Ziawasch Abedjan, John Morcos, Michael Gubanov, Ihab Ilyas, Michael Stonebraker, Paolo Papotti, Mourad Ouzzani

Age of Data Integration

DataXFormer

Leveraging the Web for Semantic Transformations

01/04/2015

CSAIL MIT, University of Waterloo, QCRI

G GENERAL AUDIENCES
ALL AGES ADMITTED

DataXFormer: Leveraging the Web for Semantic Transformations

Ziawasch Abedjan, John Morcos, Michael Gubanov, Ihab F. Ilyas, Mike Stonebraker, Paolo Papotti, Mourad Ouzzani

Integration of multiple sources

KAYAK HOTELS FLIGHTS CARS PACKAGES DEALS Login

Flights

Round-trip One-way Multi-city

Boston (BOS) | San Francisco (SFO) | Sun 1/4 | Wed 1/7 | 1 adult, Economy | Search

Include nearby | Include nearby | Show flexible dates

U·S AIRWAYS

$561
JS Airways
Select

US Airways

5:50a	BOS	→	1:15p	SFO	10h 25m	1 stop (ORD)
8:55a	SFO	→	10:25p	BOS	10h 30m	1 stop (CLT)

Show details | Only 1 seat left at this price | Eco

American Airlines operates flight 108

American Airlines

$563
erican Airlines
Select

American Airlines

5:50a	BOS	→	11:25a	SFO	8h 35m	1 stop (ORD)
12:30a	SFO	→	10:41a	BOS	7h 11m	1 stop (DFW)

Show details | Only 2 seats left at this price | Eco

Different value representations

American Airlines

Departure	BOS
Destination	SFO
Cabin	Choice
Time	5:50 am
Price	563 $

U·S AIRWAYS

Boston , MA (BOS)
San Fransisco, CA (SFO)
Coach
5:50 AM
561 $

Lufthansa
Nonstop you

Boston – Logan International
San Fransisco, CA (SFO)
Economy Restricted
14:26
613 €

KAYAK

Departure	Boston (BOS)
Destination	San Fransisco (SFO)
Cabin	Economy
Time	5:50a
…	563 $

Boston (BOS)
San Fransisco (SFO)
Economy
5:50a
561 $

Boston (BOS)
San Fransisco (SFO)
Economy
02:26p
731 $

Data Transformation Tasks

Airport code ←→ City

Airport code	City
BOS	Boston
JFK	New York
ORD	Chicago
BER	Berlin
CDG	Paris

- date format transformations
  - MM/DD/YYYY → DD/MM/YY
- currency conversion
  - 1 USD → 0.7? EUR

- model → brand
  - Iphone 6 → Apple Inc.
- ISBN→ title
  - 0-553-57340-3 → "A Game of Thrones"
- unit conversion
  - 1 Mi → 1.6 km
- long/lat → location
- language translation
- …

Problem Statement: Automatically discover transformations!

Given

airport	City
BER	?
JFK	?
ORD	?
HBE	?
IST	?
FRA	?
BOS	?
DFW	?
..	…

Find

airport	City
BER	Berlin
JFK	New York
ORD	Chicago
HBE	Alexandria
IST	Istanbul
FRA	Frankfurt
BOS	Boston
DFW	Dallas
..	…

Syntactic Transformations

US date	EU date
11/01/2014	01.11.2014
11/02/2014	02.11.2014
10/30/2014	30.10.2014
11/05/2014	05.11.2014
11/04/2014	04.11.2014

Liter	Gallon
1	0.26
5	1.04
100	26.42
34	8.98
6	1.58

GB	MB
1	1,024
0.49	500
100	102,400
2	2,048
6	6,144

Name	Last name
Michael Stonebraker	Stonebraker
Michael Bay	Bay
Michael Brodie	Brodie
Michael Jordan	Jordan

Semantic Transformations

Name	Nickname
Michael	Mike
Samuel	Sam
Ziawasch	Zia
Rebeccea	Becca

ISBN	Title
0-553-57340-3	A Game of Thrones
0-553-80202-X	Universe in a Nutshell
0-671-62964-6	The Hitchhiker's Guide to the Galaxy
0-374-53355-7	Thinking Fast and Slow
0-875-84585-1	The Innovator’s Dilemma

Airport code	City
BOS	Boston
JFK	New York
ORD	Chicago
BER	Berlin
CDG	Paris

Model	Category
Iphone 6	Mobile Phone
MacBook Air	Notebook
Logitech mouse	Accessory
Nexus 5	Mobile Phone

Problem Statement

Given

airport	City
BER	Berlin
JFK	New York
ORD	Chicago
HBE	?
IST	?
FRA	?
BOS	?
DFW	?
..	…

Example transformations

Find

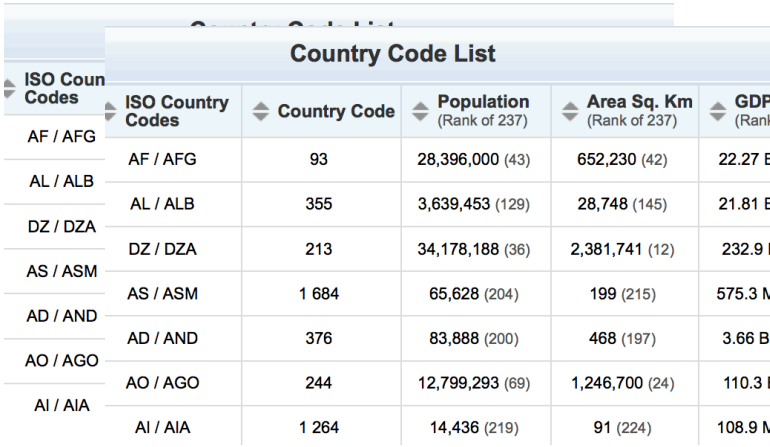
airport	City
BER	Berlin
JFK	New York
ORD	Chicago
HBE	Alexandria
IST	Istanbul
FRA	Frankfurt
BOS	Boston
DFW	Dallas
..	…

DataXFormer: The Web as general Repository

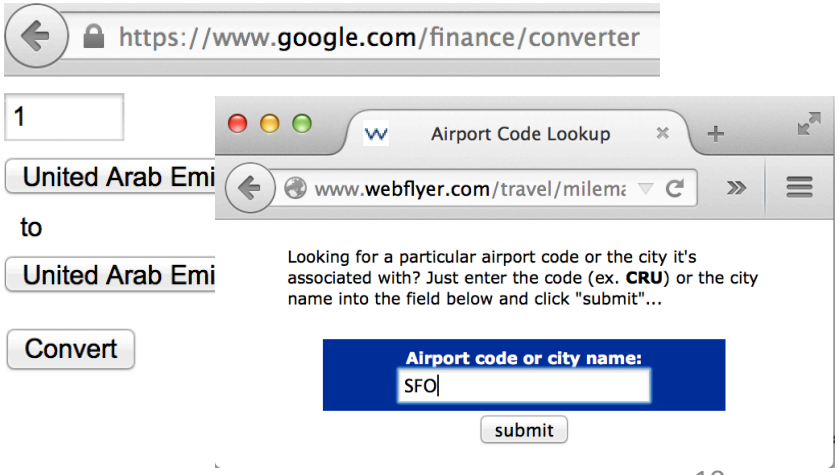
Given

airport	City
BER	Berlin
JFK	New York
ORD	Chicago
HBE	?
IST	?
FRA	?
BOS	?
DFW	?
..	…

Web Tables



Web Forms



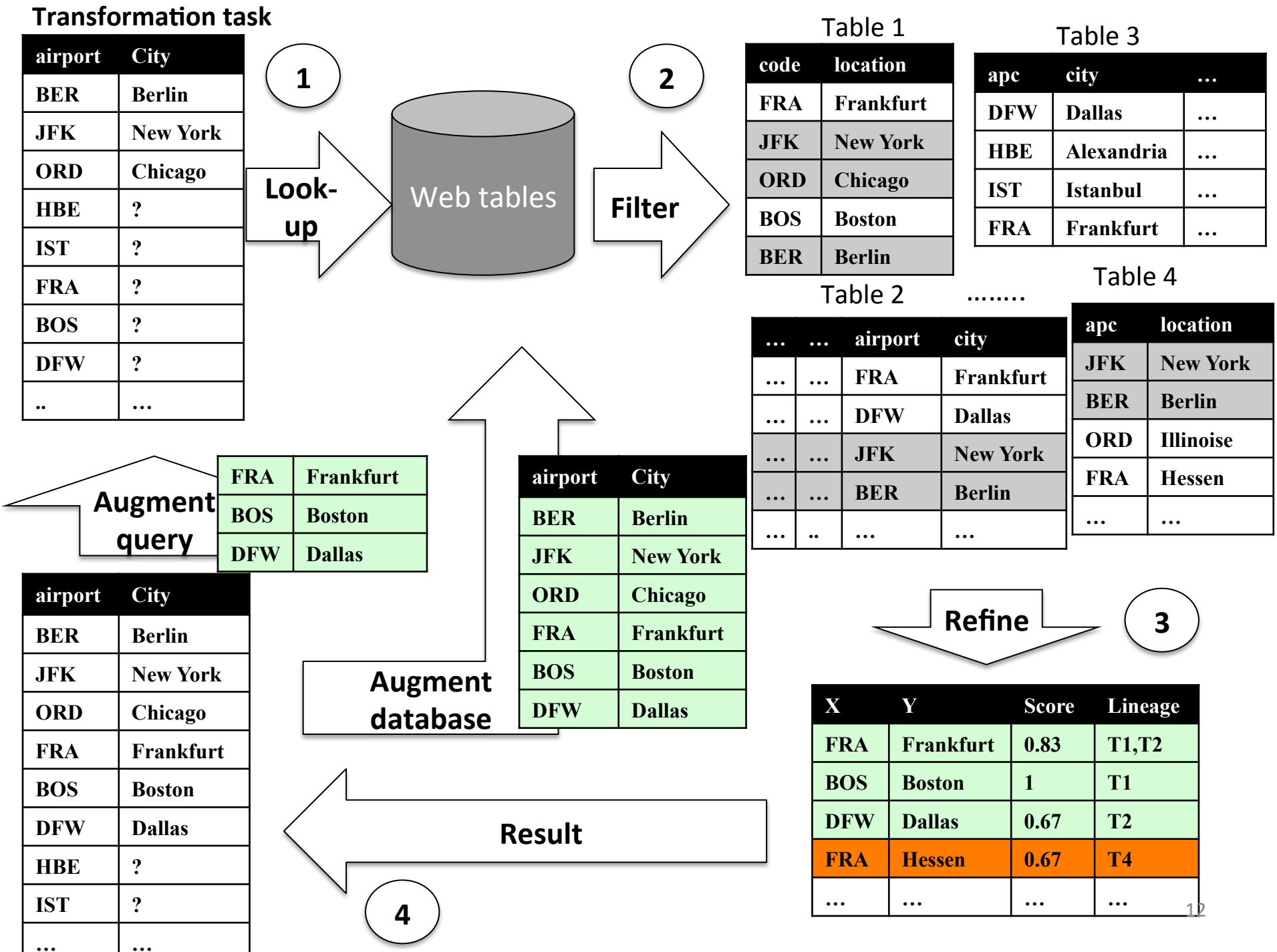
Web Tables

- Dataset
  - Dresden Web table Corpus
  - 120 Million tables
- Efficiently discovering transformation Examples:
  - Filter irrelevant tables
  - Overcome fragmentation
    - Average rowcount = 12
  - Dirty and Heterogeneous

Filter and Refine approach

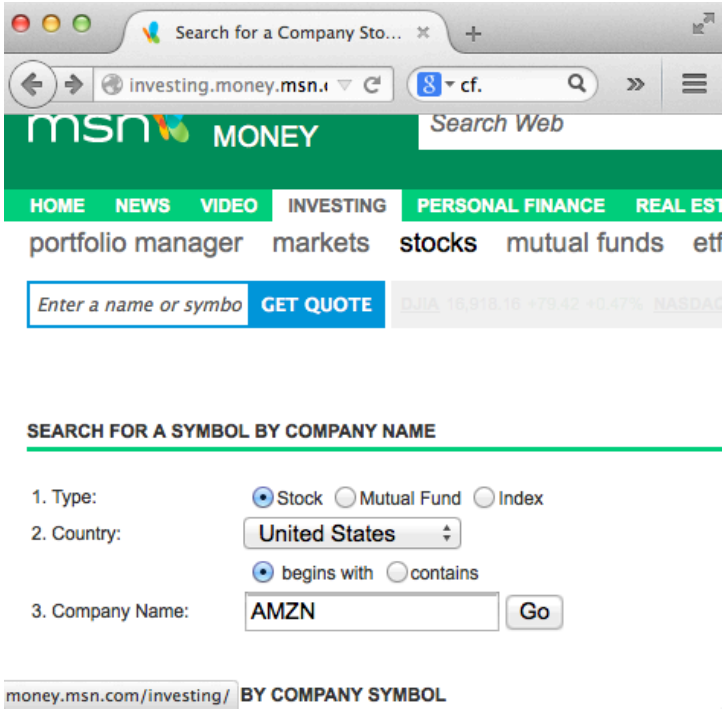
Multiple iterations and example augmentation

Rate transformations based on example hits and majority vote


Transformation task
airport City
BER Berlin
JFK New York
ORD Chicago
HBE ?
IST ?
FRA ?
BOS ?
DFW ?
.. …
1
Look-up
Web tables
2
Filter
Table 1
code location
FRA Frankfurt
JFK New York
ORD Chicago
BOS Boston
BER Berlin
Table 3
apc city …
DFW Dallas …
HBE Alexandria …
IST Istanbul …
FRA Frankfurt …
Table 2
……..
… … airport city
… … FRA Frankfurt
… … DFW Dallas
… … JFK New York
… … BER Berlin
… .. … …
Table 4
apc location
JFK New York
BER Berlin
ORD Illinoise
FRA Hessen
… …
Augment query
FRA Frankfurt
BOS Boston
DFW Dallas
airport City
BER Berlin
JFK New York
ORD Chicago
FRA Frankfurt
BOS Boston
DFW Dallas
Augment database
airport City
BER Berlin
JFK New York
ORD Chicago
FRA Frankfurt
BOS Boston
DFW Dallas
HBE ?
IST ?
… …
Refine
3
X Y Score Lineage
FRA Frankfurt 0.83 T1,T2
BOS Boston 1 T1
DFW Dallas 0.67 T2
FRA Hessen 0.67 T4
… … … …
Result
4

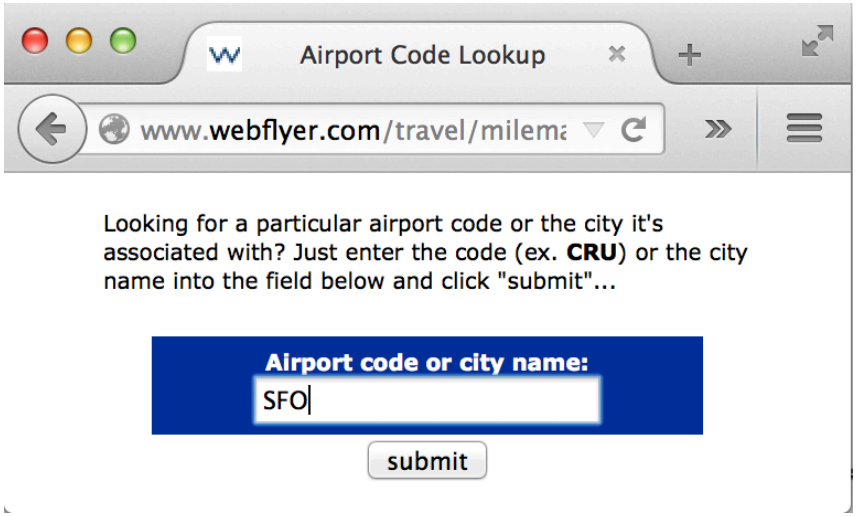
Web Forms

- How to find Web forms?
  - Use search engine
- How to use a Web form?
  - Generate a wrapper
- How to avoid high response time?
  - Cache results as new tables

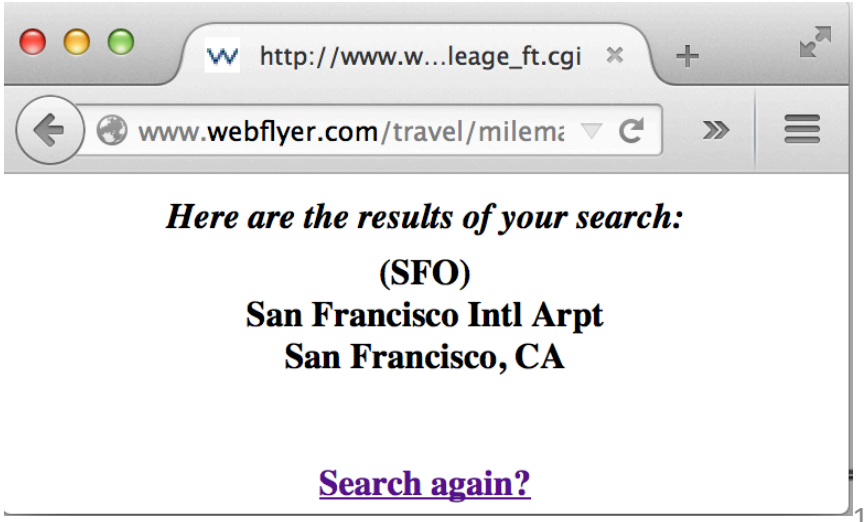


Wrapping Web forms

- Parse the HTML and find request parameters



- Locate output path by probing with examples



Expert System for Corner Cases

- Evaluate transformations
- Solve conflicts
- Create Transformations

Experiments

- Collected 50 queries from computer scientists and Tamr customers

1. Fahrenheit to Celsius
2. miles to km
3. pound to kg
4. USD to EUR
5. zip to state
6. zip to city
7. UPS tracking to address
8. english to german
9. swift code to bank
10. hex to RGB
11. ISBN to publisher
12. ISBN to title
13. ISBN to author
14. ISSN to title
15. ip adress to country
16. Domain to primary ip
17. sentence to language
18. text to encoding
19. Gregorian to Hijri
20. CUSIP to company
21. CUSIP to ticker
22. symbol to company
23. iban to bank name
24. Location to temperature
25. location to humidity
26. car plate to details
27. country code to country
28. ascii to char
29. car model to brand
30. country to demonym
31. country to language
32. country to currency
33. company to BBGID
34. patent ID to name
35. city to long/lat
36. Entity to wikipedia link
37. Entity to google graph id
38. person to twitter id
39. ip to domain
40. company to CEO
41. company to industry
42. US standard to metric
43. fractions to decimals
44. country to code
45. State to state abbrv
46. time zone to abbrv
47. city to country
48. airport code to city
49. RGB to color
50. ASCII to unicode

Coverage of the System

- Tested random input values for each query

	Web form wrapped	Web form found but not wrapped	Not found	
Covered by Tab	12	5	12	**29**
Not covered	12	5	4	**21**
	24	**10**	**16**	**50**

Covered: 24 + 29 -12= 41/50 (82%)

Precision and Recall of the Covered Transformations

$$\text{precision} = \frac{\text{number of correct transformations}}{\text{number of output values}}$$

$$\text{recall} = \frac{\text{number of correct transformations}}{\text{number of input values}}$$

- 10 Input values per query
- Average precision = 91%
- Average recall = 81.3%

Conclusion & Future Work

- DataXFormer:
  - Web tables are good at semantic transformations
  - Web forms are good at syntactic transformations
  - The expert crowd helps with difficult tasks

- Future Work
  - Extend our Web table repository
  - Apply fuzzy matching
  - Multi-column transformations
  - Collect more queries
    - **http://www.dataxformer.org**

Please Help!!!

"Humans and Transformers should be friends…"
Optimus Prime

- Give us your transformation!

http://www.dataxformer.org

- Thank you! (abedjan@csail.mit.edu)