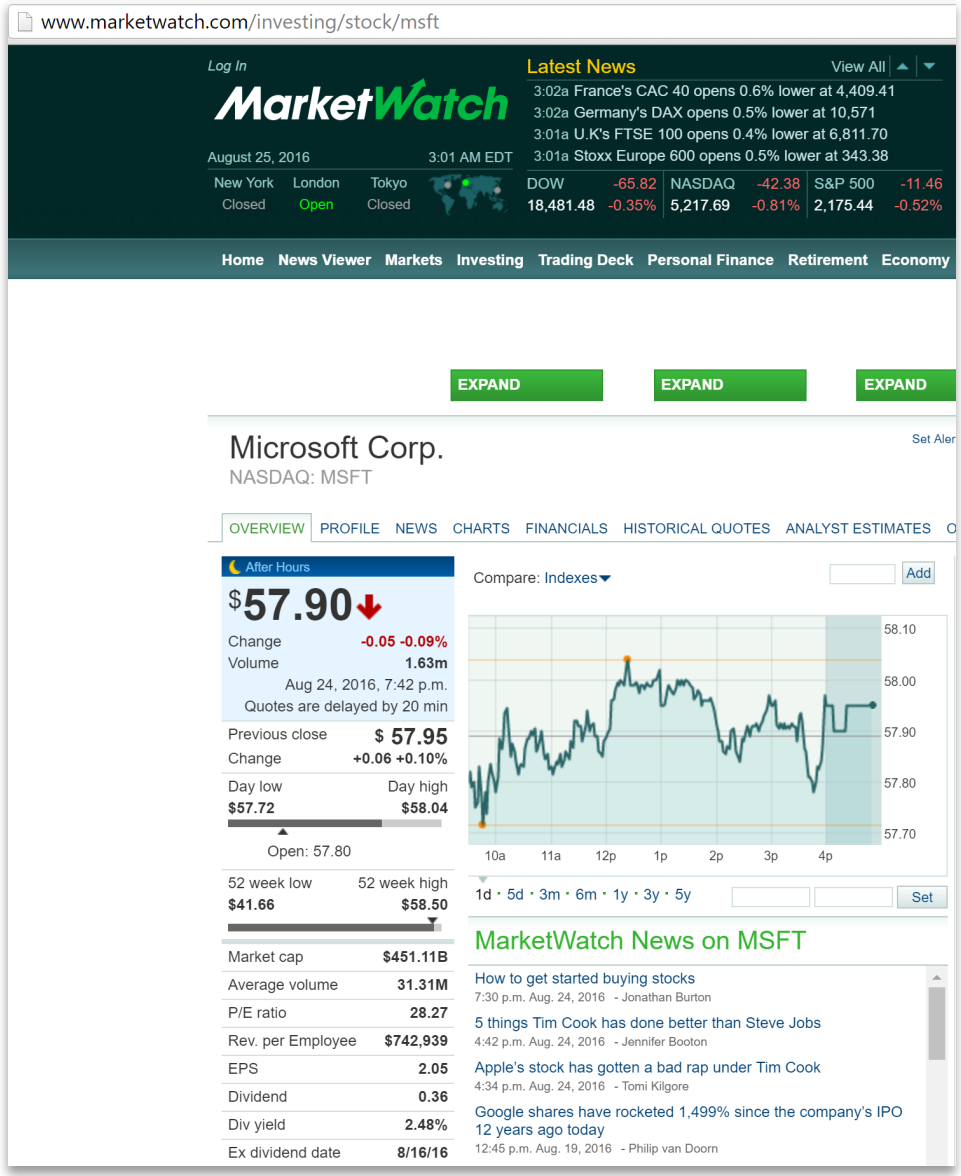
WebRelate: Integrating Web Data with Spreadsheets using Examples

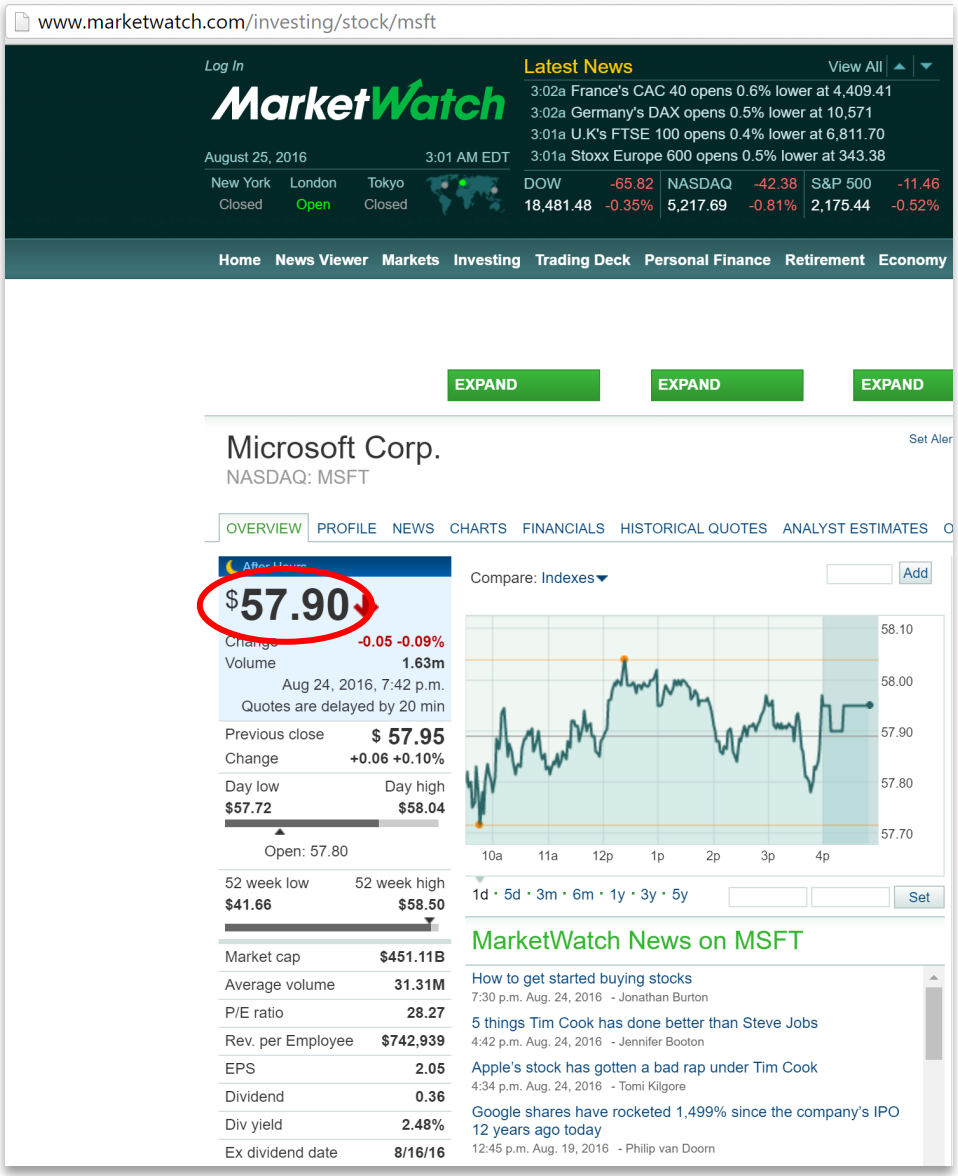
Jeevana Inala
MIT

Rishabh Singh
Microsoft Research

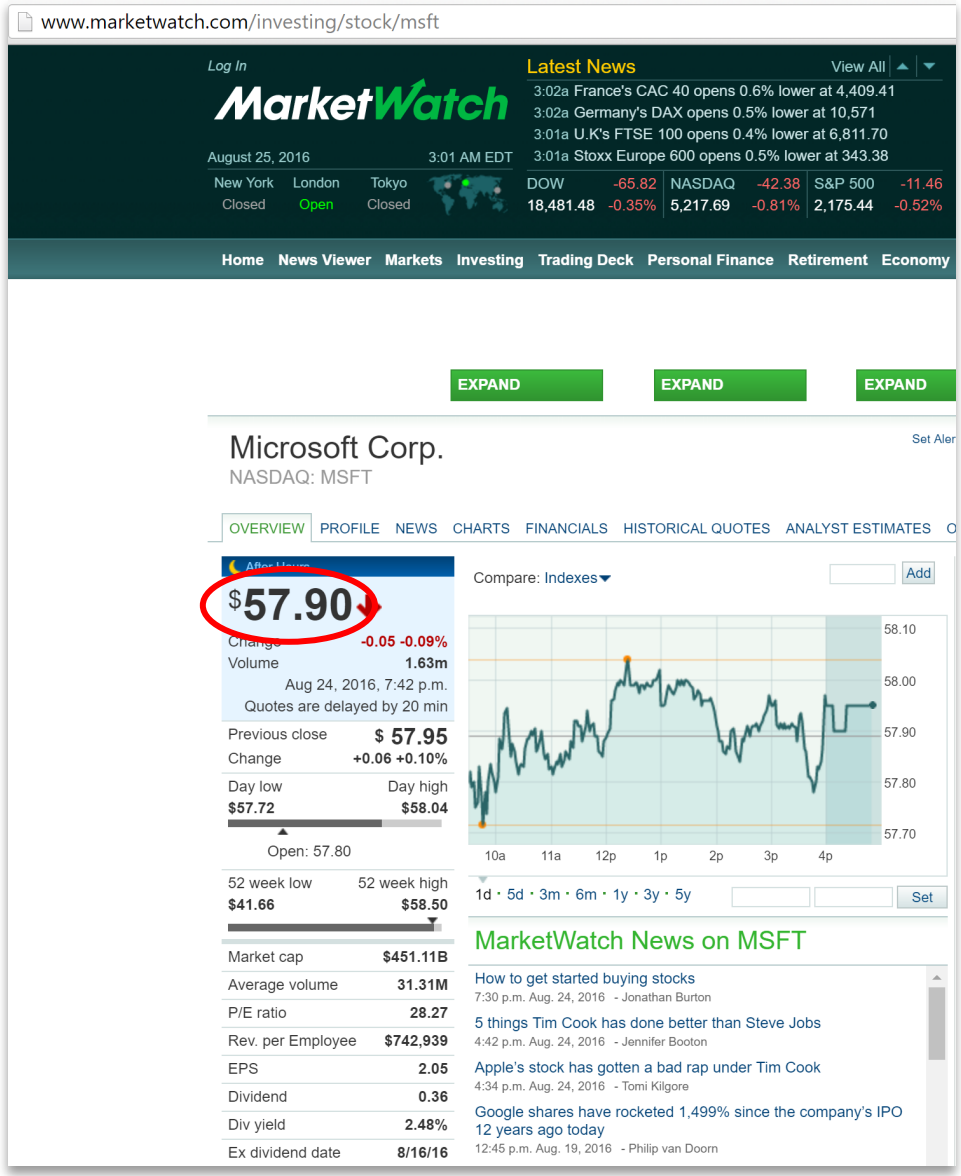
Company	Stock value
MSFT	
AMZN	
AAPL	
T	
S	



Company	Stock value
MSFT	
AMZN	
AAPL	
T	
S	



Company	Stock value
MSFT	57.90
AMZN	
AAPL	
T	
S	



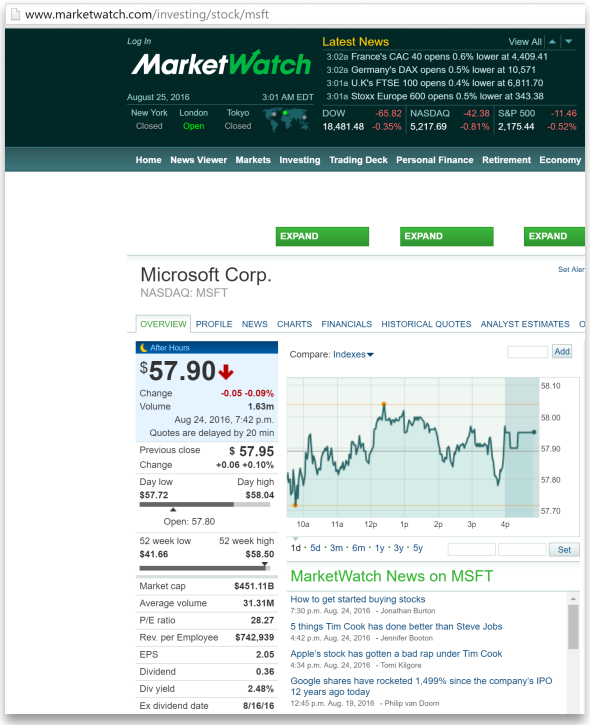
Code:

```
Public Sub Import_Yahoo_Finance_Historical()

    Dim URL As String
    Dim dateParams As String

    'Date ranges from default earliest Yahoo start date (m/d/y) to current date (m/d/y)

    dateParams = "&a=0&b=3&c=1977&d=" & Month(Date) - 1 & "&e=" & Day(Date) & "&f=" & Year(Date)

    'Daily prices

    URL = "http://ichart.finance.yahoo.com/table.csv?s=" & Sheets("Analysis").Range("C2").Value & dateParams & "&g=d&ignore=.csv"

    With Worksheets("Input")
        With .QueryTables.Add(Connection:="TEXT;" & URL, Destination:=.Range("A1"))
            .TextFileStartRow = 1
            .TextFileParseType = xlDelimited
            .TextFileCommaDelimiter = True
            .Refresh BackgroundQuery:=False
        End With
        .QueryTables(1).Delete
    End With

    'Dividends only

    URL = "http://ichart.finance.yahoo.com/table.csv?s=" & Sheets("Analysis").Range("C2").Value & dateParams & "&g=v&ignore=.csv"

    With Worksheets("Input")
        With .QueryTables.Add(Connection:="TEXT;" & URL, Destination:=.Range("I1"))
            .TextFileStartRow = 1
```

WebRelate

Company	Stock value
MSFT	
AMZN	
AAPL	
T	
S	

+



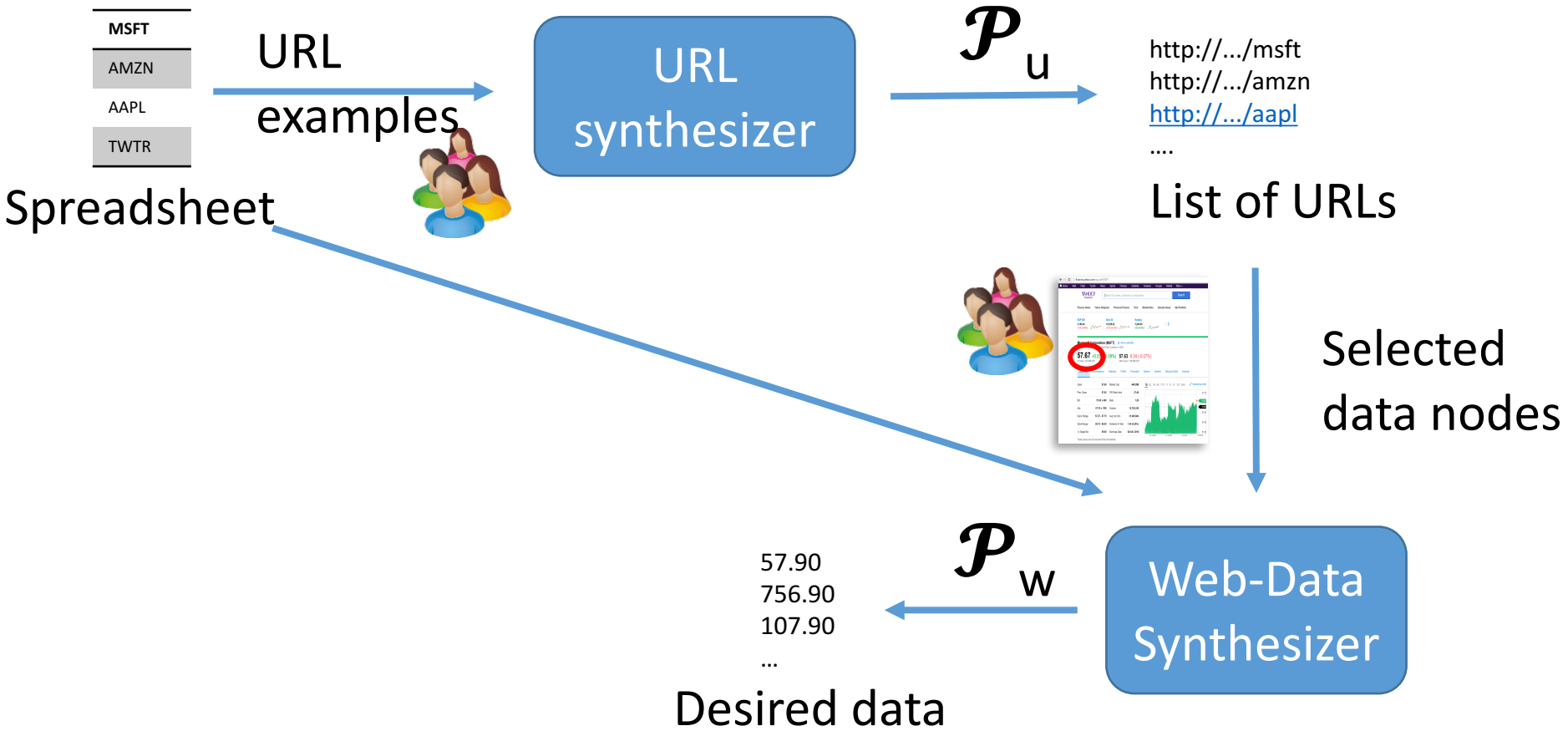
Examples

Program

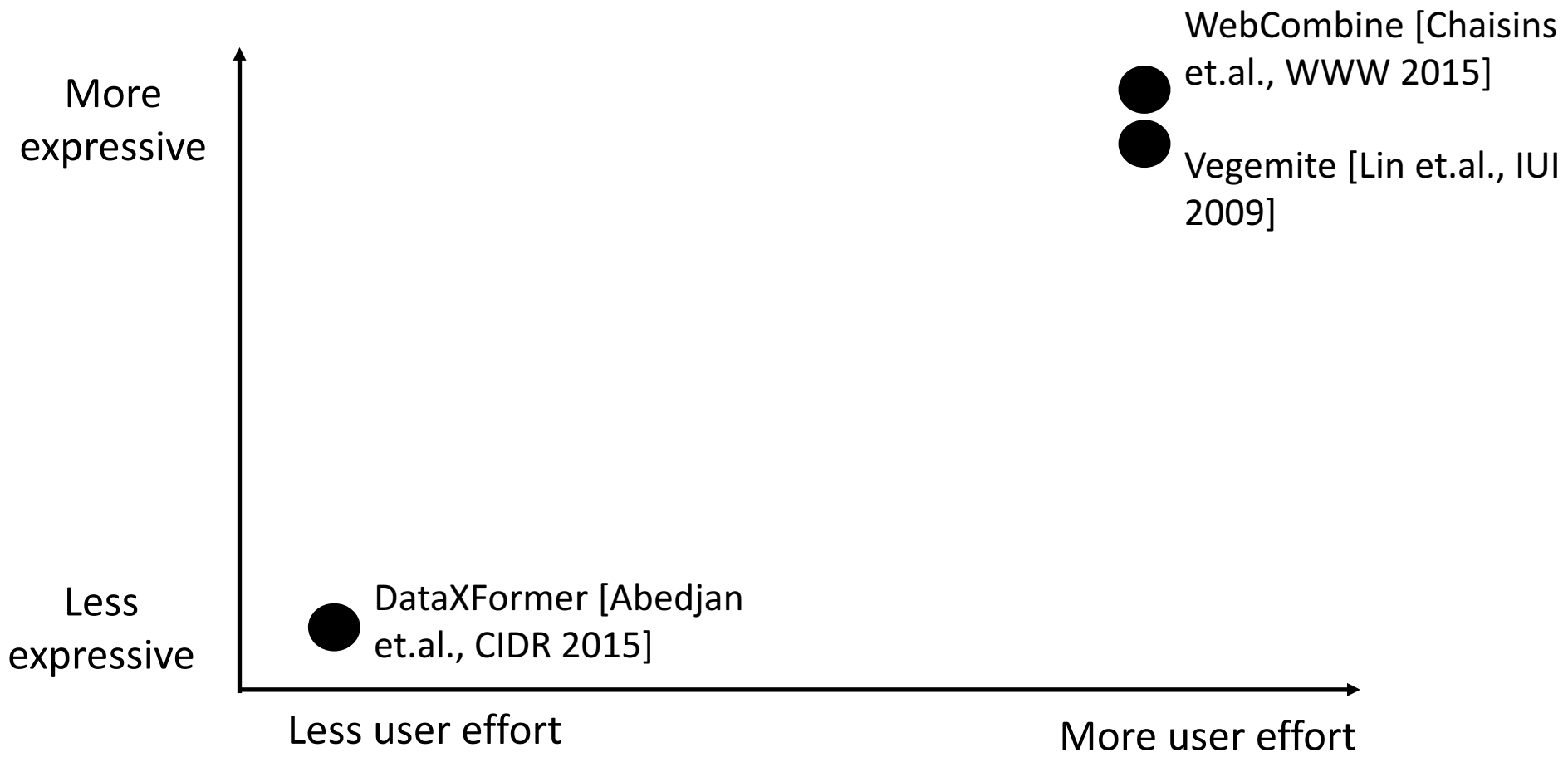
Company	Stock value
MSFT	57.90
AMZN	759.48
AAPL	108.51
T	40.91
S	6.04

Demo

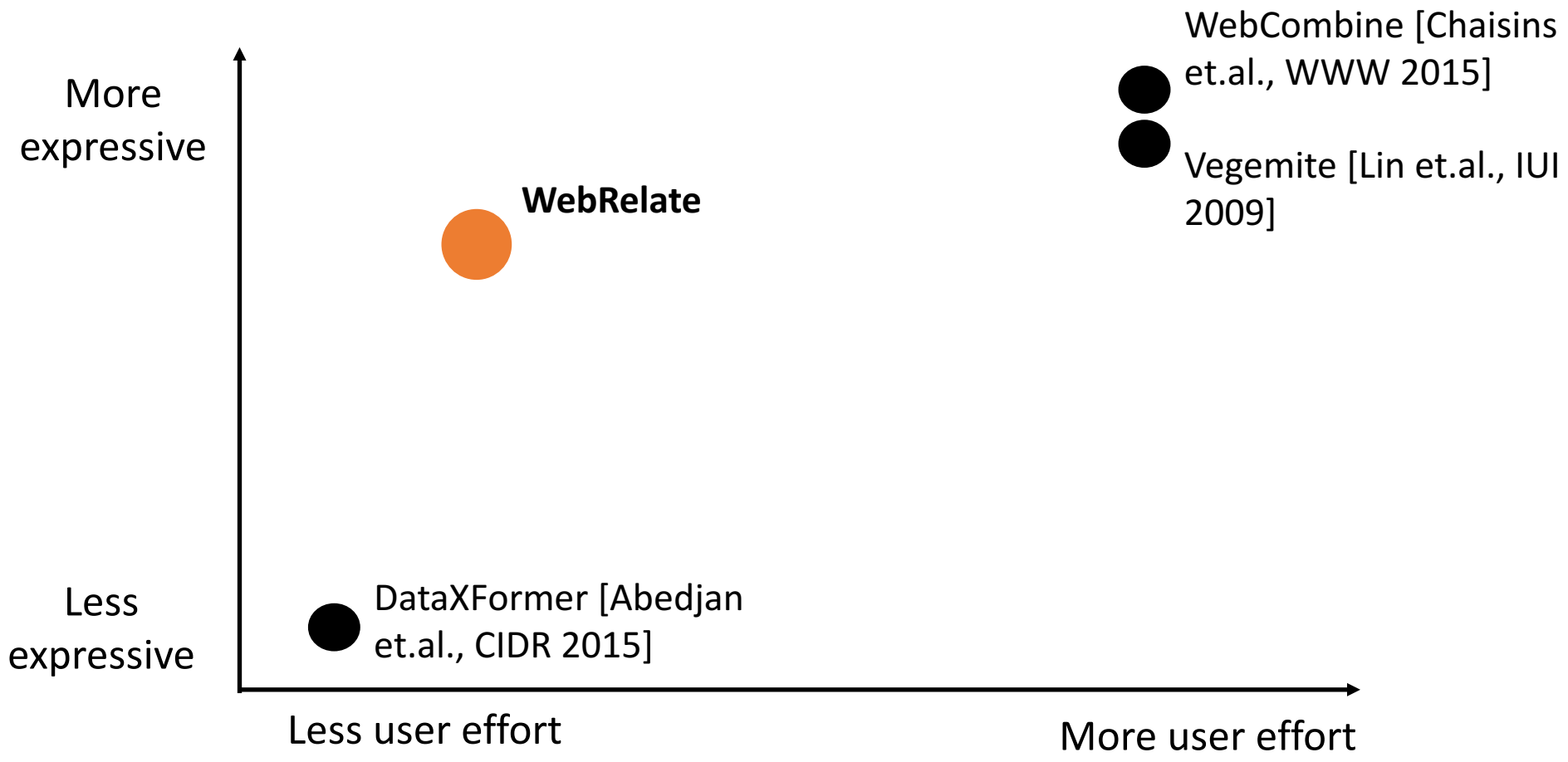
Overview


MSFT
AMZN
AAPL
TWTR
URL examples
Spreadsheet
URL synthesizer
$\mathcal{P}_u$
http://.../msft
http://.../amzn
http://.../aapl
....
List of URLs
Selected data nodes
Web-Data Synthesizer
$\mathcal{P}_w$
57.90
756.90
107.90
...
Desired data

Related Work



Related Work



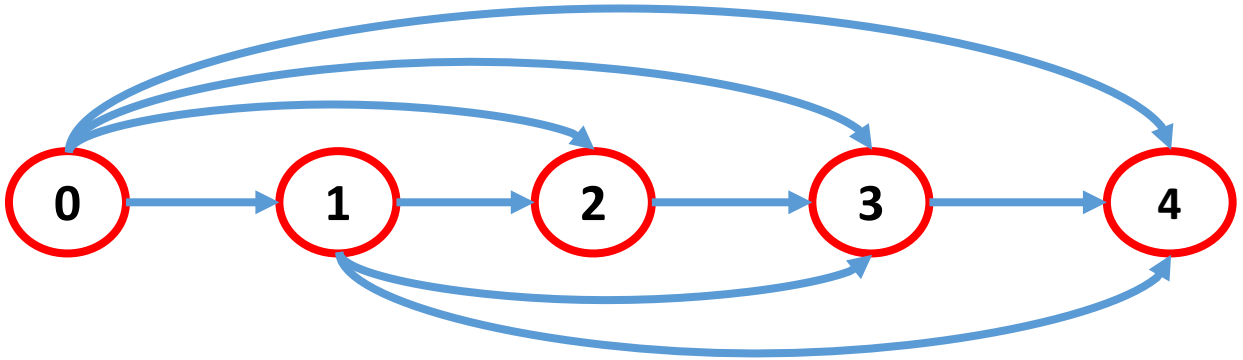
Related Work

- Learning string transformations
  - Flash Fill (Gulwani, POPL 11), Blink Fill (Singh, VLDB 16)

Rishabh Singh  R.S.

Related Work

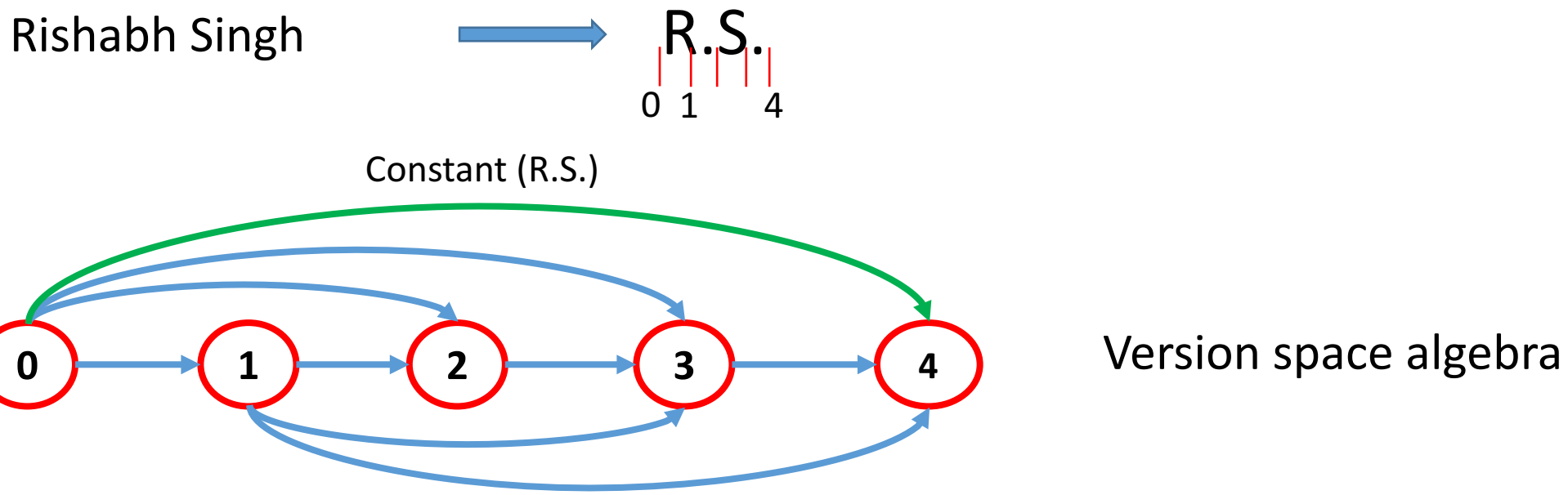
- Learning string transformations
  - Flash Fill (Gulwani, POPL 11), Blink Fill (Singh, VLDB 16)



Version space algebra

Related Work

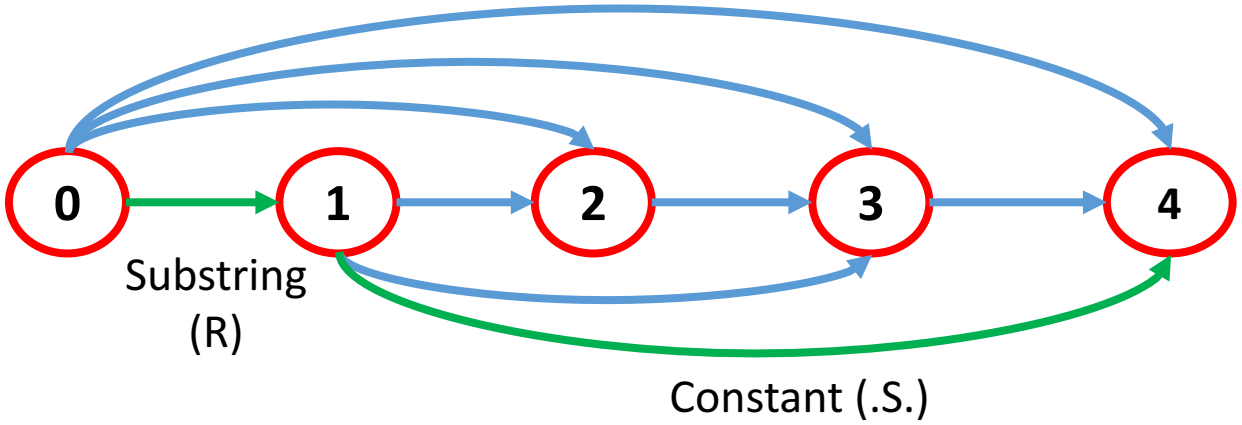
- Learning string transformations
  - Flash Fill (Gulwani, POPL 11), Blink Fill (Singh, VLDB 16)



Version space algebra

Related Work

- Learning string transformations
  - Flash Fill (Gulwani, POPL 11), Blink Fill (Singh, VLDB 16)

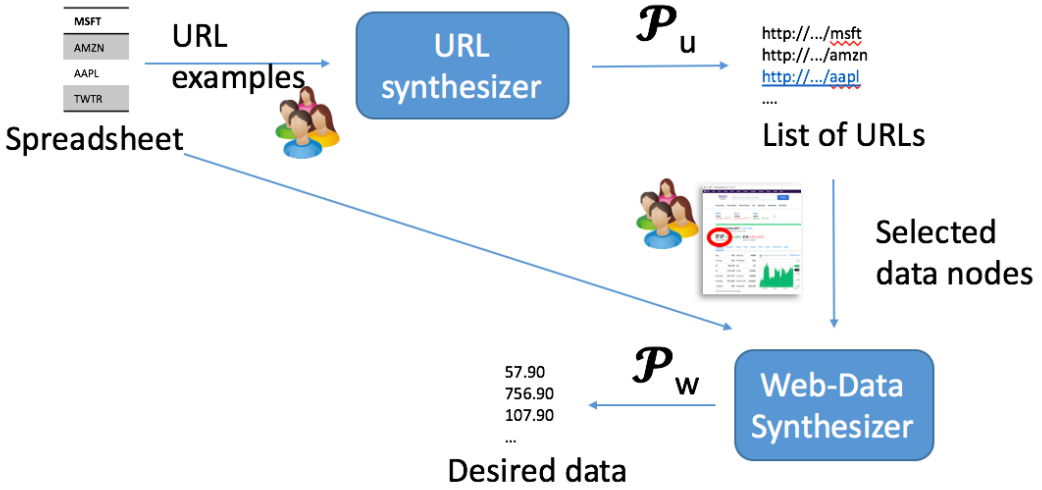


Version space algebra

Ranking: substring > constant

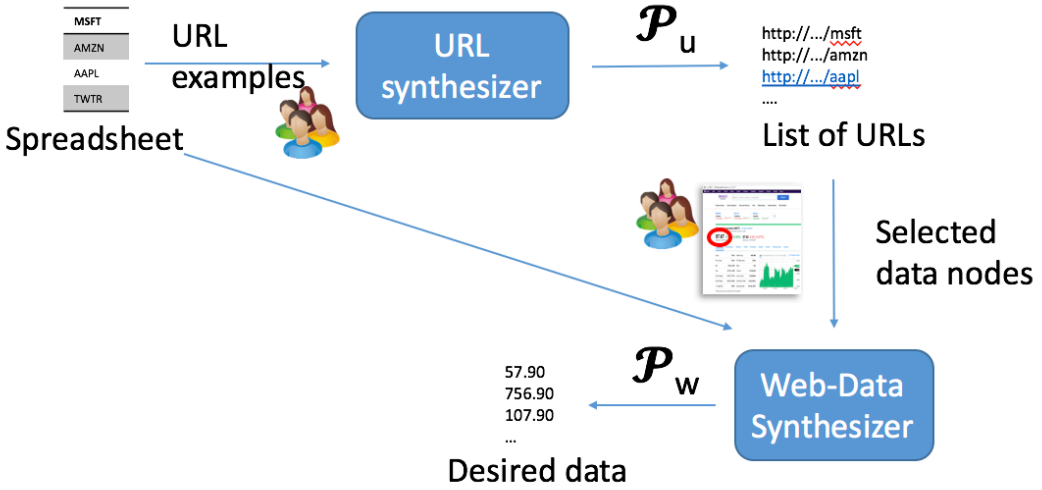
Overview

- Output-constrained PBE
- Layered Version Space Algebra
- Input-dependent Web Extraction



Overview

- ***Output-constrained PBE***
- Layered Version Space Algebra
- Input-dependent Web Extraction

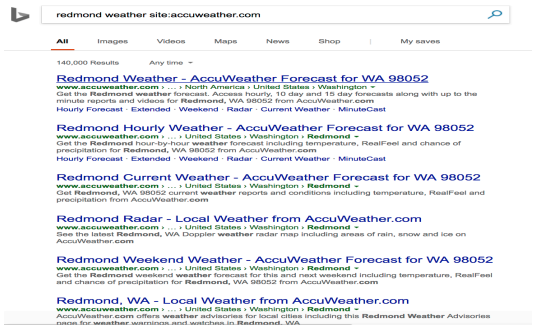


Traditional PBE

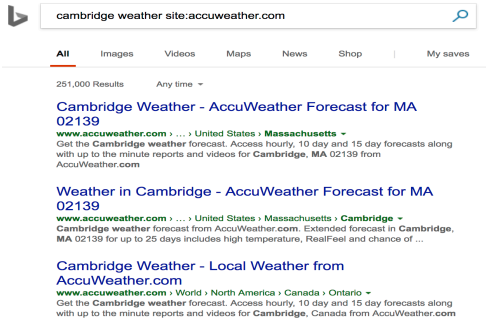
Ana Trujillo 357 21th Place SE,Redmond,WA → http://www.accuweather.com/en/us/redmond-wa/98052/weather-forecast/341347

Output-constrained PBE

Ana Trujillo 357 21th Place SE,Redmond,WA +



Charlie Gunaja 732 Memorial Drive, Cambridge,MA +



→ http://www.accuweather.com/en/us/redmond-wa/98052/weather-forecast/341347

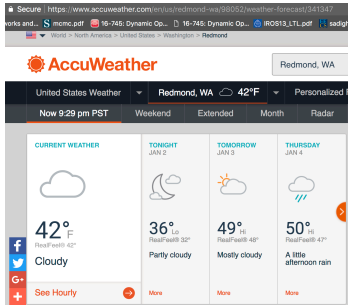
Traditional PBE

Ana Trujillo 357 21th Place SE,Redmond,WA

42°

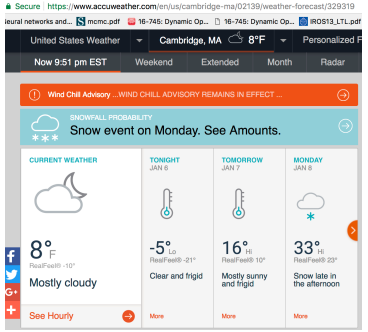
Output-constrained PBE (O-PBE)

Ana Trujillo 357 21th Place SE,Redmond,WA +



42°

Charlie Gunaja 732 Memorial Drive, Cambridge,MA +



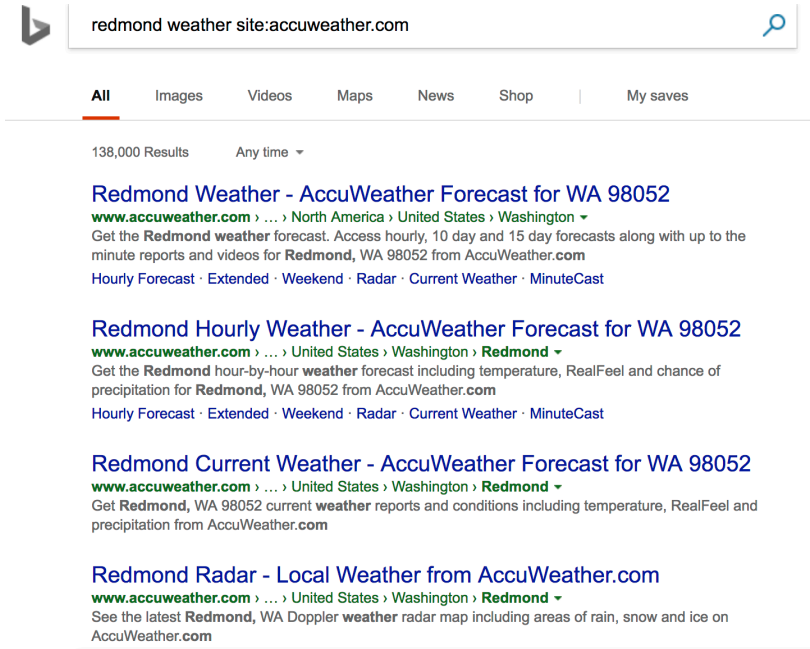
Output-constrained PBE (O-PBE)

- Generalization constraint
  - $\forall$ in $\in$ inputs. ***P(in)*** $\in$ ***list of possible of outputs for in***

- Uniqueness constraint
  - $\forall$ (in,out) $\in$ input-output examples. ***P(in) = out***

Richer class of problems

Ana Trujillo 357 21th Place SE,Redmond,WA



http://www.accuweather.com/en/us/redmond-wa/98052/weather-forecast/341347

Richer class of problems

Ana Trujillo 357 21th Place SE,Redmond,WA

https://www.accuweather.com/en/us/redmond-wa/98052/weather-forecast/341347

https://www.accuweather.com/en/us/redmond-wa/98052/hourly-weather-forecast/341347

https://www.accuweather.com/en/us/redmond-or/97756/weather-forecast/340247

http://www.accuweather.com/en/us/redmond-wa/98052/weather-forecast/341347

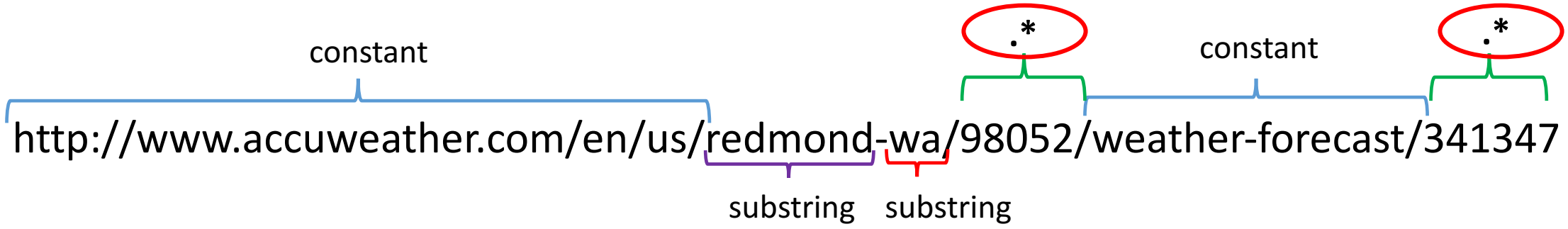
Richer class of problems

Ana Trujillo 357 21th Place SE,Redmond,WA

https://www.accuweather.com/en/us/redmond-wa/98052/weather-forecast/341347

https://www.accuweather.com/en/us/redmond-wa/98052/hourly-weather-forecast/341347

https://www.accuweather.com/en/us/redmond-or/97756/weather-forecast/340247



Richer class of problems

Ana Trujillo 357 21th Place SE,Redmond,WA

https://www.accuweather.com/en/us/redmond-wa/98052/weather-forecast/341347

https://www.accuweather.com/en/us/redmond-wa/98052/hourly-weather-forecast/341347

https://www.accuweather.com/en/us/redmond-or/97756/weather-forecast/340247

http://www.accuweather.com/en/us/redmond-wa/.*/weather-forecast/.*

Richer class of problems

Ana Trujillo 357 21th Place SE,Redmond,WA

https://www.accuweather.com/en/us/redmond-wa/98052/weather-forecast/341347

https://www.accuweather.com/en/us/redmond-wa/98052/hourly-weather-forecast/341347

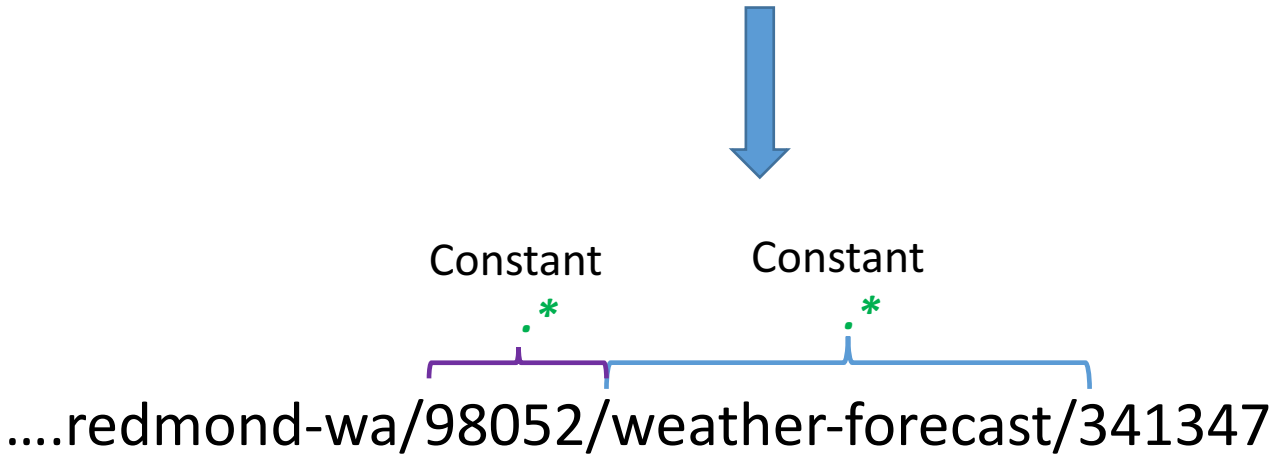
https://www.accuweather.com/en/us/redmond-or/97756/weather-forecast/340247

http://www.accuweather.com/en/us/redmond-wa/.*/weather-forecast/.*

It is sufficient to learn a program that uniquely identifies the desired output.

Better ranking

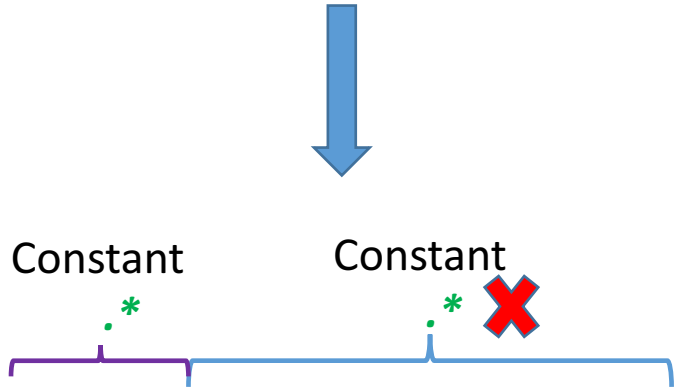
Ana Trujillo 357 21th Place SE,Redmond,WA



How do we know which part of the URL should be .*?

Better ranking

Ana Trujillo 357 21th Place SE,Redmond,WA



….redmond-wa/98052/weather-forecast/341347

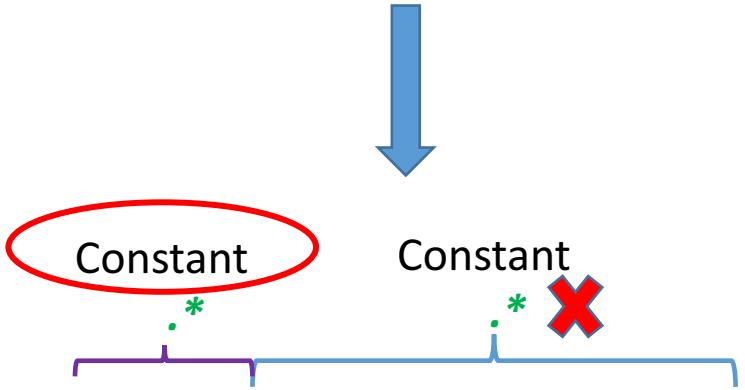
https://www.accuweather.com/en/us/redmond-wa/98052/weather-forecast/341347

https://www.accuweather.com/en/us/redmond-wa/98052/hourly-weather-forecast/341347

https://www.accuweather.com/en/us/redmond-or/97756/weather-forecast/340247

Better ranking

Ana Trujillo 357 21th Place SE,Redmond,WA



….redmond-wa/98052/weather-forecast/341347

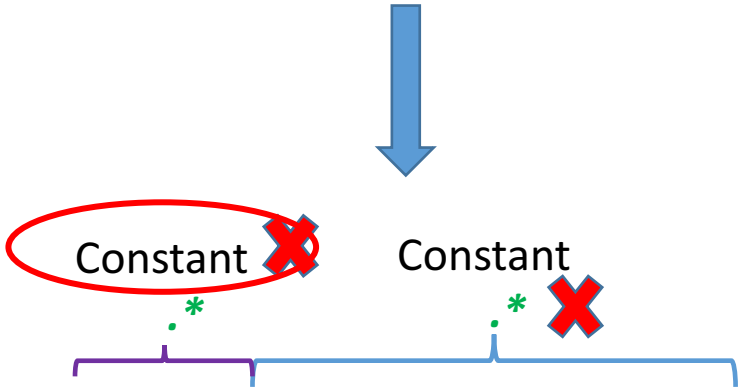
https://www.accuweather.com/en/us/redmond-wa/98052/weather-forecast/341347

https://www.accuweather.com/en/us/redmond-wa/98052/hourly-weather-forecast/341347

https://www.accuweather.com/en/us/redmond-or/97756/weather-forecast/340247

Better ranking

Ana Trujillo 357 21th Place SE,Redmond,WA



….redmond-wa/98052/weather-forecast/341347

Charlie Gunaja 732 Memorial Drive, Cambridge,MA

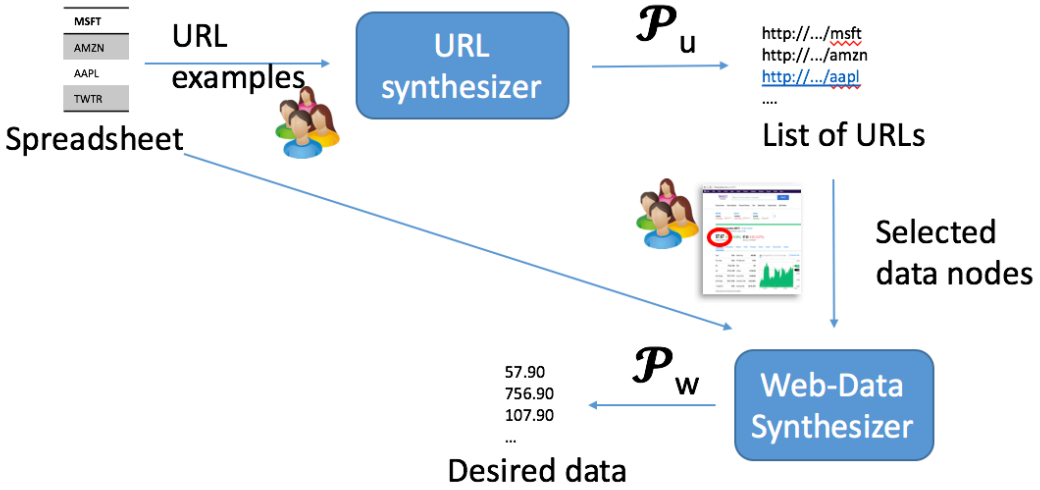
https://www.accuweather.com/en/us/cambridge-ma/02139/weather-forecast/329319

https://www.accuweather.com/en/us/cambridge-ma/02139/hourly-weather-forecast/329319

https://www.accuweather.com/en/us/cambridge-ma/02139/daily-weather-forecast/329319

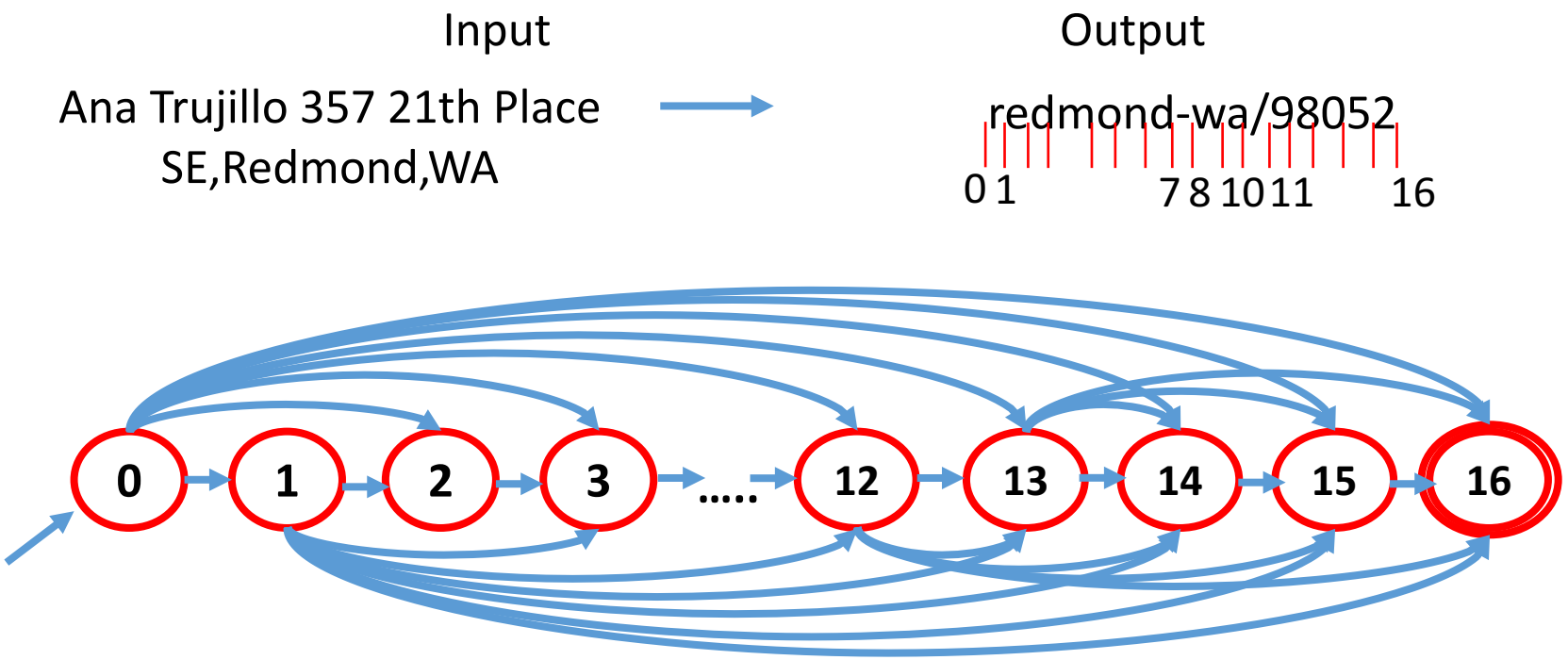
Overview

- Output-constrained PBE
- ***Layered Version Space Algebra***
- Input-dependent Web Extraction



URLs are long!

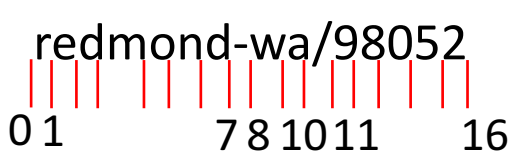
- Complexity of VSA used in Flash Fill
  - Quadratic in length of output
  - Exponential in number of examples



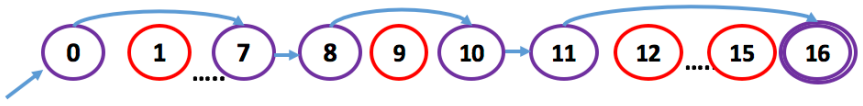
Layered Version Space Algebra

- Performs search over increasingly expressive sub-languages
  - $L_1 \subseteq L_2 \subseteq \ldots. \subseteq L_k$

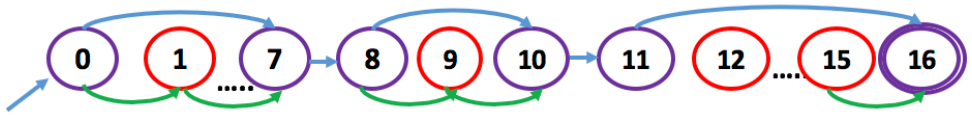
Ana Trujillo 357 21th Place
SE,Redmond,WA



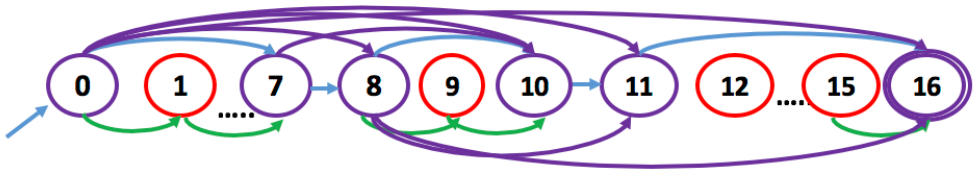
Layer 1



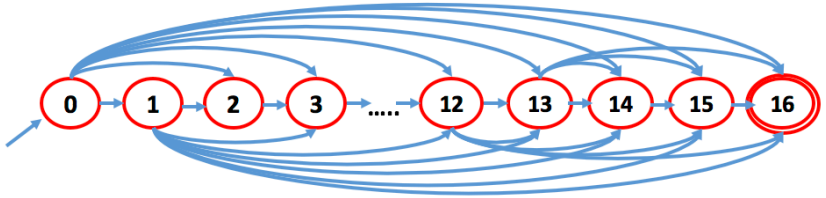
Layer 2



Layer 3

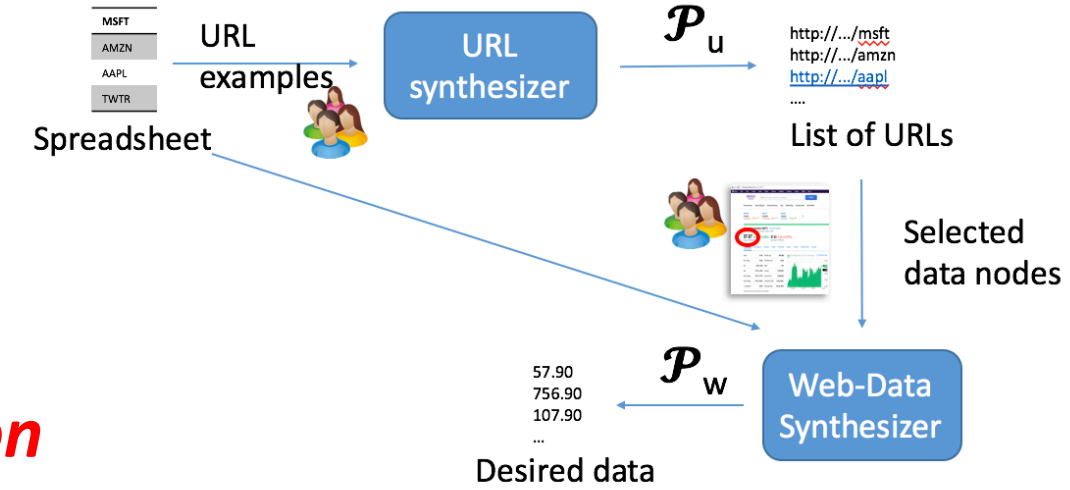


Layer 4 – Full dag

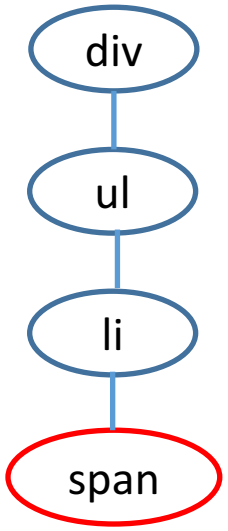


Overview

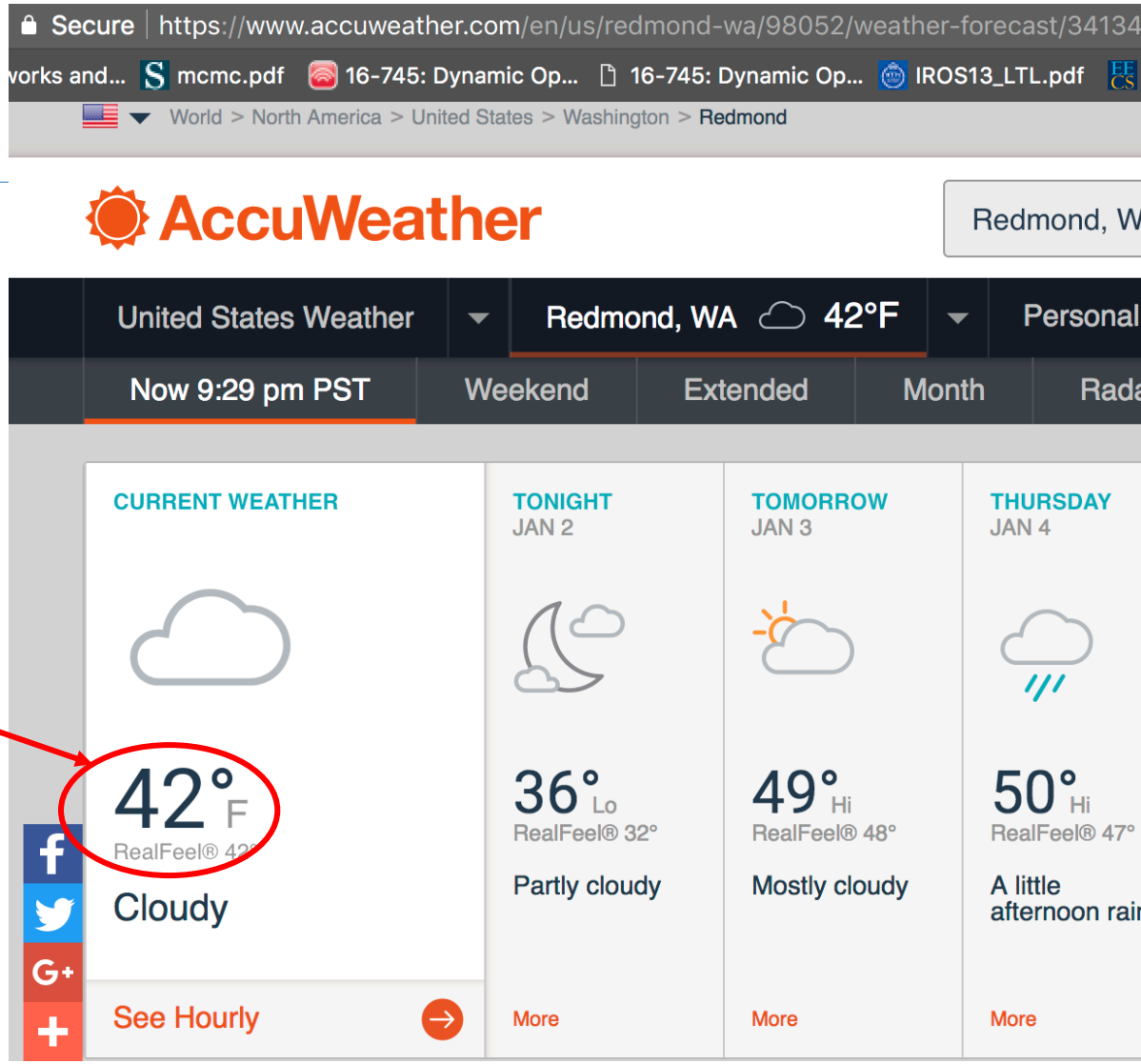
- Output-constrained PBE
- Layered Version Space Algebra
- ***Input-dependent Web Extraction***



Web-Extraction



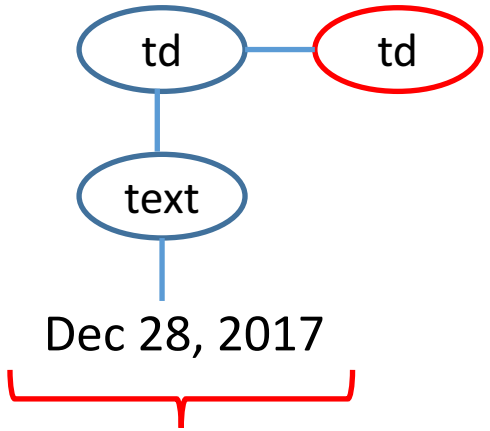
id = ”feed-tabs”



Can be expressed in Xpath query language

Web-Extraction

Cur1	Cur2	Date
USD	INR	28, December, 2017
EUR	GBP	03, January, 2018
USD	CHF	05, January, 2018



Transform(28, December, 2017)

Investing.com EUR/USD or AA

USD/INR 63.275 -0.095 (-0.15%)

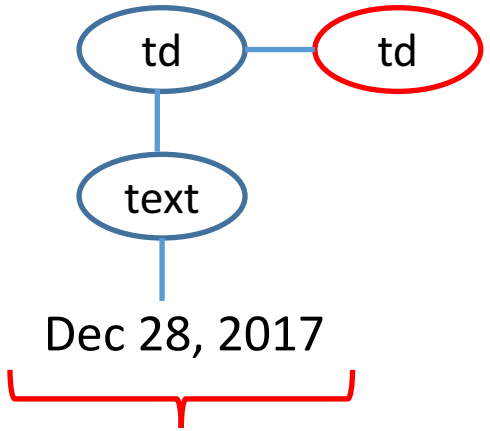
USD/INR Historical Data

Time Frame: Daily

Date	Price
Jan 05, 2018	63.340
Jan 04, 2018	63.400
Jan 03, 2018	63.505
Jan 02, 2018	63.460
Jan 01, 2018	63.680
Dec 29, 2017	63.840
Dec 28, 2017	64.080
Dec 27, 2017	64.120

Input-dependent Web-Extraction

Cur1	Cur2	Date
USD	INR	28, December, 2017
EUR	GBP	03, January, 2018
USD	CHF	05, January, 2018



Transform(28, December, 2017)

Investing.com EUR/USD or AA

USD/INR 63.275 -0.095 (-0.15%)

USD/INR Historical Data

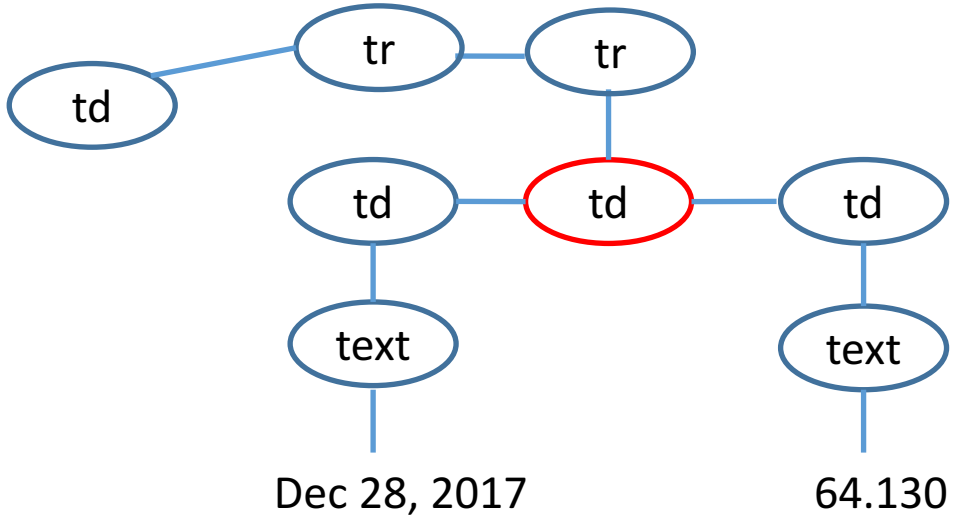
Time Frame: Daily

Date	Price
Jan 05, 2018	63.340
Jan 04, 2018	63.400
Jan 03, 2018	63.505
Jan 02, 2018	63.460
Jan 01, 2018	63.680
Dec 29, 2017	63.840
Dec 28, 2017	64.080
Dec 27, 2017	64.120

Input-dependent Web-Extraction

USD; INR; 28, December, 2017

Input



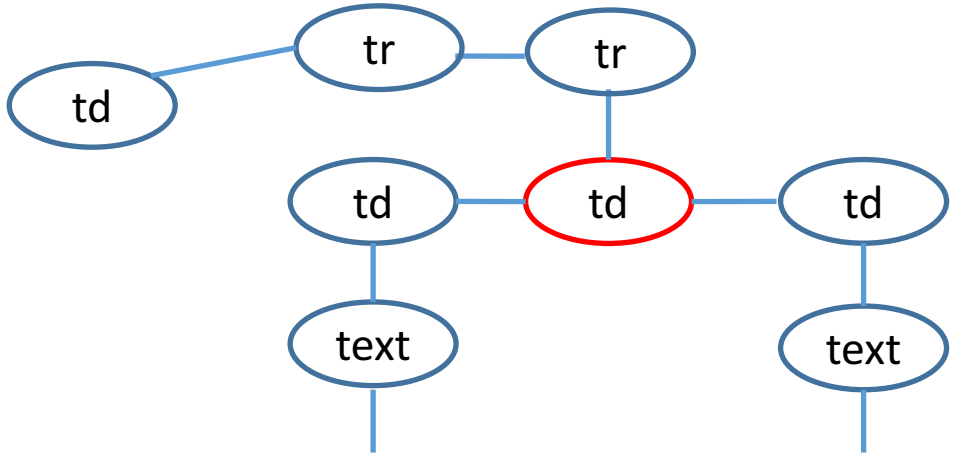
HTML tree

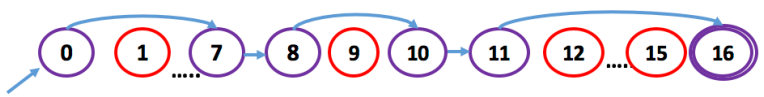
Find constraints in this tree that satisfy both *Uniqueness* and *Generalization* properties of the O-PBE problem

Input-dependent Web-Extraction

USD; INR; 28, December, 2017

Input

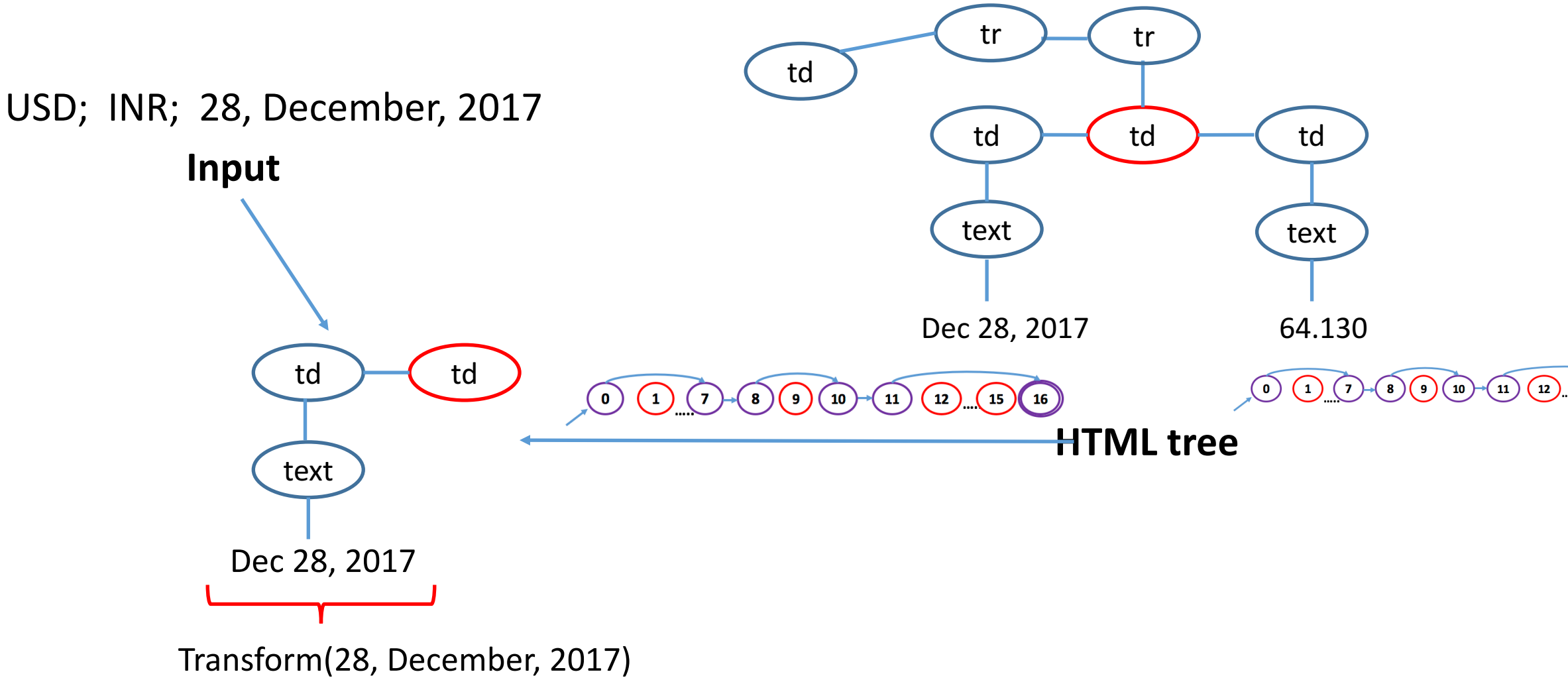




HTML tree

Find constraints in this tree that satisfy both *Uniqueness* and *Generalization* properties of the O-PBE problem

Input-dependent Web-Extraction

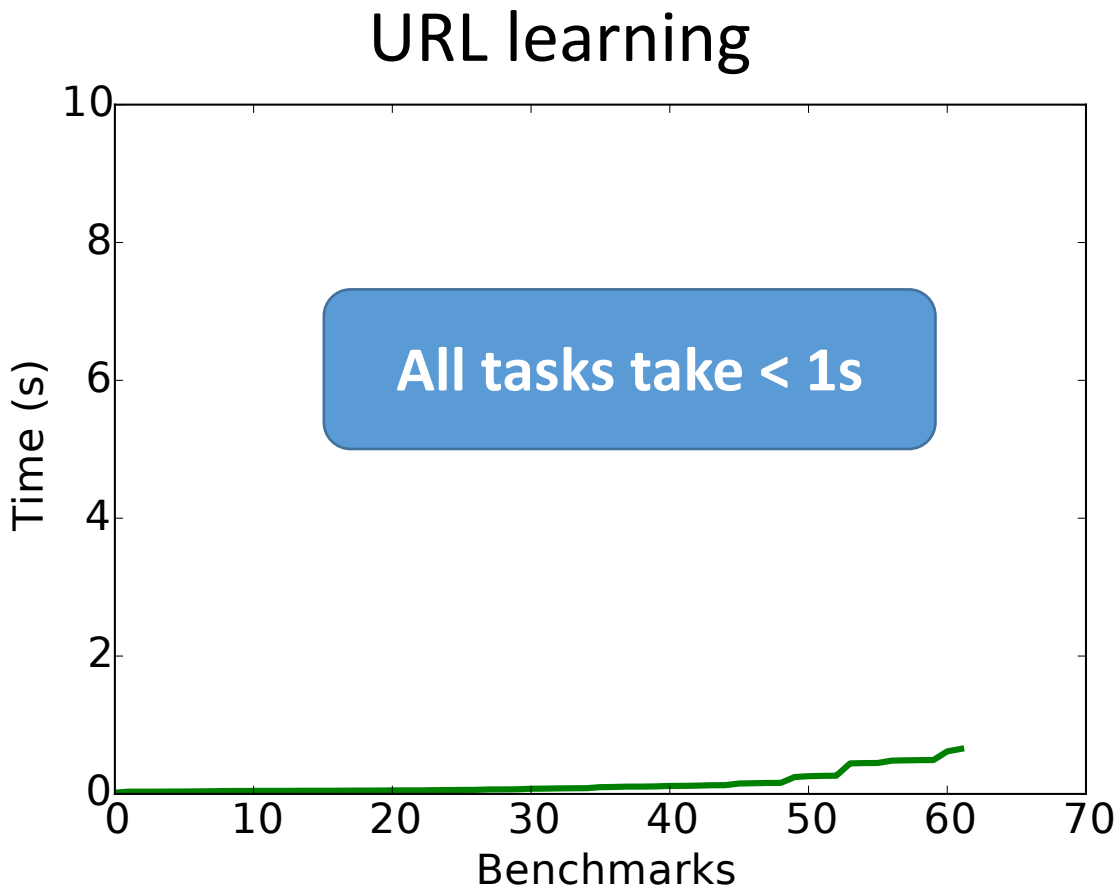


Results

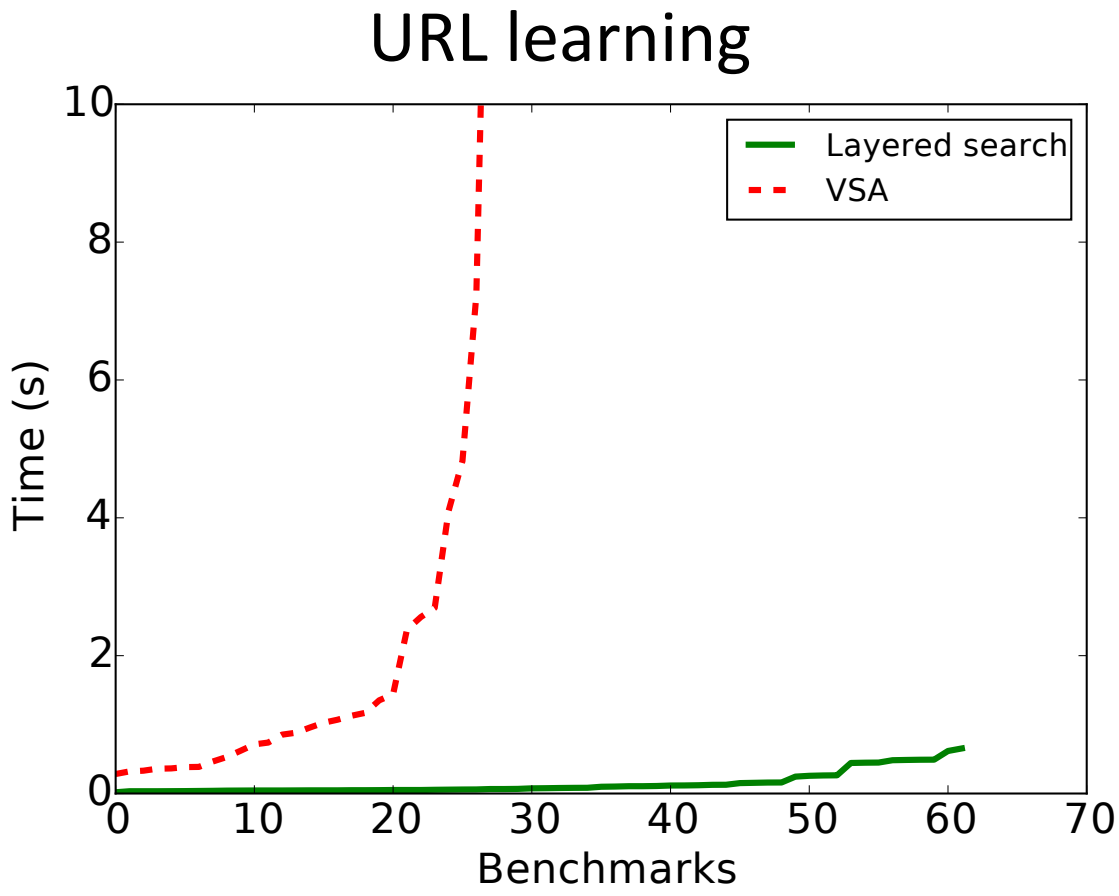
Expressive?

- 88 data integration scenarios
  - Stocks, weather, sports, currency, ….
  - 62 URL learning tasks
  - 88 Web-data extraction tasks
- 5 – 32 number of input rows in the spreadsheet
- Solves all of them correctly

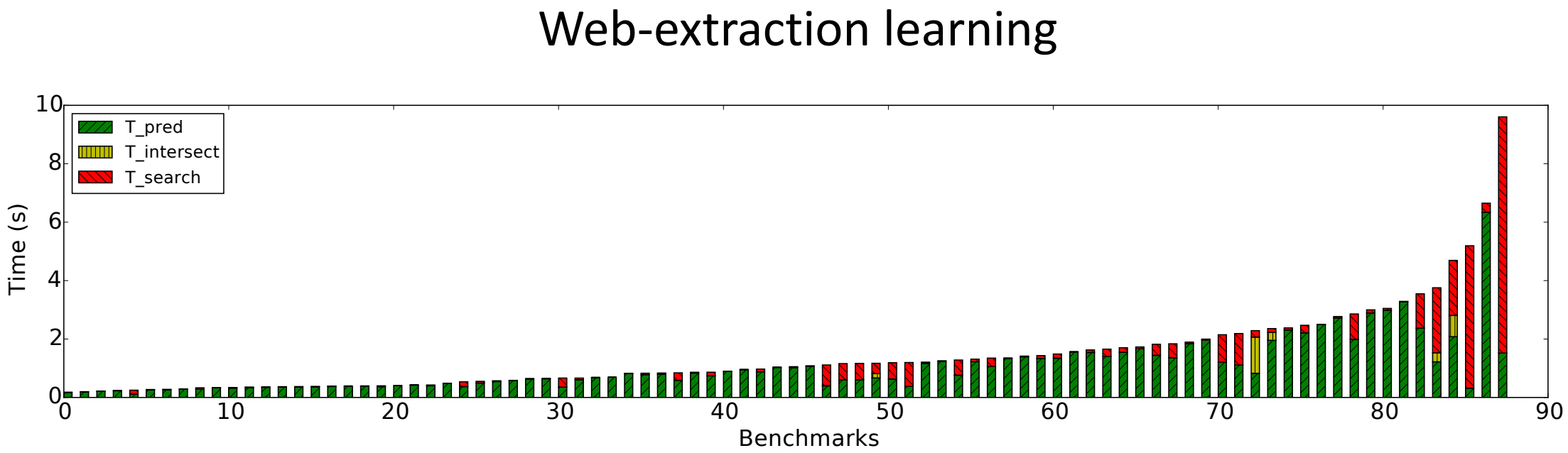
Fast?


URL learning
All tasks take < 1s
Time (s)
Benchmarks
0
2
4
6
8
10
0
10
20
30
40
50
60
70

Layered search beats VSA

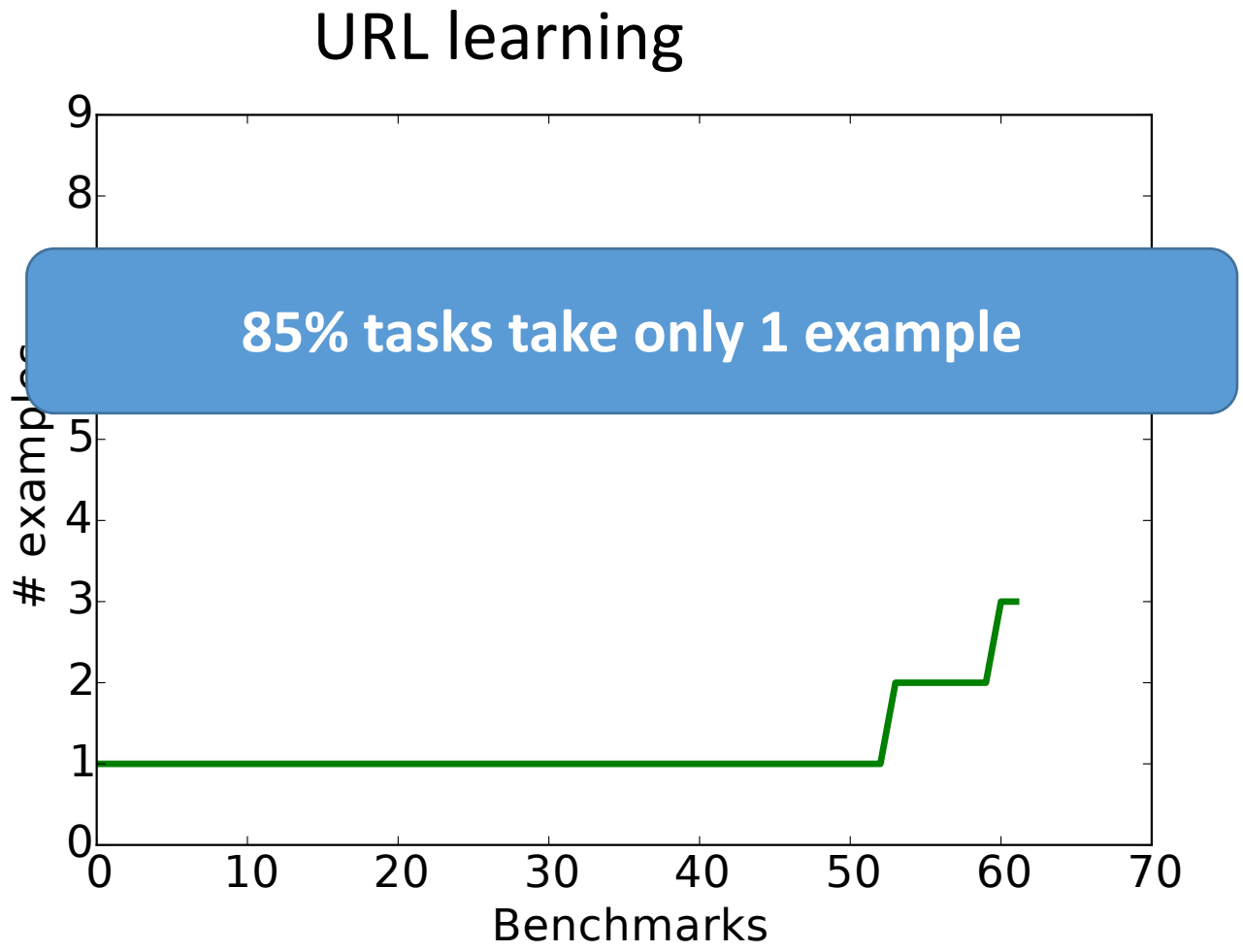


Fast?

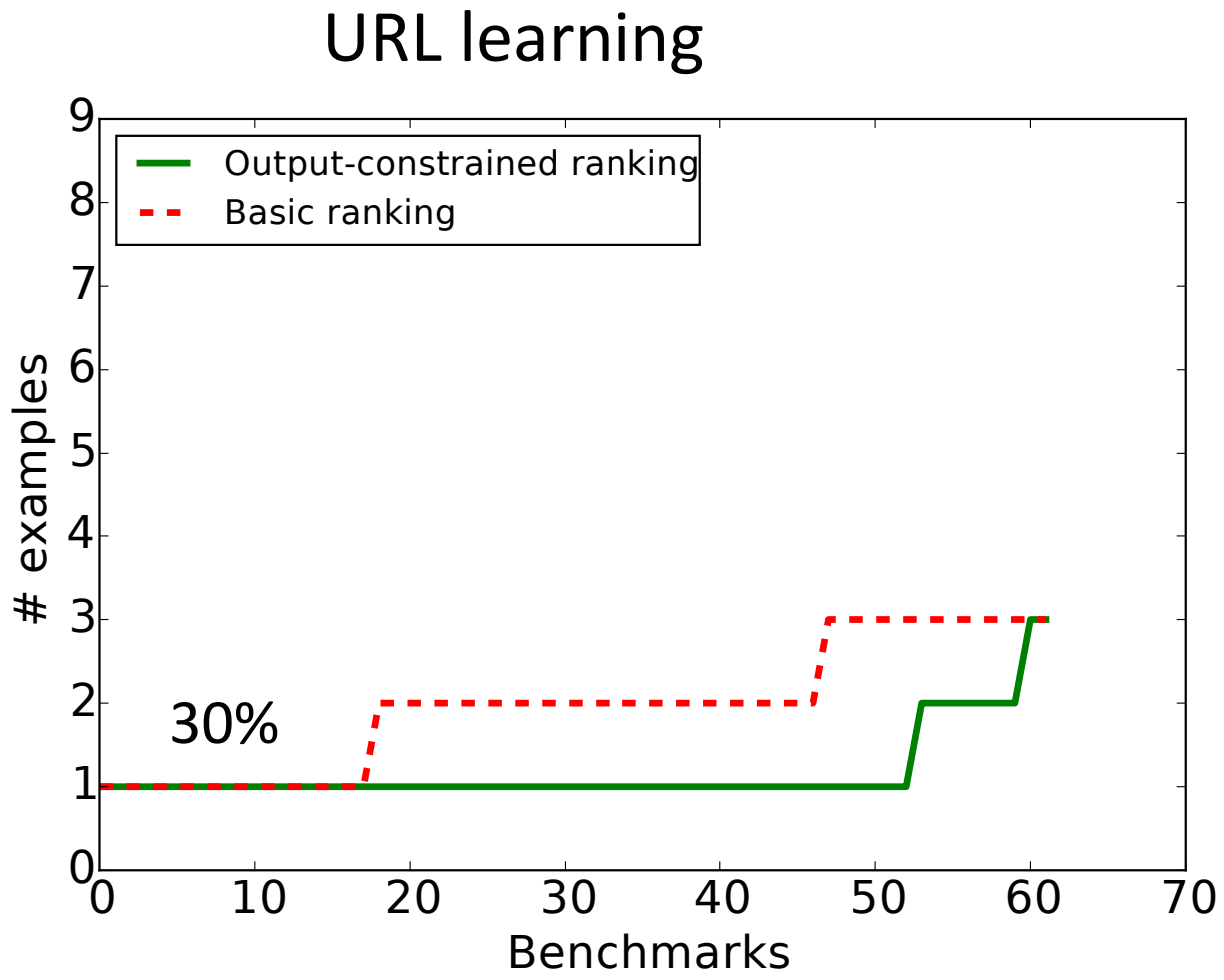


All tasks take < 10s

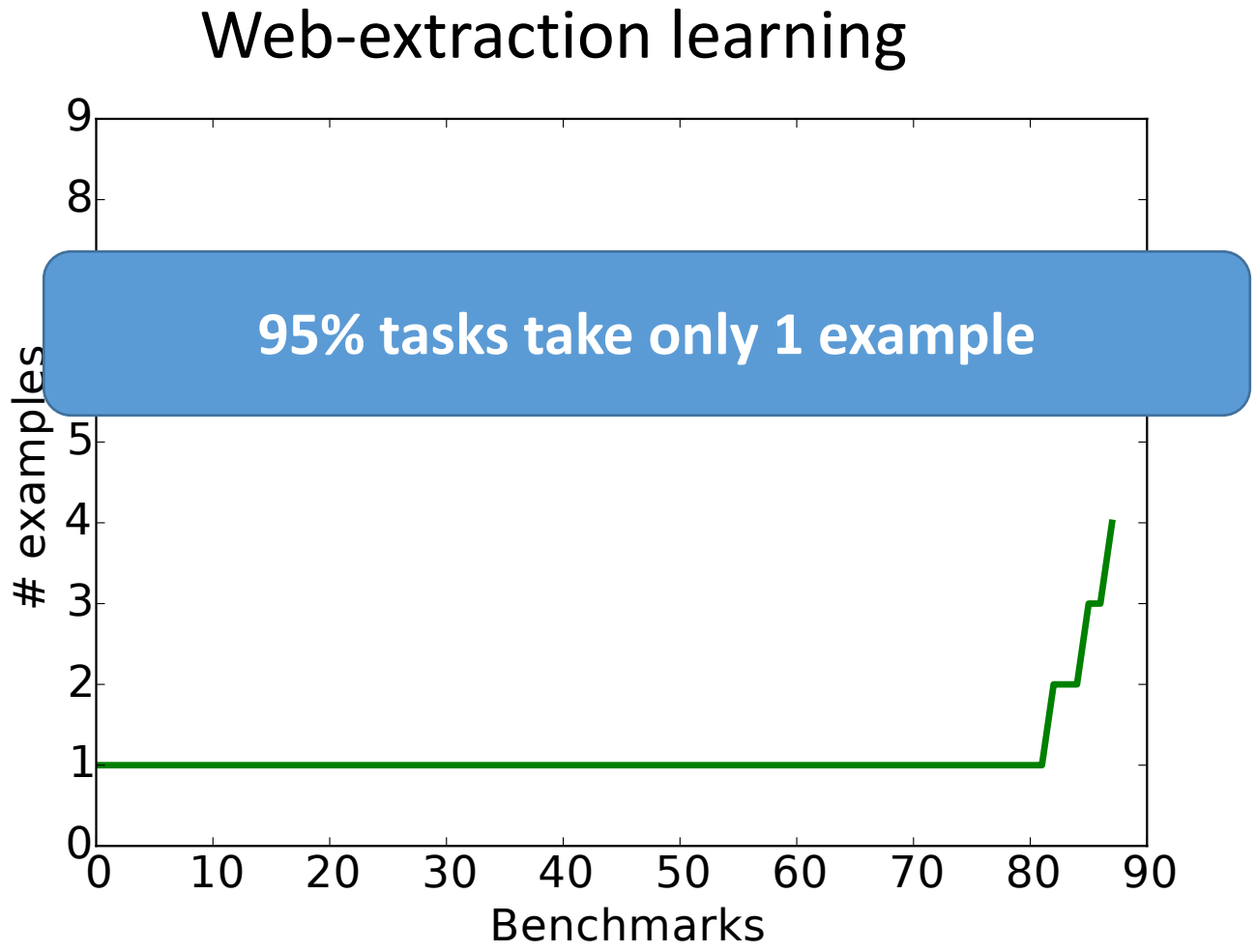
Easy to use?


URL learning
85% tasks take only 1 example
examples
0
1
2
3
4
5
8
9
0
10
20
30
40
50
60
70
Benchmarks

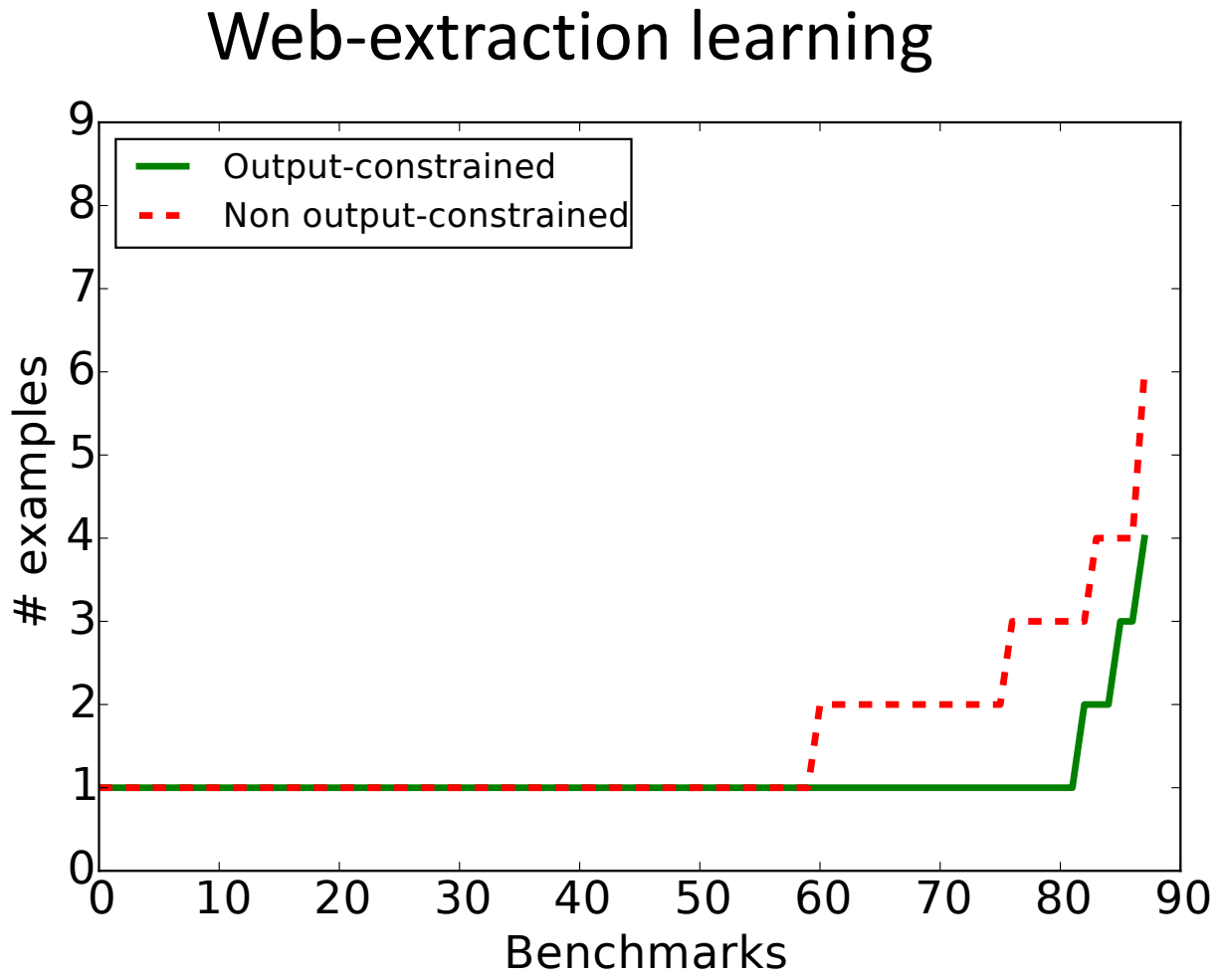
Impact of generalization constraint



Easy to use?

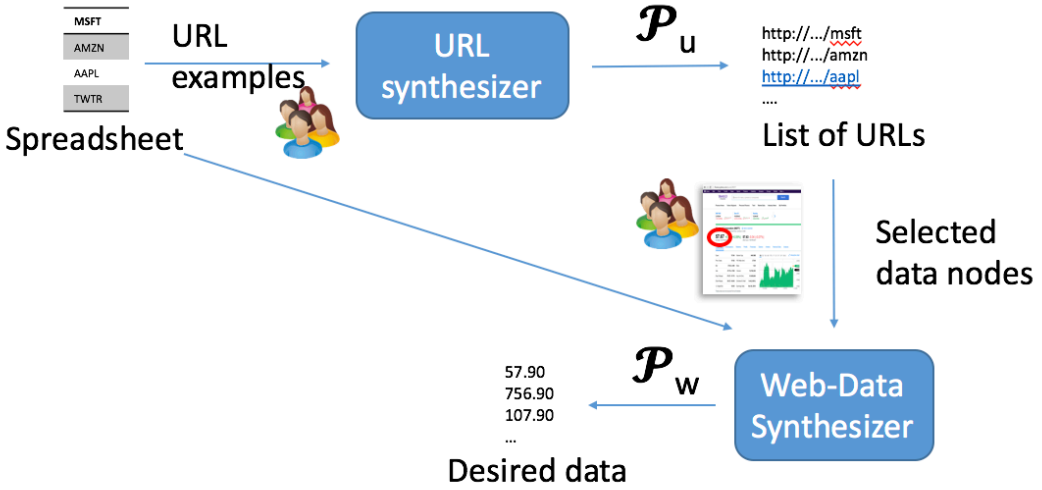

Web-extraction learning
95% tasks take only 1 example
examples
Benchmarks
0
1
2
3
4
5
8
9
0
10
20
30
40
50
60
70
80
90

Impact of generalization constraint



Summary

- Output-constrained PBE
- Layered Version Space Algebra
- Input-dependent Web Extraction



Thank You!

jinala@mit.edu