

WebRelate: Integrating Web Data with Spreadsheets using Examples

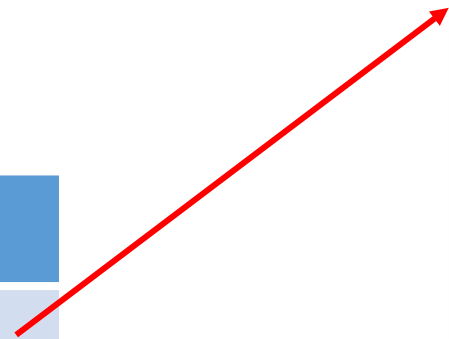
Jeevana Inala

MIT

Rishabh Singh

Microsoft Research

Company	Stock value
MSFT	
AMZN	
AAPL	
T	
S	



www.marketwatch.com/investing/stock/msft

Log In View All

MarketWatch

August 25, 2016 3:01 AM EDT

New York Closed London Open Tokyo Closed

Latest News

- 3:02a France's CAC 40 opens 0.6% lower at 4,409.41
- 3:02a Germany's DAX opens 0.5% lower at 10,571
- 3:01a U.K's FTSE 100 opens 0.4% lower at 6,811.70
- 3:01a Stoxx Europe 600 opens 0.5% lower at 343.38

DOW -65.82 -0.35% NASDAQ -42.38 -0.81% S&P 500 -11.46 -0.52%

18,481.48 5,217.69 2,175.44

Home News Viewer Markets Investing Trading Deck Personal Finance Retirement Economy

EXPAND EXPAND EXPAND

Microsoft Corp.

NASDAQ: MSFT Set Alert

OVERVIEW PROFILE NEWS CHARTS FINANCIALS HISTORICAL QUOTES ANALYST ESTIMATES

After Hours

\$57.90 ↓

Change -0.05 -0.09%
Volume 1.63m
Aug 24, 2016, 7:42 p.m.
Quotes are delayed by 20 min

Previous close **\$57.95**
Change +0.06 +0.10%
Day low \$57.72 Day high \$58.04
Open: 57.80

52 week low \$41.66 52 week high \$58.50

Market cap \$451.11B
Average volume 31.31M
P/E ratio 28.27
Rev. per Employee \$742,939
EPS 2.05
Dividend 0.36
Div yield 2.48%
Ex dividend date 8/16/16

Compare: Indexes

1d · 5d · 3m · 6m · 1y · 3y · 5y

MarketWatch News on MSFT

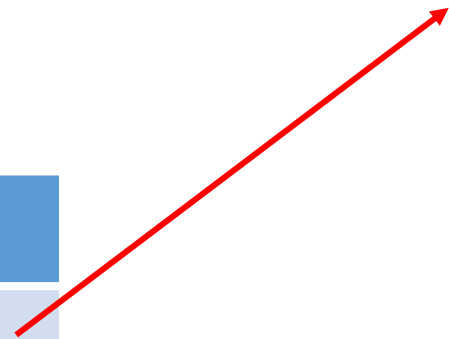
How to get started buying stocks
7:30 p.m. Aug. 24, 2016 - Jonathan Burton

5 things Tim Cook has done better than Steve Jobs
4:42 p.m. Aug. 24, 2016 - Jennifer Booton

Apple's stock has gotten a bad rap under Tim Cook
4:34 p.m. Aug. 24, 2016 - Tomi Kilgore

Google shares have rocketed 1,499% since the company's IPO 12 years ago today
12:45 p.m. Aug. 19, 2016 - Philip van Doorn

Company	Stock value
MSFT	
AMZN	
AAPL	
T	
S	



www.marketwatch.com/investing/stock/msft

Log In View All

MarketWatch

August 25, 2016 3:01 AM EDT

New York Closed London Open Tokyo Closed

Latest News

- 3:02a France's CAC 40 opens 0.6% lower at 4,409.41
- 3:02a Germany's DAX opens 0.5% lower at 10,571
- 3:01a U.K's FTSE 100 opens 0.4% lower at 6,811.70
- 3:01a Stoxx Europe 600 opens 0.5% lower at 343.38

DOW -65.82 -0.35% NASDAQ -42.38 -0.81% S&P 500 -11.46 -0.52%

18,481.48 5,217.69 2,175.44

Home News Viewer Markets Investing Trading Deck Personal Finance Retirement Economy

EXPAND EXPAND EXPAND

Microsoft Corp.

NASDAQ: MSFT Set Alert

OVERVIEW PROFILE NEWS CHARTS FINANCIALS HISTORICAL QUOTES ANALYST ESTIMATES

After Hours

\$57.90 ▼

Change -0.05 -0.09%
Volume 1.63m
Aug 24, 2016, 7:42 p.m.
Quotes are delayed by 20 min

Previous close \$ 57.95
Change +0.06 +0.10%
Day low \$57.72 Day high \$58.04
Open: 57.80

52 week low \$41.66 52 week high \$58.50

Market cap \$451.11B
Average volume 31.31M
P/E ratio 28.27
Rev. per Employee \$742,939
EPS 2.05
Dividend 0.36
Div yield 2.48%
Ex dividend date 8/16/16

Compare: Indexes

1d · 5d · 3m · 6m · 1y · 3y · 5y

MarketWatch News on MSFT

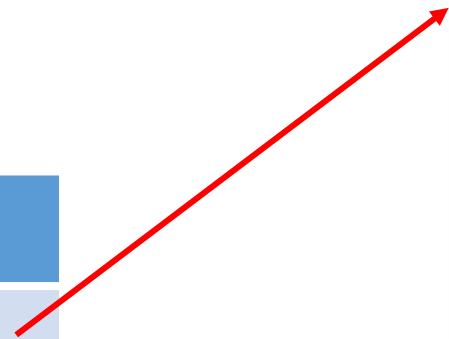
How to get started buying stocks
7:30 p.m. Aug. 24, 2016 - Jonathan Burton

5 things Tim Cook has done better than Steve Jobs
4:42 p.m. Aug. 24, 2016 - Jennifer Booton

Apple's stock has gotten a bad rap under Tim Cook
4:34 p.m. Aug. 24, 2016 - Tomi Kilgore

Google shares have rocketed 1,499% since the company's IPO 12 years ago today
12:45 p.m. Aug. 19, 2016 - Philip van Doorn

Company	Stock value
MSFT	57.90
AMZN	
AAPL	
T	
S	



www.marketwatch.com/investing/stock/msft

Log In View All

MarketWatch

August 25, 2016 3:01 AM EDT

New York Closed London Open Tokyo Closed

Latest News

- 3:02a France's CAC 40 opens 0.6% lower at 4,409.41
- 3:02a Germany's DAX opens 0.5% lower at 10,571
- 3:01a U.K's FTSE 100 opens 0.4% lower at 6,811.70
- 3:01a Stoxx Europe 600 opens 0.5% lower at 343.38

DOW -65.82 -0.35% NASDAQ -42.38 -0.81% S&P 500 -11.46 -0.52%

18,481.48 5,217.69 2,175.44

Home News Viewer Markets Investing Trading Deck Personal Finance Retirement Economy

EXPAND EXPAND EXPAND

Microsoft Corp.

NASDAQ: MSFT Set Alert

OVERVIEW PROFILE NEWS CHARTS FINANCIALS HISTORICAL QUOTES ANALYST ESTIMATES

After Hours

\$57.90 ▼

Change -0.05 -0.09%
Volume 1.63m
Aug 24, 2016, 7:42 p.m.
Quotes are delayed by 20 min

Previous close \$57.95
Change +0.06 +0.10%
Day low \$57.72 Day high \$58.04
Open: 57.80

52 week low \$41.66 52 week high \$58.50

Market cap \$451.11B
Average volume 31.31M
P/E ratio 28.27
Rev. per Employee \$742,939
EPS 2.05
Dividend 0.36
Div yield 2.48%
Ex dividend date 8/16/16

Compare: Indexes

1d · 5d · 3m · 6m · 1y · 3y · 5y

MarketWatch News on MSFT

How to get started buying stocks
7:30 p.m. Aug. 24, 2016 - Jonathan Burton

5 things Tim Cook has done better than Steve Jobs
4:42 p.m. Aug. 24, 2016 - Jennifer Booton

Apple's stock has gotten a bad rap under Tim Cook
4:34 p.m. Aug. 24, 2016 - Tomi Kilgore

Google shares have rocketed 1,499% since the company's IPO 12 years ago today
12:45 p.m. Aug. 19, 2016 - Philip van Doorn

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
```

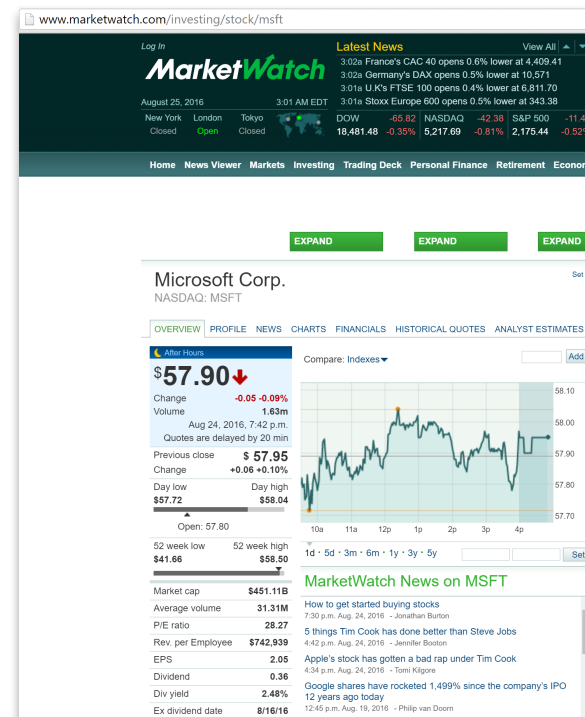


Written by an
expert user

WebRelate

Company	Stock value
MSFT	
AMZN	
AAPL	
T	
S	

+



Program

```

<code>
</code>

```

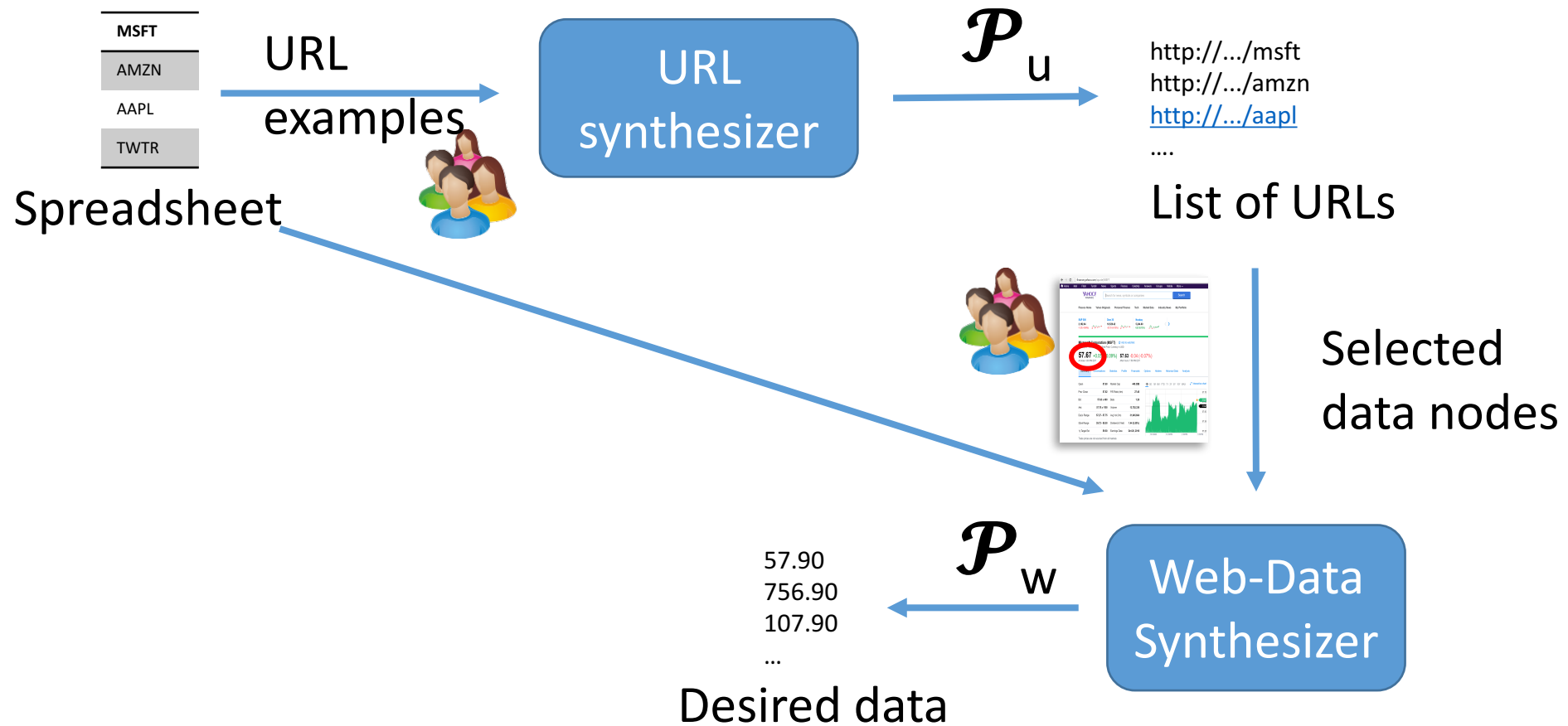


Company	Stock value
MSFT	57.90
AMZN	759.48
AAPL	108.51
T	40.91
S	6.04

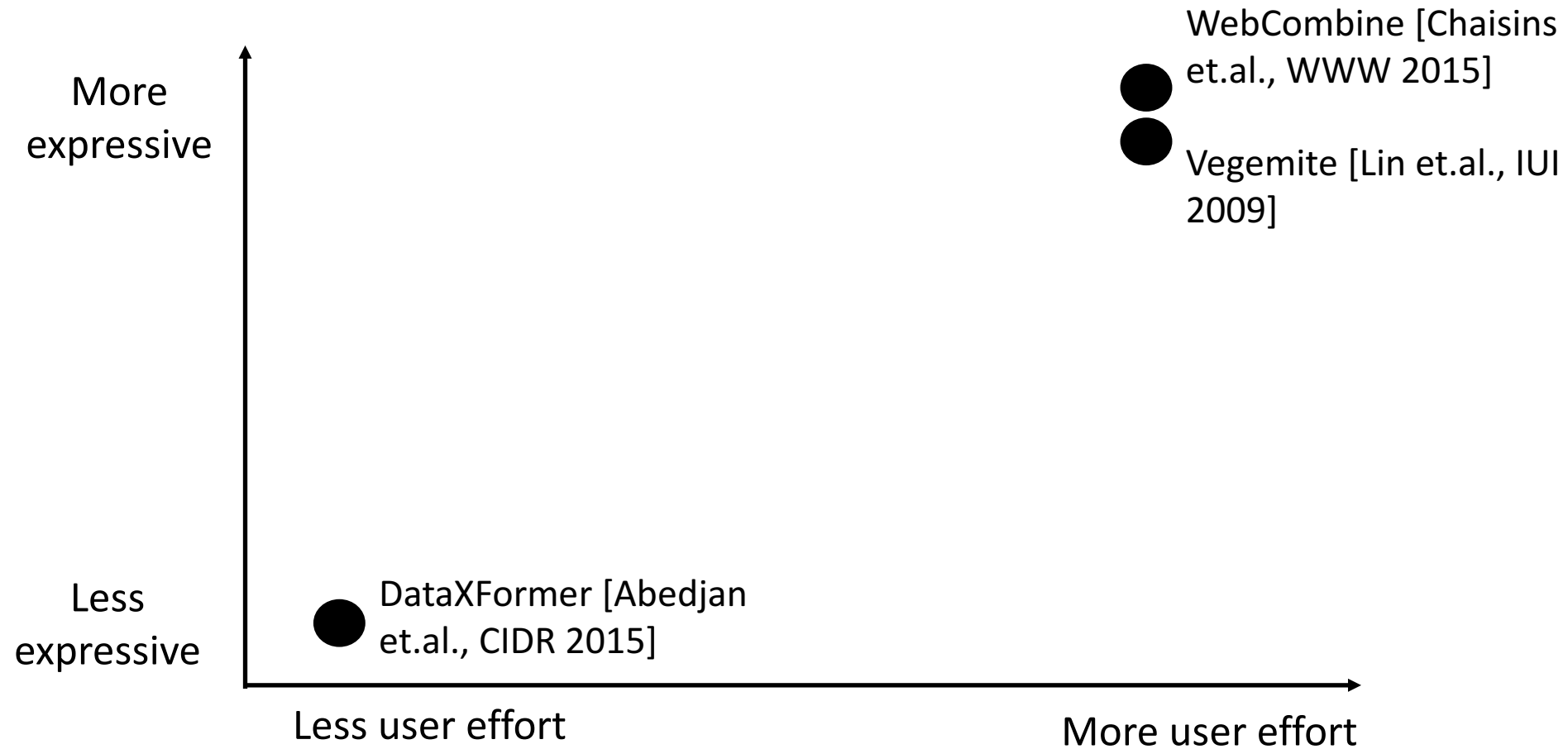
Examples

Demo

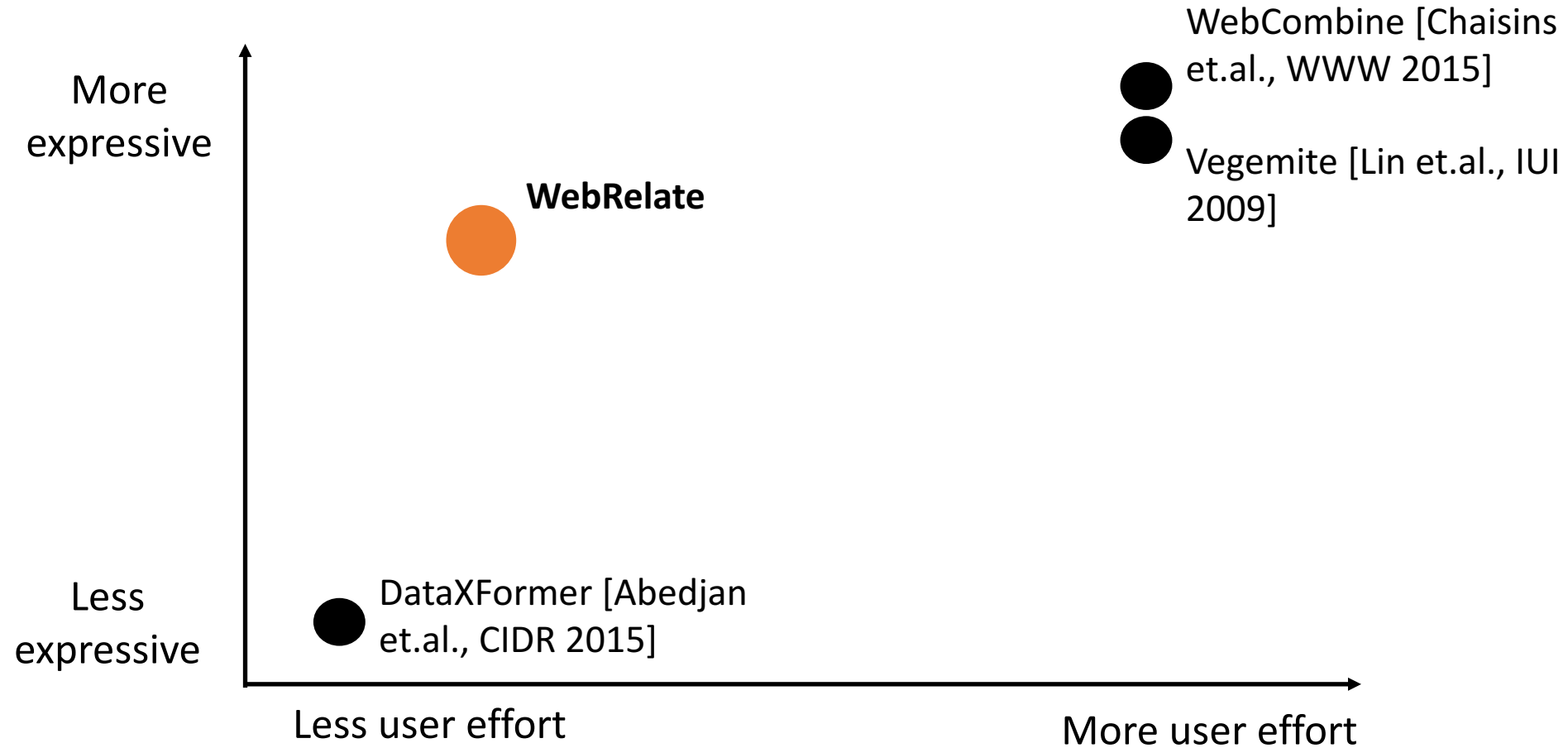
Overview



Related Work



Related Work



Related Work

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

Rishabh Singh  R.S.

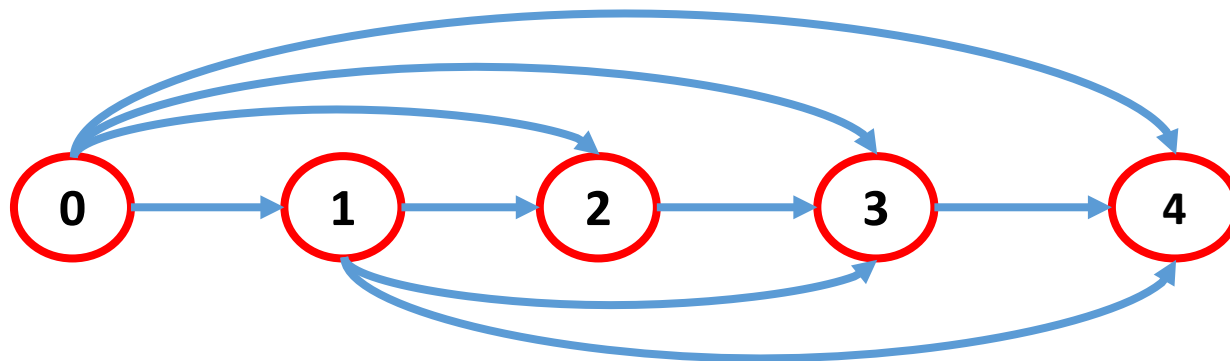
Related Work

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

Rishabh Singh



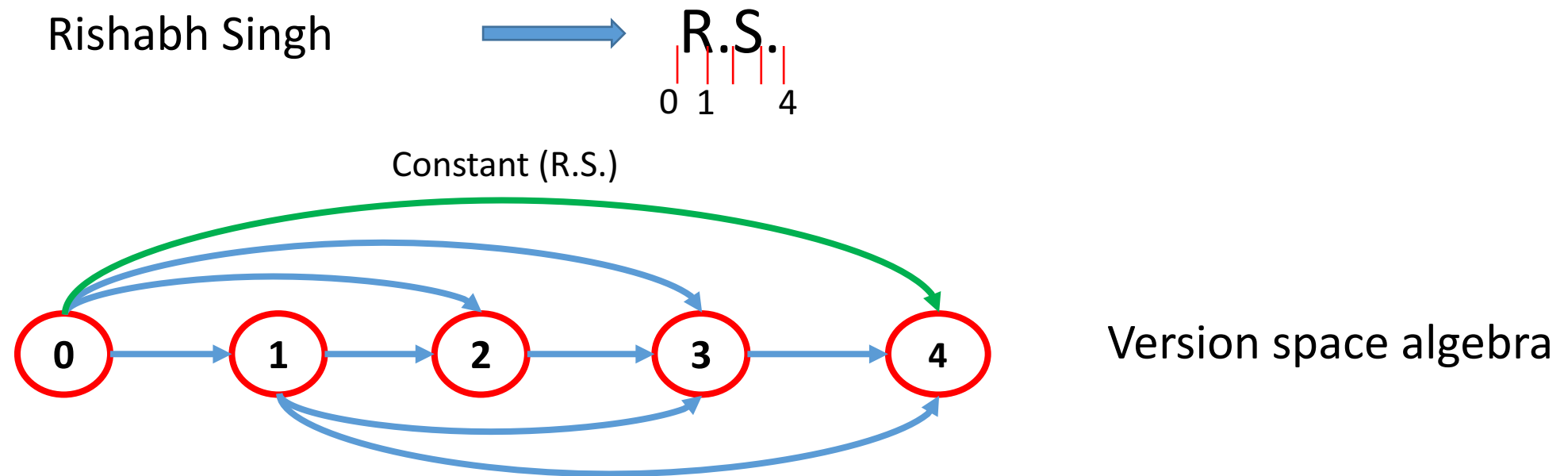
R.S.
0 1 4



Version space algebra

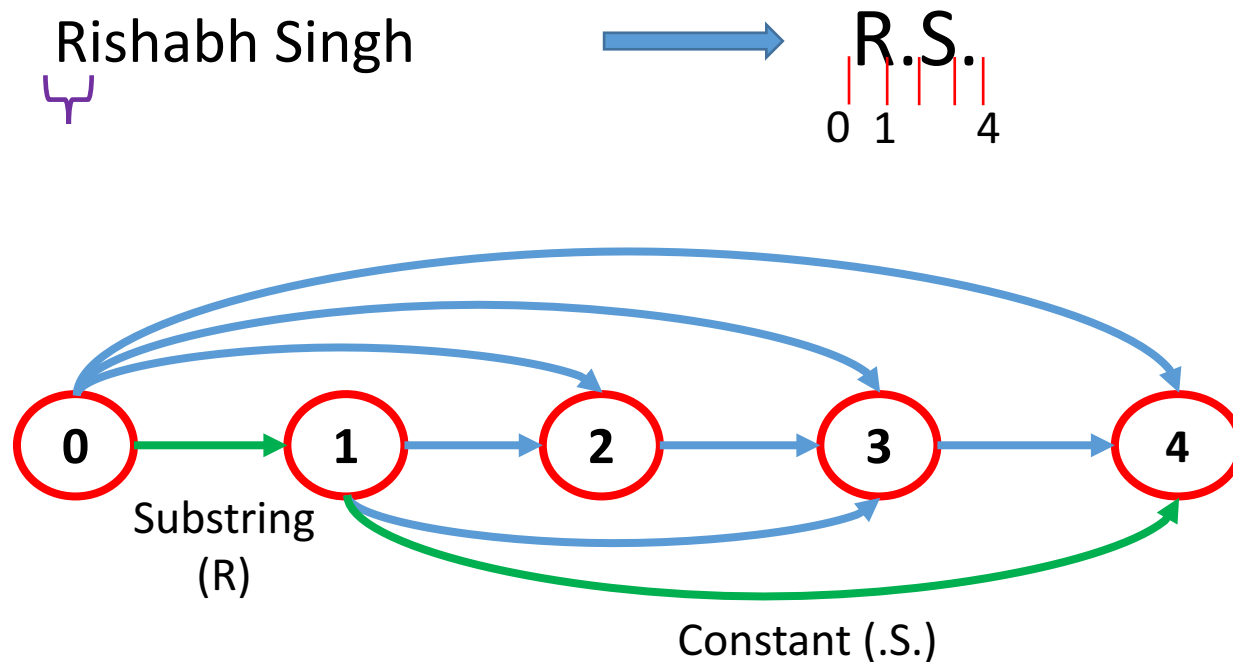
Related Work

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



Related Work

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

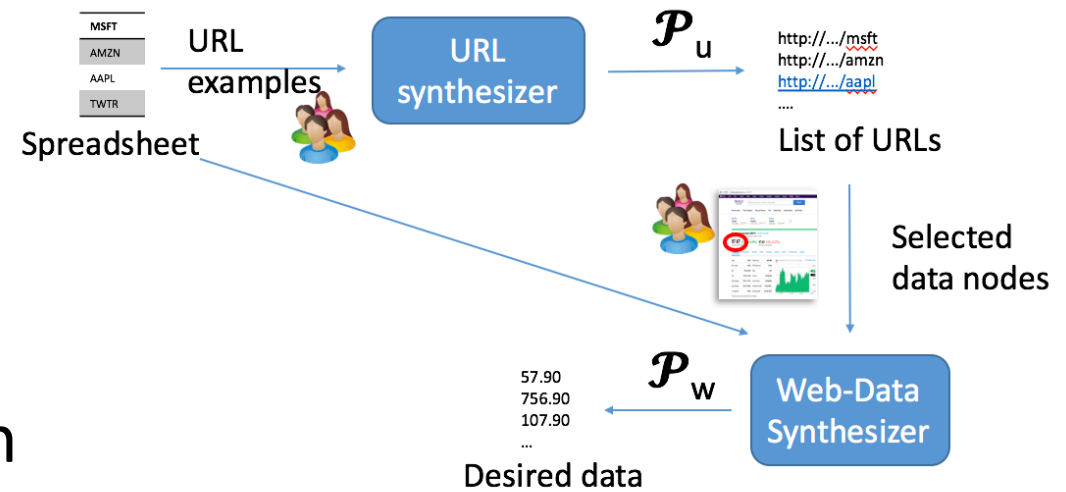


Version space algebra

Ranking: substring > constant

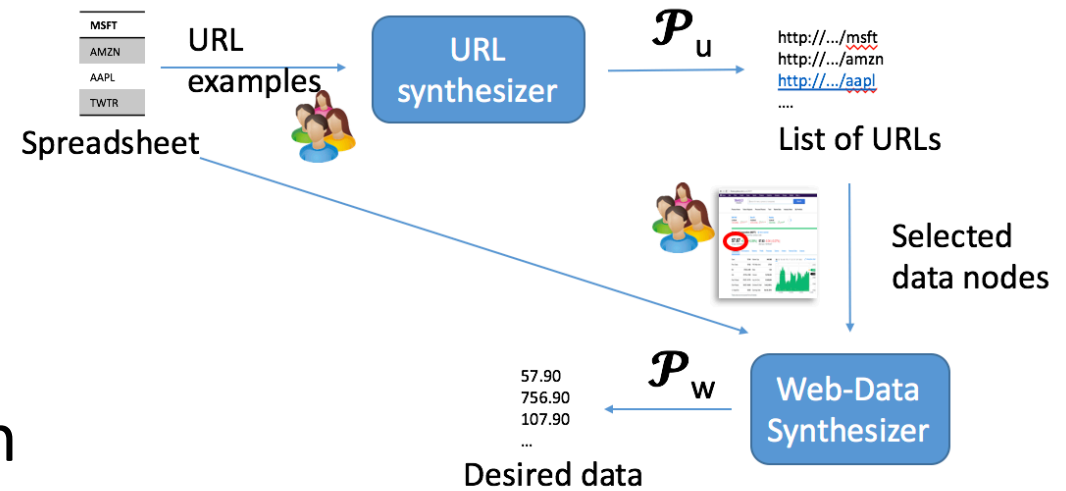
Overview

- 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

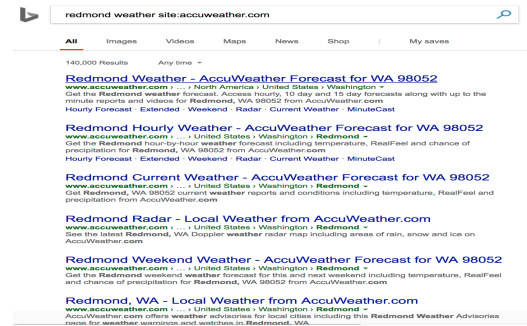


<http://www.accuweather.com/en/us/redmond-wa/98052/weather-forecast/341347>

Output-constrained PBE

Ana Trujillo 357 21th Place
SE,Redmond,WA

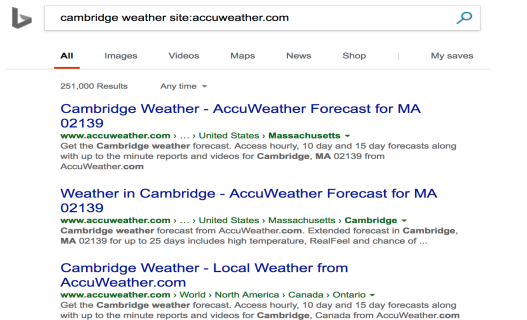
+



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

Charlie Gunaja 732 Memorial
Drive, Cambridge,MA

+



Traditional PBE

Ana Trujillo 357 21th Place
SE,Redmond,WA

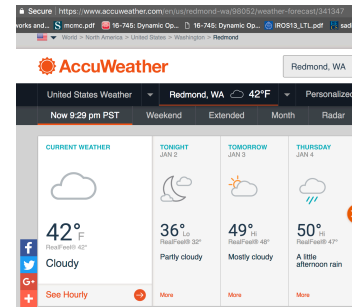


42°

Output-constrained PBE (O-PBE)

Ana Trujillo 357 21th Place
SE,Redmond,WA

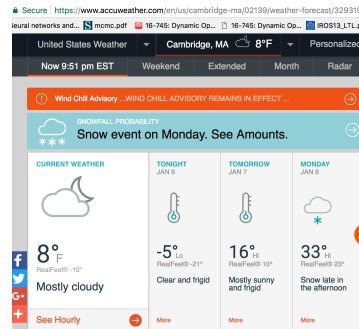
+



42°

Charlie Gunaja 732 Memorial
Drive, Cambridge,MA

+

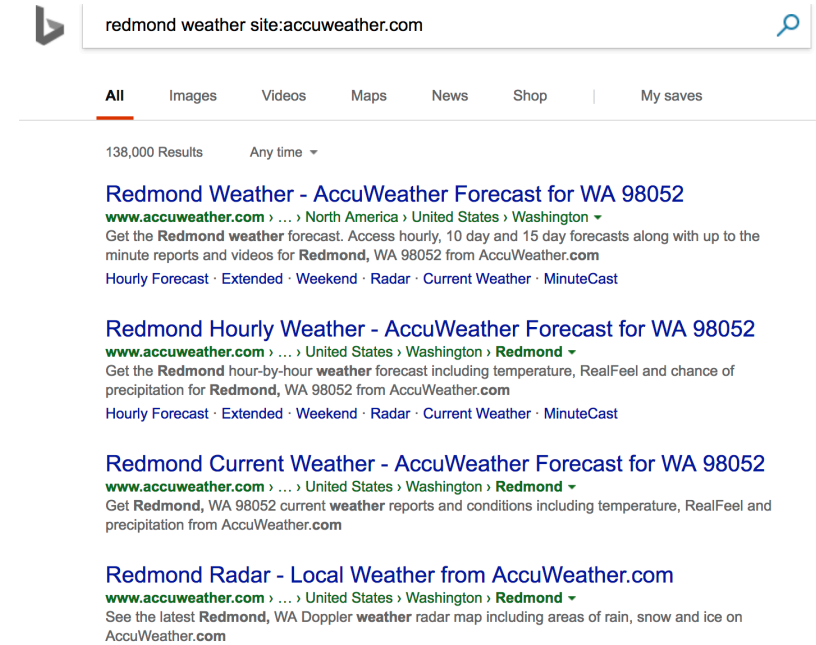


Output-constrained PBE (O-PBE)

- Generalization constraint
 - $\forall in \in \text{inputs. } P(in) \in \text{list of possible of outputs for } in$
- Uniqueness constraint
 - $\forall (in,out) \in \text{input-output examples. } P(in) = out$

Richer class of problems

Ana Trujillo 357 21th Place
SE, Redmond, WA



redmond weather site:accuweather.com

All Images Videos Maps News Shop | My saves

138,000 Results Any time ▾

Redmond Weather - AccuWeather Forecast for WA 98052
[www.accuweather.com](#) > ... > North America > United States > Washington ▾
Get the **Redmond** weather forecast. Access hourly, 10 day and 15 day forecasts along with up to the minute reports and videos for **Redmond**, WA 98052 from AccuWeather.com
Hourly Forecast · Extended · Weekend · Radar · Current Weather · MinuteCast

Redmond Hourly Weather - AccuWeather Forecast for WA 98052
[www.accuweather.com](#) > ... > United States > Washington > **Redmond** ▾
Get the **Redmond** hour-by-hour weather forecast including temperature, RealFeel and chance of precipitation for **Redmond**, WA 98052 from AccuWeather.com
Hourly Forecast · Extended · Weekend · Radar · Current Weather · MinuteCast

Redmond Current Weather - AccuWeather Forecast for WA 98052
[www.accuweather.com](#) > ... > United States > Washington > **Redmond** ▾
Get **Redmond**, WA 98052 current weather reports and conditions including temperature, RealFeel and precipitation from AccuWeather.com

Redmond Radar - Local Weather from AccuWeather.com
[www.accuweather.com](#) > ... > United States > Washington > **Redmond** ▾
See the latest **Redmond**, WA Doppler weather radar map including areas of rain, snow and ice on AccuWeather.com

<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



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

<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>

*
constant *

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



Constant

Constant

*

*

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

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

Better ranking

Ana Trujillo 357 21th Place SE,Redmond,WA



Constant

Constant

*

*



....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



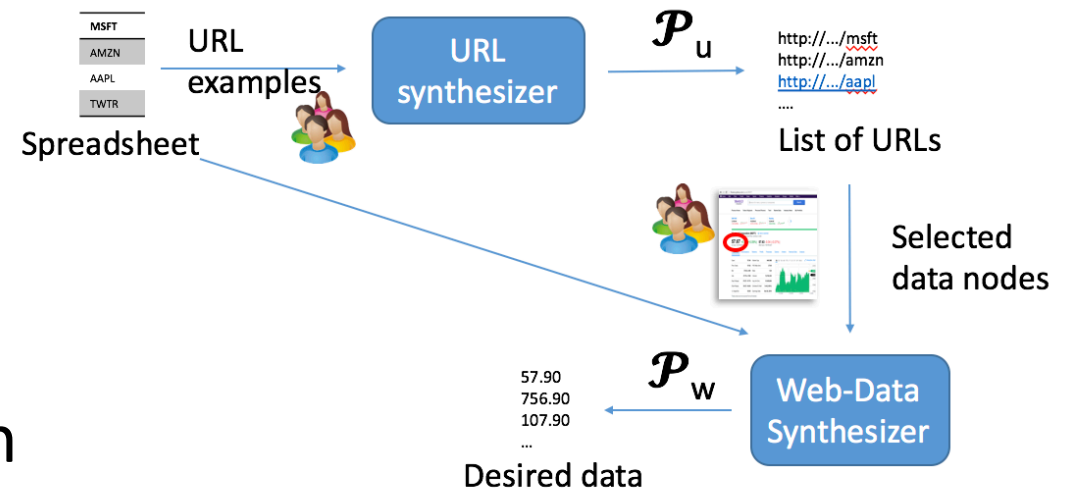
<https://www.accuweather.com/en/us/cambridge-ma/02139/weather-forecast/329319>

<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>

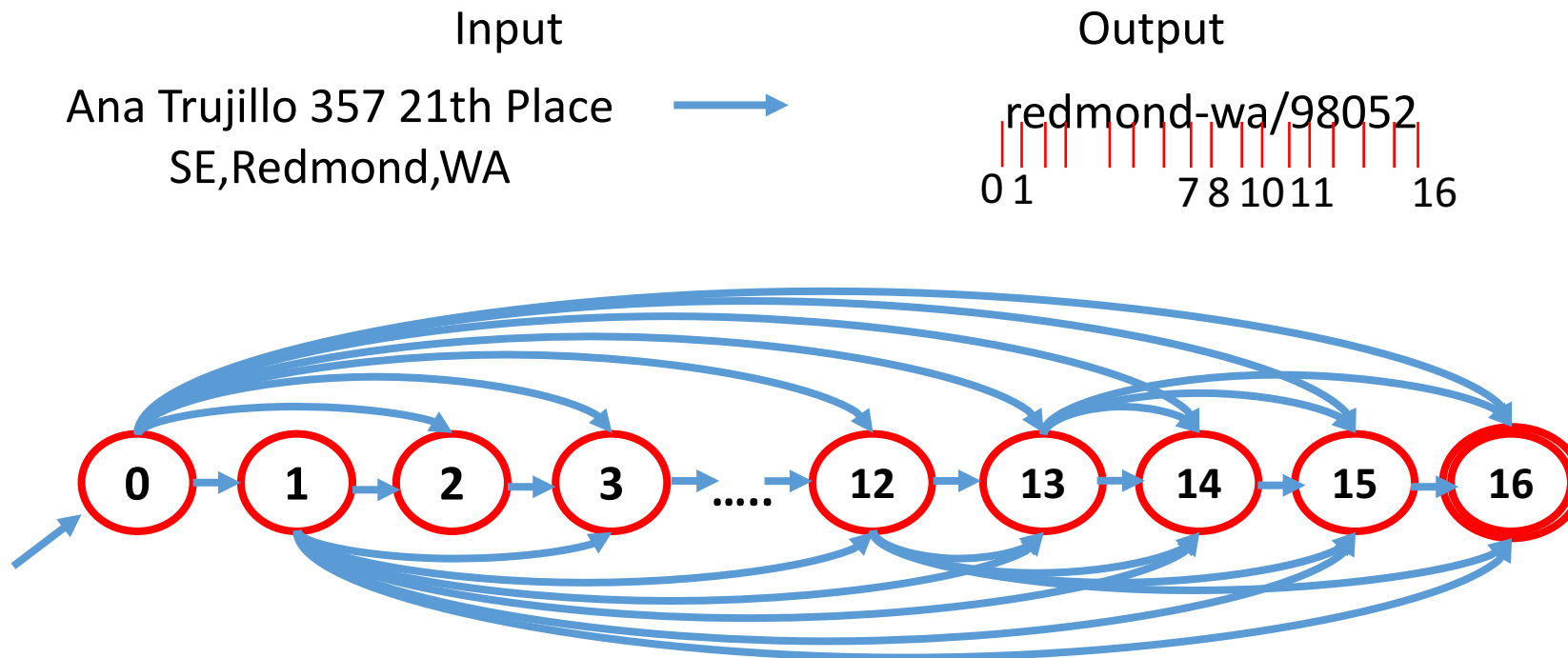
Overview

- 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

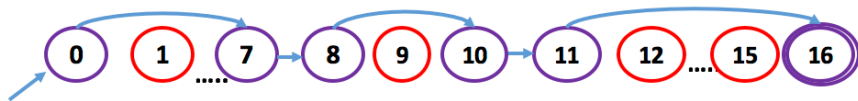
- Performs search over increasingly expressive sub-languages
 - $L_1 \subseteq L_2 \subseteq \dots \subseteq L_k$

Ana Trujillo 357 21th Place
SE,Redmond,WA



redmond-wa/98052
0 1 7 8 10 11 16

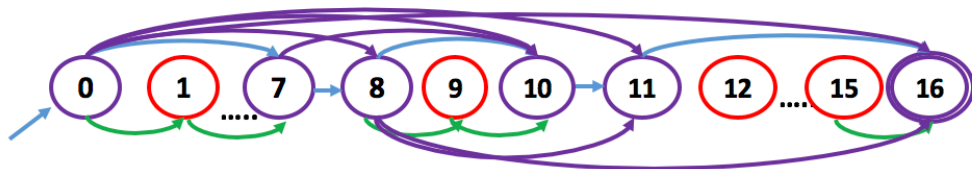
Layer 1



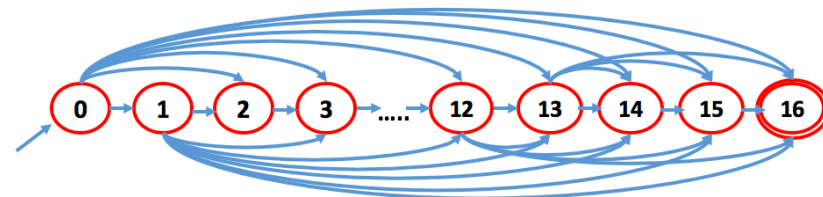
Layer 2



Layer 3

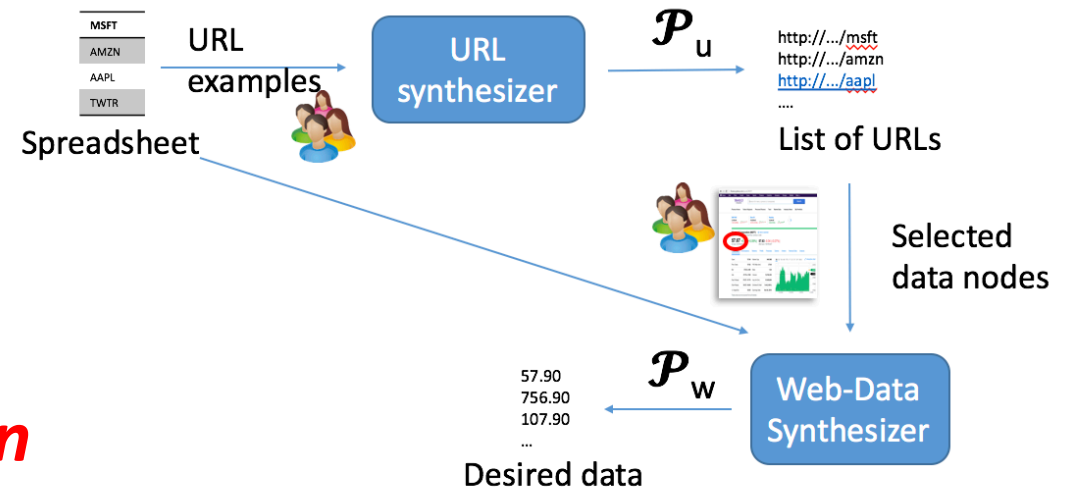


Layer 4 – Full dag

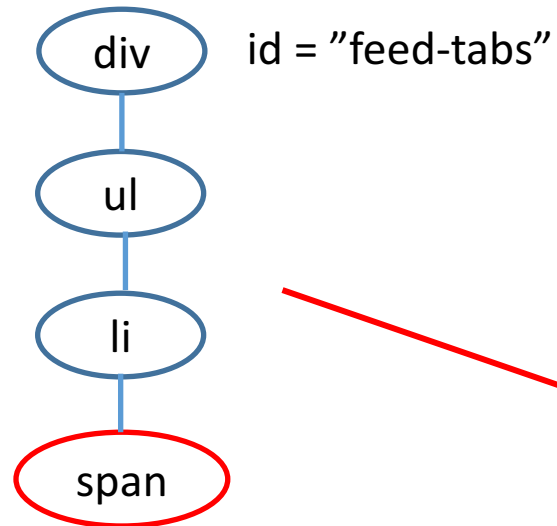


Overview

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



Web-Extraction



Can be expressed in Xpath query language

Secure | https://www.accuweather.com/en/us/redmond-wa/98052/weather-forecast/34134

works and... S mcmc.pdf 16-745: Dynamic Op... 16-745: Dynamic Op... IROS13_LTL.pdf









World > North America > United States > Washington > Redmond

AccuWeather

Redmond, WA

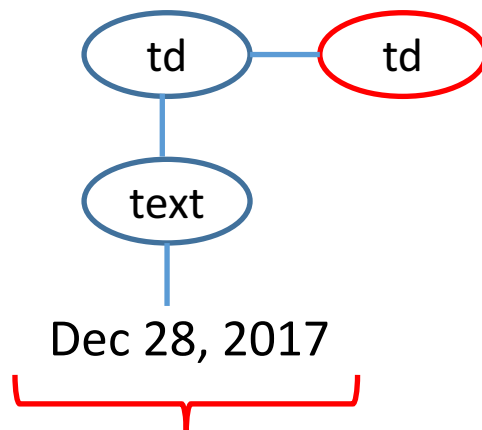
United States Weather Redmond, WA 42°F Personal

Now 9:29 pm PST Weekend Extended Month Radar

CURRENT WEATHER	TONIGHT JAN 2	TOMORROW JAN 3	THURSDAY JAN 4
 42° F RealFeel® 42° Cloudy	 36° Lo RealFeel® 32° Partly cloudy	 49° Hi RealFeel® 48° Mostly cloudy	 50° Hi RealFeel® 47° A little afternoon rain
   	See Hourly	More	More

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 AAF

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

Investing.com

EUR/USD or AAF

USD/INR 63.275 -0.095 (-0.15%)

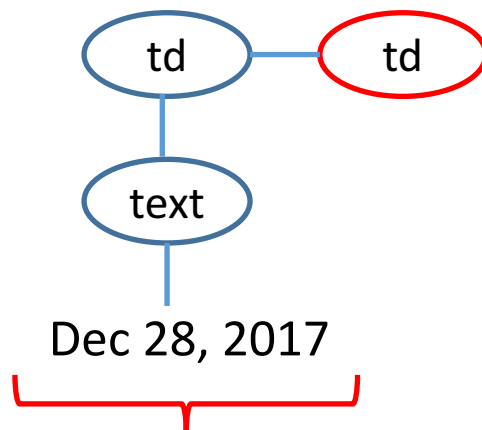
Cur1	Cur2	Date
USD	INR	28, December, 2017
EUR	GBP	03, January, 2018
USD	CHF	05, January, 2018

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

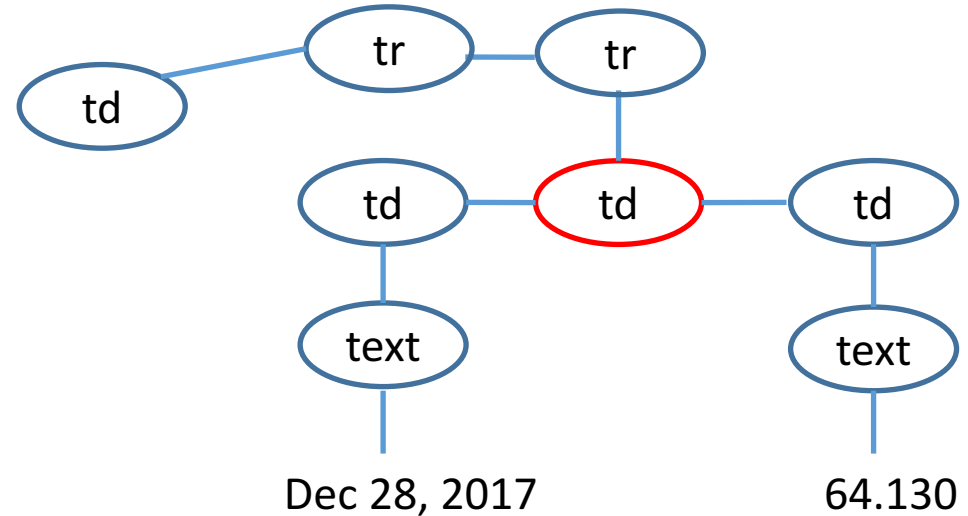


Transform(28, December, 2017)

Input-dependent Web-Extraction

USD; INR; 28, December, 2017

Input



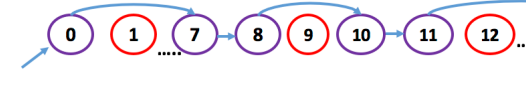
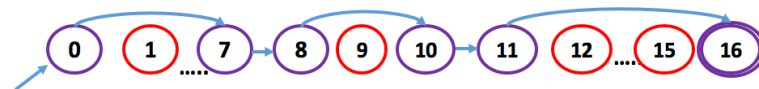
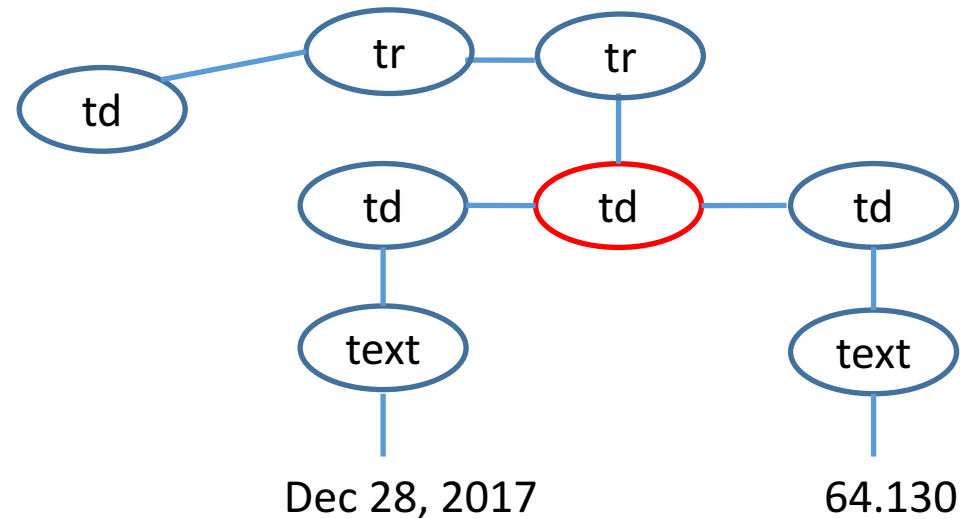
HTML tree

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



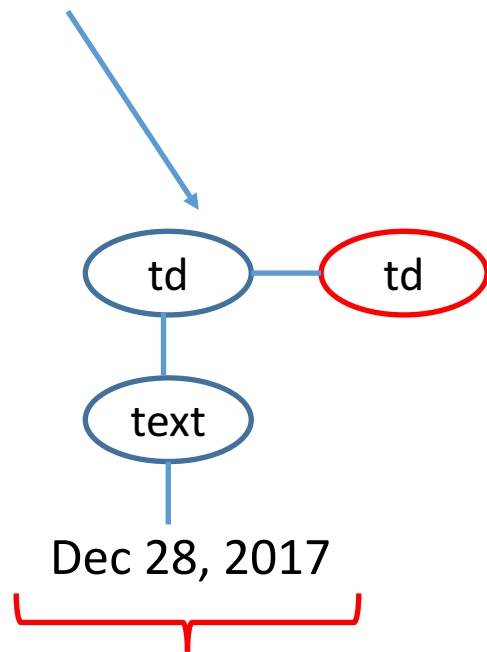
HTML tree

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

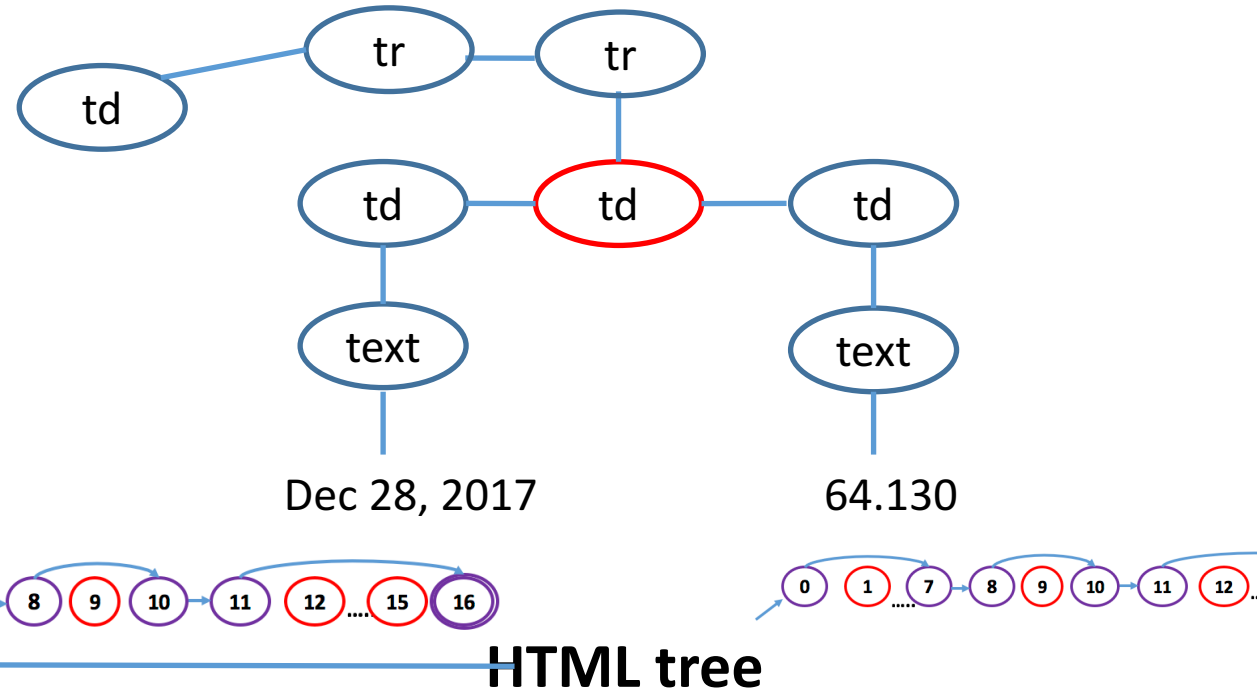
Input-dependent Web-Extraction

USD; INR; 28, December, 2017

Input



Transform(28, December, 2017)

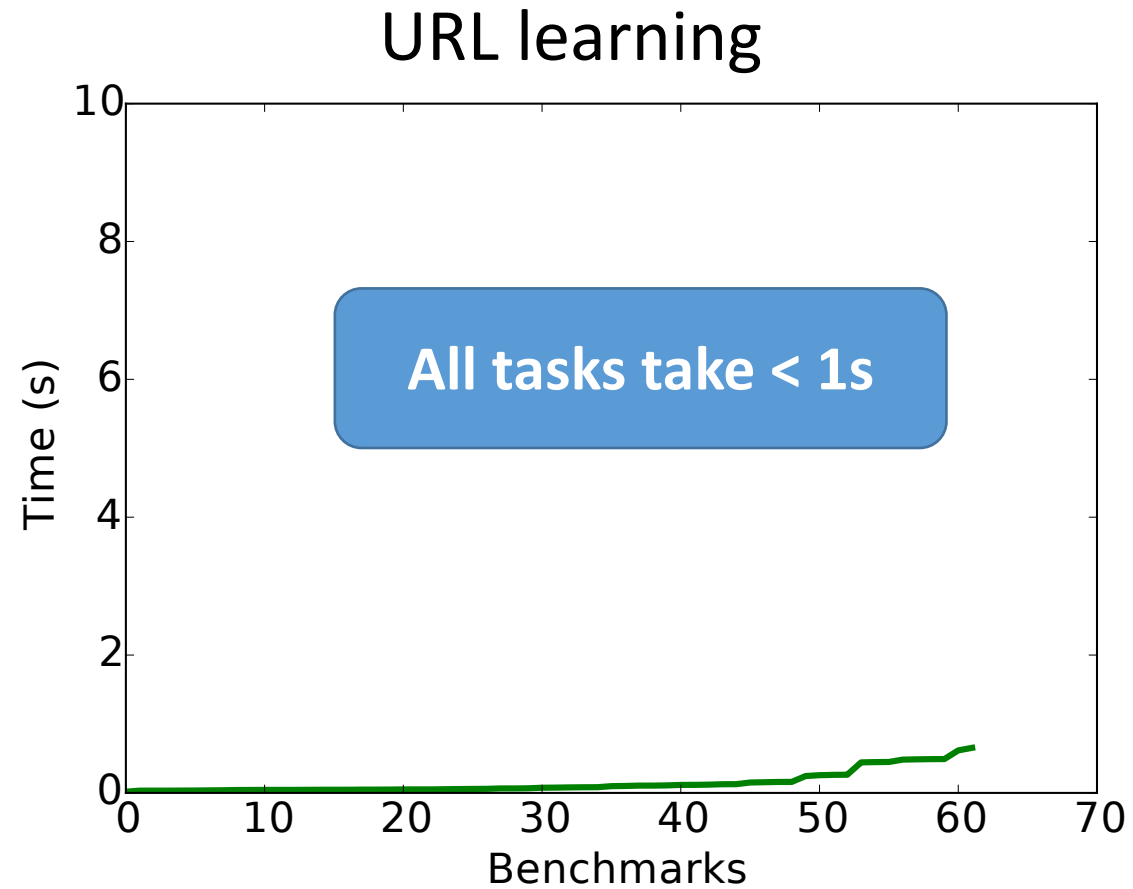


Results

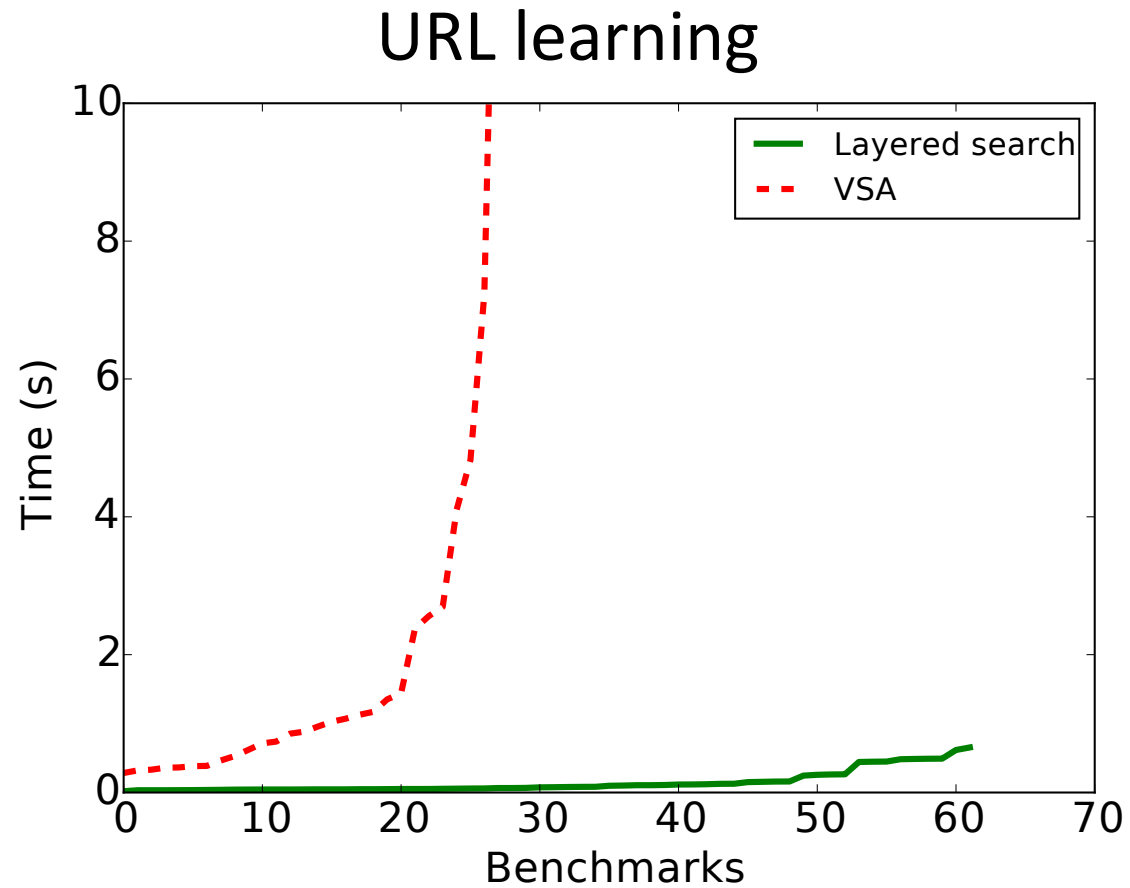
Expressive?

- 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?

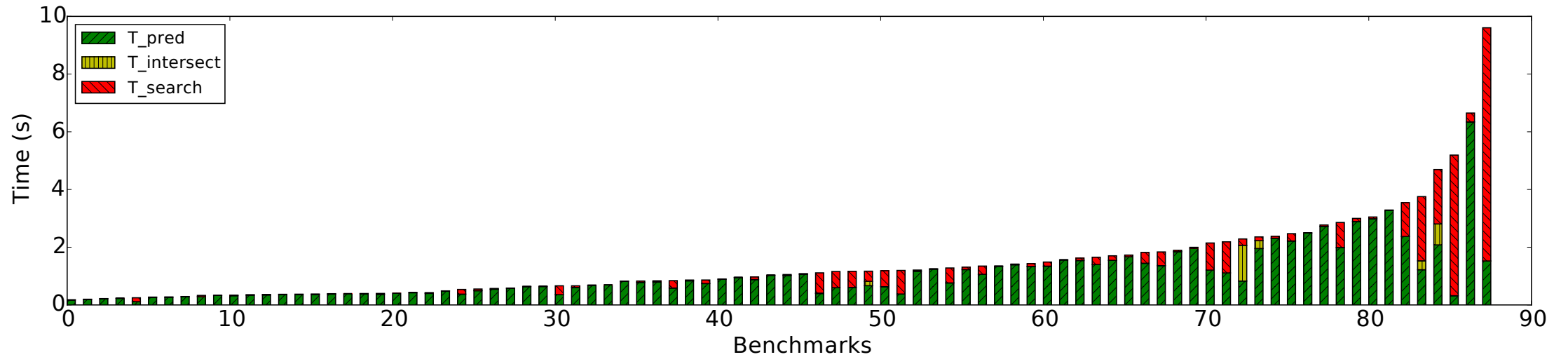


Layered search beats VSA



Fast?

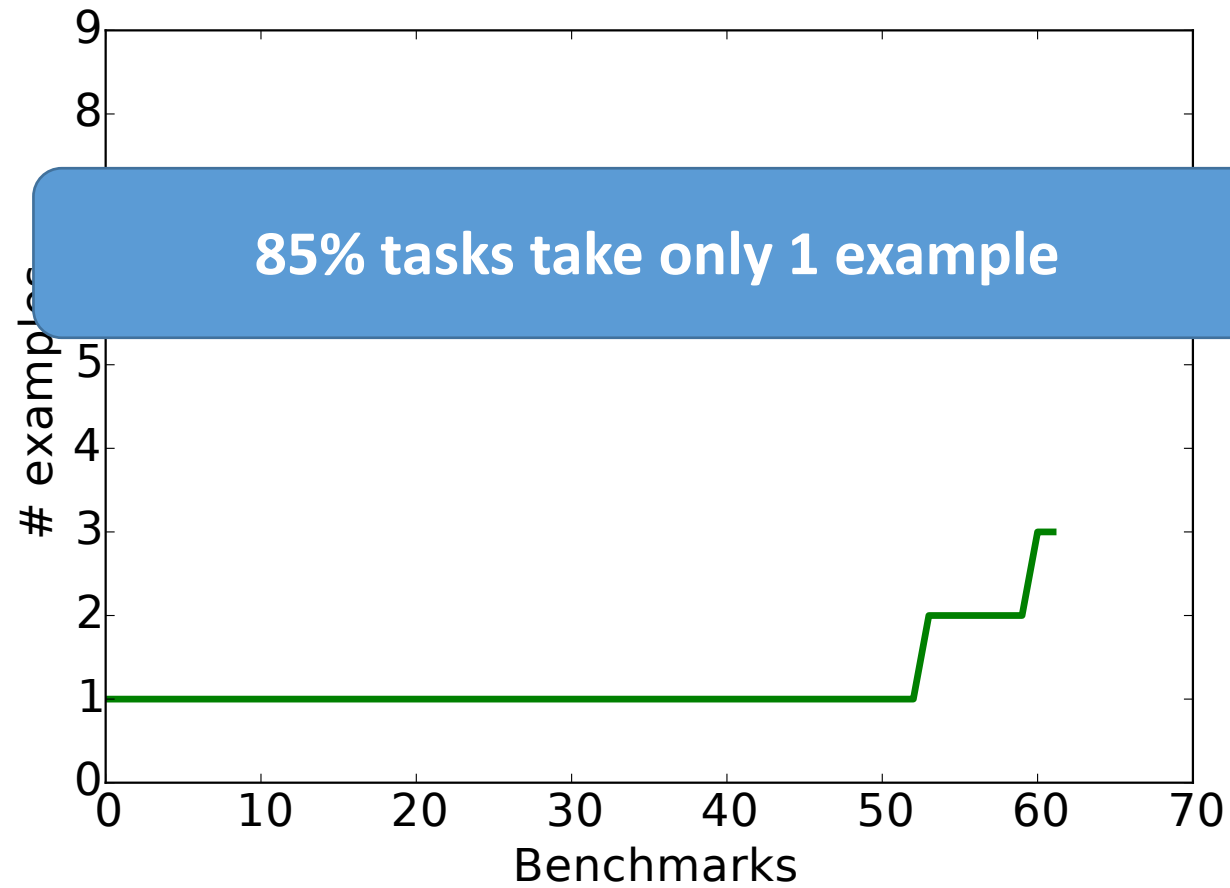
Web-extraction learning



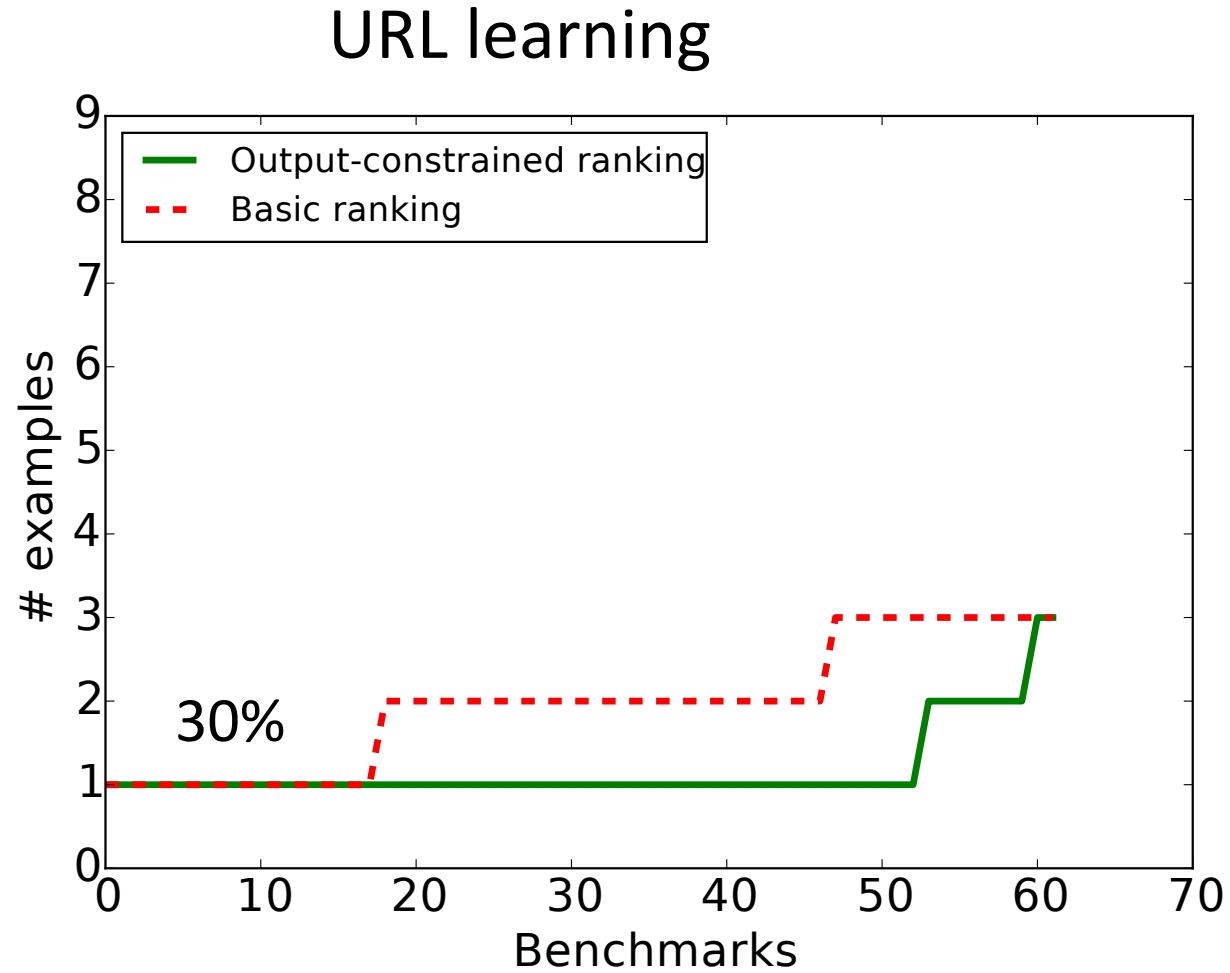
All tasks take < 10s

Easy to use?

URL learning

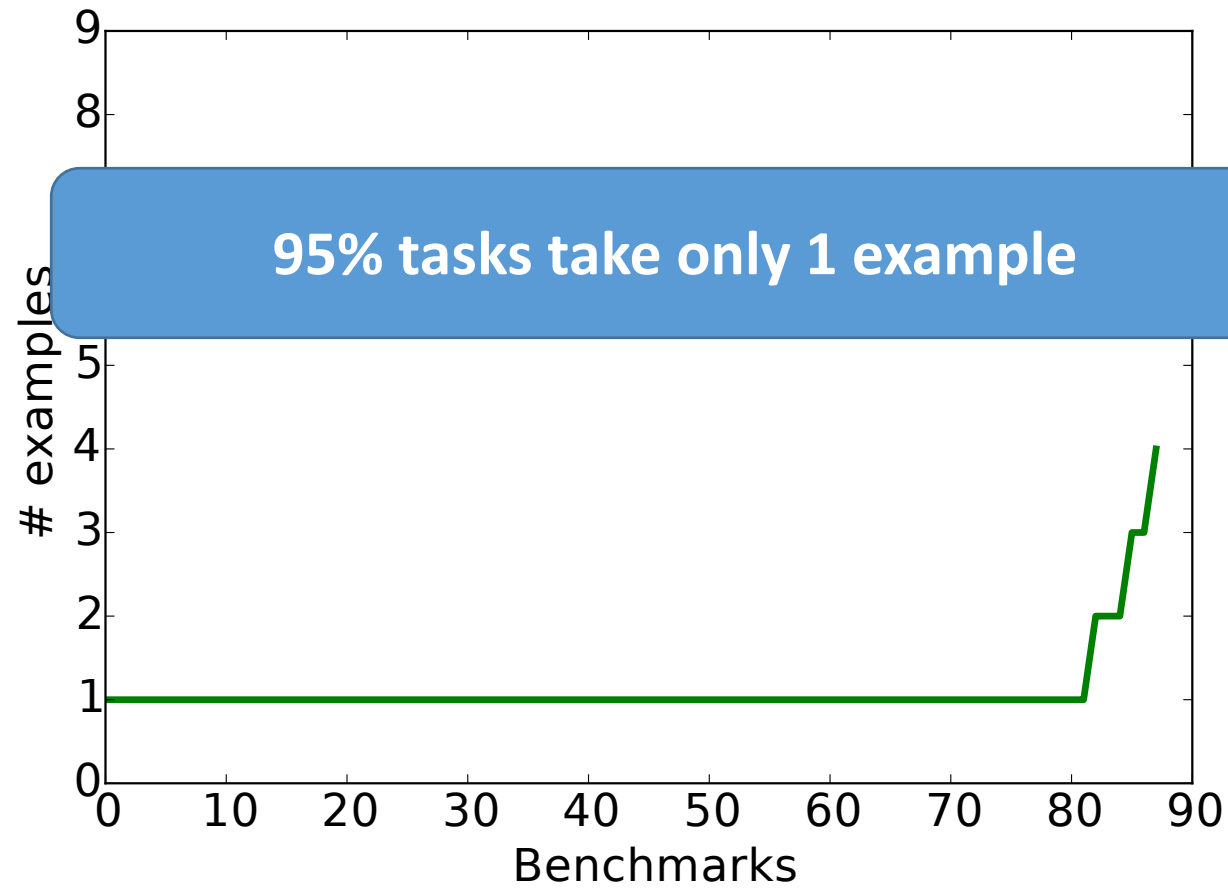


Impact of generalization constraint



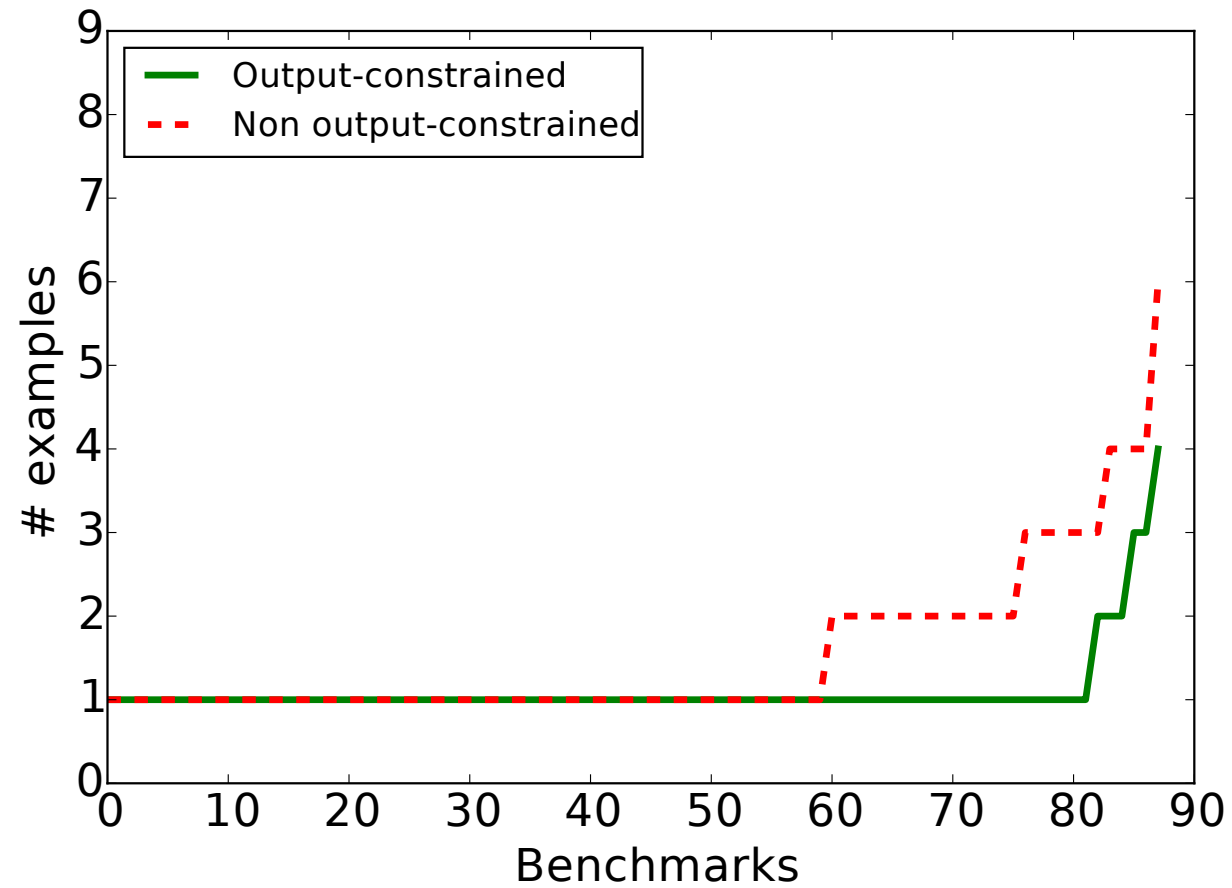
Easy to use?

Web-extraction learning



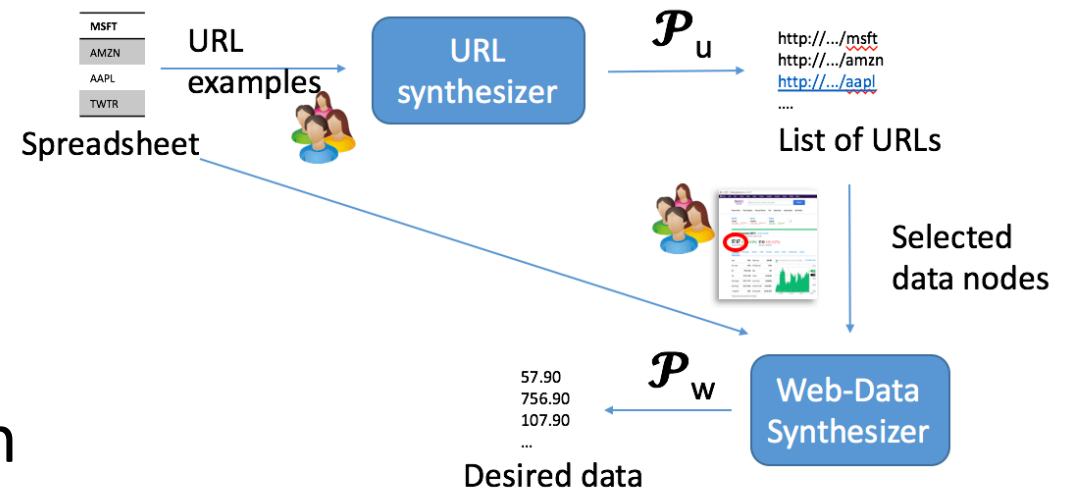
Impact of generalization constraint

Web-extraction learning



Summary

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



Thank You!

jinala@mit.edu