Lecture 17
Data Extraction and Wrangling

Nadia Polikarpova

Logistics

Project presentations

- Monday Dec 10, 3-6pm; **in this room**
- 20 min per team (15 min presentation + questions)
- Structure: motivation, **demo**, technique, evaluation

Project reports

- Due on Dec 14 (start working on them now!)
- Format: see course organization page (3-5 pages, SIGPLAN format)

Applications of synthesis

Superoptimization

Custom data structures

→ Data extraction and data wrangling

Cryptographic implementations

SQL queries

FlashExtract

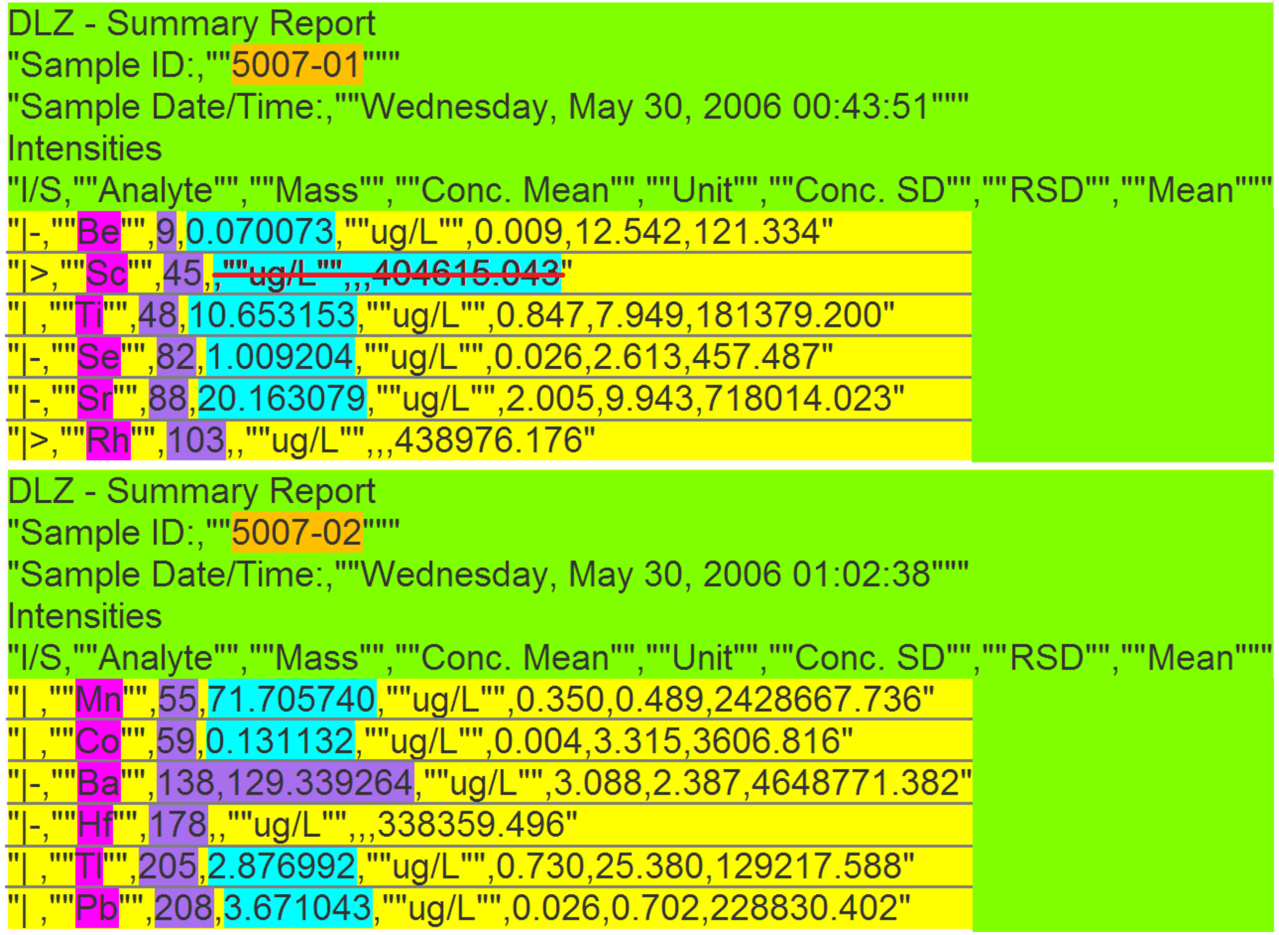
[Le, Gulwani. PLDI‘14]

Problem: extract data from semi-structured sources (e.g. log file) into a list of records

User input:

- output schema
- highlights examples of fields

Search strategy: VSA



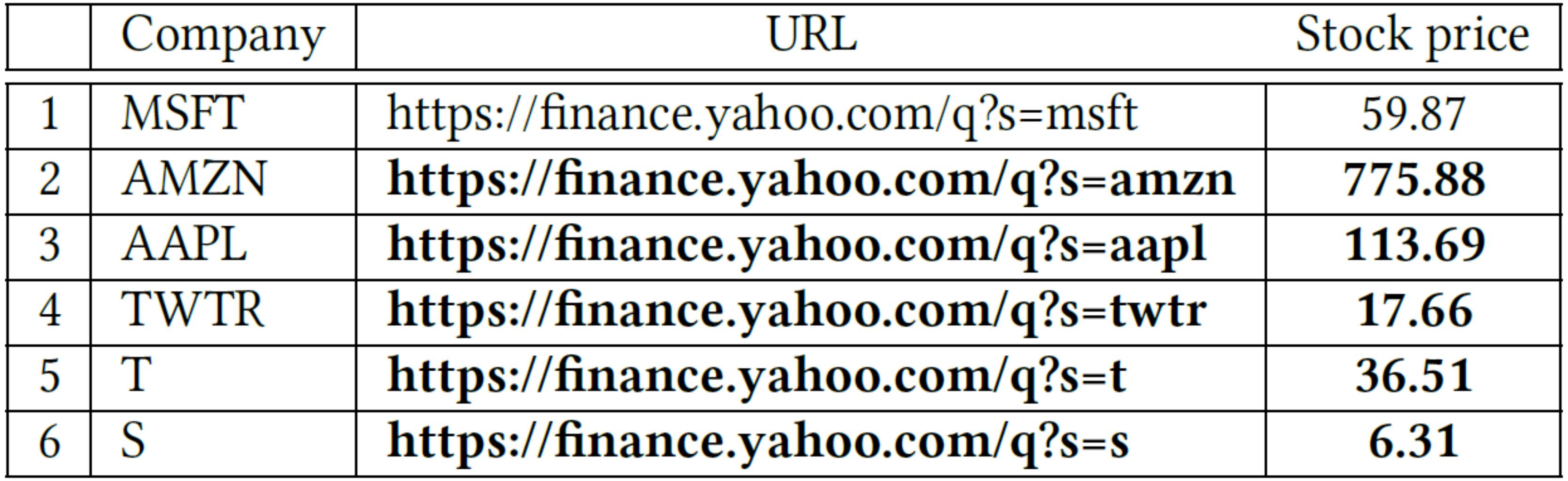
```
DLZ - Summary Report
"Sample ID:,""5007-01"""
"Sample Date/Time:,""Wednesday, May 30, 2006 00:43:51"""
Intensities
"I/S,""Analyte"",""Mass"",""Conc. Mean"",""Unit"",""Conc. SD"",""RSD"",""Mean"""
"|-,""Be"",9,0.070073,""ug/L"",0.009,12.542,121.334"
"|>,""Sc"",45,~~,""ug/L"",,,404615.043~~"
"| ,""Ti"",48,10.653153,""ug/L"",0.847,7.949,181379.200"
"|-,""Se"",82,1.009204,""ug/L"",0.026,2.613,457.487"
"|-,""Sr"",88,20.163079,""ug/L"",2.005,9.943,718014.023"
"|>,""Rh"",103,,""ug/L"",,,438976.176"
DLZ - Summary Report
"Sample ID:,""5007-02"""
"Sample Date/Time:,""Wednesday, May 30, 2006 01:02:38"""
Intensities
"I/S,""Analyte"",""Mass"",""Conc. Mean"",""Unit"",""Conc. SD"",""RSD"",""Mean"""
"| ,""Mn"",55,71.705740,""ug/L"",0.350,0.489,2428667.736"
"| ,""Co"",59,0.131132,""ug/L"",0.004,3.315,3606.816"
"|-,""Ba"",138,129.339264,""ug/L"",3.088,2.387,4648771.382"
"|-,""Hf"",178,,""ug/L"",,,338359.496"
"| ,""Tl"",205,2.876992,""ug/L"",0.730,25.380,129217.588"
"| ,""Pb"",208,3.671043,""ug/L"",0.026,0.702,228830.402"
```

WebRelate

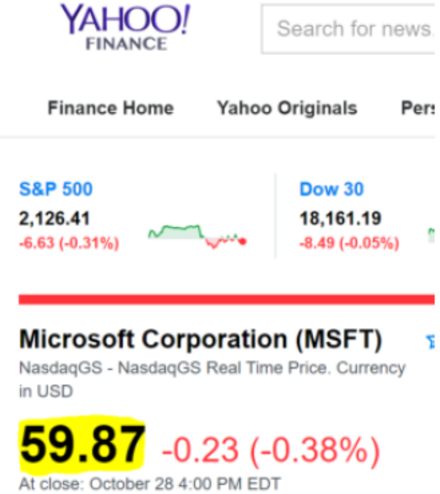
[Inala, Singh. POPL'17]

Problem: extract data from web pages into spreadsheets

User input: navigate to a webpage and select content



	Company	URL	Stock price
1	MSFT	https://finance.yahoo.com/q?s=msft	59.87
2	AMZN	**https://finance.yahoo.com/q?s=amzn**	**775.88**
3	AAPL	**https://finance.yahoo.com/q?s=aapl**	**113.69**
4	TWTR	**https://finance.yahoo.com/q?s=twtr**	**17.66**
5	T	**https://finance.yahoo.com/q?s=t**	**36.51**
6	S	**https://finance.yahoo.com/q?s=s**	**6.31**



S&P 500
2,126.41
-6.63 (-0.31%)

Dow 30
18,161.19
-8.49 (-0.05%)

Microsoft Corporation (MSFT)

NasdaqGS - NasdaqGS Real Time Price. Currency in USD

59.87 -0.23 (-0.38%)

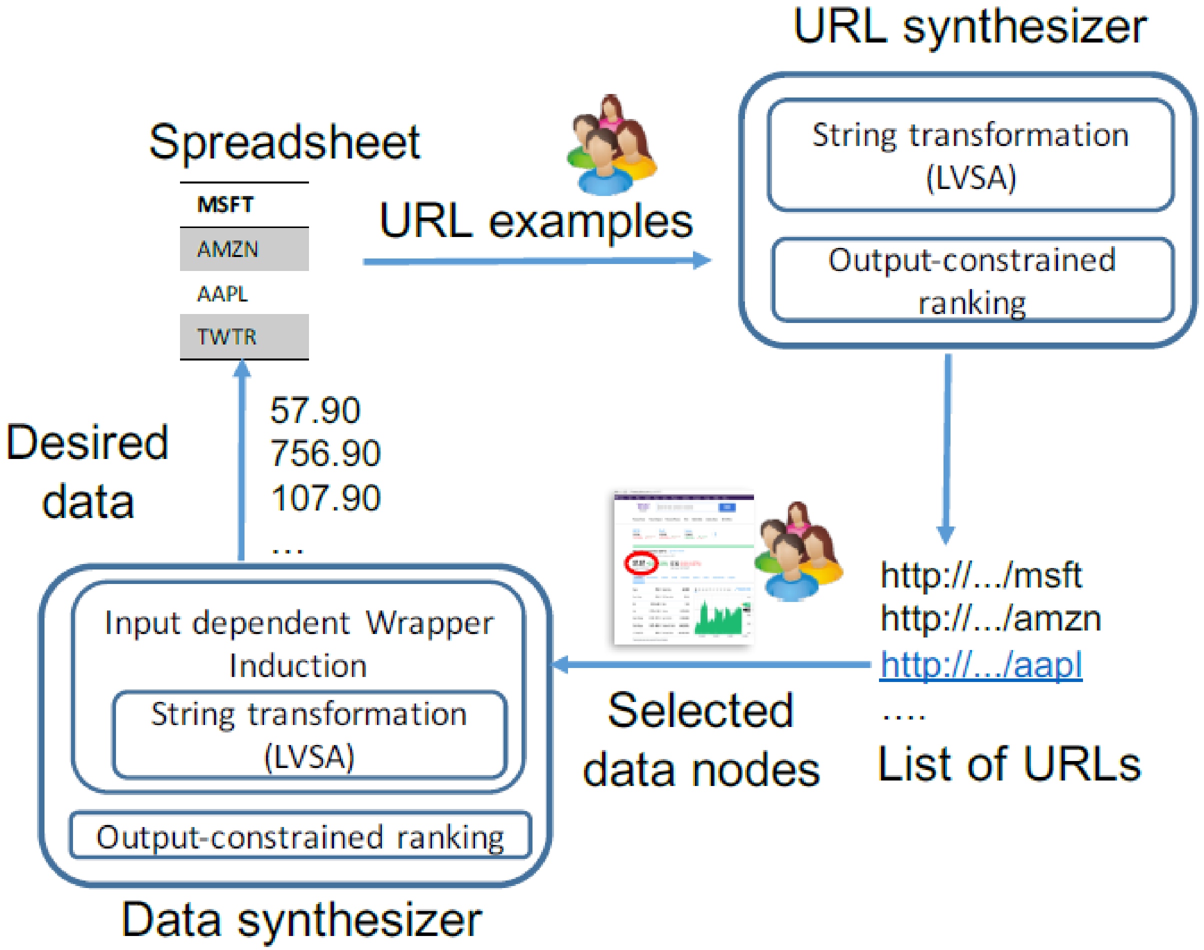
At close: October 28 4:00 PM EDT

WebRelate

Search strategy: VSA

Optimizations:

- Layered VSA (URLs are too long for FlashFill-style VSAs)
- Output-constrained synthesis: we know the space of possible outputs



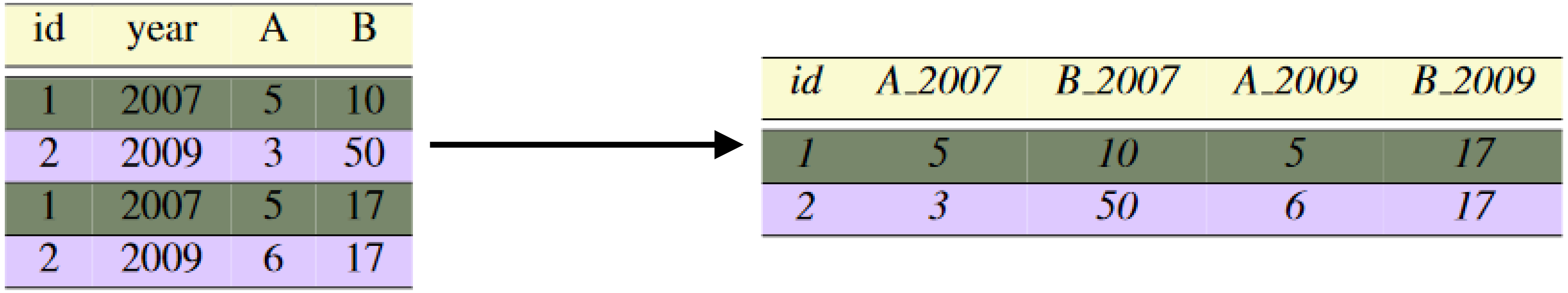
Morpheus

[Feng et al. PLDI‘17]

Problem: table data wrangling

User input: input-output examples (small tables)

Search strategy: enumerative search with deduction

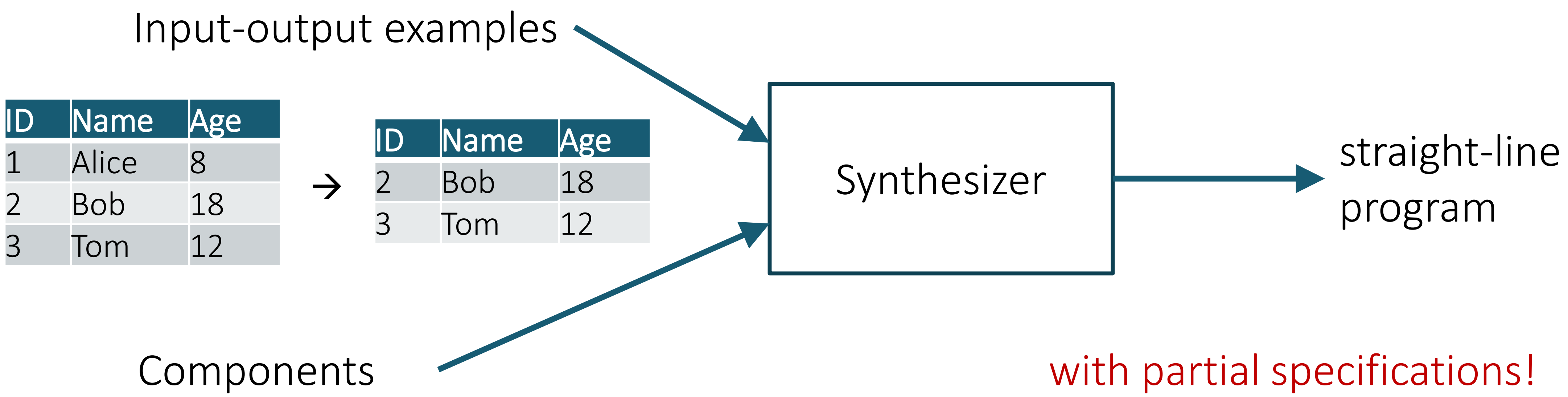


id	year	A	B
1	2007	5	10
2	2009	3	50
1	2007	5	17
2	2009	6	17

id	*A_2007*	*B_2007*	*A_2009*	*B_2009*
1	*5*	*10*	*5*	*17*
2	*3*	*50*	*6*	*17*

Morpheus: TDP with deduction

[Feng et al‘17]



ID	Name	Age
1	Alice	8
2	Bob	18
3	Tom	12

→

ID	Name	Age
2	Bob	18
3	Tom	12

```
select : Table → [Col] → Table

filter : Table → (Row → Bool) → Table
```

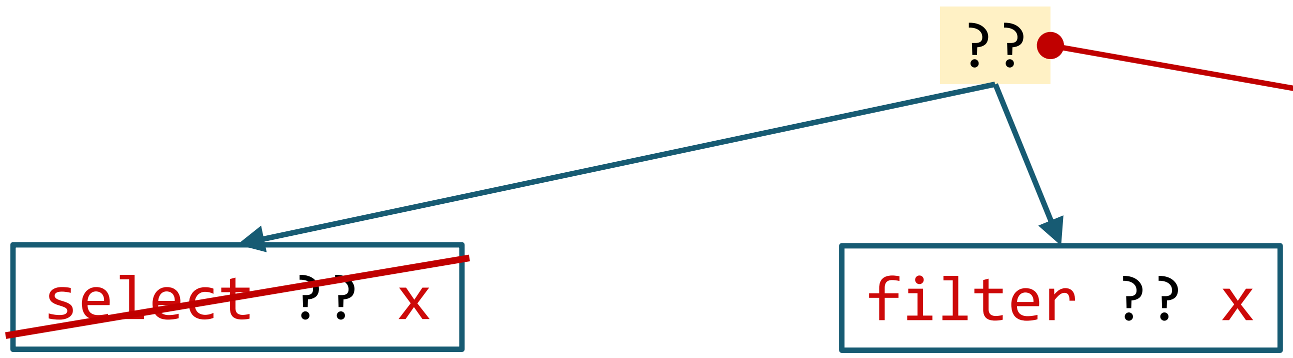
with partial specifications!

```
out.rows = in.rows
&& out.cols < in.cols

our.rows < in.rows
&& out.cols = in.cols
```

Morpheus: TDP with deduction

[Feng et al‘17]



x

ID	Name	Age
1	Alice	8
2	Bob	18
3	Tom	12

→

y

ID	Name	Age
2	Bob	18
3	Tom	12

$$\exists x\ y{:}\ y.\mathrm{rows} = x.\mathrm{rows} \land y.\mathrm{cols} < x.\mathrm{cols} \land x.\mathrm{rows} = 3 \land x.\mathrm{cols} = 3 \land y.\mathrm{rows} = 2 \land y.\mathrm{cols} = 3$$

SMT

UNSAT

$$\exists x\ y{:}\ y.\mathrm{rows} < x.\mathrm{rows} \land y.\mathrm{cols} = x.\mathrm{cols} \land x.\mathrm{rows} = 3 \land x.\mathrm{cols} = 3 \land y.\mathrm{rows} = 2 \land y.\mathrm{cols} = 3$$

SMT

SAT

```
select : Table → [Col] → Table
filter : Table → (Row → Bool) → Table
```

```
out.rows = in.rows && out.cols < in.cols
our.rows < in.rows && out.cols = in.cols
```