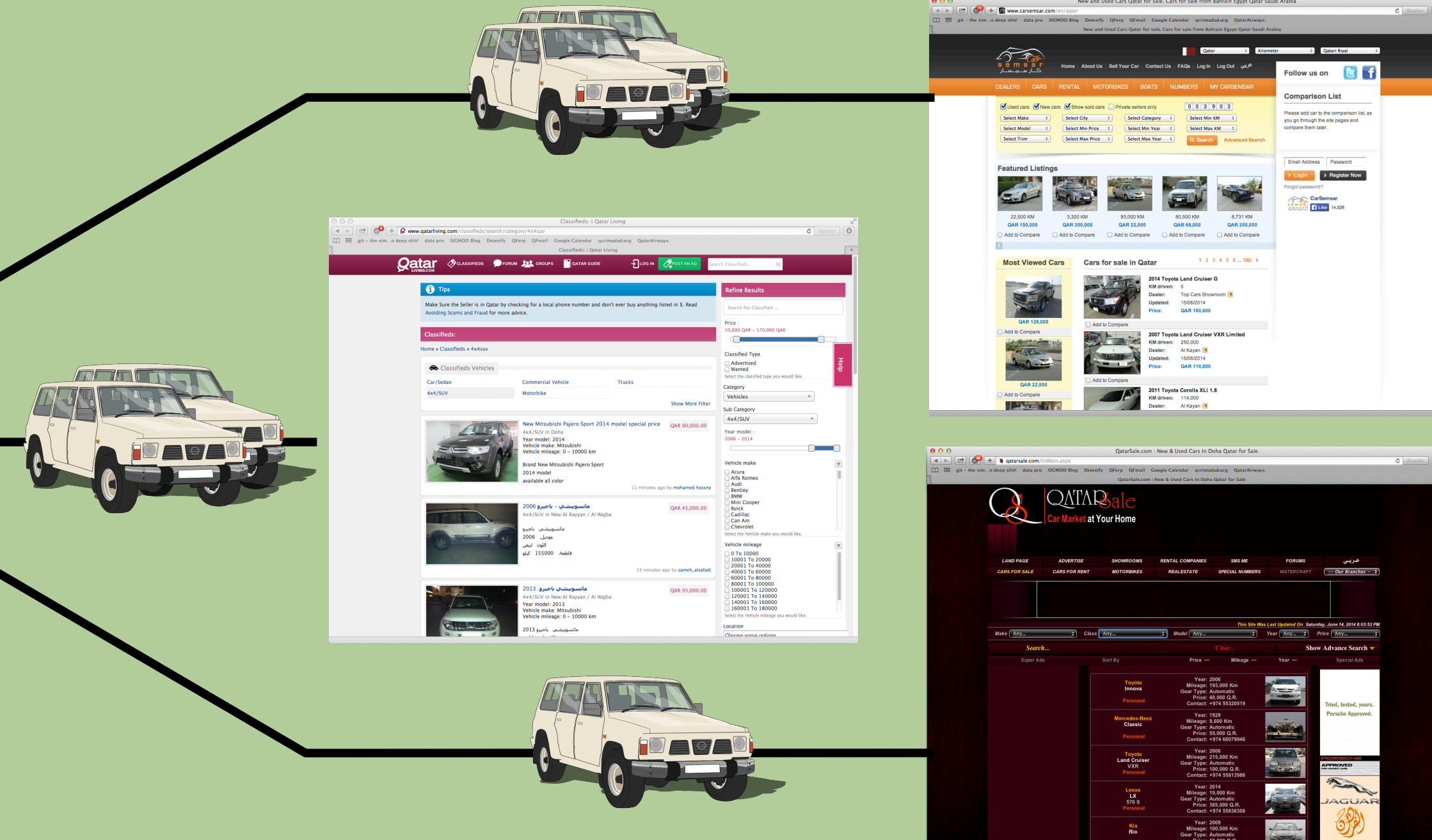
Table D is duplication rate



I want a patrol nissan

### QCars

	<b>Model:</b> patrol <b>Brand:</b> nissan <b>Mileage:</b> 10,000km <b>Year:</b> 2011 <b>Price:</b> 200,000 QAR
	<b>Model:</b> patrol <b>Brand:</b> nissan <b>Mileage:</b> 100,000km <b>Year:</b> 2004 <b>Price:</b> 57,000 QAR



## NADEEF/ER Advantages:

- (1) Generality
- (2) Ease-of-use
- (3) Data repair integration
- (4) Scalability